# Feature-level Fusion of 2D Images and 3D LiDAR Point Clouds for Semantic Segmentation

Andrey Nikolov[1[0009−0000−1125−9112]], Muhammad Shoaib Sarwar[1[0000−0001−8951−4320]], and Faizan Ahmed[1,2[0000−0002−2760−6892]]

[1] Department of Computer Science, University of Twente, Drienerlolaan 5, Enschede, The Netherlands
andrey.nikolovbg@gmail.com;{shoaib.sarwar,faizan.ahmed}@utwente.nl
[2] Ambient Intelligence Group, Saxion University of Applied Sciences, Van Galenstraat 19, Enschede, The Netherlands

**Abstract.** Semantic segmentation is an essential task in autonomous systems, including those used for driving, robot navigation, and medical diagnosis. Although there are methods for 2D segmentation using convolutional neural networks and 3D segmentation with 3D models, the complementary nature of 2D and 3D data should not be overlooked. This research explores the multimodal fusion of 2D images and 3D LiDAR point clouds for semantic segmentation in both structured and unstructured environments. Building on the DeepViewAgg framework, we aim to investigate how feature fusion impacts semantic segmentation compared to models that utilize only 2D or only 3D data. The approach involves training a model for each modality and assessing its performance. On KITTI-360, fusion raises the mean IoU from 54.20 (3D-only) and 56.70 (2D-only) to 57.53, with the most notable improvement seen in thin classes like "pole" (+21.3 points). In the WildScenes natural dataset, it reaches 33.0 mIoU, surpassing the 2D and 3D baseline models by 5.0 points. These results show that multimodal fusion can outperform single modal approaches, especially for scene elements that benefit from combined 2D-3D cues.

**Keywords:** Semantic segmentation · 2D-3D fusion · Feature fusion · Multimodal fusion

## 1 Introduction

Semantic segmentation is the pixel-level classification of different objects against a complex background [23]. This classification enables an understanding of an environment, therefore it is a fundamental requirement in fields such as robot navigation, robotic arm grasping systems, autonomous driving systems, and medical diagnosis [21]. Earlier convolutional neural network (CNN) approaches for segmentation include the use of 2D architectures such as U-NET [15] and simple models with ResNet backbones [7]. Recent approaches involve more advanced models, such as DeepLabV3 [2], which capture both fine details and the wider scene simultaneously, resulting in more accurate segmentations. These CNNs utilize RGB images (2D data) that provide rich semantic content, including color, texture, and shape, which are key aspects of object identification. A broad toolkit for enhancing per-pixel segmentation in 2D data is available for methods such as high-resolution backbones [20] and online hard example mining [16] further raising mean Intersection over Union (mIoU). However, CNNs that rely exclusively on RGB cues are vulnerable to object occlusion, changes in lighting, and the absence of geometric information, often misclassifying thin, distant, or occluded objects.

LiDAR sensors and 3D segmentation techniques address these limitations by providing precise distance and depth measurements for each point, enriching environment understanding. Significant progress has been achieved in the point cloud segmentation field, with models such as PointNet++ [13], MinkUNet [3], and KPConv [18] demonstrating the potential of point cloud based object detection on (indoor) datasets such as ScanNet [1,11] and (outdoor) KITTI 360 [10], however, often struggle with identifying semantic classes for small or occluded objects. To enhance performance, several approaches can be used to augment point clouds with color information, including the use of colorization (which requires specialized sensors or a colorization step), meshing, or true depth maps [14]. These methods are either hardware dependent or computationally expensive and may fail to capture the semantic cues available in RGB images.

In an attempt to leverage the complementary nature of both types of data, multimodal fusion techniques have emerged as a promising direction in semantic segmentation research [17,8]. Taken together, a 2D-3D pipeline can bridge the critical gaps in single modality real-world perception. Notably, 3D point cloud data is often accompanied by corresponding 2D images [22], making such fusion possible in most cases and providing a solid ground for research in the field. Current fusion frameworks such as DeepViewAgg [14] utilize 2D image features from multiple camera views and fuse them with point cloud data via an attention based mechanism to perform semantic segmentation on 3D data. This approach achieved the current state of the art performance on 3D semantic segmentation on the urban dataset KITTI-360 [10].

Although multimodal fusion has shown promising results in urban and indoor datasets, its performance in natural environments remains unclear. Current research on the intersection between multimodal fusion and natural unstructured environments remains limited. Research only explores fusion in the agricultural context or utilizes single modality architectures [9,19]. Better semantic understanding in forests, fields, and natural habitats can contribute to improved automated search and rescue applications, wildlife conservation and monitoring, and agricultural automation, for which urban datasets are not suitable [19].

This paper aims to explore the effect of multimodal feature fusion of 2D images and 3D LiDAR point clouds for semantic segmentation in urban (structured) and natural (unstructured) environments. The research is based on the outline of the DeepViewAgg framework [14]. The goal is to compare the performance of the multimodal framework against similar 2D-only and 3D-only baselines and determine its effects. To achieve this, the three segmentation paradigms will be systematically trained and evaluated using consistent splits and Intersection over Union (IoU) metrics (cf. Section 3.1). This approach will isolate the effect of fusion on overall accuracy and class-wise performance in both datasets.

This paper is organized as follows. Section 2 reviews related work on semantic segmentation and multimodal fusion. Section 3 explains the preprocessing pipeline, model architectures, and training procedures. Section 4 discusses the evaluation methods and findings, while Section 5 examines ablations conducted with different parameters. Section 6 offers an analysis of the results, emphasizing limitations and suggesting future research directions. Finally, Section 7 provides a summary of the study.

## 2 Related Work

The fusion of 2D camera imagery with 3D LiDAR data has emerged as a promising approach for semantic segmentation, with researchers exploring various architectures to leverage the complementary strengths of both modalities. Early work by Robert et al. [14] demonstrated the potential of this approach by projecting 3D LiDAR points into multiple calibrated RGB camera views. By extracting features from each image using pre-trained 2D CNNs and employing a view aware attention module for aggregation, they achieved state of the art performance on the KITTI 360 dataset [10]. Notably, this approach eliminated the need for intermediate representations like meshes or depth maps. However, while their work established the superiority of fusion over pure 3D models, it left unexplored the comparison against 2D-only architectures.

Building on this foundation, Hu et al. [8] advanced the fusion paradigm by introducing BPNet, which features a symmetric dual branch architecture that simultaneously processes data through 2D UNet and 3D MinkUNet models. Their key innovation, the Bidirectional Projection Module (BPM), enables continuous feature exchange between modalities that achieved superior mIoU scores on the indoor ScanNetV2 dataset [6]. This work provided the first comprehensive comparison showing fusion's advantages over both standalone 2D and 3D models. However, their evaluation remained confined to indoor environments, leaving the effectiveness of such approaches in complex outdoor scenes uncertain.

While urban and indoor environments have received significant attention, the application of multimodal fusion to natural environments presents unique challenges. Kragh and Underwood [9] pioneered this direction by developing a conditional random field based fusion method for agricultural obstacle detection. By establishing spatial links between 2D and 3D segments and introducing multimodal connections, they achieved impressive gains of 9 and 13 points in four class mIoU for 2D and 3D classification, respectively. Yet, their approach is limited to four basic categories: "sky", "object", "ground", and "vegetation" far simpler than the semantic richness found in urban datasets.

This progression reveals a critical gap in the literature: while multimodal fusion has proven effective in urban settings with comprehensive comparisons [14,8] and shown promise in natural environments with limited semantic categories [9], no work has thoroughly evaluated the effectiveness of 2D-3D fusion on large-scale natural datasets with rich semantic diversity. This gap motivates our comprehensive experiments on WildScenes [19], where we systematically evaluate multimodal fusion against both 2D and 3D baselines in complex natural environments.

## 3 Methodology

In the study, three types of semantic segmentation models are implemented: a 2D-3D multimodal model, a 3D-only model, and a 2D-only model. This section begins with the selected datasets and their corresponding evaluation metrics, followed by an overview of the pipeline and the specific implementation details of each model, including their configurations.

### 3.1 Datasets and Evaluation Metrics

In order to fully assess multimodal fusion methods in different environments, we carefully chose datasets that belong to both the well studied domain of structured urban scenes and the less explored natural environments. This dual dataset approach allows us to validate current results and expand the evaluation to more difficult conditions.

For structured urban environments, we employed the KITTI-360 dataset [10], which has become a benchmark for multimodal semantic segmentation in urban settings. This dataset captures the complexity of city driving through 320,000 images and 100,000 laser scans collected over 73.7 km of urban traversals in Karlsruhe, Germany. Crucially, KITTI-360 provides precise camera calibration parameters (intrinsics and extrinsics) and georegistered vehicle poses, enabling accurate 2D-3D data correspondence which is a fundamental requirement for our fusion experiments. The dataset's comprehensive coverage of urban elements across 15 semantic classes allows us to establish baseline performance in well structured environments.

However, natural environments pose fundamentally different challenges: uneven terrain, dense vegetation occlusions, and high variability in object appearance. To bridge this gap, we chose WildScenes [19], a recently released dataset that captures the complexity of Australian forests through multiple large scale traversals over six months. With 21.28 km of coverage resulting in 9,306 annotated images and 12,148 annotated point clouds, WildScenes offers the scale needed for reliable evaluation. Similar to KITTI-360, it includes camera calibration and pose data vital for multimodal fusion. The dataset's 15 classes (of which we use 13 for evaluation) reflect the semantic diversity of natural environments, from different vegetation types to terrain categories. The two classes "sky" and "water" were omitted because LiDAR returns are nonexistent or inaccurate on these classes, so we exclude them from evaluation for our 3D benchmarking.

To ensure a fair comparison across these diverse environments, we use the IoU metric. For KITTI-360, we utilize their confidence weighted IoU formulation to account for annotation ambiguity.

$$\text{IoU} = \frac{\sum_{i \in \text{TP}} c_i}{\sum_{i \in \text{TP,FP,FN}} c_i} \quad (1)$$

where TP, FP, and FN denote true positive, false positive, and false negative pixel sets respectively, and $c_i \in [0, 1]$ represents the confidence at pixel $i$. This weighting scheme is significant for KITTI-360's automatically generated annotations, where boundary regions may have inherent uncertainty. For WildScenes, which provides manually verified annotations, we use the standard IoU without confidence weighting ($c_i = 1$ for all pixels). Additionally, we report mIoU across all classes to provide a holistic view of model performance, where higher values indicate better pixel-wise agreement between predictions and ground truth.

To provide a comprehensive evaluation beyond the IoU, we computed four additional metrics from the confusion matrix for each variant of the model (cf. Table 12). The first, Overall Accuracy (OA), measures the ratio of correctly classified pixels to total pixels, offering a global performance indicator. The other three are macro-averaged metrics designed to assess performance independent of class frequency. These are Macro-averaged Precision (mPrecision)-($\text{TP}_c/(\text{TP}_c+\text{FP}_c)$) which assesses prediction correctness, Macro-averaged Recall (mRecall)-($\text{TP}_c/(\text{TP}_c+\text{FN}_c)$) which measures the ability to detect all instances of each class $c$, and Macro-averaged Dice Coefficient (mDSC) which provides a balanced performance score. These macro-averaged metrics complement IoU by weighting all classes equally and exposing precision-recall trade-offs. While this class-agnostic perspective is valuable, we retain IoU as our primary metric. This decision ensures consistency with established semantic segmentation literature [10,19] and reflects that spatial overlap quality directly impacts downstream robotics applications.

### 3.2 Pipeline

We used a three-stage pipeline to examine how the choice of modality (2D, 3D, and fusion of 2D and 3D) impacts the precision of semantic segmentation in urban environments (Figure 1). The first stage involved selecting a relevant dataset that provides synchronized LiDAR and camera data for different model types. The dataset chosen is KITTI-360. The splits defined in [14] are used for the 2D and 3D data to ensure that differences in performance among the models are due to their type and architecture, not the dataset split. Data pre-processing is performed to adapt the data for the environment and models. The next stage involved selecting models. The chosen multimodal model is the state of the art 2D-3D fusion network, previously benchmarked on KITTI 360 - Res16UNet34 + ResNet-18 with early fusion. The 3D-only setup mimics the 3D backbone of the multimodal architecture, allowing for a direct comparison. DeepLabV3+ with a ResNet-18 backbone is selected as a representative standard 2D semantic segmentation architecture. In the following stage, consistent training schedules and evaluation criteria are used to establish baselines and ensure a fair comparison. We assessed all three models using a class-wise (across all standard classes) and mIoU metric, defined in Equation 1. This approach highlighted overall performance differences and enabled a detailed analysis of classes with significant differences among the models. Additionally, ablation studies are performed and are presented in Section 5.

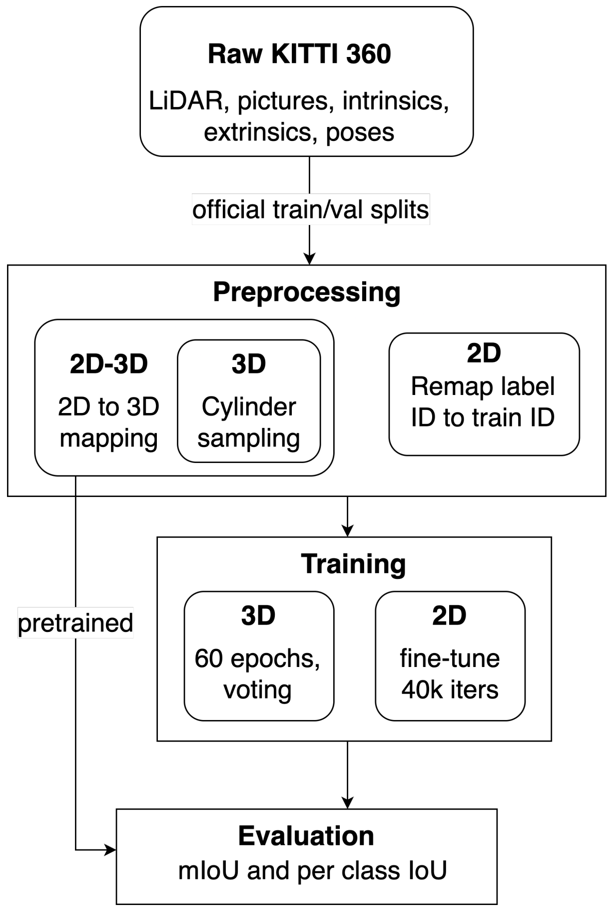



Fig. 1: Multimodal semantic segmentation pipeline

For multimodal fusion in the natural environment, the pipeline follows the same four high-level stages. However, it differs in three important aspects. First, this pipeline utilized the complete WildScenes [19] splits (train, val, and test), whereas KITTI-360 provides access only to the train and val sets. Second, different models are adopted, with the 2D-only baseline being DeepLabV3 with a ResNet-18 backbone. The 2D-3D multimodal model uses a custom Res18UNet for both the image encoder and the sparse 3D backbone. Furthermore, late logit fusion is implemented instead of early fusion, as in the previous pipeline. Unlike in KITTI-360, all models are trained on WildScenes. The 2D-only and 3D-only models are trained on the dataset and used as initializations for the multimodal architecture. Finally, each model is evaluated based on its mIoU and per-class IoU.

### 3.3 Model Architectures for KITTI-360

The environment used for implementing and evaluating the models is equipped with an NVIDIA A10 GPU (23 GB VRAM) and a CPU configuration with 65 cores and 256 GB of RAM.

**Multimodal 2D-3D model** The first model is the Res16UNet34-PointPyramid-early-cityscapes multimodal architecture, implemented in [14]. This model features a 3D-only backbone, specifically Torch-Point3D's Res16UNet34 implementation of MinkowskiNet, along with a 2D encoder, ResNet-18, pre-trained on the Cityscapes dataset [5]. It employs early fusion between the 2D and 3D features. A sampling strategy is introduced to handle the large data volume in KITTI-360 using 6m radius vertical cylinders. This involves downsampling the points to 5 cm voxels during preprocessing and selecting one image every five frames from the left perspective camera.

**3D-only Model** The second type of model is the 3D-only backbone of the multimodal model described earlier. We used the architecture and training setup established by [14]. The model is trained solely on 3D data from KITTI-360, following the official training and validation splits. Training lasted for 60 epochs, each with roughly 12,000 cylinders, using stochastic gradient descent (SGD) with an initial learning rate of 0.1. This relatively high learning rate is appropriate for sparse 3D convolution networks processing irregular point cloud data due to three factors. First, MinkowskiNet's sparse convolutions operate on substantially sparser feature maps than dense 2D convolutions, requiring larger learning rates to accumulate sufficient gradient signal from the limited number of active voxels. Second, batch normalization layers following each sparse convolution stabilize training by normalizing activations, enabling aggressive learning rates without divergence. And third, step decay ($\times 0.1$ at epochs 30 and 50) rapidly reduces the effective learning rate during later training stages (final rate: $10^{-3}$), ensuring stable convergence. Training stability was validated through monitoring: no loss spikes or gradient explosions occurred across all 60 epochs.

**2D-only Model** The third type of model is the DeepLabV3+ architecture with a ResNet-18 backbone [2]. The model is initialized from a checkpoint pre-trained on Cityscapes for 80,000 iterations [4]. For a fair comparison baseline, we fine-tuned the model on the 2D data in KITTI-360.

Images are augmented during training with standard Cityscapes style augmentations (e.g., random scale, horizontal flip). We implemented a gradient cumulative optimizer [4], which simulated a batch size of 8 and fit it within the

A10's VRAM. SGD is employed with a base learning rate of 0.005. The learning rate is linearly warmed up for the first 500 iterations from 0.000001, and then decayed according to a polynomial schedule to 0.000001 at the final iteration.

The choice of SGD over adaptive optimizers (e.g., Adam, AdamW) follows established semantic segmentation practice [4,2] for three reasons. First is that SGD with momentum demonstrates superior generalization performance for dense prediction tasks, particularly when fine-tuning from pre-trained weights, where adaptive optimizers' per-parameter learning rates may disrupt transferred feature representations [16]. Second is that SGD aligns with the baseline implementations we compare against [14,4], eliminating optimizer choice as a confounding variable. And, third is that SGD requires minimal memory overhead compared to adaptive methods that maintain per-parameter statistics critical given our GPU memory constraints (23 GB VRAM).

The learning rate warm-up strategy prevents catastrophic forgetting of pre-trained Cityscapes features during domain adaptation to KITTI-360. Specifically, the learning rate increases linearly from $10^{-6}$ (warm-up initialization) to the base rate of $5 \times 10^{-3}$ over 500 iterations, allowing gradual adaptation to the new data distribution. Following warm-up, polynomial decay $\mathrm{lr}_t = \mathrm{lr}_{\mathrm{base}} \times (1 - t/T)^{0.9}$ progressively reduces the learning rate to $10^{-6}$ at iteration $T = 40{,}000$, enabling large early updates for domain transfer followed by fine-grained convergence refinements.

### 3.4 Model Architectures for WildScenes

**Multimodal 2D-3D Model** The multimodal model features two parallel UNet branches, similar to the configuration of a ResNet18UNet. Outputs are combined at the logit level, enabling late feature fusion. The 3D branch employs sparse 3D convolutions with an initial 7 to 32 channel embedding, four downsampling stages, and a symmetric upsampling that returns to 96 channels. The 2D architecture closely mirrors this design, taking RGB inputs through the same sampling scheme to generate 96 channel per-pixel logits for the 13 dataset classes. During inference, the per-pixel image logits are pooled and aligned with each LiDAR point, then averaged and passed through a softmax to produce the final semantic label.

To address point collision, we removed the non-static mask transform, and an extra grid sampling step is added after the image mapping. The model is trained for 30 epochs, with validation every 5 epochs, with each epoch containing about 10,000 cylinders. The weights are initialized using matching layers from the 3D-only and the 2D-only models. Due to the smaller size of the training batch, SGD is used with an initial learning rate of 0.01, adjusted according to the predefined multi-step learning rate scheduler employed in the multimodal model for KITTI-360.

**3D-only Model** For the 3D-only architecture, the 3D sparse UNet model from the multimodal architecture is used. Since training this model takes less time than training the multimodal one, we conducted training for 60 epochs on the point clouds in WildScenes. To account for the difference in epochs, the number of cylinders per epoch is halved, resulting in 5,000 cylinders per epoch. However, the exact 5 cm voxel resolution is maintained. The same pre-, train, test, and validation transformations as in KITTI-360 are applied.

Table 1: Comprehensive hyperparameter settings for all model architectures across KITTI-360 experiments.

| **Configuration** | **2D-3D Fusion** | **3D-only** | **2D-only** |
|---|---|---|---|
| *Architecture* | | | |
| 3D Backbone | Res16UNet34 (MinkowskiNet) | Res16UNet34 | — |
| 2D Backbone | ResNet-18 (Cityscapes pre-trained) | — | ResNet-18 (Cityscapes pre-trained) |
| Decoder | DeepLabV3+ style | — | DeepLabV3+ |
| Fusion Strategy | Early (view-aware attention) | — | — |
| Parameters | 28.1M [14] | 21.2M | 12.5M |
| *Optimization* | | | |
| Optimizer | SGD with momentum | SGD with momentum | SGD with momentum |
| Momentum | 0.9 | 0.9 | 0.9 |
| Weight Decay | $1 \times 10^{-4}$ | $1 \times 10^{-4}$ | $1 \times 10^{-4}$ |
| Base Learning Rate | 0.01 | 0.1 | 0.005 |
| LR Schedule | Multi-step decay | Step decay | Polynomial with warmup |
| LR Decay Factors | $\times 0.1$ at epochs [30, 50] | $\times 0.1$ at epochs [30, 50] | $(1 - t/T)^{0.9}$ |
| Warmup | None | None | Linear, 500 iter. ($10^{-6}$ to base) |
| Final Learning Rate | $1 \times 10^{-5}$ | $1 \times 10^{-3}$ | $1 \times 10^{-6}$ |
| *Training Protocol* | | | |
| Training Duration | 60 epochs | 60 epochs | 40,000 iterations |
| Batch Size | 2 (effective) | 2 | 8 (via gradient accumulation) |
| Training Time (approx.) | 12 hours | 10 hours | 6 hours |
| Validation Frequency | Every 5 epochs | Every 5 epochs | Every 10,000 iterations |
| *Data Processing* | | | |
| Voxel Resolution | 5 cm | 5 cm | — |
| Cylinder Radius | 6 m | 6 m | — |
| Cylinders per Epoch | ≈12,000 | ≈12,000 | — |
| Image Frame Sampling | Every 5th frame | — | All frames |
| Image Augmentation | Random scale, flip, color jitter | — | Cityscapes-style augmentation |
| *Inference* | | | |
| Spatial Resolution | 1 (dense evaluation) | 1 (dense evaluation) | Per-pixel |
| Voting Strategy | Single vote | Single vote | — |
| Full Resolution | True | True | True |

**2D-only Model** The 2D-only architecture we implemented is DeepLabV3 with a ResNet-18 backbone [4]. Vidanapathirana et al. [19] used this model with a ResNet-50 backbone to obtain a 2D semantic segmentation baseline. We employed a smaller ResNet-18 backbone to provide a fairer comparison compared to the other architectures. The crucial difference lies in the different label mapping, with the removal of the classes "sky" and "water" to match the list of 3D classes. Furthermore, the learning rate is linearly scaled to match the batch size of two that is used, resulting in a value of 0.001. SGD is used with the same warm-up and decay configuration as in KITTI-360. We trained the model for 80,000 iterations, with evaluation every 4,000 based on mIoU.

### 3.5 Training Configuration Summary

Table 1 provides a comprehensive overview of training configurations for all model variants evaluated in this study.

## 4 Evaluation and Results

In this section, evaluation procedures and results will be discussed for all three models in the two selected datasets. Models are evaluated using IoU (cf. Section 3.1). We used the evaluation splits as defined in [10,19]. All models are implemented and evaluated in the environments outlined in Section 3.

For all models, we evaluated their performance on the validation set available in the KITTI-360 dataset. Evaluation is performed on full-resolution point clouds using a spatial resolution of one, i.e., roughly one cylinder every three meters (cf. [14]). Additionally, voting inference is done on the data with a single vote. For semantic segmentation evaluation on the 2D data, the dataset authors provide evaluation scripts [10]. No modifications are made to the scripts, except for the omission of the classes "sky" and "rider" to align the list of classes across all modalities. The complete 2D class validation is presented in Section 5. All images in the validation split are assessed.

The performance assessment of the fusion and 3D-only models for WildScenes is carried out in the same manner as for KITTI-360. For the 2D-only model, we closely followed the evaluation procedure of the WildScenes paper [19], except for the implementation of an alternative label map as mentioned in Section 3. Performance is reported on the test set of the dataset.

### 4.1 Impact of Multimodal Fusion in KITTI-360

In this section, we present quantitative results for the segmentation models on the KITTI-360 validation set. For this validation, only the 15 classes common to all models are used (Table 2). To ensure statistical robustness of our findings, we trained the 2D-only network with three independent random seeds (seeds 0, 1, and 2) to comprehensively assess run-to-run variations and model stability. Given the substantially higher computational cost of 3D models (approximately 10 hours per training run), and considering that the primary variability concern was raised for 2D models, we trained the 3D-only model once. The three-seed evaluation protocol provides sufficient statistical evidence to assess the significance and robustness of the improvements achieved by multimodal fusion.

The three seed evaluation (Table 2) reveals critical insights into performance stability across model architectures. The 2D-only baseline achieves a mean mIoU of 56.7±1.0, demonstrating relatively consistent overall performance. However, per-class analysis exposes severe instability for rare classes: "motorcycle" exhibits extreme variance (24.0±16.5 IoU), with individual seed performance ranging from 4.8 to 35.8 IoU (a 31 point spread representing 147% coefficient of variation). Similarly, "bicycle" shows high instability (13.6±12.3 IoU, range 2.8–27.1).

Table 2: KITTI-360 Val: Comprehensive comparison across all models with multiple seeds for 2D-only variants

| **Class** | **2D-3D** | **2D-only Seed 0** | **2D-only Seed 1** | **2D-only Seed 2** | **2D-only Mean ± Std** | **3D-only** |
|---|---|---|---|---|---|---|
| Avg | **57.5** | 55.7 | 57.6 | 56.7 | 56.7 ± 1.0 | 54.2 |
| road | 88.3 | 93.2 | 93.1 | 93.2 | 93.2 ± 0.1 | **92.4** |
| sidewalk | 71.4 | **76.4** | **76.4** | 76.6 | 76.5 ± 0.1 | 74.8 |
| building | **87.8** | 82.3 | 82.0 | 81.6 | 82.0 ± 0.4 | 86.9 |
| wall | 49.2 | **69.6** | **70.6** | 69.3 | 69.8 ± 0.7 | 45.9 |
| fence | 39.4 | **42.1** | 41.5 | 41.7 | 41.8 ± 0.3 | **44.9** |
| pole | **59.2** | 38.6 | 37.1 | 36.1 | 37.3 ± 1.3 | 57.3 |
| t-light | **15.4** | 0.0 | 0.0 | 0.0 | 0.0 ± 0.0 | 9.8 |
| t-sign | 46.4 | **50.2** | 50.1 | 49.9 | 50.1 ± 0.2 | 47.9 |
| vegetation | 88.8 | **91.4** | **91.6** | **91.2** | 91.4 ± 0.2 | 85.1 |
| terrain | 61.1 | **76.6** | **77.8** | **77.0** | 77.1 ± 0.6 | 54.6 |
| person | 40.5 | 45.6 | **51.4** | 44.2 | 47.1 ± 3.8 | **46.8** |
| car | **93.8** | 93.3 | 93.4 | 93.1 | 93.3 ± 0.2 | 90.9 |
| truck | 61.7 | 60.7 | **64.3** | 63.6 | 62.9 ± 1.9 | 4.5 |
| motorcycle | 40.7 | 4.8 | 31.5 | **35.8** | 24.0 ± 16.5 | **41.8** |
| bicycle | 19.2 | 11.0 | 2.8 | 27.1 | 13.6 ± 12.3 | **29.5** |

This variability directly correlates with class frequency: rare classes "motorcycle" and "bicycle" comprise only 0.04% and 0.02% of training pixels respectively (cf. Table 6), providing insufficient training signal for stable convergence. In stark contrast, the multimodal fusion model achieves 57.5 mIoU from a single training run, outperforming the 2D mean by 0.8 points while demonstrating particular strength on geometrically defined thin structures: "pole" achieves 59.2 IoU (+21.9 points over 2D mean) and "traffic light" reaches 15.4 IoU (versus 0.0 across all three 2D seeds). These improvements on shape defined classes suggest that geometric cues from 3D data provide a stabilizing inductive bias that complements the texture based learning of 2D models, yielding more robust predictions even without multi seed averaging.

The early model Res16UNet34 achieved an average mIoU of 57.5. The two 2D runs obtain 55.7 and 57.6 mIoU, resulting in a mean of $56.7 \pm 1.3$. The 3D-only model scores 54.2 mIoU, which is 3.3 points lower than the multimodal one. Compared to the 2D baseline, the fusion model gains 1.8 points over the lower run and loses 0.1 points relative to the higher run. This performance still yields a 0.9 point increase above the mean of the two runs. Although it's below the variability of the 2D model, the improvement highlights the complementary information that both modalities provide in structured urban scenes.

Small and thin classes experience a significant gain over both the 2D and 3D only models. The fusion architecture achieved 59.2 IoU for "pole" and 15.4 IoU for “traffic light”, against the 2D-only model 37.9 IoU (“pole”) and 0 IoU (“traffic light”). In contrast, the 3D-only model performs more closely to the multimodal one, with a 57.3 IoU for “pole”, but still falls short for “traffic light” with a 9.8 IoU. Notably, large surfaces and background classes show a slight decline relative to the 2D-only model.

Figure 2 provides qualitative validation of the quantitative performance differences observed in Table 2. Across the three representative validation samples, subtle but consistent differences emerge among loss rebalancing strategies. The class-weighted loss variant demonstrates marginally improved boundary delineation for certain object classes, evident in Sample 2 where the “car” boundaries (dark blue) appear more precisely defined compared to the plain baseline. However, the visual differences are modest, reflecting the relatively small quantitative gaps between variants (mean mIoU differences of 0.7–2.6 points). OHEM shows comparable qualitative performance to the plain baseline across these samples, consistent with its modest +0.7 point quantitative improvement. Particularly, all three methods exhibit similar failure modes on rare classes. In Sample 1, small distant objects remain challenging across variants, and in Sample 3, fine structural details show comparable segmentation quality regardless of loss function. These qualitative observations confirm that while loss rebalancing provides measurable improvements for specific classes (as detailed in Table 9), the overall visual segmentation quality remains largely consistent across methods for the majority of scene content.

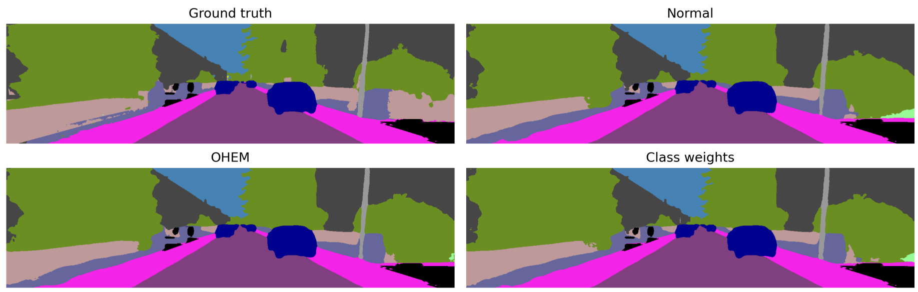
### 4.2 Impact of Multimodal Fusion in WildScenes

For the WildScenes test set, we used a list of the 13 standard classes across all models (Table 3). The 2D-3D fusion model achieved an overall mIoU of 33.0, compared to 27.9 for the 2D-only baseline and 28.0 for the 3D-only model. These results demonstrate that in the natural environment, multimodal fusion outperforms both single modalities by a substantial margin of 5.0 mIoU.

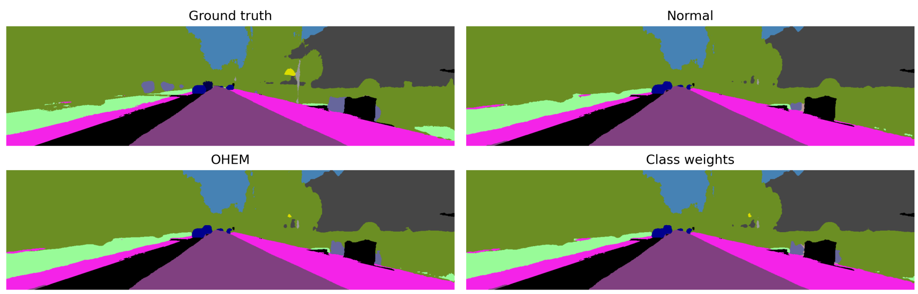
Table 3: WildScenes Test: Comparison between all models

| **Model** | **Avg** | **bush** | **dirt** | **fence** | **grass** | **gravel** | **log** | **mud** | **object** | **other-ter.** | **rock** | **structure** | **t-foliage** | **t-trunk** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2D-3D | **33.0** | **24.9** | 83.1 | 4.4 | 70.5 | 0.0 | 20.0 | **0.4** | **17.6** | **0.0** | 1.6 | **66.9** | 89.9 | 49.7 |
| 2D-only | 27.9 | 13.7 | 68.9 | 0.0 | 60.4 | **0.1** | **22.6** | 0.0 | 16.2 | 0.0 | 0.0 | 38.5 | 85.1 | **56.7** |
| 3D-only | 28.0 | 7.3 | **83.4** | **13.0** | **73.2** | 0.0 | 18.2 | 0.0 | 14.8 | 0.0 | **4.5** | 11.7 | **91.1** | 47.1 |

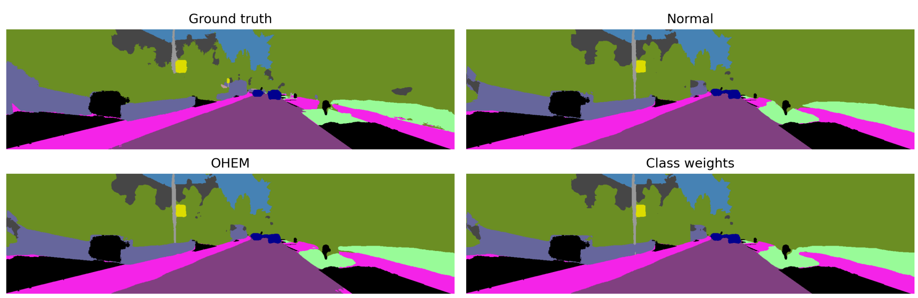
A closer look at per-class performance reveals that fusion achieves the larger score for the class “structure” with 66.9 IoU against 38.5 IoU for the 2D-only




(a) Sample 1




(b) Sample 2




(c) Sample 3

Fig. 2: Qualitative segmentation comparison of different methods (Normal, OHEM, Class weights) against Ground Truth on selected validation samples.

model and 11.7 IoU for the 3D-only model. The class "bush" experiences a significant gain, with 24.9 IoU, representing an 11.2 points increase compared to 2D-only and a 17.6 points increase compared to 3D-only. Notably, a slight increase is observed in the "mud" class, where only the fusion approach achieved a score higher than 0.0 with a 0.4 IoU.

Across both datasets, the fusion architecture attains the highest mIoU and consistently improves segmentation on small and structurally distinct classes (e.g., "pole", "structure"). A detailed interpretation of these trends and their implication can be found in Section 6.

In particular, no prior work has evaluated multimodal 2D-3D fusion on large-scale natural environment datasets. Vidanapathirana et al. [19] established single-modality baselines (ResNet-50 backbone for 2D: 31.4 mIoU) but did not explore fusion. Our results achieving 33.0 mIoU with fusion, a +5.0 point improvement over both 2D-only (27.9) and 3D-only (28.0) baselines represent the first systematic evaluation of multimodal fusion benefits in unstructured natural environments, establishing an initial benchmark for future research in this challenging domain.

# 5 Ablation Studies

## 5.1 Multimodal Model and 3D-only Model

We conducted experiments with varying values for sample resolution and complete resolution. Apart from the baseline configurations, which used sample resolution = 1 and full resolution = True, we conducted a validation with sample resolution = 3 and full resolution = False. In contrast to the method described in Section 3, these settings effectively reduced the time needed for inference by reducing the total number of evaluation locations, resulting in a roughly 80% reduction in total time. This reduction in inference time decreased the average validation mIoU by 1.86 points for the multimodal model and by 0.77 points for the 3D-only model. This run is therefore included to demonstrate that, for large urban scenes or stricter inference time requirements, a slight decrease in average mIoU can be used as a tradeoff for faster inference. The times reported in Table 4 are derived from the time needed to make a forward pass on the validation set. For the multimodal model, we omitted the time required for preprocessing the data (around 8 hours for mapping images and neighborhood based mapping features). Full class-wise performance for this experiment can be seen in Table 8.

Table 4: Comparison of models at two sampling settings

| Model | Sample Res. | Full Res. | Avg. mIoU | Time |
|---|---|---|---|---|
| 3D-only | 1 | True | **54.20** | 10h |
| 3D-only | 3 | False | 53.43 | **2h** |
| 2D-3D | 1 | True | **57.53** | 12h |
| 2D-3D | 3 | False | 55.67 | 2.5h |

## 5.2 Effect of Loss Rebalancing on 2D-only Models

In this section, the performance difference between plain DeepLabV3+ (Section 3) and the same network with OHEM and class balanced loss is explored in the context of urban semantic segmentation. The fine tuning process and validation for all models are identical to those described in Section 3. To assess run-to-run variation, we performed two training runs for each variant and reported the mean and standard deviation. Our motivation for exploring loss rebalancing is the largely skewed distribution of pixels. For example, the dominant "road" class has approximately 1,920 times more pixels than the rare "bicycle" class. More details about class pixel counts and resulting weights can be seen in Table 6.

To support the model with rare classes (e.g., traffic signs, traffic lights), we implemented an Online Hard Example Mining (OHEM) sampler [4] for the decoder head. By using it only for the decoder head, double memory usage is avoided while still allowing for gradient rebalancing. Following this default implementation, the sampler selects all pixels with model confidence below 0.7, keeping at least 100,000 pixels per crop. The use of the sampler led to a mIoU of $57.4 \pm 0.3$, an increase of 0.7 points over the plain baseline of $56.7 \pm 1.3$ mIoU. Furthermore, the rare class "motorcycle" IoU changed from $18.2 \pm 18.9$ to $26.7 \pm 4.1$ in the OHEM variant - a notable gain of 8.5 points. However, in the case of "bicycle" a significant drop is present. IoU dropped from $6.9 \pm 5.8$ IoU in the plain model to $1.7 \pm 6.4$ IoU in the OHEM one.

We explored an alternative method to address the heavily underrepresented classes' class balanced loss. For this, we used the ENet inverse log class weighting [12] with the following formula:

$$w_i = \frac{1}{\log(1 + D_i)} \tag{2}$$

where $D_i$ represents the number of pixels in class $i$, and $w_i$ is the corresponding weight for that class. The pixel count for each class is obtained from the training split. The computed class weights are used only in the decoder head of the model. The class balanced loss achieved $59.3 \pm 2.1$ mIoU, a 1.9-point increase in average mIoU compared to the OHEM variant, with the most notable improvement in the "bicycle" class. This class sees a significant $+22.3$ IoU gain compared to the plain model and a $+27.5$ IoU gain compared to the OHEM variant. Overall, mIoU improved significantly by $+2.6$ points compared to the plain DeepLabV3+. The class-wise performance of the three different 2D-only models can be seen in Table 9. Complete validation for the full list of 17 2D classes (including "sky" and "rider") is shown in Table 10.

Comparing the optimal 2D-only configuration (class-weighted loss: $59.3 \pm 2.1$ mIoU) against the multimodal fusion model (57.5 mIoU) reveals complementary strengths. Loss rebalancing achieves superior overall mIoU and dramatically improves rare classes defined by appearance: "bicycle" gains $+22.3$ points (from $6.9 \pm 5.8$ to $29.2 \pm 14.0$ IoU) through increased gradient contribution from minority class pixels. However, fusion provides orthogonal benefits for geometrically-defined classes where visual cues are inherently ambiguous: "pole" achieves 59.2 IoU with fusion versus $39.2 \pm 0.3$ IoU with class-weighted loss ($+20.0$ points), and "traffic light" reaches 15.4 IoU versus 0.0 IoU across all 2D configurations regardless of loss function. These complementary benefits suggest an optimal strategy would combine class balanced loss functions within a multimodal fusion architecture, potentially achieving both rare class improvements through statistical rebalancing and geometric understanding through 3D cues.

### 5.3 Backbone Depth Impact

We further compared multimodal fusion with the stronger single-modality baseline (the 2D model) across both datasets. By contrasting the 2D-3D architectures with deeper 2D backbones, we assess whether a smaller fusion model can match a robust deep architecture. For KITTI-360, we again used a plain DeepLabV3+ with a ResNet-101 backbone. However, due to the smaller multimodal architecture used for WildScenes, we compare it to DeepLabV3 with a ResNet-50 backbone. We trained and evaluated both models following the procedures outlined in Section 3 and Section 4.

Table 5: Comparison of models on the datasets

| **Dataset** | **KITTI-360** | **WildScenes** | **Parameters** |
|---|---|---|---|
| ResNet-18 | 56.7[1] | 27.9 | 12.5 M |
| ResNet-50 | – | 31.4 | 34.3 M |
| ResNet-101 | **62.4** | – | **53.3 M** |
| Multimodal (KITTI-360) | 57.5 | – | 28.1 M[14] |
| Multimodal (WildScenes) | – | **33.0** | 15 M |

[1] This value is the mean across the two training runs. All other numbers come from a single run.

A deeper analysis of the class-wise performance reveals that, despite a significant gap in mIoU and parameter count, the fusion model improves certain classes on KITTI-360. The classes “pole” and “traffic light” experience a notable boost in the multimodal model. The deeper 2D architecture achieved a 40 IoU for “pole” (19.2 points drop compared to multimodal) and 0 for “traffic light” (15.4 points drop compared to multimodal). This demonstrates that geometric cues can support the identification of visually challenging classes using RGB data.

Furthermore, the results highlight the difference in complexity between urban and natural environments. Even with a significantly deeper network, the 2D-only model struggles with semantic segmentation in WildScenes, with an increase of only 3.5 points in mIoU compared to ResNet-18. However, multimodal fusion outperforms both 2D models, while having 2 times fewer parameters than a DeepLabV3 with a ResNet-50 backbone. Class-wise analysis demonstrates strong increases in the classes' "structure" (20.6 points increase compared to ResNet-50), and "dirt" (15.7 points increase compared to ResNet-50). Overall, the effect on most classes is positive, except for the class "object", which suffers a 23.5 points drop compared to the 2D-model with ResNet-50. Complete comparison is presented in Table 11 and Table 7.

## 6 Discussion

The results show that early fusion of 2D images and 3D LiDAR features can significantly improve the segmentation of shape-defined or rare classes (e.g., "pole", "traffic light") in urban scenes. Due to their thin and small size, these classes are challenging for 2D segmentation because they occupy only a few pixels and can blend into the background. In contrast, the 3D-only model achieves performance similar to that of the multimodal one. The reason for this increase in scores is due to differences in how the objects are captured. For example, in point clouds, a pole appears as a thin vertical cluster of points rising from the ground, making it less likely to be mislabeled as background. When the two modalities are combined, performance for both classes reaches its peak. This indicates that the model did not rely solely on LiDAR scans but also used images to refine classification, such as distinguishing a pole from a thin tree trunk by their texture.

The benefit of fusion does not apply to every class. The fusion model underperforms on most large and background classes, except for the building class, where depth cues helped achieve a +5.6 IoU gain over the 2D baseline. For other large and background classes, early fusion may dilute visual cues and propagate projection misalignments (such as at curbs between a sidewalk and a road) into the fused representation. Roads in KITTI-360 have distinctive colors and textures, like asphalt, which a camera captures effectively, but LiDAR sensors do not, explaining the better results from the 2D modality. A key implication for systems using semantic segmentation fusion is to implement class or confidence aware fusion, such as incorporating fusion only for classes with low confidence predictions from one modality. One example where fusion degrades performance is the "bicycle" class, where fusion yields a 19.2 IoU, 10.3 points lower than the 3D baseline, indicating that noisy image features can override the more reliable geometric cues.

The WildScenes results confirm that late fusion of 2D images and 3D LiDAR data provides a significant overall benefit in unstructured (natural) environments, with a 5.0 points increase in mIoU compared to the stronger 3D baseline. This shows that, despite highly variable terrain, multimodal fusion can improve segmentation performance over single modality pipelines.

The per-class results reveal that the most significant benefit of fusion is in classes with distinctive shapes or irregular forms. For example, man-made "structure" elements and "bush", which challenged both the 2D and the 3D models, are significantly better identified in the fusion architecture. Because bushes in natural settings often share the same color and texture as other vegetation areas, such as grass or tree foliage, 2D-only models struggle. Moreover, a 3D architecture lacks the fine grained surface details needed to segment a bush cleanly. When the two modalities are combined, the fused model recovers both the shape (from geometry) and the fine boundary (from texture), yielding increases of 11.2 points over 2D and 17.6 points over 3D.

However, an instance of varying class examples that the fusion model did not correctly address is the class “fence”. The authors of WildScenes report that the class used in the train set has a single horizontal railing, while the one present in the test set has three horizontal ones [19]. Such differences confused the 2D model, which propagated this limitation in the multimodal fusion, resulting in an 8.6 point decrease in the fusion’s “fence” IoU compared to the 3D model.

### 6.1 Limitations and Future Work

We acknowledge the following limitations and outline directions for future work. The effect of class weighting should be investigated in the context of multimodal architectures. Our 2D ablation showed that class weighting significantly enhanced performance for rare classes. Additionally, to better utilize the strengths of different modalities, fusion strategies should be capable of learning to weight modalities on a per-class basis.

The current proposed architecture for multimodal fusion in WildScenes has room for improvement. Due to limited time, we restricted our experiments to late fusion only. Future work can explore different fusion approaches (e.g., early, intermediate) and their impact on performance. Additionally, we inherit several key limitations from the dataset itself. Variable lighting conditions in specific images (cf. Figure 3) degraded the output of the image encoder and, consequently, the overall fusion performance. As reported in [19], seasonal vegetation changes lead to noticeable drops of approximately 4 points in mIoU when training and testing seasons differ, and environmental domain shifts cause significantly larger drops of about 7 points, especially for man-made structures [19]. Future work can utilize multi-seasonal data over multiple years at consistent locations and explore temporal and environmental domain-adaptation techniques to mitigate these limitations.

Another limitation is backbone capacity. Robert et al. [14] report using a smaller 3D backbone than the current one, which resulted in a drop in mIoU on KITTI-360; however, they do not explore the effect of using a larger one. Future research can focus on architectures with deeper image encoders (e.g., ResNet-50/ResNet-101 or transformer backbones) or deeper 3D backbones to better compare with well-established deeper 2D architectures such as DeepLabV3 and DeepLabV3+ with a ResNet-101 backbone. Finally, extending multimodal fusion evaluation on real-time data and semantic segmentation architectures would provide key insights to improve self-driving cars, drones, and robotics.

## 7 Conclusion

This study compared multimodal fusion with 2D and 3D baselines for semantic segmentation on both urban and natural datasets. The results showed that fusion provides a clear improvement in semantic segmentation within KITTI-360 and WildScenes. In the urban KITTI-360 dataset, multimodal fusion achieved a smaller mIoU gain of 0.8 points but made significant improvements in segmenting small and thin classes that are difficult for single modalities. Across the 13 standard WildScenes classes, the fusion model gained 5.0 points in mIoU over the stronger model, supporting the idea that the two types of data are complementary. Both results show that one modality can help the other when it struggles, and that fusion works best when each modality provides consistent, complementary information to the other.

Furthermore, our ablation showed that the lightweight multimodal fusion matches or surpasses the performance of much deeper 2D architectures, demonstrating that combining modalities can be more effective than simply scaling a single model. However, the fusion approach has drawbacks. When data is noisy, too sparse, or contradictory, performance declined in the fusion architectures.

While the findings are promising, they are based mostly on single runs, and limitations exist. Future work can focus on using multiple seeds and runs to improve statistical validity and gain more insights into the field of multimodal fusion for urban and natural datasets.

## Declarations

**Code Availability.** Our code is available at `https://github.com/whid0t/multimodal-fusion`.

**Conflict of Interests.** The authors declare that they have no conflict of interest.

**Ethical and Informed Consent for Data Used.** The datasets used in this study are publicly available on the Internet. No animals or humans were victims.

**AI Statement.** The authors made use of ChatGPT and Grammarly to enhance the work's academic writing style and structure. The authors carefully reviewed every suggestion made by the tools and edited the content as needed.

## References


1. Chen, D.Z.: Pointnet2.scannet. `https://github.com/daveredrum/Pointnet2.ScanNet` (2021)
2. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European Conference on Computer Vision (ECCV) (September 2018)
3. Choy, C., Gwak, J., Savarese, S.: 4d spatio-temporal convnets: Minkowski convolutional neural networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3075–3084 (2019)

4. Contributors, M.: MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark. `https://github.com/open-mmlab/mmsegmentation` (2020)
5. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
6. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: Scannet: Richly-annotated 3d reconstructions of indoor scenes. In: Proc. Computer Vision and Pattern Recognition (CVPR), IEEE (2017)
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition (2015), `https://arxiv.org/abs/1512.03385`
8. Hu, W., Zhao, H., Jiang, L., Jia, J., Wong, T.T.: Bidirectional projection network for cross dimension scene understanding. In: CVPR (Jun 2021)
9. Kragh, M., Underwood, J.: Multimodal obstacle detection in unstructured environments with conditional random fields. Journal of Field Robotics **37**, 53–72 (03 2019). `https://doi.org/10.1002/rob.21866`
10. Liao, Y., Xie, J., Geiger, A.: KITTI-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d. Pattern Analysis and Machine Intelligence (PAMI) (2022)
11. Nekrasov, A., Schult, J., Litany, O., Leibe, B., Engelmann, F.: Mix3d: Out-of-context data augmentation for 3d scenes. In: 2021 international conference on 3d vision (3dv). pp. 116–125. IEEE (2021)
12. Paszke, A., Chaurasia, A., Kim, S., Culurciello, E.: Enet: A deep neural network architecture for real-time semantic segmentation (2016), `https://arxiv.org/abs/1606.02147`
13. Qi, C.R., Yi, L., Su, H., Guibas, L.J.: Pointnet++: Deep hierarchical feature learning on point sets in a metric space. Advances in neural information processing systems **30** (2017)
14. Robert, D., Vallet, B., Landrieu, L.: Learning multi-view aggregation in the wild for large-scale 3d semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5575–5584 (2022), `https://github.com/drprojects/DeepViewAgg`
15. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
16. Shrivastava, A., Gupta, A., Girshick, R.: Training region-based object detectors with online hard example mining. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 761–769 (2016)
17. Tan, M., Zhuang, Z., Chen, S., Li, R., Jia, K., Wang, Q., Li, Y.: Epmf: Efficient perception-aware multi-sensor fusion for 3d semantic segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence **46**(12), 8258–8273 (2024). `https://doi.org/10.1109/TPAMI.2024.3402232`
18. Thomas, H., Qi, C.R., Deschaud, J.E., Marcotegui, B., Goulette, F., Guibas, L.J.: Kpconv: Flexible and deformable convolution for point clouds. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6411–6420 (2019)
19. Vidanapathirana, K., Knights, J., Hausler, S., Cox, M., Ramezani, M., Jooste, J., Griffiths, E., Mohamed, S., Sridharan, S., Fookes, C., et al.: Wildscenes: A benchmark for 2d and 3d semantic segmentation in large-scale natural environments. The International Journal of Robotics Research **44**(4), 532–549 (2025)

20. Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X., et al.: Deep high-resolution representation learning for visual recognition. IEEE transactions on pattern analysis and machine intelligence **43**(10), 3349–3364 (2020)
21. Yang, C., Yan, Y., Zhao, W., Ye, J., Yang, X., Hussain, A., Dong, B., Huang, K.: Towards deeper and better multi-view feature fusion for 3d semantic segmentation. In: International Conference on Neural Information Processing. pp. 3–15. Springer (2023)
22. Zeid, K.A., Yilmaz, K., de Geus, D., Hermans, A., Adrian, D., Linder, T., Leibe, B.: Dino in the room: Leveraging 2d foundation models for 3d segmentation. arXiv preprint arXiv:2503.18944 (2025)
23. Zhang, R., Li, G., Li, M., Wang, L.: Fusion of images and point clouds for the semantic segmentation of large-scale 3d scenes based on deep learning. ISPRS journal of photogrammetry and remote sensing **143**, 85–96 (2018)

# Appendix

## A Pixel Count and Weights

As explained in Section 5.2, the pixel count is calculated based on the number of pixels for a corresponding class present in a ground truth image in the train split.

Table 6: Pixel count and weights for all classes[1]

| Class | Pixel count | Weight |
|---|---|---|
| road | 3847.8 | 0.818 |
| sidewalk | 1625.9 | 0.851 |
| building | 4709.6 | 0.810 |
| wall | 705.2 | 0.886 |
| fence | 557.2 | 0.896 |
| pole | 92.1 | 0.984 |
| traffic light | 0.3 | 1.437 |
| traffic sign | 31.9 | 1.045 |
| vegetation | 9157.5 | 0.787 |
| terrain | 793.9 | 0.881 |
| sky | 1949.3 | 0.844 |
| person | 11.4 | 1.111 |
| rider | 7.7 | 1.139 |
| car | 1512.0 | 0.854 |
| truck | 127.3 | 0.967 |
| bus | 4.9 | 1.172 |
| train | 6.0 | 1.156 |
| motorcycle | 6.8 | 1.147 |
| bicycle | 2.8 | 1.215 |

[1] Pixels are divided by $10^6$ and rounded to one decimal.

## B Ablation Tables

Table 7: WildScenes Test: Comparison between backbone depth

| Model | Avg | bush | dirt | fence | grass | gravel | log | mud | object | other-ter. | rock | structure | t-foliage | t-trunk |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Multimodal | **33.0** | **24.9** | **83.1** | **4.4** | **70.5** | 0.0 | 20.0 | **0.4** | 17.6 | **0.0** | **1.6** | **66.9** | **89.9** | 49.7 |
| ResNet-18 | 27.9 | 13.7 | 68.9 | 0.0 | 60.4 | 0.1 | 22.6 | 0.0 | 16.2 | 0.0 | 0.0 | 38.5 | 85.1 | 56.7 |
| ResNet-50 | 31.4 | 23.8 | 67.4 | 0.0 | 56.6 | **0.2** | **28.3** | 0.0 | **41.1** | 0.0 | 0.0 | 46.3 | 85.9 | **58.5** |

Table 8: KITTI-360 Val: Different samplings for multimodal and 3D-only

| Class | 3D-only[1] | 3D-only | 2D-3D[1] | 2D-3D |
|---|---|---|---|---|
| Avg | **54.20** | 53.43 | **57.53** | 55.67 |
| road | **92.43** | 92.07 | **88.33** | 85.62 |
| sidewalk | 74.77 | **74.96** | **71.37** | 69.73 |
| building | 86.86 | **87.42** | 87.78 | **88.13** |
| wall | **45.89** | 43.99 | **49.21** | 46.53 |
| fence | 44.85 | **45.08** | 39.44 | **39.68** |
| pole | **57.31** | 57.17 | **59.18** | 58.72 |
| t-light | **9.76** | 9.62 | 15.44 | **15.57** |
| t-sign | **47.88** | 45.83 | **46.41** | 44.11 |
| vegetation | 85.05 | **85.6** | 88.76 | **89.11** |
| terrain | 54.61 | **56.06** | **61.07** | 61.02 |
| person | **46.81** | 35.95 | **40.47** | 23.26 |
| car | 90.90 | **91.08** | **93.82** | 93.59 |
| truck | 4.52 | **5.38** | 61.74 | **64.52** |
| motorcycle | **41.82** | 41.40 | **40.67** | 36.51 |
| bicycle | 29.50 | **29.70** | **19.23** | 18.95 |

[1] Columns denoting where sample resolution = 1 and full resolution = True. In these columns validation for full resolution is used, while in the others the results from the voting run are used.

Table 9: KITTI-360 Val: 2D-Only Segmentation Results Comparison Classes

| **Class** | **plain** | **std** | **OHEM** | **std** | **class weights** | **std** |
|---|---|---|---|---|---|---|
| Avg | 56.7 | 1.3 | 57.4 | 0.3 | **59.3** | 2.1 |
| road | 93.2 | 0.1 | **93.3** | 0.5 | 92.9 | 0.1 |
| sidewalk | 76.4 | 0.0 | **76.8** | 1.0 | 75.9 | 0.3 |
| building | **82.2** | 0.2 | 82.1 | 0.1 | 82.0 | 0.3 |
| wall | **70.1** | 0.7 | 70.0 | 0.1 | 70.0 | 0.0 |
| fence | 41.8 | 0.4 | 41.3 | 0.5 | **42.1** | 0.1 |
| pole | 37.9 | 1.1 | 38.3 | 0.9 | **39.2** | 0.3 |
| t-light | 0.0 | 0.0 | 0.0 | 0.0 | **0.0** | 0.0 |
| t-sign | 50.2 | 0.1 | **50.3** | 0.1 | 49.9 | 0.9 |
| vegetation | 91.5 | 0.1 | **91.5** | 0.0 | 91.4 | 0.1 |
| terrain | 77.2 | 0.8 | 77.4 | 0.2 | **76.5** | 0.5 |
| person | 48.5 | 4.1 | 50.0 | 0.2 | **54.9** | 8.6 |
| car | 93.4 | 0.1 | **93.5** | 0.1 | 92.8 | 0.6 |
| truck | 62.5 | 2.5 | **65.3** | 0.4 | 62.2 | 2.3 |
| motorcycle | 18.2 | 18.9 | 26.7 | 4.1 | **29.4** | 9.9 |
| bicycle | 6.9 | 5.8 | 1.7 | 6.4 | **29.2** | 14.0 |

Mean values and standard deviations are reported for each model across columns.

Table 10: KITTI-360 Val: 2D-Only Segmentation Results All Classes

| **Class** | **plain** | **std** | **OHEM** | **std** | **class weights** | **std** |
|---|---|---|---|---|---|---|
| Avg | 56.7 | 1.8 | 57.6 | 0.3 | **59.3** | 2.2 |
| road | 93.2 | 0.1 | **93.3** | 0.5 | 92.9 | 0.1 |
| sidewalk | 76.4 | 0.0 | **76.8** | 1.0 | 75.9 | 0.3 |
| building | **81.8** | 0.1 | **81.8** | 0.0 | 81.6 | 0.3 |
| wall | **70.1** | 0.7 | 70.0 | 0.1 | 70.0 | 0.0 |
| fence | **41.8** | 0.4 | 41.2 | 0.4 | 42.1 | 0.1 |
| pole | 37.7 | 1.0 | 38.1 | 0.9 | **39.0** | 0.3 |
| t-light | 0.0 | 0.0 | 0.0 | 0.0 | **0.0** | 0.0 |
| t-sign | 49.9 | 0.0 | **50.0** | 0.2 | 49.5 | 0.9 |
| vegetation | **91.2** | 0.1 | **91.2** | 0.0 | 91.1 | 0.1 |
| terrain | 77.2 | 0.8 | **77.3** | 0.2 | 76.5 | 0.5 |
| sky | 94.2 | 0.1 | **94.3** | 0.0 | 94.1 | 0.1 |
| person | 45.3 | 6.2 | 48.0 | 0.6 | **52.4** | 10.1 |
| rider | 24.3 | 7.1 | 28.0 | 2.1 | **30.2** | 6.6 |
| car | 93.3 | 0.1 | **93.5** | 0.1 | 92.8 | 0.6 |
| truck | 62.0 | 3.1 | **65.0** | 0.7 | 61.5 | 2.8 |
| motorcycle | 17.8 | 18.4 | 26.3 | 4.0 | **29.1** | 9.5 |
| bicycle | 6.9 | 5.8 | 1.7 | 6.4 | **29.0** | 13.9 |

Table 11: KITTI-360 Val: Comparison between backbone depth

| Class | Multimodal | ResNet-18 | ResNet-101 |
|---|---|---|---|
| Avg | 57.5 | 56.7 | **62.4** |
| road | 88.3 | 93.2 | **95** |
| sidewalk | 71.4 | 76.4 | **81.9** |
| building | **87.8** | 82.2 | 82.5 |
| wall | 49.2 | **70.1** | 68.7 |
| fence | 39.4 | **41.8** | 39.6 |
| pole | **59.2** | 37.9 | 40.0 |
| t-light | **15.4** | 0.0 | 0.0 |
| t-sign | 46.4 | **50.2** | 48.6 |
| vegetation | 88.8 | **91.5** | 91.4 |
| terrain | 61.1 | **77.2** | 77.1 |
| person | 40.5 | 48.5 | **64.4** |
| car | 93.8 | 93.4 | **94.3** |
| truck | 61.7 | 62.5 | **71.8** |
| motorcycle | **40.7** | 18.2 | 37.5 |
| bicycle | 19.2 | 6.9 | **43.9** |

## C Additional Metrics

In Table 12, all variants achieve nearly identical OA ($\approx$91.3%), reflecting comparable performance on dominant classes ("road", "vegetation", "building") that contribute most pixels. However, macro-averaged metrics reveal significant differences. The plain model achieves the highest mRecall ($0.710\pm0.006$), indicating superior coverage across all classes including rare ones, while OHEM achieves the highest mPrecision ($0.783\pm0.013$) by aggressively focusing on hard examples, thereby reducing false positives at the cost of increased false negatives. The class weighted variant balances these extremes with intermediate precision ($0.776\pm0.004$) and recall ($0.669\pm0.001$), achieving the lowest variance across seeds. Notably, OHEM's elevated precision but diminished recall explains its mixed per-class performance in Table 10 as it correctly identifies challenging instances when predictions are made, but misses more instances overall, particularly for rare classes where the hard example mining strategy becomes overly conservative. The mDSC trends align closely with mIoU patterns, validating that our primary metric effectively captures the precision recall balance within a single interpretable measure suitable for cross study comparison.

Table 12: Comprehensive evaluation metrics for 2D-only model variants on KITTI-360 validation across three seeds.

| **Model Variant** | **OA** | **mPrecision** | **mRecall** | **mDSC** |
|---|---|---|---|---|
| *Plain DeepLabV3+* | | | | |
| Seed 0 | 0.913 | 0.772 | 0.709 | 0.705 |
| Seed 1 | 0.914 | 0.773 | 0.717 | 0.707 |
| Seed 2 | 0.913 | 0.768 | 0.705 | 0.698 |
| Mean±Std | 0.913±0.001 | 0.771±0.003 | 0.710±0.006 | 0.703±0.005 |
| *OHEM Variant* | | | | |
| Seed 0 | 0.913 | 0.777 | 0.657 | 0.668 |
| Seed 1 | 0.913 | 0.798 | 0.667 | 0.686 |
| Seed 2 | 0.913 | 0.773 | 0.667 | 0.668 |
| Mean±Std | 0.913±0.000 | 0.783±0.013 | 0.664±0.006 | 0.674±0.010 |
| *Class-weighted Loss* | | | | |
| Seed 0 | 0.915 | 0.779 | 0.670 | 0.683 |
| Seed 1 | 0.915 | 0.772 | 0.669 | 0.669 |
| Seed 2 | 0.913 | 0.777 | 0.668 | 0.668 |
| Mean±Std | 0.914±0.001 | 0.776±0.004 | 0.669±0.001 | 0.673±0.008 |

## D Lighting Conditions

Fig. 3: Example of poor lighting conditions. Image is taken from K-01 path in WildScenes [19].