Proceedings of the Semantic Scene Representations (SSR) Workshop at AIROV 2026

Christian Rauch
Chair of Cyber Physical Systems
TU Leoben, Austria
christian.rauch@unileoben.ac.at

Linus Nwankwo
Chair of Cyber Physical Systems
TU Leoben, Austria
linus.nwankwo@unileoben.ac.at

Shail Jadav
Autonomous Systems Lab
TU Wien, Austria
shail.jadav@tuwien.ac.at

Jun Zhang
Institute of Visual Computing
TU Graz, Austria
jun.zhang@tugraz.at

Abstract

Autonomous systems operating in real-world environments around humans face fundamental challenges in generalisability and explainability. As the complexity of these systems and their expected tasks grows, there is an increasing need for high-level interfaces grounded in semantic concepts from natural language. Language-grounded representations promise to facilitate scene understanding, enable generalisation to unseen environments and tasks, improve explainability of autonomous agent actions to human operators, and increase accessibility of complex systems for the general public without requiring domain-specific expertise. This workshop presents and discusses recent advances in semantic and natural-language-based 3D scene and environment representations, as well as localisation and mapping techniques, as foundational building blocks for interactive robotic tasks. The workshop brings together researchers working at the intersection of computer vision, robotics, and natural language processing to address open problems in open-set perception, multi-modal scene understanding, and semantic navigation.

1 Introduction

Autonomous robotic systems are increasingly expected to operate in unstructured, dynamic environments alongside humans. Traditional perception and control pipelines, while effective in constrained settings, often struggle with open-world generalisation, semantic reasoning, and human-interpretable decision-making [13, 19]. The research community has recognised that purely geometric or low-level visual representations are insufficient for the next generation of interactive robots that must understand and communicate about the world in human-centric terms [25, 6, 12].

The Semantic Scene Representations (SSR) workshop addresses this critical gap by focusing on the development and application of semantic and natural-language-grounded 3D scene representations. These representations serve as a bridge between low-level sensory data and high-level cognitive reasoning, enabling robots to parse environments in terms of objects, relations, and affordances that are expressible in natural language. By grounding perception in semantic concepts, we aim to facilitate scene understanding that generalises across novel environments, supports task abstraction and planning, and provides explainable interfaces between autonomous agents and human users.

The relevance of this research direction to the broader machine learning and robotics communities lies in its interdisciplinary nature: it draws upon advances in vision-language models [18], computer-

The Third Austrian Symposium on AI, Robotics, and Vision (AIROV26).

vision models [16, 23], neural scene representations [21, 11, 20], simultaneous localisation and mapping (SLAM) [22, 8, 24, 3], and human-robot interaction [7, 1, 15].

The workshop objectives are:

1. survey the state of the art in semantic scene representations for robotics,
2. identify critical methodological and theoretical gaps,
3. foster cross-disciplinary dialogue between vision, language, and robotics researchers, and
4. establish benchmarks and evaluation protocols that reflect real-world deployment requirements.

2 SSR Workshop Scope and Topics

2.1 Scope

SSR workshop scope spans theoretical foundations, algorithmic methodologies, evaluation frameworks, and application domains related to semantic scene understanding for autonomous systems.

Theoretical foundations. This encompasses the principles underlying language-grounded scene representations, including the mapping between continuous sensory signals and discrete semantic concepts, the role of foundation models in bridging perception and language, and the formal properties of representations that support compositional generalisation and causal reasoning about scene dynamics.

Algorithms and models. This forms the core technical focus, addressing open-set vision-language architectures for robotic perception, explicit and implicit 3D scene representations including neural radiance fields and Gaussian splatting [10, 9], semantic and neural-based SLAM systems [24, 2], and language-guided robot motion policies [15]. Particular emphasis is placed on self-supervised and weakly supervised learning paradigms that reduce reliance on densely annotated training data, as well as on multi-modal fusion techniques that integrate visual, linguistic, and geometric information.

Evaluation methodologies. This includes benchmarks and metrics for open-set object identification and localisation, robustness evaluation protocols for perception systems in degraded or adversarial conditions, and simulation-based testing frameworks that enable reproducible comparison of semantic mapping approaches across diverse environmental conditions.

Application domains. This covers interactive robot tasks such as manipulation and navigation in human-populated spaces, perception for safe human-robot interaction, autonomous navigation in visually challenging environments, including dense vegetation and featureless tunnels [3], and intralogistics scenarios requiring accurate trajectory prediction under dynamic contextual conditions [17]. The SSR workshop also examines the deployment of neuromorphic sensors, particularly event cameras, for visual localisation in challenging illumination conditions [4, 5, 14].

2.2 Topics

The SSR workshop topics include, but are not limited to:

- Open-set vision-language models for robotics
- Human-robot natural language interfaces
- Planning and reasoning on abstract task goals
- Object detection and state estimation
- Learning multi-modal representations
- Explicit and implicit 3D scene representations
- Interactive robot tasks (manipulation, navigation, ...)
- Semantic and neural-based SLAM

- Benchmarks and metrics for open-set object identification and localisation
- Perception for safe human-robot interaction
- Language-guided robot motion policies
- Visual localisation with neuromorphic sensor (event camera)
- Generalizable neural rendering and Gaussian Splatting
- Multi-view stereo and structure from motion

3 Organization

The SSR workshop is organised by a committee with complementary expertise spanning cyber-physical systems, autonomous robotics, and visual computing.

Organising Committee:

- Christian Rauch (Chair of Cyber Physical Systems, TU Leoben)
- Linus Nwankwo (Chair of Cyber Physical Systems, TU Leoben)
- Shail Jadav (Autonomous Systems Lab, TU Wien)
- Jun Zhang (Institute of Visual Computing, TU Graz)

The review process (Section 4) for contributed papers follows a standard double-blind peer-review protocol with at least one independent reviewer per submission. Review criteria emphasise technical novelty, methodological rigour, reproducibility of experimental results, and relevance to the workshop's focus on semantic scene representations for autonomous systems.

4 Submissions and Review Process

The SSR workshop solicits original research contributions aligned with the scope described in Section 2. Submissions undergo a double-blind peer-review process with at least one independent reviewer per paper. Review criteria assess: (i) novelty of the proposed approach or theoretical insight; (ii) technical rigour and soundness of methodology; (iii) clarity of presentation and reproducibility of results; and (iv) relevance to semantic scene representations and their application in autonomous systems.

All accepted submissions are required to present a poster, and for SSR, all papers were also allocated a 10 + 5-minute presentation slot. In addition, all camera-ready papers are part of the official proceedings. Submissions for which no camera-ready version was provided are excluded from the proceedings. Table 1 shows specific submission statistics for the SSR workshop.

5 Program Overview

The one-day SSR workshop programme balances invited keynote presentations with peer-reviewed contributed talks, providing both a broad perspective and focused technical depth.

5.1 Keynote Presentations

The programme features two keynotes:

- Prof. Dongheui Lee (TU Wien) presents on multimodal scene understanding for human and robot action monitoring, addressing the integration of perceptual cues across modalities for joint activity recognition
- Prof. Kevin Sebastian Luck (VU Amsterdam) discusses exploiting morphological intelligence and text-to-network policy synthesis, examining how robot embodiment and natural language specifications can jointly inform control policy generation

Table 1: Submission Statistics

Total Submissions: 7, **Accepted:** 6, **Rejected:** 1			
Paper ID	**Description**	**Authors**	**Camera-ready**
35	Enhanced Environmental Context Encoding for Accurate Trajectory Prediction in Intralogistics	Alexander Prutsch, Horst Possegger	Yes
39	D^2DINO: Dense Descriptors from DINO for Pixel-Level Object Understanding	Paolo Sebeto, Jean-Baptiste Weibel, Christian Hartl-Nesic, Markus Vincze	Yes
89	Overcoming Nature: Perception for Autonomous Navigation in Dense Vegetation	Lukas Wimmer, Andre Koczka, Uros Petrovic, Gerald Steinbauer-Wagner	Yes
45	A Simulation-based Benchmark for LiDAR SLAM in featureless Tunnels with a novel Robustness Evaluation	Qiuyi Cao	No
52	Event Camera Localization in LiDAR Maps via Event-to-LiDAR Depth Registration	Kuangyi Chen	No
92	Good Deep Features to Track: Self-Supervised Feature Extraction and Tracking in Visual Odometry	Sai Puneeth Reddy Gottam, Haoming Zhang	No

5.2 Contributed Oral Presentations

The peer-reviewed oral session includes six technical presentations of all the accepted papers (see Table 1). Table 2 summarises the programme schedule and technical presentations.

5.3 Session Structure

The programme is organised into two main sessions with a coffee break in between (see Table 2). The opening session (14:00 - 15:30) comprises the first keynote and three contributed talks. The afternoon session (16:00 - 17:30) comprises the second keynote and the remaining contributed talks, concluding with a closing remarks session.

6 Conclusion and Outlook

The SSR workshop at AIRoV 2026 identifies semantic and language-grounded scene representations as a critical enabling technology for the next generation of autonomous systems. By bringing together advances in vision-language modelling, 3D scene representation, and robotic perception, the workshop contributes to a research agenda that prioritises generalisability, explainability, and human accessibility. Expected impacts on the field include the establishment of evaluation benchmarks that reflect real-world operational conditions, the dissemination of self-supervised and foundation-model-based techniques that reduce reliance on manual annotation, and the cultivation of interdisciplinary collaboration between the computer vision, natural language processing, and robotics communities. Future research directions emerging from the workshop programme include: scaling semantic representations to dynamic, temporally extended scenes; developing tighter integration between language understanding and motion planning; improving robustness of open-set perception under distribution shift; and deploying these technologies on resource-constrained robotic platforms with real-time requirements. The workshop serves as a forum for defining these challenges and coordinating community efforts toward their resolution.

Table 2: Programme Schedule

Time	Title	Speaker
14:00	Opening	Organisers
14:05	Keynote: Multimodal Scene Understanding for Human and Robot Action Monitoring	Prof. Dongheui Lee (TU Wien)
14:45	D²DINO: Dense Descriptors from DINO for Pixel-Level Object Understanding	Paolo Sebeto (TU Wien)
15:00	Good Deep Features to Track: Self-Supervised Feature Extraction and Tracking in Visual Odometry	Sai Puneeth Reddy Gottam (TU Leoben)
15:15	A Simulation-based Benchmark for LiDAR SLAM in featureless Tunnels with a novel Robustness Evaluation	Qiuyi Cao (Virtual Vehicle Research GmbH)
Coffee Break — 15:30 – 16:00		
16:00	Keynote: Exploiting Morphological Intelligence and Text-to-Network Policy Synthesis	Prof. Kevin Sebastian Luck (VU Amsterdam)
16:40	Event Camera Localization in LiDAR Maps via Event-to-LiDAR Depth Registration	Kuangyi Chen (TU Graz)
16:55	Overcoming Nature: Perception for Autonomous Navigation in Dense Vegetation	Uroš Petrović (TU Graz)
17:10	Enhanced Environmental Context Encoding for Accurate Trajectory Prediction in Intralogistics	Alexander Prutsch (TU Graz)
17:25 - 17:30	Closing	Organisers

Acknowledgments and Disclosure of Funding

The organisers gratefully acknowledge the support of the respective institutions: TU Leoben, TU Wien, and TU Graz. The AIRoV 2026 conference infrastructure and logistical support are essential to the workshop's realisation.

References

[1] Anthony Brohan, Yevgen Chebotar, Chelsea Finn, Karol Hausman, Alexander Herzog, Daniel Ho, Julian Ibarz, Alex Irpan, Eric Jang, Ryan Julian, et al. Do as i can, not as i say: Grounding language in robotic affordances. In *Conference on robot learning*, pages 287–318. PMLR, 2023.

[2] Leonard Bruns, Jun Zhang, and Patric Jensfelt. Neural graph map: Dense mapping with efficient loop closure integration. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2900–2909, 2025.

[3] Qiuyi Cao, Stephanie Grubmüller, Berin Dikic, Selim Solmaz, Pamela Innerwinkler, Haris Šikić, and Eduardo Enrique Veas. A simulation-based benchmark for lidar slam in featureless tunnels with a novel robustness evaluation. In *The IEEE Intelligent Vehicles Symposium (IV)*. IEEE CS, 2026.

[4] Kuangyi Chen, Jun Zhang, and Friedrich Fraundorfer. Evloc: Event-based visual localization in lidar maps via event-depth registration. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5751–5757, 2025.

[5] Kuangyi Chen, Jun Zhang, Yuxi Hu, Yi Zhou, and Friedrich Fraundorfer. Lear: Learning edge-aware representations for event-to-lidar localization. In *2026 IEEE International Conference on Robotics and Automation (ICRA)*, 2026.

[6] Sourav Garg, Krishan Rana, Mehdi Hosseinzadeh, Lachlan Mares, Niko Sünderhauf, Feras Dayoub, and Ian Reid. Robohop: Segment-based topological map representation for open-world visual navigation. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4090–4097. IEEE, 2024.

[7] Michael A Goodrich and Alan C Schultz. Human–robot interaction: a survey. *Foundations and trends® in human–computer interaction*, 1(3):203–275, 2008.

[8] Giorgio Grisetti, Rainer Kümmerle, Cyrill Stachniss, and Wolfram Burgard. A tutorial on graph-based slam. *IEEE Intelligent Transportation Systems Magazine*, 2(4):31–43, 2011.

[9] Yuxi Hu, Jun Zhang, Kuangyi Chen, Zhe Zhang, and Friedrich Fraundorfer. C^3-gs: Learning context-aware, cross-dimension, cross-scale feature for generalizable gaussian splatting. In *36th British Machine Vision Conference 2025, BMVC 2025, Sheffield, UK, November 24-27, 2025*. BMVA, 2025.

[10] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, George Drettakis, et al. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023.

[11] Amit Pal Singh Kohli, Vincent Sitzmann, and Gordon Wetzstein. Semantic implicit neural scene representations with semi-supervised training. In *2020 International Conference on 3D Vision (3DV)*, pages 423–433. IEEE, 2020.

[12] Thang-Anh-Quan Nguyen, Amine Bourki, Mátyás Macudzinski, Anthony Brunel, and Mohammed Bennamoun. Semantically-aware neural radiance fields for visual scene understanding: A comprehensive review. *International Journal of Computer Vision*, 134(3):109, 2026.

[13] Linus Nwankwo, Bjoern Ellensohn, Vedant Dave, Peter Hofer, Jan Forstner, Marlene Villneuve, Robert Galler, and Elmar Rueckert. Envodat: A large-scale multisensory dataset for robotic spatial awareness and semantic reasoning in heterogeneous environments. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 153–160. IEEE, 2025.

[14] Linus Nwankwo and Elmar Rueckert. Understanding why slam algorithms fail in modern indoor environments. In *International Conference on Robotics in Alpe-Adria Danube Region*, pages 186–194. Springer, 2023.

[15] Linus Nwankwo and Elmar Rueckert. The conversation is the command: Interacting with real-world autonomous robots through natural language. In *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, pages 808–812, 2024.

[16] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research Journal*, 2024.

[17] Alexander Prutsch, Matthias Wess, and Horst Possegger. Lanes are not enough: Enhancing trajectory prediction in intralogistics through detailed environmental context. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 8134–8141, 2025.

[18] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[19] Christian Rauch, Björn Ellensohn, Linus Nwankwo, Vedant Dave, and Elmar Rueckert. Real-time 3d vision-language embedding mapping. *arXiv preprint arXiv:2508.06291*, 2025.

[20] Vincent Sitzmann, Semon Rezchikov, Bill Freeman, Josh Tenenbaum, and Fredo Durand. Light field networks: Neural scene representations with single-evaluation rendering. *Advances in Neural Information Processing Systems*, 34:19313–19325, 2021.

[21] Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. *Advances in neural information processing systems*, 32, 2019.

[22] Cyrill Stachniss, John J Leonard, and Sebastian Thrun. Simultaneous localization and mapping. In *Springer handbook of robotics*, pages 1153–1176. Springer, 2016.

[23] Ao Wang, Hui Chen, Lihao Liu, Kai Chen, Zijia Lin, Jungong Han, and Guiguang Ding. Yolov10: Real-time end-to-end object detection. *Advances in neural information processing systems*, 37:107984–108011, 2024.

[24] Thomas Whelan, Stefan Leutenegger, Renato F Salas-Moreno, Ben Glocker, and Andrew J Davison. Elasticfusion: Dense slam without a pose graph. In *Robotics: science and systems*, volume 11. Rome, Italy, 2015.

[25] Shuaifeng Zhi, Tristan Laidlow, Stefan Leutenegger, and Andrew J Davison. In-place scene labelling and understanding with implicit scene representation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15838–15847, 2021.