Jurnal Ilmu Komputer dan Informasi (Journal of Computer Science and Information)
19/2 (2026), 179-195. DOI: http://dx.doi.org/10.21609/jiki.v19i2.1757

Efficient and Robust Crosswalk Segmentation under Adverse Weather Using ConvNeXt-Enhanced DeepLabv3

Muhammad Faqih, Ridho Aulia Rahman, and Khadijah Fahmi Hayati Holle

Department of Informatics Engineering, Universitas Islam Negeri Maulana Malik Ibrahim Malang, Malang, Indonesia

E-mail: khadijah.holle@uin-malang.ac.id

Abstract

Reliable crosswalk perception is crucial for first-person vision (FPV) navigation in assistive guidance and intelligent transportation, but segmentation accuracy often decreases under glare, rain reflections, nighttime illumination, and worn low-contrast markings. This study proposes ConvNeXt-Enhanced DeepLabv3 (CEDL), a pixel-level segmentation architecture that integrates DeepLabv3 atrous multi-scale encoding with the modern convolutional design of ConvNeXt-Tiny. Experiments were conducted on the FPVCrosswalk2025 dataset, containing synthetic and real FPV images captured under sunny, cloudy, rainy, and night conditions. The proposed model was compared with DeepLabv3 using ResNet-50 and MobileNetV3-L backbones under the same training and evaluation protocol. CEDL achieved the best overall performance, with 0.946 mean IoU and 0.972 Dice, while maintaining strong per-condition robustness and improved boundary preservation for thin crosswalk structures. It also achieved practical inference speed at 20.6 ms per frame, nearly five times faster than ResNet-50, despite having more parameters than MobileNetV3-L. Qualitative results show more continuous crosswalk stripes and fewer missed segments under adverse conditions. These findings indicate that CEDL provides a robust and computationally practical solution for FPV crosswalk segmentation on a mixed synthetic-real benchmark.

Keywords: *Crosswalk segmentation, semantic segmentation, adverse weather conditions, ConvNeXt backbone, DeepLabv3.*

1. Introduction

Pedestrian safety and reliable perception of road infrastructure are central requirements for intelligent transportation systems and autonomous driving. Crosswalks function as critical visual cues that regulate vehicle behavior, yet pedestrian injuries remain high, with pedestrians accounting for around 22% of global road traffic deaths [1]. For visually impaired individuals, detecting tactile guidance paths (often referred to as blind roads) and crosswalks is particularly challenging, motivating the development of robust vision-based segmentation methods [2]. Despite their importance, traffic landmark segmentation, such as crosswalk detection, has received comparatively limited scholarly attention relative to other road-scene perception tasks [3].

Early crosswalk detection relied on classical image-processing techniques such as illuminant-invariant color spaces, Gaussian modeling, and co-occurrence matrices. While computationally lightweight, these methods remain highly sensitive to illumination changes, perspective distortions, and background clutter [2]. Traditional pipelines based on vanishing-point estimation, edge extraction, or template matching likewise exhibit reduced reliability under varying lighting conditions, degraded markings, or nonuniform block shapes [4]. Collectively, these limitations have driven a shift toward deep learning–based semantic segmentation paradigms that can learn discriminative features directly from data and operate more robustly in unconstrained scenes.

Deep convolutional networks have substantially advanced dense prediction. DeepLabv3 demonstrates that atrous convolutions and multi-rate context aggregation mitigate resolution loss while enhancing multi-scale representation learning [5]. Concurrently, modern convolutional backbones such as ConvNeXt show that carefully redesigned ConvNeXt architectures

can match or exceed transformer-based models on large-scale benchmarks while retaining architectural simplicity and strong feature extraction capacity [6]. These developments form a compelling foundation for crosswalk segmentation, where maintaining spatial precision and modeling multi-scale stripe patterns are essential.

Several works have examined crosswalk or traffic landmark recognition. A lightweight network for blind-road and crosswalk segmentation based on depthwise separable convolutions and contextual feature aggregation was proposed in [2]. Mask R-CNN with ResNet-101 can detect crosswalks with high accuracy, whereas CDNet adapts YOLOv5 for real-time landmark detection across diverse illumination conditions [4]. A large-scale traffic landmark dataset containing challenging conditions such as shadows, glare, and pavement deterioration was introduced in [3], along with a comprehensive benchmarking of various semantic segmentation models. Although these studies collectively demonstrate the promise of deep learning for crosswalk-related perception, most focus on object detection, instance segmentation, or specific landmark subsets rather than comprehensive pixel-level modeling of crosswalk structures.

Achieving robust segmentation under adverse environmental conditions remains insufficiently explored. Scenes affected by nighttime glare, heavy shadows, wet surfaces, and worn markings introduce low-contrast boundaries that challenge segmentation models. Despite the capabilities of atrous-based context modeling and modern convolutional backbones, none of the reviewed studies explicitly integrate these designs for pixel-level crosswalk segmentation under systematically varied weather and illumination conditions [5], [6]. Related research in UAV-based road or crack segmentation reinforces the importance of capturing subtle texture transitions in complex backgrounds yet these findings have not been translated into ground-level crosswalk scenarios [7], [8]. Existing work also tends to emphasize either lightweight design, detection pipelines, or dataset-level benchmarking, leaving unresolved whether multi-scale atrous encoders paired with strong convolutional backbones can deliver improved robustness in adverse conditions [1], [2], [3], [4].

Unlike object detection or region-based approaches, crosswalk perception in first-person-view scenarios is better formulated as pixel-wise semantic segmentation because downstream navigation requires an accurate dense occupancy map rather than coarse localization. Crosswalks are characterized by thin, repetitive stripe patterns whose spatial continuity and precise boundaries are critical for estimating the traversable footprint, crosswalk width, and orientation in the ground plane. Detection-based methods or patch-wise representations tend to localize crosswalks at coarse spatial granularity, which is insufficient for capturing fine structural details, partial occlusions, worn markings, or boundary degradation commonly observed under adverse illumination and weather conditions [1], [4]. Moreover, because crosswalk markings are often low-contrast, small spatial errors can translate into larger geometric deviations after perspective projection. Pixel-level modeling enables explicit supervision at each spatial location, allowing the network to preserve stripe geometry and maintain structural consistency even when contrast is severely reduced [2], [5]. From a safety perspective, missing crosswalk stripes (false negatives) is more critical than small boundary over-segmentation, because a missed crossing cue can delay yielding behavior or misguide assistive navigation. Pixel-wise segmentation provides a dense, geometry-preserving map that supports conservative safety reasoning (e.g., maintaining the predicted traversable footprint and avoiding boundary breaks under low contrast), which is difficult to obtain reliably from coarse bounding-box detection or proposal-driven instance segmentation.

Furthermore, this work adopts a convolutional neural network (CNN)–based architecture rather than a purely transformer-based formulation. While transformer models have demonstrated strong global context modeling capability, their patch-wise tokenization and high computational overhead can be suboptimal for real-time, pixel-accurate segmentation of thin road markings in first-person-view settings. Modern CNN designs, such as ConvNeXt, have shown that carefully redesigned convolutional architectures can match or surpass transformer-based models in dense prediction tasks while retaining strong spatial inductive bias and computational efficiency [6]. These properties make CNN-based encoders particularly suitable for pixel-wise crosswalk segmentation, where fine boundary preservation, robustness to local photometric distortion, and real-time feasibility are essential. Accordingly, this study formulates crosswalk perception as a pixel-wise segmentation problem and employs a ConvNeXt-enhanced DeepLabv3 (CEDL) framework to explicitly address these requirements.

This study addresses these limitations by integrating the multi-scale context modeling of DeepLabv3 with the modernized convolutional design of ConvNeXt to improve crosswalk segmentation robustness across adverse weather

and illumination scenarios. DeepLabv3 provides an atrous convolution framework that enlarges the receptive field while maintaining spatial detail, and ConvNeXt offers strong hierarchical feature extraction shown to be effective in challenging visual tasks [5], [6]. Their combination aims to enhance pixel-level discrimination of crosswalk structures under cloudy, rainy, night, and standard daylight scenes, extending prior work that largely focuses on detection accuracy, efficiency, or selective landmark categories [1], [2], [3], [4].

To clarify the motivation, the research gaps identified from existing literature are as follows:

1. Lack of dedicated pixel-level crosswalk segmentation methods, as most studies emphasize detection or instance segmentation.
2. Limited evaluation under adverse weather and illumination, despite real-world relevance.
3. Absence of architectures combining atrous-based multiscale encoders with modern convolutional backbones specifically for crosswalk segmentation.
4. Insufficient exploration of robustness against degraded markings and low-contrast conditions, which persist in real road environments.

By addressing these gaps, the present work contributes an efficient and robust formulation for crosswalk segmentation under challenging conditions. The integration of DeepLabv3 and ConvNeXt offers a principled pathway for enhancing spatial precision, contextual reasoning, and feature robustness, complementing existing detection-oriented and lightweight segmentation approaches while advancing the state of research in adverse-condition road-scene perception.

2. Related Works

Semantic segmentation for road scenes and pedestrian infrastructure has undergone rapid advances over the past decade, largely driven by improvements in deep convolutional architectures, multi-scale context modeling, and the evolution of backbone networks. Foundational work such as DeepLabv3 [5] demonstrated that atrous convolution and the Atrous Spatial Pyramid Pooling (ASPP) module effectively preserve spatial resolution while expanding receptive fields. By adjusting the output stride (OS=16 or 8) and employing cascaded dilated convolutions, DeepLabv3 achieved 85.7% mIoU on PASCAL VOC 2012 without DenseCRF post-processing. This capability to densely encode fine structural details makes the design well-suited for repetitive, thin crosswalk patterns. This line of research was further advanced by the introduction of the Pyramid Pooling Module (PPM), which aggregates global-to-local contextual features across four pooling scales, achieving 85.4% mIoU on the PASCAL VOC dataset and 80.2% on Cityscapes [9]. Together, these works established multi-scale contextual encoding as a core requirement for robust segmentation of structured outdoor landmarks subject to complex backgrounds, occlusion, and illumination variability.

Progress in backbone network design further strengthened segmentation performance through the introduction of residual learning with identity shortcuts, enabling very deep architectures of up to 152 layers. This approach achieved a top-5 ImageNet error rate of 3.57% and became a widely adopted encoder for segmentation frameworks such as DeepLab, PSPNet, Mask R-CNN, and various U-Net variants due to its training stability and strong generalization capability [10]. Parallel research on lightweight segmentation addressed computational constraints in embedded perception systems through the integration of hardware-aware neural architecture search, inverted bottlenecks, squeeze-and-excitation modules, and the h-swish activation function. In this framework, the LR-ASPP decoder achieved a 34% inference speed improvement over the R-ASPP decoder of MobileNetV2 on the Cityscapes dataset while maintaining comparable segmentation accuracy [11]. Subsequent lightweight variants, such as Zhang and Chen, introduced leaky-ReLU6, skip connections, and depthwise separable convolutions to improve early-stage representation, though performance details were not explicitly reported [12]. Despite improved efficiency, lightweight backbones consistently exhibit performance degradation under adverse visual conditions including nighttime illumination, rain streaks, specular reflections, and low-contrast road surfaces, conditions where crosswalk boundaries often become visually ambiguous.

A major shift in backbone design was marked by the introduction of a modernized pure convolutional network aimed at competing with Transformer-based models while preserving the efficiency of convolutional operations [6]. ConvNeXt achieved 87.8% ImageNet top-1 accuracy and outperformed Swin Transformer on COCO detection and ADE20K segmentation. Its superior hierarchical feature extraction has motivated adoption in several dense-prediction tasks. Pan and Xiao integrated ConvNeXt within DeepLabv3+ for brain glioma segmentation, observing a performance increase from 70.15% to 70.33% mIoU, modest but meaningful within highly irregular lesion boundaries. Further evidence of the effectiveness of modern

convolutional backbones was demonstrated in a UAV-based pavement crack segmentation study, where a ConvNeXt-Large–UPerNet architecture achieved 79.73% mIoU, 76.03% F1-score, and 61.71% Crack-IoU, outperforming seven CNN- and ViT-based baseline models [7]. The model exhibited robustness in detecting extremely thin and low-contrast structures, directly paralleling the challenges encountered in zebra-cross segmentation under rain or low-light scenes. ConvNeXt to multispectral land-cover segmentation through a dual-branch encoder with Convolutional Block Attention Module (CBAM) and pyramidal progressive fusion, reaching 90.83% overall accuracy and 77.69% mIoU, surpassing U-Net, PSPNet, DeepLabv3, and DeepLabv3+ . Similarly, embedded ConvNeXt in a U-Net variant augmented with a residual deformable convolution pyramid, SIIM, and CBAM, yielding 10.3% and 13.12% mIoU improvements over baselines on Cityscapes . Collectively, these studies establish ConvNeXt as a resilient backbone for environments involving heterogeneous modalities, adverse illumination, and thin structures.

In parallel to modernized ConvNets, the state of the art in semantic segmentation has increasingly shifted toward transformer-based encoders and decoders. SegFormer introduces a hierarchical Mix-Transformer encoder and a lightweight multi-layer perceptron (MLP) decoder, reporting strong accuracy-efficiency trade-offs in dense prediction tasks [21]. Earlier patch-wise transformer segmenters such as SETR [23] and Segmenter [24] treat segmentation as sequence-to-sequence prediction over image patches, highlighting the benefit of global context modeling. More recently, foundation models such as the Segment Anything Model (SAM) demonstrate promptable segmentation with strong zero-shot generalization, but typically require prompts and can be computationally heavy for embedded real-time deployment [22]. Given the real-time constraints and the need for stable training on a task-specific first person vision (FPV) dataset, this study focuses on a CNN-based design (ConvNeXt + DeepLabv3) while positioning transformer-based segmenters as important baselines for future comparative evaluation.

Research explicitly targeting crosswalks and traffic landmarks reveals a different trajectory, with many studies emphasizing object detection or lightweight CNN pipelines rather than dense segmentation. Based on YOLOv5 with squeeze-and-excitation (SE) attention, Negative Samples Training, synthetic fog augmentation, and region of interest (ROI) processing, achieved 94.83% F1-score at 33.1 frames per second (FPS) on Jetson Nano across sunny, cloudy, rainy, foggy, and nighttime scenes [4]. However, its ROI acceleration induced a 7.69-point accuracy drop and the method focuses on detection rather than pixel-wise segmentation. Mask R-CNN was used by Malbog for instance-level segmentation, achieving 95.6-99.8% accuracy on a small dataset with limited illumination diversity [1]. A lightweight segmentation framework for blind roads and crosswalks was proposed using dense atrous spatial pyramid pooling and a context feature fusion module, demonstrating superior performance over MobileNet-based baselines, although explicit mIoU or F1-score metrics were not reported in the available evaluation results [2]. Although these works demonstrate feasibility, they do not incorporate high-capacity backbones such as ConvNeXt nor do they evaluate robustness under severe weather or nighttime visibility degradation.

Dataset availability also shapes progress. Kim and Shim's RSSO dataset includes crosswalks, stop lines, lane boundaries, bumps, and arrows, achieving 98.5% precision and 80.7% mIoU with BiSeNet, but illumination is constrained to daylight conditions (500–2000 lux) [13]. Studies using Mapillary Vistas with U-Net and U-Net-Inception report 85.07% and 88.41% pixel accuracy, but rely on pixel accuracy, which is an insensitive metric for minority classes, and do not isolate crosswalk performance [14]. Complementary but still limited evidence was provided by a dense hierarchical fusion U-Net variant incorporating multiscale attention mechanisms, which achieved 92.97% mIoU at 38.26 FPS for blind-road segmentation [15]. A segmentation framework combining super-resolution generative adversarial networks with a U-Net backbone was introduced to address low-resolution semantic segmentation on the BDD100K dataset; however, specific quantitative performance metrics were not reported in the available evaluation results [16]. A field-programmable gate array (FPGA)-based spiking neural network approach was investigated for real-time crosswalk segmentation, reporting an accuracy of 74.6% using the Izhikevich neuron model at 0.726 W power consumption and 62.86% accuracy with the Integrate-and-Fire model at 0.287 W [17]. While efficient, these hardware-oriented approaches fall significantly below the segmentation fidelity achieved by modern CNNs or ConvNeXt-based frameworks.

Synthesizing these results reveals several limitations that motivate further research. First, no existing study combines DeepLabv3's atrous multi-scale encoder with a ConvNeXt backbone for crosswalk segmentation, despite evidence that each component independently improves

robustness in complex environments. Pan and Xiao explored ConvNeXt within DeepLabv3+ but only in medical imagery rather than structured outdoor patterns. Second, although adverse-condition segmentation is a recurring challenge, no prior work evaluates pixel-level crosswalk segmentation across all major weather and illumination categories, namely sunny, cloudy, rainy, and nighttime conditions. Only CDNet covers a comparable range, but in a detection setting. Third, lightweight backbones such as MobileNetV3 deliver efficiency but repeatedly underperform in adverse conditions, suggesting a need for a stronger representation model. Fourth, widely used datasets such as RSSO provide limited nighttime or weather diversity, restricting generalization. Fifth, although ConvNeXt has shown success in thin-structure segmentation, such as pavement cracks, its suitability for zebra-cross patterns under weather-induced degradation has not been studied. Sixth, efficiency metrics such as inference latency, video RAM (VRAM) usage, and parameter count are seldom reported in crosswalk segmentation research, leaving an unresolved trade-off between robustness and deployability.

Given these gaps, the literature motivates a unified segmentation framework that integrates DeepLabv3's ASPP-based multi-scale context modeling with the representational capacity of ConvNeXt. The FPVCrosswalk2025 dataset, which includes sunny, cloudy, rainy, and nighttime conditions, provides an appropriate benchmark for evaluating this robustness in a manner not previously addressed in the literature [18]. Furthermore, by examining computational efficiency alongside segmentation quality, this work contributes to addressing long-standing limitations in the crosswalk segmentation domain, where robustness under weather variability and efficiency for real-world deployment remain insufficiently explored.

3. Method

This section explains the research methodology, including the dataset, preprocessing, model architecture, backbone variants, proposed CEDL framework, training configuration, and evaluation metrics used in this study.

3.1 Dataset

This study employs the FPVCrosswalk2025 dataset, a publicly released first-person-view (FPV) crosswalk segmentation dataset designed for diverse weather and lighting conditions [18]. All images and masks share a 640×360 resolution, are stored in .jpg format, and follow a strict one-to-one structure where each image name.jpg corresponds to name_mask.jpg. The dataset is organized into Synthetic and Real-world domains, each with four condition subfolders (sunny, cloudy, rainy, and night) containing paired images/ and masks/ directories.

The synthetic dataset contains 3,000 images generated using a fine-tuned Stable Diffusion model. It comprises 1,500 standard images created with the prompt "a crosswalk image", and 1,500 condition-specific images generated using environmental modifiers (sunny, cloudy, rainy, night). These subsets contain 600 sunny, 300 cloudy, 300 rainy, and 300 night samples.

The real-world dataset consists of 300 FPV frames extracted from chest-mounted smartphone recordings in urban environments: 120 sunny, 60 cloudy, 60 rainy, and 60 night. Each physical crosswalk contributes at most two images captured from different approach directions, ensuring scene diversity while avoiding redundancy.

Both synthetic and real images were annotated via a polygon-based interface, where annotators defined the crosswalk using four-point quadrilaterals. Masks employ white (255) for crosswalk regions and black (0) for background, ensuring consistent supervision across domains. A separate Fine-tuning seed images folder includes the 150 real FPV images used to fine-tune the diffusion model, supporting transparent and reproducible synthetic data generation.

The overall distribution of images across domains and conditions is summarized in Table 1, while Figure 1 illustrates representative image–mask pairs for cloudy, night, sunny, and rainy conditions, highlighting the appearance variations that our ConvNeXt-enhanced DeepLabv3 (CEDL) model must handle.

Table 1. Dataset composition across domains and environmental conditions.

Domain	Condition	Images	Description
Synthetic	Standard	1500	Base prompt ("a crosswalk image")
	Sunny	600	Prompt with sunny weather modifier
	Cloudy	300	Prompt with cloudy weather modifier
	Rainy	300	Prompt with rainy weather modifier
	Night	300	Prompt with night-time modifier
Real-world	Sunny	120	FPV frames from chest-mounted camera
	Cloudy	60	Real cloudy-condition samples
	Rainy	60	Real rainy-condition samples
	Night	60	Real night-condition samples
Fine-tuning Seeds	–	150	Real images used to fine-tune SD model

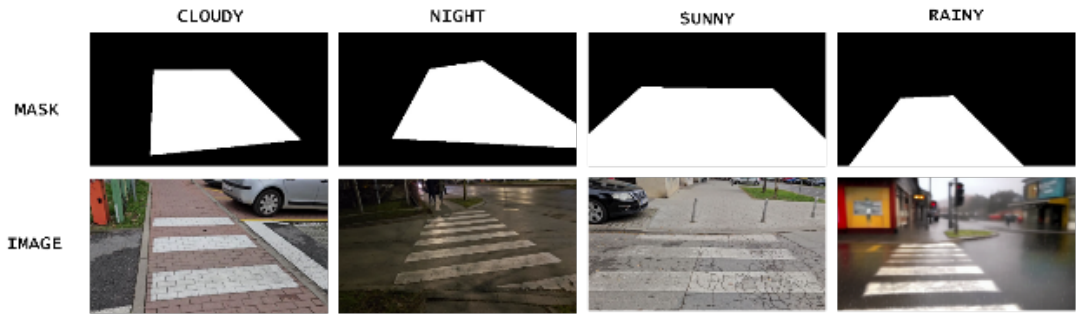

Figure 1. Examples of image–mask pairs under cloudy, night, sunny, and rainy conditions (top: masks; bottom: images).

3.2 Preprocessing and Data Augmentation

All images in the FPVCrosswalk2025 dataset are provided at a fixed resolution of 640×360 pixels, eliminating the need for spatial resizing during preprocessing [18]. The preprocessing pipeline therefore focuses solely on appearance-based normalization and augmentation. For the training set, brightness-contrast modulation ($p = 0.5$) and random rotation within $\pm 15^{\circ}$ ($p = 0.5$) are applied to simulate illumination fluctuations and natural viewpoint perturbations encountered in first-person navigation scenarios. All images are then normalized using ImageNet mean and standard deviation to align with ConvNeXt's pretrained feature space, followed by tensor conversion [20].

The validation and test sets undergo deterministic preprocessing consisting only of normalization and tensor conversion, ensuring unbiased performance evaluation.

The dataset is partitioned into 70% training, 15% validation, and 15% test splits using a fixed random seed (42) for reproducibility. The split is performed by randomly shuffling all images regardless of their synthetic or real origin and weather condition, without applying any stratification. All subsets are loaded using a batch size of 16 for all experiments. Because the split is performed over the combined pool, each split (train/val/test) contains both synthetic and real-world images, with synthetic samples being the majority (3000 vs 300). Unless stated otherwise, all quantitative results in this paper are computed on the mixed-domain test split (synthetic + real). For transparency, qualitative examples in Figures 7–8 are selected from the real-world portion of the test split to reflect realistic FPV conditions.

3.3 Model Architecture Overview

DeepLabv3 is employed as the core segmentation architecture due to its ability to preserve spatial detail while capturing long-range contextual information through atrous (dilated) convolution [5]. Unlike conventional CNNs that repeatedly downsample feature maps using pooling or strided convolutions, DeepLabv3 replaces late-stage downsampling with atrous convolutions, enabling control over the output stride and allowing the model to maintain high-resolution representations crucial for delineating thin crosswalk markings. The atrous convolution operation enlarges the receptive field by spacing kernel elements according to a dilation rate r. For an input feature map x and convolution kernel w, the output feature at location i is computed using atrous convolution. The atrous convolution operation enlarges the receptive field by spacing kernel elements according to a dilation rate r. The mathematical operation of the atrous convolution is expressed in equation (1).

$$\mathrm{y(i)} = \sum_{\mathrm{k}\in\mathcal{K}} \mathrm{w(k)}\ \mathrm{x(i} + r\,\mathrm{k)} \qquad (1)$$

where x is the input feature map, y is the output feature map, w is the convolution kernel, i denotes a 2D spatial location in the output feature map, $\mathrm{k} \in \mathcal{K}$ indexes kernel offsets (e.g., $\mathcal{K} = \{-1,0,1\}^2$ for a 3×3 kernel), and r is the atrous (dilation) rate that controls the spacing between sampled input locations. The term $\mathrm{i} + r\mathrm{k}$ denotes the dilated sampling position in the input feature map.

Equation (1) expands the effective field of view without increasing parameter count or computational burden, allowing the network to capture broader context while retaining spatial detail. This characteristic is particularly beneficial in first-person-view crosswalk segmentation, where stripe visibility is frequently affected by scale variation, perspective distortion, and adverse lighting conditions.

To further improve multi-scale representation, DeepLabv3 incorporates the Atrous Spatial Pyramid Pooling (ASPP) module, which applies multiple parallel atrous convolutions using different dilation rates alongside an image-level

pooling branch as shown in Figure 2. The resulting feature maps encode complementary contextual information: smaller dilation rates capture fine structural boundaries, while larger rates incorporate broader semantic cues. These multi-scale outputs are concatenated and projected through a 1×1 convolution before being upsampled to the original spatial resolution.

DeepLabv3 is inherently backbone-agnostic, allowing seamless integration with various encoders. Throughout this paper, ConvNeXt-Enhanced DeepLabv3 (CEDL) denotes DeepLabv3 equipped with a ConvNeXt-Tiny encoder, unless stated otherwise. In this study, DeepLabv3 + ResNet-50 and DeepLabv3 + MobileNetV3-L are used as baseline configurations, while CEDL is evaluated as the proposed model. All backbones supply hierarchical feature maps to the ASPP head without modifying the DeepLabv3 segmentation head.

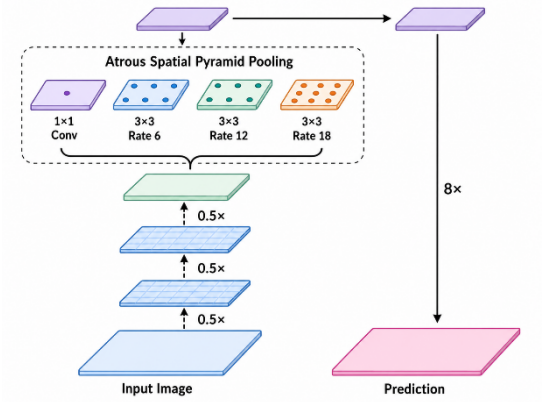

Figure 2. General architecture of DeepLabv3 showing feature extraction, spatial pyramid pooling, and final upsampling. (adapted from Chen et al. [5]).

The targeted failure modes in FPV crosswalk scenes include (i) glare and specular reflections that suppress local contrast, (ii) heavy shadows and non-uniform illumination that distort stripe appearance, (iii) wet surfaces that introduce texture-like noise patterns, and (iv) worn markings that fragment stripe boundaries. CEDL mitigates these issues through complementary mechanisms: ConvNeXt-Tiny improves local texture discrimination and boundary sensitivity via large-kernel depthwise spatial mixing and stable LayerNorm-based feature normalization, reducing sensitivity to photometric shifts; ASPP provides multi-scale aggregation that preserves thin stripe geometry at small dilation while leveraging broader context at larger dilation to maintain continuity when markings are partially occluded or faded; and appearance-focused augmentation (brightness/contrast modulation) exposes the model to illumination variability during training, improving generalization under glare, shadows, and wet-surface reflections. This design rationale is consistent with the observed improvements in boundary-preservation metrics (Boundary IoU) and the qualitative reduction of boundary fragmentation and missed stripe segments under rainy and night conditions.

3.4 Backbone Architectures

DeepLabv3 is a backbone-agnostic architecture, meaning that its performance is highly dependent on the quality of the encoder that supplies hierarchical feature representations. In dense prediction tasks such as crosswalk segmentation under adverse visual conditions, backbone selection affects spatial precision, contextual capacity, and computational efficiency. This study evaluates three backbones ResNet-50, MobileNetV3, and ConvNeXt-Tiny representing standard, efficiency-oriented, and modernized convolutional designs, respectively.

ResNet-50 is employed as the standard baseline due to its stable optimization behavior and strong hierarchical feature extraction capability. The architecture is built around residual bottleneck blocks, where each block learns a residual function that is added back to its input through an identity shortcut. This residual mapping can be expressed mathematically as shown in equation (2).

$$y = x + F(x; W) \qquad (2)$$

When dimensions differ, a projection W_s is used: $y = F(x; W) + W_s x$, where x is the input feature tensor to a residual block, y is the output tensor, $F(\cdot; W)$ denotes the residual mapping implemented by the block's stacked layers (e.g., convolutions, normalization, and nonlinearity), and W represents the learnable parameters of the residual branch. The skip connection $\mathbf{x}$ (identity shortcut) facilitates gradient flow and stabilizes optimization.

The skip connection in equation (2) preserves gradient flow and stabilizes optimization, which is beneficial for training deep encoders. As illustrated in Figure 3, the ResNet bottleneck block comprises a 1×1 convolution for channel reduction, a 3×3 convolution for spatial feature extraction, and a 1×1 convolution for channel expansion. These layers are typically paired with Batch Normalization and non-linear activation (with the final activation applied after the residual addition), enabling the network to capture both local spatial cues and higher-level semantics. This design makes ResNet-50 a widely adopted backbone for segmentation tasks.

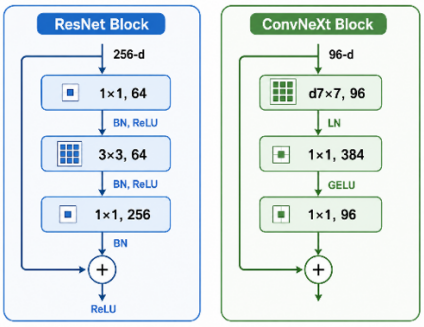

Figure 3. Comparison between the ResNet bottleneck block and the ConvNeXt block, illustrating the structural differences in normalization strategy, kernel design, and channel expansion used in the backbone architectures.

However, the architecture relies heavily on aggressive downsampling and large channel widths. These factors increase computational cost and may produce overly coarse representations, which can impair the delineation of thin crosswalk stripes under low contrast or adverse illumination. Despite this limitation, ResNet-50 serves as a strong and reliable baseline for evaluating modern backbone designs.

MobileNetV3 is used as the efficiency-oriented baseline due to its low computational footprint and hardware-aware design. It employs inverted residual blocks with depthwise separable convolutions, squeeze-and-excite (SE) modules, and the h-swish non-linearity to maximize accuracy per FLOP. As illustrated in Figure 4, each MobileNetV3 block begins with a pointwise 1×1 expansion, followed by a lightweight 3×3 depthwise convolution, and finally a 1×1 projection layer. The computational advantage of depthwise separable convolution can be understood from its cost formulation shown in equation (3).

$$FLOPs_{DW} = K^2HWC + HWC \tag{3}$$

Where K is the kernel size, $H \times W$ is the spatial dimensionality, and C is the number of channels. This reduction allows MobileNetV3 to maintain real-time performance on resource-constrained hardware while still preserving spatial structure important for segmentation.

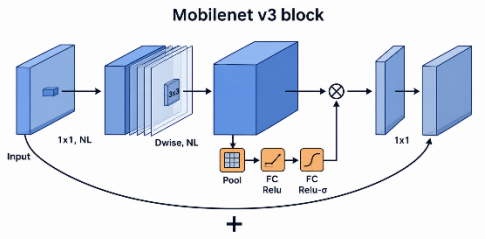

Figure 4. Structure of a MobileNetV3 inverted residual block. (adapted from Howard et al. [11]).

A key element of MobileNetV3 is the Squeeze-and-Excite (SE) attention module, which performs channel-wise feature recalibration. First, global average pooling computes a channel descriptor as shown in equation (4).

$$z_c = \frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W} x_c(i,j) \tag{4}$$

This descriptor is passed through two fully connected layers with non-linear activations to generate the excitation weights, as written in equation (5).

$$s = \sigma\big(W_2, \delta(W_1 z)\big) \tag{5}$$

The recalibrated output is then obtained through channel-wise multiplication, as shown in equation (6)

$$\widehat{x_c} = s_c \cdot x_c \tag{6}$$

The SE mechanism in (4)–(6) enhances channel discrimination, enabling MobileNetV3 to emphasize informative features such as stripe edges or high-contrast transitions while suppressing irrelevant noise in adverse visual conditions. Additionally, the use of h-swish activation improves non-linearity efficiency, supporting deeper expressiveness without increasing inference cost. However, the compact channel depth of MobileNetV3 restricts its representational capacity, which may limit performance when segmenting thin or low-contrast crosswalk markings. For this reason, MobileNetV3 serves as an efficiency benchmark rather than the primary backbone for robustness-driven segmentation tasks.

ConvNeXt-Tiny is adopted as the proposed backbone due to its modernized convolutional design, which incorporates several architectural refinements inspired by Vision Transformers while retaining the efficiency and inductive priors of CNNs. As illustrated in Figure 3, the ConvNeXt block replaces the classic ResNet bottleneck with a streamlined sequence consisting of a large-kernel depthwise convolution, Layer Normalization, a pointwise channel expansion, a Gaussian Error Linear Unit (GELU) activation, and a pointwise projection back to the original dimensionality.

The block begins with a 7×7 depthwise convolution, which performs spatial mixing independently across channels. This operation enlarges the effective receptive field more smoothly than dilated convolution and can be expressed as shown in equation (7).

$$u_c(i,j) = \sum_{(m,n)\in\Omega_{7\times7}} x_c(i+m, j+n)\,, w_c(m,n) \tag{7}$$

Here, x_c and u_c denote the input and output of channel c, and w_c represents channel-wise depthwise filters. This formulation preserves computational efficiency while enabling the model to capture long-range structural cues essential for crosswalk stripe segmentation under adverse weather. Following depthwise convolution, Layer Normalization (LN) stabilizes channel statistics and improves convergence when training with small batch sizes an important property in high-resolution segmentation. The transformation applied to each vectorized feature u is shown in equation (8).

$$\mathrm{LN}(u) = \frac{u-\mu}{\sqrt{\sigma^2+\epsilon}} \cdot \gamma + \beta \tag{8}$$

where μ and σ^2 are per-sample statistics, and γ, β are learnable affine parameters. LN provides more stable behavior than BatchNorm under batch-size constraints often encountered in FPV segmentation training.

Next, ConvNeXt applies a pointwise expansion–projection MLP, which increases the channel dimensionality by a factor of four before projecting it back to the original dimension. This channel mixing process is expressed as shown in equation (9).

$$y = W_2,\mathrm{GELU}(W_1, v) \tag{9}$$

where $v = \mathrm{LN}(u)$, W_1 expands the channel width ($4C_{\text{in}}$), and W_2 projects it back to C_{in}. The GELU activation enhances nonlinearity with smoother gradients, improving the model's ability to represent subtle texture and boundary variations in challenging crosswalk scenes.

Due to its hierarchical four-stage design (with channel dimensions 96–192–384–768 in the Tiny variant), ConvNeXt generates progressively richer mid-level and high-level semantic features compared to ResNet-50 and MobileNetV3. The combination of large-kernel depthwise convolution, stable normalization, and expressive MLP-style transformations allows ConvNeXt-Tiny to better capture elongated stripe boundaries, worn markings, and low-contrast patterns even under rain, nighttime glare, and motion blur. This enhanced representational capacity makes ConvNeXt-Tiny the most robust and best-performing encoder among the evaluated backbones. Moreover, its full compatibility with the DeepLabv3 ASPP and decoder pipeline enables seamless integration without modifying the segmentation head.

3.5 Proposed ConvNeXt-Enhanced DeepLabv3 (CEDL)

The proposed architecture integrates ConvNeXt-Tiny as the backbone of DeepLabv3, forming a unified segmentation framework designed to enhance robustness under adverse illumination and weather conditions. As illustrated in Figure 5, the model combines a hierarchical ConvNeXt encoder with the DeepLabv3 segmentation head equipped with multi-scale Atrous Spatial Pyramid Pooling (ASPP). In this work, leveraging modern convolutional modeling refers to using ConvNeXt's redesigned convolutional blocks that improve feature expressiveness while preserving strong spatial inductive bias, properties that are beneficial for thin and repetitive crosswalk stripes that require sharp boundary sensitivity. Specifically, large-kernel depthwise convolutions expand the effective receptive field at low cost, while Layer Normalization, GELU activation, and inverted-bottleneck channel mixing stabilize training and enhance discrimination of faint stripe textures. In combination with ASPP's multi-dilation aggregation, the model preserves sharp local boundaries while incorporating broader contextual cues across crosswalk scales and perspectives.

ConvNeXt-Tiny first processes the input image through four sequential stages that progressively increase the channel dimensionality (96–192–384–768) while reducing the spatial resolution. Each stage consists of ConvNeXt blocks employing large-kernel 7×7 depthwise convolutions to perform spatial mixing as described in equation (7), followed by Layer Normalization as given in equation (8), a 4× channel expansion with GELU activation, and a final 1×1 projection as formalized in equation (9). This architectural refinement allows ConvNeXt to capture fine structural details while modeling long-range spatial patterns supporting better feature expressiveness for thin-structure segmentation. Such capabilities are essential for first-person-view crosswalk segmentation, where markings may appear faint, irregular, or partially occluded under challenging environmental conditions, including glare, shadows, rain, or nighttime illumination.

The high-level feature representation produced by ConvNeXt is then processed by the DeepLabv3 ASPP module, which applies parallel atrous convolutions with different dilation rates and incorporates an image-level pooling branch as illustrated by Figure 5. These multi-scale pathways enhance the model's ability to handle variations in crosswalk width, perspective distortion, and global scene context. The outputs of the atrous branches

are concatenated to form the aggregated multi-scale representation based on equation (10).

$$f_{\text{ASPP}} = \text{Concat}\left(f_{1\times1},, f_{r_1},, f_{r_2},, f_{r_3},, f_{\text{img}}\right) \quad (10)$$

This concatenated feature is projected into a single-channel segmentation logit map through a 1×1 convolution, as expressed in equation (11):

$$o = W_{\text{cls}}, f_{\text{ASPP}} \quad (11)$$

By unifying ConvNeXt's modernized convolutional blocks with DeepLabv3's multi-scale ASPP segmentation head, the proposed model achieves both robustness and computational efficiency. ConvNeXt provides stable and expressive encoder features through enlarged receptive fields and reliable normalization, while ASPP contributes strong multi-scale discrimination crucial for crosswalk geometry understanding. The resulting synergy produces a segmentation system well-suited for first-person-view scenarios where environmental variability significantly affects boundary visibility and prediction stability.

3.6 Training Configuration

All experiments are conducted using three backbone variants ConvNeXt-Tiny, ResNet-50, and MobileNetV3 which are integrated into the DeepLabv3 segmentation framework. Each backbone is initialized with ImageNet-1K pretrained weights, and all layers are fully fine-tuned to adapt to the FPV crosswalk segmentation domain. The training is performed on an NVIDIA T4 GPU system (15 GB VRAM, 12.7 GB system RAM), providing a consistent environment across all model variants.

Training is executed for 25 epochs with a batch size of 16 using the AdamW optimizer (learning rate = 3×10^{-4}). To mitigate severe pixel-level class imbalance in binary crosswalk segmentation, where foreground crosswalk pixels occupy only a small fraction of the image compared to background, we use a hybrid loss combining Binary Cross-Entropy and Dice loss. The hybrid loss is formulated as $L = \alpha L_{BCE} + (1-\alpha)L_{Dice}$, with $\alpha = 0.5$ used consistently across all experiments. This equal weighting was selected to balance the complementary strengths of the two loss components: BCE supports stable pixel-wise optimization and classification accuracy, while Dice loss enhances region-overlap learning and reduces the negative effect of foreground–background imbalance. Similar balanced combinations of cross-entropy-based and overlap-based objectives have been reported to improve performance in class-imbalanced segmentation tasks [25]. Since the main objective of this study is to compare backbone architectures rather than optimize loss-weight hyperparameters, the same loss configuration is applied to all models to ensure a fair and controlled comparison. To improve computational efficiency, mixed-precision training using torch.cuda.amp is employed to reduce memory usage and accelerate training. During inference, the predicted logits are passed through a sigmoid activation function and thresholded at 0.5 to generate the final binary segmentation masks.

During each epoch, predictions are evaluated on the validation split to compute IoU (Jaccard Index) and Dice Score, which serve as the primary segmentation metrics. The model achieving the highest validation IoU is stored and later restored as the final checkpoint. A fixed random seed (42) is applied to ensure determinism in data splitting, augmentation, and loader sampling.

This training configuration is applied uniformly across the three backbone variants, ensuring that performance differences reflect architectural characteristics rather than inconsistencies in the optimization process.

3.7 Evaluation Metrics

Model performance is assessed using a combination of region-based, boundary-based, and system-level metrics to comprehensively characterize segmentation quality and computational efficiency. The primary metrics monitored during training and validation are Intersection over Union (IoU) and Dice Score, both computed on binary masks obtained by thresholding the predicted logits at 0.5. IoU quantifies the overlap between the predicted mask $P$and the ground truth Gas shown in equation (12).

$$\text{IoU} = \frac{|P \cap G|}{|P \cup G|} \quad (12)$$

The Dice Score emphasizes agreement in foreground regions—useful in class-imbalanced scenarios—and is defined in equation (13).

$$\text{Dice} = \frac{2|P \cap G|}{|P| + |G|} \quad (13)$$

For final testing and comparative analysis, additional metrics are incorporated. Boundary IoU evaluates mask quality near object boundaries by restricting the IoU computation to a narrow band around the ground-truth contour. Let B_P and B_G denote boundary regions of prediction and ground truth; Boundary IoU is computed as equation (14).

$$\text{Boundary IoU} = \frac{|B_P \cap B_G|}{|B_P \cup B_G|} \quad (14)$$

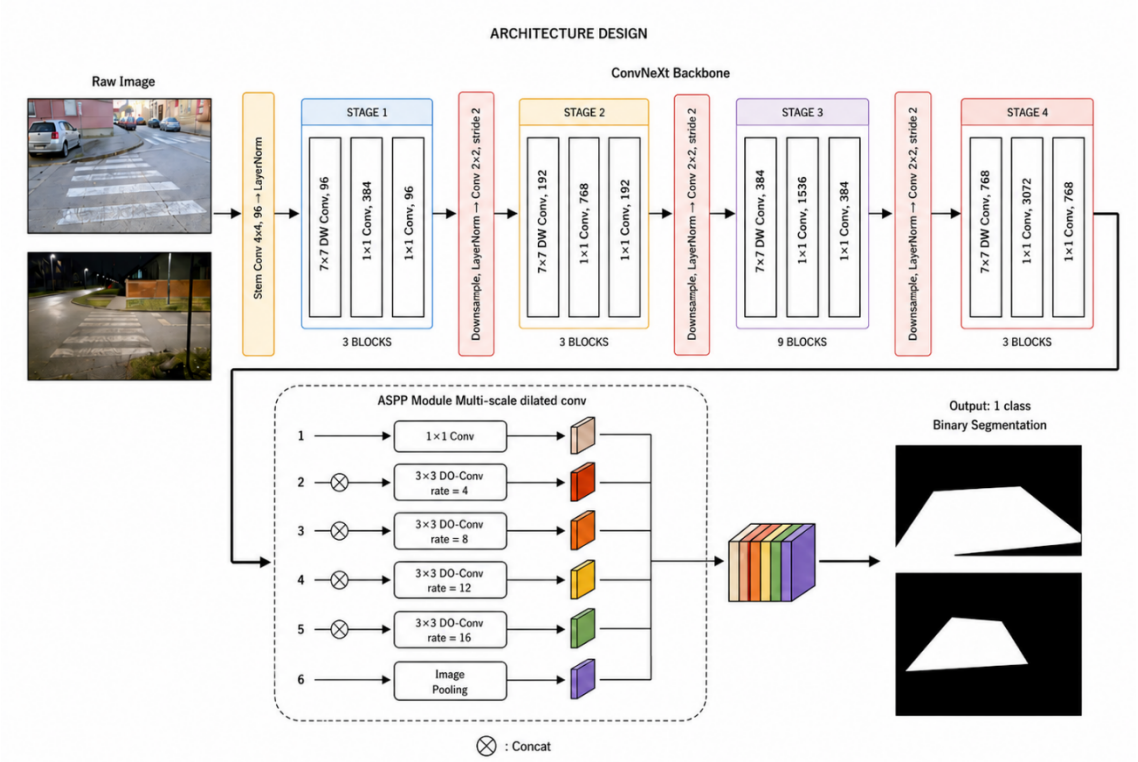

Figure 5. Overall pipeline of the proposed ConvNeXt-Enhanced DeepLabv3 (CEDL) model, consisting of a hierarchical ConvNeXt-Tiny backbone and a multi-scale DeepLabv3 ASPP head producing 1-class binary segmentation outputs.

This metric captures the ability to preserve crosswalk stripe edges, which often degrade under night glare, motion blur, shadows, or wet surfaces. False Negative Rate (FNR) measures the proportion of ground-truth crosswalk pixels that are missed by the model, as written in equation (15).

$$\mathrm{FNR} = \frac{|G \setminus P|}{|G|} \tag{15}$$

A low FNR is essential for safety-critical applications, where missing crosswalk regions is more problematic than oversegmenting them.

To assess computational efficiency and deployment feasibility, system-level metrics are reported. Inference Time is computed as the average forward-pass latency per image, while Peak VRAM Usage records the maximum GPU memory consumption during inference. Parameter Count characterizes model complexity and enables fair comparisons across different backbone–head configurations. Together, these metrics evaluate not only segmentation accuracy but also the operational practicality of deploying the model on real-world, resource-constrained devices.

In addition to mean IoU, we report Min IoU (worst-case) to characterize model behavior on the most challenging samples. Min IoU is defined as the minimum per-image IoU across all test images: IoU is computed independently for each test image by comparing the predicted mask to its corresponding ground-truth mask, and the lowest value is selected. This metric is therefore computed at the single-sample level, not over batches. While Min IoU is not intended to represent typical performance and may be sensitive to outliers, it serves as a robustness-oriented indicator that highlights failure severity under extreme visual conditions (e.g., strong glare, heavy reflections, or very low contrast).

4. Results and Discussion

This section presents and discusses the experimental results, including quantitative performance, robustness under adverse conditions, computational efficiency, error analysis, and the overall implications of the proposed model.

4.1 Quantitative Results

All metrics reported in Tables 2–5 are evaluated on the mixed-domain test split (synthetic + real-world). Real-world behavior is additionally illustrated using qualitative examples selected from the real-world portion of the test split (Figures 7–8). We compare the proposed method against two standard CNN backbones used as segmentation

baselines: ResNet-50 [10] and MobileNetV3-Large [11]. Both baselines are trained and evaluated under the same protocol as the proposed method. The quantitative performance of all backbone variants is summarized in Table 2, reporting final test-set metrics including mean IoU, Dice, Precision, Recall, and Min IoU (worst-image IoU). The proposed CEDL achieves the highest overall segmentation accuracy (mean IoU = 0.946; Dice = 0.972) and the most balanced precision–recall trade-off, whereas ResNet-50 attains slightly higher recall but lower precision, indicating a tendency to over-segment crosswalk regions. While MobileNetV3-L yields the highest Min IoU, its mean IoU is the lowest, suggesting that its average performance is less competitive despite a comparatively better worst-image case. Min IoU (worst-case) is reported as a robustness-oriented indicator to expose failure severity on the hardest samples and to complement mean metrics in safety-relevant segmentation analysis.

Table 2. Final test metrics across backbone variants.

Model	Mean IoU	Mean Dice	Precision	Recall	Worst-case IoU
Proposed CEDL	**0.946**	**0.972**	**0.972**	0.972	0.467
ResNet-50	0.942	0.970	0.966	**0.974**	0.439
MobileNetV3-L	0.936	0.967	0.962	0.972	**0.512**

Although the mean IoU and Dice improvements are numerically around 1–2%, their practical impact should be interpreted in the context of thin-structure segmentation, where small pixel-level differences can translate into boundary breaks, stripe discontinuities, or missed segments under adverse illumination. Therefore, modest average gains may correspond to operationally meaningful improvements in crosswalk mask integrity, particularly when supported by robustness-focused indicators such as worst-case IoU, per-condition IoU, Boundary IoU, and qualitative comparisons. Although formal statistical significance testing was not performed, the observed improvements are consistently reflected across multiple evaluation criteria, including boundary-level robustness, adverse-condition performance, worst-case behavior, and visual error analysis.

Further insight into optimization behavior is illustrated in Figure 6, which shows the training and validation loss curves for all models. CEDL demonstrates the most stable convergence, with smooth validation loss trends after epoch 10 and minimal divergence from the training curve, indicating strong generalization capability. ResNet-50 exhibits moderate fluctuations in validation loss throughout the training process, suggesting sensitivity to illumination and texture variations. MobileNetV3-L converges rapidly during early epochs due to its lightweight architecture but exhibits persistent noise in validation loss, reflecting limited representational capacity for fine-grained crosswalk structures. Together, these trends confirm that the proposed ConvNeXt-based model consistently maintains high performance while preserving training stability.

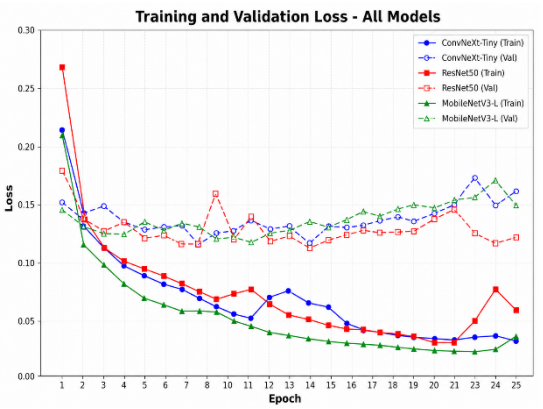

Figure 6. Training and validation loss curves for all backbone variants.

4.2 Robustness Against Adverse Conditions

Environmental robustness was evaluated using per-condition IoU across five subsets, namely sunny, cloudy, rainy, night, and standard, as summarized in Table 3. As shown in the table, the CEDL achieves the strongest performance under cloudy (0.947), rainy (0.941), and standard conditions (0.948), while maintaining competitive results under night (0.938) and sunny (0.931) scenes. These trends indicate that the large-kernel depthwise convolutions and enhanced normalization in ConvNeXt enable the model to preserve crosswalk geometry even when exposed to low-contrast illumination, reflections, or weather-induced distortions. ResNet-50 performs similarly on cloudy and rainy inputs but shows more pronounced degradation in night and sunny conditions, whereas MobileNetV3-L exhibits the largest performance drop in low-light scenarios due to its reduced channel capacity.

Table 3. Per-condition IoU across environmental conditions.

Condition	CEDL	ResNet-50	MobileNetV3
cloudy	**0.947**	0.942	0.928
night	**0.938**	0.930	0.908
rainy	**0.941**	0.941	0.929
standard	**0.948**	0.944	0.944
sunny	**0.931**	0.929	0.918

Boundary-level robustness is further assessed through Boundary IoU and False Negative Rate

(FNR), reported in Table 4. CEDL attains the highest Boundary IoU (0.073), demonstrating its superiority in preserving thin crosswalk edges under glare, shadow boundaries, and uneven road textures. Although ResNet-50 yields the lowest FNR (0.026), the difference relative to CEDL (0.028) is marginal and does not offset its weaker boundary preservation. MobileNetV3-L shows the largest degradation in both metrics, confirming that lightweight architectures struggle to maintain spatial fidelity in complex scenes.

Table 4. Boundary-level robustness metrics.

Model	Boundary IoU	FNR
CEDL	**0.073**	0.028
ResNet-50	0.072	**0.026**
MobileNetV3-L	0.064	0.028

The robustness trends observed numerically are further illustrated through qualitative comparisons in Figure 7 and Figure 8, which display segmentation results for sunny–rainy and cloudy–night conditions, respectively. As evident from the figures, CEDL produces the most stable and continuous crosswalk masks across all weather scenarios. In sunny conditions, where strong shadows and specular highlights often disrupt edge continuity, CEDL maintains consistent stripe boundaries, whereas MobileNetV3 frequently undersegments bright regions. Under rainy scenes, CEDL demonstrates resilience to reflections and wet-surface distortions, avoiding the fragmented outputs observed in the MobileNetV3 baseline. For cloudy scenes, both CEDL and ResNet-50 maintain high-quality predictions, though MobileNetV3 occasionally misidentifies painted road artifacts. The night condition presents the most severe challenge; nevertheless, CEDL retains structural consistency while the baselines exhibit substantial misalignment and false positives. These qualitative findings align with the numerical metrics and confirm that the proposed model achieves superior robustness across adverse environmental conditions.

4.3 Efficiency Evaluation

Computational efficiency was assessed using inference time, peak VRAM consumption, and parameter count across the three backbone variants. The quantitative summary is presented in Table 5, which highlights the trade-off between model complexity and real-time feasibility. As shown in the table, MobileNetV3-L achieves the fastest inference time at 11.36 ms per image and the lowest VRAM usage (383.9 MB), consistent with its lightweight design and reduced channel depth. ResNet-50, in contrast, exhibits the slowest inference time (104.47 ms) and the highest memory footprint, reflecting its large bottleneck layers and BatchNorm-heavy structure. The proposed CEDL strikes a favorable balance, achieving a fast inference time of 20.61 ms—nearly 5× faster than ResNet-50—while maintaining a moderate VRAM footprint of 404.98 MB.

Table 5. Efficiency metrics of all backbone variants.

Model	Inference Time (ms/img)	Peak VRAM (MB)	Parameter Count
CEDL	20.608	404.98	34,441,569
ResNet-50	104.475	450.69	41,998,934
MobileNetV3-L	11.357	383.89	11,024,188

Despite having a higher parameter count than MobileNetV3-L, CEDL demonstrates significantly stronger efficiency–accuracy trade-offs. The architecture's large depthwise kernels and LayerNorm-based normalization reduce memory overhead during forward propagation, enabling faster execution than ResNet-50 even with comparable parameter magnitude. Importantly, CEDL achieves the highest segmentation accuracy while remaining viable for real-time or near–real-time deployment scenarios, especially in FPV settings where processing each frame within 30–40 ms is typically sufficient for pedestrian navigation or driver-assistance systems. These findings indicate that CEDL delivers not only improved robustness and segmentation quality but also practical computational efficiency suitable for edge devices with moderate GPU capability.

4.4 Error Analysis

To provide a deeper understanding of model behavior under varying levels of visual difficulty, qualitative predictions across all environmental conditions are examined using Figure 7 and Figure 8. Each figure contains three representative samples per condition best-case, typical-case, and worst-case predictions arranged from top to bottom. These samples were selected based on per-image IoU ranking and allow a detailed inspection of how each backbone responds to clean scenes, moderately challenging inputs, and severe visual degradation.

In the best-case examples across sunny, rainy, cloudy, and night scenes, all models generally produce high-quality masks, with CEDL achieving the most accurate boundary alignment and the least edge erosion. ResNet-50 also performs reliably, whereas MobileNetV3-L tends to produce slightly thinner or partially eroded stripe boundaries in

regions with strong brightness transitions or fine texture details. These best-case cases reflect scenarios where environmental conditions pose minimal interference, and all backbones operate near their performance ceiling.

In the typical-case examples characterized by moderate lighting variations, mild reflections, or partial occlusions, differences between the backbones become more pronounced. MobileNetV3-L often undersegments crosswalk edges or misinterprets adjacent pavement patterns. ResNet-50 maintains structural coherence but shows small gaps near stripe borders when contrast decreases. CEDL remains the most stable, preserving global shape and edge continuity even when visual quality is reduced. These mid-level samples confirm the role of large depthwise kernels and hierarchical feature mixing in enhancing local and contextual consistency.

The worst-case examples, located in the bottom rows of each condition block, highlight the most critical failure modes. These often arise from intense specular reflections in rainy scenes, headlight glare in night scenes, cast shadows in sunny scenes, or subtle texture blending in cloudy scenes. In these difficult inputs, MobileNetV3-L frequently collapses, producing fragmented masks or hallucinating bright areas as crosswalks. ResNet-50 performs better but still struggles with structural collapse near thin stripe segments. CEDL, although not immune to severe degradation, retains better overall topology, exhibiting fewer false positives and maintaining partial crosswalk continuity. These observations align with the worst-case IoU values reported earlier and reinforce CEDL's superior resilience to extreme visual interference.

Overall, Figures 7 and 8 demonstrate that while all models perform adequately in clean scenes, only CEDL consistently preserves crosswalk geometry across varying levels of visual difficulty. The persistent failure patterns across baselines, such as fragmentation, boundary collapse, and reflection sensitivity, highlight the challenges inherent in FPV crosswalk segmentation, especially under adverse lighting and weather. The qualitative evidence thus supports the conclusion that the proposed CEDL provides the most robust and reliable segmentation performance among the evaluated architectures.

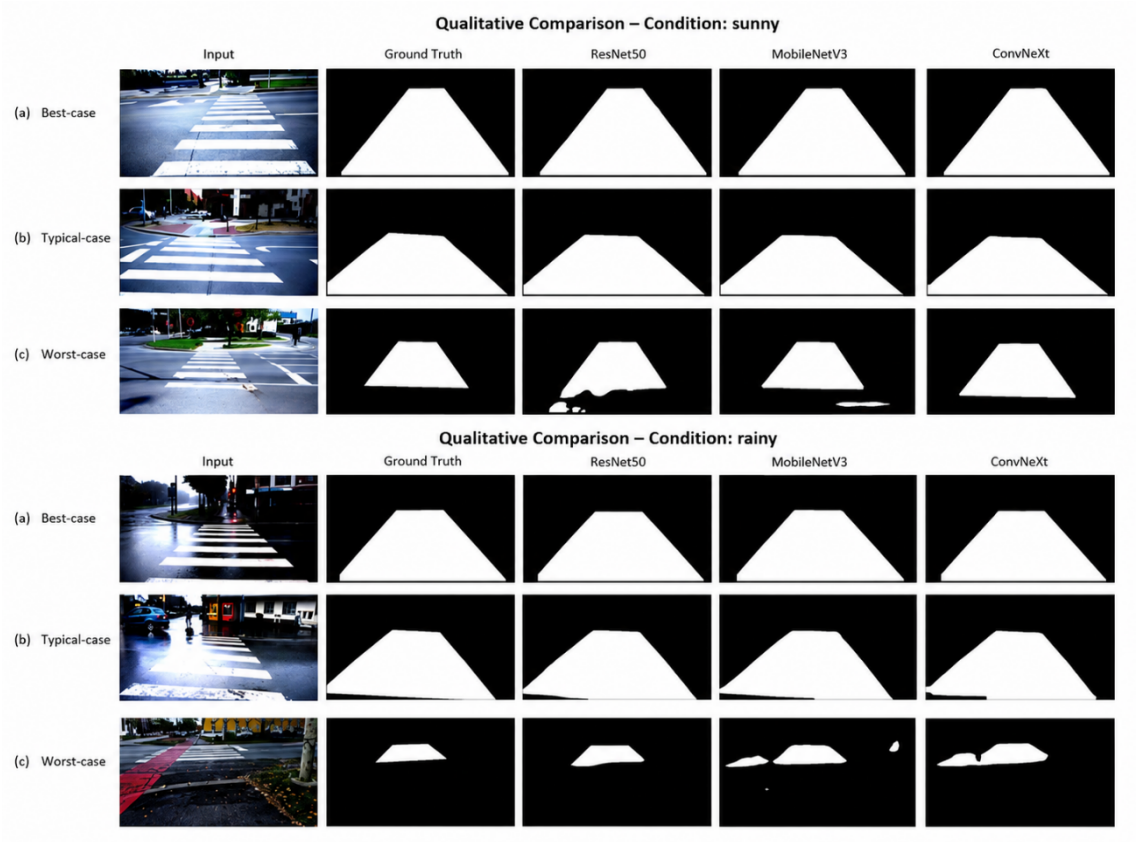

Figure 7. Qualitative comparison under sunny and rainy conditions: (a) Best, (b) Typical, and (c) Worst case.

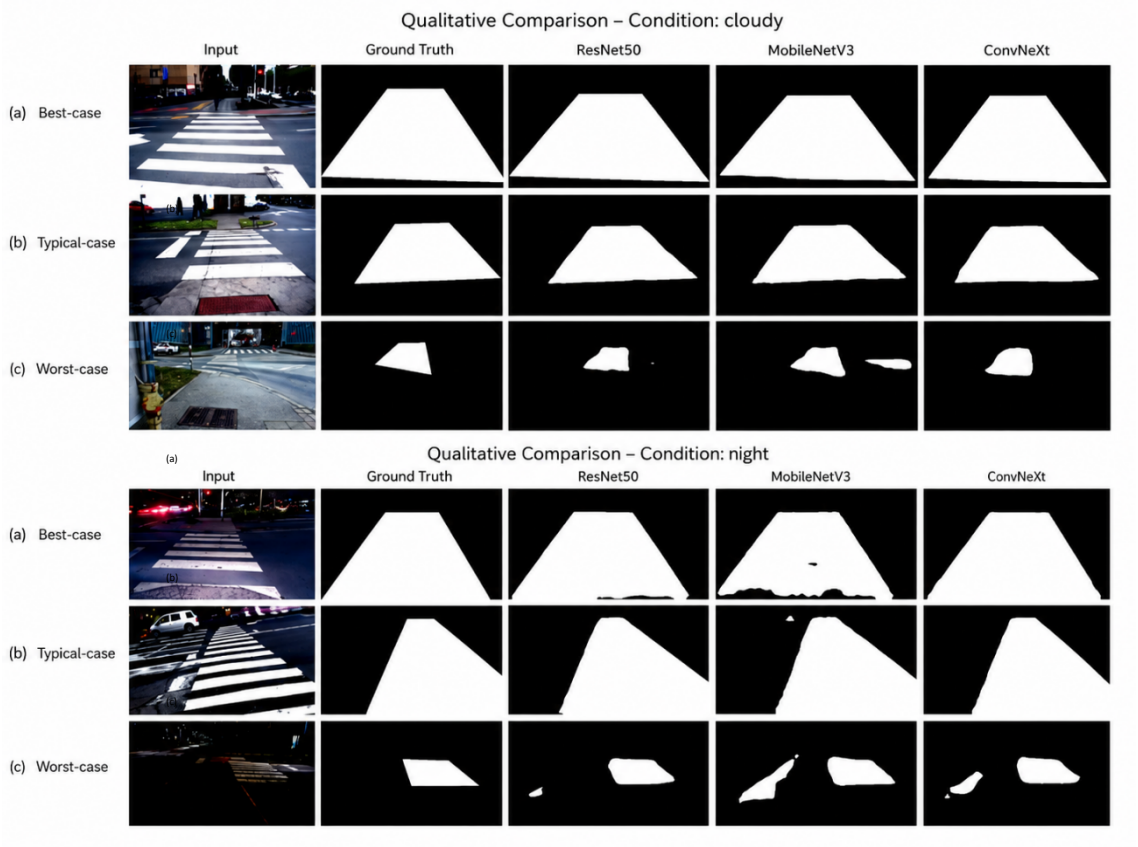

Figure 8. Qualitative comparison under cloudy and night conditions: (a) Best, (b) Typical, and (c) Worst case.

4.5 Discussion

The results collectively show that the CEDL offers a substantially more reliable solution for FPV crosswalk segmentation than both ResNet-50 and MobileNetV3-L, particularly in adverse environments. The most significant finding is the model's robustness across all environmental conditions, including sunny, cloudy, rainy, and night where it consistently maintains higher per-condition IoU and superior Boundary IoU. This outcome directly addresses the primary gap identified in earlier sections: the lack of segmentation-oriented evaluations under diverse weather and illumination settings. The ability of CEDL to preserve crosswalk geometry under low contrast, reflections, glare, and nighttime illumination indicates that its modern convolutional design produces feature representations that remain stable despite severe photometric distortion.

Another important insight concerns efficiency, where the results reveal a non-trivial relationship between parameter count and actual runtime performance. Although ConvNeXt-Tiny contains more parameters than MobileNetV3-L and a comparable scale to ResNet-50, it operates nearly five times faster than ResNet-50 while remaining well within real-time bounds (~20.6 ms per frame). This demonstrates that the architectural refinements in ConvNeXt, such as large depthwise kernels, simplified channel mixing, and LayerNorm-based normalization, reduce computational bottlenecks that traditionally slow down classical CNNs with similar parameter sizes. This finding challenges the common assumption that parameter count alone determines latency and highlights the importance of architectural modernity in achieving deployable segmentation models.

A third key observation is the accuracy–robustness–efficiency trade-off achieved by the proposed model. MobileNetV3-L remains the fastest backbone but collapses in challenging scenes due to its limited representational capacity. ResNet-50 produces more stable predictions than MobileNetV3-L but suffers from thin-structure degradation and significant inference delays, making it unsuitable for real-time FPV deployment. ConvNeXt, in contrast, provides accuracy comparable to heavy backbones, robustness superior to both baselines, and latency close to lightweight models, positioning it uniquely for safety-critical FPV navigation or driver-assistance systems.

The qualitative analyses further reinforce these insights by showing how failure modes differ across architectures. Under strong glare, heavy rain, or low-light conditions, MobileNetV3-L frequently undersegments or misclassifies crosswalk stripes, while ResNet-50 exhibits boundary collapse and spatial drift. CEDL degrades more gracefully, retaining partial topology and reducing false positives even in worst-case scenarios. This graceful degradation property is particularly important for real-world deployment, where unpredictable environmental variation is unavoidable.

Overall, the findings demonstrate that combining ConvNeXt with DeepLabv3's multi-scale ASPP module produces a robust segmentation framework capable of maintaining structural fidelity across diverse FPV conditions without sacrificing computational practicality. Remaining limitations, such as residual errors under extreme glare or nearly invisible road markings, point to potential future directions involving temporal feature stabilization, multimodal input (e.g., NIR or depth), or illumination-invariant representation learning. Nonetheless, the current results support CEDL as a strong and reliable backbone for crosswalk segmentation under adverse conditions on the mixed synthetic–real FPV benchmark, while further real-world-only evaluation remains necessary to confirm broader deployment robustness.

A domain-wise quantitative breakdown (synthetic-only vs real-only) may provide additional insight into real-world-only performance and domain generalization. However, given the limited size of the real-world subset and its distribution across multiple conditions, such estimates may be unstable when further split. We consider expanded real-world benchmarking and domain-wise reporting as future work.

5. Conclusion

This study presents a ConvNeXt-Enhanced DeepLabv3 (CEDL) architecture for crosswalk segmentation under adverse weather and illumination conditions, addressing a key gap in existing research where robust pixel-level segmentation across diverse environments has been largely overlooked. By integrating the ConvNeXt-Tiny backbone with DeepLabv3's ASPP module, the proposed model effectively combines large-kernel depthwise convolutions, stable LayerNorm-based normalization, and multi-scale contextual encoding to achieve superior representation of both global structure and fine boundary details.

Extensive experiments on the FPVCrosswalk2025 dataset demonstrate that the proposed model delivers the highest overall performance across all evaluated metrics, achieving 0.946 mean IoU and 0.972 Dice Score, while maintaining strong boundary preservation and resilience in challenging scenes such as heavy rain, low-light nighttime conditions, and high-contrast sunny environments. Compared to the ResNet-50 and MobileNetV3-L baselines, the ConvNeXt backbone consistently exhibits better robustness, fewer catastrophic failures, and more coherent predictions in worst-case scenarios, supporting its potential suitability for FPV crosswalk perception under real-world adverse conditions.

Despite its higher parameter count than lightweight backbones, the proposed model remains computationally practical, achieving 20.6 ms inference time and moderate VRAM usage, making it significantly faster than ResNet-50 and well within real-time constraints for FPV navigation or on-device driver-assistance systems. These findings highlight the strength of combining modern convolutional design with multi-scale segmentation frameworks to achieve both accuracy and efficiency.

Overall, this work establishes a high-performing and resilient solution for crosswalk segmentation under complex adverse conditions in a mixed synthetic–real FPV benchmark. Future research may extend this approach by incorporating temporal modeling for video-based FPV, exploring illumination-invariant feature learning, or leveraging multi-modal inputs such as depth or infrared information. Nonetheless, the present results clearly demonstrate that CEDL offers a reliable and efficient foundation for robust environmental perception in intelligent transportation and human-centered navigation systems.

Acknowledgement

The authors used a generative AI tool to assist with English language editing (grammar and clarity). All technical content, experiments, analyses, and conclusions were produced and verified by the authors, who take full responsibility for the manuscript.

References

[1] M. A. Malbog, "Mask R-CNN for pedestrian crosswalk detection and instance segmentation," in *Proc. 2019 IEEE 6th International Conference on Engineering Technologies and Applied Sciences (ICETAS)*, 2019, pp. 1–5, doi: 10.1109/ICETAS48360.2019.9117217.

[2] Z. Cao, X. Xu, B. Hu, and M. Zhou, "Rapid detection of blind roads and crosswalks by using a lightweight semantic segmentation network," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 10, pp. 6188–6197, 2021, doi: 10.1109/TITS.2020.2989129.

[3] M. Boudissa, H. Kawanaka, and T. Wakabayashi, "Surveying semantic segmentation models using traffic landmark dataset," in *Proc. 2022 Joint 12th International Conference on Soft Computing and Intelligent Systems and 23rd International Symposium on Advanced Intelligent Systems (SCIS&ISIS)*, 2022, pp. 1–7, doi: 10.1109/SCISISIS55246.2022.10002013.

[4] Z.-D. Zhang, M.-L. Tan, Z.-C. Lan, H.-C. Liu, L. Pei, and W.-X. Yu, "CDNet: A real-time and robust crosswalk detection network on Jetson Nano based on YOLOv5," *Neural Comput. Appl.*, vol. 34, pp. 10719–10730, 2022, doi: 10.1007/s00521-022-07007-9.

[5] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv:1706.05587*, 2017.

[6] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," in *Proc. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 11966–11976, doi: 10.1109/CVPR52688.2022.01167.

[7] H. Taha, H. El-Habrouk, W. Bekheet, S. El-Naghi, and M. Torki, "Pixel-level pavement crack segmentation using UAV remote sensing images based on the ConvNeXt-UPerNet," *Alexandria Eng. J.*, vol. 124, pp. 147–169, 2025, doi: 10.1016/j.aej.2025.03.072.

[8] M. N. Mahmud, M. K. Osman, A. P. Ismail, F. Ahmad, K. A. Ahmad, and A. Ibrahim, "Road image segmentation using unmanned aerial vehicle images and DeepLab V3+ semantic segmentation model," in *Proc. 2021 11th IEEE International Conference on Control System, Computing and Engineering (ICCSCE)*, 2021, pp. 1–6, doi: 10.1109/ICCSCE52189.2021.9530950.

[9] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2881–2890, doi: 10.1109/CVPR.2017.660.

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.

[11] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, Q. V. Le, and H. Adam, "Searching for MobileNetV3," in *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 1314–1324, doi: 10.1109/ICCV.2019.00140.

[12] Y. Zhang and X. Chen, "Lightweight semantic segmentation algorithm based on MobileNetV3 network," in *Proc. 2020 International Conference on Intelligent Computing and Automation Systems (ICICAS)*, 2020, pp. 429–433, doi: 10.1109/ICICAS51530.2020.00095.

[13] T. Kim and H. Shim, "Road semantic segmentation oriented dataset for autonomous driving," in *Proc. 2020 IEEE International Conference on Consumer Electronics–Asia (ICCE-Asia)*, 2020, pp. 1–3, doi: 10.1109/ICCE-Asia49877.2020.9277270.

[14] S. M. F. Hussain, S. M. Hamza, and A. Samad, "Image segmentation for autonomous driving using U-Net Inception," in *Proc. 2022 7th International Conference on Signal and Image Processing (ICSIP)*, 2022, pp. 426–431, doi: 10.1109/ICSIP55141.2022.9885809.

[15] X. Yang, Z. Dong, Y. Wang, J. Li, Q. Zhang, and J. Gao, "DHF-UNet: Enhancing blind road segmentation in complex environments via multi-scale feature extraction and fusion," *J. Real-Time Image Process.*, vol. 22, no. 2, art. no. 104, 2025, doi: 10.1007/s11554-025-01680-4.

[16] Y. Agarwal, A. Yadav, D. K. Dewangan, and K. Arya, "SR-SegDrive: Enhancing UNet-based semantic segmentation for autonomous driving with super-resolution GAN," in *Proc. 2025 International Conference on Networks and Cryptology (NETCRYPT)*, 2025, pp. 1699–1706, doi: 10.1109/NETCRYPT65877.2025.11102793.

[17] A. J. Rad, A. Monazzam, A. R. Showkatabad, and S. Safari, "Real-time crosswalk segmentation using FPGA-based spiking neural networks: A comparison of integrate-and-fire and Izhikevich hardware neural models," in *Proc. 2024 IEEE East-West Design & Test Symposium (EWDTS)*, 2024, pp. 1–6, doi: 10.1109/EWDTS63723.2024.10873691.

[18] K. Romić, H. Leventić, M. Habijan, and I. Galić, "Synthetic and real-world datasets for crosswalk segmentation under diverse weather and lighting conditions," *Data Brief*, vol. 61, art. no. 111755, 2025, doi: 10.1016/j.dib.2025.111755.

[19] Y. Zhang, Q. Zhang, and Y. Zhang, "Semantic segmentation of traffic scene based on DeepLabv3+ and attention mechanism," in *Proc. 2023 3rd International Conference on Neural Networks, Information and Communication Engineering (NNICE)*, 2023, pp. 542–548, doi: 10.1109/NNICE58320.2023.10105805.

[20] S. Yang, Z. Zhang, Y. Yin, T. Wang, and M. Wan, "CA-Seaformer: A road image segmentation algorithm based on improved Seaformer," in *Proc. 2025 IEEE 5th International Conference on Electronic Technology, Communication and Information (ICETCI)*, 2025, pp. 1174–1180, doi: 10.1109/ICETCI64844.2025.11084166.

[21] 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*, vol. 34, 2021, pp. 12077–12090.

[22] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick, "Segment anything," in *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 4015–4026.

[23] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P. H. S. Torr, and L. Zhang, "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 6877–6886, doi: 10.1109/CVPR46437.2021.00681.

[24] R. Strudel, R. Garcia, I. Laptev, and C. Schmid, "Segmenter: Transformer for semantic segmentation," in *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 7262–7272.

[25] S. A. Taghanaki, Y. Zheng, S. K. Zhou, B. Georgescu, P. Sharma, D. Xu, D. Comaniciu, and G. Hamarneh, "Combo loss: Handling input and output imbalance in multi-organ segmentation," *Comput. Med. Imaging Graph.*, vol. 75, pp. 24–33, 2019, doi: 10.1016/j.compmedimag.2019.04.005.