RangeSAM: On the Potential of Visual Foundation Models for Range-View represented LiDAR segmentation

Anonymous WACV Applications Track submission

Paper ID

Abstract

LiDAR point cloud segmentation is central to autonomous driving and 3D scene understanding. While voxel- and point-based methods dominate recent research due to their compatibility with deep architectures and ability to capture fine-n geometry, they often incur high computational cost, irregular memory access, and limited runtime efficiency due to scaling issues. In contrast, range-view methods, though relatively underexplored - can leverage mature 2D semantic segmentation techniques for fast and accurate predictions. Motivated by the rapid progress in Visual Foundation Models (VFMs) for captioning, zero-shot recognition, and multimodal tasks, we investigate whether SAM2, the current state-of-the-art VFM for segmentation tasks, can serve as a strong backbone for LiDAR point cloud segmentation in the range view representations. We present ***RangeSAM****, to our knowledge, the first range-view framework that adapts SAM2 to 3D segmentation, coupling efficient 2D feature extraction with projection/back-projection to operate on point clouds. To optimize SAM2 for range-view representations, we implement several architectural modifications to the encoder: (1) a novel* ***Stem*** *module that emphasizes horizontal spatial dependencies inherent in LiDAR range images, (2) a customized configuration of* ***Hiera Blocks*** *tailored to the geometric properties of spherical projections, and (3) an adapted* ***Window Attention*** *mechanism in the encoder backbone specifically designed to capture the unique spatial patterns and discontinuities present in range-view pseudo-images. Our approach achieves competitive performance on SemanticKITTI while benefiting from the speed, scalability, and deployment simplicity of 2D-centric pipelines. This work highlights the viability of VFMs as general-purpose backbones for point cloud segmentation and opens a path toward unified, foundation-model-driven LiDAR segmentation. Results let us conclude that range-view segmentation methods using VFMs lead to promising results.*

1. Introduction

Reconstructing urban scenes with 2D and 3D semantics is crucial for autonomous systems (vehicles, drones, robots), requiring real-time viewpoint synthesis, semantic segmentation, depth generation, and dynamic object tracking. Among these, LiDAR point cloud semantic segmentation is fundamental for differentiating vehicles, pedestrians, road signs, and infrastructure. Recent research focuses on voxel- and point-based methods [26, 43, 63, 76, 77, 95], which achieve strong performance but impose substantial computational and memory costs on large-scale outdoor data [18, 46, 47, 49, 64, 67, 70, 75, 85] and struggle with irregular, unordered point clouds. In contrast, range-view segmentation [2, 13, 33, 41, 48] projects 3D point clouds into dense 2D representations, enabling reuse of mature 2D models with reduced memory and faster inference. Though previously overlooked due to limitations in handling occlusions and resolution loss, recent advances in attention mechanisms [66], multi-scale fusion [93], and context-aware architectures [40] warrant reassessing the range-view paradigm. We propose *RangeSAM*, which adapts the Segment Anything Model 2 (SAM2) [50] for range-view LiDAR segmentation (Figure 1). Our approach leverages SAM2's robust *zero-shot capabilities* for 2D understanding and extends them to 3D via range-view representation. The methodology includes: (1) range projection preprocessing to transform unordered LiDAR scans, (2) a multi-component architecture incorporating Receptive Field Blocks [37], (3) postprocessing with k-NN label propagation, and (4) a composite loss function addressing class imbalance and boundary accuracy. Our main contributions are:

- We introduce RangeSAM, the first framework adapting SAM2 for LiDAR point cloud segmentation via range-view representations.
- We have designed a multi-component encoder architecture with a pretrained Hiera backbone, custom Stem module, novel embedding matrix, and Hiera blocks utilizing

localized and global Multi-Head Attention (Figure 1).

- We demonstrate competitive performance on SemanticKITTI [4], validating our approach's viability, and conduct ablation studies on training strategies and data augmentation.

2. Related Work

Visual Foundation Models

Recent large-scale vision models have revolutionized image tasks [6, 38, 79, 88]. The Segment Anything Model (SAM) series achieved promptable segmentation with zero-shot transferability [30, 36, 53, 56], while SAM2 extends to videos and complex scenarios [12, 28, 39, 50, 80, 87], and has been adapted to 3D [7, 83, 86]. Other VFMs include self-supervised transformers (DINO series [9, 42, 57]), multimodal models [45, 97], and the EVA series [20, 21, 60, 61], offering strong generalization and label-efficient learning. For instance, DINOv2 features improve 3D semantic segmentation [83, 89], while CLIP's image-text embeddings enable zero-shot 3D understanding [11, 45, 62, 90].

2.1. Methods on Projection-Based representations

Convolutional Neural Networks (CNNs): Projection-based methods transform point clouds into structured 2D representations (range-, bird's-eye-, or perspective-view) to leverage mature 2D CNN architectures [4, 5, 10, 16, 27, 31, 34, 41, 51, 64, 65, 81, 91, 96]. The SqueezeSeg series [68, 72, 73, 81] pioneered real-time segmentation with lightweight CNNs and CRF post-processing [17]. Methods like DarkNetSeg [4], SalsaNext [16], KPRNet [31], RangeNet++ [41], Lite-HDSeg [51], and SqueezeSegV3 [81] employ a two-stage approach: range projection followed by U-Net-style architectures [54]. RangeSeg works [13, 69] incorporated multi-scale context modeling for spherical projections.

Vision Transformers (ViTs): ViT architectures have driven significant advances: RangeViT and FRNet [2, 84] introduced hybrid CNN-Transformer backbones combining convolutional inductive biases with global attention, benefiting from 2D transfer learning [15], while RangeFormer [33] established a fully transformer-based architecture achieving state-of-the-art performance on outdoor benchmarks [4, 5, 8, 44, 59].

3. Methodology

We introduce **RangeSAM**, a projection-based LiDAR segmentation model leveraging VFMs. We detail the range projection preprocessing, architectural modifications, and postprocessing procedures. Figure 1 illustrates the pipeline, including the stem module, SAM2-based encoder, and decoder components.

3.1. Range Projection

Each LiDAR scan is initially represented as an unordered point set $\mathcal{P} = (x, y, z, f)$, where (x, y, z) denotes the Cartesian coordinates and f represents auxiliary sensor measurements such as intensity or remission values. Subsequently, the point cloud undergoes transformation to a range view representation via projection onto the sensors spherical coordinate system [41]:

$$\theta = \arctan\left(\frac{y}{x}\right), \phi = \arcsin\left(\frac{z}{r}\right), r = \sqrt{x^2 + y^2 + z^2} \tag{1}$$

Specifically, the 3D points are discretized through rasterization into a 2D cylindrical projection with dimensions $H \times W$:

$$\begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} \frac{1}{2}\left[1 - \arctan(y, x)\pi^{-1}\right] w \\ \left[1 - \left(\arcsin\left(zr^{-1}\right) + \mathfrak{f}_{\text{up}}\right)\mathfrak{f}^{-1}\right] h \end{pmatrix}, \tag{2}$$

where (u, v) represent the projected coordinates of point p_n in the range image coordinate system, and h and w denote the image height and width parameters, respectively. $r = \sqrt{x^2 + y^2 + z^2}$ denotes the range between a point and its sensor, the sensors vertical field of view $\mathfrak{f}$ comprises the sum of $(\mathfrak{f}_{\text{up}} + \mathfrak{f}_{\text{down}})$ viewing angles. For our method, we configured the imaging system to operate at the commonly used spatial resolution of 64×2048 pixels [2, 13, 33]. Multiple points projecting to the same pixel were resolved by retaining the minimum-range feature. Unprojected pixels were zero-filled.

3.2. Model Architecture

SAM2 is a state-of-the-art VFM trained on the SA-V dataset [50]. SAM2-UNet [80] achieves superior performance across multiple segmentation tasks while maintaining simplicity. RangeSAM adopts a similar paradigm (Figure 1), incorporating Receptive Field Blocks (RFB) [37] for enhanced feature decoding. **Stem:** Transforms input tensors from (B,6,H,W) to (B,96,H,W) via linear transformation, layer normalization [3], and GELU activation [25] (Figure 1 top), then features were partitioned into overlapping 7×7 patches with unit stride. We replace SAM2's positional embedding with a novel (4,128) embedding matrix to enhance horizontal spatial sensitivity. **Encoder:** Uses the pretrained Hiera [50, 55] backbone with a multi-component architecture (Figure 1 right) comprising: **Hiera block:** Each stage of the backbone consists of a customized number of Hiera blocks, and each block includes two modules: *1) Multi-Head Attention module*, in early stages, attention is restricted within a window, while later stages may also use global attention. The operation can be written as

$$X_{out} = X + \text{DropPath}(\text{MHA}(\text{LayerNorm}(X))) \tag{3}$$

$$\text{MHA}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O \tag{4}$$

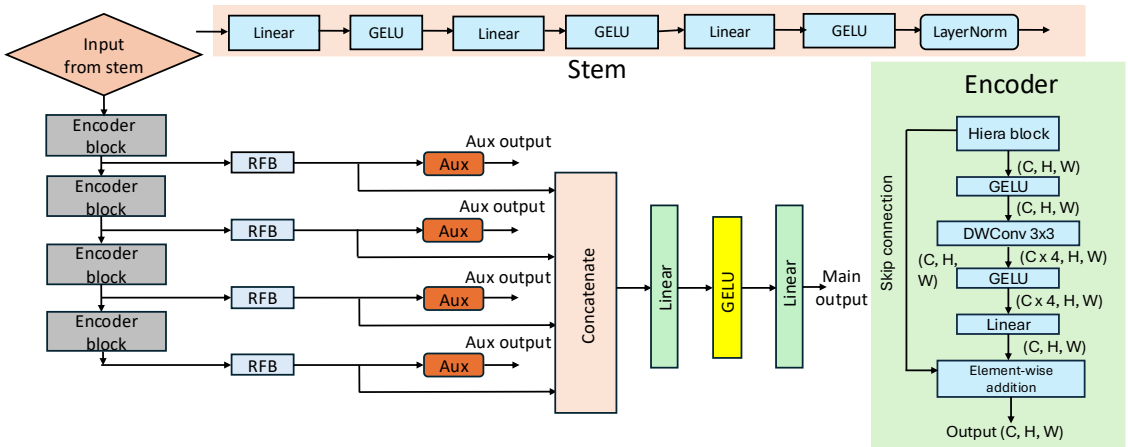

Figure 1. Overview of our SAM2-based point-cloud segmentation model. The stem reshapes range-view point clouds into a tensor suitable for the encoder. The encoder is built from stacked Hiera blocks—each containing a multi-head self-attention module and a feed-forward network [50]. The decoder comprises of Receptive Field Blocks [37] (RFB) with LayerNorm [3] and GELU [25], concatenates multi-scale features, and projects them to N_{classes} while also adding auxiliary head (Aux) on corresponding output.

where

$$\begin{aligned} \text{head}_i &= \text{Attention}\left(QW_Q^i, KW_K^i, VW_V^i\right), \\ \text{Attention}(Q,K,V) &= \sigma\left(\frac{QK^\top + M}{\sqrt{d_{\text{head}}}}\right)V \end{aligned} \tag{5}$$

where DropPath represents the stochastic depth regularization technique [35], and LayerNorm denotes the layer normalization method [3], which provides batch size-independent normalization capabilities. *2) Feed-forward network:*

$$O = \text{DropPath}(\text{MLP}(\text{LayerNorm}(O))) \oplus X_{\text{out}}\,, \tag{6}$$

$$F = \text{Lin}\left(\text{GELU}\left(\text{DWConv 2D}_{3\times3}(\text{GELU(O)})\right)\right) + \text{O} \tag{7}$$

where $\text{DWConv2D}_{3\times3}$ denotes the *3×3* depthwise convolution operation that introduces spatial locality inductive bias while maintaining computational efficiency through minimal parameter overhead. The SAM2-tiny backbone architecture comprises four Hierarchical stages containing $[1, 2, 7, 2]$ Hiera blocks respectively. Global attention mechanisms are implemented at blocks $[5, 7, 9, 11]$ to capture long-range spatial dependencies.

Attention Window: In the early stages, Hiera employs local windowed attention with a masking strategy where $M_{ij} = 0$ for tokens i and j within the same mask unit, and $M_{ij} = -\infty$ otherwise. Later stages utilize global attention with $M = 0$. Given the range-view image resolution of 64×2048, we propose an asymmetric attention window design that emphasizes horizontal spatial relationships: 8×64 for the first and fourth stages, and 16×128 for the second and third stages. This horizontally-elongated window configuration, while empirically motivated, demonstrates superior performance compared to conventional square attention windows, effectively capturing the inherent horizontal structure of range-projected LiDAR data.

Decoder: While the outputs of the encoder are multiscale features F_1, F_2, F_3, F_4 with spatial size,

$$\left[(H,W), \left(\frac{H}{2}, \frac{W}{2}\right), \left(\frac{H}{4}, \frac{W}{4}\right), \left(\frac{H}{8}, \frac{W}{8}\right)\right] \tag{8}$$

and channel dimensions $[96, 192, 384, 768]$. Following [80], we standardized the output channel dimensions to 256 using a modified Receptive Field Block [37] architecture that substitutes LayerNorm and GELU activations for the conventional BatchNorm [29] and ReLU, thereby achieving better compatibility with contemporary transformer architectures. Consistent with [22], we concatenate these four normalized feature maps and progressively reduce the dimensionality to match the number of target classes while incorporating auxiliary classification heads at corresponding feature levels to enhance gradient flow during training.

Postprocessing: Dense evaluation on datasets like SemanticKITTI requires label propagation from processed points to the full-resolution point cloud via *k-NN* interpolation with majority voting. We use $k = 7$, balancing efficiency and noise robustness within the typical range of 3-7 [2, 13, 33, 69].

Loss: We follow the training objective of the proposed SAM2-UNet [80] model which incorporates a multi-component loss function comprising weighted cross-entropy ($\mathcal{L}_{WCE}$), Dice loss ($\mathcal{L}_{\text{Dice}}$), boundary loss ($\mathcal{L}_{\text{Boundary}}$) and Jaccard index loss ($\mathcal{L}_{\text{IoU}}$). This composite loss formulation addresses class imbalance through weighted cross-entropy [98], enhances region-level segmentation accuracy via Dice and Jaccard losses [58, 71],

Table 1. Comparisons among different Hiera backbones on the test set of SemanticKITTI [4]. Interestingly, the SAM2-tiny model outperforms its larger counterparts despite having fewer parameters, demonstrating that model capacity does not necessarily correlate with performance in this task domain.

Backbone	Params(M)	mIoU	car	bicy	moto	truc	o.veh	ped	b.list	m.list	road	park	walk	o.gro	build	fenc	veg	trun	terr	pole	sign
Hiera-tiny	63.3	**61.5**	95.0	44.3	47.3	60.0	55.2	68.2	79.6	2.6	94.8	49.6	82.1	1.2	88.6	62.0	85.1	68.2	71.3	64.8	48.2
Hiera-small	70.2	60.5	94.7	47.2	47.7	72.2	39.7	63.1	82.5	0.0	95.3	32.1	80.8	0.1	88.3	60.0	88.4	68.6	78.4	64.9	45.5

Table 2. Experiments with and without training augmentations. We observe a strong performance gain when introducing the proposed training augmentations from [33]. Comparisons among SAM2-tiny on SemanticKITTI test set [4].

Augs.	mIoU	car	bicy	moto	truc	o.veh	ped	b.list	m.list	road	park	walk	o.gro	build	fenc	veg	trun	terr	pole	sign
off	55.0	93.1	36.2	37.2	53.9	36.7	53.7	59.7	0.0	93.4	35.3	79.3	0.0	85.5	55.8	86.4	59.3	76.7	57.1	45.3
on	**61.5**	95.0	44.3	47.3	60.0	55.2	68.2	79.6	2.6	94.8	49.6	82.1	1.2	88.6	62.0	85.1	68.2	71.3	64.8	48.2

Table 3. Experiments with and without transfer-learning techniques using Cityscapes dataset [15]. We observe that overall our model does not benefit from extensive transfer-learning strategies introduced by [2]. Comparisons among SAM2-tiny on SemanticKITTI [4] test set.

Pretrain	mIoU	car	bicy	moto	truc	o.veh	ped	b.list	m.list	road	park	walk	o.gro	build	fenc	veg	trun	terr	pole	sign
yes	59.7	94.2	43.2	38.0	82.2	45.3	56.4	65.5	0.0	94.8	43.7	81.9	0.1	87.9	60.7	86.0	68.10	73.7	64.3	46.0
no	**61.5**	95.0	44.3	47.3	60.0	55.2	68.2	79.6	2.6	94.8	49.6	82.1	1.2	88.6	62.0	85.1	68.2	71.3	64.8	48.2

Table 4. Comparisons among state-of-the-art LiDAR range view semantic segmentation approaches on the validation sequence of SemanticKITTI [4]. We compare our best performing model using SAM2 with the tiny Hiera backbone. Notably, our model is the first and only approach that leverages VFMs for this task.

Method	VFM	mIoU	car	bicy	moto	truc	o.veh	ped	b.list	m.list	road	park	walk	o.gro	build	fenc	veg	trun	terr	pole	sign
KPRNet[32] (2020)	×	63.1	95.5	54.1	47.9	23.6	42.6	65.9	65.0	16.5	93.2	73.9	80.6	30.2	91.7	68.4	85.7	69.8	71.2	58.7	64.1
LiteHDSeg[52] (ICRA21)	×	63.8	92.3	40.0	55.4	37.7	39.6	59.2	71.6	54.3	93.0	68.2	78.3	29.3	91.5	65.0	78.2	65.8	65.1	59.5	67.7
MPF[1] (WACV21)	×	55.5	93.4	30.2	38.3	26.1	28.5	48.1	46.1	18.1	90.6	62.3	74.5	30.6	88.5	59.7	83.5	59.7	69.2	49.7	58.1
FIDNet[94] (IROS21)	×	59.5	93.9	54.7	48.9	27.6	23.9	62.3	59.8	23.7	90.6	59.1	75.8	26.7	88.9	60.5	84.5	64.4	69.0	53.3	62.8
RangeViT[82] (ICCV21)	×	64.0	95.4	55.8	43.5	29.8	42.1	63.9	58.2	38.1	93.1	70.2	80.0	32.5	92.0	69.0	85.3	70.6	71.2	60.8	64.7
CENet[14] (ICME22)	×	64.7	91.9	58.6	50.3	40.6	42.3	68.9	65.9	43.5	90.3	60.9	75.1	31.5	91.0	66.2	84.5	69.7	70.0	61.5	67.6
MaskRange[24] (2022)	×	66.1	94.2	56.0	55.7	59.2	52.4	67.6	64.8	31.8	91.7	70.7	77.1	29.5	90.6	65.2	84.6	68.5	69.2	60.2	66.6
RangeFormer[92] (NeurIPS22)	×	73.3	96.7	69.4	73.7	59.9	66.2	78.1	75.9	58.1	92.4	73.0	78.8	42.4	92.3	70.1	86.6	73.3	72.8	66.4	66.6
RangeSAM(Ours)	✓	**60.9**	91.7	42.0	44.3	41.2	36.1	55.9	54.4	30.2	92.2	67.5	77.2	28.6	89.8	62.8	84.5	66.9	69.8	58.9	62.4

and promotes precise boundary delineation through the boundary loss [74] component. The combined loss function is expressed as a weighted sum:

$$\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_{WCE} + \lambda_2 \mathcal{L}_{\text{Dice}} + \lambda_3 \mathcal{L}_{\text{Boundary}} + \lambda_4 \mathcal{L}_{\text{IoU}} , \quad (9)$$

where λ_i represent empirically determined weighting coefficients. In our experimental configuration, we set $\lambda_i = 1$, assigning equal importance to each loss component.

4. Dataset and implementation details

We now describe the datasets used, the training configurations applied, and the evaluation metrics employed.

4.1. Dataset selection and metrics:

We used SemanticKITTI [4] and nuScenes [8] as our primary training datasets. SemanticKITTI consisted of urban driving scenes captured with a 64-beam LiDAR, split into 19,130 training scans (sequences 00-07, 09-10), 4,071 validation scans (sequence 08), and 20,351 test scans (sequences 11-21) across 19 semantic classes. NuScenes contained 1,000 urban scenes from Boston and Singapore captured with a 32-beam LiDAR, providing 28,130 training and 6,019 validation frames across 16 categories. We converted nuScenes to SemanticKITTI format for consistency and reported mean Intersection over Union (mIoU) [19] on the SemanticKITTI validation split.

4.2. Data Augmentations:

We applied standard augmentations including global rotation, coordinate jittering, flipping, and random point elimination (all with probability 1.0), along with range-view-specific augmentations from [33], mixing, union, shifting, and copy-paste with probabilities [0.9, 0.1, 0.9, 1.0], respectively.

4.3. Training strategy:

All experiments used 8 A100 GPUs (40GB) with a per-GPU batch size of 2 (effective batch size of 16). Training em-

ployed a 5-epoch linear warm-up followed by cosine annealing [23]. Following recent works [2, 33, 77, 78] show that transformer-based architectures benefit from pretraining, we adopted the training protocol from [2, 33] and pretrain on nuScenes [8] (converted to SemanticKITTI format) for 60 epochs. We selected SAM2-tiny as our backbone, as heavier SAM2 variants yielded negligible performance gains (Table 1) while significantly increasing computational cost and inference latency. The final model contains approximately 63 million parameters. **Optimization:** We use AdamW with differentiated learning rates: the Hiera backbone uses $lr = 0.0004$ and weight decay $\alpha = 0.001$, while the rest of the model uses $lr = 0.001$ and $\alpha = 0.0001$.

4.4. Evaluation

We present RangeSAM's experimental results on SemanticKITTI [4]. Table 4 shows validation results on sequence 08, while Tables 1, 2, and 3 report test set performance. Our approach achieves competitive results compared to state-of-the-art methods.

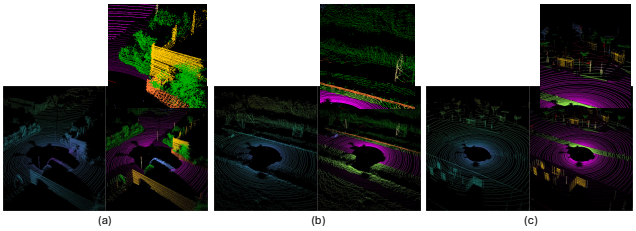

Figure 2. Qualitative segmentation examples of increasing difficulty using our model. (a) Urban intersection, (b) Suburban environment and (c) Highly cluttered scene.

4.5. Quantitative evaluation

Table 1 compares SAM2 Hiera backbones: tiny and small variants show comparable mIoU, with tiny offering computational advantages through reduced parameters and faster inference.

Table 4 benchmarks our model against baselines. While recent state-of-the-art achieves 65–75% mIoU, our model reaches 60.9%, successfully integrating the SAM2 VFM [50]. Performance stratifies across three tiers:

- High IoU (80–90%): Large, frequent classes (cars, road, building, vegetation) perform competitively with state-of-the-art.
- Mid IoU (60–70%): Mid-frequency categories (trucks, fences, terrain) remain competitive.
- Low IoU (29–47%): Rare, small objects (motorcycles, bicycles, persons) prove challenging, consistent with modern approaches.

While lower performance on long-tail categories reflects limited training examples relative to our 63M parameters, results remain competitive on well-represented categories (highlighted in Table 4). Dataset expansion and design refinement should close the gap on tail classes.

4.6. Qualitative evaluation

Figure 2 illustrates three segmentation scenarios: **(a) Urban intersection:** The model accurately segments vehicles, static elements (trees, walls, fences), and fine structures (fence posts). **(b) Suburban environment:** Dominant classes (cars, vegetation, fences) are correctly segmented, but the model struggles with unseen classes like transmission towers. **(c) Highly cluttered scene:** Despite overlapping objects (trees, fences, poles, signs), major classes are captured. Small or distant structures exhibit noise, but overall segmentation remains robust.

4.7. Ablation Study

We conduct a comprehensive ablation study to systematically evaluate the impact of different training strategies on model performance.

Transfer learning: We followed a transfer learning approach established by other baseline methods[2] in this area of research. For the experiments in Table 3, we pretrained our model on the Cityscapes [15] dataset for 25 epochs to enhance feature learning before training on the target dataset. However, this approach led to reduced performance. We believe this may be due to the SAM2 model being extensively pretrained on large-scale image datasets, which could have caused a domain mismatch affecting transfer learning. Further investigation is needed to confirm this hypothesis and explore other adaptation strategies.

Range View Augmentations: The results (see Table 2) indicate that adding augmentations introduced by [33] to the training process improves performance across nearly all datasets, achieving a 10% gain in mIoU.

5. Conclusion

This work, to our knowledge, is the first to adapt VFMs for range-view LiDAR point cloud segmentation. Despite the domain gap between RGB images and range-view representations, our targeted architectural modifications to the SAM2 encoder and decoder effectively bridged this modality shift and achieved competitive performance. We have integrated augmentation strategies and training protocols from prior work [2, 13, 33, 77]. Notably, we find that multi-dataset training across diverse 3D benchmarks outperforms extensive 2D pretraining as in [2]. We will release source code and model weights.

Future Work: The primary limitation is computational complexity. The convolution-free Hiera backbone [55] struggles with point cloud sparsity, requiring RFBs [37] for competitive performance. However, RFBs constitute the main computational bottleneck, currently preventing real-time deployment—a key target for future optimization.

References

[1] Yara Ali Alnaggar, Mohamed Afifi, Karim Amer, and Mohamed Elhelw. Multi-Projection Fusion for Real-Time Semantic Segmentation of 3D LiDAR Point Clouds. In *Winter Conference on Applications of Computer Vision (WACV)*, 2021. 4

[2] Angelika Ando, Spyros Gidaris, Andrei Bursuc, Gilles Puy, Alexandre Boulch, and Renaud Marlet. Rangevit: Towards vision transformers for 3d semantic segmentation in autonomous driving. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5240–5250, 2023. 1, 2, 3, 4, 5

[3] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization, 2016. 2, 3

[4] J. Behley, M. Garbade, A. Milioto, J. Quenzel, S. Behnke, C. Stachniss, and J. Gall. SemanticKITTI: A Dataset for Semantic Scene Understanding of LiDAR Sequences. In *Proc. of the IEEE/CVF International Conf. on Computer Vision (ICCV)*, 2019. 2, 4, 5

[5] J. Behley, M. Garbade, A. Milioto, J. Quenzel, S. Behnke, J. Gall, and C. Stachniss. Towards 3D LiDAR-based semantic scene understanding of 3D point cloud sequences: The SemanticKITTI Dataset. *The International Journal on Robotics Research*, 40(8-9):959–967, 2021. 2

[6] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ B. Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri S. Chatterji, Annie S. Chen, Kathleen Creel, Jared Quincy Davis, Dorottya Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren E. Gillespie, Karan Goel, Noah D. Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark S. Krass, Ranjay Krishna, Rohith Kuditipudi, and et al. On the opportunities and risks of foundation models. *CoRR*, abs/2108.07258, 2021. 2

[7] Nhat-Tan Bui, Dinh-Hieu Hoang, Minh-Triet Tran, Gianfranco Doretto, Donald Adjeroh, Brijesh Patel, Arabinda Choudhary, and Ngan Le. Sam3d: Segment anything model in volumetric medical images, 2024. 2

[8] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. *CoRR*, abs/1903.11027, 2019. 2, 4, 5

[9] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. *CoRR*, abs/2104.14294, 2021. 2

[10] Jiale Chen, Fei Xia, Jianliang Mao, Haoping Wang, and Chuanlin Zhang. Siesef-fusionnet: Spatial inter-correlation enhancement and spatially-embedded feature fusion network for lidar point cloud semantic segmentation, 2024. 2

[11] Runnan Chen, Youquan Liu, Lingdong Kong, Xinge Zhu, Yuexin Ma, Yikang Li, Yuenan Hou, Yu Qiao, and Wenping Wang. Clip2scene: Towards label-efficient 3d scene understanding by clip, 2023. 2

[12] Tianrun Chen, Ankang Lu, Lanyun Zhu, Chaotao Ding, Chunan Yu, Deyi Ji, Zejian Li, Lingyun Sun, Papa Mao, and Ying Zang. Sam2-adapter: Evaluating & adapting segment anything 2 in downstream tasks: Camouflage, shadow, medical image segmentation, and more, 2024. 2

[13] Tzu-Hsuan Chen and Tian Sheuan Chang. Rangeseg: Range-aware real time segmentation of 3d lidar point clouds. *IEEE Transactions on Intelligent Vehicles*, 7(1):93–101, 2022. 1, 2, 3, 5

[14] Hui-Xian Cheng, Xian-Feng Han, and Guo-Qiang Xiao. Cenet: Toward concise and efficient lidar semantic segmentation for autonomous driving. In *2022 IEEE International Conference on Multimedia and Expo (ICME)*, pages 01–06. IEEE, 2022. 4

[15] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2, 4, 5

[16] Tiago Cortinhal, George Tzelepis, and Eren Erdal Aksoy. Salsanext: Fast, uncertainty-aware semantic segmentation of lidar point clouds for autonomous driving, 2020. 2

[17] Aashish Dhawan, Pankaj Bodani, and Vishal Garg. Post processing of image segmentation using conditional random fields. In *2019 6th International Conference on Computing for Sustainable Global Development (INDIACom)*, pages 729–734, 2019. 2

[18] Francis Engelmann, Martin Bokeloh, Alireza Fathi, Bastian Leibe, and Matthias Nießner. 3d-mpa: Multi proposal aggregation for 3d semantic instance segmentation. *CoRR*, abs/2003.13867, 2020. 1

[19] Mark Everingham, S. M. Eslami, Luc Gool, Christopher K. Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *Int. J. Comput. Vision*, 111(1):98–136, 2015. 4

[20] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale. *arXiv preprint arXiv:2211.07636*, 2022. 2

[21] Yuxin Fang, Quan Sun, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva-02: A visual representation for neon genesis. *Image and Vision Computing*, 149:105171, 2024. 2

[22] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3354–3361, 2012. 3

[23] Akhilesh Gotmare, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. A closer look at deep learning heuris-

tics: Learning rate restarts, warmup and distillation. *CoRR*, abs/1810.13243, 2018. 5

[24] Yi Gu, Yuming Huang, Chengzhong Xu, and Hui Kong. Maskrange: A mask-classification model for range-view based lidar segmentation, 2022. Under review. 4

[25] Dan Hendrycks and Kevin Gimpel. Bridging nonlinearities and stochastic regularizers with gaussian error linear units. *CoRR*, abs/1606.08415, 2016. 2, 3

[26] Yuenan Hou, Xinge Zhu, Yuexin Ma, Chen Change Loy, and Yikang Li. Point-to-voxel knowledge distillation for lidar semantic segmentation, 2022. 1

[27] Qingyong Hu, Bo Yang, Linhai Xie, Stefano Rosa, Yulan Guo, Zhihua Wang, Niki Trigoni, and Andrew Markham. RandLA-Net: Efficient Semantic Segmentation of Large-Scale Point Clouds . In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11105–11114, Los Alamitos, CA, USA, 2020. IEEE Computer Society. 2

[28] Guohao Huo, Ruiting Dai, and Hao Tang. Samba-unet: Synergizing sam2 and mamba in unet with heterogeneous aggregation for cardiac mri segmentation, 2025. 2

[29] Sergey Ioffe and Christian Szegedy. Batch normalization: accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, page 448–456. JMLR.org, 2015. 3

[30] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything, 2023. 2

[31] Deyvid Kochanov, F. Karimi Nejadasl, and Olaf Booij. Kprnet: Improving projection-based lidar semantic segmentation. *ArXiv*, abs/2007.12668, 2020. 2

[32] Deyvid Kochanov, Fatemeh Karimi Nejadasl, and Olaf Booij. KPRNet: Improving projection-based LiDAR semantic segmentation. *arXiv preprint arXiv:2007.12668*, 2020. 4

[33] Lingdong Kong, Youquan Liu, Runnan Chen, Yuexin Ma, Xinge Zhu, Yikang Li, Yuenan Hou, Yu Qiao, and Ziwei Liu. Rethinking range view representation for lidar segmentation, 2023. 1, 2, 3, 4, 5

[34] Loïc Landrieu and Martin Simonovsky. Large-scale point cloud semantic segmentation with superpoint graphs. *CoRR*, abs/1711.09869, 2017. 2

[35] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Fractalnet: Ultra-deep neural networks without residuals. *CoRR*, abs/1605.07648, 2016. 3

[36] Feng Li, Hao Zhang, Peize Sun, Xueyan Zou, Shilong Liu, Jianwei Yang, Chunyuan Li, Lei Zhang, and Jianfeng Gao. Semantic-sam: Segment and recognize anything at any granularity. *arXiv preprint arXiv:2307.04767*, 2023. 2

[37] Songtao Liu, Di Huang, and Yunhong Wang. Receptive field block net for accurate and fast object detection. *CoRR*, abs/1711.07767, 2017. 1, 2, 3, 5

[38] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 2

[39] Yuyuan Liu, Yuanhong Chen, Chong Wang, Junlin Han, Junde Wu, Can Peng, Jingkun Chen, Yu Tian, and Gustavo Carneiro. Auralsam2: Enabling sam2 hear through pyramid audio-visual feature prompting, 2025. 2

[40] Liane-Marina Messmer, Christoph Reich, and Djaffar Ould Abdeslam. Context-aware machine learning: A survey. In *Proceedings of the Future Technologies Conference (FTC) 2024, Volume 1*, pages 252–272, Cham, 2024. Springer Nature Switzerland. 1

[41] Andres Milioto, Ignacio Vizzo, Jens Behley, and Cyrill Stachniss. Rangenet ++: Fast and accurate lidar semantic segmentation. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4213–4220, 2019. 1, 2

[42] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Q. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russ Howes, Po-Yao (Bernie) Huang, Shang-Wen Li, Ishan Misra, Michael G. Rabbat, Vasu Sharma, Gabriel Synnaeve, Huijiao Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *ArXiv*, abs/2304.07193, 2023. 2

[43] Xuran Pan, Zhuofan Xia, Shiji Song, Li Erran Li, and Gao Huang. 3d object detection with pointformer. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7459–7468, 2021. 1

[44] Yancheng Pan, Biao Gao, Jilin Mei, Sibo Geng, Chengkun Li, and Huijing Zhao. Semanticposs: A point cloud dataset with large quantity of dynamic instances. *CoRR*, abs/2002.09147, 2020. 2

[45] Songyou Peng, Kyle Genova, Chiyu Jiang, Andrea Tagliasacchi, Marc Pollefeys, and Thomas Funkhouser. Openscene: 3d scene understanding with open vocabularies. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–824, 2023. 2

[46] Quang-Hieu Pham, Duc Thanh Nguyen, Binh-Son Hua, Gemma Roig, and Sai-Kit Yeung. JSIS3D: joint semantic-instance segmentation of 3d point clouds with multi-task pointwise networks and multi-value conditional random fields. *CoRR*, abs/1904.00699, 2019. 1

[47] Charles Ruizhongtai Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. *CoRR*, abs/1612.00593, 2016. 1

[48] Charles Ruizhongtai Qi, Wei Liu, Chenxia Wu, Hao Su, and Leonidas J. Guibas. Frustum pointnets for 3d object detection from RGB-D data. *CoRR*, abs/1711.08488, 2017. 1

[49] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J. Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *CoRR*, abs/1706.02413, 2017. 1

[50] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junt-

ing Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos, 2024. 1, 2, 3, 5

[51] Ryan Razani, Ran Cheng, Ehsan Taghavi, and Liu Bingbing. Lite-hdseg: Lidar semantic segmentation using lite harmonic dense convolutions. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9550–9556, 2021. 2

[52] Ryan Razani, Ran Cheng, Ehsan Taghavi, and Liu Bingbing. Lite-HDSeg: LiDAR Semantic Segmentation Using Lite Harmonic Dense Convolutions. *arXiv preprint arXiv:2103.08852*, 2021. 4

[53] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, Zhaoyang Zeng, Hao Zhang, Feng Li, Jie Yang, Hongyang Li, Qing Jiang, and Lei Zhang. Grounded sam: Assembling open-world models for diverse visual tasks, 2024. 2

[54] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *CoRR*, abs/1505.04597, 2015. 2

[55] Chaitanya Ryali, Yuan-Ting Hu, Daniel Bolya, Chen Wei, Haoqi Fan, Po-Yao Huang, Vaibhav Aggarwal, Arkabandhu Chowdhury, Omid Poursaeed, Judy Hoffman, Jitendra Malik, Yanghao Li, and Christoph Feichtenhofer. Hiera: a hierarchical vision transformer without the bells-and-whistles. In *Proceedings of the 40th International Conference on Machine Learning*. JMLR.org, 2023. 2, 5

[56] Mohsen Shahraki, Ahmed Elamin, and Ahmed El-Rabbany. Samnet++: A segment anything model for supervised 3d point cloud semantic segmentation. *Remote Sensing*, 17(7), 2025. 2

[57] Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. Dinov3, 2025. 2

[58] Carole H. Sudre, Wenqi Li, Tom Vercauteren, Sébastien Ourselin, and M. Jorge Cardoso. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. *CoRR*, abs/1707.03237, 2017. 3

[59] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, Vijay Vasudevan, Wei Han, Jiquan Ngiam, Hang Zhao, Aleksei Timofeev, Scott Ettinger, Maxim Krivokon, Amy Gao, Aditya Joshi, Yu Zhang, Jonathon Shlens, Zhifeng Chen, and Dragomir Anguelov. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[60] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023. 2

[61] Quan Sun, Jinsheng Wang, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, and Xinlong Wang. Eva-clip-18b: Scaling clip to 18 billion parameters. *arXiv preprint arXiv:2402.04252*, 2023. 2

[62] Zhiyu Tan, Zichao Dong, Cheng Zhang, Weikun Zhang, Hang Ji, and Hao Li. Ovo: Open-vocabulary occupancy, 2023. 2

[63] Haotian Tang, Zhijian Liu, Shengyu Zhao, Yujun Lin, Ji Lin, Hanrui Wang, and Song Han. Searching efficient 3d architectures with sparse point-voxel convolution. *CoRR*, abs/2007.16100, 2020. 1

[64] Hugues Thomas, C. Qi, Jean-Emmanuel Deschaud, Beatriz Marcotegui, François Goulette, and Leonidas J. Guibas. Kpconv: Flexible and deformable convolution for point clouds. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6410–6419, 2019. 1, 2

[65] Larissa T. Triess, David Peter, Christoph B. Rist, and J. Marius Zöllner. Scan-based semantic segmentation of lidar point clouds: An experimental study. In *2020 IEEE Intelligent Vehicles Symposium (IV)*, pages 1116–1121, 2020. 2

[66] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017. 1

[67] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks, 2018. 1

[68] Bernie Wang, Virginia Wu, Bichen Wu, and Kurt Keutzer. Latte: accelerating lidar point cloud annotation via sensor fusion, one-click annotation, and tracking. In *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, pages 265–272. IEEE, 2019. 2

[69] Song Wang, Jianke Zhu, and Ruixiang Zhang. Metarangeseg: Lidar sequence semantic segmentation using multiple feature aggregation. *IEEE Robotics and Automation Letters*, 7(4):9739–9746, 2022. 2, 3

[70] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E. Sarma, Michael M. Bronstein, and Justin M. Solomon. Dynamic graph CNN for learning on point clouds. *CoRR*, abs/1801.07829, 2018. 1

[71] Zifu Wang, Xuefei Ning, and Matthew B. Blaschko. Jaccard metric losses: Optimizing the jaccard index with soft labels, 2024. 3

[72] Bichen Wu, Alvin Wan, Xiangyu Yue, and Kurt Keutzer. Squeezeseg: Convolutional neural nets with recurrent crf for real-time road-object segmentation from 3d lidar point cloud. In *ICRA*, 2018. 2

[73] Bichen Wu, Xuanyu Zhou, Sicheng Zhao, Xiangyu Yue, and Kurt Keutzer. Squeezesegv2: Improved model structure and unsupervised domain adaptation for road-object segmentation from a lidar point cloud. In *ICRA*, 2019. 2

[74] Dongyue Wu, Zilin Guo, Aoyan Li, Changqian Yu, Changxin Gao, and Nong Sang. Conditional boundary loss for semantic segmentation. *IEEE Transactions on Image Processing*, 32:3717–3731, 2023. 4

[75] Wenxuan Wu, Zhongang Qi, and Li Fuxin. Pointconv: Deep convolutional networks on 3d point clouds. In *2019*

IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 9613–9622, 2019. 1

[76] Xiaoyang Wu, Yixing Lao, Li Jiang, Xihui Liu, and Hengshuang Zhao. Point transformer v2: Grouped vector attention and partition-based pooling. In *NeurIPS*, 2022. 1

[77] Xiaoyang Wu, Li Jiang, Peng-Shuai Wang, Zhijian Liu, Xihui Liu, Yu Qiao, Wanli Ouyang, Tong He, and Hengshuang Zhao. Point transformer v3: Simpler, faster, stronger. In *CVPR*, 2024. 1, 5

[78] Xiaoyang Wu, Zhuotao Tian, Xin Wen, Bohao Peng, Xihui Liu, Kaicheng Yu, and Hengshuang Zhao. Towards large-scale 3d representation learning with multi-dataset point prompt training. In *CVPR*, 2024. 5

[79] Bin Xiao, Haiping Wu, Weijian Xu, Xiyang Dai, Houdong Hu, Yumao Lu, Michael Zeng, Ce Liu, and Lu Yuan. Florence-2: Advancing a unified representation for a variety of vision tasks, 2023. 2

[80] Xinyu Xiong, Zihuang Wu, Shuangyi Tan, Wenxue Li, Feilong Tang, Ying Chen, Siying Li, Jie Ma, and Guanbin Li. Sam2-unet: Segment anything 2 makes strong encoder for natural and medical image segmentation. *arXiv preprint arXiv:2408.08870*, 2024. 2, 3

[81] Chenfeng Xu, Bichen Wu, Zining Wang, Wei Zhan, Peter Vajda, Kurt Keutzer, and Masayoshi Tomizuka. Squeezesegv3: Spatially-adaptive convolution for efficient point-cloud segmentation. In *European Conference on Computer Vision*, pages 1–19. Springer, 2020. 2

[82] Chenfeng Xu, Bichen Wu, Wei Zhan, Peter Vajda, Kurt Keutzer, and Masayoshi Tomizuka. Rangevit: Vision transformer for lidar range image segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 27069–27078, 2021. 4

[83] Jianyun Xu, Song Wang, Ziqian Ni, Chunyong Hu, Sheng Yang, Jianke Zhu, and Qiang Li. Sam4d: Segment anything in camera and lidar streams, 2025. 2

[84] Xiang Xu, Lingdong Kong, Hui Shuai, and Qingshan Liu. Frnet: Frustum-range networks for scalable lidar segmentation. *IEEE Transactions on Image Processing*, 34:2173–2186, 2025. 2

[85] Xu Yan, Chaoda Zheng, Zhen Li, Sheng Wang, and Shuguang Cui. Pointasnl: Robust point clouds processing using nonlocal neural networks with adaptive sampling. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5588–5597, 2020. 1

[86] Yunhan Yang, Xiaoyang Wu, Tong He, Hengshuang Zhao, and Xihui Liu. Sam3d: Segment anything in 3d scenes, 2023. 2

[87] Haobo Yuan, Xiangtai Li, Tao Zhang, Zilong Huang, Shilin Xu, Shunping Ji, Yunhai Tong, Lu Qi, Jiashi Feng, and Ming-Hsuan Yang. Sa2va: Marrying sam2 with llava for dense grounded understanding of images and videos, 2025. 2

[88] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, Ce Liu, Mengchen Liu, Zicheng Liu, Yumao Lu, Yu Shi, Lijuan Wang, Jianfeng Wang, Bin Xiao, Zhen Xiao, Jianwei Yang, Michael Zeng, Luowei Zhou, and Pengchuan Zhang. Florence: A new foundation model for computer vision. *CoRR*, abs/2111.11432, 2021. 2

[89] Karim Abou Zeid, Kadir Yilmaz, Daan de Geus, Alexander Hermans, David Adrian, Timm Linder, and Bastian Leibe. Dino in the room: Leveraging 2d foundation models for 3d segmentation, 2025. 2

[90] Junbo Zhang, Runpei Dong, and Kaisheng Ma. Clip-fo3d: Learning free open-world 3d scene representations from 2d dense clip. In *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 2040–2051, 2023. 2

[91] Yang Zhang, Zixiang Zhou, Philip David, Xiangyu Yue, Zerong Xi, Boqing Gong, and Hassan Foroosh. Polarnet: An improved grid representation for online lidar point clouds semantic segmentation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9598–9607, 2020. 2

[92] Zhen Zhang, Yin Zhou, Hassan Foroosh, Mikko Lauri, Shiguang Li, Hongkai Xu, and Vincent Lepetit. Rangeformer: Point cloud semantic segmentation using range image based transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 22397–22410, 2022. 4

[93] Rongzhen Zhao, Vivienne Wang, Juho Kannala, and Joni Pajarinen. Multi-scale fusion for object representation, 2025. 1

[94] Yiming Zhao, Lin Bai, and Xinming Huang. FIDNet: LiDAR Point Cloud Semantic Segmentation with Fully Interpolation Decoding. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2021. 4

[95] Hui Zhou, Xinge Zhu, Xiao Song, Yuexin Ma, Zhe Wang, Hongsheng Li, and Dahua Lin. Cylinder3d: An effective 3d framework for driving-scene lidar semantic segmentation. *CoRR*, abs/2008.01550, 2020. 1

[96] Zhuangwei Zhuang, Rong Li, Kui Jia, Qicheng Wang, Yuanqing Li, and Mingkui Tan. Perception-aware multi-sensor fusion for 3d lidar semantic segmentation. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 16260–16270, 2021. 2

[97] Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Wang, Lijuan Wang, Jianfeng Gao, and Yong Jae Lee. Segment everything everywhere all at once, 2023. 2

[98] Özgür Özdemir and Elena Battini Sönmez. Weighted cross-entropy for unbalanced data with application on covid x-ray images. In *2020 Innovations in Intelligent Systems and Applications Conference (ASYU)*, pages 1–6, 2020. 3