GANav: Group-wise Attention for Classifying Navigable Regions in Unstructured Outdoor Environments

Tianrui Guan, Divya Kothandaraman, Rohan Chandra and Dinesh Manocha

(Code and Videos at https://gamma.umd.edu/offroad)

***Abstract*— We present a new learning-based method for identifying safe and navigable regions in off-road terrains and unstructured environments from RGB images. Our approach consists of classifying groups of terrain classes based on their navigability levels using coarse-grained semantic segmentation. We propose a bottleneck transformer-based deep neural network architecture that uses a novel group-wise attention mechanism to distinguish between navigability levels of different terrains. Our group-wise attention heads enable the network to explicitly focus on the different groups and improve the accuracy. In addition, we propose a dynamic weighted cross-entropy loss function to handle the long-tailed nature of the dataset. We show through extensive evaluations on the RUGD and RELLIS-3D datasets that our learning algorithm improves visual perception accuracy in off-road terrains for navigation. We compare our approach with prior work on these datasets and achieve an improvement over the state-of-the-art mIoU by 6.74-39.1% on RUGD and 3.82-10.64% on RELLIS-3D. Supplementary materials including code, videos, and a full technical report are available at gamma.umd.edu/offroad.**

I. Introductions

The problem of autonomous navigation [1] has been an active area of research in robotics. There is considerable progress in terms of navigating indoor or structured environments due to advances in computer vision and sensor-based planning algorithms [2], [3], [4]. Recent developments in autonomous driving have significantly increased the interest in outdoor navigation [5], [6], where the main focus is on driving on well paved and clearly defined roadways or finding obstacles on the road. On the other hand, many applications corresponding to mining, disaster relief [7], agricultural robotics [8], or environmental surveying require the capability to navigate in uneven terrains or scenarios that lack structure or well-identified navigation features

A key issue in developing autonomous navigation capabilities in off-road terrain environments is to find safe and navigable regions that can be used by a robot, which may correspond to an autonomous vehicle, an autonomous wheelchair, a surveillance vehicle, or a mobile robot. For instance, some terrains like concrete or asphalt are highly navigable, while rocks or gravel may not be navigable. The underlying navigation system needs to be able to detect such regions. Moreover, the notion of classifying a safe, navigable region depends on the size, weight, and dynamics characteristics of the robot. While a large transport vehicle can drive on dirt roads or grass, a small mobile robot may not operate well on natural surfaces covered by grass or vegetation. In this paper, our goal is to identify such safe and navigable regions in unstructured outdoor environments from RGB images.

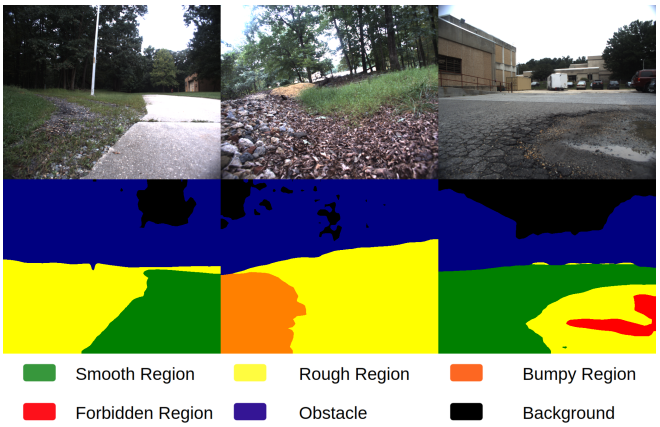


Fig. 1: We highlight the performance of our algorithm on unstructured outdoor terrains. Our algorithm classifies regions filled with gravel and sand as rough regions, roads as smooth regions, and water as a forbidden region. On the RUGD dataset, our model achieves an accuracy of 83.24, which is an improvement of $6.74 - 39.1\%$ over prior algorithms. We classify the navigable regions *(bottom)* using our approach on raw RGB images *(top)*.

A key aspect of classifying appropriate navigable regions in unstructured environments is to use visual perception capabilities such as semantic segmentation [9], [10]. Semantic segmentation is a pixel-level task that assigns a label for every pixel in the image. Prior works in segmentation [11], [12] have been limited to structured environments [5], [6] and may not work well, in general, in off-road terrains for several reasons. First, the distribution of the semantic classes (rocks, gravel, concrete, asphalt, etc.) in terrains can be long-tailed, i.e., the number of instances of a large number of semantic classes is very low. As a consequence, learning-based segmentation algorithms are unable to learn representations of these categories with few instances. Furthermore, different terrain classes can be highly similar in appearance (*e.g.*, concrete and asphalt; water and a puddle), with overlapping boundaries. This results in similar feature embeddings for multiple classes that can ultimately result in wrong classifications.

It has been shown that in unstructured environments, navigation systems may not require fine-grained semantic segmentation [13], [14]. For example, it is sufficient to recognize trees and poles as obstacles for collision avoidance rather than segmenting them individually as different types of objects, which is a harder problem. A coarser approach that classifies various types of objects based on their navigation

characteristics is more appropriate. In general, performing coarse-grained segmentation could also decrease the error rate due to its ability to accommodate misclassifications between similar classes, which tend to be in the same group.

Main Results: We present a new learning-based approach for identifying the navigability of different terrains from RGB images or videos. Our approach is general and designed for off-terrain and unstructured outdoor environments. Our learning algorithm uses a novel deep neural network with bottleneck transformers as the underlying architecture. In order to perform coarse-grained segmentation and address the issue of long-tailed data distribution, we group terrain classes according to their navigability levels. We also introduce a novel group-wise attention mechanism, which consists of several attention heads that help in learning semantically meaningful representations of these different groups of terrains. We use a novel, dynamically weighted cross-entropy loss function that helps in tackling the long-tailed distribution. The key contributions of our work include:

1) A new strategy for coarse-grained semantic segmentation for off-road terrains. We show that training deep neural networks with coarse-grained labels addresses the issue of long-tailed data distribution.
2) A novel deep neural network architecture using Group-wise Attention based on bottleneck transformers and multi-head self-attention. Our method achieves SOTA performance on classifying different terrain class grouping.
3) A dynamically weighted cross-entropy loss term to alleviate the bias in training caused by long-tailed distributions.

We evaluate our approach on the RUGD [15] and the RELLIS-3D [16] datasets, which contain 24 and 20 semantic classes with 8000 and 6000 images, respectively. We outperform prior methods for semantic segmentation and navigable regions classifications [10], [17], [9], [18], [19] by $6.74-39.1\%$ on RUGD and $3.82-10.64\%$ on RELLIS-3D. We also conduct several ablation experiments that highlight the benefits of various components of our method.

II. Related work

We give a brief overview of prior work on robot navigation, semantic segmentation, and outdoor datasets.

A. Navigation in Outdoor Scene

Prior work in uneven terrain navigation includes techniques for mobile robots [13], [1], [20], [21] to large vehicles [14], [22], [23]. In these algorithms, learning the navigation characteristics of the terrain and its traversability is a crucial step. At a broad level, there are three types [24] of approaches that are used to determine the terrain features: 1) proprioceptive-based methods, 2) geometric methods, and 3) appearance-based methods. The proprioceptive-based methods [25], [26] use frequency domain vibration information gathered by the robot sensors to classify the terrains using machine learning techniques. However, these methods require the robot to navigate through the region to collect data and assume that a mobile robot has the ability to navigate over the entire terrain. The geometric-based methods [23], [22], [27] generally use Lidar and stereo cameras to gather 3D point cloud and depth information of the environment. This information can be used to detect the elevation, slope, and roughness of the environment as well as obstacles in the surrounding area. Nevertheless, the accuracy of the detection outcome is usually governed by the range of the sensors. Appearance-based methods [28] usually extract road features with SIFT or SURF features and use MLP to classify a set of terrain classes. [29], [30] perform road material classifications. [31], [32] utilize deep neural network models for fine-grained semantic segmentation of urban scenarios to determine traversability of indoor scenarios. Additionally, there has been extensive work on the binary classification problem of segmenting roads into traversable or non-traversable regions [33], [34]. Our goal is to compute the navigable regions from RGB images and is complimentary to these methods.

B. Semantic Segmentation

Semantic segmentation is an important task in computer vision that involves assigning a label to each pixel in an image. The problem has been widely studied in the literature [35]. Many deep learning architectures have been proposed, including OCRNet [18], DeepLab [9], and PSP-Net [17]. While most of these architectures work well on structured datasets like CityScapes, they do not work well for off-road datasets due to boundaries that are not well defined or long-tailed distribution. Segmentation methods like Global Convolution Networks [36] are designed to deal with complex boundaries and do not scale well in terms of performance due to convoluted and overlapping boundaries that are generally not found in structured datasets. In contrast, we propose a new method that uses bottleneck transformers with group-wise attention.

C. Off-road Datasets

Recent developments in semantic segmentation have achieved high accuracy on object datasets like PASCAL VOC [37], COCO [38], and ADE20K [39], as well as driving datasets like Cityscape [5] and KITTI [6]. On the other hand, there has not been much work on recognition or segmentation in unstructured off-road scenes, which is important for navigation. Perception in an unstructured environment is more challenging since many object classes (e.g., puddles or asphalt) lack clear boundaries. In addition, the number of instances of multiple categories is very low, which leads to a "long-tail distribution." RUGD [15] and RELLIS-3D [16] are two recent datasets available for off-road semantic segmentation. The RUGD dataset consists of various scenes like trails, creeks, parks, and villages with fine-grained semantic segmentation annotations. The RELLIS-3D dataset is derived from RUGD and includes unique terrains like water puddles. In addition, RELLIS-3D includes lidar data and 3D lidar annotations. We use these datasets to evaluate the performance of our algorithm.

TABLE I: **Texture-based Terrain Classification:** We show an example of a grouping of terrain classes based on their texture and navigability. Our approach is general and designed for dynamic grouping in which different terrains may be re-classified into a different hierarchy level depending on the type of robot.

Hierarchy level	Classes
Smooth	Concrete, Asphalt
Rough	Gravel, Grass, Dirt, Sand
Bumpy	Rock, Rock Bed
Forbidden	Water, Bushes
Obstacles	Trees, Poles, Logs, etc.
Background	Void, Sky, Sign

III. Unstructured Environments: Levels of Navigability

There are certain characteristics, *e.g.*, texture, color, temperature, etc., that highlight the differences between various terrains. For robot navigation, it is important to identify different terrains based on the navigability in that terrain, which is indicated by its texture [13], [2], [22]. Furthermore, there may be terrains with similar textures and navigability that similarly affect navigation capabilities, but may pose challenges to visual perception systems such as semantic segmentation (due to the long-tailed distribution problem). In such cases, it is advantageous to group terrains with similar navigability into a single category. Given a dataset with C different semantic classes, we regroup them into G classes using the following criteria.

Terrain navigability varies for different robot systems. For instance, large autonomous vehicles may be better able to navigate on dirt roads than smaller mobile robots that would find such dirt roads a rough surface to traverse. Therefore, our approach is generally designed to identify different groupings of terrain categories and is not restricted to a fixed set of groupings. In Table I, we show an example of a grouping based on terrain texture and other characteristics. Below, we describe in detail the characteristics of different terrains based on their texture and navigability.

- *Concrete, Asphalt (Smooth)*: Concrete and asphalt are smooth terrains and are commonly found in urban roads. These terrains correspond to the "flat" category in the grouping provided by the CityScapes [5] autonomous driving dataset. These terrains are navigable in outdoor environments [13], and most mobile robots should be able to navigate in these terrains.
- *Gravel, Grass, Dirt, Sand (Rough)*: These terrains have been termed as rough on account of the increased friction while traversing these terrains [2], [41]. Most existing perception modules in off-road environments either [22], [23] do not consider these factors or may conservatively avoid such regions [13], [21] resulting in a sub-optimal solution, particularly in off-road navigation environments. In order to handle a large class of navigation methods, we want to be able to explicitly distinguish between smooth surfaces and rough surfaces. For example, in the presence of a smooth navigable region, a planning algorithm could prioritize it over a rough navigable region to reduce the energy loss.
- *Small rocks, Rock-bed (Bumpy)*: Autonomous vehicles or large robots may be able to navigate through rocks, while smaller robots with weaker off-road capabilities may face issues in such terrains [13], [22], [23], [3]. We provide a safe and flexible option where the planning scheme can be customized according to different scenes and different hardware characteristics of the robot. Specifically, this level of grouping can be ignored for off-road robots or vehicles by re-allocating these terrains to a different navigability group such as rough (for larger robots) or forbidden (for smaller robots). Our approach is general and handles dynamic grouping.
- *Water, Bushes, etc. (Forbidden Regions)*: These are regions that the agent must avoid to prevent causing damage to the robot hardware.
- *Obstacles*: Detecting obstacles such as trees, poles, etc. is critical for safe navigation. There has been a lot of work [3], [23], [22] on obstacle and hazardous terrain detection.
- *Background*: We use this buffer group to include background and non-navigable classes, including void, sky, and sign, that are not commensurate with any of the earlier definitions.

IV. Overview

We present GANav, an approach for classifying the navigability of different terrains in off-road environments via coarse-grained segmentation. The architecture of our deep neural network is motivated by the Bottleneck Transformer due to its superior performance over prior state-of-the-art methods in instance segmentation [42], [43], image classification [44], and object detection [38], [40].

To address the limitations of long-tailed distributions of terrain classes in such environments, our method leverages group segmentation via a group-wise attention mechanism. Our mechanism is a combination of group-wise attention heads and a novel attention loss function. In addition, we introduce a dynamically weighted cross-entropy loss function that can alleviate the long-tailed distribution issue via the notion of class weights. Our approach is general and can handle dynamic groupings depending on the robot hardware and terrain classes.

The rest of this section is structured as follows. We begin by stating our problem in Section IV-A. We then discuss the various components of our approach, namely the Bottleneck Transformer-based architecture (Section IV-B), group-wise attention head (Section IV-C), and an attention-based loss function (Section IV-D).

A. Problem Definition

The input consists of an RGB image $I \in \mathbb{R}^{3\times H\times W}$ and the corresponding ground-truth semantic segmentation labels $Y \in \mathbb{Z}^{H\times W}$ (provided by the dataset) denoting the category to which each pixel belongs among C different classes. After grouping these C classes into G groups, we obtain new group-wise labels $Y_G \in \mathbb{Z}^{H\times W}$. For each group, we also compute the binary mask $Y_{Bi} \in \{0,1\}^{H\times W}$ for $i = 1, ..., G$.

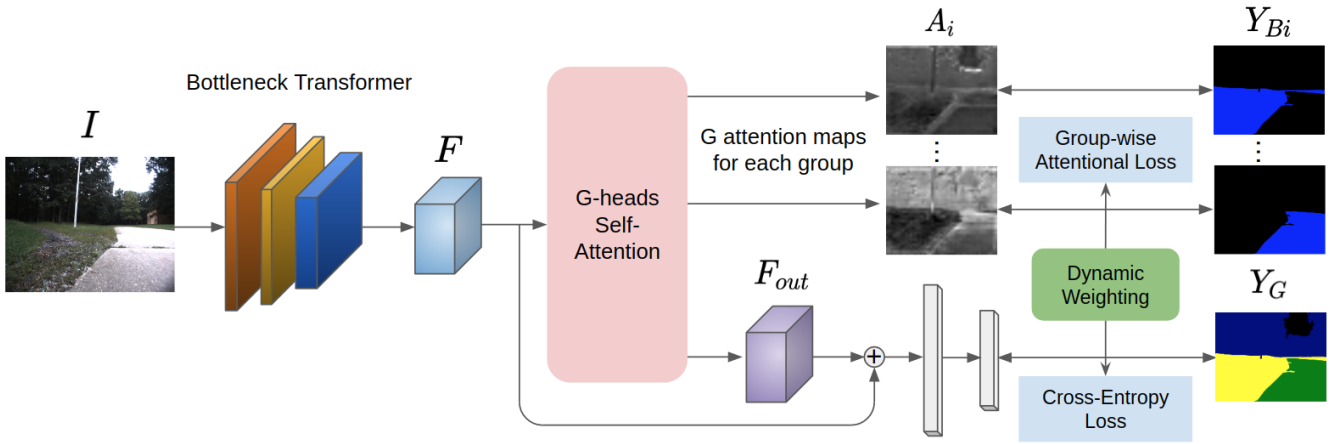


Fig. 2: **Architecture of our proposed network GANav:** We build on the transformer attention backbone [40] and introduce a novel group-wise attention head. The group-wise attention loss acts between the multi-head attention outputs corresponding to each group and the corresponding group-wise binary ground truth to explicitly guide the network towards accurate predictions of the different groups. In addition, we incorporate a dynamically weighted cross entropy-loss to handle the long-tailed nature of the dataset.

Our goal is to perform semantic segmentation on I and compute probability maps $P \in \mathbb{R}^{G \times H \times W}$. Each entry in P is a probability distribution of a pixel in that entry location belonging to one of the G classes. G could be dependent on the nature of the task; for example, an off-road navigation task should have more fine-grained classes on different types of road. In contrast, in the context of collision avoidance, the focus should be on different kinds of obstacles for location and moving capability.

B. Bottleneck Transformer

Bottleneck Transformers (BoTs) [40] are a type of deep neural network (DNN) for feature extraction and are designed as an alternative to the widely used ResNet [40] architecture via the Multi-Head Self-Attention (MHSA). MHSA, or self-attention applied on an image feature map, is the main component of BoTs, contributing to much of their success over ResNets in computer vision-based tasks. Our approach uses a BoT-based backbone architecture and leverages the MHSA module to individually apply self-attention to each group of terrain classes. Given an input feature vector A_{in}, the output self-attention map A_{out} is computed as follows:

$$A_{out} = softmax(f(A_{in})^T \bigodot g(A_{in}))^T \bigodot h(A_{in})$$

, where $f(A_{in})$, $g(A_{in})$, $h(A_{in})$ represent the key, query, and value feature maps in the self-attention [10] literature.

The RGB image I is passed through the BoT feature extractor to obtain intermediate feature maps, denoted by $F \in \mathbb{R}^{D_{in} \times H_f \times W_f}$. This feature map F is reshaped (and transpositioned) into $F_{flat} \in \mathbb{R}^{H_f W_f \times D_{in}}$, which is passed through the MHSA block with G attention heads. The MHSA component generates G attention maps (one for each group), $A_1, A_2, \ldots, A_G \in [0,1]^{H_f W_f \times H_f W_f}$, as well as the new feature maps $F_{out} \in \mathbb{R}^{D_{out} \times H_f \times W_f}$. The final output $P \in \mathbb{R}^{G \times H \times W}$ is obtained through a standard procedure of a segmentation network by a series of 1×1 convolutions and up-sampling from F_{out}. In the next section, we will explain our proposed detection head.

C. Group-wise Attention Mechanism

Attention maps can capture the relevancy between two pixel locations. Given an attention map A_i, its main diagonal represents the relevance of each location with respect to the attention head h_i. In our detection head, we use MHSA to extract features F (blue cuboid in Figure 2) and G attention maps $A_1, A_2, \ldots, A_G$. As demonstrated in Figure 2, we use A_i as an additional branch in the detection head and train the detection head to resemble the group distribution by explicitly guiding each attention map towards a corresponding category using a binary cross-entropy loss. Intuitively, $A_{i,[x,y]}$ represents the amount of attention that the pixel at the x^{th} position needs to pay to the pixel at the y^{th} position. Attention networks thus strive to learn feature maps that emphasize the relevance between a pair of locations, i.e., locations with high relevance have a high softmax score and vice versa.

D. Attention Loss Function

Based on the structure of our group-wise attention head, we propose group-wise attention loss. For each attention head h_i, we have its corresponding attention map $A_i \in [0,1]^{L \times L}$, where $L = H_f \times W_f$. We take the main diagonal of A_i and reshape it to $B_i \in [0,1]^{H \times W}$ using bi-linear image resizing. Each pixel in B_i represents the self-attention score with respect to h_i.

To guide each attention-map in the multi-head self-attention module, we apply a binary cross-entropy loss function:

$$\mathcal{L}_{\text{CA}} = -\sum_{h,w} y_{\text{G}} \log(B_i), \tag{1}$$

where $y_G \in Y_G$. This equation calculates loss between the predictions of the self-attention output and the corresponding group's binary ground truth with respect to a group.

The purpose of group-wise attention loss is not using attention maps to accurately predict the group's distribution; instead, as an intermediate layer, it aims to provide guidance

on which region an attention head should focus on for further classification.

In concurrence with the tradition in segmentation models, we optimize our network with a multi-class cross-entropy Loss and an auxiliary loss:

Cross-Entropy Loss: This is the standard semantic segmentation cross-entropy loss defined as follows:

$$\mathcal{L}_{\text{CE}} = -\sum_{h,w}\sum_{c\in\mathcal{C}} y_{\text{GT}}\log(P), \tag{2}$$

where h, w represent the dimensions of the image, C represents the number of groups, P denotes the output probability map, and y_{GT} corresponds to the ground-truth annotations.

Auxiliary Loss via Deep Supervision: Deep supervision was first proposed in [45] and has been widely used in the task of segmentation [17], [19], [18]. Adding an existing good-performing segmentation decoder head like FCN [46] in parallel to our GA head during training not only provides strong regularization but also reduces training time.

E. Dynamic weighting

Models trained on datasets with long-tailed distribution of classes can be biased towards the most frequent classes. There are many techniques [47], [48] to re-weight classes or effective samples to deal with label corruption and long-tail distribution by giving priority to rare classes. To assist the model in learning feature representations for all classes uniformly, we introduce a weighted cross-entropy loss function in which the weights are determined by a new dynamic weighting strategy based on class error rate.

Let the weights at the first epoch be $W_{init} \in \mathbb{R}^G$ and the weights at the i^{th} epoch be $W_i \in \mathbb{R}^G$. After d more epochs, the weight update rule is given as:

$$\mathcal{W}_{i+d} = m * W_i + (1-m) * (W_{init} + W^d_{error}), \tag{3}$$

where m is the momentum and $W^d_{error} \in [0,1]^G$ is the error rate of each group during the intermediate d epochs.

The momentum term assures that the change in weights is not drastic. Thus, the dynamic weighting scheme increments the weights in small steps. Additionally, our method can adaptively assign weights in accordance with the stage of training. The usage of the initial weights in the second term prevents bias towards a specific class. Our dynamic weighting scheme aims to make small adjustment based on the initial weighting from the hyper-parameters.

V. Results and Analysis

A. Implementation Details

We pre-train the bottleneck transformer backbone on RUGD [15] and RELLIS-3D [16] and use the weights to initialize the model for all our experiments. Optimization is done using a stochastic gradient descent optimizer with a learning rate of 0.01 and a decay of 0.0005. We adopt the polynomial learning rate policy with a power of 0.9. We augment the data with horizontal random flip and random crop. In the ablations and comparisons, for a fair study, we benchmark all our models at 80K iterations. For the RUGD dataset, we use a batch size of 8 and a crop size of 300×375 (the resolution of the original image is 688×550). On the other hand, for the RELLIS-3D dataset, due to the high resolution of the image, we train models with a batch size of 2 and crop size 375×600 (the resolution of the original image is 1920×1600).

Groups (Accuracy)	OCRNet [18]	OCR-G	GANet
6 Groups			
mean IoU	64.05	76.5	83.24
mean accuracy (per class)	71.77	84.97	89.84
All pixels accuracy	91.47	93.46	94.26
4 Groups			
mean IoU	77.95	81.0	82.8
mean accuracy (per class)	84.04	88.0	89.14
All pixels accuracy	93.11	93.7	94.25

TABLE II: **Effect of grouping on the RUGD test set:** The first column shows the results of current SOTA method OCRNet [18] on the 24-class setting and the second column shows results of OCRNet [18] on our newly proposed taxonomy. We show how training by grouping can improve the performance of segmentation models, thus proving that grouping is indeed beneficial. In the third column, we show that our architecture results in improved performance over SOTA.

B. Evaluation Metrics and Baselines

We evaluate our models on the standard segmentation metrics: Intersection over Union (IoU), mean IoU (mIoU, IoU averaged over all classes), and average pixel accuracy (aAcc). We compare our method with the following baseline methods:

- PSPNet [17] (2017): Pyramid Scene Parsing Network (PSPNet) uses a pyramid pooling module to extract global and local feature and makes predictions based on the aggregated information.
- DeepLabv3+ [9] (2018): The latest version of DeepLab is DeeepLabv3+, which utilizes an encoder-decoder structure based on multi-scale atrous convolution operations and progressive up-sampling during the decoding stage.
- PSANet [19] (2018): Point-wise Spatial Network (PSANet) utilizes bi-directional attention maps to aggregate point information.
- DANet [10] (2019): Dual Attention Network (DANet) uses ResNet as a backbone and proposes a position attention module and a channel attention module in the detection head.
- OCRNet [18] (2020): Object-Contextual Representation Network (OCRNet) uses relational context information and a soft object region as an augmentation on the original features for a better representation of objects and areas.

C. Effect of Grouping

In this section, we analyze the effect of evaluating models based on our new taxonomy of classes. The empirical results

6 Groups									
Dataset	Methods (IoU)	Smooth Region	Rough Region	Bumpy Region	Forbidden Region	Obstacle	Background	mIoU	aAcc
RUGD	PSPNet [17]	48.62	88.92	69.45	29.07	87.98	78.29	67.06	92.85
	DeepLabv3+ [9]	5.86	84.99	50.40	25.04	87.50	81.47	55.88	91.51
	DANet [10]	2.26	81.47	8.69	15.00	82.54	74.86	44.14	88.81
	OCRNet [18]	66.29	89.47	76.15	59.14	88.77	79.17	76.50	93.46
	PSANet [19]	34.92	87.70	35.64	8.66	86.95	78.97	55.47	92.13
	GANav (ours)	**86.00**	**90.88**	**80.26**	**72.82**	**89.67**	**79.78**	**83.24**	**94.26**
RELLIS-3D	PSPNet [17]	69.21	80.99	8.89	53.7	60.7	94.67	61.36	86.01
	DeepLabv3+ [9]	65.76	79.84	19.72	47.52	64.88	95.92	62.27	85.84
	DANet [10]	72.93	85.18	13.10	60.60	70.53	**95.95**	66.38	89.11
	OCRNet [18]	**74.67**	83.04	27.76	60.44	62.35	92.58	66.81	86.95
	PSANet [19]	64.06	75.29	17.08	47.45	61.74	94.31	59.99	83.17
	GANav (ours)	73.63	**86.72**	**29.08**	**67.33**	**71.49**	95.52	**70.63**	**90.23**

4 Groups							
Dataset	Methods (IoU)	Navigable Region	Forbidden Region	Obstacle	Background	mIoU	aAcc
RUGD	PSPNet [17]	90.22	60.56	88.82	79.06	79.66	93.91
	DeepLabv3+ [9]	90.47	57.64	**89.35**	82.33	79.94	**94.27**
	DANet [10]	90.0	66.84	89.19	79.33	81.56	94.18
	OCRNet [18]	89.94	66.29	88.24	**79.54**	81.0	93.7
	PSANet [19]	90.48	68.92	88.96	79.39	81.94	94.02
	GANav (ours)	**91.1**	**71.37**	89.28	79.44	**82.8**	94.25
RELLIS-3D	PSPNet [17]	79.11	46.94	72.18	94.11	73.09	85.13
	DeepLabv3+ [9]	70.36	42.71	**74.57**	95.39	70.76	82.28
	DANet [10]	75.89	43.75	69.27	94.5	70.85	83.26
	OCRNet [18]	80.18	44.57	73.99	95.17	73.48	85.73
	PSANet [19]	77.3	46.52	72.91	94.94	72.91	84.56
	GANav (ours)	**85.9**	**59.67**	69.15	**95.56**	**77.57**	**89.8**

TABLE III: **Comparison with state-of-the-art methods on RUGD and RELLIS-3D test sets:** We compare the performance of our learning method with other methods. Our GANav algorithm improves the state-of-the-art mIoU by $6.74-39.1\%$ on RUGD and $3.82-10.64\%$ on RELLIS-3D on 6 groups (top). Further, we show that our grouping method can improve accuracy of classes like Smooth Region, Bumpy Region, and Forbidden regions by large margins, which is crucial for safe navigation. We also highlight the results and improvements with using 4 groups (bottom), the overall difficulty of the task is reduced.

are presented in Table II. We measure the effect of grouping with two grouping strategies relevant to navigation:

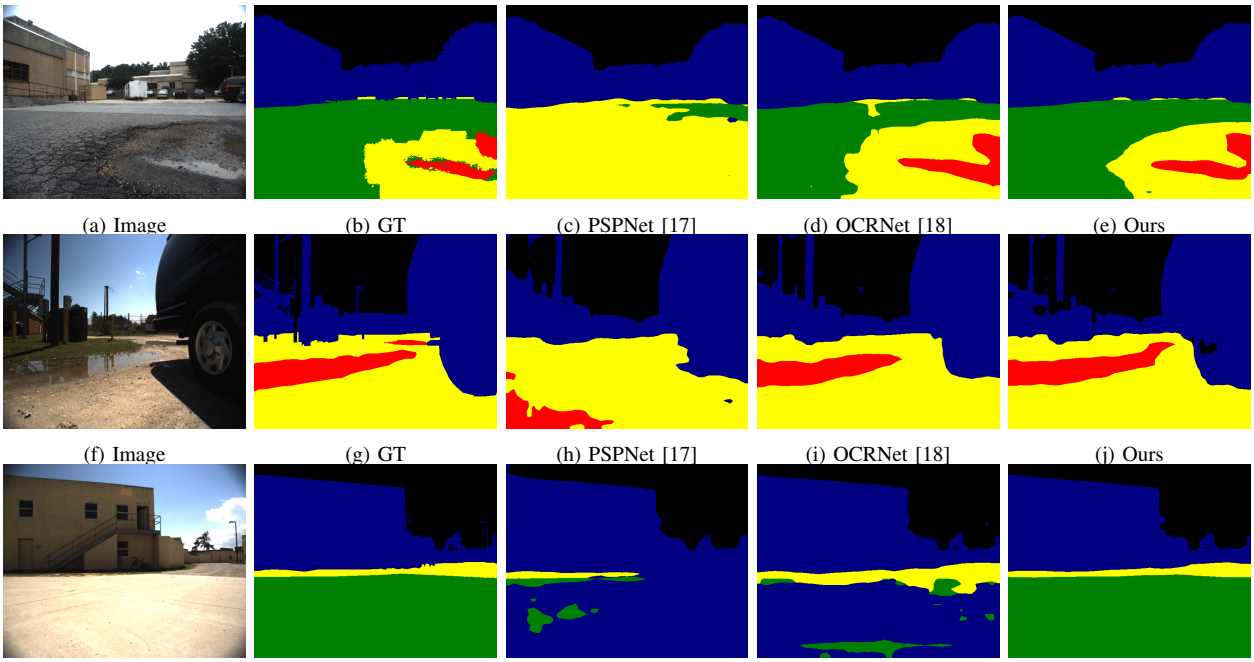
- 6 groups: Smooth region, Rough region, Bumpy region, Forbidden region, Obstacle, and Background
- 4 groups: Navigable region (encompassing smooth regions, rough regions, and bumpy regions defined in the 6 groups case), Forbidden region, Obstacle, and Background

a) Experimental setup: We train one previous SOTA method, OCRNet [18], and GANav (ours), under two settings: 6 groups and 4 groups. We evaluate ungrouped scenario with OCRNet and the grouped scenario on both models. Note that our method only works in grouped scenario, as group-wise attention requires extensive GPU memory with original 24 or 20 classes.

b) Analysis: We show that such a grouping strategy improves the performance of current semantic segmentation approaches in off-road terrains by reducing the error rate (or improving accuracy) in long-tail distributed data. We wish to emphasize the intuition behind this through an example. Consider the classes "concrete" and "asphalt", which are highly similar. Under the ungrouped setting, classifying them separately can lead to errors due to homogeneous feature embeddings. However, under the grouped setting, "asphalt" and "concrete" are categorized within the same group, thus providing the network with more flexibility and better margins for learning feature mappings. More error analysis of grouping with other methods can be find in [49].

D. Results and state-of-the-art comparisons

The results on 6 groups and 4 groups are presented in Table III. Our method works extremely well on more complicated grouping (6 groups), on which we will analyze the performance. Our architecture, which uses multi-head self-attention to explicitly focus on different groups through a group-wise attention loss, improves the state-of-the-art mIoU and aAcc by $6.74-39.1\%$ and $1.41-5.45\%$ on RUGD. We notice an improvement of 13.68% on the "forbidden regions" group, which is critical for safety applications. Additionally, improvements of 4.11% and 19.71% on "bumpy regions" and "smooth regions," respectively, provide our network with the ability to facilitate better navigation. Similarly, on the RELLIS dataset, GANav demonstrates an improvement of $4.82-10.64\%$ and $1.12-7.06\%$ on mIoU and aAcc. Some visualizations on our method and comparison methods are provided in figure 3.



(a) Image (b) GT (c) PSPNet [17] (d) OCRNet [18] (e) Ours

(f) Image (g) GT (h) PSPNet [17] (i) OCRNet [18] (j) Ours

(k) Image (l) GT (m) PSPNet [17] (n) OCRNet [18] (o) Ours

Fig. 3: **Qualitative Results:** Each row highlights the original image, ground truth (GT) label, prediction of two SOTA methods and prediction of our method. We observe that our method is able to detect different navigable regions and is robust with respect to objects with similar color, as shown in image (k).

Dataset	CW-Attn.	D.W.	mIoU	mAcc	aAcc
RUGD			74.32	80.41	93.53
	✓		83.24	89.84	94.26
	✓	✓	83.01	89.46	94.12
RELLIS-3D			63.01	68.7	92.2
	✓		68.27	77.56	88.77
	✓	✓	70.63	81.32	90.23

TABLE IV: Ablation studies on the RUGD and RELLIS-3D test sets. We observe that the group-wise attention mechanism is successful in improving performance by 8.92% and 7.62% on the RUGD and RELLIS datasets, respectively. We also notice that DW is effective in datasets, like RELLIS-3D, that have a long-tailed distribution even after grouping.

E. Ablation Studies

The ablation studies are presented in Table IV. We conduct experiments on both RUGD and RELLIS to show the effects of group-wise attention and dynamic grouping. We observe that group-wise attention improves performance (in terms of mIoU) by 8.92% and 7.62%, respectively. Using the smaller number of groups, as opposed to training on all 24 classes, enables us to define separate attention heads for each of the groups, which improves the overall performance. Dynamic weighting used in our group-wise attention model results in an improvement of 2.36% on the RELLIS-3D dataset. Dynamic weighting does not lead to an improvement in performance for RUGD, because it is more effective when the dataset has a long-tailed distribution.

VI. Conclusions, Limitations, and Future Work

In this paper, we present a learning-based method for discovering various terrain types in off-road environments as well as the possibility of a general framework for improving perception accuracy by grouping classes with close semantic meaning in a specific task. We also propose a novel segmentation network that can provide robust prediction and distinguish regions with different semantic meaning through use of group-wise attention. We present a new techniques called dynamic class weighting to handle the long-tailed distribution. We demonstrate considerable improvement in accuracy over SOTA on complex, unstructured terrain datasets.

Our approach has some limitations. We need to make sure the grouping scheme is appropriate for a specific navigation task and dataset. Our current performance analysis is based on mIoU metric and we may need to evaluate these methods with appropriate navigation metrics. In addition, we need to make sure that the data is labeled correctly and consistently. For example, trees can either be an obstacle or an object in the background. An annotated dataset might not label them separately.

As part of our future work, we need to consider how to measure the accuracy in terms of navigation metrics. We want to explore how this learning method can be applied and deployed to real-world robot navigation, and design other meaningful groups for different tasks. We also need to extend our approach to different sensors or multiple sensor inputs.

Groups (Accuracy)	PSPNet [17]			DeepLabv3+ [9]			OCRNet [18]		
	w/o. Grouping	w. Grouping	↑(%)	w/o. Grouping	w. Grouping	↑(%)	w/o. Grouping	w. Grouping	↑(%)
6 Groups									
Smooth Navigable Region	8.36	50.95	**42.59**	4.88	5.94	1.06	58.94	70.65	**11.71**
Rough Navigable Region	95.4	95.05	−0.35	92.21	95.98	3.77	93.32	96.1	2.78
Bumpy Region	19.81	81.46	**61.65**	15.67	74.67	**59**	51.65	91.56	**39.91**
Forbidden Region	46.35	65.47	**19.12**	68.7	71.74	3.04	51.12	71.85	**20.73**
Obstacle	93.84	94.03	0.19	95.61	92.16	−3.45	93.08	93.57	0.49
Background	84.8	85.73	0.93	85.62	86.93	1.31	82.49	86.11	3.62
mean IoU	50.16	67.06	**16.9**	47.14	55.88	8.74	64.05	76.5	**12.45**
mean accuracy (per class)	58.09	78.78	**20.69**	60.45	71.24	**10.79**	71.77	84.97	**13.2**
All pixels accuracy	91.8	92.85	1.05	91.48	91.51	0.03	91.47	93.46	1.99
4 Groups									
Navigable Region	95.36	95.17	−0.19	89.57	96.03	6.46	92.47	96.88	4.41
Forbidden Region	46.35	76.35	**30**	68.7	79.42	**10.72**	67.8	75.71	7.91
Obstacle	93.84	94.49	0.65	95.61	93.72	−1.89	95.83	92.32	−3.51
Background	84.8	84.65	−0.15	85.62	89.38	3.76	80.05	87.09	7.04
mean IoU	70.03	79.66	9.63	66.41	79.94	**13.53**	77.95	81.0	3.05
mean accuracy (per class)	80.09	87.66	7.57	84.88	89.64	4.76	84.04	88.0	3.96
All pixels accuracy	93.66	93.91	0.25	92.31	94.27	1.96	93.11	93.7	0.59

TABLE V: **Effect of grouping on the RUGD test set:** The first column under each method shows the results on the 24-class setting, the second column shows results on the 6-class setting and the third column shows the improvement that the grouping scheme imparts to the network.

APPENDIX

VII. More Analysis on Effect of Grouping

In this section, we include more comparison results on the effect of grouping in V. In most of the case, the grouping can improve the accuracy drastically up to 61.65% in some groups. It's inevitable that for some groups, the performance degraded within a margin of 3.5%. Overall, with grouping, we have a higher accuracy during training without losing the necessary information, as we distinguish all semantically different labels that are relevant to the task.

VIII. More Visualizations in Different Scenario

We include more visual results from our methods on RUGD in figure 4 and RELLIS-3D in figure 5.

References

[1] M. J. Procopio, J. Mulligan, and G. Grudic, "Learning terrain segmentation with classifier ensembles for autonomous robot navigation in unstructured environments," *Journal of Field Robotics*, vol. 26, no. 2, pp. 145–175, 2009.

[2] K. Zhao, M. Dong, and L. Gu, "A new terrain classification framework using proprioceptive sensors for mobile robots," *Mathematical Problems in Engineering*, vol. 2017, pp. 1–14, 09 2017.

[3] R. A. Hewitt, A. Ellery, and A. de Ruiter, "Training a terrain traversability classifier for a planetary rover through simulation," *International Journal of Advanced Robotic Systems*, vol. 14, no. 5, p. 1729881417735401, 2017.

[4] L. Rummelhard, J. Lussereau, J.-A. David, C. Laugier, S. Dominguez, G. Garcia, and P. Martinet, "Perception and Automation for Intelligent Mobility in Dynamic Environments," in *ICRA 2017 Workshop on Robotics and Vehicular Technologies for Self-driving cars*, Singapore, Singapore, June 2017.

[5] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[6] H. Alhaija, S. Mustikovela, L. Mescheder, A. Geiger, and C. Rother, "Augmented reality meets computer vision: Efficient data generation for urban driving scenes," *International Journal of Computer Vision (IJCV)*, 2018.

[7] R. R. Murphy, *Disaster robotics*. MIT press, 2014.

[8] R. R Shamshiri, C. Weltzien, I. A. Hameed, I. J Yule, T. E Grift, S. K. Balasundram, L. Pitonakova, D. Ahmad, and G. Chowdhary, "Research and development in agricultural robotics: A perspective of digital farming," 2018.

[9] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[10] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[11] M. D. Zapata, O. Erkent, and C. Laugier, "Instance segmentation with unsupervised adaptation to different domains for autonomous vehicles," in *ICARCV 2020-16th International Conference on Control, Automation, Robotics and Vision*, 2020.

[12] A. Paigwar, O. Erkent, D. S. Gonzalez, and C. Laugier, "Gndnet: Fast ground plane estimation and point cloud segmentation for autonomous vehicles," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020.

[13] S. Matsuzaki, K. Yamazaki, Y. Hara, and T. Tsubouchi, "Traversable region estimation for mobile robots in an outdoor image," *J. Intell. Robotics Syst.*, vol. 92, no. 3–4, p. 453–463, Dec. 2018.

[14] R. Manduchi, A. Castano, A. Talukder, and L. Matthies, "Obstacle detection and terrain classification for autonomous off-road navigation," *Autonomous Robots*, vol. 18, no. 1, pp. 81–102, jan 2005.

[15] M. Wigness, S. Eum, J. G. Rogers, D. Han, and H. Kwon, "A rugd dataset for autonomous navigation and visual perception in unstructured outdoor environments," in *International Conference on Intelligent Robots and Systems (IROS)*, 2019.

[16] P. Jiang, P. Osteen, M. Wigness, and S. Saripalli, "Rellis-3d dataset: Data, benchmarks and analysis," *arXiv preprint arXiv:2011.12954*, 2020.

[17] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2881–2890.

[18] Y. Yuan, X. Chen, and J. Wang, "Object-contextual representations for semantic segmentation," in *16th European Conference Computer Vision (ECCV 2020)*, August 2020.

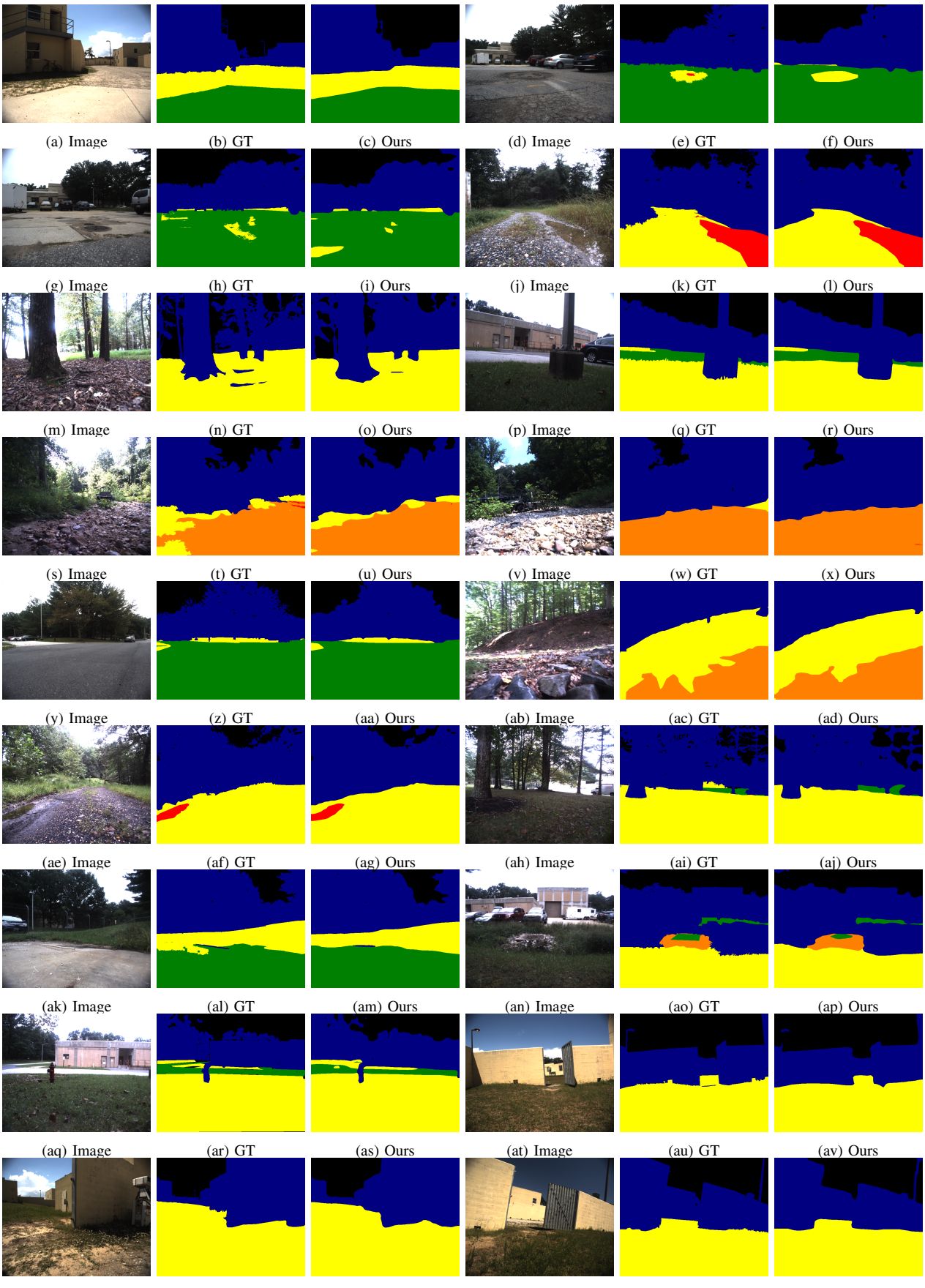
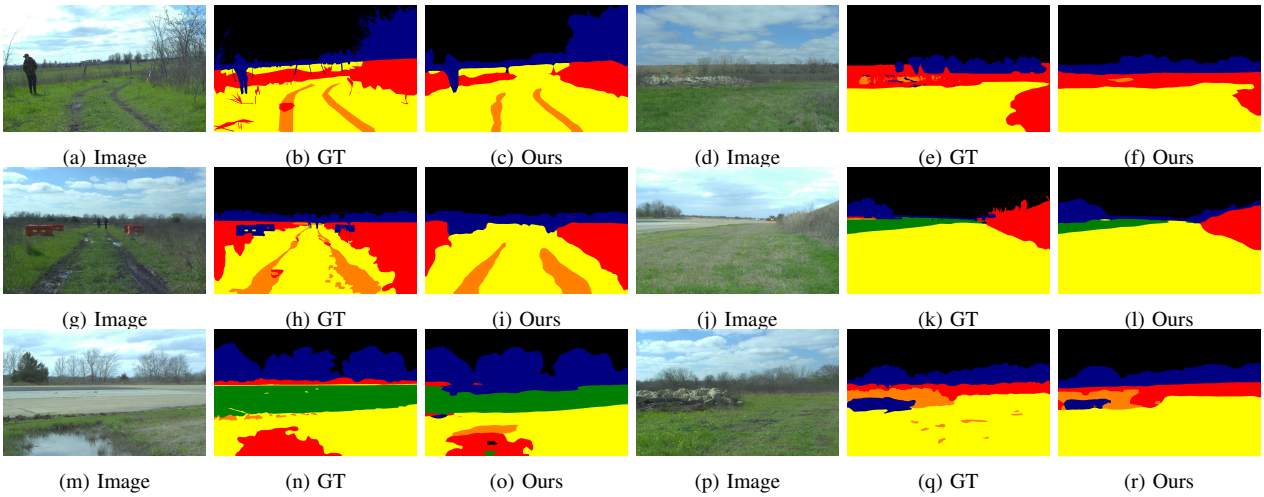


Fig. 4: **More Qualitative Results on RUGD**



(a) Image (b) GT (c) Ours (d) Image (e) GT (f) Ours

(g) Image (h) GT (i) Ours (j) Image (k) GT (l) Ours

(m) Image (n) GT (o) Ours (p) Image (q) GT (r) Ours

Fig. 5: **More Qualitative Results on RELLIS-3D**

[19] H. Zhao, Y. Zhang, S. Liu, J. Shi, C. C. Loy, D. Lin, and J. Jia, "Psanet: Point-wise spatial attention network for scene parsing," in *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[20] G. Kahn, P. Abbeel, and S. Levine, "Badgr: An autonomous self-supervised learning-based navigation system," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 1312–1319, 2021.

[21] G. Pereira, L. Pimenta, L. Chaimowicz, A. Fonseca, D. Almeida, L. Corrêa, R. Mesquita, and M. Campos, "Robot navigation in multi-terrain outdoor environments," vol. 28, 01 2006, pp. 331–342.

[22] G. Reina, A. Milella, and R. Rouveure, "Traversability analysis for off-road vehicles using stereo and radar data," in *2015 IEEE International Conference on Industrial Technology (ICIT)*, 2015, pp. 540–546.

[23] J. Larson, M. Trivedi, and M. Bruch, "Off-road terrain traversability analysis and hazard avoidance for ugvs," 2011.

[24] G. Wilson, "Towards speed selection for high-speed operation of autonomous ground vehicles on rough off-road terrains," 2018.

[25] A. Vicente, Jindong Liu, and Guang-Zhong Yang, "Surface classification based on vibration on omni-wheel mobile base," in *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2015, pp. 916–921.

[26] E. DuPont, C. Moore, E. Collins, and E. Coyle, "Frequency response method for terrain classification in autonomous ground vehicles," *Autonomous Robots*, vol. 24, pp. 337–347, 05 2008.

[27] G. Wilson, "Terrain roughness identification for high-speed ugvs," *Journal of Automation and Control Research*, vol. 1, pp. 11–21, 10 2014.

[28] S. Matsuzaki, K. Yamazaki, Y. Hara, and T. Tsubouchi, "Traversable region estimation for mobile robots in an outdoor image," *J. Intell. Robotic Syst.*, vol. 92, no. 3-4, pp. 453–463, 2018.

[29] J. Xue, H. Zhang, K. J. Dana, and K. Nishino, "Differential angular imaging for material recognition," *CoRR*, vol. abs/1612.02372, 2016.

[30] P. Filitchkin and K. Byl, "Feature-based terrain classification for littledog," in *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2012, pp. 1387–1392.

[31] N. Hirose, A. Sadeghian, M. Vázquez, P. Goebel, and S. Savarese, "Gonet: A semi-supervised deep learning approach for traversability estimation," 2018.

[32] A. Valada, J. Vertens, A. Dhall, and W. Burgard, "Adapnet: Adaptive semantic segmentation in adverse environmental conditions," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*, 2017, pp. 4644–4651.

[33] T. Ort, L. Paull, and D. Rus, "Autonomous vehicle navigation in rural environments without detailed prior maps," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, 2018, pp. 2040–2047.

[34] G. Varma, A. Subramanian, A. Namboodiri, M. Chandraker, and C. V. Jawahar, "Idd: A dataset for exploring problems of autonomous navigation in unconstrained environments," 2018.

[35] S. Hao, Y. Zhou, and Y. Guo, "A brief survey on semantic segmentation with deep learning," *Neurocomputing*, vol. 406, pp. 302–321, 2020.

[36] C. Peng, X. Zhang, G. Yu, G. Luo, and J. Sun, "Large kernel matters–improve semantic segmentation by global convolutional network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4353–4361.

[37] M. Everingham, L. Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *Int. J. Comput. Vision*, vol. 88, no. 2, p. 303–338, June 2010.

[38] T. Lin, M. Maire, S. J. Belongie, L. D. Bourdev, R. B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: common objects in context," *CoRR*, vol. abs/1405.0312, 2014.

[39] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, "Scene parsing through ade20k dataset," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[40] A. Srinivas, T.-Y. Lin, N. Parmar, J. Shlens, P. Abbeel, and A. Vaswani, "Bottleneck transformers for visual recognition," *arXiv preprint arXiv:2101.11605*, 2021.

[41] C. Bai, J. Guo, G. Linli, and J. Song, "Deep multi-layer perception based terrain classification for planetary exploration rovers," *Sensors*, vol. 19, p. 18, 07 2019.

[42] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2980–2988.

[43] J. Dai, K. He, and J. Sun, "Instance-aware semantic segmentation via multi-task network cascades," in *CVPR*, 2016.

[44] H. Pham, Z. Dai, Q. Xie, M.-T. Luong, and Q. V. Le, "Meta pseudo labels," 2021.

[45] C.-Y. Lee, S. Xie, P. Gallagher, Z. Zhang, and Z. Tu, "Deeply-Supervised Nets," in *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, vol. 38. PMLR, 2015.

[46] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 4, pp. 640–651, 2017.

[47] Z. Zhang and M. Sabuncu, "Generalized cross entropy loss for training deep neural networks with noisy labels," in *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[48] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, and S. Belongie, "Class-balanced loss based on effective number of samples," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9268–9277.

[49] T. Guan, D. Kothandaraman, R. Chandra, and D. Manocha, "Ganav: Group-wise attention network for classifying navigable regions in unstructured outdoor environments," 2021.