Volume 11, Number 2, 2025 210

Transformer-Based Network for Robust 3D Industrial Environment Understanding in Autonomous UAV Systems

Oleksii Kuchkin, Artem Sazonov*, Iryna Cherepanska, Anatoliy Zhuchenko

National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", 37 Peremohy Ave., Kyiv, 03056, Ukraine

Received: October 15, 2025. Revised: December 11, 2025. Accepted: December 18, 2025.

Abstract

Autonomous navigation of unmanned aerial vehicles (UAVs) in unstructured industrial environments remains challenging due to irregular geometry, dynamic obstacles and sensor uncertainty. Classical Simultaneous Localization and Mapping (SLAM) systems, though geometrically consistent, often fail under poor initialization, textureless areas or reflective surfaces. To overcome these issues, this work proposes a hybrid transformer-geometric framework that fuses learned scene priors with keyframe-based SLAM. A TinyViT encoder and lightweight multi-task decoder jointly estimate inverse depth, surface normals and semantic segmentation, providing dense geometric and semantic cues that stabilize localization and mapping. These priors are incorporated into the SLAM optimization to enhance convergence, reject dynamic objects and improve relocalization. The system operates near real-time (~1 FPS) on a Raspberry Pi 5 CPU, suitable for keyframe-level inference. Experiments show robust localization and consistent mapping in cluttered, reflective and dynamic industrial scenes, confirming that transformer-based dense perception effectively complements classical SLAM for resource-efficient UAV navigation.

Keywords: UAV robust control; computer vision; SLAM; transformer-based neural network; autonomous navigation; 3D environment understanding.

1. Introduction

Autonomous navigation of unmanned aerial vehicles (UAVs) i.e. quadcopters in unstructured industrial environments remains a vital task for modern mobile robotics. In such settings, the surrounding scene is typically characterized by irregular geometry, varying illumination, dynamic obstacles and sensor noise. These factors lead to incomplete or inconsistent sensory data, making real-time localization, mapping, and trajectory planning highly uncertain. Moreover, the limited computational resources available on embedded systems further constrain the use of computationally intensive algorithms for reliable navigation.

Conventional Simultaneous Localization and Mapping (SLAM) methods, though widely used, remain limited when applied to stochastic and unstructured environments. Traditional feature-based or keyframe-based SLAM approaches rely heavily on precise initial conditions and accurate sensor calibration. Without sufficiently accurate initialization, the optimization problem behind SLAM may diverge or converge to suboptimal local minima. Furthermore, these methods typically assume static or near-static environments, while industrial spaces often contain reflective, transparent or dynamic objects that produce unstable key-points and degrade the quality of the generated map. As a result, the robustness and reliability of autonomous navigation are significantly reduced under real-world conditions.

* Corresponding author. Email address: a.sazonov@kpi.ua

This paper should be cited as: O. Kuchkin, A. Sazonov, I. Cherepanska, A. Zhuchenko. (2025). Transformer-based network for robust 3D industrial environment understanding in autonomous UAV systems. *Energy Engineering and Control Systems*, Vol. 11, No. 2, pp. 210 – 216. https://doi.org/10.23939/jeecs2025.02.210

To address these challenges, recent developments in neural scene understanding offer a promising alternative. Transformer-based neural architectures demonstrate exceptional capabilities when capturing long-range dependencies and global spatial relationships in both two-dimensional and three-dimensional data. Unlike convolutional networks, transformers process visual information through self-attention mechanisms, enabling them to model context and uncertainty across the entire visual field. This property is extremely beneficial for unstructured scenes, where local geometric cues alone are insufficient for reliable understanding and further autonomous navigation.

In this work, we propose a transformer-based approach for predicting 3D scene parameters – such as depth, surface normal orientation, semantic segmentation, and object class distributions – directly from visual observations. The predicted scene structure serves as a prior for keyframe-based SLAM, providing an informed initialization that improves convergence stability and reduces localization errors. Additionally, semantic segmentation allows the system to exclude points associated with reflective or dynamically moving objects from the optimization process, preventing instability in pose estimation. When trained on large-scale datasets encompassing diverse visual contexts, including segmentation supervision derived from EfficientSAM, the model demonstrates robust generalization across a wide range of industrial environments and illumination conditions.

The integration of transformer-based scene prediction with traditional SLAM approaches enables a hybrid navigation framework that combines the interpretability of geometric methods with deep learning adaptiveness. Such a hybrid approach enhances localization robustness, reduces computational overhead and enables more efficient decision-making for autonomous quadcopter control in stochastic and dynamically changing unstructured industrial environments.

2. Analysis of related works

Autonomous UAV navigation in unstructured and visually challenging industrial environments lies at the convergence of geometric SLAM, learning-based perception, and resource-constrained onboard control. Classical geometric SLAM-both key-frame and feature-based pipelines (e.g., ORB-style systems, direct dense methods)-remains widely used due to well-understood models of camera geometry, principled optimization, and interpretability. However, such systems exhibit characteristic failure modes under real-world conditions: poor or ambiguous initialization, reflective or transparent surfaces, dynamic obstacles, or high sensor noise [1], [2]. Critically, bundle adjustment and pose-graph optimization depend on reasonably accurate initial poses and reliable correspondences. When these prerequisites are violated, the resulting non-linear optimization may diverge or converge to incorrect local minima, producing tracking dropouts or severely distorted maps [2], [3].

To alleviate these limitations of purely geometric pipelines, a second class of approaches augments SLAM with learned priors. These hybrid geometric-neural methods integrate compact depth codes, CNN-based dense prediction models, or latent scene representations into the optimization loop [3], [4]. By providing coarse but globally consistent scene structure, such priors improve robustness in weakly textured or ambiguous regions and regularize dense mapping. Yet the majority of these approaches still rely on convolutional backbones or low-dimensional latent codes, which inherently limit the ability to model long-range spatial dependencies and global context – an important aspect of complex industrial scenes with clutter, large structures, and variable lighting.

A third category, neural SLAM and neural-implicit mapping, replaces explicit geometric representations with learned continuous fields. Neural radiance field (NeRF)-inspired methods, neural implicit mapping (e.g., NICE-SLAM) and end-to-end learned tracking-mapping architectures [5], [6] can recover high-fidelity scene geometry and demonstrate resilience to noise or partial observations. Nevertheless, because these systems typically maintain dense neural fields or multi-level feature grids, they suffer from scalability constraints and are difficult to deploy on resource-limited UAV platforms that require strict real-time guarantees. Although hierarchical latent grids and local implicit representations partially mitigate this overhead, integrating these dense neural structures with real-time control pipelines and onboard estimation remains an open challenge [7].

A more recent lineage of SLAM-relevant research investigates transformer-based perception and multimodal fusion as a means to enhance mapping inputs. Vision Transformers (ViTs), hierarchical variants such as Swin, as well as transformer-based dense predictors and point-cloud transformers (e.g., DPT, Point Transformer) demonstrate strong global reasoning capabilities for depth estimation, surface normal prediction, and semantic segmentation [8]–[11]. Furthermore, multimodal transformers-such as BEV/3D fusion models (e.g., BEVFormer, TransFusion)-leverage global cross-attention to align heterogeneous inputs (RGB, depth, LiDAR, inertial cues) and temporal information, which directly aligns with UAV sensing conditions under calibration uncertainty [12]. For tasks central to SLAM, transformer-augmented depth and segmentation networks (e.g., AdaBins, DPT, BinsFormer) provide improved

accuracy and uncertainty estimation, while large pre-trained models such as SAM and EfficientSAM supply reliable masks that enable filtering of dynamic or reflective objects during geometric optimization [13]–[16].

Finally, learned optimization frameworks represent a distinct category that reinterprets tracking and mapping as differentiable iterative refinement. Systems such as DROID-SLAM show that learned, recurrent update rules can improve resilience to poor initialization and reduce catastrophic failures without fully replacing the geometric pipeline [16]. These frameworks highlight an emerging direction: using learned modules not as end-to-end replacements, but as components that supply robust priors or corrections to traditional SLAM optimization.

Across these SLAM paradigms-geometric, hybrid geometric–neural, neural-implicit, transformer-based perception, and learned optimization – emerges a common insight: geometric SLAM remains efficient, explainable, and suited to real-time UAV deployment, but its performance degrades in visually challenging conditions; meanwhile, fully neural alternatives offer richer scene understanding but often impose prohibitive computational and scalability costs. This motivates a balanced strategy in which learning contributes high-level global scene priors while geometric optimization maintains accuracy and real-time reliability. The approach proposed in this work follows precisely this principle: a transformer-based dense perception module is executed only on keyframes to predict depth, surface normals, semantic masks, and uncertainty maps, which then provide informed initialization and robust correspondence selection for a conventional key-frame SLAM backend. By restricting learned inference to keyframes and leveraging segmentation-aware rejection of unstable points, the system preserves the efficiency and interpretability of geometric SLAM while substantially improving convergence stability and robustness in industrial UAV environments characterized by reflective materials, dynamic objects and variable lighting.

3. Algorithms and methods

This section outlines the proposed hybrid perception and mapping framework, which integrates transformer-based dense prediction with traditional keyframe-based SLAM for robust autonomous navigation in unstructured industrial environments. The overall system comprises a lightweight transformer encoder-decoder network for multi-task scene understanding and a tightly coupled SLAM module that exploits the predicted geometric and semantic priors to improve localization stability and mapping fidelity.

3.1. Overview

In this work, we propose a hybrid perception and mapping framework designed for autonomous UAV navigation in unstructured industrial environments. The framework leverages a transformer-based encoder-decoder architecture to generate dense predictions of inverse depth, surface normals, and semantic segmentation from monocular RGB inputs. These dense priors are subsequently integrated into a keyframe-based VIO backend to enhance robustness and reduce the risk of divergence during bundle adjustment. The high-level architecture is depicted in Fig. 1. A TinyViT backbone serves as the lightweight visual encoder, providing multi-scale features to a multi-head decoder.

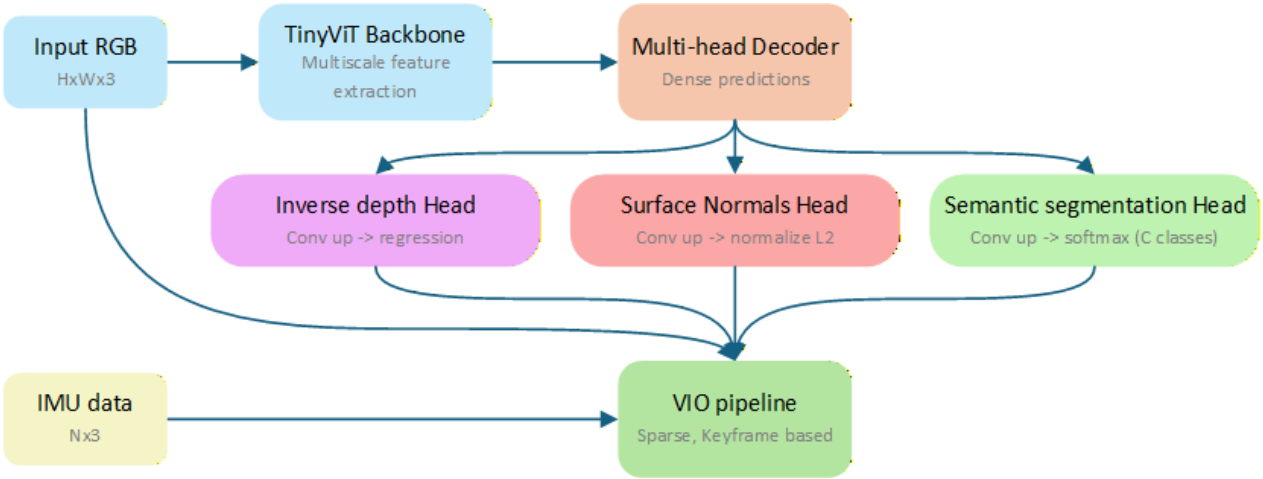


Fig. 1. High-level architecture of the proposed framework.

The decoder outputs three complementary modalities:

1) Inverse depth maps, providing scene geometry cues even in textureless regions.

2) Surface normals, enforcing local geometric consistency.

3) Semantic segmentation masks, facilitating the rejection of dynamic or reflective surfaces.

The predicted outputs are fused into the SLAM module to guide pose-graph optimization and improve convergence reliability.

3.2. TinyViT backbone and decoder design

The TinyViT backbone [1] was selected for its balance between representational capacity and computational efficiency, enabling deployment on resource-constrained UAV hardware. It processes monocular RGB images to extract hierarchical feature representations at multiple spatial resolutions.

The decoder, shown in Fig. 2, adopts a multi-task structure with three dedicated output branches for inverse depth, surface normals, and semantic segmentation. Each branch employs convolutional upsampling with skip connections from the corresponding TinyViT layers, thereby maintaining high-resolution spatial information while minimizing computational cost.

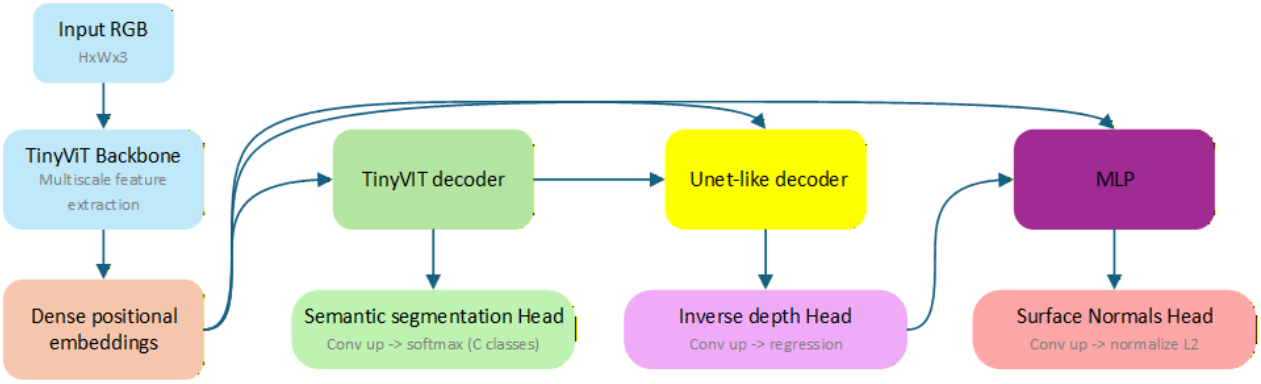


Fig. 2. Transformer-based network architecture.

This design enables joint optimization across complementary geometric and semantic cues, which has been shown to improve prediction stability and cross-task consistency in dense scene understanding.

3.3. Data and training

The model is trained on widely used 3D perception datasets that provide synchronized RGB, depth, normal, and semantic labels:

- NYUv2 [17]: indoor scenes with dense depth and semantic labels.
- ScanNet [18]: large-scale indoor scans with rich geometry and object classes.
- TUM-RGBD [19]: sequences suitable for evaluating SLAM performance in dynamic and cluttered environments.

To improve generalization, data augmentations include random cropping, photometric distortion, and horizontal flipping. The total loss function is defined as:

$$L = \lambda_d L_{depth} + \lambda_n L_{normals} + \lambda_s L_{segmentation}, \quad (1)$$

where L_{depth} is the L1 loss on inverse depth; $L_{normals}$ is the cosine distance between predicted and ground-truth normals; $L_{segmentation}$ is the standard cross-entropy loss; $\lambda_d, \lambda_n, \lambda_s$ are the weights (they are tuned empirically to balance geometric and semantic supervision).

3.4. Integration with SLAM

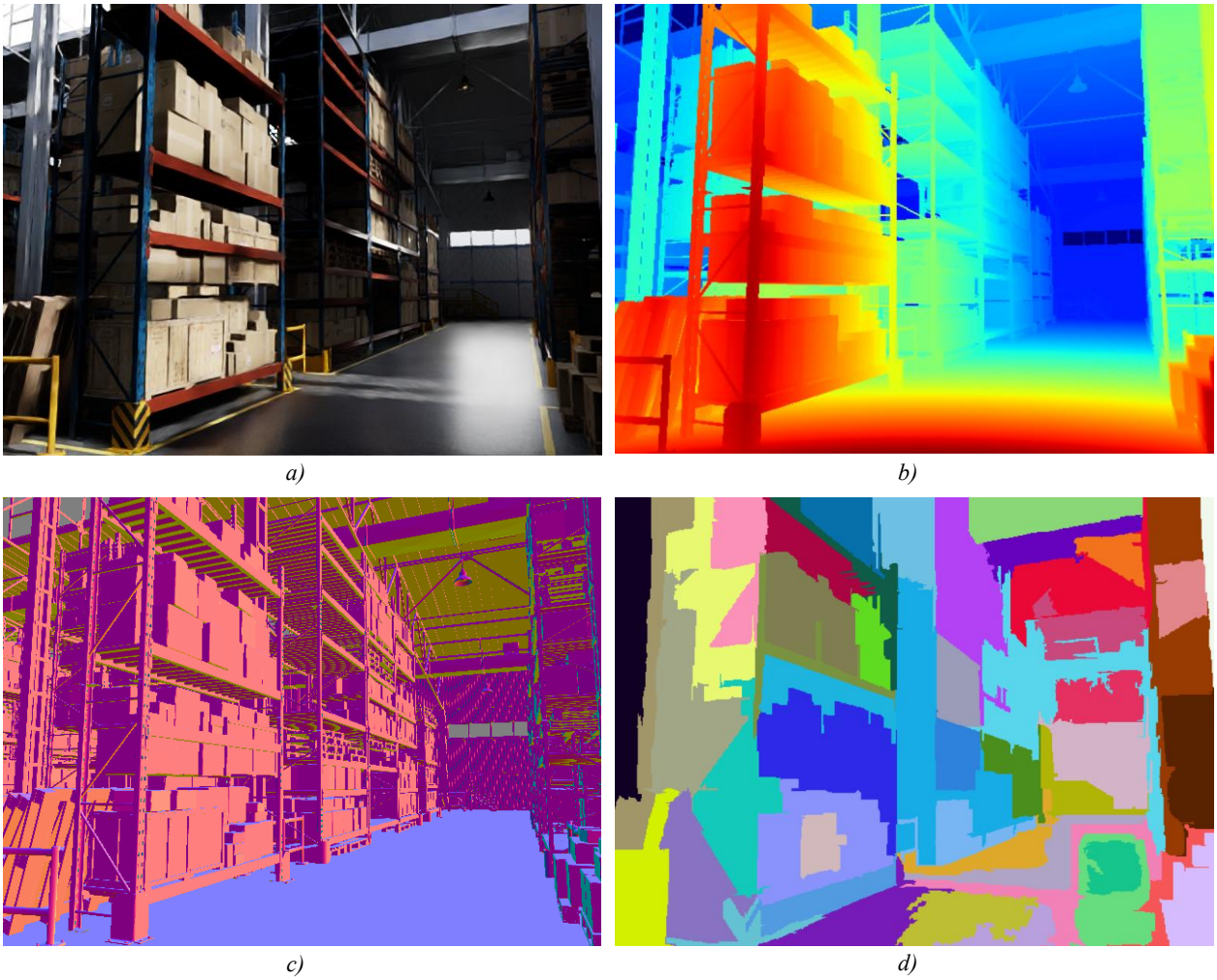
The integration between the transformer-based predictor and the geometric SLAM module is central to the proposed system. The predicted priors are incorporated at multiple stages of the SLAM pipeline:

1) Depth and normal priors are used to initialize 3D landmarks in each keyframe, providing improved geometric consistency for bundle adjustment.
2) Segmentation masks filter out features corresponding to reflective, transparent, or dynamic objects (e.g., machinery in motion or operator presence), preventing them from corrupting the optimization process.
3) Uncertainty weighting is introduced by leveraging per-pixel confidence maps from the network to modulate residuals in the optimization objective, effectively guiding the solver toward reliable measurements.
4) Pose refinement employs a modified cost function:

$$E = \sum_i w_i \|r_i(T)\|^2, \tag{2}$$

where w_i are confidence weights derived from the neural priors; $r_i(T)$ are photometric and geometric residuals.

By combining learned priors with geometric optimization, the hybrid system achieves improved convergence from poor initial conditions, enhanced map consistency, and increased robustness in the presence of environmental uncertainty. An illustration of the fused predictions-depth, normals, and segmentation is shown in Fig. 3, highlighting how geometric and semantic cues complement each other to produce reliable scene representations even under challenging lighting and cluttered conditions.



a) *b)* *c)* *d)*

Fig. 3. Example predictions from the multi-task decoder: (*a*) input RGB image, (*b*) inverse-depth map, (*c*) surface-normal map, and (*d*) semantic segmentation.

3.5. Computational considerations

The proposed architecture is designed with the computational limitations of embedded onboard systems in mind. In particular, the entire inference pipeline is capable of running on a Raspberry Pi 5 using CPU-only processing, without the need for any dedicated GPU accelerator. Despite the modest computational resources, the transformer-based predictor achieves an average throughput of approximately one frame per second.

Although this frame rate might appear low compared to high-performance computing setups, it remains entirely adequate for the target application. The model is invoked exclusively on keyframes, primarily during relocalization or loop-closure events within the SLAM pipeline – operations that typically occur at a temporal frequency close to one second. Consequently, the computational burden imposed by the neural predictor does not hinder real-time navigation or control.

The TinyViT encoder ensures efficient feature extraction through a compact hybrid convolution-attention design, while the multi-task decoder reuses shared feature hierarchies to minimize redundant computation across depth, normal and segmentation outputs. As a result, the overall system maintains a favorable trade-off between semantic richness, geometric reliability, and energy efficiency, enabling practical onboard deployment on resource-constrained aerial platforms.

4. Conclusion

This paper introduced a hybrid perception-localization framework designed for UAV navigation in unstructured industrial environments, where classical geometric SLAM systems often fail due to poor initialization, textureless regions, reflective surfaces or dynamic elements. By combining a lightweight transformer-based encoder (TinyViT) with a shared multi-task decoder predicting inverse depth, surface normals and semantic segmentation, the proposed method provides rich geometric and semantic priors that guide and stabilize subsequent keyframe-based SLAM optimization.

Unlike purely geometric pipelines, the presented system maintains reliable localization even under partial occlusions, illumination changes or non-Lambertian surfaces – conditions common in industrial halls, workshops or storage areas. The learned priors supply accurate depth and normal structure in regions where visual features are sparse, while segmentation enables selective rejection of unstable or dynamic regions during bundle adjustment. This synergy between learned perception and geometric optimization significantly enhances map consistency, pose convergence, and relocalization reliability.

Despite being executed on a low-power embedded platform (Raspberry Pi 5, CPU-only), the neural module achieves approximately 1 frame per second, which is fully sufficient given its role in keyframe-level correction rather than per-frame tracking. This makes the overall system practical for onboard deployment, satisfying real-time constraints while preserving robustness and interpretability.

In summary, the proposed hybrid approach demonstrates that combining transformer-based dense perception with traditional SLAM backends enables robust localization in highly unstructured, cluttered and noisy environments, bridging the gap between data-driven scene understanding and precise geometric mapping. This line of research opens the way toward fully autonomous UAV systems capable of long-term, stable operation in complex industrial scenarios without dependence on external infrastructure or high-end hardware.

References

[1] Pistun, Y. , Lesovoy, L. , Matiko, F., Fedoryshyn, R. (2014). Computer Aided Design of Differential Pressure Flow Meters. *World Journal of Engineering and Technology*, **2**, 68-77. doi: 10.4236/wjet.2014.22009.

[2] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly and J. Uszkoreit. (2020). An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale. *arXiv preprint* arXiv:2010.11929.

[3] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin and B. Guo. (2021). Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, Montreal, QC, Canada, Oct. 10-17 2021, pp. 10012-10022.

[4] Z. Teed and J. Deng. (2021). DROID-SLAM: Deep Visual SLAM for Monocular, Stereo, and RGB-D Cameras. *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 34/35. arXiv:2108.10869

[5] Z. Zhu, S. Peng, et al. (2022). NICE-SLAM: Neural Implicit Scalable Encoding for SLAM. *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. arXiv:2112.12130

[6] J. Czarnowski, T. Laidlow, R. Clark and A. J. Davison. (2020). DeepFactors: Real-Time Probabilistic Dense Monocular SLAM. *IEEE Robot. Autom. Lett.* arXiv:2001.05049

[7] X. Zhai, J. Wu, Y. Wang, K. Ye, S. Ruan, et al. (2021). Scaling Vision Transformers. *arXiv preprint* arXiv:2106.04560.

[8] Y. Chen, C.–F. Chen, Z. Dong, T. Wu, et al. (2021). CrossViT: Cross-Attention Multi-Scale Vision Transformer for Image Classification. *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*. https://doi.org/10.48550/arXiv.2103.14899

[9] X. Dong, J. Bao, D. Chen, W. Zhang, N. Yu, L. Yuan, D. Chen and B. Guo. (2021). CSWin Transformer: A General Vision Transformer Backbone with Cross-Shaped Windows. *arXiv preprint* arXiv:2107.00652.

[10] X. Zhang, Y. Tian, W. Huang, Q. Ye, L. Xie and Q. Tian. (2022). HiViT: Hierarchical Vision Transformer Meets Masked Image Modeling. *arXiv preprint* arXiv:2205.14949.

[11] A. Hassani and H. Shi. (2022). Dilated Neighborhood Attention Transformer. *arXiv preprint* arXiv:2209.15001.

[12] X. Yu, et al. (2023). Mix-ViT: Mixing Attentive Vision Transformer for Ultra-Fine-Grained Visual Classification. *Signal Processing*, vol. 215. https://doi.org/10.1016/j.patcog.2022.109131

[13] X. Bai, Z. Hu, X. Zhu, et al. (2022). TransFusion: Robust LiDAR-Camera Fusion for 3D Object Detection with Transformers. *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. arXiv:2203.11496

[14] Z. Li, W. Wang, E. Xie, et al. (2022). BEVFormer: Learning Bird's-Eye-View Representation from Multi-Camera Images via Spatio-Temporal Transformers. *Proc. European Conference on Computer Vision (ECCV)*. arXiv:2203.17270

[15] A. Kirillov, E. Mintun, N. Ravi, et al. (2023). Segment Anything (SAM). *arXiv preprint* arXiv:2304.02643.

[16] Y. Xiong, B. Varadarajan, et al. (2024). EfficientSAM: Leveraged Masked Image Pretraining for Efficient Segment Anything. *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. arXiv:2312.00863

[17] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus. (2012). NYU Depth Dataset V2. [Online]. Available: https://cs.nyu.edu/~silberman/datasets/nyu_depth_v2.html

[18] A. Dai, A. X. Chang, M. Savva, et al. (2017). ScanNet: Richly-annotated 3D Reconstructions of Indoor Scenes. [Online]. Available: http://www.scan-net.org/

[19] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers. (2012). A Benchmark for the Evaluation of RGB-D SLAM Systems. [Online]. Available: https://vision.in.tum.de/data/datasets/rgbd-dataset

Нейромережа трансформерного типу для робастного розуміння тривимірного промислового середовища в автономних системах БПЛА

Олексій Кучкін, Артем Сазонов, Ірина Черепанська, Анатолій Жученко

Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського», просп. Перемоги, 37, м. Київ, 03056, Україна

Анотація

Автономна навігація безпілотних літальних апаратів (БПЛА) в неструктурованих промислових середовищах залишається складним завданням через нерегулярну геометрію, динамічні перешкоди та невизначеність сенсорних даних. Класичні системи SLAM, попри геометричну узгодженість, часто виявляються нестійкими за умов поганої ініціалізації, відсутності текстури або наявності віддзеркалювальних поверхонь. Щоб подолати ці обмеження, у роботі запропоновано гібридний трансформерно-геометричний підхід, який поєднує навчальні апріорні уявлення сцени з ключовим SLAM-конвеєром. Енкодер TinyViT та легковаговий мультизадачний декодер спільно оцінюють зворотну глибину, нормалі поверхні та семантичну сегментацію, формуючи густі геометричні й семантичні підказки, що стабілізують локалізацію й побудову карти. Ці апріорні дані інтегруються в оптимізацію SLAM для прискорення збіжності, відкидання динамічних об'єктів та покращення релокалізації. Система працює майже в реальному часі (~1 FPS) на CPU Raspberry Pi 5, що робить її придатною для покадрового інференсу. Експерименти демонструють стійку локалізацію та консистентне картографування у захаращених, віддзеркалювальних і динамічних промислових сценах, підтверджуючи, що трансформерна густинна перцепція ефективно доповнює класичний SLAM для ресурсоощадної навігації БПЛА.

Ключові слова: робастне керування БПЛА; комп'ютерний зір; SLAM; нейронна мережа на основі трансформера; автономна навігація; тривимірне розуміння середовища.