Applied Intelligence (2025) 55:328
https://doi.org/10.1007/s10489-024-06191-6

STViT+: improving self-supervised multi-camera depth estimation with spatial-temporal context and adversarial geometry regularization

Zhuo Chen[1] · Haimei Zhao[2] · Xiaoshuai Hao[3] · Bo Yuan[4] · Xiu Li[1]

Accepted: 11 December 2024

Abstract

Multi-camera depth estimation has gained significant attention in autonomous driving due to its importance in perceiving complex environments. However, extending monocular self-supervised methods to multi-camera setups introduces unique challenges that existing techniques often fail to address. In this paper, we propose **STViT+**, a novel Transformer-based framework for self-supervised multi-camera depth estimation. Our key contributions include: 1) the **Spatial-Temporal Transformer (STTrans)**, which integrates local spatial connectivity and global context to capture enriched spatial-temporal cross-view correlations, resulting in more accurate 3D geometry reconstruction; 2) the **Spatial-Temporal Photometric Consistency Correction (STPCC)** strategy that mitigates the impact of varying illumination, ensuring brightness consistency across frames during photometric loss calculation; 3) the **Adversarial Geometry Regularization (AGR)** module, which employs Generative Adversarial Networks to impose spatial constraints by using unpaired depth maps, enhancing performance under adverse conditions such as rain and nighttime driving. Extensive evaluations on large-scale autonomous driving datasets, including Nuscenes and DDAD, confirm that STViT+ sets a new benchmark for multi-camera depth estimation.

Keywords Multi-camera perception · Depth estimation · Spatial-temporal transformer · Adversarial geometry regularization

1 Introduction

Depth estimation is the process of determining the distance of objects in the scene from the camera by assigning depth values to each pixel in input RGB images, thereby reconstructing the 3D geometric structure of the environment. This task is essential for perceiving spatial relationships and serves as a foundation for several critical technologies. In particular, depth estimation is critical for key applications in fields such as autonomous driving, robotics, drone navigation, and virtual/augmented reality [1–6]. In the context of autonomous driving, accurate depth estimation enables vehicles to effectively interpret their surroundings in three dimensions, which is crucial for safety and navigation. Multi-camera depth estimation, which leverages multiple camera viewpoints, plays a pivotal role in these systems, as it provides a more comprehensive and precise understanding of the environment compared to monocular methods. By combining the information from multiple angles, multi-camera setups can generate richer spatial data, facilitating enhanced 3D scene reconstruction. This is particularly valuable for tasks such as object detection [7, 8], path planning [9], and collision avoidance [10], which require a detailed understanding of both the vehicle's immediate surroundings and distant obstacles.

With the advent of the deep learning techniques [11, 12], supervised depth estimation has garnered significant attention. These approaches typically rely on high-precision devices such as LiDAR to generate ground truth depth from 3D point clouds, which are subsequently used to supervise network training. Depth estimation is often framed as a regression [13, 14] or classification [15, 16] task and these methods have exhibited impressive performance, thereby propelling advancements in 3D perception. However, due to the difficulty and high cost of obtaining LiDAR devices, accurate ground truth depth is rarely available in practical applications, posing challenges to the widespread adoption of supervised depth estimation methods.

Consequently, a growing body of research has shifted towards self-supervised depth estimation. These methods leverage photometric consistency across consecutive frames as a supervisory signal, allowing for the simultaneous optimization of depth and pose estimation. In a typical pipeline, a

Zhuo Chen and Haimei Zhao contributed equally to this work

Extended author information available on the last page of the article

Published online: 16 January 2025

depth network and a pose network are employed to predict the corresponding depth maps and pose transformations. These predictions are used to warp the source frame to adjacent frames, achieving network optimization by minimizing the photometric difference between the original images and the warped images. Notably, these approaches [17–19] utilize multiple frames during the training phase for loss computation, and a single monocular image is required during inference. As such, they are categorized as Single-frame Self-supervised Monocular Depth Estimation methods, as shown in Fig. 1 A.

To leverage the abundant sequential image data effectively, some methods [20–22] propose utilizing multi-frame images as input during both training and inference stages, as shown in Fig. 1 B. The inter-frame geometric correlations are usually exploited by constructing cost volumes or correlation layers. These approaches significantly enhance the performance of self-supervised depth estimation methods by harnessing the temporal multi-frame correlations.

In addition to the ongoing advancements in self-supervised monocular depth estimation techniques, recent methods [23–25] have extended monocular methods to multi-camera configurations to fulfill the perceptual demands of autonomous driving cars with 360-degree surround-view cameras. These approaches enhance the monocular framework by enabling cross-camera feature interaction and fusion, leveraging the overlap among adjacent cameras to boost the representation learning, as shown in Fig. 1 C. Furthermore, by taking multi-camera sequence as input, the overlap of field-of-view (FoV) not only exists in adjacent cameras but also in adjacent temporal frames. This comprehensive integration of spatial and temporal data facilitates more robust depth representation learning, as shown in Fig. 1 D.

While adapting monocular self-supervised methods to the multi-camera setup has demonstrated promise in previous methods, several challenges specific to multi-camera setups remain unaddressed, impeding further performance improvement. Self-supervised depth estimation methods highly rely on the co-visible regions across different frames to compute reprojection errors. Additionally, they assume that the corresponding pixels of the same 3D point in different images exhibit identical intensities, an assumption frequently violated due to factors such as variations in illumination, extreme weather, and occlusion. For the multi-camera setups in large-scale autonomous driving datasets collected in real-world environments, such as NuScenes [26] and DDAD (the Dense Depth for Autonomous Driving dataset [27]), the challenges include 1) The overlap between adjacent cameras (e.g., the front camera w.r.t. the front-left or front-right cameras) is too small (as low as 10% [25]) to conduct effective image or feature matching for accurate 3D geometry recovery; 2) Various challenging condition, such as adverse weather or low-light scenarios (e.g., driving in rainy days or at night), hinder the ability to provide accurate photometric supervision, which is critical for self-supervised depth estimation.

In this paper, we explore the Multi-frame Self-supervised Multi-camera Depth Estimation paradigm and propose novel techniques to mitigate these challenges and enhance performance. First, we propose a Spatial-Temporal Transformer that effectively captures both local connectivity and the global context from image features, while learning enriched spatial-temporal cross-view correlations for improved 3D geometry reconstruction. As shown in Fig. 2, our approach leverages not only cross-camera correlations within the same frame (yellow arrows) and cross-frame correlations of the same camera (views with the corresponding same color)

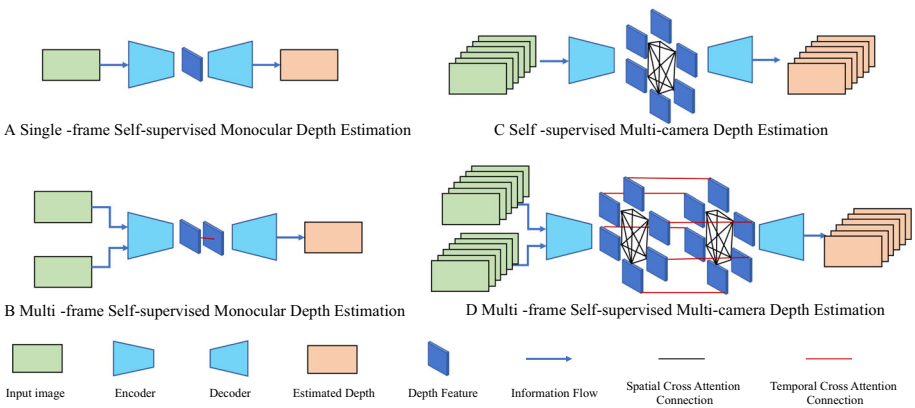

Fig. 1 Comparison of self-supervised depth estimation pipelines under different settings (A): Single-frame Self-supervised Monocular Depth Estimation; (B): Multi-frame Self-supervised Monocular Depth Estimation; (C): Self-supervised Multi-camera Depth Estimation; (D): Multi-frame Self-supervised Multi-camera Depth Estimation). For simplicity, only depth networks are illustrated and the corresponding pose networks are omitted

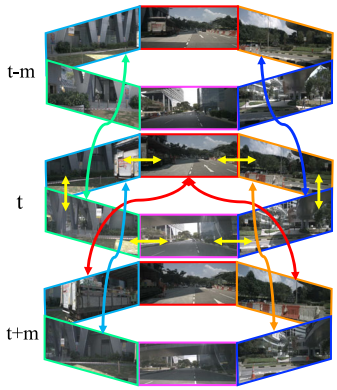

Fig. 2 Illustration of simultaneous cross-camera and cross-frame correlations

but also cross-camera and cross-frame correlations (different cameras in different frames, shown as colorful arrows across different temporal views). This strategy maximizes the use of co-visible overlap among images, enhancing both feature matching and network training. However, varying illumination and brightness across cameras and frames during driving can negatively impact both image correlation acquisition and projection error calculation in the self-supervised learning process. To mitigate this, we employ a spatial-temporal photometric consistency correction strategy to adjust image intensities and maintain brightness consistency. Additionally, we introduce a Generative Adversarial Network-based geometry regularization module to address prediction anomalies in challenging conditions such as rainy and nighttime scenarios. This paper builds upon and significantly expands the preliminary ideas and concepts initially presented in our previous two-page conference abstract [28].

In summary, the main contributions of this paper are fourfold:

- We tackle the challenging task of self-supervised multi-camera depth estimation by introducing STViT+, a novel Transformer-based framework.
- We develop the Spatial-Temporal Transformer (STTrans) for comprehensive feature extraction, capturing both cross-camera and cross-frame geometric correlations. Together with the Spatial-Temporal Photometric Consistency Correction (STPCC), our method effectively leverages the spatial-temporal context to enhance depth and pose learning.
- We propose the Adversarial Geometry Regularization (AGR) module, which imposes spatial positional constraints on predicted depth maps, mitigating prediction anomalies in challenging cases such as rainy and nighttime conditions.
- We conduct comprehensive evaluations and ablation studies, demonstrating the effectiveness of our method. STViT+ achieves state-of-the-art results on two large-scale self-supervised multi-camera depth estimation benchmarks: NuScenes [26] and DDAD [27].

2 Related work

2.1 Monocular single-frame self-supervised depth estimation

Research into self-supervised depth estimation initially began with monocular settings, wherein researchers employed monocular image sequences as training data and estimated depth maps for individual monocular frames during inference. SfMLearner [17] is one of the first attempts to explore monocular depth estimation in a self-supervised manner. It exploits predicted depth and pose to warp source images to reconstruct its adjacent images thereby formulating the learning as a projection error minimization process. Many subsequent works further improve this paradigm by additionally introducing 3D constraint [29], imposing feature-level consistency [30], integrating uncertainty learning [31, 32] and incorporating related tasks [33, 34], e.g., optical flow estimation [35–37] and semantic segmentation [38, 39]. Monodepth2 [18] proposes several schemes to improve the effectiveness of photometric loss, including a minimum reprojection loss and an auto-masking strategy, yielding more accurate results. Recently, many works (DIFFNet [36], MonoFormer [40], MonoViT [19] and SRD [41]) explore stronger network architectures to enhance the representation learning ability including PackNet [27], HRNet [42] and Vision Transformer [43], further improves the prediction accuracy. Besides, there is a line of work devoted to addressing illumination issues in adverse conditions such as nighttime driving scenarios [44–46]. Some methods [44, 45] utilize domain adaptation techniques to adapt the daytime training estimation network to be applicable for the nighttime scenes. STEPS [46] proposes to jointly learn a nighttime image enhancer and a depth estimator to overcome the low illumination problems in the depth estimation task, but additional illumination estimation and calibration networks are imposed, increasing computation burdens.

2.2 Monocular multi-frame self-supervised depth estimation

Given the availability of image sequences as training data, researchers then embarked on investigating how to leverage temporal information to further enhance the efficacy of monocular depth estimation. TC-Depth [47] fused the multi-frame features with proposed spatial and temporal attention modules to create a multi-frame depth estimation network, which improves the temporal depth stability and accuracy

by combining modules with photometric cycle consistency. Inspired by multi-frame stereo methods [48, 49], ManyDepth [20] is introduced as an innovative self-supervised multi-frame depth estimation model that capitalizes on the synergies between monocular and multi-view depth estimation, incorporating multiple frames during the testing phase. DepthFormer [21] proposed a novel end-to-end transformer, which generates cost volume through multi-view feature matching via cross- and self-attention with depth-discretized epipolar sampling. IterDepth [50] further improves the multi-frame monocular depth estimation approach with the proposed iterative residual refinement network, incorporating a gated recurrent depth fusion unit to enable iterative feature fusion and inverse depth prediction. DS-Depth [51] presents a dynamic cost volume leveraging residual optical flow to improve occlusion handling, further enhanced by a fusion module. Additionally, pyramid distillation and adaptive photometric error losses are proposed for accuracy improvement.

2.3 Multi-camera self-supervised depth estimation

Multi-camera depth estimation is a long-standing topic, which is usually solved by multi-view stereo, *i.e.*, reconstructing 3D information of the scene from pictures of different angles. Multi-view stereo usually needs a large overlap to conduct image matching and cost volume construction, which is not suitable for driving scenes. FSM [23] extends self-supervised monocular depth estimation to the surrounding multi-camera setting to meet the increasing demand in autonomous driving scenarios. FSM focuses on enhancing the self-supervision signal by leveraging spatial and temporal contexts to enrich the photometric consistency supervision and imposing pose consistency constraints to learn robust pose estimation. SurroundDepth [24] utilizes a shared encoder to extract high-level feature maps for each view with a cross-view transformer to fuse features and capture cross-view interactions. MCDP [25] formulate the depth estimation as a weighted combination of depth basis to iteratively update and propagate to maintain a consistent structure of depth predictions. EGA-Depth [52] simplifies the cross-attention mechanism employed in SurroundDepth by limiting cross-attention to adjacent cameras for each individual camera. This refinement enables cross-attention to be conducted on higher-resolution features, further improving the accuracy (Fig. 3).

2.4 Self-supervised depth estimation with generative adversarial networks

Generative Adversarial Networks (GANs) [53] have garnered significant attention in various vision tasks, including style transfer [54, 55], image-to-image translation [56, 57], image editing [58–60], and cross-domain image generation [61, 62]. Since our proposed Geometry Regularization Module is based on a GAN, we give a brief review of previous works on self-supervised depth estimation with GAN. One line of research [63–65] utilizes GANs as a robust loss item to differentiate between warped images and original images within the self-supervised depth estimation pipeline. Some other approaches [66–68] leverage the image-to-image translation ability of GANs to enhance input image quality or transfer synthetic and realistic images, utilizing additional synthetic datasets [69] for domain adaptation. Wu et al. [70] and Wang et al. [71] also designed a GAN-based module for regularization and refinement. However, the former is dedicated to distinguishing the ground-truth depth map and the

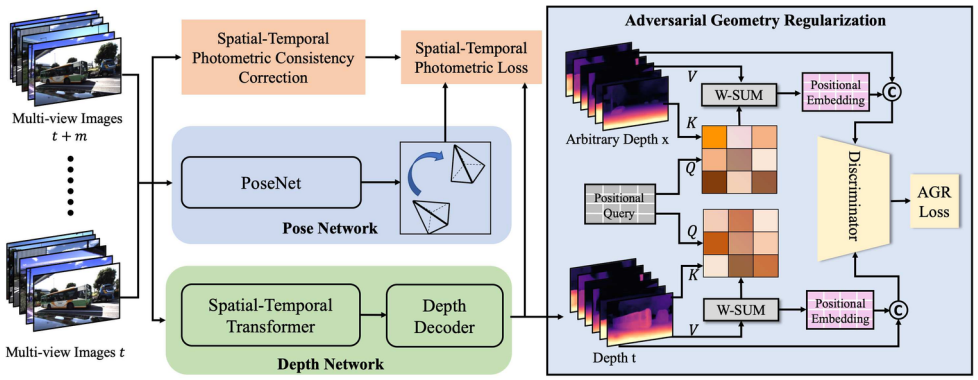

Fig. 3 Overview of our STViT+ framework. Our STViT+ is composed of a Depth Network, a Pose Network, and an Adversarial Geometry Regularization Module. The Depth Network consists of a Spatial-Temporal Transformer Encoder and a Depth Decoder. The Pose Network is implemented by a lightweight ResNet. The 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 Adversarial Geometry Regularization Module

predicted depth map while the latter seeks to constrain the incorrect depth estimations during nighttime using daytime prediction in an adversarial manner. In contrast, our work diverges from these studies in two key aspects: 1) we utilize arbitrary depth maps from other scenes to regularize the depth maps of the corresponding camera without relying on specific ground truth or predictions under varying illumination conditions, and 2) we design a novel depth-aware positional embedding that, along with predicted depth maps, serves as the input for the discriminator instead of the corresponding RGB frames or coordinates.

3 Method

3.1 Network architecture of STViT+

3.1.1 Motivation

In self-supervised depth estimation algorithms, explicit ground truth information is absent, and supervisory signal relies solely on photometric consistency across different viewpoints. In a multi-camera setup with six cameras capturing temporal sequences, a wealth of data is available for both training and inference. Consequently, depth estimation networks must effectively extract both local and global features from the input images. This process involves comprehensive feature extraction from individual frames, as well as the acquisition of geometric features across temporally sequential frames and co-observable regions among different camera viewpoints. Previous multi-camera self-supervised depth estimation methods typically employed Convolutional Neural Networks (CNNs) to extract features from input images and subsequently performed cross-attention operations between these features. However, the localized nature of convolution operations often limits CNNs in capturing long-range context similarity and dependencies effectively. Due to the localized nature of the extracted features, which tend to excessively focus on individual objects or semantic categories, subsequent attempts to capture inter-frame correlations through subsequent cross-attention mechanisms have been rendered ineffective. This has hindered the accurate recovery of the geometric information for the entire scene [19]. Therefore, to enhance the extraction of both global and local geometric features and leverage correlations across different viewpoints and sequential frames, we introduce the Spatial-Temporal Transformer Framework, referred to as STViT+, which is specifically designed for multi-camera self-supervised depth estimation. It follows the typical self-supervised depth estimation structure, consisting of a Depth Network and a Pose Network. The Depth Network integrates a Spatial-Temporal Transformer and a Decoder.

3.1.2 Depth network

Similar to prior works, our Depth Network is designed following the encoder-decoder architecture. We will explain the details of Depth Network in the following sections.

Spatial-Temporal Transformer (STTrans) Previous studies [19, 27] have highlighted the importance of extracting effective features to improve the performance of depth estimation. Therefore, we enhance the encoder architecture for multi-camera self-supervised depth estimation by employing powerful vision transformer models. We propose a Spatial-Temporal Transformer to not only leverage the transformer's ability to model long-range dependencies, overcoming the locality issue in feature extraction seen in previous works [18, 24], but also introduce Spatial-Temporal Cross-Correlation to fully exploit the co-visibility regions across cameras and temporal frames for geometric structure recovery. Inspired by recent transformer models such as MPViT [72] which introduces the concept of a Multi-Path Transformer Block, we devise a Depth Encoder to capture both local and global context within images and further exploit the spatial-temporal cross correlations.

As shown in Fig. 4, our Depth Encoder consists of Conv-Stem and Spatial-Temporal Transformer Layers. Each Transformer layer contains Multi-Scale Patch Embedding, Transformer Blocks, a Convolutional Block, a Global-to-Local Feature Interaction, and a Spatial-Temporal Cross Correlation Module. The input multi-camera sequence is fed to a Conv-Stem and then Spatial-Temporal Transformer Layers to obtain the depth feature. The Spatial-Temporal Transformer layer first embeds the extracted features into different-sized visual tokens in Multi-Scale Patch Embedding which is formed by several parallel convolutional patch embedding layers with different kernel sizes, to exploit both fine- and coarse-grained visual tokens at the same feature level following MPViT [72]. After that, parallel Transformer Blocks and Convolutional Block are leveraged to further process the embedded tokens. As shown in Fig. 4, there are three Transformer Blocks to capture the long-range dependencies and global context. Each Block contains M Transformer Layers, which consists of a Layer Normalization (LayerNorm) module, a Factorized Multi-head Self Attention (MHSA) layer [72], another Layer Normalization, and a Feed-forward Network (FFN). Parallel to the Transformer Blocks, a Convolutional Block is used to exploit local connectivity from translation invariance. The Convolutional Block comprises a sequence of 1×1, 3×3 depth-wise, and 1×1 convolutions. By combining the advantages of Transformer Blocks and Convolutional Blocks, the modeled feature can capture both local connectivity and global context simultaneously. A subsequent Global-to-Local Feature Interaction is further used to enhance the local and global feature interactions to obtain

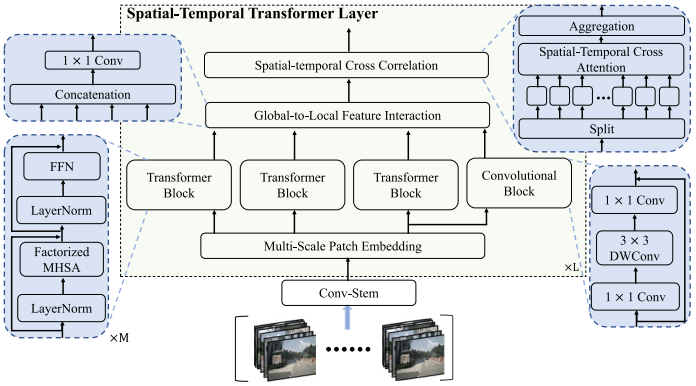

Fig. 4 The architecture of Depth Encoder. It is composed of Conv-Stem and Spatial-Temporal Transformer Layers. Each Transformer layer contains Multi-Scale Patch Embedding, Transformer Blocks, a Convolutional Block, a Global-to-Local Feature Interaction, and a Spatial-Temporal Cross Correlation Module. The structure of each module is illustrated in blue blocks

enriched representations. We use Enc_{depth} to represent the feature extraction part of the model:

$$F = Enc_{depth}([I_t, ..., I_{t+m}]), \tag{1}$$

where I_t represents the surrounding six images at timestamp t and $m \in \{-1, 0, 1\}$ means we take three temporal frames as input in our default setting. Although 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 a Spatial-Temporal Cross Correlation module to facilitate correlation learning and geometry recovery.

Spatial-Temporal Cross Correlation As shown in Fig. 4, the interacted features are first split into different cameras and different temporal frames, e.g. F_t^i denoting the feature of the ith camera in timestamps t. For each feature F_t^i, we pre-define the list of views that share overlap regions with feature F_t^i. 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, as shown in Fig. 2. Thus, Spatial-Temporal Cross Correlation is able to learn enriched spatial-temporal cross-view correlations. Specifically, feature F_t^i is leveraged to compute queries, and the features of overlapped views G_{t+m}^x are used to obtain keys and values:

$$Q_t^i = F_t^i W_{q_t}^i,\ K_t^i = G_{t+m}^x W_{k_t}^i,\ V_t^i = G_{t+m}^x W_{v_t}^i. \tag{2}$$

Here, $x \in [1, N]$ and N is the number of cameras. $W_{q_t}^i$, $W_{k_t}^i$, and $W_{v_t}^i$ are the learnable projections for query, key, and value. Thus, the feature after cross correlation is:

$$\hat{F}_t^i = softmax(\frac{Q_t^i K_t^i}{\sqrt{d}}) V_t^i, \tag{3}$$

where d denotes the embedding dimension. By interacting with features of both cross-camera and cross-frame views, features $\hat{F}_t^i$ can learn enriched cross-view context and correspondence, which is beneficial to accurately inferring 3D geometry.

Depth decoder Utilizing multi-scale features obtained from the depth encoder, our depth decoder incorporates cross-layer and cross-scale connections following [19]. Recognizing the contextual distinctions among features at various scales, such as the preference for higher-resolution features for fine-grained details, we bolster cross-scale feature integration. To achieve this, we employ both spatial and channel attention mechanisms. Ultimately, the disparity (inverse depth) predictions D under different resolutions are handled by four heads comprising two convolutional layers and a Sigmoid activation function. The prediction of depth maps can be formulated as:

$$D_t^i = Dec_{depth}(\hat{F}_t^i), \tag{4}$$

where Dec_{depth} is the decoder of depth network, i denotes the ith camera and t denotes the tth frame.

3.1.3 Pose network

Following the common practice of self-supervised depth estimation, we adopt a lightweight ResNet18 [12] as the encoder Enc_{pose} of Pose Network and subsequent convolution layers as the decoder Dec_{pose} to regress the 6 DoF relative poses P between adjacent temporal frames. Specifically, we first take N pairs of adjacent frames as the input and output a universal pose $P_{t+m\to t}$ for all N cameras [24]. Then, the predicted universal pose $P_{s\to t}$ is transformed to each specific camera with its known camera extrinsic matrix. The whole process

of predicting pose can be formulated as:

$$\begin{aligned} h^i &= Enc_{pose}([I_t^i, I_{t+m}^i]), \\ P_{t+m\to t} &= Dec_{pose}(\frac{1}{N}\sum_{i=1}^{N} h^i), \\ P_{t+m\to t}^i &= (T^i)^{-1} P_{t+m\to t} T^i, \end{aligned} \tag{5}$$

where $P_{t+m\to t}^i$ is the learned pose for the ith camera and T^i is its corresponding extrinsic matrix. The universal pose prediction manner can naturally ensure geometry consistency among cameras.

3.2 Self-supervised training

The self-supervised depth estimation problem is formulated to a projection error minimization process, where the depth network and pose network are jointly optimized. Given the input images, depth maps D and relative pose transformation P are predicted with depth network and pose network. Then the depth map and pose are utilized to reproject the source image to reconstruct the target image. The networks are optimized by minimizing the difference between the synthesized target image and the original target image.

3.2.1 Spatial-temporal photometric loss

The target image can be each frame e.g., I_t^i denoting the image captured by the ith camera at the timestamp t. To fully exploit the spatial-temporal consistency, the source images include not only the spatial neighborhood, the temporal neighborhood and also the cross-frame and cross-camera views with overlapped regions. Similar to the Spatial-Temporal Cross Correlation part, we pre-define a list of views that can be observed co-visible regions with the target image, e.g., the length of the ith correlation image list is CI. Thus, the photometric loss ℓ_p can be formulated as:

$$\ell_p = \sum_{i=1}^{N}\sum_{ci=1}^{CI} \ell_{ph}(I_t^i, I_{s\to t}^{ci}). \tag{6}$$

The reconstructed target image $I_{s\to t}^{ci}$ is obtained via reprojection with the predicted depth map D_t^i and pose $P_{t\to s}^{ci}$:

$$I_{s\to t}^{ci} = Proj(K, P_{t\to s}^{ci}, D_t^i, K^{-1}, I_t^i), \tag{7}$$

where K is the camera intrinsic matrix. The typical photometric loss in prior works comprises an SSIM [73] metric and L1 Loss term:

$$\ell_{ph}(I_t^i, I_{s\to t}^{ci}) = \alpha\frac{1-SSIM(I_t^i, I_{s\to t}^{ci})}{2} + (1-\alpha)||I_t^i - I_{s\to t}^{ci}||. \tag{8}$$

Moreover, an edge-aware smoothing term is often incorporated to add a regularization on depth maps in many previous works [18, 74]:

$$\ell_{sm} = |\partial_x \mu_{D_t}| e^{-|\partial_x I_t|} + |\partial_y \mu_{D_t}| e^{-|\partial_y I_t|}, \tag{9}$$

where μ_{D_t} is the inverse depth normalized by mean depth. $\partial_x \mu_{D_t}$ and $\partial_x \mu_{D_t}$ denote the disparity gradient among two directions.

3.2.2 Spatial-temporal photometric consistency correction (STPCC)

The photometric loss is designed based on the assumption that the same 3D points have the same intensity in diverse projected views. However, in practical outdoor driving scenarios, the illumination among different cameras and different timestamps can vary severely, which impedes network learning. Therefore, we propose Spatial-Temporal Photometric Consistency Correction (STPCC) to enforce the brightness consistency of diverse views before the calculation of photometric loss.

Inspired by Contrast Limited Histogram Equalization (CLHE) [75], we leverage a common mapping function ψ to correct image brightness and make the image color spatially and temporally consistent. We first compute the histograms H of input images, which are the frequency distributions of L intensity levels (usually $L \in \{0, 1, ..., 255\}$) of images. The histograms of spatially and temporally adjacent images, (taking temporal images as examples, H_{t-1}, H_t, H_{t+1}), are then processed by a normalization operation, $H = avg(H_{t-1}, H_t, H_{t+1})$. Based on the normalized frequency distribution, by setting a threshold ω, we assume that if a certain intensity level in the histogram exceeds the threshold, it will be clipped, and the portions exceeding the threshold will be evenly distributed among the various intensity levels, as shown in Fig. 5. After adjusting the Histograms consistently, the mapped figure (taking I_t' as an example) can be obtained:

$$\bar{I}_t = \psi(H(I_t)) = \frac{CDF(H(I_t)) - CDF_{min}}{CDF_{max} - CDF_{min}} \times (|L|), \tag{10}$$

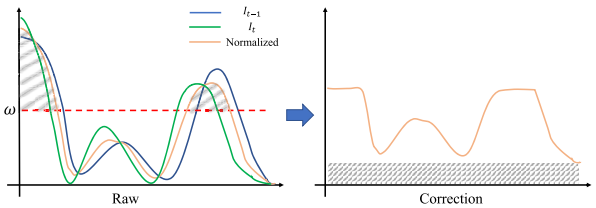

Fig. 5 Illustration of the step of histogram adjusting in Spatial-Temporal Photometric Consistency Correction

where CDF represents the Cumulative Distribution Function. CDF_{max} and CDF_{min} are the corresponding maximum and minimum values of CDF.

In this way, the intensities distribution of spatially and temporally adjacent images can be aligned consistently. Moreover, the brightness of images in adverse illustration conditions *i.e.*, night or dark driving scenarios can be adjusted with higher visibility. The color correction effect is illustrated in Fig. 6. The top two rows show the correction effect of three temporally adjacent images, which makes the brightness more consistent. The bottom two rows show the correction effect of spatially adjacent images in nighttime scenarios, which adjusts and improves the visibility. Note that STPCC is only applied to images before the photometric loss computation rather than the input for network learning. Therefore, the final photometric loss is:

$$\ell_p = \sum_{i=1}^{N} \sum_{ci=1}^{CI} \ell_{ph}(\bar{I}_t^i, \bar{I}_{s\to t}^{ci}). \tag{11}$$

3.3 Adversarial geometry regularization module (AGR)

In real-world outdoor driving scenarios, adverse conditions such as rainy weather and nighttime driving are frequently encountered. Under such extreme circumstances, the effectiveness of photometric loss diminishes, thereby significantly affecting the performance of depth estimation. Therefore, we propose a GAN-based Adversarial Geometry Regularization Module (AGR) to further constrain the depth estimation, as shown in the right part of Fig. 3. 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 [76]. Thus, we use the positional query to scan over the depth map which serves as key and value. So that we can obtain the depth-aware positional embedding e_t^i by calculating the dot product similarity between the query and keys. In this way, the depth-aware positional embedding can provide soft geometric correspondence between query positions and depth maps. After that, the positional embedding is concatenated with the normalized predicted depth maps, denoted as $[e_t^i, \mu(D_t^i)]$. Similarly, the arbitrary depth maps are also concatenated with the corresponding positional embedding, denoted as $[e_t^{iR}, \mu(D_t^{iR})]$. We use the PatchGAN [56] discriminator Θ_{Dis} to distinguish $[e_t^i, \mu(D_t^i)]$ and $[e_t^{iR}, \mu(D_t^{iR})]$, while the depth network tries to make the prediction $[e_t^i, \mu(D_t^i)]$ indistinguishable with the regularization reference $[e_t^{iR}, \mu(D_t^{iR})]$. The PatchGAN network consists of 5 layers, which progressively extract features from the input image. Each convolutional layer is followed by a LeakyReLU activation function, introducing non-linearity to the network. Batch normalization layers are inserted after every other convolutional layer to stabilize and speed up training. The final layer of the network is a separate convolutional layer and the output patch size is 1/8 times the original depth predictions. Overall, the network

Fig. 6 Qualitative comparison of images before and after Spatial-temporal Photometric Consistency Correction. The top two rows illustrate the correction effect on three temporally adjacent images. The bottom two rows demonstrate the correction effect on spatially adjacent images captured in nighttime conditions. For optimal details, viewing with zoom-in is recommended

gradually reduces the spatial dimensions of the input while increasing the number of feature channels, culminating in a classification output with two units corresponding to the desired classes.

The optimization objective for AGR can be formulated as:

$$\begin{aligned} L_{Dis} =& \frac{1}{2}\mathbb{E}_{D_t^{iR}}[(\Theta_{Dis}([e_t^{iR}, \mu(D_t^{iR})]) - 1)^2] \\ &+ \frac{1}{2}\mathbb{E}_{D_t^{i}}[\Theta_{Dis}([e_t^{i}, \mu(D_t^{i})])^2], \\ L_{Gen} =& \frac{1}{2}\mathbb{E}_{D_t^{i}}[(\Theta_{Dis}([e_t^{i}, \mu(D_t^{i})]) - 1)^2], \\ L_{AGR} =& \min_{Gen}\max_{Dis} L_{Dis} + L_{Gen}. \end{aligned} \tag{12}$$

The AGR module acts as a discriminator, applying spatial constraints to the depth estimation network (the generator) during training. The discriminator loss minimizes prediction errors by discouraging outliers caused by these challenging scenarios, thereby refining the spatial consistency of depth estimation. During inference, AGR is bypassed to prevent additional computational overhead, maintaining the model's efficiency. Our ablation studies confirm that AGR significantly improves depth accuracy, especially in difficult environments, without affecting inference performance.

3.4 Training loss

To sum up, the final training loss consists of the photometric loss ℓ_p (11), the smoothing loss ℓ_{sm} (9) and the AGR regularization loss ℓ_{AGR} (12):

$$Loss = \ell_p + 10^{-3}\ell_{sm} + 5 \times 10^{-4}\ell_{AGR}. \tag{13}$$

Here, the loss weights of the photometric loss and the smoothing loss are kept the same as the monocular depth estimation methods [18, 74] while the parameter of the AGR regularization loss is obtained by empirical experiments.

4 Experiment

4.1 Datasets

Following the common practice in previous multi-camera depth estimation methods, we adopted NuScenes [26] and DDAD [27] to evaluate our method. These two recently released autonomous driving datasets are both with six surrounded cameras and relatively small overlaps among cameras, which are more challenging than the prior monocular datasets.

NuScenes The NuScenes dataset [26] encapsulates urban driving contexts and is characterized by a coordinated assemblage of imagery acquired from a sextuple-camera configuration. This compilation encompasses 1,000 distinct scenes and boasts an extensive repository of 1.4 million images. Renowned for its role as a benchmark for diverse tasks encompassing 2D and 3D object detection, alongside semantic and instance segmentation, this dataset assumes a pivotal position in the domain. Particularly pertinent to the self-supervised depth estimation task, the NuScenes dataset poses inherent challenges attributed to the relatively modest image resolution, constrained spatial inter-camera overlap, variegated weather conditions, diurnal temporal variations, and complex, unstructured settings. The raw image dimensions are specified as 1600×900, subsequently downscaled to a resolution of 640×352. Captured at a frequency of 30Hz, dataset samples are annotated at a reduced 2Hz cadence, dictated by keyframes. The temporal interval between these key frames is appreciably large, precluding the training of deep networks through conventional self-supervision techniques. Consequently, annotated Sweep data emerge as a viable recourse, furnishing pivotal supervisory signals in the training process.

DDAD The Dense Depth for Automated Driving (DDAD) dataset [27] encompasses urban driving scenarios and has been meticulously recorded through six synchronized cameras, displaying limited spatial overlap. It is distinguished by its provision of highly precise dense ground-truth depth maps for evaluative purposes, extending up to an impressive maximum depth range of 250 meters. This dataset comprises a training subset encompassing 12,650 instances (comprising 63,250 images) and a validation subset containing 3,950 instances (consisting of 15,800 images). In the training set, the utilization of ground-truth depth maps is eschewed. Notably, the image resolution is denoted as $1{,}936 \times 1{,}216$, following which, in consonance with the methodology delineated in [16], input images undergo a downsampling procedure to achieve a resolution of 640×384. Subsequently, during the evaluation phase, image resolution is restored to its original dimensions through bilinear interpolation.

4.2 Evaluation metrics

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:

- $Abs\ Rel = (1/n)\sum_{i\in n}((|d_i - d_i^*|)/d_i)$,
- $Sq\ Rel = (1/n)\sum_{i\in n}((||d_i - d_i^*||^2)/d_i)$,
- $RMSE = ((1/n)\sum_{i\in n}||d_i - d_i^*||^2)^{1/2}$,

- $RMSE\ log = ((1/n)\sum_{i\in n} ||log(d_i) - log(d_i^*)||^2)^{1/2}$
- Accuracy: % of d_i s.t. $max((d_i/d_i^*), (d_i^*/d_i)) = \delta < \delta_n$,

where n is the total number of pixels in the ground truth depth map, d_i and d_i^* represent the predicted and ground truth depth value of pixel i. δ_n denotes a threshold, which is usually set to 1.25^1, 1.25^2 and 1.25^3 (Fig. 7).

4.3 Implementation details

We implement our STViT+ in Pytorch. The model is trained for 5 epochs on the NuScenes dataset [26] and 20 epochs on the DDAD dataset [27] using AdamW as the optimizer and a batch size set to 6. The initial learning rate for PoseNet and depth decoder is 10^{-4}, while the Transformer-based depth encoder is trained with an initial learning rate of 5×10^{-5}. Both the pose encoder and depth encoder are pre-trained on ImageNet [43]. We use 4 A100 GPUs for the experiments on Nuscenes and 8 GPUs for experiments on DDAD. In our experiments, we adopt the same data augmentation detailed in [18, 19]. For our default setting, we use 3 temporal frames as input and we also test the version with a single temporal input.

4.4 Comparison with the state-of-the-arts

We conduct extensive quantitative evaluations on two large-scale autonomous driving datasets, *i.e.*, Nuscenes [26] and DDAD [27] datasets. Our method is compared with two approaches adapted from monocular depth estimation methods [18, 27] and four state-of-the-art multi-camera-based methods [23–25, 52]. The detailed evaluation results are presented in Tables 1 and 2. In comparison with recent state-of-the-art methods [24, 25, 52], our approach demonstrates superior performance across most evaluation metrics, achieving the best results in five out of seven metrics on Nuscenes and four out of seven on DDAD. Our method leverages multiple temporal sequences input in the Spatial-Temporal Transformer, and for completeness, we also showcase its performance with a single temporal input (six camera figures at the same timestamp). Despite a slight performance degradation without temporal input and modeling, our single-input version still delivers promising results compared to other advanced methods (Tables 1 and 2).

4.5 Ablation study

4.5.1 Performance of individual cameras

To provide a comprehensive understanding of inference performance, we present an extensive presentation of evaluation results concerning the six individual cameras in both the Nuscenes and DDAD datasets, detailed in Tables 3 and 4, respectively. The experiment reveals that self-supervised depth estimation performs exceptionally well on front views compared to back views. Furthermore, the inference results in the left view significantly outperform their right counterpart. This divergence might be attributed to the inherent dissimilarities in scenes captured on opposing sides, signifying the sensitivity of the model to the specific spatial characteristics within its field of vision. This detailed examination and analysis may shed light on the intricacies of its responses to diverse perspectives, contributing valuable insights for future refinement in model designation and learning strategies.

4.5.2 Ablation study for proposed contributions

To demonstrate the effectiveness of each component of our methods, we conduct thorough ablation studies on both the Nuscenes and DDAD datasets, with detailed findings presented in Table 5. Utilizing SurroundDepth as our baseline, we systematically introduce and evaluate each augmentation, including the Spatial-Temporal Transformer (STTrans), Spatial-Temporal Photometric Consistency Correction (STPCC), and the Adversarial Geometry Regularization module (AGR). The performance trends observed across the datasets exhibit consistent variations. The Spatial-Temporal Transformer notably enhances depth estimation outcomes, leveraging improved feature extraction and spatial-temporal cross-view feature interaction. STPCC, augmenting the photometric loss calculation through adjustments in the alignment of multiple spatial-temporal input images, brings further enhancements, as evidenced in Table 5. Moreover, the Adversarial Geometry Regularization module, denoted as AGR, significantly reduces prediction errors, val-

Example scenes from NuScenes Example scenes from DDAD

Fig. 7 Example scenes display for two datasets Nuscenes [26] and DDAD [27]

Table 1 Quantitative evaluation of self-supervised multi-camera depth estimation on nuScenes [26]

Methods	Resolution	Abs Rel ↓	Sq Rel ↓	RMSE ↓	RMSE log ↓	$\delta < 1.25$ ↑	$\delta < 1.25^2$ ↑	$\delta < 1.25^3$ ↑
Monodepth2 [18]	352 × 640	0.287	3.349	7.184	0.345	0.641	0.845	0.925
PackNet-SfM [27]	352 × 640	0.309	2.891	7.994	0.345	0.547	0.796	0.899
FSM* [23]	352 × 640	0.334	2.845	7.786	0.406	0.508	0.761	0.894
SurroundDepth [24]	352 × 640	0.245 ±0.002	3.067 ±0.006	6.835 ±0.004	0.321 ±0.001	0.719 ±0.002	0.878 ±0.001	0.935 ±0.001
MCDP [25]	448 × 768	0.237	3.030	6.822	–	0.719	–	–
EGA-Depth [52]	352 × 640	0.239	**2.357**	6.801	0.317	0.723	**0.880**	0.936
STViT+ (single)	352 × 640	0.235 ±0.001	2.934 ±0.005	6.736 ±0.003	0.315 ±0.001	**0.724** ±0.001	0.877 ±0.001	0.936 ±0.001
STViT+	352 × 640	**0.233** ±0.001	2.815 ±0.004	**6.681** ±0.003	**0.312** ±0.001	**0.724** ±0.001	0.878 ±0.001	**0.937** ±0.001

The best results are highlighted in bold. The row of FSM* shows the results of FSM reproduced by [24]. The best results in each column are highlighted in **bold**, while the second-best ones are underlined. The error bar is displayed in red color, summarized from 5 times inference

Table 2 Quantitative evaluation of self-supervised multi-camera depth estimation on DDAD [27]

Methods	Resolution	Abs Rel ↓	Sq Rel ↓	RMSE ↓	RMSE log ↓	$\delta < 1.25$ ↑	$\delta < 1.25^2$ ↑	$\delta < 1.25^3$ ↑
Monodepth2 [18]	384 × 640	0.217	3.641	12.962	0.323	0.699	0.877	0.939
PackNet-SfM [27]	384 × 640	0.234	3.802	13.253	0.331	0.672	0.860	0.931
FSM* [23]	384 × 640	0.229	4.589	13.520	0.327	0.677	0.867	0.936
SurroundDepth [24]	384 × 640	0.200 ±0.002	3.392 ±0.004	12.270 ±0.004	0.301 ±0.002	0.740 ±0.001	0.894 ±0.001	0.947 ±0.001
MCDP [25]	384 × 640	<u>0.193</u>	3.111	12.264	–	**0.811**	–	–
EGA-Depth [52]	384 × 640	0.195	3.211	**12.117**	0.297	<u>0.743</u>	**0.896**	0.947
STViT+ (single)	384 × 640	<u>0.193</u> ±0.001	<u>3.093</u> ±0.002	12.206 ±0.002	<u>0.295</u> ±0.001	0.735 ±0.001	<u>0.895</u> ±0.001	<u>0.948</u> ±0.001
STViT+	384 × 640	**0.192** ±0.001	**2.965** ±0.002	<u>12.156</u> ±0.003	**0.293** ±0.001	0.734 ±0.001	<u>0.895</u> ±0.001	**0.949** ±0.001

The best results are highlighted in bold. The row of FSM* shows the results of FSM reproduced by [24]. The best results in each column are highlighted in **bold**, while the second-best ones are <u>underlined</u>. The error bar is displayed in red color, summarized from 5 times inference

Table 3 Quantitative evaluation of corresponding six cameras of self-supervised multi-camera depth estimation on Nuscenes [26]

Cameras	Resolution	Abs Rel ↓	Sq Rel ↓	RMSE ↓	RMSE log ↓	$\delta < 1.25$ ↑	$\delta < 1.25^2$ ↑	$\delta < 1.25^3$ ↑
Front	352 × 640	0.153	1.845	7.108	0.245	0.803	0.928	0.968
Front-Left	352 × 640	0.231	2.186	6.322	0.313	0.710	0.868	0.931
Back-Left	352 × 640	0.231	2.233	5.825	0.312	0.727	0.869	0.930
Back	352 × 640	0.193	2.277	7.286	0.292	0.741	0.901	0.954
Back-Right	352 × 640	0.304	4.372	6.569	0.358	0.676	0.846	0.918
Front-Right	352 × 640	0.286	3.980	6.974	0.352	0.688	0.858	0.922
All	352 × 640	0.233	2.815	6.681	0.312	0.724	0.878	0.937

Table 4 Quantitative evaluation results of corresponding six cameras of self-supervised multi-camera depth estimation on DDAD [27]

Cameras	Resolution	Abs Rel ↓	Sq Rel ↓	RMSE ↓	RMSE log ↓	$\delta < 1.25$ ↑	$\delta < 1.25^2$ ↑	$\delta < 1.25^3$ ↑
Front	384 × 640	0.130	2.699	13.219	0.216	0.845	0.945	0.977
Front-Left	384 × 640	0.186	2.745	11.845	0.294	0.745	0.898	0.948
Back-Left	384 × 640	0.199	2.885	11.419	0.301	0.729	0.891	0.944
Back	384 × 640	0.188	3.062	14.027	0.292	0.717	0.900	0.956
Back-Right	384 × 640	0.224	3.021	10.874	0.331	0.683	0.866	0.935
Front-Right	384 × 640	0.224	3.377	11.552	0.327	0.684	0.867	0.934
All	384 × 640	0.192	2.965	12.156	0.293	0.734	0.895	0.949

Table 5 Ablation study on Nuscenes [26] and DDAD [27]

Methods	Ablation study on Nuscenes				Ablation study on DDAD			
	Abs Rel	Sq Rel	RMSE	RMSE log	Abs Rel	Sq Rel	RMSE	RMSE log
Baseline	0.245	3.067	6.835	0.321	0.200	3.392	12.270	0.301
+ STTrans	0.238	2.889	6.732	0.316	0.195	3.126	12.204	0.297
+ STTrans + STPCC	0.236	2.864	6.709	0.315	0.194	3.103	12.189	0.295
+ STTrans + STPCC + AGR	**0.233**	**2.815**	**6.681**	**0.312**	**0.192**	**2.965**	**12.156**	**0.293**

STTrans denotes the Spatial-Temporal Transformer framework and AGR represents our Adversarial Geometry Regularization module

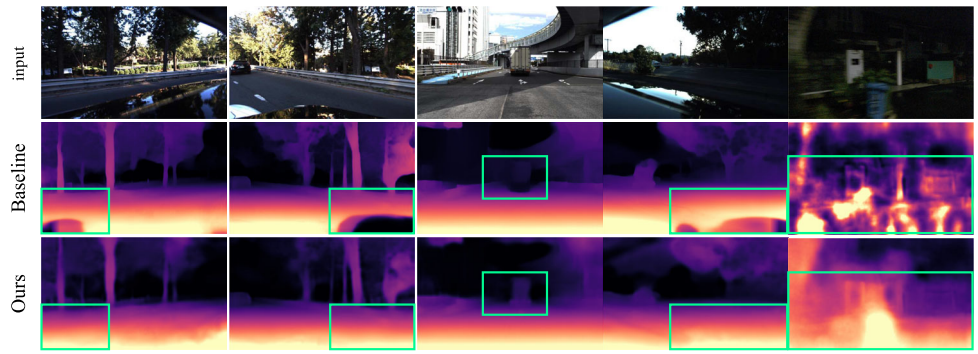

Fig. 8 Qualitative ablation of AGR. Regions with large differences are highlighted with green boxes. The visualization comparison can demonstrate the effectiveness of AGR in constraining prediction weirdness in low-illumination and nighttime driving scenarios

Table 6 Ablation study of Spatial-Temporal Transformer (STTrans) on Nuscenes [26]

Methods	Abs Rel ↓	Sq Rel ↓	RMSE ↓	RMSE log ↓
CNN Path only	0.256	3.418	8.675	0.301
Trans. Path only	0.248	3.272	8.016	0.398
1 Trans. Path	0.246	3.165	7.693	0.346
2 Trans. Path	0.243	3.134	7.238	0.325
STTrans	**0.233**	**2.815**	**6.681**	**0.312**

CNN Path and Trans. Path denotes the Convolutional Block and Transformer Block, respectively

idating its efficacy. To offer a more vivid illustration of the impact of AGR, we conduct a qualitative ablation by visualizing predicted depth maps both with and without the inclusion of AGR in the model, as depicted in Fig. 8. The comparison showcases that the model without AGR tends to generate artifacts in challenging conditions such as low-illumination regions. In contrast, our complete model incorporating AGR effectively mitigates these issues, underscoring the crucial role of AGR in enhancing the robustness of the model, especially in adverse conditions.

4.5.3 Ablation study for spatial-temporal transformer (STTrans)

Ablation study of structure The ablation study conducted on the Spatial-Temporal Transformer (STTrans) structure, detailed in Table 7, provides insights into the critical components influencing its performance. The variants explored include modifications to the structure components as illustrated in Fig. 4, including Convolutional Block (CNN Path), Transformer Block (Trans. Path), adjustments in the number of Transformer Block Paths, and alterations in the structure of the Spatial-Temporal Cross-Correlation (STCC). Examining Table 6, it is evident that both the Convolutional Block and Transformer Block significantly contribute to the feature extraction process.

Ablation study of spatial-temporal cross correlation In Table 7, specific analyses involve the removal of the full STCC, spatial cross-correlation (SCC), and temporal cross-correlation (TCC). The outcomes underscore the indispensability of both spatial and temporal cross-correlation mechanisms. Notably, the Convolutional Block and Transformer Block act as pivotal elements in shaping the feature representation, while the inclusion of spatial and temporal cross-correlation mechanisms enhances the model's capacity for capturing intricate spatial-temporal dependencies. These findings emphasize the separate effectiveness and interplay of components within the STTrans architecture, highlighting its holistic design for effective multi-camera depth estimation in driving scenarios.

Table 7 Ablation study of Spatial-Temporal Transformer (STTrans) on Nuscenes [26]

Methods	Abs Rel ↓	Sq Rel ↓	RMSE ↓	RMSE log ↓
w/o STCC	0.242	2.986	6.985	0.321
w/o SCC	0.238	2.956	6.893	0.318
w/o TCC	0.235	2.934	6.736	0.315
STTrans	**0.233**	**2.815**	**6.681**	**0.312**

SCC and TCC denote the spatial cross-correlation and temporal cross-correlation, respectively

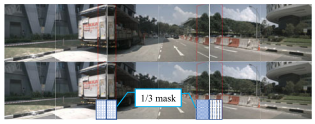

Fig. 9 Illustration of spatial overlap regions (red lines) and the applied masks (blue block)

Ablation study of overlapping proportion Overlap regions are very critical in self-supervised depth estimation in two aspects, cross-view correlation and photometric loss calculation. To explore the significance of overlapping regions, we conduct an ablation experiment by applying a mask to exclude different proportions of the overlap area. As shown in Fig. 9, we illustrate the overlap region with red lines and the applied mask with blue blocks, taking 1/3 masking in Front-Left, Front, and Front-right cameras as an example. According to the experiment results in Table 8, model performance degrades as the proportion of the overlap area decreases, verifying the value of view overlaps.

4.5.4 Ablation study for adversarial geometry regularization (AGR)

We extend our exploration to the position embedding approach within the Adversarial Geometry Regularization (AGR) module, conducting an insightful ablation study. In our analysis, we introduce a variant denoted as AGR (w/ concat), inspired by the methodology presented in the work by [71]. This variant integrates arbitrary depth maps and

Table 8 Ablation study of remained overlapping proportions, including 0%, 1/3, 2/3 and 100%, after being excluded with masks

Overlap	Abs Rel ↓	Sq Rel ↓	RMSE ↓	RMSE log ↓
0	0.248	3.028	7.013	0.336
1/3	0.240	2.962	6.906	0.320
2/3	0.236	2.952	6.738	0.317
1	**0.233**	**2.815**	**6.681**	**0.312**

Table 9 Ablation study of AGR on Nuscenes [26]

Methods	Abs Rel ↓	Sq Rel ↓	RMSE ↓	RMSE log ↓
AGR (w/ concat)	0.235	2.851	6.697	0.314
AGR	**0.233**	**2.815**	**6.681**	**0.312**

AGR (w/ concat) means directly using the concatenation of depth maps and positions

2D pixel coordinates through a concatenation process. The ablation results, outlined in Table 9, showcase the distinct performances of these approaches. Notably, our proposed depth-aware positional embedding operation demonstrates superior efficacy compared to the simpler concatenation strategy, affirming the significance of our design choice in enhancing the overall performance of the AGR module. This observation reinforces the critical role of thoughtful positional embedding strategies in optimizing depth estimation under adverse conditions within the self-supervised multi-camera context.

4.6 Qualitative evaluation results

In Fig. 10, we present the qualitative evaluation results, showcasing the effectiveness of our proposed method. The top four rows depict input images and the corresponding predicted depth maps from the Nuscenes dataset, while the bottom four rows showcase analogous results from the DDAD dataset. The visual inspection of these results underscores the capability of our method to generate high-quality depth maps. Notably, our approach excels in capturing fine contextual details and delineating clear borders around objects. This qualitative assessment provides a compelling visual demonstration of the robustness and accuracy of our depth estimation method across diverse scenes and datasets.

4.7 Model computational efficiency

To investigate the impact of model computational requirements on model performance, we compare our method with other state-of-the-art methods in Table 10. "SurroudDepth-T" and "EGA-Depth-T" are the corresponding variants of state-of-the-art methods utilizing multiple temporal frames. According to the results in Table 10, our method can achieve better performance (4% and 36% improvement on SurroundDepth and SurroundDepth-T; 1.6% and 1.7% improvement on EGA-Depth and EGA-Depth-T) with comparable computation requirements. We also test the inference time of one batch (6 images) on a single RTX 4090. According to the results, our method can achieve a better trade-off between performance and efficiency.

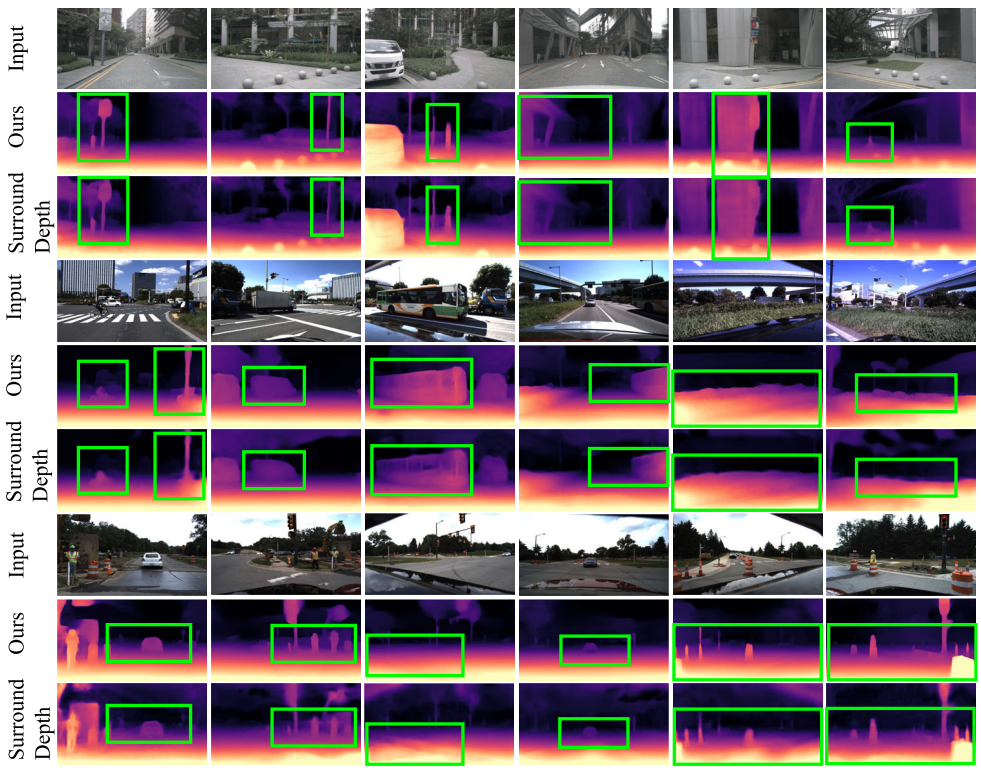

Fig. 10 Qualitative evaluation results and comparison with other state-of-the-art methods on Nuscenes (top three rows) and DDAD (bottom six rows). For each scene, we show the front, front-left, back-left, back, back-right, and front-right camera views from left to right. The predicted depth maps of our methods on both datasets display flatter ground, clearer object contour, and finer texture details, as highlighted in green boxes

Table 10 Model computational efficiency comparison

Methods	Abs Rel ↓	RMSE ↓	GFLOPs ↓	Inference time(s)↓
SurroudDepth	0.245	6.835	132.32	0.086
SurroudDepth-T	0.368	7.315	220.15	0.132
EGA-Depth	0.239	6.801	**64.94**	**0.044**
EGA-Depth-T	0.237	6.769	91.56	0.062
Ours(single)	0.235	6.736	68.66	0.048
Ours	**0.233**	**6.681**	96.80	0.068

5 Conclusion, limitations, and future directions

In this work, we introduced STViT+, a Transformer-based framework designed to tackle the complex challenges of self-supervised multi-camera depth estimation, specifically improving depth prediction in autonomous driving systems. The core of our approach lies in the Spatial-Temporal Transformer (STTrans), which captures both local spatial relationships and global contextual information, enabling more precise 3D geometry reconstruction across multiple camera views. To enhance stability in depth estimation, we incorporated the Spatial-Temporal Photometric Consistency Correction (STPCC), which effectively mitigates the impact of illumination variability across frames. Furthermore, the Adversarial Geometry Regularization (AGR) module imposes stronger spatial constraints, significantly boosting performance under challenging conditions such as nighttime driving. Extensive evaluations on large-scale datasets, including NuScenes and DDAD, confirmed the robustness and efficiency of our approach, positioning STViT+ as a leading solution for multi-camera depth estimation. Our ablation studies further underscore the contributions of each component, highlighting the comprehensive strength of the framework.

Despite these promising improvements, STViT+ has certain limitations. While STPCC efficiently handles illumination variability, it struggles in extreme conditions such as complete darkness or intense glare, where photometric cues are unreliable. These scenarios, particularly in nighttime driving, remain challenging for accurate depth estimation. Additionally, the lack of diverse, high-quality training data continues to constrain the model's generalization. Although NuScenes and DDAD offer valuable resources, the limited coverage of extreme conditions like poor lighting, nighttime driving, and adverse weather hampers broader applicability.

Looking forward, future research should focus on enhancing the robustness of depth estimation models, especially in challenging and long-tailed cases such as nighttime driving, low-light conditions, and extreme weather. Addressing these challenges will require more extensive and diverse datasets, along with comprehensive benchmarks tailored to these scenarios.

Author Contributions Conceptualization, Zhuo Chen and Haimei Zhao; data curation, Zhuo Chen and Haimei Zhao; experiments, Zhuo Chen, Haimei Zhao, and Xiaoshuai Hao; paper writing, Zhuo Chen, Haimei Zhao, Xiaoshuai Hao, Bo Yuan, and Xiu Li; investigation, Bo Yuan and Xiu Li. All authors have read and agreed to the published version of the manuscript.

Funding Open Access funding enabled and organized by CAUL and its Member Institutions.

Data Availability The datasets used in this paper were derived from sources in the public domain: https://www.cvlibs.net/datasets/kitti/ and https://github.com/TRI-ML/DDAD, reference number [17] and [18].

Declarations

Competing Interests: There is NO Competing Interest.

Ethical and informed consent for data used: This article does not contain any studies with human participants performed by any of the authors. The data used in this study were obtained from publicly available autonomous driving datasets (e.g., Nuscenes, DDAD). All data were anonymized to protect the privacy of individuals. The use of these datasets complies with the terms and conditions set by the data providers and adheres to ethical guidelines for data usage in research.

References

1. Wang Y, Chao W-L, Garg D, Hariharan B, Campbell M, Weinberger KQ (2019) Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 8445–8453

2. Dong X, Garratt MA, Anavatti SG, Abbass HA (2022) Towards real-time monocular depth estimation for robotics: A survey. IEEE Trans Intell Transp Syst 23(10):16940–16961
3. Yang X, Chen J, Dang Y, Luo H, Tang Y, Liao C, Chen P, Cheng K-T (2019) Fast depth prediction and obstacle avoidance on a monocular drone using probabilistic convolutional neural network. IEEE Trans Intell Transp Syst 22(1):156–167
4. El Jamiy F, Marsh R (2019) Distance estimation in virtual reality and augmented reality: A survey. In: 2019 IEEE International conference on electro information technology (EIT), IEEE, pp 063–068
5. Abed A, Akrout B, Amous I (2024) Deep learning-based few-shot person re-identification from top-view rgb and depth images. Neural Comput & Applic pp 1–18
6. Wei Q, Shan J, Cheng H, Yu Z, Lijuan B, Haimei Z (2016) A method of 3D human-motion capture and reconstruction based on depth information. In: 2016 IEEE International conference on mechatronics and automation. IEEE, pp 187–192
7. Li Y, Ge Z, Yu G, Yang J, Wang Z, Shi Y, Sun J, Li Z (2023) Bevdepth: Acquisition of reliable depth for multi-view 3d object detection. In: Proceedings of the AAAI conference on artificial intelligence, vol 37, pp 1477–1485
8. Zhao H, Zhang Q, Zhao S, Chen Z, Zhang J, Tao D (2024) Simdistill: Simulated multi-modal distillation for bev 3d object detection. In: Proceedings of the AAAI conference on artificial intelligence, vol 38, pp 7460–7468
9. Reda M, Onsy A, Haikal AY, Ghanbari A (2024) Path planning algorithms in the autonomous driving system: A comprehensive review. Robot Auton Syst 174:104630
10. Müller H, Niculescu V, Polonelli T, Magno M, Benini L (2023) Robust and efficient depth-based obstacle avoidance for autonomous miniaturized uavs. IEEE Trans Robot
11. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems, pp 1097–1105
12. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778
13. Eigen D, Puhrsch C, Fergus R (2014) Depth map prediction from a single image using a multi-scale deep network. In: Advances in neural information processing systems, pp 2366–2374
14. Yin W, Liu Y, Shen C, Yan Y (2019) Enforcing geometric constraints of virtual normal for depth prediction. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 5684–5693
15. Fu H, Gong M, Wang C, Batmanghelich K, Tao D (2018) Deep ordinal regression network for monocular depth estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2002–2011
16. Bhat SF, Alhashim I, Wonka P (2021) Adabins: Depth estimation using adaptive bins. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 4009–4018
17. Zhou T, Brown M, Snavely N, Lowe DG (2017) Unsupervised learning of depth and ego-motion from video. In: Proceedings of the IEEE conference on computer vision and pattern recognition
18. Godard C, Mac Aodha O, Firman M, Brostow GJ (2019) Digging into self-supervised monocular depth estimation. In: Proceedings of the IEEE international conference on computer vision, pp 3828–3838
19. Zhao C, Zhang Y, Poggi M, Tosi F, Guo X, Zhu Z, Huang G, Tang Y, Mattoccia S (2022) Monovit: Self-supervised monocular depth estimation with a vision transformer. arXiv preprint arXiv:2208.03543
20. Watson J, Mac Aodha O, Prisacariu V, Brostow G, Firman M (2021) The temporal opportunist: Self-supervised multi-frame monocular depth. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 1164–1174
21. Guizilini V, Ambruș R, Chen D, Zakharov S, Gaidon A (2022) Multi-frame self-supervised depth with transformers. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 160–170
22. Zhang S, Zhao C (2023) Dyna-depthformer: Multi-frame transformer for self-supervised depth estimation in dynamic scenes. arXiv preprint arXiv:2301.05871
23. Guizilini V, Vasiljevic I, Ambrus R, Shakhnarovich G, Gaidon A (2022) Full surround monodepth from multiple cameras. IEEE Robot Autom Lett 7(2):5397–5404
24. Wei Y, Zhao L, Zheng W, Zhu Z, Rao Y, Huang G, Lu J, Zhou J (2023) Surrounddepth: Entangling surrounding views for self-supervised multi-camera depth estimation. In: Conference on robot learning, PMLR, pp 539–549
25. Xu J, Liu X, Bai Y, Jiang J, Wang K, Chen X, Ji X (2022) Multi-camera collaborative depth prediction via consistent structure estimation. In: Proceedings of the 30th ACM international conference on multimedia, pp 2730–2738
26. Caesar H, Bankiti V, Lang AH, Vora S, Liong VE, Xu Q, Krishnan A, Pan Y, Baldan G, Beijbom O (2020) nuscenes: A multimodal dataset for autonomous driving. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 11621–11631
27. Guizilini V, Ambrus R, Pillai S, Raventos A, Gaidon A (2020) 3d packing for self-supervised monocular depth estimation. In: IEEE Conference on computer vision and pattern recognition (CVPR)
28. Chen Z, Zhao H, Yuan B, Li X (2024) Stvit: Improving self-supervised multi-camera depth estimation with spatial-temporal context and adversarial geometry regularization (student abstract). In: Proceedings of the AAAI conference on artificial intelligence, vol 38, pp 23460–23461
29. Mahjourian R, Wicke M, 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
30. Shu C, Yu K, Duan Z, Yang K (2020) Feature-metric loss for self-supervised learning of depth and egomotion. In: European conference on computer vision, Springer, pp 572–588
31. Poggi M, Aleotti F, Tosi F, Mattoccia S (2020) On the uncertainty of self-supervised monocular depth estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 3227–3237
32. Yang N, Stumberg Lv, Wang R, 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
33. Zhao H, Zhang J, Zhang S, Tao D (2022) Jperceiver: joint perception network for depth, pose and layout estimation in driving scenes. In: European conference on computer vision. Springer, pp 708–726
34. Zhao H, Bian W, Yuan B, Tao D (2020) Collaborative learning of depth estimation, visual odometry and camera relocalization from monocular videos. In: IJCAI, pp 488–494
35. Yin Z, Shi J (2018) Geonet: Unsupervised learning of dense depth, optical flow and camera pose. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1983–1992
36. Zhou H, Greenwood D, Taylor S (2021) Self-supervised monocular depth estimation with internal feature fusion. In: British machine vision conference (BMVC)
37. Zhao W, Liu S, Shu Y, Liu Y-J (2020) Towards better generalization: Joint depth-pose learning without posenet. In: Proceedings of IEEE conference on computer vision and pattern recognition
38. Klingner M, Termöhlen J-A, Mikolajczyk J, Fingscheidt T (2020) Self-supervised monocular depth estimation: Solving the dynamic

object problem by semantic guidance. In: European conference on computer vision, Springer, pp 582–600
39. Jung H, Park E, Yoo S (2021) Fine-grained semantics-aware representation enhancement for self-supervised monocular depth estimation. In: Proceedings of the IEEE international conference on computer vision, pp 12642–12652
40. Bae J, Moon S, Im S (2022) Deep digging into the generalization of self-supervised monocular depth estimation. arXiv preprint arXiv:2205.11083
41. Liu Z, Li R, Shao S, Wu X, Chen W (2023) Self-supervised monocular depth estimation with self-reference distillation and disparity offset refinement. IEEE Trans Circ Syst Vid Technol
42. Sun K, Xiao B, Liu D, Wang J (2019) Deep high-resolution representation learning for human pose estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 5693–5703
43. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S et al (2020) An image is worth 16x16 words: Transformers for image recognition at scale. In: International conference on learning representations
44. Vankadari M, Garg S, Majumder A, Kumar S, Behera A (2020) Unsupervised monocular depth estimation for night-time images using adversarial domain feature adaptation. In: Computer vision–ECCV 2020: 16th European conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVIII 16, Springer, pp 443–459
45. Wang W, Xu Z, Huang H, Liu J (2022) Self-aligned concave curve: Illumination enhancement for unsupervised adaptation. In: Proceedings of the 30th ACM international conference on multimedia, pp 2617–2626
46. Zheng Y, Zhong C, Li P, Gao H-a, Zheng Y, Jin B, Wang L, Zhao H, Zhou G, Zhang Q et al (2023) Steps: Joint self-supervised nighttime image enhancement and depth estimation. Proceedings of the IEEE international conference on robotics and automation
47. Ruhkamp P, Gao D, Chen H, Navab N, Busam B (2021) Attention meets geometry: Geometry guided spatial-temporal attention for consistent self-supervised monocular depth estimation. In: 2021 International conference on 3d vision (3DV), IEEE, pp 837–847
48. Kendall A, Martirosyan H, Dasgupta S, Henry P, Kennedy R, Bachrach A, Bry A (2017) End-to-end learning of geometry and context for deep stereo regression. In: Proceedings of the IEEE international conference on computer vision, pp 66–75
49. Sun D, Yang X, Liu M-Y, Kautz J (2018) Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 8934–8943
50. Feng C, Chen Z, Zhang C, Hu W, Li B, Lu F (2023) Iterdepth: Iterative residual refinement for outdoor self-supervised multi-frame monocular depth estimation. IEEE Trans Circ Syst Vid Technol
51. Miao X, Bai Y, Duan H, Huang Y, Wan F, Xu X, Long Y, Zheng Y (2023) Ds-depth: Dynamic and static depth estimation via a fusion cost volume. IEEE Trans Circ Syst Vid Technol
52. Shi Y, Cai H, Ansari A, Porikli F (2023) Ega-depth: Efficient guided attention for self-supervised multi-camera depth estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 119–129
53. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville, A, Bengio Y (2014) Generative adversarial nets. Adv Neural Inform Process Syst 27
54. Jing Y, Yang Y, Feng Z, Ye J, Yu Y, Song M (2019) Neural style transfer: A review. IEEE Trans Visual Comput Graphics 26(11):3365–3385
55. Xu W, Long C, Wang R, Wang G (2021) Drb-gan: A dynamic resblock generative adversarial network for artistic style transfer. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 6383–6392
56. Isola P, Zhu J-Y, Zhou T, Efros AA (2017) Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1125–1134
57. Zhu J-Y, Park T, Isola P, Efros AA (2017) Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision, pp 2223–2232
58. Zhu J-Y, Krähenbühl P, Shechtman E, Efros AA (2016) Generative visual manipulation on the natural image manifold. In: Computer vision–ECCV 2016: 14th European conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14, Springer, pp 597–613
59. Chen Z, Wang C, Yuan B, Tao D (2020) Puppeteergan: Arbitrary portrait animation with semantic-aware appearance transformation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 13518–13527
60. Chen Z, Wang C, Zhao H, Yuan B, Li X (2022) D2animator: Dual distillation of stylegan for high-resolution face animation. In: Proceedings of the 30th ACM international conference on multimedia, pp 1769–1778
61. Bousmalis K, Silberman N, Dohan D, Erhan D, Krishnan D (2017) Unsupervised pixel-level domain adaptation with generative adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3722–3731
62. Deng W, Zheng L, Ye Q, Kang G, Yang Y, Jiao J (2018) Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 994–1003
63. CS Kumar A, Bhandarkar SM, Prasad M (2018) Monocular depth prediction using generative adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp 300–308
64. Zhao C, Yen GG, Sun Q, Zhang C, Tang Y (2020) Masked gan for unsupervised depth and pose prediction with scale consistency. IEEE Trans Neural Netw Learn Syst
65. Xu Y, Wang Y, Huang R, Lei Z, Yang J, Li Z (2022) Unsupervised learning of depth estimation and camera pose with multi-scale gans. IEEE Trans Intell Transp Syst 23(10):17039–17047
66. Zheng C, Cham T-J, Cai J (2018) T2net: Synthetic-to-realistic translation for solving single-image depth estimation tasks. In: Proceedings of the European conference on computer vision (ECCV), pp 767–783
67. Zhao S, Fu H, Gong M, Tao D (2019) Geometry-aware symmetric domain adaptation for monocular depth estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 9788–9798
68. Sun Q, Yen GG, Tang Y, Zhao C (2023) Learn to adapt for self-supervised monocular depth estimation. IEEE Trans Neural Netw Learn Syst
69. Gaidon A, Wang Q, Cabon Y, Vig E (2016) Virtual worlds as proxy for multi-object tracking analysis. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4340–4349
70. Wu Z, Wu X, Zhang X, Wang S, Ju L (2019) Spatial correspondence with generative adversarial network: Learning depth from monocular videos. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 7494–7504
71. Wang K, Zhang Z, Yan Z, Li X, Xu B, Li J, Yang J (2021) Regularizing nighttime weirdness: Efficient self-supervised monocular depth estimation in the dark. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 16055–16064
72. Lee Y, Kim J, Willette J, Hwang SJ (2022) Mpvit: Multi-path vision transformer for dense prediction. In: Proceedings of the IEEE/CVF

conference on computer vision and pattern recognition, pp 7287–7296
73. Wang Z, Bovik AC, Sheikh HR, Simoncelli EP et al (2004) Image quality assessment: from error visibility to structural similarity. IEEE Trans Image Process 13(4):600–612
74. Godard C, Mac Aodha O, Brostow GJ (2017) Unsupervised monocular depth estimation with left-right consistency. In: Proceedings of the IEEE conference on computer vision and pattern recognition
75. Pizer SM, Amburn EP, Austin JD, Cromartie R, Geselowitz A, Greer T, Haar Romeny B, Zimmerman JB, Zuiderveld K (1987) Adaptive histogram equalization and its variations. Comput Vision, graph Image Process 39(3):355–368
76. Dijk Tv, Croon Gd (2019) How do neural networks see depth in single images? In: Proceedings of the IEEE/CVF international conference on computer vision, pp. 2183–2191

Zhuo Chen received the B.Eng. degree and the Ph.D. degree in automation from Tsinghua University, China in 2017 and 2024. His research interests include computer vision, image generation, video generation, multi-modal large model.

Haimei Zhao received the Ph.D. degree in computer science from the University of Sydney, Australia in 2024, and the M.Phil. degree from Tsinghua University, China in 2020. She is currently a Postdoc research fellow in at the University of Sydney, Australia.

Her research interest is computer vision, autonomous driving, AI for biomedical science, and digital health.

Xiaoshuai Hao obtained his Ph.D. from the Institute of Information Engineering at the Chinese Academy of Sciences in 2023. He is currently a researcher specializing in embodied multimodal large models at the Beijing Academy of Artificial Intelligence. His research interests encompass multimedia retrieval, multimodal learning, and embodied intelligence.

Bo Yuan (Senior Member, IEEE) received the B.E. degree in Computer Science from the Nanjing University of Science and Technology, Nanjing, China, in 1998, and the M.Sc. and Ph.D. degrees in Computer Science from the University of Queensland (UQ), St Lucia, QLD, Australia, in 2002 and 2006, respectively. From 2006 to 2007, he worked as a Research Officer on a project funded by the Australian Research Council at UQ. From 2007 to 2021, he was a Faculty Member at the Division of Informatics, Tsinghua Shenzhen International Graduate School, Shenzhen, China, where he served as a Lecturer (2007 - 2009) and Associate Professor (2009 - 2021). During this period, he also held the role of Deputy Director of the Office of Academic Affairs (2013 - 2020). He has authored more than 140 papers in refereed international conferences and journals. His research interests include data science, evolutionary computation, and intelligent decision making.

Xiu Li (Member, IEEE) received the Ph.D. degree in computer integrated manufacturing from Nanjing University of Aeronautics and Astronautics in 2000. From then to 2002, she was a Postdoctoral Fellow with the Department of Automation, Tsinghua University, Beijing, China. From 2003 to 2010, she was an Associate Professor with the Department of Automation, Tsinghua University, Beijing, China. Since 2016, She has been a Full Professor at Shenzhen International Graduate School, Tsinghua University. In recent years, she has authored more than 100 papers in peer-reviewed journals and conferences. Among them, more than 20 papers have been published in top journals and conferences, such as the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, IEEE Transactions on Neural Networks and Learning Systems, and CVPR, ICCV. Her research interests include computer vision, pattern recognition, and image processing.

Authors and Affiliations

Zhuo Chen[1] · Haimei Zhao[2] · Xiaoshuai Hao[3] · Bo Yuan[4] · Xiu Li[1]

✉ Haimei Zhao
hzha7798@uni.sydney.edu.au

Zhuo Chen
z-chen17@tsinghua.org.cn

Xiaoshuai Hao
xshao@baai.ac.cn

Bo Yuan
boyuan@ieee.org

Xiu Li
li.xiu@sz.tsinghua.edu.cn

1 Shenzhen International Graduate School, Tsinghua University, Shenzhen 518055, Guangdong, China

2 School of Computer Science, The University of Sydney, Sydney 2008, NSW, Australia

3 Beijing Academy of Artificial Intelligence, Beijing, China

4 School of Electrical Engineering and Computer Science, The University of Queensland, Brisbane, QLD 4072, Australia