OccRWKV: Rethinking 3D Semantic Occupancy Prediction with Linearly Scalable Inference

Junming Wang[1,2], Wei Yin[1], Xiaoxiao Long[2], Xingyu Zhang[1], Zebin Xing[1]

***Abstract*— 3D semantic occupancy prediction networks have demonstrated remarkable ability in reconstructing the geometric and semantic structure of 3D scenes, providing crucial information for robot navigation and autonomous driving systems. However, existing networks face challenges in balancing accuracy and real-time performance due to their dense network structure designs. In this paper, we introduce OccRWKV, the first RWKV-based 3D semantic occupancy network that addresses these challenges by leveraging novel network structures and insights from the sparse nature of real-world 3D occupancy. OccRWKV separates semantic and occupancy predictions into distinct branches, facilitating specialized learning and enhancing prediction accuracy. We integrate novel Sem-RWKV, Geo-RWKV, and BEV-RWKV blocks into these branches to capture long-distance dependencies critical for semantic accuracy and occupancy prediction. By projecting features into the bird's-eye view (BEV) space, we reduce fusion latency, enabling real-time inference without compromising performance. Experimental results demonstrate that OccRWKV achieves state-of-the-art performance on benchmark datasets while maintaining linear complexity, making it a promising solution for real-world deployment in robot navigation and autonomous driving systems. Finally, we will release our code for the reference of the community.**

I. INTRODUCTION

3D semantic occupancy prediction networks [1]–[3] have garnered significant attention in recent years due to their remarkable ability to reconstruct the geometric and semantic structure of 3D scenes, providing comprehensive occupancy maps and semantic information crucial for robot navigation tasks [4] and autonomous driving systems [2], [5], [6]. While existing single modality (i.e., LiDAR-based [4], [7], [8] and Camera-based [1], [3], [9]) and multi-modal networks [6] have made substantial progress in 3D semantic occupancy predictions utilizing 3D CNN [1] and transformer [10] architectures, their dense network structure designs, such as expensive feature fusion and global attention, have hindered their real-world deployment. Although some methods employ 2D convolution [4], [7] to reduce network complexity, the simplified network structure sacrifices prediction accuracy in favour of real-time performance, leaving room for improvement in achieving a balance between accuracy and efficiency.

To address these challenges, our key insights lie in rethinking and designing novel network structures that enable 3D semantic occupancy prediction networks to strike a balance between accuracy and real-time performance. Firstly, we observe that 3D occupancy in the real world is sparse, with most voxels being empty, suggesting the potential benefits of migrating dense feature fusion to the bird's-eye view (BEV) space [11]–[13]. Moreover, the recent RWKV [14], [15] model, which demonstrates efficient text processing capabilities in natural language processing (NLP) and shows promise for real-world deployment in image generation [16] with low memory usage, inspires us to explore its potential in 3D semantic occupancy prediction. This leads us to the question: *Can we design a 3D semantic occupancy network that achieves high performance while maintaining linear complexity?*

Building upon these insights, we introduce **OccRWKV**, the first RWKV-based 3D semantic occupancy network. In contrast to previous networks that jointly learn semantics and occupancy, OccRWKV separates these predictions into distinct branches. This separation facilitates specialized learning within each domain, enhancing prediction accuracy and fully leveraging the complementary properties of semantic and geometric features in the subsequent feature fusion stage. We integrate novel Sem-RWKV, Geo-RWKV, and BEV-RWKV blocks into these branches to capture long-distance dependencies critical for semantic accuracy and occupancy prediction. Furthermore, by projecting features into the BEV space, we reduce fusion latency, enabling real-time inference without compromising performance.

We first assessed OccRWKV on the SemanticKITTI benchmark, comparing its accuracy and inference speed to some leading occupancy networks. We also deployed OccRWKV on a real robot to test its efficiency in field deployment. Our evaluation reveals:

- **OccRWKV is high-performence.** OccRWKV achieves state-of-the-art performance (mIoU = 25.0) on the SemanticKITTI benchmark. (§ **??**)
- **OccRWKV is efficient.** OccRWKV enables high-speed inference (i.e., 22.1 FPS) and reduces the FLOPs. (§ **??**)
- **OccRWKV is scalable.** The RWKV block tailored for 3D semantic occupancy prediction can be easily integrated into single-mode/multi-mode networks. (§ **??**)

II. RELATED WORK

A. 3D Semantic Occupancy Prediction

3D semantic occupancy prediction is crucial for interpreting occluded environments, as it discerns the spatial layout beyond visual obstructions by merging geometry with semantic clues. This process enables autonomous systems to anticipate hidden areas, crucial for safe navigation and decision-making. Research on 3D semantic occupancy pre-

*Corresponding author.
[1]Horizon Robotics. [2]The University of Hong Kong, Hong Kong SAR, China. jmwang@cs.hku.hk

diction can be summarized into three main streams: *Camera-based* approaches capitalize on visual data, with pioneering works like MonoScene by Cao et al. [1] exploiting RGB inputs to infer indoor and outdoor occupancy. Another notable work by *Li et al.* [3] is VoxFormer, a transformer-based semantic occupancy framework capable of generating complete 3D volume semantics using only 2D images. *LiDAR-based* approaches like S3CNet by Cheng et al. [17], JS3C-Net by Yan et al. [18], and SSA-SC by Yang et al. [19], which adeptly handle the vastness and variability of outdoor scenes via point clouds. *Fusion-based* approaches aim to amalgamate the contextual richness of camera imagery with the spatial accuracy of LiDAR data. The Openoccupancy benchmark by Wang et al. [20] is a testament to this synergy, providing a platform to assess the performance of integrated sensor approaches.

III. MATH

Before you begin to format your paper, first write and save the content as a separate text file. Keep your text and graphic files separate until after the text has been formatted and styled. Do not use hard tabs, and limit use of hard returns to only one return at the end of a paragraph. Do not add any kind of pagination anywhere in the paper. Do not number text heads-the template will do that for you.

Finally, complete content and organizational editing before formatting. Please take note of the following items when proofreading spelling and grammar:

A. Abbreviations and Acronyms

Define abbreviations and acronyms the first time they are used in the text, even after they have been defined in the abstract. Abbreviations such as IEEE, SI, MKS, CGS, sc, dc, and rms do not have to be defined. Do not use abbreviations in the title or heads unless they are unavoidable.

B. Units

- Use either SI (MKS) or CGS as primary units. (SI units are encouraged.) English units may be used as secondary units (in parentheses). An exception would be the use of English units as identifiers in trade, such as Ò3.5-inch disk driveÓ.
- Avoid combining SI and CGS units, such as current in amperes and magnetic field in oersteds. This often leads to confusion because equations do not balance dimensionally. If you must use mixed units, clearly state the units for each quantity that you use in an equation.
- Do not mix complete spellings and abbreviations of units: ÒWb/m2Ó or Òwebers per square meterÓ, not Òwebers/m2Ó. Spell out units when they appear in text: Ò. . . a few henriesÓ, not Ò. . . a few HÓ.
- Use a zero before decimal points: Ò0.25Ó, not Ò.25Ó. Use Òcm3Ó, not ÒccÓ. (bullet list)

C. Equations

The equations are an exception to the prescribed specifications of this template. You will need to determine whether or not your equation should be typed using either the Times New Roman or the Symbol font (please no other font). To create multileveled equations, it may be necessary to treat the equation as a graphic and insert it into the text after your paper is styled. Number equations consecutively. Equation numbers, within parentheses, are to position flush right, as in (1), using a right tab stop. To make your equations more compact, you may use the solidus (/), the exp function, or appropriate exponents. Italicize Roman symbols for quantities and variables, but not Greek symbols. Use a long dash rather than a hyphen for a minus sign. Punctuate equations with commas or periods when they are part of a sentence, as in

$$\alpha + \beta = \chi \tag{1}$$

Note that the equation is centered using a center tab stop. Be sure that the symbols in your equation have been defined before or immediately following the equation. Use Ò(1)Ó, not ÒEq. (1)Ó or Òequation (1)Ó, except at the beginning of a sentence: ÒEquation (1) is . . .Ó

D. Some Common Mistakes

- The word ÒdataÓ is plural, not singular.
- The subscript for the permeability of vacuum ?0, and other common scientific constants, is zero with subscript formatting, not a lowercase letter ÒoÓ.
- In American English, commas, semi-/colons, periods, question and exclamation marks are located within quotation marks only when a complete thought or name is cited, such as a title or full quotation. When quotation marks are used, instead of a bold or italic typeface, to highlight a word or phrase, punctuation should appear outside of the quotation marks. A parenthetical phrase or statement at the end of a sentence is punctuated outside of the closing parenthesis (like this). (A parenthetical sentence is punctuated within the parentheses.)
- A graph within a graph is an ÒinsetÓ, not an ÒinsertÓ. The word alternatively is preferred to the word ÒalternatelyÓ (unless you really mean something that alternates).
- Do not use the word ÒessentiallyÓ to mean ÒapproximatelyÓ or ÒeffectivelyÓ.
- In your paper title, if the words Òthat usesÓ can accurately replace the word ÒusingÓ, capitalize the ÒuÓ; if not, keep using lower-cased.
- Be aware of the different meanings of the homophones ÒaffectÓ and ÒeffectÓ, ÒcomplementÓ and ÒcomplimentÓ, ÒdiscreetÓ and ÒdiscreteÓ, ÒprincipalÓ and ÒprincipleÓ.
- Do not confuse ÒimplyÓ and ÒinferÓ.
- The prefix ÒnonÓ is not a word; it should be joined to the word it modifies, usually without a hyphen.

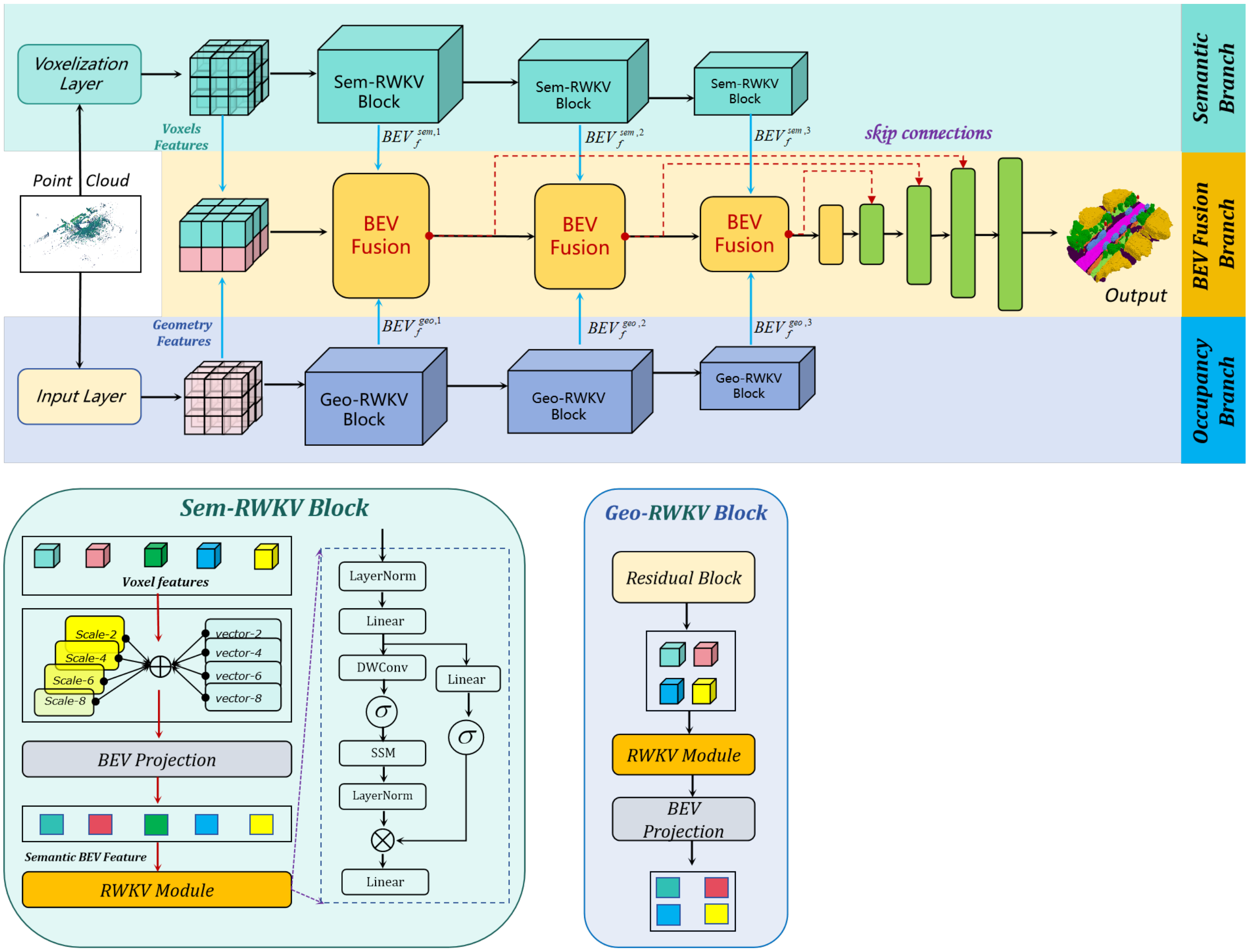

Fig. 1: OccRWKV system architecture. The perception network (i.e., OccMamba) and AGR-planner run asynchronously on the onboard computer, connected through a query-based map update method from [4] to ensure real-time local map updates with prediction results.

TABLE I: 3D Semantic occupancy prediction results on SemanticKITTI test set. The C and L denote Camera and LiDAR.

Method	Modality	IoU ↑	mIoU ↑	road (15.30%)	sidewalk (11.13%)	parking (1.12%)	other-grnd (0.56%)	building (14.1%)	car (3.92%)	truck (0.16%)	bicycle (0.03%)	motorcycle (0.03%)	other-veh. (0.20%)	vegetation (39.3%)	trunk (0.51%)	terrain (9.17%)	person (0.07%)	bicyclist (0.07%)	motorcyclist. (0.05%)	fence (3.90%)	pole (0.29%)	traf.-sign (0.08%)	FPS
MonoScene [1]	C	34.2	11.1	54.7	27.1	24.8	5.7	14.4	18.8	3.3	0.5	0.7	4.4	14.9	2.4	19.5	1.0	1.4	0.4	11.1	3.3	2.1	1.1
OccFormer [21]	C	34.5	12.3	55.9	30.3	31.5	6.5	15.7	21.6	1.2	1.5	1.7	3.2	16.8	3.9	21.3	2.2	1.1	0.2	11.9	3.8	3.7	1.8
VoxFormer [3]	C	43.2	13.4	54.1	26.9	25.1	7.3	23.5	21.7	3.6	1.9	1.6	4.1	24.4	8.1	24.2	1.6	1.1	13.1	6.6	5.7	8.1	1.5
TPVFormer [5]	C	34.3	11.3	55.1	27.2	27.4	6.5	14.8	19.2	3.7	1.0	0.5	2.3	13.9	2.6	20.4	1.1	2.4	0.3	11.0	2.9	1.5	1.0
LMSCNet [7]	L	55.3	17.0	64.0	33.1	24.9	3.2	38.7	29.5	2.5	0.0	0.0	0.1	40.5	19.0	30.8	0.0	0.0	0.0	20.5	15.7	0.5	21.3
SSC-RS [8]	L	59.7	24.2	**73.1**	44.4	38.6	17.4	**44.6**	36.4	5.3	10.1	5.1	**11.2**	44.1	26.0	41.9	4.7	2.4	0.9	30.8	15.0	7.2	16.7
SCONet [4]	L	56.1	17.6	51.9	30.7	23.1	0.9	39.9	29.1	1.7	0.8	0.5	4.8	41.4	27.5	28.6	0.8	0.5	0.1	18.9	21.4	8.0	20.0
M-CONet [20]	C&L	55.7	20.4	60.6	36.1	29.0	13.0	38.4	33.8	4.7	3.0	2.2	5.9	41.5	20.5	35.1	0.8	2.3	**0.6**	26.0	18.7	15.7	1.4
Co-Occ [6]	C&L	56.6	24.4	72.0	43.5	42.5	10.2	35.1	**40.0**	**6.4**	4.4	3.3	8.8	41.2	**30.8**	40.8	1.6	**3.3**	0.4	**32.7**	**26.6**	**20.7**	1.1
OccRWKV (Ours)	L	**59.9**	**25.0**	72.9	**44.8**	**42.7**	**18.1**	44.2	36.1	3.5	**12.3**	**6.0**	10.1	**44.6**	29.5	**42.1**	**5.9**	2.9	0.4	32.2	17.6	8.1	**22.1**

- There is no period after the ÒetÓ in the Latin abbreviation Òet al.Ó.
- The abbreviation Òi.e.Ó means Òthat isÓ, and the abbreviation Òe.g.Ó means Òfor exampleÓ.

IV. USING THE TEMPLATE

Use this sample document as your LaTeX source file to create your document. Save this file as **root.tex**. You have to make sure to use the cls file that came with this distribution. If you use a different style file, you cannot expect to get required margins. Note also that when you are creating your

TABLE II: 3D occupancy results on SemanticKITTI [22] validation set.

Method	IoU (%)	mIoU (%)	Prec. (%)	Recall (%)	Params(M)	FLOPs(G)	Mem. (GB)
MLP/CNN-based							
Monoscene [1]	37.1	11.5	52.2	55.5	149.6	501.8	20.3
NDC-Scene [23]	37.2	12.7	-	-	-	-	20.1
Symphonies [24]	41.9	14.9	62.7	55.7	59.3	611.9	20.0
Transformer-based							
OccFormer [21]	36.5	13.5	47.3	60.4	81.4	889.0	21.0
VoxFormer [3]	57.7	18.4	69.9	76.7	57.8	-	15.2
TPVFormer [5]	35.6	11.4	-	-	48.8	946.0	20.0
CGFormer [25]	45.9	16.9	62.8	63.2	122.4	**314.5**	19.3
Mamba-based (Ours)							
OccMamba	**58.6**	**25.2**	**77.8**	70.5	**23.8**	505.1	**3.5**

out PDF file, the source file is only part of the equation. *Your TEX → PDF filter determines the output file size. Even if you make all the specifications to output a letter file in the source - if your filter is set to produce A4, you will only get A4 output.*

It is impossible to account for all possible situation, one would encounter using TEX. If you are using multiple TEX files you must make sure that the “MAIN“ source file is called root.tex - this is particularly important if your conference is using PaperPlaza’s built in TEX to PDF conversion tool.

A. Headings, etc

Text heads organize the topics on a relational, hierarchical basis. For example, the paper title is the primary text head because all subsequent material relates and elaborates on this one topic. If there are two or more sub-topics, the next level head (uppercase Roman numerals) should be used and, conversely, if there are not at least two sub-topics, then no subheads should be introduced. Styles named ÒHeading 1Ó, ÒHeading 2Ó, ÒHeading 3Ó, and ÒHeading 4Ó are prescribed.

B. Figures and Tables

Positioning Figures and Tables: Place figures and tables at the top and bottom of columns. Avoid placing them in the middle of columns. Large figures and tables may span across both columns. Figure captions should be below the figures; table heads should appear above the tables. Insert figures and tables after they are cited in the text. Use the abbreviation ÒFig. 1Ó, even at the beginning of a sentence.

TABLE III: An Example of a Table

One	Two
Three	Four

Figure Labels: Use 8 point Times New Roman for Figure labels. Use words rather than symbols or abbreviations when writing Figure axis labels to avoid confusing the reader. As an example, write the quantity ÒMagnetizationÓ, or ÒMagnetization, MÓ, not just ÒMÓ. If including units in the label, present them within parentheses. Do not label axes only with units. In the example, write ÒMagnetization (A/m)Ó or ÒMagnetization A[m(1)]Ó, not just ÒA/mÓ. Do not label axes with a ratio of quantities and units. For example, write ÒTemperature (K)Ó, not ÒTemperature/K.Ó

We suggest that you use a text box to insert a graphic (which is ideally a 300 dpi TIFF or EPS file, with all fonts embedded) because, in an document, this method is somewhat more stable than directly inserting a picture.

Fig. 2: Inductance of oscillation winding on amorphous magnetic core versus DC bias magnetic field

V. CONCLUSIONS

A conclusion section is not required. Although a conclusion may review the main points of the paper, do not replicate the abstract as the conclusion. A conclusion might elaborate on the importance of the work or suggest applications and extensions.

APPENDIX

Appendixes should appear before the acknowledgment.

ACKNOWLEDGMENT

The preferred spelling of the word ÒacknowledgmentÓ in America is without an ÒeÓ after the ÒgÓ. Avoid the stilted expression, ÒOne of us (R. B. G.) thanks . . .Ó Instead, try ÒR. B. G. thanksÓ. Put sponsor acknowledgments in the unnumbered footnote on the first page.

References are important to the reader; therefore, each citation must be complete and correct. If at all possible, references should be commonly available publications.

REFERENCES

[1] A.-Q. Cao and R. de Charette, "Monoscene: Monocular 3d semantic scene completion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 3991–4001.

[2] P. Tang, Z. Wang, G. Wang, J. Zheng, X. Ren, B. Feng, and C. Ma, "Sparseocc: Rethinking sparse latent representation for vision-based semantic occupancy prediction," *arXiv preprint arXiv:2404.09502*, 2024.

[3] Y. Li, Z. Yu, C. Choy, C. Xiao, J. M. Alvarez, S. Fidler, C. Feng, and A. Anandkumar, "Voxformer: Sparse voxel transformer for camera-based 3d semantic scene completion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 9087–9098.

[4] J. Wang, Z. Sun, X. Guan, T. Shen, Z. Zhang, T. Duan, D. Huang, S. Zhao, and H. Cui, "Agrnav: Efficient and energy-saving autonomous navigation for air-ground robots in occlusion-prone environments," *arXiv preprint arXiv:2403.11607*, 2024.

[5] Y. Huang, W. Zheng, Y. Zhang, J. Zhou, and J. Lu, "Tri-perspective view for vision-based 3d semantic occupancy prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 9223–9232.

[6] J. Pan, Z. Wang, and L. Wang, "Co-occ: Coupling explicit feature fusion with volume rendering regularization for multi-modal 3d semantic occupancy prediction," *IEEE Robotics and Automation Letters*, 2024.

[7] L. Roldao, R. de Charette, and A. Verroust-Blondet, "Lmscnet: Lightweight multiscale 3d semantic completion," in *2020 International Conference on 3D Vision (3DV)*. IEEE, 2020, pp. 111–119.

[8] J. Mei, Y. Yang, M. Wang, T. Huang, X. Yang, and Y. Liu, "Ssc-rs: Elevate lidar semantic scene completion with representation separation and bev fusion," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2023, pp. 1–8.

[9] Y. Wei, L. Zhao, W. Zheng, Z. Zhu, J. Zhou, and J. Lu, "Surroundocc: Multi-camera 3d occupancy prediction for autonomous driving," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 21 729–21 740.

[10] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[11] M. Li, Y. Zhang, X. Ma, Y. Qu, and Y. Fu, "Bev-dg: Cross-modal learning under bird's-eye view for domain generalization of 3d semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 11 632–11 642.

[12] Z. Li, W. Wang, H. Li, E. Xie, C. Sima, T. Lu, Y. Qiao, and J. Dai, "Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers," in *European conference on computer vision*. Springer, 2022, pp. 1–18.

[13] Z. Liu, H. Tang, A. Amini, X. Yang, H. Mao, D. L. Rus, and S. Han, "Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation," in *2023 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2023, pp. 2774–2781.

[14] Y. Duan, W. Wang, Z. Chen, X. Zhu, L. Lu, T. Lu, Y. Qiao, H. Li, J. Dai, and W. Wang, "Vision-rwkv: Efficient and scalable visual perception with rwkv-like architectures," *arXiv preprint arXiv:2403.02308*, 2024.

[15] B. Peng, E. Alcaide, Q. Anthony, A. Albalak, S. Arcadinho, S. Biderman, H. Cao, X. Cheng, M. Chung, M. Grella *et al.*, "Rwkv: Reinventing rnns for the transformer era," *arXiv preprint arXiv:2305.13048*, 2023.

[16] Z. Fei, M. Fan, C. Yu, D. Li, and J. Huang, "Diffusion-rwkv: Scaling rwkv-like architectures for diffusion models," *arXiv preprint arXiv:2404.04478*, 2024.

[17] R. Cheng, C. Agia, Y. Ren, X. Li, and L. Bingbing, "S3cnet: A sparse semantic scene completion network for lidar point clouds," in *Conference on Robot Learning*. PMLR, 2021, pp. 2148–2161.

[18] X. Yan, J. Gao, J. Li, R. Zhang, Z. Li, R. Huang, and S. Cui, "Sparse single sweep lidar point cloud segmentation via learning contextual shape priors from scene completion," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 4, 2021, pp. 3101–3109.

[19] X. Yang, H. Zou, X. Kong, T. Huang, Y. Liu, W. Li, F. Wen, and H. Zhang, "Semantic segmentation-assisted scene completion for lidar point clouds," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021, pp. 3555–3562.

[20] X. Wang, Z. Zhu, W. Xu, Y. Zhang, Y. Wei, X. Chi, Y. Ye, D. Du, J. Lu, and X. Wang, "Openoccupancy: A large scale benchmark for surrounding semantic occupancy perception," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 17 850–17 859.

[21] Y. Zhang, Z. Zhu, and D. Du, "Occformer: Dual-path transformer for vision-based 3d semantic occupancy prediction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 9433–9443.

[22] 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 *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 9297–9307.

[23] J. Yao, C. Li, K. Sun, Y. Cai, H. Li, W. Ouyang, and H. Li, "Ndc-scene: Boost monocular 3d semantic scene completion in normalized device coordinates space," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE Computer Society, 2023, pp. 9421–9431.

[24] H. Jiang, T. Cheng, N. Gao, H. Zhang, W. Liu, and X. Wang, "Symphonize 3d semantic scene completion with contextual instance queries," *arXiv preprint arXiv:2306.15670*, 2023.

[25] Z. Yu, R. Zhang, J. Ying, J. Yu, X. Hu, L. Luo, S. Cao, and H. Shen, "Context and geometry aware voxel transformer for semantic scene completion," *arXiv preprint arXiv:2405.13675*, 2024.