STViT: Improving Self-Supervised Multi-Camera Depth Estimation with Spatial-Temporal Context and Adversarial Geometry Regularization (Student Abstract)

Zhuo Chen[1]*, Haimei Zhao[2]*, Bo Yuan[3], Xiu Li[1]
[1]Shenzhen International Graduate School, Tsinghua University
[2]University of Sydney
[3]University of Queensland
z-chen17@mails.tsinghua.edu.cn, hzha7798@uni.sydney.edu.au, boyuan@ieee.org, li.xiu@sz.tsinghua.edu.cn

Abstract

Multi-camera depth estimation has recently garnered significant attention due to its substantial practical implications in the realm of autonomous driving. In this paper, we delve into the task of self-supervised multi-camera depth estimation and propose an innovative framework, STViT, featuring several noteworthy enhancements: 1) we propose a Spatial-Temporal Transformer to comprehensively exploit both local connectivity and the global context of image features, meanwhile learning enriched spatial-temporal cross-view correlations to recover 3D geometry. 2) to alleviate the severe effect of adverse conditions, e.g., rainy weather and nighttime driving, we introduce a GAN-based Adversarial Geometry Regularization Module (AGR) to further constrain the depth estimation with unpaired normal-condition depth maps and prevent the model from being incorrectly trained. Experiments on challenging autonomous driving datasets Nuscenes and DDAD show that our method achieves state-of-the-art performance.

Introduction

In the realm of self-supervised depth estimation, techniques leveraging photometric consistency across consecutive frames have achieved considerable success. Recently, the self-supervised depth estimation paradigm has been extended to multi-camera settings, due to the need for autonomous driving. However, this extension is not trivial and poses unique challenges. Self-supervised methods rely heavily on co-visible regions among frames to recover 3D geometry and compute reprojection errors. However, in large-scale autonomous driving datasets like NuScenes and DDAD, challenges arise from very small overlaps between adjacent cameras and diverse weather and illumination conditions. To address these challenges, we propose a Spatial-Temporal Transformer that comprehensively exploits both local and global context features of images while leveraging cross-camera and cross-frame geometric correlations using cross-attention layers. This approach maximizes the utilization of co-visible regions, improving feature matching and network training. To handle challenging scenarios like rainy and night conditions, we introduce a Generative Adversarial Network-based geometry regularization module to regularize the prediction weirdness and mitigate the side effects.

*These authors contributed equally.

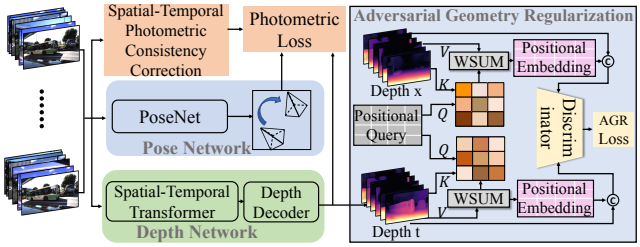

Figure 1: Overview of our STViT framework.

Method

Proposed Framework. Our STViT is composed of a Depth Network, a Pose Network, and an Adversarial Geometry Regularization (AGR) module. The Depth Network consists of a Spatial-Temporal Transformer Encoder and a Depth Decoder. The Pose Network is implemented by a lightweight ResNet. The Depth Network and Pose Network are jointly optimized via the minimization of Spatial-Temporal Photometric Loss. After predicted depth maps are obtained, they are further regularized and refined in the AGR module.

Spatial-Temporal Transformer. We propose enhancements to the encoder architecture and build a Spatial-Temporal Transformer. It not only leverages the Transformer's ability to model long-range dependencies, overcoming the locality issue in feature extraction seen in previous works (Godard and et al 2019; Wei and et al 2023), but also introduces Spatial-Temporal Cross-Correlation to fully exploit the co-visibility regions across cameras and temporal frames for geometric structure recovery. Taking inspiration from a recent Transformer model (Lee and et al 2022), we construct our Depth Encoder by introducing a Multi-Path Transformer Block to capture both local and global context within images simultaneously. As shown in Figure 2, It consists of a Conv-Stem and L Spatial-Temporal Transformer Layers. Each Transformer layer contains Multi-Scale Patch Embedding, Transformer Blocks, Convolutional Block, Global-to-Local Feature Interaction, and Spatial-Temporal Cross Correlation Module.

Spatial-Temporal Cross Correlation (STCC). Al-

Methods	Resolution	Abs Rel ↓	Sq Rel ↓	RMSE ↓	RMSE log ↓	Resolution	Abs Rel ↓	Sq Rel ↓	RMSE ↓	RMSE log ↓
Monodepth2 (Godard and et al 2019)	352 × 640	0.287	3.349	7.184	0.345	384 × 640	0.217	3.641	12.962	0.323
FSM* (Guizilini and et al 2022)	352 × 640	0.334	2.845	7.786	0.406	384 × 640	0.200	3.392	12.270	0.301
SurroundDepth (Wei and et al 2023)	352 × 640	0.245	3.067	6.835	0.321	384 × 640	0.200	3.392	12.270	0.301
EGA-Depth (Shi and et al 2023)	352 × 640	<u>0.239</u>	**2.357**	<u>6.801</u>	<u>0.936</u>	384 × 640	0.195	3.211	**12.117**	<u>0.297</u>
Ours	352 × 640	**0.233**	<u>2.815</u>	**6.681**	**0.312**	384 × 640	**0.192**	**2.965**	<u>12.156</u>	**0.293**

Table 1: Quantitative evaluation of self-supervised multi-camera depth estimation on nuScenes (left part) and DDAD (right part). FSM* denotes reproduced results. The best results are highlighted in bold and the second-best ones are underlined.

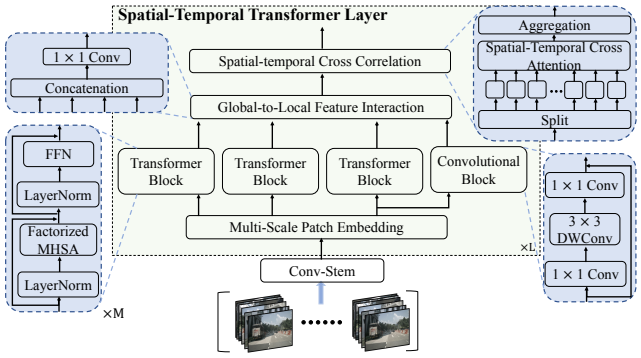

Figure 2: The architecture of Depth Encoder.

though we can effectively acquire the image feature, the cross-view correlation among different cameras and different temporal frames is still not exploited. Thus, we introduce an STCC Module to facilitate correlation learning and geometry recovery. As shown in Figure 2, the interacted features are first split into different cameras and different temporal frames. For each feature, we pre-define the list of views that share overlapping regions. The overlapped views contain adjacent cameras at the same timestamp, adjacent temporal frames of the same camera, and simultaneously cross-camera and cross-frame views as well (e.g., the front image at timestamps t and the left-front image at $t+1$). Thus, STCC can learn enriched spatial-temporal cross-view correlations for accurately inferring 3D geometry.

Adversarial Geometry Regularization. In real-world outdoor driving scenes, adverse conditions such as rainy weather and nighttime driving are frequently encountered, which significantly affects network learning and estimation performance. Thus, we propose a GAN-based Adversarial Geometry Regularization (AGR) module to further constrain the depth estimation, as shown in Figure 1. Specifically, we consider the Depth Network as a generator to provide depth map predictions. And adopt the depth predictions of an arbitrary normal-condition frame as a reference to regularize the depth distribution. It is observed the depth value distribution has a close relationship with the pixel positions in prior research. Thus, we use the positional query to scan over the depth map which serves as key and value. So that we can obtain the positional embedding e by calculating the dot product similarity between the query and keys. After that, the positional embedding is concatenated with the normalized predicted depth maps, denoted as $[e, \mu(D)]$. Similarly, the arbitrary depth maps D^R are also concatenated with the corresponding positional embedding as the regularization $[e^R, \mu(D^R)]$. A discriminator is used to distinguish $[e, \mu(D)]$ and $[e^R, \mu(D^R)]$, while the depth network tries to make predictions indistinguishable with regularization references. The optimization item L_{AGR} is the GAN loss.

Self-supervised Training Loss. The final training loss consists of the typical photometric loss ℓ_p and smoothing loss ℓ_{sm} from self-supervised monocular methods (Godard and et al 2019) and addtional AGR regularization loss ℓ_{AGR}:

$$Loss = \ell_p + 10^{-3}\ell_{sm} + 5 \times 10^{-4}\ell_{AGR} \quad (1)$$

Results and Conclusion

The evaluation metrics for multi-camera depth estimation are the same as its monocular counterpart. Four error metrics: **Abs Rel** for Absolute Relative Error, **Sq Rel** for Square Relative Error, **RMSE** for Root Mean Square Error, **RMSE log** for Root Mean Square Logarithmic Error and three accuracy metrics are included. We evaluate STViT using large-scale benchmarks Nuscene and DDAD, as shown in Table 1. Compared with existing methods, our method achieves state-of-the-art performance. Overall, we introduce a new self-supervised multi-camera depth estimation framework in this paper, with the proposed Spatial-Temporal Transformer to comprehensively exploit image features and spatial-temporal cross-view correlation and a GAN-based Adversarial Geometry Regularization to regularize the side effects of adverse conditions for training.

Acknowledgments

This research was partly supported by Shenzhen Key Laboratory of next generation interactive media innovative technology (Grant No: ZDSYS20210623092001004).

References

Godard, C.; and et al. 2019. Digging into self-supervised monocular depth estimation. In *ICCV*.

Guizilini, V.; and et al. 2022. Full surround monodepth from multiple cameras. *RAL*, 7(2): 5397–5404.

Lee, Y.; and et al. 2022. Mpvit: Multi-path vision transformer for dense prediction. In *CVPR*.

Shi, Y.; and et al. 2023. EGA-Depth: Efficient Guided Attention for Self-Supervised Multi-Camera Depth Estimation. In *CVPRW*.

Wei, Y.; and et al. 2023. Surrounddepth: Entangling surrounding views for self-supervised multi-camera depth estimation. In *CoRL*.