2024 24th International Conference on Control, Automation and Systems (ICCAS)
Jeju Shinhwa World, Jeju, Korea. Oct 29 - Nov 1, 2024

Attention Regulation for Efficient Semantic Segmentation on Unstructured Terrain

Xiujin Liu†, Feng Xue†, and Xiaoxiao Du*

Department of Robotics, University of Michigan, Ann Arbor, MI, USA 48109
{jeanliu, fengxe, xiaodu}@umich.edu
† Equal Contribution * Corresponding author

Abstract: We present AR-Net, an efficient semantic segmentation pipeline for unstructured terrains. For applications such as autonomous navigation, it is essential to accurately and efficiently understand the unstructured scenes in outdoor and urban environments. Given RGB images as inputs, the AR-Net uses an encoder backbone to extract multi-scale features and a novel Attention-Regulation layer as part of the decoder to predict the pixel-level segmentation results for unstructured terrains. Our AR-Net model achieved superior segmentation performance and fast inference on two real-world outdoor terrain datasets. We also provide detailed ablation studies and analyses on model parameter selections.

Keywords: Deep learning, Semantic segmentation, Attention regulation, Encoder-decoder, Transformers, Unstructured environment, Terrain classification, real-time

1. INTRODUCTION

Unstructured terrains are challenging operational environments for mobile robots, often characterized by complex and diverse landscapes such as gravel paths, marshy terrain, and rugged mine pits. mobile robots operating in these conditions require traversability perception methods that are highly accurate, robust, and efficient [1].

Deep learning (DL) techniques, such as Convolution Neural Networks (CNNs) [2], [3] and Transformers [4], [5], [6], have been developed for semantic segmentation and terrain traversability classification [7]. However, previous methods often show relatively rough handling of edges and lacking robustness and generalizability when testing on a different scene or dataset. Besides, previous DL methods often require a large number of parameters and can be slow in inference.

To address these challenges, we propose an attention-regulation-based encoder-decoder neural network, *AR-Net*, that accurately and efficiently predicts pixel-level segmentation results across various unstructured terrain types. We develop a novel Attention-Regulation (AR) layer and design an efficient encoder-decoder network, *AR-Net*. We show that the AR layer can improve segmentation accuracy, especially on boundary areas, and can be easily integrated in other networks. We present experimental results on two real-world outdoor unstructured terrain datasets and demonstrates that our model can achieve high accuracy and efficiency in terrain segmentation, with improved robustness and generalizability when testing on a new dataset (unseen in training). We also provide comprehensive ablation studies on various parameters in the Attention-Regulation blocks, providing insights for model parameter selection and optimization.

2. RELATED WORK

Deep neural networks have seen significant progress in semantic segmentation. Fully Convolutional Networks (FCNs) [8] use stacked convolution layers to learn hierarchical features from images and make dense predictions for per-pixel tasks like semantic segmentation. U-Net [9], SegNet [10], and Fast-SCNN [11] are classic encoder-decoder models following FCNs, where the encoder uses convolutional and pooling layers to extract latent features from input images, and the decoder uses transposed convolutions to restore detailed features and output segmentation maps. PSPNet [12] introduces the pyramid pooling module to obtain global and local information through pooling operations at different scales. DeepLab models [13], [14], [15] integrate CNNs and probabilistic graphical models to refine segment boundaries and address multi-scale objects. GCNet [16] proposed a global context (GC) block based on 1 convolutions, which is lightweight and can effectively model the global context. ANNNet [17] uses asymmetric pyramid and fusion blocks for efficient semantic segmentation.

In addition to CNN-based methods, transformers [4] have also been used. SETR [18], based on the Vision Transformer (ViT) [19], models semantic segmentation as a sequence-to-sequence problem and replaced the existing CNN encoder with transformer layers. SegFormer [6] is another simple and efficient model, which integrated transformers with lightweight multi-layer perceptron (MLP)-based decoders. For terrain segmentation and navigable region identification, GANav [20] uses group-wise attention mechanism in transformers to identify safe and navigable regions in off-road terrains and unstructured environments.

Several gaps remain. First, prior work often show inaccurate segmentation predictions along edges and boundary areas, especially in complex terrains or at lower resolutions. Second, prior methods often lack robustness and do not generalize well when tested in new environments unseen in training. Third, efficiency (smaller model parameters, faster inference speed) is necessary for real-world applications such as autonomous navigation. In this work, we develop an encoder-decoder architecture with a novel attention-regulation layer to address the above gaps for improved terrain segmentation in both accuracy and efficiency.

3. METHOD

In this work, we propose a novel lightweight network architecture, named *AR-Net*, that integrates an attention-regulation layer after the feature fusion step for better segmentation performance. We also developed a novel loss function accounting for both cross-entropy loss and attention region loss.

3.1 Input and Output

The goal of *AR-Net* is to classify terrain traversability in outdoor environments via semantic segmentation. The inputs for the *AR-Net* are RGB images of size $3 \times H \times W$, where H and W are the height and width of the images. The images contain scenes of unstructured terrains in outdoor environments. The outputs of the *AR-Net* are predicted pixel-level segmentation probabilities for N_c semantic classes in the image. To associate the model performance with different terrain types and to enable the generalization of semantic labels across multiple datasets, we categorized the semantic labels into four coarse groups corresponding to different levels of traversability: Background/Obstacles (sky, people, sign, building, vehicle, etc.), Stable (smooth surface such as concrete and asphalt), Granular (sand, gravel, mud) and High resistance (bush, grass, log). Thus, $N_c = 4$.

3.2 Encoder Backbone

Fig. 1 illustrates the network architecture for *AR-Net*. The *AR-Net* has an encoder-decoder structure. The encoder part of the proposed *AR-Net* leverages Transformers [4] as backbones to extract and combine multi-scale features. In this work, we use the Mix Transformer (MiT) encoders from SegFormer [6] as our backbone, as it has shown good results on dense prediction tasks by using small patches and extracting multi-scale features. In our implementation, the MiT transformer blocks use image patches of size 7×7, stride 4 and padding 3 in the first block, and size 3×3, stride 2 and padding 1 for the next three blocks. All the blocks contribute to the spatial resolution reduction. The multi-scale features obtained after each block have spatial resolution of $H_i \times W_i = \frac{H}{2^{i+1}} \times \frac{W}{2^{i+1}}$ and feature channel $C_i = \{64, 128, 320, 512\}$,, where block index $i \in \{1, 2, 3, 4\}$.

3.3 Decoder Segmentation Head

The multi-scale encoder features are passed through a decoder to perform semantic segmentation. The decoder of the *AR-Net* consists of two parts. The first part uses MLPs to unify the channels from the encoder features, upsample them to the largest feature size, and concatenate the features. Then, another MLP layer is used to fuse the features together. The fused feature will be sent to two branches. The first branch uses another MLP layer to generate the segmentation result with resolution $\frac{H}{4} \times \frac{W}{4} \times N_c$. The second branch downsamples the fused features with resolution $F_H \times F_W \times C_e$, where $F_H = H/32$, $F_W = W/32$, C_e is the embedding dimension.

In the second down-sampling branch, a novel Attention-Regulation (AR) layer was proposed to generate attention map for the fused features. The AR layer's dimension is the same as embedding dimension C_e and has N_c heads (number of semantic classes). Each head corresponds to one semantic label. First, the fused features are flattened to F with shape $B \times C_e \times N$, where B is batch size and $N = F_H \times F_W$. Then, the attention map M can be derived from key and query matrices and the position embedding information following

$$M = \frac{Q_F K_F^T}{\sqrt{C_e/N_c}} + P. \tag{1}$$

The first term is based on the attention calculation in the original Transformer [4], where Q and K are the query and keys. The second term, P, is the position embedding. Empirically, summing the two terms together improved segmentation results compared with using only the attention term. The output attention map has shape $N_c \times (F_H * F_W) \times (F_H * F_W)$.

From the attention maps, we extract what we call "attention masks". The mask has shape $N_c \times F_H \times F_W$ and its values come from the main diagonal of the attention map. The c^{th} layer of the mask represents the prediction associated with each terrain category, $c \in \{1, ..., N_c\}$. The size of the attention mask, F_H and F_W, is determine by a parameter named "AR-Ratio" (the attention-regulation size ratio), where AR-Ratio= $\frac{F_H}{H}$, the ratio of the size of mask and raw image. We will provide detailed ablation study on the effect of AR-Ratio in Section 5.Also note that the proposed AR layer is a general module and can be easily added to other networks. We show improved segmentation performance when adding the AR layer to other networks in Section 4.

3.4 Loss Function

Our loss function has two terms. The first is cross-entropy loss between pixel-level semantic segmentation prediction result and ground truth labels, written as

$$L_1 = -\sum_{H,W} \sum_{N_c} Y log(\hat{Y}) \tag{2}$$

where Y denotes the ground truth semantic labels at each pixel and $\hat{Y}$ is the predicted segmentation output of the *AR-Net* model. The second term is a binary cross-entropy loss for the attention-regulation mask. This computes the loss between each channel in the mask and the corresponding semantic class, written as

$$L_2^c = -\sum_{H,W} Y log(B_c) \tag{3}$$

Where B_c is the attention score of the pixel corresponding to the c^{th} semantic class.

This novel binary cross-entropy loss term encourages the focus area of the attention-regulation layer to overlap as much as possible with the region associated with the ground truth semantic class, thus enhancing the segmentation accuracy. The binary cross-entropy loss also sets a

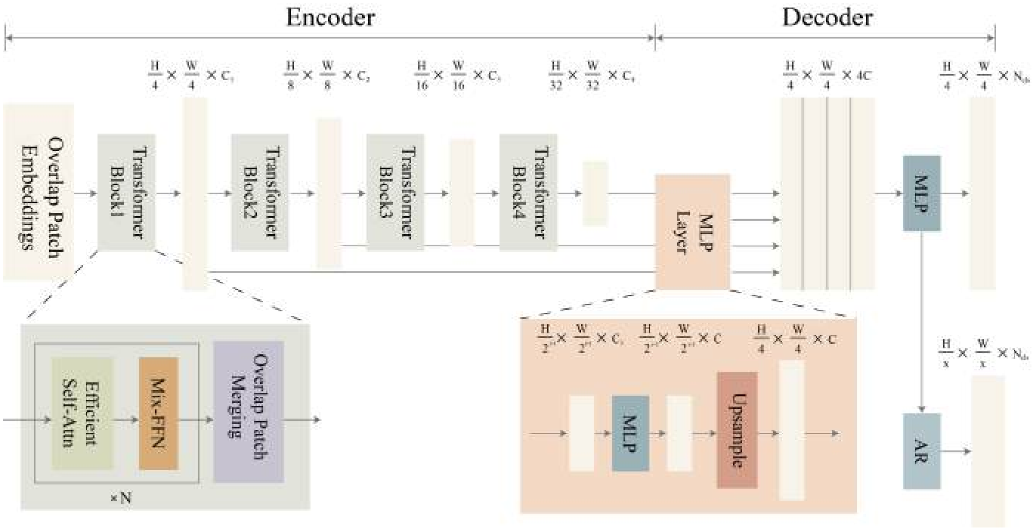

Fig. 1. The architecture of our proposed *AR-Net*. AR-Net has an encoder-decoder struture. In the encoder, SegFormer mix transformer (MiT) blocks were used as backbone. In the decoder, a novel attention-regulation block is used to extract semantic class-specific attention masks and produce segmentation results.

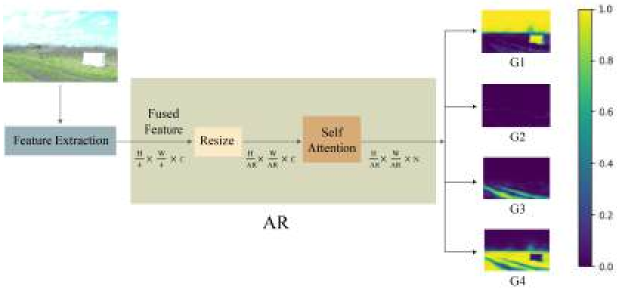

Fig. 2. Illustration of Attention-Regulation (AR) block described in Section 3.3. The four images (G1-G4) illustrate the attention masks associated with the four terrain types.

tighter restriction especially along the edges and boundary areas, where semantic class labels changes. The attention masks also enables visual interpretability.

The final loss function for the *AR-Net* is written as

$$L = -\sum_{H,W}\sum_{N_c} Y log(\hat{Y}) - \sum_{H,W}\sum_{c=1}^{N_c} Y log(B_c) \tag{4}$$

4. EXPERIMENTS

4.1 Datasets

Two datasets, RELLIS-3D and RUGD, were used in this work. RELLIS-3D [21] is a multi-modal off-road dataset collected on the Rellis Campus of Texas A&M University. We only use the RGB camera images in this work. The raw image size of RELLIS-3D is 1920×1200 and we resize it to 600×375 before passing it into the *AR-Net* network for efficient processing. The training/testing data size is 3302/1672 frames. Additionally, we test on 1500 images from another outdoor benchmark dataset with unstructured scenes, RUGD [22]. RUGD is a video dataset captured from a camera onboard a mobile robot platform in unstructured outdoor environments, and has applications in off-road autonomous navigation.

4.2 Implementation Details

The image height and width $H = 600, W = 375$. The number of semantic classes $N_c = 4$. AdamW was used as the optimizer with the initial learning rate of $6\text{e}{-5}$. The coefficients used for computing running averages of gradient and its square (the betas) is equal to $(0.9, 0.999)$. Weight decay parameter is 0.01. We used linear learning rate decay and all models were trained for 160K iterations. All models were tested on a laptop with an Intel i9-13900H Processor, a RTX-4070 GPU, and 32GB RAM.

4.3 Evaluation Metrics

Three metrics were used to evaluate the pixel-level semantic segmentation performance, the mIoU (mean Intersection over Union), mean accuracy (mAcc, average accuracy across four semantic classes), and all accuracy (aAcc, average accuracy over all pixels). We also report MParams (millions of parameters), GFlops (billion floating-point operations), and FPS (inference frames per second) to evaluate model complexity and efficiency. Lower MParams and GFlops and Higher FPS corresponds to less complexity and higher efficiency. All the models were trained on the RELLIS-3D dataset and we report evaluation results on both RELLIS-3D and RUGD datasets. In the following tables, the suffix "-ar$n$" indicates the AR module having an AR-Ratio of $1/n$ in the model decoder. The "b0-b5" corresponds to the "MiT-B0" to "MiT-B5" models in the encoder backbone used in Segformer (MiT-B0 is the lightweight model for fast inference, while MiT-B5 is the largest model for the more accurate performance [6]).

Table 1. Comparison with State-of-the-arts methods: BG/O stands for background/Obstacle, S for stable, G for granular, HR for high resistance. The suffix -ar$n$ means we added our novel attention regulation layer in other models' decoders with AR-Ratio equals $1/n$.

Dataset	Model	BG/O IoU	BG/O Acc	S IoU	S Acc	G IoU	G Acc	HR IoU	HR Acc	mIoU ↑	mAcc ↑	aAcc ↑
RELLIS-3D	Segformer-b1[6]	89.82	91.36	77.32	84.28	33.21	60.33	90.61	97.66	72.74	83.41	94.18
	GANav[20]	89.92	92.60	76.15	78.19	26.90	51.65	90.46	96.99	70.86	79.86	94.09
	FastSCNN[11]	86.66	96.06	47.65	48.10	7.66	9.72	86.34	92.02	57.08	61.48	91.60
	FastSCNN-ar8	89.45	94.45	58.09	60.09	12.14	16.73	89.34	95.83	62.33	66.77	93.97
	PSPNet[12]	93.25	95.23	79.53	82.22	30.15	52.41	93.35	97.86	74.07	81.93	95.82
	PSPNet-ar8	93.47	96.38	82.21	85.83	**40.55**	55.24	93.37	97.14	77.40	83.65	96.08
	PSPNet-ar32	93.42	96.28	79.23	84.16	36.42	56.81	93.46	97.11	75.64	83.59	95.98
	Deeplabv3[15]	93.36	95.48	81.89	85.33	31.19	51.16	93.59	**97.79**	75.01	82.44	95.99
	Deeplabv3-ar8	93.24	96.24	81.45	87.27	34.96	63.13	93.55	96.72	75.80	85.84	95.92
	GCNet[16]	93.06	95.59	79.87	81.41	34.84	61.06	93.15	97.37	75.23	83.86	95.76
	GCNet-ar32	93.14	95.34	83.38	86.94	30.94	64.94	93.74	97.38	75.30	86.15	95.88
	ANNNet[17]	93.08	95.29	80.23	84.00	28.65	63.81	93.35	97.19	73.83	85.07	95.65
	ANNNet-ar8	93.59	96.42	81.31	86.10	36.80	75.34	93.22	96.35	76.23	88.55	95.86
	ours-b0-ar32	93.50	95.32	80.42	87.68	31.51	**72.86**	92.74	96.63	74.54	88.12	95.57
	ours-b1-ar32	93.25	95.25	84.11	89.91	35.57	72.48	93.03	96.95	76.49	88.65	95.77
	ours-b2-ar16	**94.05**	**96.53**	**84.39**	**91.46**	38.97	71.78	**93.82**	96.72	**77.81**	**89.12**	**96.26**
RUGD	Segformer-b1	29.72	36.34	7.73	8.64	23.51	**75.33**	14.29	15.31	18.81	33.9	38.63
	GANav	50.13	99.31	0.47	0.52	0.02	0.02	6.33	6.55	14.24	26.6	50.59
	FastSCNN	53.94	**99.73**	1.47	1.48	2.35	2.37	24.48	25.98	20.56	32.39	56.49
	FastSCNN-ar8	55.15	99.66	5.08	5.16	6.17	6.40	25.44	27.10	22.96	34.58	57.72
	PSPNet	64.57	89.61	6.63	6.74	6.52	6.72	41.51	67.46	29.81	42.63	63.62
	PSPNet-ar8	63.95	92.19	5.58	5.64	2.18	2.37	40.53	58.14	28.06	39.59	63.44
	PSPNet-ar32	63.86	86.63	6.48	6.54	4.45	4.50	38.42	67.92	28.30	41.40	61.78
	Deeplabv3	66.93	86.81	9.60	9.84	13.18	14.65	39.29	67.23	32.25	44.63	63.99
	Deeplabv3-ar8	65.18	93.65	10.95	11.11	4.81	5.06	41.55	62.42	30.62	43.06	64.01
	GCNet	65.35	89.82	5.22	5.37	19.61	21.85	42.51	63.45	33.17	45.12	65.93
	GCNet-ar32	**70.18**	96.06	8.97	9.24	23.46	25.11	43.38	61.34	36.50	47.94	69.25
	ANNNet	62.69	80.86	9.71	10.23	17.35	19.50	38.81	**68.60**	32.14	44.80	62.50
	ANNNet-ar8	62.89	96.20	10.52	11.03	6.90	7.00	37.69	52.53	29.50	41.69	63.05
	ours-b0-ar32	67.97	92.93	27.44	28.60	**38.32**	43.40	47.44	60.12	**45.30**	**56.26**	**71.85**
	ours-b1-ar32	64.63	97.75	21.56	22.52	25.10	27.33	44.32	51.70	38.90	49.83	68.32
	ours-b2-ar16	65.50	95.26	**43.47**	**45.47**	19.69	21.01	**47.69**	60.93	44.09	55.67	68.70

4.4 Quantitative Results

Table 1 shows the quantitative results of our proposed *AR-Net* model compared with other state-of-the-art semantic segmentation methods. We report mIoU and accuracy for each semantic group, as well as overall mIoU and accuracy. As shown, our model, especially the "ours-b2-ar16" variant (AR-ratio 1/16, using MiT-B2 backbone), achieved superior performance in both IoU and Accuracy compared with other methods. The lightweight model variants (ours-b0 and ours-b1), also achieved moderate to high accuracy. Additionally, the evaluation results on the RUGD dataset demonstrates the robustness and generalizability of our proposed *AR-Net*, with substantial improvements across all metrics over comparison methods. Additionally, if we compare a previous model with and without our proposed AR layer (e.g., PSPNet versus PSPNet-ar8), we observed that adding the proposed novel attention regulation layer to other networks has generally helped improve segmentation performance.

Table 2 shows the computational complexity analysis of three variations of our model compared with other segmentation methods. The "ours-b0" is the most lightweight variant of our model, with the least number of parameters and GFlops (lower complexity and smaller GPU memory cost) and fastest inference time (highest FPS). On the other hand, "ours-b2" variant requires the higher number of parameters but produces the most accurate segmentation results (see Table 1). With a 35-46FPS inference speed, our model is well-suited for real-time semantic tasks such as autonomous navigation or simultaneous localization and mapping (real-time SLAM).

Fig. 3 shows some visual examples of the segmentation outputs of our *AR-Net* model on the RELLIS-3D dataset. For clearer visualization, the pixel-level segmentation prediction results are overlaid with the raw RGB image, where red stands for background/obstacles semantic group (sky, people, etc.), blue for stable/smooth surface (concrete and asphalt) , dark green for granular group (sand, gravel, mud) and light green for high resistance surface (bushes, grass, log). As shown, our model produced accurate segmentation predictions, especially along the boundary areas, and can correctly interpret the traversability information in the unstructured scene.

Table 2. Model complexity analysis.

Models	MParams ↓	GFlops ↓	FPS ↑
segformer-b1	13.68	12.99	44.8
GANav	6.21	11.12	15.9
FastSCNN	1.21	0.82	45.2
FastSCNN-ar8	1.28	0.99	36.9
PSPNet	48.97	153.82	22.7
PSPNet-ar8	116.32	331.24	11.4
PSPNet-ar32	116.14	165.29	21.0
Deeplabv3	68.11	232.31	17.4
Deeplabv3-ar8	94.48	301.61	12.4
GCNet	49.62	130.78	21.1
GCNet-ar32	50.68	170.56	20.4
ANNNet	46.22	159.34	18.9
ANNNet-ar8	47.30	162.11	15.2
ours-b0-ar32	3.96	6.46	46.2
ours-b1-ar32	13.92	12.85	43.6
ours-b2-ar16	24.96	21.31	35.3

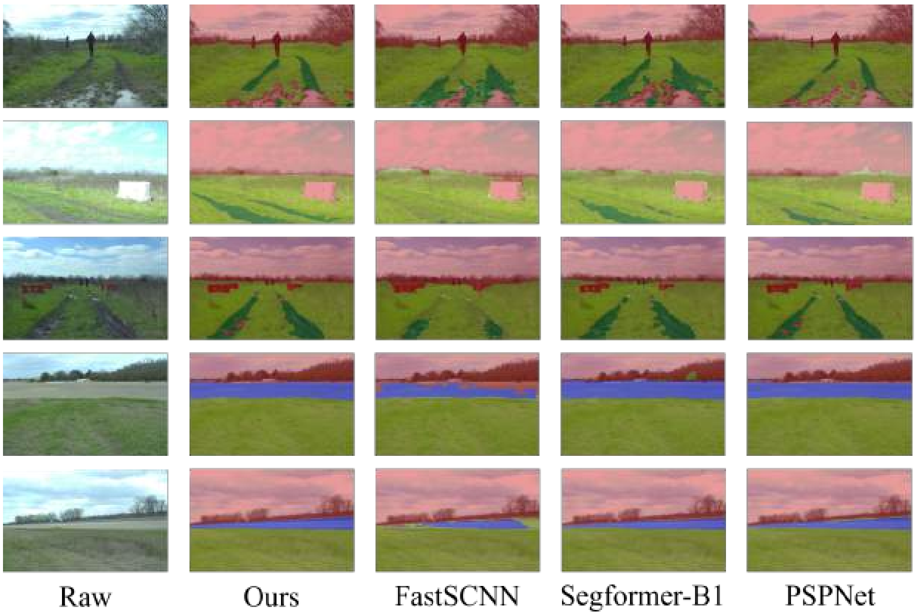

Fig. 3. The segmentation results overlaid with raw RGB image, where red for background/obstacles, blue for stable, dark green for granular and light green for high resistance.

Table 3. Ablation study on varying model parameters on the segmentation performance. Trained on RELLIS-3D dataset, and tested on both RELLIS-3D and RUGD datasets.

Chn	C_e	AR-Ratio	B	mIoU ↑	mIoU ↑	mAcc ↑	mAcc ↑	aAcc ↑	aAcc ↑
				RELLIS-3D	RUGD	RELLIS-3D	RUGD	RELLIS-3D	RUGD
384	150	1/32	b1	76.48	40.03	88.53	51.67	95.88	68.77
384	200	1/32	b1	75.78	39.45	90.56	51.91	95.89	65.86
384	250	1/32	b1	76.49	38.90	88.65	49.83	95.77	68.32
256	250	1/32	b1	76.97	42.10	88.90	53.19	95.97	69.33
128	250	1/32	b1	77.27	38.62	87.76	49.64	96.22	67.58
384	250	1/32	b0	74.54	45.30	88.12	56.26	95.57	71.85
384	250	1/32	b5	77.61	38.98	86.40	50.79	96.21	68.98
384	250	1/16	b5	78.48	38.34	85.01	50.38	96.44	68.27
384	250	1/8	b5	78.65	38.51	86.15	49.81	96.31	69.12

5. ABLATION STUDIES

We conduct additional ablation studies to evaluate the effect of model parameters on the segmentation performance. We vary the number of channels (Chn), embedding dimension (C_e), backbone (B), and attention regulation size ratio (AR-Ratio) and we report their segmentation accuracy and model complexity results.

Table 3 shows that selecting a more complex backbone (e.g., b5) and a larger AR-ratio (e.g., 1/16 or 1/8) can increase the segmentation accuracy, but it seems to decrease the model's generalization performance when evaluated on RUGD dataset. We did not observe a significant trend regarding the number of channels and embedding dimensions and we recommend choosing appropriate values based on testing.

Table 4 shows the model complexity analysis when changing encoder backbones and AR-Ratios. We observed that the inference speed dropped considerably when AR-Ratio changes from 1/16 to 1/8, as the larger masks require more cache and involves more read-write operations, thus reducing the inference efficiency.

Accounting for both segmentation accuracy, robustness in generalization, and model complexity, we select the following optimal values for each parameter in our experiments: 384 for channel size, 250 for embedding dimensions, and 1/16 for AR-ratio. Usually, a more complex decoder backbone tends to improve the overall performance of the model. Therefore, we choose MiT-B0, b1, and b2 as the backbones for our models of different scales. This is demonstrated in Table 1 where we indeed observed the "ours-b2-ar16" model (our *AR-Net*, with MiT-B2 backbone, AR-ratio of 1/16) outperforms the other model variants.

Table 4. Ablation study on model complexity when changing backbone and AR-Ratio.

AR-Ratio	B	MParams ↓	GFlops ↓	FPS ↑
1/32	b5	82.21	63.77	19.4
1/32	b1	13.92	12.85	43.6
1/32	b0	3.96	6.46	46.2
1/16	b1	13.92	12.98	42.6
1/8	b1	13.92	13.47	26.3

6. CONCLUSION

This paper proposes a new Attention-Regulation-based network (*AR-Net*) for accurate and efficient semantic segmentation models for outdoor unstructured terrains. The proposed attention-regulation layer can be easily integrated into various encoder-decoder architectures and allows for flexible adjustment of the spatial ratio. We report evaluation results on two outdoor unstructured benchmark datasets, RELLIS-3D and RUGD, and demonstrated superior performance in semantic segmentation accuracy and robustness when generalizing to another dataset. We also provided detailed ablation studies and analyses on model parameter selection.

Future work include refining the attention module to further improve its performance, addressing model inconsistency with inaccurate labels, and integrating the method into downstream tasks such as real-time SLAM.

ACKNOWLEDGEMENT

This material is based upon work supported by the National Science Foundation under Grant IIS-2153171-CRII: III: Explainable Multi-Source Data Integration with Uncertainty.

REFERENCES

[1] W. Zhang, S. Lyu, C. Yao, F. Xue, Z. Zhu, and Z. Jia, “Analysis of robot traversability over unstructured terrain using information fusion,” in *IEEE Int. Conf. Adv. Robotics and Mechatronics (ICARM)*, 2022, pp. 413–418.

[2] H.-W. Kim, J.-H. Huh, and M.-C. Lee, “A study on occlusion removal technology using integral imaging with semantic segmentation and predictive labeling,” in *23rd IEEE Int. Conf. Control, Automation and Systems (ICCAS)*, 2023, pp. 1694–1699.

[3] K. Sato, H. Madokoro, T. Nagayoshi, S. Chiyonobu, P. Martizzi, S. Nix, H. Woo, T. K. Saito, and K. Sato, “Semantic segmentation of outcrop images using deep learning networks toward realization of carbon capture and storage,” in *21st IEEE Int. Conf. Control, Automation and Systems (ICCAS)*, 2021, pp. 436–441.

[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” *Adv. Neural info. Proc. Syst. (NIPS)*, vol. 30, 2017.

[5] K.-H. Choi and J.-E. Ha, “Random swin transformer,” in *22nd IEEE Int. Conf. Control, Automation and Systems (ICCAS)*, 2022, pp. 1611–1614.

[6] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, “Segformer: Simple and efficient design for semantic segmentation with transformers,” *Adv. Neural info. Proc. Syst. (NIPS)*, vol. 34, pp. 12 077–12 090, 2021.

[7] M. Park and D.-o. Kang, “Urban scene editing with diffusion model using segmentation mask,” in *23rd IEEE Int. Conf. Control, Automation and Systems (ICCAS)*, 2023, pp. 1881–1884.

[8] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in *CVPR*, 2015, pp. 3431–3440.

[9] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *Medical image computing and computer-assisted intervention (MICCAI)*, 2015, pp. 234–241.

[10] V. Badrinarayanan, A. Kendall, and R. Cipolla, “Segnet: A deep convolutional encoder-decoder architecture for image segmentation,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[11] R. P. Poudel, S. Liwicki, and R. Cipolla, “Fast-scnn: Fast semantic segmentation network,” *arXiv preprint arXiv:1902.04502*, 2019.

[12] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, “Pyramid scene parsing network,” in *CVPR*, 2017, pp. 2881–2890.

[13] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, “Semantic image segmentation with deep convolutional nets and fully connected crfs,” *arXiv preprint arXiv:1412.7062*, 2014.

[14] ——, “Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2017.

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

[16] Y. Cao, J. Xu, S. Lin, F. Wei, and H. Hu, “Gcnet: Non-local networks meet squeeze-excitation networks and beyond,” in *CVPRW*, 2019.

[17] Z. Zhu, M. Xu, S. Bai, T. Huang, and X. Bai, “Asymmetric non-local neural networks for semantic segmentation,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 593–602.

[18] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P. H. Torr *et al.*, “Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers,” in *CVPR*, 2021, pp. 6881–6890.

[19] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, “An image is worth 16x16 words: Transformers for image recognition at scale,” *Int. Conf. Learning Representations (ICLR)*, 2021.

[20] T. Guan, D. Kothandaraman, R. Chandra, A. J. Sathyamoorthy, K. Weerakoon, and D. Manocha, “Ga-nav: Efficient terrain segmentation for robot navigation in unstructured outdoor environments,” *IEEE Robotics and Automation Letters*, vol. 7, no. 3, pp. 8138–8145, 2022.

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

[22] 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.