Enhancing Monocular Depth Estimation with Self-Supervised Learning and Transformer-Based Architectures

Madhu Vanthi Sankar Ganesh
School of Computing
National College of Ireland
Dublin, Ireland
x23310928@student.ncirl.ie

Abstract

This paper presents a self-supervised learning framework for monocular depth estimation in dynamic scenes using transformers. Accurate depth estimation in dynamic regions remains challenging due to spatial-temporal inconsistency and scene clutter. While recent approaches have made progress by separating static and dynamic regions and understanding object-level motion cues, architectures often rely on convolutional neural networks, which can limit the ability to capture long-term dependencies. This paper proposes a transformer- based architecture that models well for both static and dynamic scenes using global attention mechanisms. This approach starts off with an object-centric module for estimating the depth of moving regions under a rigid motion assumption, along with a scale alignment strategy to resolve scale inconsistencies across regions. The pseudo-labels generated from this step are then used for self-training a depth estimation network in an end-to-end manner. Experiments conducted on Cityscapes dataset demonstrate that this method proves to be better than other CNN-based counterparts in terms of performance, particularly in dynamic environments, highlighting the effectiveness of transformer-based representations for this task.

Index Terms— Monocular depth estimation, Self-supervised learning, Transformers, Computer vision, Deep learning, Convolutional Neural Networks.

I. Introduction

Monocular depth estimation is an emerging computer vision task with applications in autonomous driving, robotics and augmented reality. The major challenge in understanding three-dimensional scene geometry from a single RGB image has been extensively studied and major advancements have been achieved with the help of various deep learning methodologies Aleotti et al. (2018); Pilzer et al. (2018); Xue et al. (2020).

While Supervised methods achieve high accuracy when a large amount of annotated data is available, the cost of generating said annotated data is too high and not always practical Liu et al. (2021). For this reason, there has been an increased interest in self-supervised approaches in various applications, including monocular depth estimation Aleotti et al. (2018); Pilzer et al. (2018); Wang et al. (2023); Zheng et al. (2023); Jia et al. (2022).One of the recent advances in this industries is a self supervised architecture Monodepth2 Godard et al. (2019), which overcomes the data limitation by assuming static scenes and consistent camera motion and use photometric losses to learn depth and pose jointly. However, recent works have shown that handling dynamic objects explicitly Nguyen et al. (2024) can significantly improve performance in real world scenarios.

Just as self-supervised learning, transformers-based architectures have revolutionized computer vision through their superior attention mechanisms and context modeling capabilities in the global context performing even better than convolutional neural networks Han et al. (2022); Verma et al. (2025); Lv et al. (2022); Wu et al. (2021); Sowmya et al. (2023); Agarwal and Arora (2022). Several papers have also integrated transformers into depth estimation pipelines. DPT shows the effectiveness of vision transformers for dense prediction, while DepthFromer shows that transformer-based encoders can capture long-range dependencies important for accurate depth estimation Harish et al. (2025). The combination of transformers with self-supervised learning presents unique opportunities. The integration of transformers, especially its attention mechanisms, is advantageous for modeling spatial relationships and object boundaries Han et al. (2022), while including self-supervised training provides the scalability needed for large-scale deployment and addresses issues such as cost in generating accurate labels Liu et al. (2021). However current approaches only focus on either self-supervised learning in depth estimation with traditional CNN architecture Wang et al. (2023); Zheng et al. (2023); Jia et al. (2022); Elazab et al. (2025); Nguyen et al. (2024) or transformer-based models Agarwal and Arora (2022); Harish et al. (2025) for depth estimation with limited exploration of their combination.

This paper presents a multi-stage self-supervised learning framework that combines transformer-enhanced architecture with explicit object motion modeling. This approach leverages the global attention capabilities of transformers while at the same time maintains the advantages of self-supervised learning in terms of scalability and ability to understand features. The framework is

divided into 4 main interconnected stages, starting with a pixel-wise depth training using a DiffNet architecture that incorporates a transformer-based model into traditional CNN backbone.

The key contributions include: (1) A novel integration of a transformer-based model within a self-supervised depth estimation framework that handles dynamic objects explicitly. (2) A multi-stage pipeline from pixel depth training to depth scale alignment and a final training module with weights from all the previous stages. This architecture demonstrates significant improvements in depth accuracy and temporal consistency, particularly for dynamic urban scenes.

The paper follows a structure which starts with a related work , moves on to the methodology that goes into the architecture in detail and finally presents the results from the model training and future directions.

II. Literature Review

A. Self Supervised Learning

Self supervised learning has gained significant traction in machine learning applications such as computer vision, natural language processing and graph learning Liu et al. (2021); Gui et al. (2024); Zhang et al. (2024) . SSL leverages the data itself as a supervision, and has proven its capability due to its superior performance in learning representations. SSL can be broadly classified into Contrastive, Generative and Generative-contrastive or adversarial based on their objectives. Contrastive learning aims to learn representations by pulling similar instances closer and pushing dissimilar instances farther apart while generative learning aims to train an encoder to reconstruct the input to learn the representations.Because of Generative SSLs ability at recovering the original data distribution without making assumptions, it makes a widely applicable method of training for monocular depth estimation.Various methods within generative SSLs were explored for this task Aleotti et al. (2018); Pilzer et al. (2018).Yang et al. (2020) aims to enhance the depth estimation using generative SSL by normalizing lighting conditions across input images as well as accounts for photometric uncertainties at the pixel level.

There are various papers that discuss self supervised learning in the realm of depth estimation and one of them introduces PlaneDepth. Unlike earlier methods that use fronto-parallel planes, PlaneDepth uses a mixture of Laplacian distributions based on orthogonal planes which include vertical as well as ground planes.Wang et al. (2023) PlaneDepth further enhances the performance by addressing data augmentation transformations, such as resizing and cropping as well as incorporates a self distillation loss, enhanced with occlusion aware masks, resulting in improved accuracy in depth predictions even in visually complex environments.This paper incorporates this self supervised learning technique for monocular depth estimation.

B. Depth Estimation

Processing data from event cameras for depth estimation is very popular in various applications including robotics, augmented reality and autonomous driving.Various papers introduce several architectures for depth estimation for each application Wu et al. (2024); Zheng et al. (2023); Jia et al. (2022); Elazab et al. (2025). In the domain of autonomous driving,Xue et al. (2020) proposes a self-supervised monocular depth estimation model for calculating absolute depth, which is in case is the distance between the camera and other obstacles on the road.The model is tested on KITTI dataset and shows better performance as compared to existing models, yet it assumes that the ground is a flat,static surface. This assumption may not hold true in real world scenarios, resulting in the model making inaccurate estimations of depth. Hence, Nguyen et al. (2024) introduces a self-supervised monocular depth estimation architecture, specifically designed for dynamic scenes. This method first starts with a scene depth estimation, specifically for static regions and then performs dynamic object depth estimation and moves on to depth scale alignment and finally trains the model for depth prediction. Self-supervised learning is performed on all steps except for depth scale alignment module training.

C. Transformers

Papers like Han et al. (2022) highlight the applications of vision transformers in various fields and its advantages over the traditional convolutional neural networks across various visual benchmarks. It also indicates their potential to complement existing CNN architectures. Verma et al. (2025) aims to improve the performance of ViT without relying on CNNs in small datasets, but the proposed architecture works exclusively on smaller datasets, and the performance diminishes as the dataset size increases. The combination of CNNs and transformers can lead to improved performance by leveraging global and local connections as proven in Lv et al. (2022). This paper introduces a novel transformer-based survival analysis model designed

for colorectal cancer prognosis. It leverages the complementary strengths of both CNN and transformers, and the results show that this combination performs better than the individual counterparts.

Various papers Wu et al. (2021); Sowmya et al. (2023); Agarwal and Arora (2022) prove the efficiency of this combination and introduce various combinations of CNN and transformers for their respective field of research. For example, in the case of Tanvir et al. (2024), which introduces a architecture that integrates ResNet50v2 and Vision transformer to enhance handwritten character recognition, leverages the strengths of both models to gain understanding of input images across various scales and abstraction levels.Harish et al. (2025) uses transformer based model for Monocular depth estimation, this paper also sees the incorporation of ResNet and Vision Transformers and shows better results than other existing architectures but could be enhanced by using self-supervised learning.

III. Methodology

A. Transformer-Enhanced DiffNet Architecture

The main aspect of the depth estimation pipeline is built upon a DiffNet transformer-based architecture, which integrates spatial attention mechanisms into a ResNet-based encoder-decoder framework. The core of the model uses a DiffNet Resnet Encoder with 18 layers and incorporates a HRNet multi-scale feature extraction which is enhanced with transformer blocks Wang et al. (2020). The encoder is for processing input RGB images through a hierarchical feature pyramid, enabling multi-resolution spatial understanding.

Transformers replace standard ResNet bottleneck blocks in deeper layers for enhancement. To capture long-range spatial dependencies for accurate depth estimation, these blocks incorporate multi-head spatial attention with 8 attention heads Vaswani et al. (2017). The model is able to reason with geometric relationships throughout the entire image while maintaining computational efficiency as the attention mechanism operates on flattened spatial features with positional encoding.

The HRNet based deconder processes multi-scale features to generate depth predictions at four different scales, allowing coarse-to-fine depth estimation.The decoder is designed to maintain high-resolution features throughout the decoding process which is essential for preserving finer details in complex urban scenes Harish et al. (2025).

This architecture, due to its transformer-based enhancement, ensures consistent spatial reasoning capabilities throughout the training process.

B. Stage 1: Pixel-wise Scene Depth Learning with Self Supervised Training

1) Network Architecture and Training Paradigm: The foundation of the depth estimation capability is through a comprehensive self-supervised learning framework which is established in stage 1. There are three interconnected networks in this training stage: the DiffNet transformer-enhanced depth network, a pose estimation network and the RAFT optical flow network Teed and Deng (2020). Both the depth network and the pose network employ transformer-based architecture for camera motion estimation.

There are multiple steps within this stage, the first step is training both the depth and pose networks jointly, establishing basic scene understanding and camera motion estimation capabilities. Then the data is trained on the RAFT optical flow network and finally all three networks are optimized simultaneously. This strategy is to ensure smooth integration of optical flow constraints with depth and pose estimation Nguyen et al. (2024).

2) Self-Supervised training: The self-supervised training aspect of this stage depends on photometric consistency across temporal sequences, implementing a loss function that minimizes reprojection errors between consecutive frames using the predicted depth and camera poses.

This model also integrates a Structural Similarity Index (SSIM) loss that preserves quality by maintaining local structure consistency.

3) Temporal Consistency and Smoothness Regularization: Penalizing inconsistent depth estimates for static scene elements done by temporal consistency constraints ensure coherent depth predictions across video sequences. This model implements

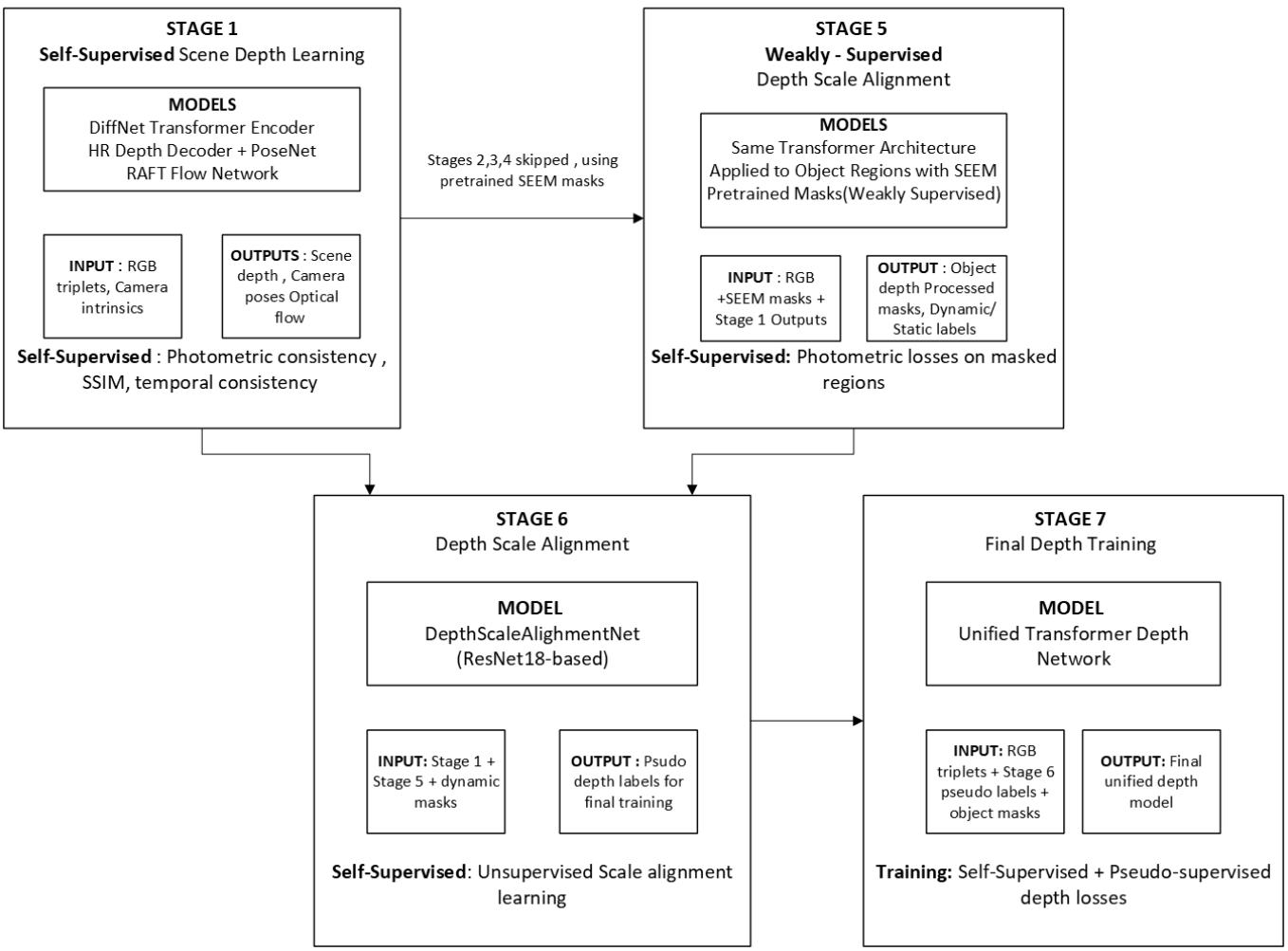

Fig. 1. Self-Supervised Depth Estimation Pipeline with Transformer Architecture

depth smoothness regularization that is also edge-aware to preserve depth irregularities at object boundaries while encouraging smooth depth transitions within homogeneous regions Mahjourian et al. (2018).

C. Stage 5 : Object-Specific Depth Learning with Weakly Supervised Masks

1) SEEM Mask Integration and Architectural Consistency: In stage 5, a similar transformer model from stage 1 is applied specifically to object regions identified by pretrained SEEM masks Nguyen et al. (2024). The architecture is a combination of both weakly supervised learning of object segmentation masks with self-supervised depth estimation. This is to ensure that object-specific depth learning benefits from the same spatial attention mechanisms that proved effective for scene-level depth estimation.

SEEM masks are used for precise object boundaries without requiring manual annotation. This enables the system to focus computational resources on object regions while still maintaining the self-supervised learning aspect of the architecture.

2) Dynamic and Static Object Classification: One of the major components of this stage is a classification module that distinguishes between moving and stationary objects using temporal consistency analysis. The classification employs a threshold-based approach where objects showing motion above 1.0 pixels are classified as dynamic and the rest with minimal motion are classified as static.

This classification is in turn used for the next stage which is depth scale alignment as static and dynamic objects require different approaches in scale estimation. Dynamic objects need specialized handling for their motion induced depth changes Luo et al. (2019).

D. Stage 6 : Depth Scale Alignment with Unsupervised Learning

The training in this stage is to address the scale ambiguity, which is prevalent in self-supervised depth estimation, where networks can only predict relative depth relationships to an unknown scale factor Nguyen et al. (2024). This module uses ResNet based architecture with group normalization layers, which is specifically designed to learn scale alignment between scene depth and object depth without labels.

This stage takes RGB images, object masks and static depth from stage 1, dynamic depth from stage 5 as inputs, to predict per-object scale factors. This makes the final system robust in scale estimation as it leverages complementary information sources.

1) Unsupervised Scale Learning Framework: The unsupervised aspect of this stage exploits the concept that static scene elements and objects always maintain consistent depth relationships. The network aims to learn to minimize depth inconsistencies between overlapping regions while preserving geometric integrity of both scene and depth predictions by employing Huber loss for robust scale regression. This loss function is robust due to its ability to work with outliers while maintaining gradient stability during optimization.

2) Pseudo Label Generation: Once the training is completed , the system generates pseudo depth labels for both the training and validation data by applying scale factors that combine scene and object depth predictions. These labels are used as weakly supervised targets for the final unified training for depth estimation bridging the gap between supervised and self supervised learning in terms of results.

E. Stage 7 : Final Unified Depth Network Training

1) End-to-End Architecture and Curriculum Learning: The final stage is end-to-end training of a unified transformer depth network that integrates scene understanding, object awareness and scale consistency into a single depth estimation model. The training uses the same DiffNet transformer, while incorporating knowledge and results from all previous stages.The compressed training process involves the usage of initialization from stage 1 and pseudo-supervised targets generated in stage 6 enabling rapid convergence and optimal performance. Learning rate is reduced in regular intervals as the training goes on to prevent overfitting to the pseudo-supervised targets Smith (2017).

2) Multi Model Loss Function Integration: The training uses a loss combination that integrates multiple supervision signals, namely the self-supervised photometric losses in stage 1 and the pseudo-supervised depth losses from stage 6, providing depth supervision while at the same time maintaining the self-supervised training.

3) Motion Sparsity Regularization and Geometric Consistency: The model includes motion sparsity regularization to facilitate sparse motion fields consistent with geometric scene assumptions. This implementation encourages minimal motion assumptions as well as accommodating necessary object motion. This maintains geometric consistency across the unified depth estimation.

The final model enables real-time inference while maintaining the advantages of transformer attention mechanisms. This model is the amalgamation of multiple supervised signals within a coherent learning framework. This ensures that the final model achieves state-of-the-art performance effectively combining scene-level understanding, object specific reasoning and scale-consistent depth estimation Nguyen et al. (2024).

IV. Experiments and Results

The evaluation of the proposed approach is conducted on the CityScapes dataset with 6849 images for training and 796 images for testing.To enable a more fine grained analysis, the training and evaluation includes the moving object annotations provided by Nguyen et al. (2024) allowing the evaluation of model performance on static and dynamic regions separately. This training framework is particularly important as depth estimation in dynamic regions remain a challenging problem in autonomous driving scenarios.

For the motion estimation, the RAFT framework is employed to provide dense, pixel-wise motion fields.At the same time, the DSA modules are integrated with a ResNet18 backbone to enhance spatial feature representation.In this setup, two depth estimation architectures are investigated. The first is DiffNet which follows a conventional CNN-based design, while the second Transformer-DiffNet which extends the current DiffNet architecture by incorporating transformer layers, to capture long-range dependencies and contextual information.

Both models are trained for exactly 10 epochs under the same experimental settings to ensure fair comparison. The results for both the models is stored in Table 1 for comparison. DiffNet demonstrates stronger performance across the overall and static regions, highlighting the effectiveness of CNN-based local feature extraction for stable environments. The improvement by using transformer based architecture is shown in the dynamic regions, which consistently outperforms DiffNet, due to the advantage of transformers in modeling motion and capturing contextual cues in scenes with moving objects.

The primary objective of this study is to improve the depth estimation in dynamic areas using transformers, the performance of the proposed model proves the objective right. Although DiffNet maintains slightly better results in static regions, the superior performance of Transformer-DiffNet on dynamic objects makes it the more efficient architecture for the targeted problem domain. The findings of this paper is in support of the adoption of transformer-based approaches as a promising direction for advancing depth estimation in complex, real-world scenarios.

TABLE I
Comparison of DiffNet and Transformer_DiffNet Performance Across Regions

Region / Metric	**abs_rel**	**sq_rel**	**rmse**	**rmse_log**	**a1**	**a2**	**a3**
All Region							
DiffNet	0.282	**2.538**	**8.940**	**0.325**	0.539	**0.820**	**0.954**
Transformer_DiffNet	**0.269**	2.817	10.349	0.341	**0.549**	0.815	0.941
Static Region							
DiffNet	0.276	**2.473**	**9.005**	**0.323**	0.541	**0.822**	**0.956**
Transformer_DiffNet	**0.267**	2.834	10.502	0.342	**0.548**	0.814	0.940
Dynamic Region							
DiffNet	0.339	3.651	7.741	0.323	0.534	0.820	0.918
Transformer_DiffNet	**0.285**	**2.906**	**7.217**	**0.294**	**0.576**	**0.846**	**0.937**

V. Conclusion

This paper presents a novel fully self-supervised pipeline with transformers for depth estimation in dynamic regions.Transformers have already proved to perform better in global context compared to CNN based models and the comparison between DiffNet model which is CNN based and Transformer-DiffNet which is a combination of CNN and Transformer based architecture proves the same. This transformer model outperforms previous works in depth predictions for dynamic regions on Cityscapes dataset. The one limitation in this method is the high computational requirements that come from using transformer based models.

REFERENCES

Agarwal, A. and Arora, C. (2022), Depthformer: Multiscale vision transformer for monocular depth estimation with global local information fusion, *in* '2022 IEEE international conference on image processing (ICIP)', IEEE, pp. 3873–3877.

Aleotti, F., Tosi, F., Poggi, M. and Mattoccia, S. (2018), Generative adversarial networks for unsupervised monocular depth prediction, *in* 'Proceedings of the European conference on computer vision (ECCV) workshops', pp. 0–0.

Elazab, G., Gräber, T., Unterreiner, M. and Hellwich, O. (2025), Monopp: Metric-scaled self-supervised monocular depth estimation by planar-parallax geometry in automotive applications, *in* '2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)', IEEE, pp. 2777–2787.

Godard, C., Mac Aodha, O., Firman, M. and Brostow, G. J. (2019), Digging into self-supervised monocular depth estimation, *in* 'Proceedings of the IEEE/CVF international conference on computer vision', pp. 3828–3838.

Gui, J., Chen, T., Zhang, J., Cao, Q., Sun, Z., Luo, H. and Tao, D. (2024), 'A survey on self-supervised learning: Algorithms, applications, and future trends', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **46**(12), 9052–9071.

Han, K., Wang, Y., Chen, H., Chen, X., Guo, J., Liu, Z., Tang, Y., Xiao, A., Xu, C., Xu, Y. et al. (2022), 'A survey on vision transformer', *IEEE transactions on pattern analysis and machine intelligence* **45**(1), 87–110.

Harish, M., Balaji, V., Dannie Jerome, E., Amutha Yalini, E. et al. (2025), Transformer-based approach for monocular depth estimation using dpt and depthformer, *in* '2025 3rd International Conference on Artificial Intelligence and Machine Learning Applications Theme: Healthcare and Internet of Things (AIMLA)', IEEE, pp. 1–5.

Jia, S., Pei, X., Yao, W. and Wong, S. C. (2022), 'Self-supervised depth estimation leveraging global perception and geometric smoothness', *IEEE Transactions on Intelligent Transportation Systems* **24**(2), 1502–1517.

Liu, X., Zhang, F., Hou, Z., Mian, L., Wang, Z., Zhang, J. and Tang, J. (2021), 'Self-supervised learning: Generative or contrastive', *IEEE transactions on knowledge and data engineering* **35**(1), 857–876.

Luo, C., Yang, Z., Wang, P., Wang, Y., Xu, W., Nevatia, R. and Yuille, A. (2019), 'Every pixel counts++: Joint learning of geometry and motion with 3d holistic understanding', *IEEE transactions on pattern analysis and machine intelligence* **42**(10), 2624–2641.

Lv, Z., Lin, Y., Yan, R., Wang, Y. and Zhang, F. (2022), 'Transsurv: transformer-based survival analysis model integrating histopathological images and genomic data for colorectal cancer', *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **20**(6), 3411–3420.

Mahjourian, R., Wicke, M. and Angelova, A. (2018), Unsupervised learning of depth and ego-motion from monocular video using 3d geometric constraints, *in* 'Proceedings of the IEEE conference on computer vision and pattern recognition', pp. 5667–5675.

Nguyen, H. C., Wang, T., Alvarez, J. M. and Liu, M. (2024), Mining supervision for dynamic regions in self-supervised monocular depth estimation, *in* 'Proceedings of the IEEE/CVF conference on computer vision and pattern recognition', pp. 10446–10455.

Pilzer, A., Xu, D., Puscas, M., Ricci, E. and Sebe, N. (2018), Unsupervised adversarial depth estimation using cycled generative networks, *in* '2018 international conference on 3D vision (3DV)', IEEE, pp. 587–595.

Smith, L. N. (2017), Cyclical learning rates for training neural networks, *in* '2017 IEEE winter conference on applications of computer vision (WACV)', IEEE, pp. 464–472.

Sowmya, T. S., Narasimhulu, T., Sunitha, G., Manikanta, T. and Venkatesh, T. (2023), Vision transformer based resnet model for pneumonia prediction, *in* '2023 4th International Conference on Electronics and Sustainable Communication Systems (ICESC)', IEEE, pp. 316–321.

Tanvir, K., Al Huda, M. S., Kabir, M. S., Obayed, R., Attef, M. F. and Ali, M. A. (2024), Resvit: An integrated approach using resnet50v2 and vision transformer for enhanced bangla handwritten character recognition, *in* '2024 2nd International Conference on Information and Communication Technology (ICICT)', IEEE, pp. 170–174.

Teed, Z. and Deng, J. (2020), Raft: Recurrent all-pairs field transforms for optical flow, *in* 'European conference on computer vision', Springer, pp. 402–419.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. and Polosukhin, I. (2017), 'Attention is all you need', *Advances in neural information processing systems* **30**.

Verma, A., Dubey, S. R. and Singh, S. K. (2025), Patch attention excitation based vision transformer for small-sized datasets, *in* 'ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)', IEEE, pp. 1–5.

Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X. et al. (2020), 'Deep high-resolution representation learning for visual recognition', *IEEE transactions on pattern analysis and machine intelligence* **43**(10), 3349–3364.

Wang, R., Yu, Z. and Gao, S. (2023), Planedepth: Self-supervised depth estimation via orthogonal planes, *in* 'Proceedings of the IEEE/CVF conference on computer vision and pattern recognition', pp. 21425–21434.

Wu, H., Xiao, B., Codella, N., Liu, M., Dai, X., Yuan, L. and Zhang, L. (2021), Cvt: Introducing convolutions to vision

transformers, *in* 'Proceedings of the IEEE/CVF international conference on computer vision', pp. 22–31.

Wu, X., He, W., Yao, M., Zhang, Z., Wang, Y., Xu, B. and Li, G. (2024), 'Event-based depth prediction with deep spiking neural network', *IEEE Transactions on Cognitive and Developmental Systems* .

Xue, F., Zhuo, G., Huang, Z., Fu, W., Wu, Z. and Ang, M. H. (2020), Toward hierarchical self-supervised monocular absolute depth estimation for autonomous driving applications, *in* '2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)', IEEE, pp. 2330–2337.

Yang, N., Stumberg, L. v., Wang, R. and Cremers, D. (2020), D3vo: Deep depth, deep pose and deep uncertainty for monocular visual odometry, *in* 'Proceedings of the IEEE/CVF conference on computer vision and pattern recognition', pp. 1281–1292.

Zhang, K., Wen, Q., Zhang, C., Cai, R., Jin, M., Liu, Y., Zhang, J. Y., Liang, Y., Pang, G., Song, D. et al. (2024), 'Self-supervised learning for time series analysis: Taxonomy, progress, and prospects', *IEEE transactions on pattern analysis and machine intelligence* **46**(10), 6775–6794.

Zheng, M., Luo, L., Zheng, H., Ye, Z. and Su, Z. (2023), 'A dual encoder–decoder network for self-supervised monocular depth estimation', *IEEE Sensors Journal* **23**(17), 19747–19756.