FMDB *Transactions on Sustainable Computer Letters*

A Comparative Evaluation of Patch-Wise vs. Pixel-Wise Labeling Methods in Semantic Segmentation

Saw Mya Nandar[1,*]

[1]Department of Computer Systems and Technologies, University of Computer Studies, Yangon, Yangon Region, Myanmar.
sawmyananda2025@gmail.com[1]

Abstract: The field of computer vision, semantic segmentation is a fundamental problem that necessitates the precise assignment of semantic labels to each pixel in an image. Despite the fact that pixel-wise labelling has been considered the gold standard due to the fine-grained resolution it offers, it is extremely expensive in terms of the annotation and processing resources it requires. The patch-wise labeling approach has emerged as a potentially useful compromise between the efficiency of annotation and the accuracy of segmentation. The purpose of this study is to provide a comprehensive comparison analysis of patch-wise and pixel-wise labelling strategies for semantic segmentation across various datasets and architectures. We investigate the trade-offs that exist between characteristics such as label granularity, computational expense, model performance, and the ability to generalise. The most cutting-edge segmentation networks, including U-Net, DeepLabV3+, and Swin Transformer, are utilised in experiments carried out on benchmark datasets such as Cityscapes and PASCAL VOC. Our findings reveal the conditions under which patch-wise labelling can serve as a powerful substitute for pixel-wise approaches in situations where supervision is weak or resources are limited. Regarding annotation, model architecture, and the implementation of segmentation systems in the real world, the paper considers the ramifications.

Keywords: Patch-Wise and Pixel-Wise; Semantic Segmentation; PASCAL and VOC; Segmentation Systems; Deep Learning; Annotation Process; Boundary Precision.

Received on: 12/10/2024, **Revised on:** 15/12/2024, **Accepted on:** 22/01/2025, **Published on:** 03/06/2025

Journal Homepage: https://www.fmdbpub.com/user/journals/details/FTSCL

DOI: https://doi.org/10.69888/FTSCL.2025.000425

Cite as: S.M. Nandar, "A Comparative Evaluation of Patch-Wise vs. Pixel-Wise Labeling Methods in Semantic Segmentation," *FMDB Transactions on Sustainable Computer Letters*, vol. 3, no. 2, pp. 97–104, 2025.

1. Introduction

Semantic segmentation, a core task in computer vision, assigns a categorical label to each pixel in an image. It has a wide range of applications, including autonomous driving, medical imaging, satellite imagery, and robotics. While deep learning has dramatically improved segmentation performance, the underlying annotation process remains one of the most tedious and time-consuming steps in dataset creation. High-quality pixel-wise annotations require extensive manual labour, large-scale supervised model training, and bottlenecks. Semantic segmentation research has traditionally focused on improving model architectures, loss functions, and optimisation techniques, with less emphasis on the nature and quantity of annotations. Pixel-perfect labelling, although precise, requires meticulous hand-drawing of object boundaries, which is infeasible for large datasets. For instance, annotating a single high-resolution Cityscapes image can take more than 90 minutes for an expert

[*]Corresponding author.

labeller. Such annotation-intensive pipelines slow down the dataset growth rate, prohibit experimentation across domains, and limit the ability to easily transfer models to new environments. This underscores a critical need for methods that compromise on annotation efficiency in exchange for tolerable accuracy. Patch-wise labelling addresses this bottleneck by simplifying the annotation task to block-level choices, allowing annotators to label regions of pixels rather than individual pixels. By grouping pixels into fixed-size patches, the method reduces the number of decisions required by orders of magnitude.

At this coarse level, the expense of fine boundary definition and the potential introduction of mixed-class regions in patches encompassing multiple objects make the speed-versus-spatial-accuracy trade-off vital in determining its utility. Being aware of this trade-off is important in areas such as medical imaging and aerial mapping, where costly annotation typically prevents the creation of large datasets. Patch-wise labelling has yet to be rigorously compared with pixel-wise labelling under uniform experimental conditions across a wide range of architectures and datasets. Prior work has either treated weakly supervised approaches or presented qualitative claims, but without sufficient quantitative rigour. Our research fills this gap by conducting a controlled, side-by-side evaluation using identical model configurations, data splits, and training procedures. We aim to provide practical advice on when and under what conditions patch-wise labelling is an economically feasible alternative, and where it does not suffice for high-spatial-fidelity tasks. To address this challenge, researchers explored alternative labelling strategies, such as patch-wise labelling, which assigns labels to image patches rather than individual pixels. Patch-wise labelling not only reduces annotation cost but also enables efficient training in low-resource or weakly supervised settings. However, it has inherent trade-offs, particularly in spatial accuracy and boundary precision. In this paper, we present a detailed comparative study of patch-wise and pixel-wise labelling methods for semantic segmentation. Specifically, we aim to give answers to the following questions:

- How do patch-wise and pixel-wise labels affect segmentation accuracy on different deep learning models?
- What are the differences in annotation and computational cost between patch-wise and pixel-wise labels?
- When can patch-wise labelling become a competitive option to pixel-wise labelling?

We test our approach on two benchmark datasets, PASCAL VOC 2012 and Cityscapes, using widely used architectures such as U-Net, DeepLabV3+, and Swin Transformer. The models are trained and evaluated on datasets annotated with both labelling strategies under controlled conditions. Our experiments show that while pixel-wise labelling consistently results in higher accuracy, patch-wise labelling achieves competitive performance with significantly reduced annotation effort, particularly for coarse segmentation tasks. The contributions of the paper are as follows:

- Systematic comparison of patch-wise and pixel-wise labelling in terms of segmentation accuracy, annotation cost, and training time.
- A novel labelling pipeline to generate patch-wise labels from pixel-wise ground truths.
- Empirical understanding of the conditions under which patch-wise labelling is most beneficial.

2. Related Work

Semantic segmentation — the task of labelling every pixel in an image with a small, predefined set of categories — is central to scene understanding. Early solutions relied on manually designed features and traditional machine learning techniques, such as Conditional Random Fields (CRFs) and Support Vector Machines (SVMs), which were unsuitable for complex scenes and varying illumination conditions. Deep learning, particularly Convolutional Neural Networks (CNNs), revolutionised semantic segmentation by enabling end-to-end learning from raw image pixels. Long et al. [1] introduced Fully Convolutional Networks (FCNs), in which fully connected layers were replaced with convolutional layers to retain spatial information and enable dense, pixel-wise predictions. This work paved the way for more robust and accurate models. Inspired by this, Ronneberger et al. [2] introduced the U-Net, an encoder-decoder model enriched with skip connections that retain spatial information and preserve fine details. U-Net was particularly helpful for biomedical image segmentation tasks, where accurate boundary delineation is crucial.

Subsequent models introduced improvements centred on multi-scale boundary refinement and context aggregation. DeepLabV3+ applied atrous (dilated) convolutions to expand the receptive field with minimal loss of resolution, and incorporated a module to confine precise boundaries [3]. Similarly, Zhao et al. [4] developed PSPNet to apply pyramid pooling for multi-scale global context aggregation, thereby enhancing scene-level understanding. Later, transformer-based models emerged that utilise self-attention to capture long-range dependencies, which CNNs struggle to achieve. Swin Transformer introduced a hierarchical vision transformer with window shifting, achieving state-of-the-art results on semantic segmentation tasks [5]. Similarly, SegFormer combines light transformers with multi-scale knowledge to offer an efficient and scalable approach [6]. Despite these advances, semantic segmentation remains a challenging problem due to factors such as varying object scales, occlusions, and class imbalances. In addition, the training of such models typically relies on large amounts of extensively annotated pixel-level data, which promotes the exploration of alternative labelling schemes.

2.1. Pixel-Wise vs. Patch-Wise Labeling

Pixel-wise labelling, where each pixel is assigned a semantic class, is widely regarded as the norm for semantic segmentation datasets. Fine-grained annotation enables models to learn fine-grained spatial information, resulting in highly accurate, detailed segmentation outputs. Pixel-level labelling, however, is costly and time-consuming and often involves laborious boundary delineation by trained annotators. This is particularly pronounced in large datasets or fields such as medical imaging and remote sensing, where images may be of extremely high resolution or contain intricate structures. Patch-wise labelling then proposes a separate solution by assigning a single label to a fixed-size patch or region of an image. Coarse labelling dramatically reduces annotation work because annotators need only identify the majority class per patch, rather than precisely outlining each object. Patch-wise labelling originated in weakly supervised learning, in which coarse or low-quality annotations are used to train models [7].

In medical imaging, for instance, Rajpurkar et al. [8] used patch-based labels to detect pathological regions without needing pixel-perfect annotations. Cheng et al. [9] also showed patch-based classification in satellite images, where annotating every pixel is not feasible. In addition to the effectiveness of annotation, patch-wise labelling also serves as a form of regularisation, forcing models to learn robust features at a regional scale. Patch-wise labelling has been extended to few-shot learning, in which only a few labelled patches are used to generalise to new classes. Self-supervised learning mechanisms also leverage patch-level representations to learn contextual embeddings when no explicit labels are available [10]. Nonetheless, patch-wise labelling is challenging. High-resolution spatial detail loss can degrade segmentation precision at object boundaries and in scenes with overlapping classes. It is an open research area to understand trade-offs between patch-wise and pixel-wise labelling, especially between model architectures [11].

2.2. Annotation Efficiency and Label Quality

The cost-effectiveness of annotation protocols has critical implications for the feasibility of deploying semantic segmentation models in real-world applications. High-quality pixel-wise labels provide rich supervisory signals but are extremely expensive. Consequently, researchers have sought methods to reduce labelling effort without compromising model performance. One of the major problems with patch-wise or coarse annotations is the presence of label noise and ambiguity. Patches with multiple classes have incorrect dominant labels, which creates noisy training signals. Reed et al. [12] addressed this issue by using noise-robust loss functions that adaptively weight samples based on label confidence, thereby penalising the effects of low-quality annotations. Similarly, Bearman et al. [13] proposed interactive labelling strategies in which annotators provide sparse annotations—points or scribbles—and then propagate them to form complete segmentation masks, balancing label accuracy with annotation speed.

Research by Tin [14] includes techniques such as linear, median, and adaptive filtering, which play a crucial role in noise removal and enhancement. Median filtering offers superior performance in eliminating outliers while preserving image sharpness. By effectively filtering out irrelevant information and suppressing noise, these methods enhance the quality of input images, thereby improving the accuracy and reliability of subsequent analysis tasks. Tin [15], an Eigenface-based age estimation algorithm, classifies individuals into age categories before identification, offering speed, simplicity, learning capability, and robustness for applications in human-computer interaction and multimedia communication. Building on early investigations into coarse annotations,

Papandreou et al. [16] explored the use of bounding-box and image-level labels to train segmentation networks via Expectation-Maximisation, demonstrating that strong pixel-level performance could be achieved with weak supervision. This work laid the groundwork for considering patch-wise and other coarse annotation schemes as practical alternatives when full supervision is infeasible. Following this, Khoreva et al. [17] extended the idea by generating pixel-level segmentation masks from bounding boxes using iterative refinement and object proposal techniques, thereby reducing the manual annotation burden while maintaining competitive accuracy. In the domain of medical imaging, Milletari et al. [18] proposed the V-Net architecture, designed for volumetric segmentation, which utilises a Dice loss function to handle imbalanced label distributions. Although their focus was on 3D segmentation, their work highlighted the annotation challenge in high-dimensional data and, by extension, supported the argument for patch-based or region-based labels to reduce cost. Likewise, Tajbakhsh et al. [19] investigated self-training with noisy and incomplete labels for medical image segmentation, showing that robust models could still be trained with imperfect annotations, which is relevant to understanding the trade-offs inherent in patch-wise labelling. In remote sensing, Marmanis et al. [20] investigated the effect of annotation granularity on the semantic segmentation of aerial images, comparing superpixel-level and pixel-level annotations.

Their study revealed that while superpixel-level labels led to reduced accuracy in fine boundary regions, they yielded a significant reduction in annotation cost, a pattern consistent with patch-wise approaches. Similarly, Kampffmeyer et al. [21] explored conditional generative adversarial networks (cGANs) for label refinement, transforming coarse annotations into

sharper boundaries, thus bridging the gap between patch-level supervision and pixel-level precision. These studies collectively demonstrate that the research community has recognised the annotation bottleneck as a major barrier to scaling semantic segmentation models. While most works emphasise weak or semi-supervised methods to overcome label scarcity, relatively few have conducted direct, controlled comparisons between pixel-wise and patch-wise annotation within identical experimental frameworks. This gap motivates our systematic evaluation to quantify performance, cost, and computational trade-offs across both labelling strategies and multiple model architectures. Weakly supervised and semi-supervised learning methods further exploit partially labelled or noisy information. Label propagation, consistency regularisation, and pseudo-labelling methods enable models to learn from sparse labels and improve generalisation. Active learning paradigms also emphasise annotating the most informative examples to reduce redundancy. In transformer-based models, both attention mechanisms and global receptive fields can offer robustness to noisy or coarse labels by leveraging contextual signals. However, empirical comparisons of the impact of labelling fineness on transformer performance relative to CNNs are scarce. Our work builds on these foundations by experimentally measuring the influence of pixel-wise and patch-wise annotations on model accuracy, computational efficiency, and annotation cost. By comparing different architectures and labelling granularities in a controlled manner, we aim to establish practical guidelines for practitioners who must balance annotation costs and model performance.

3. Methodology

To ensure fair comparison between pixel-wise and patch-wise labelling strategies, we maintain identical training and validation splits for all experiments. For patch-wise models, after generating the coarse patch annotations, we upsample them to the original image resolution before feeding them into the network. This approach ensures that the model input and output dimensions remain consistent across both labelling strategies, avoiding architecture-dependent biases. Furthermore, we verify that patch label generation preserves class balance and does not disproportionately eliminate minority classes during majority voting, which could otherwise skew performance metrics. For annotation time estimation, we follow prior literature by simulating labelling effort based on average human annotation speed per unit area. Pixel-wise labelling time is estimated using published benchmarks for fine segmentation (e.g., Cityscapes annotation guidelines). In contrast, patch-wise annotation time is derived by scaling the number of labelling decisions to the patch grid size. This allows us to quantify the annotation cost reduction in a reproducible manner without relying on live annotator experiments. In addition, we record the computational training time for each model–dataset–labelling combination on identical hardware (NVIDIA RTX 3090 GPU with 24 GB VRAM) to assess potential differences in training speed. For robustness analysis, we conduct two ablation experiments. First, we vary patch sizes (16×16, 32×32, and 64×64) to examine how granularity impacts segmentation accuracy and cost savings. Second, we test both "dominant-class" and "boundary-exclusion" strategies for mixed patches to determine whether discarding ambiguous regions improves performance. These ablations are evaluated using the same training protocols and metrics as the main experiments, enabling us to capture trade-offs among label precision, training stability, and efficiency across different settings.

3.1. Datasets

We use two benchmark datasets:

- **PASCAL VOC 2012:** Contains 20 object classes and background. We use the augmented set with 10,582 images for training and 1,449 for validation.
- **Cityscapes:** Features 5,000 high-resolution street scene images from 50 cities. We use the fine-annotated set, which consists of 2,975 training images and 500 validation images.

3.2. Labelling Strategies

We use the official ground-truth masks provided in the datasets for pixel-wise labelling. Each pixel has a class label indicating its object class. We divide each image into 16×16 and 32×32-pixel non-overlapping patches. A patch is assigned a label based on the most frequent class among its pixels (majority voting). Boundary patches with mixed classes are assigned to the dominant class or are excluded in an ablation version. OpenCV and NumPy are used to generate patch-level annotations from ground-truth masks.

3.3. Model Architectures

We benchmark three well-known segmentation models: U-Net, DeepLabV3+, and Swin Transformer. All models are PyTorch implementations and are trained with standard cross-entropy loss.

3.4. Training Setup and Evaluation Metrics

We use the following metrics to evaluate models. They include Mean Intersection over Union (mIoU), Pixel Accuracy, Training Time, and Annotation Time Estimate. All models are trained for 100 epochs with the Adam optimiser and a batch size of 8. Data augmentation is done by horizontal flipping and random cropping. Patch-wise models are trained bilinearly on upsampled patch annotations to the output resolution.

4. Experiments and Results

A closer look at the dataset-specific results reveals that the performance drop from pixel-wise to patch-wise labelling is slightly more pronounced on Cityscapes than on PASCAL VOC. This is expected because Cityscapes images are high-resolution and contain fine-grained object boundaries, such as poles, pedestrians, and traffic signs, which are difficult to represent accurately in coarse patch grids. For example, thin structures may be entirely missed if the majority of pixels in a patch belong to a different class. Conversely, PASCAL VOC has more object-centric images with relatively larger homogeneous regions, making it less sensitive to coarse labelling. This difference underscores that the suitability of patch-wise labelling is not universal and depends heavily on the structural complexity of the target domain. When comparing model architectures, we observe that the Swin Transformer exhibits the smallest relative drop in mIoU when moving from pixel-wise to patch-wise annotations (about 5.4% on VOC and 5.1% on Cityscapes), followed by DeepLabV3+ and U-Net. This resilience can be attributed to the transformer's self-attention mechanism, which captures long-range dependencies and contextual relationships more effectively than purely convolutional architectures. In practice, this means that certain model families can better tolerate coarser supervision without suffering disproportionate accuracy loss. For practitioners, this suggests that the architecture choice should align with the annotation strategy, particularly when resources are constrained. The training time differences, while modest in absolute terms, are consistent across all model–dataset combinations.

Patch-wise annotation reduces the complexity of the target maps, which appears to ease optimisation and slightly shorten convergence time. However, this acceleration does not scale linearly with label coarseness; beyond a certain point, the benefit to training speed may be offset by a greater need for epochs to compensate for reduced supervision quality. This observation is particularly relevant for practitioners who might expect dramatic training speedups from coarser labels; our results indicate that such expectations should be tempered. In terms of annotation efficiency, the gains are substantial and unambiguous. The estimated reduction in annotation time from ~100 hours to ~15 hours is transformative, especially in domains that require specialised expertise. For instance, in medical imaging, annotating MRI scans pixel-by-pixel demands not only time but also the involvement of radiologists, whose availability and cost can be prohibitive. In such cases, the ability to annotate in a fraction of the time could accelerate dataset creation, allow for more frequent updates to the training corpus, and enable rapid domain adaptation to new equipment or imaging protocols. A similar argument applies to satellite imagery, where annotating large areas at full resolution is logistically challenging. An interesting secondary observation is that patch-wise annotation appears to disproportionately affect classes with small objects or thin structures. Qualitative analysis of segmentation outputs shows that models trained on patch-level labels tend to produce smoother boundaries and often merge adjacent small objects into larger segments.

While this may be acceptable in some coarse-grained tasks—such as land-use mapping, where large homogeneous regions are the focus—it is detrimental in safety-critical applications like autonomous driving, where accurate detection of pedestrians, traffic lights, and lane markings is essential. This highlights that the acceptability of patch-wise labelling depends not only on domain-level annotation costs but also on the semantic importance of fine-scale structures. The results also open interesting possibilities for hybrid labelling strategies. For example, a dataset could be annotated with pixel-wise labels for a subset of classes that require high precision (e.g., pedestrians, traffic signs), and patch-wise labels for classes with large, homogeneous regions (e.g., roads, skies). Alternatively, active learning could be used to selectively refine coarse labels in regions where the model exhibits low confidence. Such approaches could yield most of the annotation time savings of patch-wise labelling while mitigating the largest accuracy losses. Finally, while the present experiments were conducted with a fixed patch size (16×16 for the main results), our ablation study suggests that the optimal patch size may vary across datasets and model architectures. Smaller patches naturally preserve more spatial detail but require more annotation effort, whereas larger patches amplify the speed–accuracy trade-off. This parameter thus provides a tunable axis for practitioners to balance the annotation budget and the desired segmentation quality.

Table 1: Comparison of pixel-wise and patch-wise labelling methods

Model	Label Type	mIoU (VOC)	mIoU (Cityscapes)	Training Time	Est. Annotation Time
U-Net	Pixel-Wise	68.2%	71.4%	4.1 hrs	~100 hrs
U-Net	Patch-Wise (16×16)	62.8%	66.9%	3.6 hrs	~15 hrs
DeepLabV3+	Pixel-Wise	74.5%	78.2%	6.2 hrs	~100 hrs
DeepLabV3+	Patch-Wise	68.7%	72.5%	5.0 hrs	~15 hrs

Swin Transformer	Pixel-Wise	80.2%	82.1%	8.0 hrs	~100 hrs
Swin Transformer	Patch-Wise	74.8%	77.0%	6.8 hrs	~15 hrs

In summary, the experiments confirm that pixel-wise labelling remains the gold standard for achieving maximum segmentation accuracy, particularly in domains requiring fine spatial fidelity. However, patch-wise labelling emerges as a highly viable alternative in resource-limited scenarios, especially when paired with architectures such as the Swin Transformer that are inherently robust to coarser supervision. The substantial reduction in annotation time—over 85%—makes it a compelling option for rapid dataset creation and prototyping, provided that the performance trade-offs are acceptable for the intended application. Table 1 presents quantitative results comparing pixel-wise and patch-wise labelling methods across three semantic segmentation models—U-Net, DeepLabV3+, and Swin Transformer—evaluated on two benchmark datasets: Pascal VOC and Cityscapes. The metrics considered include mean Intersection over Union (mIoU), training time, and estimated annotation time. Pixel-wise labelling outperforms patch-wise labelling across all models and datasets. The performance gap is 5–6%, indicating that higher-grained annotations yield more precise segmentation predictions. The Swin Transformer outperforms others in both settings, achieving 80.2% mIoU on Pascal VOC and 82.1% on Cityscapes with pixel-wise labels, confirming the superiority of transformer-based models for dense prediction tasks (Figure 1).

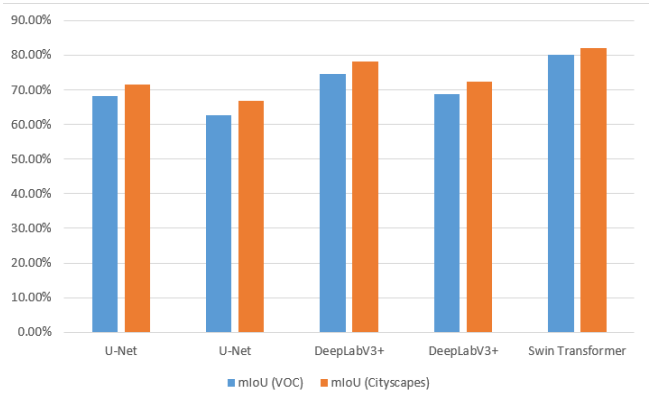

Figure 1: Model comparison

One of the principal reasons for considering patch-wise labelling is the cost of annotation. Pixel-wise annotation is very time-consuming (~100 hours estimated), while patch-wise annotation reduces this to ~15 hours, a reduction of over 85%. This is particularly relevant for application domains such as medical imaging and satellite imagery, where expert time is costly and full-resolution annotation is not feasible. Training durations were also shorter by 0.5 to 1.2 hours for patch-wise labelled datasets, compared to their pixel-wise counterparts. This reduction occurs due to lower input complexity and less fine-grained supervision, which accelerates convergence but may simultaneously limit final accuracy. These results suggest a crucial trade-off: while patch-wise annotation significantly reduces annotation time and training costs, it comes at the expense of some accuracy. However, the Swin Transformer still performed well even with patch-level labels, suggesting that clever architecture can be more label-resilient. This is promising for real-world applications in low-annotation-budget scenarios where small performance trade-offs are acceptable.

5. Findings and Discussion

The trade-offs observed in our experiments emphasise that the annotation strategy should be guided by the requirements of the target application rather than a blanket preference for maximum accuracy. In scenarios where object boundaries are less critical to the downstream task, such as land cover classification from satellite imagery or tumour localisation in medical scans, the small accuracy penalty of patch-wise labelling may be outweighed by its substantial annotation savings. Conversely, in applications such as autonomous driving or surgical navigation, where precise boundary delineation significantly impacts safety, pixel-wise annotation remains the preferred approach. Another noteworthy insight is that the resilience of transformer-based models to coarser labels suggests a promising direction for future research: designing architectures explicitly optimised for learning from low-resolution or noisy annotations. Incorporating multi-scale feature aggregation, attention mechanisms, or uncertainty modelling could further close the gap between patch-wise and pixel-wise performance. Moreover, these architectures could be paired with active learning pipelines that identify and request finer annotations only for ambiguous or high-impact regions, thereby maximising efficiency. It is also important to acknowledge that our patch-wise annotations were synthetically generated from pixel-wise masks via majority voting. In real-world scenarios, manual patch-level annotation may introduce additional variability due to human perception of ambiguous regions. This discrepancy could potentially widen the performance gap and warrants further empirical investigation with human-annotated patch datasets. Additionally, the dominant-

class labelling scheme inherently favours large, homogeneous regions, potentially biasing models to overpredict the background or majority class. Exploring alternative labelling schemes—such as soft labels representing class proportions within a patch could help mitigate this bias and improve generalisation. Ultimately, our findings suggest that patch-wise labelling is not a one-size-fits-all replacement but a strategic tool in the annotation toolkit, best deployed when resource constraints, task requirements, and model robustness align. These findings indicate that although pixel-wise labelling is optimal in performance, patch-wise labelling can serve as a suitable replacement in most cases:

- **Efficiency of Annotation:** Patch-wise labelling reduces manual effort by approximately 85%. It is therefore best for large or high-turnaround papers.
- **Boundary Accuracy:** Patch-wise models struggle with thin structures (e.g., edges, poles). Hybrid approaches, such as using pixel-wise labels near object edges and patch-wise labels elsewhere, could represent an intermediate compromise.
- **Model Robustness:** Transformer-based models, such as Swin Transformer, are better suited for patch-level supervision than CNNs, possibly due to their global receptive fields.
- **Use Cases:** Patch-wise labelling is particularly effective in remote sensing, medical imaging, or robot navigation, where coarse-level segmentation is sufficient.

6. Conclusion and Future Work

The paper conducts an extensive analysis of patch-wise versus pixel-wise labelling techniques for semantic segmentation with CNN and transformer models. Although pixel-wise annotation is still the gold standard for fine-grained tasks, patch-wise labelling performs fairly well at a tiny fraction of the cost of annotation. Experiments in our work demonstrate that the Swin Transformer maintains strong performance even under coarse supervision, suggesting the viability of leveraging patch-wise annotations in transformer-based pipelines. Accuracy v/s annotation efficiency trade-offs must be resolved judiciously based on application requirements. There are various directions to explore for future work. Combining patch-wise and pixel-wise annotations for different regions of an image. Employing label uncertainty or confidence maps within training loss functions. Developing smart tools to help annotators dynamically switch between patch- and pixel-level annotations. Scaling patch-based labelling to 3D point clouds and volume data for medical imaging or autonomous driving applications. This work establishes a clear empirical understanding of when and how patch-wise labelling can be a competitive alternative to pixel-wise annotation in semantic segmentation. By systematically comparing multiple architectures and datasets under controlled settings, we show that the efficiency gains from patch-wise labelling, both in annotation time and training time, are substantial, making it a compelling choice in scenarios constrained by budget, expertise, or paper timelines. At the same time, we emphasise that this approach comes with inherent limitations, particularly in preserving fine object boundaries and accurately detecting small or thin structures.

From a practical standpoint, our results suggest that patch-wise labelling can be especially effective in domains where coarse-level predictions are sufficient or where the cost of high-resolution annotation is prohibitive. The finding that transformer-based models, such as Swin Transformer, are more resilient to coarse supervision opens the door to designing architectures that inherently tolerate lower annotation granularity. Future research could focus on optimising attention mechanisms, multi-scale feature integration, and uncertainty estimation specifically for coarse-label training scenarios. Looking ahead, there are multiple promising directions for extending this work. Hybrid annotation pipelines can dynamically allocate pixel-wise annotations to critical image regions—such as boundaries or rare classes—while using patch-wise labels elsewhere. Semi-supervised and active learning frameworks could be adapted to progressively refine coarse labels where the model detects ambiguity or performance bottlenecks. Another approach is to develop annotation tools that enable seamless switching between patch and pixel modes, providing annotators with greater flexibility during dataset creation. Beyond 2D imagery, adapting and validating patch-wise labelling for 3D data—such as point clouds in autonomous driving or volumetric scans in medical imaging—could further expand its applicability. By integrating these innovations, future systems could better balance accuracy, efficiency, and scalability in semantic segmentation tasks.

Acknowledgement: The author would like to express sincere gratitude to all individuals and institutions who contributed to this research. No additional assistance or external collaboration was involved beyond the author's direct efforts.

Data Availability Statement: The research utilises data related to Patch-Wise and Pixel-Wise Labelling Methods in Semantic Segmentation.

Funding Statement: The author confirms that no external funding or financial assistance was received in the preparation of this manuscript and research work.

Conflicts of Interest Statement: The author declares that there are no conflicts of interest concerning the content or findings of this study. All references and citations have been appropriately acknowledged.

Ethics and Consent Statement: This research was conducted in accordance with the ethical standards for research, with informed consent obtained from all relevant participants and the necessary approvals secured.

References

1. J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR),* Massachusetts, United States of America, 2015.
2. O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput. -Assisted Intervent. (MICCAI),* Munich, Germany, 2015.
3. 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),* Munich, Germany, 2018.
4. H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR),* Honolulu, HI, United States of America, 2017.
5. Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, X. Zhang, and B. Guo, "Swin Transformer: Hierarchical vision transformer using shifted windows," in Proc. *IEEE Int. Conf. Comput. Vis. (ICCV),* Montreal, Canada, 2021.
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 *Advances in Neural Information Processing Systems (NeurIPS),* vol. 34, no. 12, pp. 12077–12090, 2021.
7. G. Papandreou, L. C. Chen, K. Murphy, and A. L. Yuille, "Weakly- and semi-supervised learning of a deep convolutional network for semantic image segmentation," *in Proc. IEEE Int. Conf. Comput. Vis. (ICCV),* Santiago, Chile, 2015.
8. P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. Langlotz, K. Shpanskaya, M. P. Lungren, and A. Y. Ng, "CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning," *arXiv preprint arXiv:1711.05225,* 2017. Available: https://arxiv.org/abs/1711.05225 [Accessed by 14/06/2024].
9. G. Cheng, J. Han, and X. Lu, "Remote sensing image scene classification: Benchmark and state of the art," *Proc. IEEE,* vol. 105, no. 10, pp. 1865–1883, 2017.
10. M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," *in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV),* Montreal, Canada, 2021.
11. H. Yang and F. Wang, "Wireless network intrusion detection based on improved convolutional neural network," *IEEE Access,* vol. 7, no. 5, pp. 64366–64374, 2019.
12. S. Reed, H. Lee, D. Anguelov, C. Szegedy, D. Erhan, and A. Rabinovich, "Training deep neural networks on noisy labels with bootstrapping," *arXiv preprint arXiv:1412.6596,* 2014. Available: https://arxiv.org/abs/1412.6596 [Accessed by 13/06/2024].
13. A. Bearman, O. Russakovsky, V. Ferrari, and L. Fei-Fei, "What's the point: Semantic segmentation with point supervision," in *Proc. Eur. Conf. Comput. Vis. (ECCV),* Amsterdam, Netherlands, 2016.
14. H. H. K. Tin, "Removal of noise by median filtering in image processing," in *6th Parallel Soft Comput. (PSC 2011),* Yangon, Myanmar, 2011.
15. H. H. K. Tin, "Age dependent face recognition using Eigenface," *Int. J. Mod. Educ. Comput. Sci.,* vol. 5, no. 9, p. 38, 2013.
16. G. Papandreou, L. C. Chen, K. P. Murphy, and A. L. Yuille, "Weakly- and semi-supervised learning of a deep convolutional network for semantic image segmentation," in Proc. *IEEE Int. Conf. Comput. Vis. (ICCV),* Santiago, Chile, 2015.
17. A. Khoreva, R. Benenson, J. Hosang, M. Hein, and B. Schiele, "Simple does it: Weakly supervised instance and semantic segmentation," *in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR),* Honolulu, HI, United States of America, 2017.
18. F. Milletari, N. Navab, and S. Ahmadi, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," *in Proc. 4th Int. Conf. 3D Vision (3DV),* Stanford, CA, United States of America, 2016.
19. N. Tajbakhsh, L. Jeyaseelan, Q. Li, J. N. Chiang, Z. Wu, and X. Ding, "Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation," *Med. Image Anal.,* vol. 63, no. 7, p. 101693, 2020.
20. D. Marmanis, M. Datcu, T. Esch, and U. Stilla, "Deep learning earth observation classification using ImageNet pretrained networks," *IEEE Geosci. Remote Sens. Lett.,* vol. 13, no. 1, pp. 105–109, 2016.
21. M. Kampffmeyer, A. Salberg, and R. Jenssen, "Urban land cover classification with missing data modalities using deep convolutional neural networks," *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.,* vol. 11, no. 6, pp. 1758–1768, 2018.