Feature-level fusion of 2D images and 3D LiDAR point clouds for semantic segmentation

ANDREY NIKOLOV, University of Twente, The Netherlands

Fig. 1. Images from the KITTI-360[10] and WildScenes[19] datasets used in the paper.

Semantic segmentation is a crucial task in autonomous systems, including those used in driving, robot navigation, and medical diagnosis. While there are methods for 2D segmentation using convolutional neural networks (CNN) and 3D segmentation using 3D models, the complementary nature of 2D data and 3D data should not be ignored. This research investigates multimodal fusion of 2D images and 3D LiDAR point clouds for semantic segmentation in structured and unstructured environments. Building on the DeepViewAgg framework, we aim to investigate the impact of feature fusion on semantic segmentation compared to 2D- and 3D-only models. The methodology involves training a model for each modality and evaluating its performance. On KITTI-360, fusion improves mean IoU from 54.20 (3D-only) and 56.70 (2D-only) to 57.53, with the largest gain on thin classes such as 'pole' (+21.3 points). In the WildScenes natural dataset, it achieves 33.0 mIoU, outperforming 2D and 3D baselines with a margin of 5.0 points. These trends demonstrate that multimodal fusion can outperform single modalities, particularly in scene elements with complementary 2D-3D cues.

Additional Key Words and Phrases: semantic segmentation, 3D point clouds, 2D images, 2D-3D fusion, feature fusion, multimodal fusion, natural environments, urban environments

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 captures 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 - methods such as high-resolution backbones[20] and online hard example mining[16] further raise 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[8, 17]. 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

TScIT 43, July 4, 2025, Enschede, The Netherlands

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].

1.1 Objective and goals

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 in [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 as defined in Section 3.1. This approach will isolate the effect of fusion on overall accuracy and class-wise performance in both datasets.

1.2 Research questions

In order to address the defined problem statement and achieve these goals, the following research question is defined:

RQ1: How does multimodal feature fusion of 2D images and 3D LiDAR point clouds impact the performance of semantic segmentation compared to 3D-only and 2D-only models?

To examine the effects of fusion, the question was further broken down into the following sub-questions:

- **RQ1.1**: How does multimodal feature fusion affect segmentation accuracy in a structured urban environment compared to 3D-only and 2D-only models?
- **RQ1.2**: Which object classes (e.g., car, person, building) are affected the most by the fusion in a structured environment?
- **RQ1.3**: How does multimodal feature fusion affect segmentation accuracy in natural, unstructured scenes compared to 3D-only and 2D-only models?
- **RQ1.4**: Which object classes (e.g., bush, dirt, fence) are affected the most by the fusion in an unstructured environment?

The remainder of this paper is organized as follows. Section 2 reviews related work on semantic segmentation and multimodal fusion. Section 3 details the pipeline used for preprocessing the data, model architectures, and their training procedures. Section 4 presents the paper's evaluation procedures and results, and Section 5 discusses ablations conducted with different parameters. Finally, Section 6 presents analysis of the results, highlighting limitations and outlining directions for future research. Section 7 concludes the research.

2 RELATED WORK

Research has been conducted to identify relevant works in the field of data fusion for semantic segmentation. This section will review existing literature on the topic, discuss the methods and techniques used, while highlighting limitations in current research.

2.1 Attention-based multi-view framework

In [14], 3D LiDAR points are projected into multiple calibrated RGB camera views. Features are extracted from each image using a pre-trained 2D CNN, and a view-aware attention module is used to aggregate features from different perspectives before feeding them into a 3D segmentation backbone by early fusion. This strategy eliminates the need for meshes or depth maps while achieving the state-of-the-art performance on 3D semantic segmentation in the KITTI 360 dataset[10]. The research demonstrates the higher performance of fusion compared to 3D models however, it does not compare the proposed fusion model against a 2D architecture.

2.2 Bidirectional feature projection

In [8], a new model, BPNet, is proposed. This model employed a symmetric dual-branch architecture, simultaneously running 2D UNet and 3D MinkUNet models, thereby allowing for feature exchange between them. This feature exchange facilitated by the Bidirectional Projection Module (BPM), enables both models to benefit from the features of the other model. Such an exchange is a form of intermediate fusion that leads to higher mIoU for the ScanNetV2 dataset[6] compared to the other models. The paper highlights the performance boost of fusion compared to 3D and 2D models but does not assess its performance on an outdoor dataset.

2.3 Multimodal obstacle detection

The authors in [9] propose a method for fusing camera and LiDAR sensing with a conditional random field to perform obstacle detection in agricultural fields using a moving ground vehicle. Adding spatial links between segments in 2D and 3D, and further including a multimodal link between them, resulted in performance gains of 9 points for 2D classification and 13 points for 3D classification in four-class mIoU. Although the study demonstrates the potential of using combined 2D-3D data for natural environments, its experiments were limited to the use of only four simple classes: 'sky', 'object', 'ground', and 'vegetation'.

In summary, prior studies have focused on urban datasets without complete comparison or have utilized limited obstacle categories in natural scenes. The effectiveness of multimodal fusion in a large natural dataset remains unexplored, motivating our experiments in WildScenes[19].

3 METHODOLOGY

To answer the main research questions of the study, we implemented three types of semantic segmentation models: a 2D-3D multimodal model, a 3D-only model, and a 2D-only model. This section begins with the chosen datasets and their corresponding evaluation metrics, followed by a high-level explanation of the pipeline, as well as the specific implementation of the three models, each with its own configuration.

3.1 Datasets and evaluation metric

First, performance in structured urban environments was evaluated using the KITTI-360[10] dataset, which consists of 320,000 images and 100,000 laser scans across a driving distance of 73.7 km in urban

areas in Karlsruhe, Germany. The data was captured with a multi-sensor mobile platform. The camera calibration, which provides the intrinsics and extrinsics of the sensors, and accurate georegistered vehicle poses are provided to enable mapping between the 2D and 3D data in the dataset.

The selected unstructured natural dataset is WildScenes[19]. The dataset comprises multiple large-scale traversals of forests in Australia, collecting multimodal data in the span of 6 months. The distance covered totals 21.28 km, resulting in the collection of 9,306 annotated images and 12,148 annotated point clouds. Furthermore, it provides camera calibration and poses used for mapping the two types of data. It provides 15 total classes, of which 13 will be used to assess the performance of the models.

The KITTI-360 dataset allows for confidence-weighted Intersection-over-Union (IoU) performance evaluation for each of the 15 semantic classes. This metric shows the overlap between a prediction and the ground truth for a pixel set and is calculated as follows:

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

where TP, FP, and FN denote the number of true positives, false positives, and false negatives pixel sets for each class, respectively. The evaluation of the KITTI-360 dataset uses confidence weighting to account for ambiguity in their automatically generated annotations, where c_i denotes confidence at pixel i, and $c_i \in [0; 1]$. Additionally, we used a mean Intersection-over-Union (mIoU) to show the average performance across all classes for a model. For both metrics, a higher value corresponds to a better pixel-wise overlap between the prediction mask and the ground truth annotation. The metric that will be used for WildScenes is IoU, as defined in Equation 1. However, for this dataset, confidence weighting will not be employed.

3.2 Pipeline

We employed a three-stage pipeline to investigate how modality selection (2D, 3D, and fusion of 2D and 3D) affects semantic segmentation accuracy in urban environments. The first stage involved selecting of a relevant dataset that offers synchronized LiDAR and camera data for use by the different types of models. The dataset selected was KITTI-360. The officially defined splits for the 2D and 3D data were used to ensure that any differences in performances among the models are due to their type and architecture rather than the dataset split. Data preprocessing was used to adjust the data for the environment and models used. The next stage involved selecting models. The selected 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 configuration mirrors the 3D backbone of the multimodal architecture, allowing for a direct comparison. DeepLabV3+ with a ResNet-18 backbone was selected as a representative for a standard, comparable 2D semantic segmentation architecture. In the next stage, consistent training schedules and evaluation criteria were employed to establish baselines and a fair comparison. We evaluate all three models using a class-wise (across all common classes) and mean Intersection-over-Union (IoU) metric, defined in Equation 1. This helped highlight the overall contrast in performance and provided a more in-depth analysis of the classes that experienced substantial differences between the models. Additionally, ablation studies were conducted and are presented in Section 5.

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 were adopted - the 2D-only baseline is DeepLabV3 with a ResNet-18 backbone. For the 2D-3D multimodal model, we selected a lighter architecture due to the resources and time needed for training it. The multimodal architecture uses a custom Res18UNet architecture for both the image encoder and the sparse 3D backbone. Furthermore, late logit fusion is implemented, rather than early fusion, as in the previous pipeline. Unlike in KITTI-360, for WildScenes all models were trained. The 2D-only and 3D-only models were trained on the dataset and used as initializations for the multimodal architecture. Finally, each model was evaluated based on its mean IoU and per-class IoU.

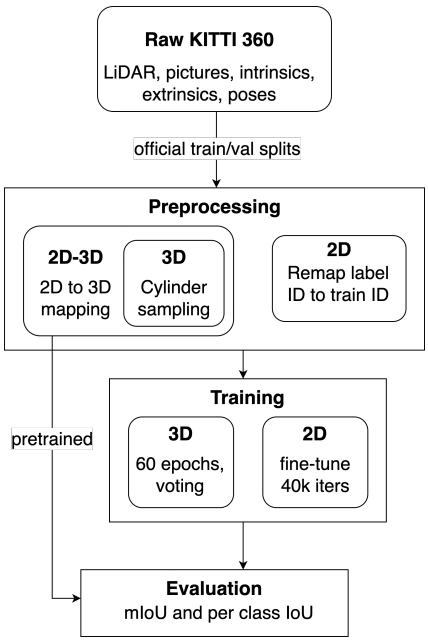

Fig. 2. High-level overview of urban dataset pipeline

3.3 Model architectures for KITTI-360

The environment used for implementing and evaluating the models was a Jupyter server available to the university, equipped with an NVIDIA A10 GPU (23 GB VRAM) and a CPU configuration with 65 cores and 256 GB of RAM.

3.3.1 Multimodal 2D-3D model. The first kind of model is the multimodal architecture Res16UNet34-PointPyramid-early-cityscapes implemented by the authors in [14]. This model consists of a 3D-only backbone, namely Torch-Point3D's Res16UNet34 implementation of MinkowskiNet and a 2D encoder, ResNet-18, pre-trained on the

Cityscapes dataset[5]. Early fusion between the 2D and 3D features is employed in the model. The authors of the paper provide the publicly available code and pre-trained weights for the model. They define a sampling strategy for managing the large amount of data present in KITTI-360 with the help of 6 m-radius vertical cylinders. This approach involves downsampling the points in preprocessing to 5 cm voxels, as well as selecting one image every five from the left perspective camera of the dataset. Except for adding CUDA acceleration for computing the 3D-2D mapping between a window and all images of the sequence (to decrease the time needed for preprocessing of data), we made no changes to the configuration used by the authors of the paper.

3.3.2 3D-only model. The second kind of model is the 3D-only backbone of the multimodal model previously described. We employed the architecture and training configuration made by the authors of [14]. The model was trained on 3D-only data from KITTI-360, respecting the official training and validation splits. The training was conducted over 60 epochs, each consisting of approximately 12,000 cylinders, using stochastic gradient descent (SGD) with an initial learning rate of 0.1. The learning rate was adjusted according to the predefined multi-step learning rate scheduler, multi-kitti-360.

3.3.3 2D-only model. The third type of model is the architecture DeepLabV3+ with a ResNet-18 backbone[2]. The model is initialized from a checkpoint pre-trained on Cityscapes for 80,000 iterations by the authors of [4]. For a fair comparison baseline, we fine-tuned the model on the 2D data in KITTI-360. The data used were according to the officially defined train split. Based on the configuration files for the framework mmsegmenation[4], a custom configuration file was created to provide the structure and pipelines of the fine-tuned model, as well as a custom dataset configuration file. In the model and training configuration file, we adopt the standard Encoder-Decoder framework of DeepLabV3+ with a ResNet-18 backbone.

For better training, all data were converted from the original label IDs to training IDs using an official label map. Images are augmented during training with standard Cityscapes-style augmentations (e.g., random scale, horizontal flip).

To better simulate the original training conditions of the model within the memory constraint of the available GPU, we implemented a gradient cumulative optimizer[4], which simulated a batch size of 8 and fit it within the A10's VRAM. Stochastic gradient descent (SGD) was employed with a base learning rate of 0.005. The learning rate was 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 model was fine-tuned for 40,000 iterations, with evaluation every 10,000 based on the mIoU metric.

3.4 Model architectures for WildScenes

For the natural dataset, the environment used was a high-performance cluster available to the university, equipped with an NVIDIA A40 GPU (48 GB VRAM) and a CPU configuration with 16 cores and 64 GB of RAM.

3.4.1 Multimodal 2D-3D model. Due to resource and time constraints, we adopt a lighter model architecture, utilizing the framework defined by [14]. The multimodal model consists of two parallel UNet branches, resembling the configuration of a ResNet18UNet. Outputs are combined at the logit level, producing a late feature fusion. The 3D branch uses sparse 3D convolutions with an initial 7 to 32-channel embedding, four downsampling stages, and a symmetric upsampling back to 96 channels. The 2D part of the architecture closely mirrors this design, taking RGB inputs through the same sampling scheme to produce 96-channel per-pixel logits over the 13 dataset classes. During inference, the per-pixel image logits are pooled and aligned to each LiDAR point, and the set of logits is averaged and passed through a softmax to obtain the final semantic label.

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

3.4.2 3D-only model. For the 3D-only architecture, the 3D sparse UNet model from the multimodal architecture was used. Since the training for this model is less time-consuming than that of the multimodal one, we conducted training for 60 epochs on the point clouds present in WildScenes. To account for the difference in epochs, the number of cylinders per epoch was reduced by half, resulting in 5,000 cylinders per epoch. However, the same 5 cm voxel resolution was used. The same pre-, train, test, and validation transformations as in KITTI-360 were used.

3.4.3 2D-only model. The 2D-only architecture we implemented is DeepLabV3 with a ResNet-18 backbone[4]. The dataset authors used this model with a ResNet-50 backbone to obtain a 2D semantic segmentation baseline in their paper[19]. We employed a smaller ResNet-18 backbone to provide a fairer comparison compared to the other architectures. The training closely followed the procedures outlined by the dataset authors. 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 was linearly scaled to match the batch size of 2 that was used, resulting in a value of 0.001. SGD was 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.

4 EVALUATION AND RESULTS

In this section, evaluation procedures and results will be discussed for all three models in the two selected datasets. Models were evaluated using IoU (cf. Section 3.1). We used the officially provided evaluation splits as defined by the datasets authors[10, 19]. All models were 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 was performed on full-resolution point clouds using a spatial resolution of 1, i.e., roughly one cylinder every 3 meters cf. [14]. Additionally, voting

inference was 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 were 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. During evaluation the images were remapped back from train IDs to label IDs. All images in the validation split are assessed.

The performance assessment of the fusion and 3D-only models for WildScenes was 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 report 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 4). Due to its lightweight nature, we trained the 2D-only network twice with two independent random seeds to measure run-to-run variations. Due to time and resource constraints, we trained the 3D-only model only once.

The early-fusion model Res16UNet34 achieved an average mIoU of 57.5. The two 2D runs obtain 55.7 and 57.6 mIoU, giving a mean of 56.7 ± 1.3. The 3D-only model scores 54.2 mIoU, which is 3.3 points lower than the multimodal one. Against the 2D baseline, the fusion model performs with a 1.8 points gain over the lower run and a -0.1 points loss compared to the higher run. This performance still results in a +0.9 gain above the mean of the two runs. Although this is below the 2D model's variability, the improvement indicates the complementary information that both modalities convey in structured urban scenes.

Small and thin classes experience a large 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.

4.2 Impact of multimodal fusion in WildScenes

For the WildScenes test set, we used a list of the 13 common classes across all models (Table 5). 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.

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 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.

5 ABLATION STUDIES

5.1 Multimodal model and 3D-only model

We conducted experiments with varying values for sample resolution and full 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 was 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 trade-off for faster inference. The times reported in Table 1 are derived from the time needed to make a forward pass on the validation set. For the multimodal model, we omitted the time needed for preprocessing the data (around 8 hours for mapping images and neighborhood-based mapping features).

Table 1. 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

Full class-wise performance for this experiment can be seen in Table 6.

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 are explored in the context of urban semantic segmentation. The fine-tuning process and validation for all models were identical to those described in Section 3. To assess run-to-run variation, we performed two training runs for each variant and reported 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 3.

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 was avoided while still allowing for gradient rebalancing. Following mmsegmentation's 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 ± 0.3 - an increase of 0.7 points over the plain baseline of 56.7 ± 1.3 mIoU. Furthermore, the rare class 'motorcycle' IoU changed from 18.2 ± 18.9 to 26.7 ± 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 ± 5.8 IoU in the plain model to 1.7 ± 6.4 IoU in the OHEM one.

We explored another approach to deal with the largely 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 denotes the number of pixels belonging to class i and w_i is the corresponding weight for that class. The pixel count for each class was obtained from the training split after mapping label IDs to train IDs. The computed class weights were used only in the decoder head of the model. The class-balanced loss resulted in 59.3 ± 2.1 mIoU, a 1.9 points increase in average mIoU compared to the OHEM variant, with the most notable gain in the 'bicycle' class. This class achieves 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+.

All model training took approximately 1.5 hours to complete on one NVIDIA A10 GPU. The time needed shows that there is no notable difference in training time when using sampling strategies. The class-wise performance between the three different 2D-only models can be seen in Table 7. Complete validation on the full list of 17 2D classes (including 'sky' and 'rider') is presented in Table 8.

5.3 Backbone depth impact

We further explored the comparison of multimodal fusion with the stronger single-modality baseline (the 2D model) across both datasets. By comparing the 2D-3D architectures with those of 2D architectures having deeper backbones, we investigate whether a smaller fusion model is comparable to a strong deep architecture. For KITTI-360 we again used a plain DeepLabV3+ but 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 defined in Section 3 and Section 4.

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.

Table 2. 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.

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 less 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 9 and Table 10.

6 DISCUSSION

6.1 Answering sub-question 1 and sub-question 2

The results show that early fusion of 2D images and 3D LiDAR features can noticeably improve 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, as they occupy only a few pixels and can blend into the background. In contrast, the 3D-only model achieves a performance comparable to that of the multimodal one. The reason for this increase in scores is due to the differences in capturing the objects. For example, in point clouds, a pole appears as a thin vertical cluster of points rising from the ground, which is less likely to be mislabeled for background. When the two modalities are combined, performance for both classes reaches its highest level. This suggests that the model did not rely only on the LiDAR scans but also used images to refine the classification, for example, in distinguishing a pole from a thin tree trunk by their texture.

The benefit of fusion does not extend to every class - the fusion model underperforms on most large and background classes, with the exception of the building class, where depth cues helped achieve a +5.6 IoU gain over the 2D baseline. For the 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 (such as asphalt appearance) that a

camera effectively captures, but LiDAR sensors do not, explaining the better results achieved from the 2D modality. A key implication for systems utilizing semantic segmentation fusion is to implement class- or confidence-aware fusion, e.g., using fusion only for classes with low confidence predictions from one of the modalities. One example where fusion also 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.

6.2 Answering sub-question 3 and sub-question 4

The WildScenes results confirm that late fusion of 2D images and 3D LiDAR data yields a clear overall advantage in unstructured (natural) environments, achieving a 5.0 points mIoU increase compared to the stronger 3D baseline. This synergy demonstrates that, regardless of the highly variable terrain, multimodal fusion can enhance segmentation performance compared to 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 points decrease in the fusion's 'fence' IoU compared to the 3D model.

6.3 Limitations and future work

We acknowledge the following limitations and outline directions for future work. Due to time constraints we did not measure statistical significance using different runs for multimodal and 3D models. In addition, because of high standard deviation for some classes for the 2D-only ablation, additional runs would be needed for a better statistical significance.

The effect of class weighting should be explored in the context of multimodal architectures. Our 2D ablation demonstrated that class weighting greatly improved the performance in rare classes. In addition, to better leverage the strengths of the different modalities, fusion strategies should be able to learn weight modalities on a per class basis.

The current proposed architecture for multimodal fusion in WildScenes does have room for improvement. Given the 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. Moreover, 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, thus, the overall fusion performance. As reported in the WildScenes paper, seasonal vegetation change lead to noticeable drops of ≈ 4 points in mIoU when train and test seasons differ, and environmental domain shift result in considerably mIoU drops of ≈ 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 environment domain-adaptation techniques to mitigate these limitations.

Another limitation is backbone capacity. The authors in [14] report about using a smaller 3D backbone than the current one, that resulted in a drop in mIoU on KITTI-360, however they do not explore the effect of using a bigger 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/DeepLabV3+ with a ResNet-101 backbone. Finally, extending multimodal fusion evaluation on real-time data and semantic segmentation architectures would provide key findings 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 demonstrated that the fusion yields a clear improvement in semantic segmentation within KITTI-360 and WildScenes. In the context of the urban KITTI-360 dataset, multimodal fusion achieved a smaller mIoU gain of 0.8 points yet improved significantly in segmentation for small and thin classes that are challenging for single modalities. Across the 13 common WildScenes classes, the fusion model achieved a 5.0 points mIoU gain over the stronger model, reinforcing the claim about the complementary nature of the two types of data. Both results demonstrate that one modality can support the other when struggling and that fusion is most effective when each modality supplies consistent, complementary information to the other.

Furthermore, our ablation revealed that the lightweight multimodal fusion matches or exceeds the performance of much deeper 2D architectures, demonstrating that combining modalities can be more effective than simply scaling a single model. Nevertheless, the fusion approach has drawbacks. In cases where data was noisy, too sparse, or contradicting, performance degraded in the fusion architectures.

While the findings are promising, they are a result of mostly single runs, and limitations are present. Future work can focus on using multiple seeds and runs to ensure better statistically validity and provide more insights into the field of multimodal fusion for urban and natural datasets.

8 AI STATEMENT

The author made use of ChatGPT and Grammarly to enhance the work's academic writing style and structure. The AI tool CursorAI

was used to learn about frameworks' features used for implementation. The author carefully reviewed every suggestion made by the tools and edited the content as needed.

REFERENCES

[1] D. Z. Chen. 2021. Pointnet2.scannet. https://github.com/daveredrum/Pointnet2.ScanNet.

[2] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. 2018. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

[3] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 2019. 4D Spatio-Temporal ConvNets: Minkowski Convolutional Neural Networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 3070–3079. https://doi.org/10.1109/CVPR.2019.00319

[4] MMSegmentation Contributors. 2020. MMSegmentation: OpenMMLab Semantic Segmentation Toolbox and Benchmark. https://github.com/open-mmlab/mmsegmentation.

[5] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. 2016. The Cityscapes Dataset for Semantic Urban Scene Understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[6] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. 2017. ScanNet: Richly-annotated 3D Reconstructions of Indoor Scenes. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep Residual Learning for Image Recognition. arXiv:1512.03385 [cs.CV] https://arxiv.org/abs/1512.03385

[8] W. Hu, H. Zhao, L. Jiang, J. Jia, and T.-T. Wong. 2021. Bidirectional Projection Network for Cross Dimension Scene Understanding. In *CVPR*.

[9] Mikkel Kragh and James Underwood. 2019. Multimodal obstacle detection in unstructured environments with conditional random fields. *Journal of Field Robotics* 37 (03 2019), 53–72. https://doi.org/10.1002/rob.21866

[10] Yiyi Liao, Jun Xie, and Andreas Geiger. 2022. KITTI-360: A Novel Dataset and Benchmarks for Urban Scene Understanding in 2D and 3D. *Pattern Analysis and Machine Intelligence (PAMI)* (2022).

[11] Alexey Nekrasov, Jonas Schult, Or Litany, Bastian Leibe, and Francis Engelmann. 2021. Mix3D: Out-of-Context Data Augmentation for 3D Scenes. arXiv:2110.02210 [cs.CV] https://arxiv.org/abs/2110.02210

[12] Adam Paszke, Abhishek Chaurasia, Sangpil Kim, and Eugenio Culurciello. 2016. ENet: A Deep Neural Network Architecture for Real-Time Semantic Segmentation. arXiv:1606.02147 [cs.CV] https://arxiv.org/abs/1606.02147

[13] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. 2017. PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space. In *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/d8bf84be3800d12f74d8b05e9b89836f-Paper.pdf

[14] Damien Robert, Bruno Vallet, and Loic Landrieu. 2022. 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*. 5575–5584. https://github.com/drprojects/DeepViewAgg

[15] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi (Eds.). Springer International Publishing, Cham, 234–241.

[16] Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick. 2016. Training Region-based Object Detectors with Online Hard Example Mining. arXiv:1604.03540 [cs.CV] https://arxiv.org/abs/1604.03540

[17] Mingkui Tan, Zhuangwei Zhuang, Sitao Chen, Rong Li, Kui Jia, Qicheng Wang, and Yuanqing Li. 2024. EPMF: Efficient Perception-Aware Multi-Sensor Fusion for 3D Semantic Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46, 12 (2024), 8258–8273. https://doi.org/10.1109/TPAMI.2024.3402232

[18] Hugues Thomas, Charles R. Qi, Jean-Emmanuel Deschaud, Beatriz Marcotegui, Francois Goulette, and Leonidas J. Guibas. 2019. KPConv: Flexible and Deformable Convolution for Point Clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

[19] Kavisha Vidanapathirana, Joshua Knights, Stephen Hausler, Mark Cox, Milad Ramezani, Jason Jooste, Ethan Griffiths, Shaheer Mohamed, Sridha Sridharan, Clinton Fookes, and Peyman Moghadam. 2024. WildScenes: A benchmark for 2D and 3D semantic segmentation in large-scale natural environments. *The International Journal of Robotics Research* 44, 4 (Sept. 2024), 532–549. https://doi.org/10.1177/02783649241278369

[20] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, Wenyu Liu, and Bin Xiao. 2020. Deep High-Resolution Representation Learning for Visual Recognition. arXiv:1908.07919 [cs.CV] https://arxiv.org/abs/1908.07919

[21] C. Yang, Y. Yan, W. Zhao, J. Ye, X. Yang, A. Hussain, and K. Huang. 2022. Towards Deeper and Better Multi-view Feature Fusion for 3D Semantic Segmentation. arXiv preprint arXiv:2212.06682.

[22] Karim Abou Zeid, Kadir Yilmaz, Daan de Geus, Alexander Hermans, David Adrian, Timm Linder, and Bastian Leibe. 2025. DINO in the Room: Leveraging 2D Foundation Models for 3D Segmentation. arXiv:2503.18944 [cs.CV] https://arxiv.org/abs/2503.18944

[23] R. Zhang, G. Li, M. Li, and L. Wang. 2018. 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 (May 2018), 85–96.

A PIXEL COUNT AND WEIGHTS

As explained in Section 5.2, the pixel count was calculated based on the number of pixels for a corresponding class present in a ground truth image in the train split. The images used were previously converted into train ID instead of label ID and the count is based on train ID.

Table 3. 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 RESULTS TABLE

Table 4. KITTI-360 Val: Comparison between all models

Model	Avg	road	sidewalk	building	wall	fence	pole	t-light	t-sign	vegetation	terrain	person	car	truck	motorcycle	bicycle
2D-3D	**57.5**	88.3	71.4	**87.8**	49.2	39.4	**59.2**	**15.4**	46.4	88.8	61.1	40.5	**93.8**	61.7	40.7	19.2
2D-only A	55.7	**93.2**	**76.4**	82.3	69.6	42.1	38.6	0.0	**50.2**	91.4	76.6	45.6	93.3	60.7	4.8	11
2D-only B	**57.6**	93.1	76.4	82	**70.6**	41.5	37.1	0.0	50.1	**91.6**	**77.8**	**51.4**	93.4	**64.3**	31.5	2.8
3D-only	54.2	92.4	74.8	**86.9**	45.9	**44.9**	57.3	9.8	47.9	85.1	54.6	46.8	90.9	4.5	**41.8**	**29.5**

Table 5. 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

C ABLATION TABLES

Table 6. KITTI-360 Val: Different samplings for mutlimodal and 3D-only

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

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

Table 7. KITTI-360 Val: 2D-Only Segmentation Results Comparison Classes[1]

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

[1] Mean values are reported for the models in each row.

Table 8. KITTI-360 Val: 2D-Only Segmentation Results All Classes

Model	Avg	road	sidewalk	building	wall	fence	pole	t-light	t-sign	vegetation	terrain	sky	person	rider	car	truck	motorcycle	bicycle
plain	56.7	93.2	76.4	81.8	**70.1**	**41.8**	37.7	0.0	49.9	91.2	77.2	94.2	45.3	24.3	93.3	62	17.8	6.9
std	1.8	0.1	0.0	0.1	0.7	0.4	1.0	0.0	0.0	0.1	0.8	0.1	6.2	7.1	0.1	3.1	18.4	5.8
OHEM	57.6	**93.3**	**76.8**	**81.8**	70.0	41.2	38.1	0.0	**50.0**	**91.2**	**77.3**	**94.3**	48.0	28.0	**93.5**	**65.0**	26.3	1.7
std	0.3	0.5	1.0	0.0	0.1	0.4	0.9	0.0	0.2	0.0	0.2	0.0	0.6	2.1	0.1	0.7	4.0	6.4
class weights	**59.3**	92.9	75.9	81.6	70.0	42.1	**39.0**	**0.0**	49.5	91.1	76.5	94.1	**52.4**	**30.2**	92.8	61.5	**29.1**	**29.0**
std	2.2	0.1	0.3	0.3	0.0	0.1	0.3	0.0	0.9	0.1	0.5	0.1	10.1	6.6	0.6	2.8	9.5	13.9

Table 9. KITTI-360 Val: Comparison between backbone depth

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

Table 10. 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**

D LIGHTING CONDITIONS

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