IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 7 November 2023, accepted 30 November 2023, date of publication 4 December 2023,
date of current version 13 December 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3339152

RESEARCH ARTICLE

SAU-Net: Monocular Depth Estimation Combining Multi-Scale Features and Attention Mechanisms

WEI ZHAO, YUNQING SONG, AND TINGTING WANG
China College of Vehicle and Traffic Engineering, Henan University of Science and Technology, Luoyang 471023, China
Corresponding author: Wei Zhao (zhaowei@haust.edu.cn)

ABSTRACT Monocular depth estimation technology is widely utilized in autonomous driving for sensing and obstacle avoidance. Recent advancements in deep-learning techniques have resulted in significant progress in monocular depth estimation. However, monocular depth estimation is mainly optimized for the luminosity error of pixels, mostly disregarding the related problems of result ambiguity and boundary artifacts in the image. To address these issues, we developed an improved network model called SAU-Net. The superposition of excessive convolutional layers in conventional convolution networks impairs the network's timeliness and results in the loss of primary information. Therefore, we propose a convolution-free stratified transformer as an image feature extractor at the network's coding end, which limits self-attention to innumerable windows and leverages sliding windows for characterization to reduce the network delay. This study also addresses the issue of critical information loss. We connect each feature map directly to another from a different scale. In addition, an attention module is introduced to focus on the effective features, which increases the amount of target information in the depth map. We employ the gradient loss function during the training stage to improve the segmentation accuracy of the network and the smoothness of the output image. Training and testing were conducted using the KITTI dataset. To ensure the robustness of the algorithm in practical applications, we also validated the algorithm using a campus dataset that we collected. The experimental results indicated that the accuracy of the algorithm was 89.1%, 96.4%, and 98.5% under three proportional thresholds. The estimated depth map was clear in details and edges, with fewer artifacts.

INDEX TERMS Autonomous driving, monocular depth estimation, SAU-net, stratified transformer.

I. INTRODUCTION

Monocular depth estimation is becoming increasingly important in autonomous driving [1]. Accurately predicting the depth information of the vehicle's surrounding environment is crucial for various autonomous driving functions, including autonomous navigation, obstacle avoidance, and automatic parking [2], [3]. Traditional autonomous driving systems mostly use light detection and ranging (LiDAR) datasets or stereovision technology to achieve depth perception [4]. These methods often require expensive equipment and incur high energy costs. Furthermore, they can only be used in particular environments. In contrast, monocular depth estimation technology can predict the depth using the images captured by a single camera, improving the environmental awareness and decision-making ability of the autonomous driving system without other hardware. In addition, the accuracy and robustness of monocular depth estimation technology have significantly improved [5]. Studies have indicated that in particular scenarios, it can outperform traditional stereovision methods [6]. Therefore, it is necessary to study algorithms for monocular depth estimation.

Monocular depth estimation using deep-learning techniques can be categorized into supervised and self-supervised methods. Regarding supervised monocular depth estimation, the Eigen team [7] made significant contributions by utilizing depth maps to train models. They developed a convolutional neural network (CNN) structure with coarse and fine scales, achieving accurate estimation results and pioneering deep learning in monocular image depth estimation. Li et al. [8] developed a deep network for multilayer conditional random fields (CRFs). Their approach involved a two-stage network

The associate editor coordinating the review of this manuscript and approving it for publication was Jiankang Zhang.

for depth map estimation and refinement. Superpixel technology was applied to the input image in the first stage, with image patches extracted around these super pixels. The depth map was refined, and the superpixel depth map was modified to the pixel level in the second stage using a multilayer CRF. This significantly increased the resolution of the output depth map. Qi et al. [9] employed two networks to estimate depth maps and the surface normal from a single depth image. These two networks facilitated depth-to-normal and normal-to-depth conversion, resulting in improved depth maps and surface normal accuracy. Summarily, supervised depth estimation methods require large annotated datasets for training models. However, high-resolution public labeled datasets require substantial equipment and intensive labor, while commonly used datasets for sparse-depth information labeling remain prevalent. Consequently, supervised deep-learning methods are unsuitable for application scenarios that require dense estimation.

In contrast to supervised depth estimation, the self-supervised learning method of monocular depth estimation involves learning depth information directly from geometric constraints and does not require numerous densely labeled datasets. Typically, stereo-paired images or monocular image sequences are used for training, transforming the depth estimation problem into an image reconstruction problem. Network-predicted binocular images and disparity maps are used to achieve self-supervised monocular depth estimation. Garg et al. [10] proposed a self-supervised monocular depth estimation network based on stereovision. The left view in the binocular image is considered as the input; in addition, the corresponding disparity map is predicted by a CNN. Subsequently, the right view is input to reconstruct the left view according to the cross-reconstruction principle, and the composite loss function, including photometric reconstruction loss and depth gradient loss, is used to constrain the network weight. Finally, the best-predicted disparity map is obtained. However, they employed Taylor expansion for linear optimization in deformation reconstruction. The result of this method is not completely differentiable; thus, the model may fall into a local optimal solution, which makes the prediction result not ideal. To overcome this difficulty, Godard et al. [11] introduced differentiable difference functions, proposed left–right consistency constraints to train self-supervised networks, and reconstructed left–right views. In addition, they optimized the loss function, proposing surface matching losses and parallax smoothing losses. Their experiment confirmed that the addition of the new loss function increases the accuracy of each view prediction. Watson et al. [12] introduced depth hints to alleviate the impact of reprojection in the reconstruction process. These hints increase the current luminosity loss function and play a substantive role in training the current leading self-monitoring model. For the problem of occlusion and artifacts, the introduction of depth hints may help to identify and deal with partially occluded situations. However, there are still difficulties in highly complex or dense occlusion situations, and obtaining high-quality depth cues may require complex camera equipment or additional sensors, which may be expensive and unsuitable for certain applications.

Thus, we developed a self-supervised monomial depth estimation model called SAU-net based on StratifiedTransformer and an improved U-Net [13] framework. By improving the algorithm and optimizing the loss function, the depth estimation can be improved without the use of expensive equipment. The problems of occlusion and artifacts in the image are solved to the greatest extent possible, leading to significantly enhanced accuracy in the predicted depth map. Experimental results indicated that the algorithm can highlight the edge of the depth map and play a role in blocking. The contributions of this study are as follows:

1. Using a stratified transformer-based layered feature extractor to replace the commonly used depth residual network (Res-Net) reduces the number of layers in the network, expands the receptive field, and provides greater flexibility in extracting feature maps of different sizes.

2. An improvement was made to U-Net, in which the layered feature map is connected to the decoding layer employing a jump connection, and the attention mechanism is introduced to focus attention on the primary information in the feature map, which alleviates the over-segmentation and increases the segmentation precision.

3. A loss function based on luminosity reprojection and automatic masking loss, supplemented by gradient loss, was constructed. In addition, it is employed in the depth map to reduce boundary artifacts and highlight detailed features.

4. The proposed algorithm was evaluated using the KITTI dataset and a self-collected campus dataset in comparison with many advanced algorithms.

The remainder of this paper is organized as follows. Section II briefly reviews previous studies involving the attention mechanism, transformer, and U-Net. In Section III, we describe the network model of SAU-net. The experiments and their results are presented in Section IV. Finally, Section V concludes the paper.

II. RELATED WORK

In this section, we introduce attention mechanisms, transformers, and U-Net.

A. ATTENTION MECHANISM

The attention mechanism focuses on local information. It was initially used in the field of natural language processing (NLP) to extract contextual semantic information for improving model performance [14], and since it was introduced into the field of vision, it has been widely used in image classification, object detection, semantic segmentation, face recognition, and other tasks [15], [16]. Recently, several studies have been performed on the use of the attention mechanism in the field of vision. Hu et al. [17] proposed SE Net, focusing on the channel part and recalibrating the channel feature response by superimposing an SE block, thus improving the representation capability of the traditional CNN

network and creating a precedent for channel attention. Yang et al. [18] proposed the concept of the Gated Signal (GCT) on this basis. They replaced the FC layer in SE Net with a normalized module in the processing of channel-wise embeddings in GCT, modeled the feature relationship between channels, and increased the efficiency of information collection. However, the addition of excessive modules increases the network complexity. To reduce the model complexity while ensuring the quality of the output results, Wang et al. [19] proposed an efficient channel attention module (ECA), which replaces conventional dimensionality reduction with inter-channel interactions. ECA transforms the multilayer perceptron (MLP) modules in the SE block into one-dimensional convolution, which significantly reduces the number of parameters. This method addresses the problem of the increase in computation due to module superposition.

The attention mechanism can aid the model by assigning different weights to various parts of the input image [20], extracting critical information, and allowing the model to make more accurate judgments without incurring additional computational and storage costs [21], [22]. Because of these advantages, this method has a strong generalization ability. In the present study, the attention-mechanism module is introduced in the compressed channel part to make it focus on the detailed features, which can increase the prediction accuracy without excessively increasing the calculation amount and ensure the real-time performance.

B. TRANSFORMER

A transformer is a neural-network structure based on an attention mechanism that can perform sequence-to-sequence conversion without convolutional or cyclic layers. The transformer first achieved major success in the NLP field and was subsequently introduced into the field of computer vision. The vision transformer (VIT) was proposed by Dosovitskiy et al. [23] as the first pure transformer structure for image processing. The algorithm proves that the pure attention network is superior to the most advanced CNN in image classification. Subsequently, numerous tasks based on the VIT network architecture have emerged. Lin et al. [24] proposed a VIT controller using VIT as the backbone network for adaptive fusion and feature selection in semantic segmentation. Rizzoliet al. [25] injected depth information from multiple stages into a segmentation module based on the VIT architecture for passive semantic segmentation. Deng et al. [26] combined linear attention with a U-Net network to obtain a T-former for inpainting tasks.

Although the application and development of the transformer has brought revolutionary improvements to the field of computer vision, the conventional transformer has a lower computing efficiency than the StratifiedTransformer proposed in this study. The transformer necessitates the calculation of every element of the entire sequence once, resulting in a substantial increase in computational cost when the input image is extremely large. The Stratified Transformer uses a sliding-window mechanism to split large images into smaller pieces and exchange information between these pieces, avoiding the computational burden of processing the entire image simultaneously.

C. U-NET

U-Net is a type of CNN with a simple and symmetric structure. Different from the first two hot in the NLP field [27], U-Net is widely used in medical image segmentation. Researchers have developed numerous U-Net-based network models [28], [29] to improve the quality of medical images and the accuracy of automated medical systems. In depth estimation and optimization models based on U-Net have achieved significant breakthroughs. Liang et al. [30] constructed an attention feature fusion module based on the original U-Net, which is called AFFM and consists of channel and spatial attention modules. The improved U-Net model increased the segmentation accuracy of the image. Cui et al. [31] fused an improved U-Net network with a pose network to construct a lightweight depth estimation model called MonoDA for unmanned agricultural vehicles to obtain the depth information of the surrounding environment. This model achieves a good balance between accuracy and computation time. Duong et al. [32] constructed a depth estimation model called UR-Net, which adds attention to the decoder and replaces the transmission block in the conventional U-Net with spatial pyramid pool blocks (ASPP). In experiments, their model outperformed U-Net with regard to the error rate and accuracy.

U-Net has numerous advantages in the field of segmentation—particularly for the processing of detailed features. However, because of problems in the convolutional structure, such as translation invariance and an insufficient ability to capture long-term dependence, it cannot capture sufficient image features when processing large datasets. The transformer network has a large receptive field that can address these two problems well; however, it has shortcomings in processing fine-grained information, resulting in inaccurate positioning. Therefore, the combination of the two can produce better results, which is explained in detail in Section III.

III. SAU-NET

Starting with the construction of the network model, this section introduces the Stratified Transformer, the improved AU-Net network architecture, and the selection of the loss function to illustrate the proposed self-supervised monocular depth estimation method.

A. DEEP NETWORK FRAMEWORK

The self-supervised depth estimation strategy used in Monodepth2 [33] is employed in this study to estimate the depth value and relative pose of monocular image sequences. The overall network structure comprises a deep network and a camera pose network. The SAU-net section uses an architecture that combines the improved AU-Net network with the

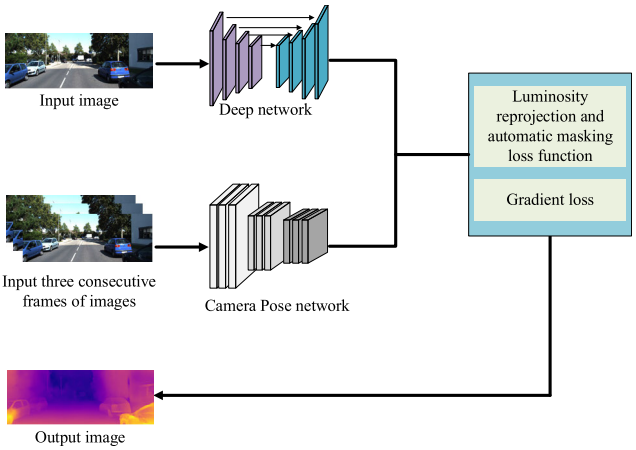

FIGURE 1. Network architecture.

Stratified Transformer. The overall network is an end-to-end structure, with the Stratified Transformer feature extractor as the network's coding part and the improved U-Net network as the decoding part. A single-frame image at a particular moment is taken as the input of the deep network, and this moment and the image sequence of two adjacent frames are taken as the inputs of the pose network. The pose network is a standard CNN, the encoding part is Resnet34 [34], and the decoding part is an AU-Net network. The structure of Resnet34 is as follows: Conv (3 × 3)–BN–RELU–Conv (3 × 3)–BN, where ''BN'' represents a batch normalization layer and ''RELU'' represents the rectified linear unit activation function. In the whole model, the convolution step is set to 2, and the output part of the convolution kernel adopts the RELU activation function. The loss function has a structure of one primary loss and one auxiliary loss. The primary loss follows the double loss function in Monodepth2, and the auxiliary loss is the gradient loss [35]. The proposed network architecture is depicted in Fig. 1, with images from the KITTI dataset [36].

B. STRATIFIED TRANSFORMER STRUCTURE

The Swin transformer [37] uses a hierarchical construction method similar to that of CNNs. The Stratified Transformer structure used in this study adopts several architecture designs employed in the Swin transformer. In contrast to VIT, which generates a single feature map of input features with a similar resolution, the Stratified Transformer generates multiple feature maps with different resolutions. We first convolve the original input data to obtain the feature map and then divide the feature map into small patches. Because the feature map's size and resolution will be halved after the downsampling module, the patch and network architecture pattern module in Fig. 1 splice features to reduce the space size by a factor of 2 and expand the feature dimension by a factor of 4. As illustrated in Fig. 2(a), four hierarchical blocks are used to generate feature maps of four different resolutions, i.e., 1/2, 1/4, 1/8, and 1/16.

The sliding-window mechanism allows the Swin transformer to increase the extraction efficiency for image information features; however, segmentation errors or missing segmentation may still occur when feature mapping segmentation is performed for the first time. We reclassified each pixel and utilized a single Swin block to optimize the original segmentation map, mitigating such problems. The Swin block is a self-attention module and an essential component of the network structure depicted in Fig. 2(b). We replaced the multi-head self-attention (MSA) mechanism in the transformer with window multi-head self-attention (W-MSA), and each window counts only its attention. This simplifies the calculation to linear complexity, significantly reducing the amount of computation. Furthermore, the sliding window solves the problem of information isolation induced by the independence of each window. It can increase the recognition accuracy for each pixel and lead to a space dependence between adjacent regions of the image.

A Swin block consists of W-MSA, SW-MSA, and an MLP. Each module is preceded by a layer normalization (LN) layer, and the remaining connections are applied after each module. The MLP module consists of two layers with a Gaussian Error Linear Units (GELU) nonlinear activation function. Multiresolution feature maps are generated by computing successive stratified transformer blocks, as follows:

$$\hat{d}^n = W - MSA[LN(d^{n-1})] + d^{n-1} \quad (1)$$

$$d^n = MLP[LN(\hat{d}^n)] + \hat{d}^n \quad (2)$$

$$\hat{d}^{n+1} = SW - MSA[LN(d^n)] + d^n \quad (3)$$

$$d^{n+1} = MLP[LN(\hat{d}^{n+1})] + \hat{d}^{n+1} \quad (4)$$

where $\hat{d}^n$ and d^n represent the characteristic outputs of (S) W-MSA and MLP in module n, respectively; similarly, $\hat{d}^{n+1}$ and d^{n+1} represent the characteristic outputs of $n+1$; and d^{n-1} represents the corresponding characteristic output of layer n–1. According to these features of the Stratified Transformer, the jump join strategy generates intensive feature predictions, resulting in a model with few parameters and low computational costs.

C. AU -NET STRUCTURE

The original U-Net network is a lightweight full CNN, which includes a contraction path and an expansion path, which correspond to the coding and decoding parts, respectively, of the network. Its advantage is that the network structure is simple and efficient, and good training results can be obtained even in the case of a small data volume.

However, the skip connection module in the U-Net network is inadequate, and its function is overly simple to improve the connectivity among the internal features of the deep decoder. The layered image features of four different resolutions (1/2, 1/4, 1/8, and 1/16) are passed from the encoder to each stage of the deep decoder via the jump connection between the encoder and the deep decoder to overcome the shortcomings of this section. Subsequently, the attention module is used to locate the most critical information from the compressed

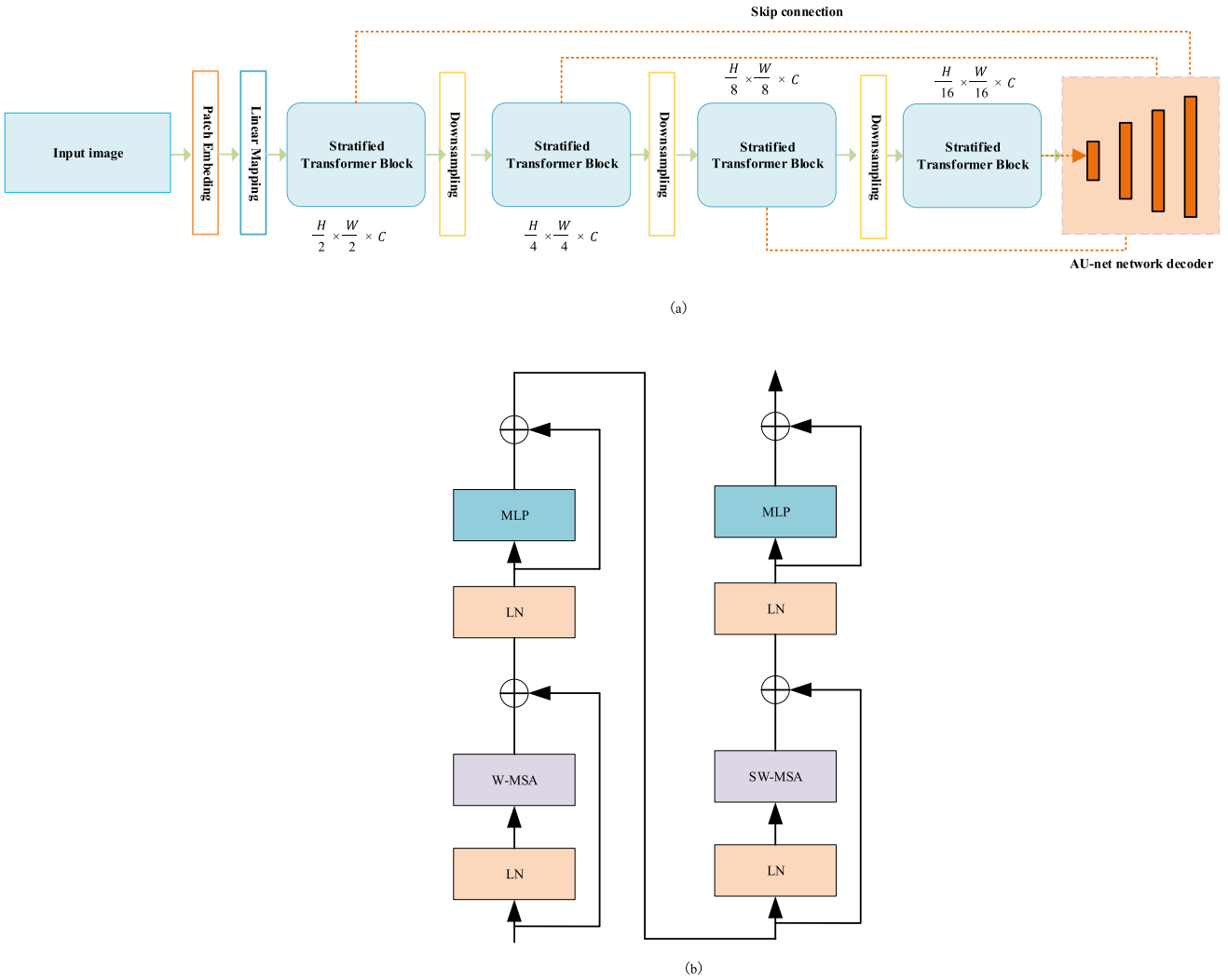

FIGURE 2. Stratified transformer structure diagrams. (a) network structure diagram; (b) block structure diagram.

channel, reducing the degree of loss of target features and significantly improving the performance of the network. The network architecture is illustrated in Fig. 3.

Once the layered image features are passed to the decoder, they are projected onto the same dimensional space; subsequently, the features of different stages are joined using the method of adding elements. Subsequently, the cascade path is interpolated bilinearly from the bottom up:

$$x_s = \sum_{k=1}^{s} (f_k) \uparrow \tag{5}$$

where f_k and $() \uparrow$ represent the projection features of the stage k and the two-line interpolation, respectively. $\sum_{k=1}^{s} (f_k) \uparrow$ refers to the superposition of k from 1 to the s stage. The nonlinear feature mapping module is used to map the spliced features to the same dimensional space, and the aliasing effect is alleviated by bilinear upsampling. Double upsampling is performed to improve the versatility of the network.

Convolutional layers of 3×3 and 1×1 are used in the depth decoder part, whereas a single convolutional layer of 3×3 and 1×1 is used in the projection module and feature mapping module, respectively. In the first layer of the depth decoder, the crop operation is performed when the skip connection is executed. First, the size of the feature map is reduced to half the original size, crucial features are captured through the attention mechanism, and a $2\times$ upsampling operation is performed to restore the size of the feature map. Finally, the feature map of the restored size is combined with the layered image features in the same dimension to form a feature map with a $2\times$ size. The second and subsequent convolutional layers have no cropping or upsampling operations.

D. LOSS FUNCTION

The loss function constructed in this study consists of a primary loss function and an auxiliary loss function. Considering the network structure used for extracting image feature information, we decompose the whole network optimization into two subproblems. The primary loss function is used to eliminate the effects of object movement and occlusion. The auxiliary loss function solves the problems of missing local detail and deep holes in the depth map.

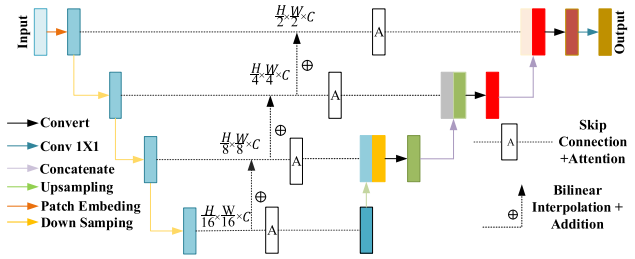

FIGURE 3. AU-Net architecture.

First, we use the double loss function in Monodepthh2 as our primary function L_1, which is obtained by multiplying the luminosity reprojection function L_a and the automatic masking loss μ. The photometric reprojection is expressed as follows:

$$L_a = \sum_{t'} pe(H_t, H_{t'\to t}) \tag{6}$$

$$H_{t'\to t} = H_{t'}[proj(Z_t, D_{t\to t'}, k)] \tag{7}$$

where H_t denotes the target image, $H_{t'}$ denotes the source view, and $H_{t\to t'}$ represents the part where the reprojection error of the abnormal image is smaller than that of the original image. In stereo matching, $H_{t'}$ is the second image of H_t but the pose of $H_{t'}$ is not known in the monocular sequence; thus, it is necessary to calculate $H_{t\to t'}$ at the two-dimensional level. *pe* represents the luminosity reprojection error, which is composed of the loss function L' and *SSIM* (structural similarity index measure), as given by (8); Z_t denotes the generated depth map corresponding to H_t; K is the internal function; *proj*() is the projection coordinate of Z_t at the two-dimensional level, and $D_{t\to t'}$ represents the camera pose of each source view relative to the target image.

$$p_e(H_a, H_b) = \partial.\frac{1 - SSIM(H_a - H_b)}{2} + L' \tag{8}$$

$$L' = (1 - \partial)||H_a - H_b|| \tag{9}$$

Here, $(1-\partial)||H_a-H_b||$ is the loss function L', which is used to calculate the absolute-value error of pixels in the target image and the estimated image, that is, $\partial = 0.85$. The value refers to the setting of Godard et al. [11]. *SSIM* indicates the similarity between two images. To reduce projection errors and the number of artifacts in depth maps, we select the minimum error of luminosity reprojection for calculation. Therefore, the final expression of luminosity reprojection is

$$L_a = \min_{t'} pe(H_t - H_{t'\to t}) \tag{10}$$

Using the automatic masking loss function μ, the main loss L_1 can be determined as

$$\mu = [\min_{t'} pe(H_t, H_{t'}) < \min_{t'} pe(H_t, H_{t'\to t})] \tag{11}$$

where $\mu = 1$ when the conditions within the parentheses are met; otherwise, $\mu = 0$. [] denotes the Iverson parentheses. Because the reprojection error of $H_{t'\to t}$ exceeds that of H_t, μ is used to ignore the pixel loss of the original luminosity reprojection error. Finally, the primary loss function is

$$L_1 = \mu.L_a \tag{12}$$

The main reason why we designed the auxiliary function L_2 is that when the loss function is only the primary loss, it is easy to ignore a large number of background and object detail pixels; thus, the convergence is extremely fast. However, it is easy to have large gradient-value changes, leading to unstable training, and the output results also have the problem of missing local details. The gradient loss function can enhance local details—particularly at the depth boundary—and make the gradient decline more smoothly by measuring the similarity between the ground truth (GT) and the model prediction. Through the aforementioned analysis, we added gradient loss as a supplement during the training process, with the expression

$$L_2 = \frac{1}{N}\sum_{i}^{N} |g_{h,i} - m(g^*)_{h,i}| + |g_{v,i} - m(g^*)_{v,i}| \tag{13}$$

Here, N represents the total number of pixels; $g_{h,i}$ and $m(g^*)_{h,i}$ represent the i^{th} gradient value in the depth map and the interpolated GT value in the horizontal direction, respectively; $g_{v,i}$ and $m(g^*)_{v,i}$ represent the corresponding values in the vertical direction; and || is the absolute-value symbol. In calculating the loss, we only consider the differences between the predicted and GT values in the horizontal and vertical directions.

The loss function used in this study is

$$L = \alpha L_1 + \beta L_2 \tag{14}$$

where α and β represent the equilibrium factors of L_1 and L_2, respectively, whose values are analyzed via a trial-and-error method in this study. During adjustment, the batch size is fixed, the proportion of parameter values is increased and reduced, the training curve of the loss rate and accuracy is examined, and the optimal value is selected. Experiments indicate that the value of the α/β ratio gradually converges during the process of increasing from small to large. When the value of ratio is too large, the convergence is too fast, resulting in inadequate training and overfitting. In the experiments, the convergence effect is best when the parameter ratio is in the range of 95–105; thus, α and β are set as 10 and 0.1, respectively, in this study. The training curves presented in Figs. 4 and 5 were obtained when $\alpha = 10$ and $\beta = 0.1$.

Different loss functions were added to SAU-net for training. As illustrated in Fig. 4, with an increase in the training period, the gradient of the loss function of one primary loss and one auxiliary loss was stable when the rate of convergence was almost the same as that for the single primary loss function. As indicated by Fig. 5, the composite loss function also had excellent performance with regard to accuracy. In the first 10 epochs, the accuracies of the two were almost identical. Following the 20$^{\text{th}}$ epoch, the accuracy of the proposed loss function significantly exceeded that of the single loss

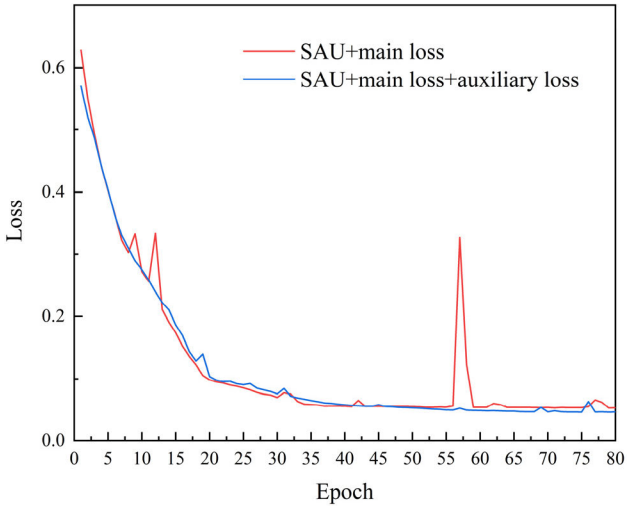

FIGURE 4. Loss rate curves for the SAU-net training set.

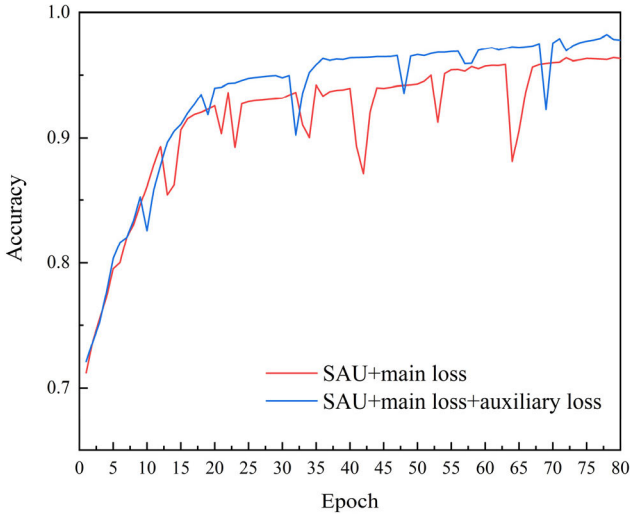

FIGURE 5. Accuracy curves for the SAU-net training set.

function. Additionally, the volatility of the curve was lower with regard to data stability.

IV. EXPERIMENTAL PROCESS AND RESULTS

We evaluated the prediction performance of the SAU-net algorithm on the KITTI dataset and a self-collected campus dataset. We conducted several ablation experiments for optimizing the loss function and algorithm structure. The algorithm was compared with conventional monocular depth estimation algorithms. Additionally, the original algorithm was compared with the optimized algorithm in the same scenario, and the effectiveness of the proposed algorithm was verified.

A. ACQUISITION OF DATASETS

We conducted training using the KITTI dataset, which is the world's largest computer-vision algorithm evaluation dataset for automatic-driving scenarios. After removing the static frames in the monocular sequence, we used the method suggested in Eigen to split the fusion dataset, selecting 39915 images for training and 4525 for testing. Furthermore, to verify the robustness of the algorithm in practical applications, we collected 200 images on campus and added them to the KITTI dataset for evaluating the performance of the model. The resolution of the images was the most widely used resolution for evaluating the depth estimation performance on the KITTI dataset (640×190).

For collecting campus data, we used a monocular camera produced by BYD Han the original factory, which has the characteristics of small distortion and clear images. A BYD HanDM-I car was used as the mobile platform for the experiment. To ensure the accuracy of the camera's field of view and data acquisition, we installed the camera at the center of the car's front windshield, with the lens kept horizontal and 120 cm above the ground. The car was moving at a speed of 20 km/h during the data collection.

B. IMPLEMENTATION DETAILS

To expand the dataset and prevent overfitting, data enhancement operations such as translation and inversion were performed on three adjacent input images. The network model used in this experiment is based on the Pytorch framework, and four Nvidia RTX 3090 graphics processing units (GPUs) were used for training. The model used the Adam optimizer [38], and the parameters were set to $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The proposed model had an initial learning rate of 1×10^{-4}, 80 training epochs, and a batch size of 12 per epoch. Following the 20^{th} epoch, the loss rate decreased slowly to nearly 0. To solve this problem, we attenuated the learning rate to one-tenth of the original at 60 epochs after training. To reduce the training time of the network and improve the overall performance of the model, we trained the model using the pre-training weights on ImageNet-1k [39].

C. EVALUATION INDICES

The evaluation indices used in this study were commonly used error indices and threshold accuracy indices (δ). Among these, the quantitative standard of the error indicators were the absolute relative error (Abs Rel), root-mean-square error (RMS), logarithmic root-mean-square error (log RMS), and squared relative error (Sq Rel). They are used to measure the error between the prediction result and the real depth. The smaller the value of the error index, the better the experimental result. The threshold accuracy index δ was used to measure the accuracy of the model; a larger value corresponded to a higher prediction accuracy. The advantage of this index is that it directly reflects the accuracy of the prediction results. The corresponding formulas are as follows:

$$\%\ of\ dis.t.max(\frac{D_i}{D_i^*}, \frac{D_i^*}{D_i}) = \delta < thr \tag{15}$$

$$Abs\ Rel = \frac{1}{N}\sum_{i=1}^{N}\frac{|D_i - D_i^*|}{D_i^*} \tag{16}$$

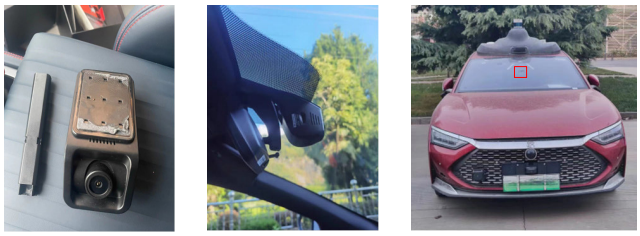

FIGURE 6. Campus data acquisition equipment.

$$RMS = \sqrt{\frac{1}{N}\sum_{i=1}^{N} |D_i - D_i^*|^2} \quad (17)$$

$$logRMS = \sqrt{\frac{1}{N}\sum_{i=1}^{N} |lgD_i - D_i^*|^2} \quad (18)$$

$$Sq\ Rel = \frac{1}{N}\sum_{i=1}^{N} \frac{|D_i - D_i^2|^2}{D_i^*} \quad (19)$$

where D_i represents the predicted depth value of the i^{th} pixel, D_i^* represents the true depth value corresponding to the i^{th} pixel, and N represents the total number of pixels. The threshold was $thr = 1.25^i$, with $i = 1, 2, 3$.

D. ANALYSIS OF RESULTS

First, we performed ablation experiments on two different loss functions to verify their impact on the output of the algorithm. When we replace the backbone of the original convolutional network with the stratified transformer, we discover that when only the loss function of a single structure is used to train the network, the output depth map appears similar to wrinkles in the top area of the image. For clarity, this phenomenon is indicated by white boxes in Fig. 7. The broad receptive field of the stratified transformer allows the network to extract a large amount of image information in an extremely short time, while the original loss function often ignores many background pixels in the calculation process, resulting in a loss of image depth information. Later, we incorporated gradient loss as an auxiliary loss to alleviate the negative impact caused by the imbalance between foreground and background areas, which addressed this problem.

Subsequently, we selected five scenarios from the KITTI dataset: pedestrian, vehicle, crowded, open, and other working conditions. For these scenarios, we compared our results with the results of other advanced algorithms, as shown in Fig. 8.

Fig. 8 shows a comparison of the results of the proposed method, Monodepth2 [33], and Midas [40]. Monodepth2 [33]and Midas [40] are classical algorithms for monocular depth estimation. As indicated by the figure, the depth map produced by our method was more precise in both distant and close-range regions. In the distant part, for the other two algorithms, there were black areas in the image that could not be recognized, while the estimated depth map showed the edge contour of the distant part. Our algorithm was more effective for restoring edges and details.

To verify the reliability of the SAU-net algorithm and its robustness in practical applications, we also collected scenes on the campus as test data. For the campus data, we selected five scene diagrams under different lighting conditions, such as cloudy day, sunny day, and shadow. Similar to the above experiments, we compared the results of SAU-net with those of other advanced algorithms.

Fig. 9(a) presents an image collected on a cloudy day, Figs. 9(b)–(d) present images of sunny scenes, and Fig. 9(e) presents an image of a shadow scene. As illustrated, from left to right, the image estimated by Monodepth2 [33] had the worst effect in the results of horizontal comparison, and distant scenes could not be displayed in the depth map, resulting in the problem of depth loss, which is particularly obvious in Figs. 9(d) and (e). Compared with Monodepth2 [33], Midas [40] paid more attention to the segmentation effect of the edge parts of objects; thus, the objects in the scene were clearly displayed in the depth map. However, for the detailed features of objects and the prediction of distant scenes, the effect was not ideal, and there was still a problem of depth loss. Compared with the other two networks, SAU-net compensated for the problem of depth loss in the details and distant features of images and significantly increased the prediction accuracy. For example, the electric bicycle in Fig. 9(b), the tree in the distant shadow in Fig. 9(d), and the white car in Fig. 9(e) are all clearly represented in our depth map. As indicated by the longitudinal comparison, the results of the proposed network were slightly worse on cloudy days, and the electric bicycle in Fig. 9(a) was less visible in the depth map, which was caused by the inability of the monocular camera to collect good experimental data under poor lighting conditions.

Thus, we evaluated the depth maps estimated by different algorithms and verified the effectiveness of SAU-net from a subjective viewpoint. Next, to confirm the robustness of the proposed algorithm, we compared it with relevant algorithms proposed in recent years using the KITTI dataset. We analyzed the effectiveness of the proposed algorithm from an objective viewpoint using the aforementioned evaluation indices. The results are presented in Table 1.

As indicated by Table 1, among the previously proposed monocular depth estimation methods based on supervised learning, the deep ordered regression network proposed by Fu et al. [44] performed the best, and it has been significantly improved with regard to both error and accuracy. In the field of self-supervised learning methods, the proposed SAU-net model performed the best. The Midas [40] algorithm is one of the most advanced algorithms in the field of supervised monocular depth estimation. However, compared with SAU-net, its log RMS was only 0.1% smaller, and its results for the other three error indices were inferior to those of the proposed algorithm. This is because we add the attention-mechanism module on the decoding side, which significantly reduces the loss of feature information, and the design of the composite loss function addresses the problems of depth loss and the fuzzy target edge, reducing the values of the four error indicators. From the perspective of the threshold accuracy δ, the image prediction accuracy within the three

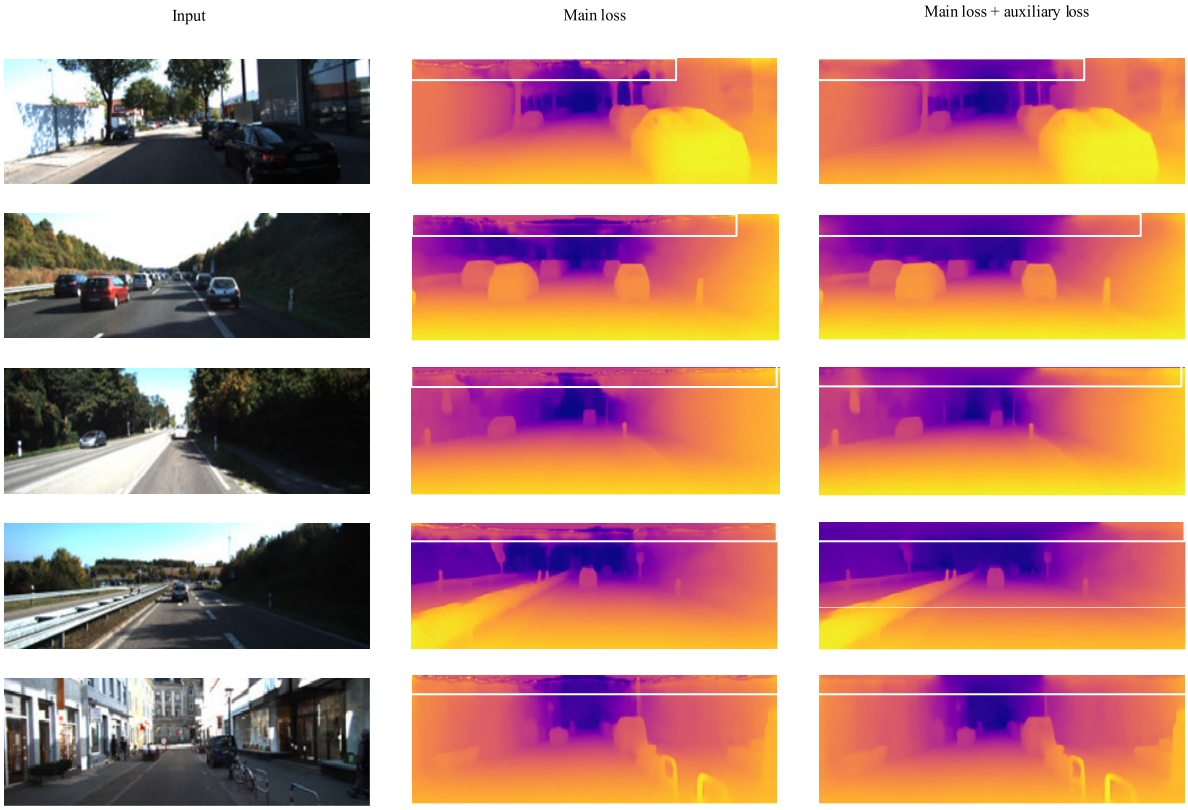

FIGURE 7. Experimental comparison of loss functions.

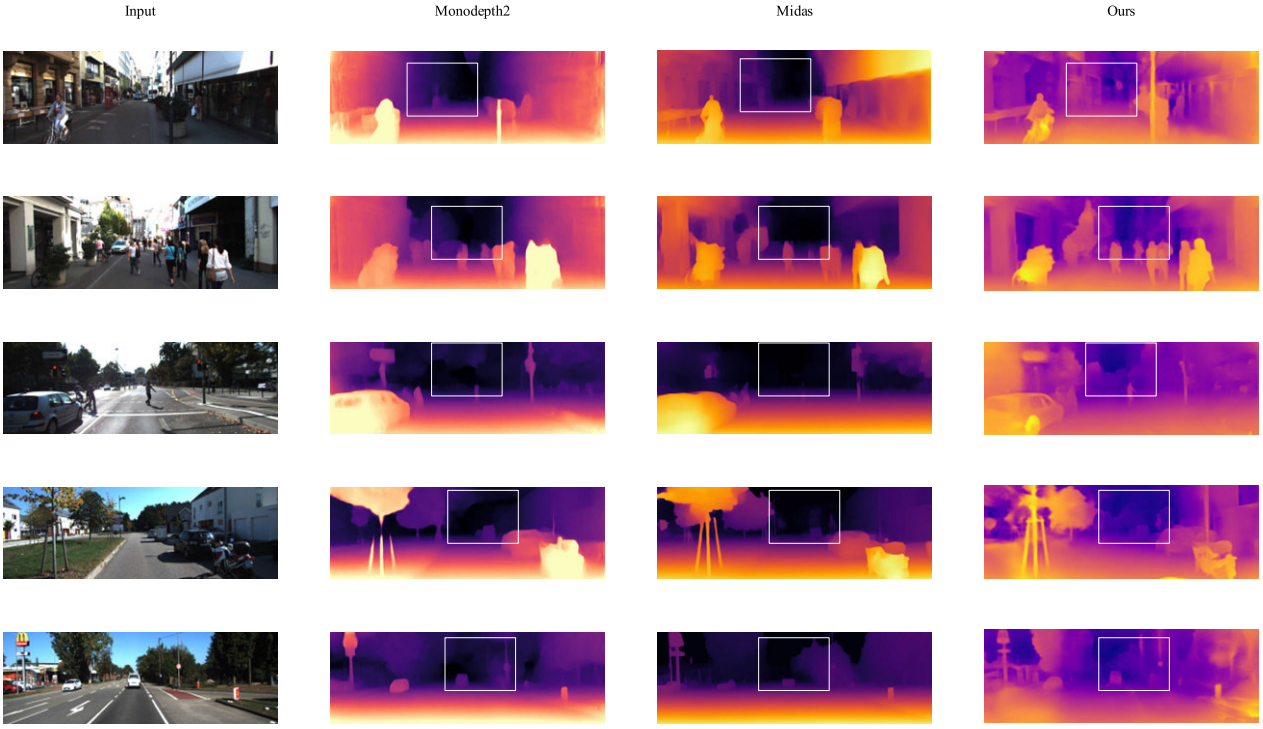

FIGURE 8. Comparison of experimental results.

recognized threshold ranges of 1.25, 1.25^2, and 1.25^3 reached 89.1%, 96.4%, and 98.5%, respectively, which were 0.7%, 0.2% and 0.2% higher than those of the Midas algorithm [40]. In summary, the self-supervised learning method SAU-net outperformed the conventional algorithms for almost all the evaluation indicators. Of course, compared with labeled

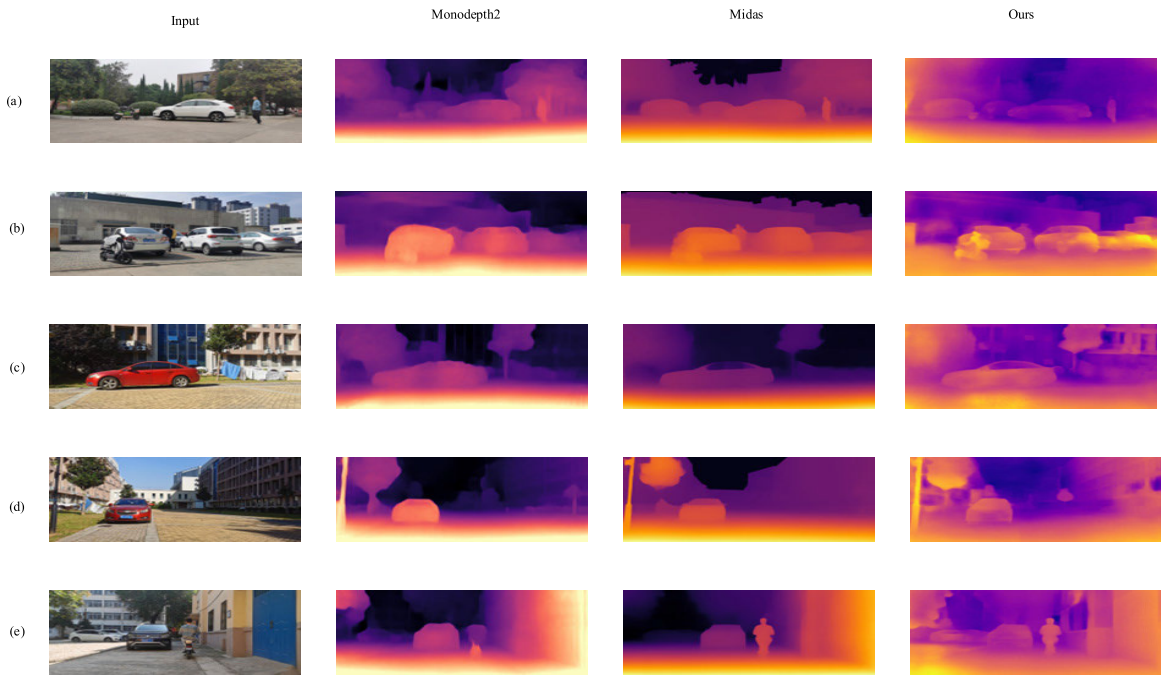

FIGURE 9. Experimental comparison for campus data.

supervised learning methods, there was a gap in the experimental results owing to the lack of sufficient GT values.

Real-time performance is crucial index for judging the performance of an algorithm; thus, the computational complexity of the proposed algorithm was evaluated. We compared the computational efficiency of the network and the number of parameters of the model with those of the self-supervised monocular depth estimation algorithms in Table 1. The results are presented in Fig. 10.

Here, the horizontal axis indicates the threshold accuracy $\delta < 1.25$, which can be used to measure the accuracy of the algorithm, and the vertical axis indicates the processing time required by the test set to test a single image (in seconds). The experiments were performed on an AGX Orin edge computing embedded system, which is specifically designed to detect the real-time responses of algorithms and is often used in autonomous driving, smart-factory machines, etc. As indicated by Fig. 10, our algorithm was undoubtedly the best with regard to accuracy, reaching 89.1%. Regarding the processing time, MonoR18 [33] was the fastest, with a single-image processing time of 0.0028 s; however, it had the lowest accuracy (87.7%). This is because the algorithm uses an 18-layer Res-Net model, and fewer network layers extract fewer features. In contrast, Monodepth2R50 [33] increases the number of network layers in the coding part, which increased the accuracy of the output results but also increased the computation time. For Midas [40], the accuracy ranked second, the processing time ranked second-to-last, and the cost was relatively high. Pack-Net [46] had the longest processing time, and its accuracy was 87.8%, which was 1.3% lower than that of the proposed algorithm; thus, its performance was inferior overall.

The number of model parameters of the proposed algorithm was compared with those of other algorithms, as shown in Table 2. Compared with Pack-Net [46], the number of parameters of the proposed algorithm was reduced by 50.4%. Among the five algorithms evaluated, the proposed algorithm had the second-fewest parameters. This is because instead of expanding the sensitivity field by stacking layers of the network, we use StratifiedTransformer, which reduces the complexity of the network, and the sliding-window mechanism design further increases the computational efficiency. Compared with MonoR18 [33], the number of parameters was slightly increased; however, compared with Midas [40] and MonoR50 [33], the number of parameters was reduced by 13.5Mb and 9.9Mb, respectively. As indicated by the results in Tables 1 and 2 and Fig. 10, our algorithm not only guarantees real-time performance but also has a higher accuracy and better cost performance than conventional algorithms.

E. ABLATION EXPERIMENT

To highlight the innovative aspects of this study, we conducted ablation experiments on the algorithm, and the results are presented in Fig. 11, where "STF-T" refers to the stratified transformer.

In the figure, red boxes highlight the effects of algorithm improvement. To minimize the influence of other factors on the ablation experiment results, we selected five scenarios similar to those in Fig. 8 for comparison. As demonstrated, the depth map estimated by the unmodified original network model was blurry, with missing details, and there was a problem of artifacts in the estimation of the distant view. For example, the columns in scene (a) and the tree supports in scene (d) both have a lack of depth information, which

TABLE 1. Verification results for the KITTI dataset obtained using the Eigen split set. In the "Pattern" column, D represents the supervised learning method, M represents the monocular self-supervision method, and S represents the self-supervision method based on stereoscopic image pairs. R18 and R50 indicate that the backbone is Resnet18 and Resnet50, respectively. The best results are underlined, and the results of the proposed method are presented in bold.

Method	Pattern	Abs Rel	RMS	Log RMS	Sq Rel	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Eigen [7]	D	0.203	6.307	0.282	1.548	0.702	0.891	0.891
Liu [41]	D	0.201	6.471	0.273	1.584	0.680	0.898	0.967
Klodt [42]	D+M	0.166	5.998	-	1.490	0.778	0.919	0.966
Garg [10]	S	0.152	5.849	0.246	1.226	0.784	0.921	0.967
Guo [43]	D+S	0.096	4.095	0.168	0.641	0.892	0.967	0.986
Dorn [44]	D	0.072	2.727	0.120	0.307	0.932	0.984	0.994
EPC++[45]	M	0.141	5.350	0.216	1.029	0.816	0.941	0.976
MonoR18[33]	M	0.115	4.863	0.193	0.903	0.877	0.959	0.981
MonoR50[33]	M	0.111	4.642	0.188	0.831	0.883	0.962	0.982
Pack-Net[46]	M	0.111	4.601	0.189	0.785	0.878	0.961	0.982
Midas[40]	M	0.109	4.632	0.182	0.792	0.884	0.962	0.983
Ours	M	**0.108**	**4.598**	**0.183**	**0.745**	**0.891**	**0.964**	**0.985**

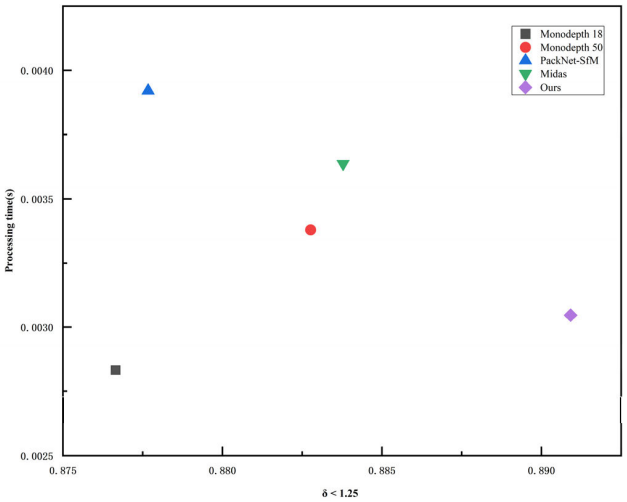

FIGURE 10. Computational efficiencies of different algorithms.

TABLE 2. Comparison of the parameters of different algorithm models. "Mb" is unit of measurement for the number of parameters.

Method	Parameter quantity/Mb
MonoR18 [33]	52.7
MonoR50 [33]	68.7
Pack-Net [46]	118.5
Midas [40]	72.3
Ours	58.8

our method compensated for this effectively. Regarding the problem of occlusion, for example, the crowd in scene (b), bicycles behind the flower bed, and pedestrians and vehicles in scene (c), this algorithm incorporates an attention module based on U-Net, which improves the capture of object edge features and significantly improves the model's performance. In addition, the stratified transformer has better scalability and a better receptive field than the Res-Net network; therefore, our method also achieved good results in the estimation of distant objects. For example, the car and McDonald's logo in scene (e) exhibited significant improvements in image blur and artifacts.

To demonstrate the effectiveness of the proposed method, we conducted ablation experiments in which the experimental images of the feature extractor Fig. 11 components were compared with the those of the dense decoder of the depth estimation network. The error and precision values are presented in Table 3.

As indicated by Table 3, the error indices RMS and Sq Rel were reduced by 0.221 and 0.072, respectively, after the backbone was changed. Following the addition of the attention module, their values were further reduced; the RMS and Sq Rel decreased by 0.044 and 0.041, respectively. Thus, replacing the backbone of Res-Net with a Stratified Transformer structure more significantly reduced the error values. The addition of attention also increased the accuracy, but the effect was slightly weaker. However, from the perspective of the threshold accuracy, the addition of the attention module played a very important role in improving the accuracy. At $thr = 1.25$, the addition of attention increased the precision of the network by 0.8%, whereas changing the backbone only increased the precision by 0.6%. This is because when the backbone was changed, to ensure the timeliness of the network, we also performed subsampling operations, during which features were lost. Therefore, although the accuracy is increased when only the backbone is changed, the improvement effect is limited. To solve this problem, we added the attention module in the compressed channel part.

V. CONCLUSION

We developed an algorithm for increasing the accuracy of depth maps in self-supervised monocular depth estimation, which is reduced by object occlusion and artifacts. 1) The core idea of this algorithm is that a StratifiedTransformer is used to replace the backbone of traditional CNNs such as Res-Net, and multi-scale features are captured in a hierarchical

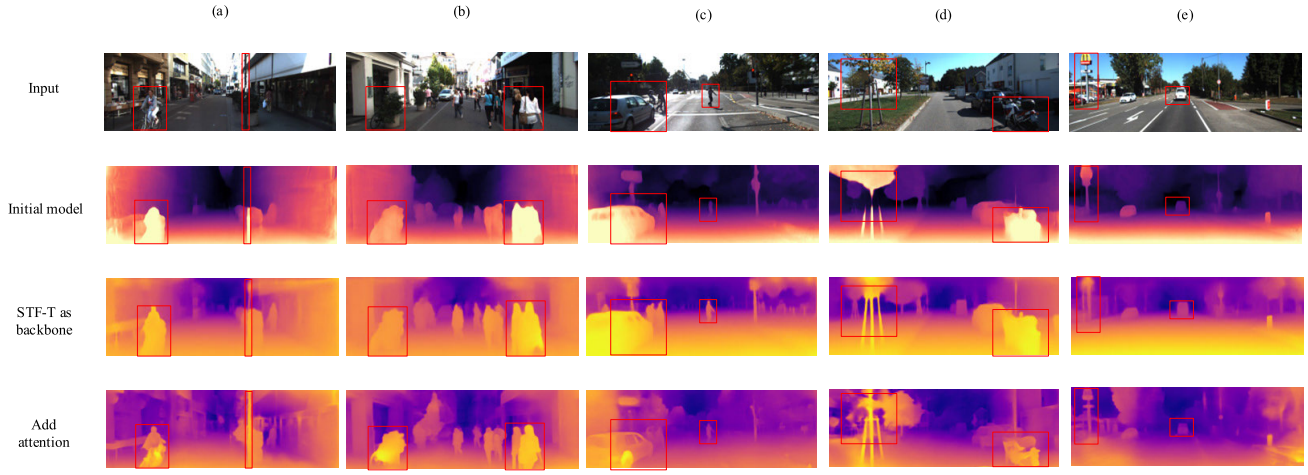

FIGURE 11. Component comparison experimental diagram.

TABLE 3. Ablation results for the KITTI dataset. In the "Method" column, "+" refers to the encoding layer, and "−" refers to the decoding layer. Resnet50+UNet is the original network model. In "STF-T+UNet," the backbone of the original ResNet network is replaced with the Stratified Transformer. "STF-T+AUNet" is the propsoed algorithm.

Method	STF-T	Attention	Abs Rel	RMS	Log RMS	Sq Rel	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
ResNet50+UNet			0.115	4.863	0.193	0.903	0.877	0.959	0.981
STF-T+UNet	√		0.111	4.642	0.188	0.831	0.883	0.962	0.982
STF-T+AUNet	√	√	0.108	4.598	0.183	0.745	0.891	0.964	0.985

manner to obtain a large sensitivity field, increasing the accuracy of the model. 2) On the basis of photometric reprojection and an automatic masking loss function, a gradient function is added to form main and auxiliary structures for improving the quality of the depth map. Moreover, an attention-mechanism module is added to U-Net, which makes the network focus on the object occlusion and artifacts and increases the segmentation accuracy. 3) Comparative experimental results for the KITTI dataset and a campus dataset indicated that the depth map boundary estimated by the proposed algorithm was clearer and had less artifacts than those estimated by conventional algorithms. Thus, SAU-net can better solve the problems of missing details and depth loss in the depth map, indicating its effectiveness and feasibility.

REFERENCES

[1] Y. B. Can, A. Liniger, D. P. Paudel, and L. Van Gool, "Improving online lane graph extraction by object-lane clustering," 2023, *arXiv:2307.10947*.

[2] M. Rokonuzzaman, N. Mohajer, and S. Nahavandi, "Human-tailored data-driven control system of autonomous vehicles," *IEEE Trans. Veh. Technol.*, vol. 71, no. 3, pp. 2485–2500, Mar. 2022.

[3] M. Itkina, Y.-J. Mun, K. Driggs-Campbell, and M. J. Kochenderfer, "Multi-agent variational occlusion inference using people as sensors," in *Proc. Int. Conf. Robot. Autom. (ICRA)*, May 2022, pp. 4585–4591.

[4] C. Min, J. Xu, L. Xiao, D. Zhao, Y. Nie, and B. Dai, "Attentional graph neural network for parking-slot detection," *IEEE Robot. Autom. Lett.*, vol. 6, no. 2, pp. 3445–3450, Apr. 2021.

[5] A. Zhanabatyrova, C. Souza Leite, and Y. Xiao, "Automatic map update using dashcam videos," 2021, *arXiv:2109.12131*.

[6] A. J. C. Alveyra, Y. Edgar Foronda, and G. V. Magwili, "Pseudo binocular stereo vision camera realignment using NCC," in *Proc. 9th Int. Conf. Adv. Comput. Commun. Syst. (ICACCS)*, vol. 1, Mar. 2023, pp. 1020–1025, doi: 10.1109/ICACCS57279.2023.10112912.

[7] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," 2014, *arXiv:1406.2283*.

[8] B. Li, C. Shen, Y. Dai, A. van den Hengel, and M. He, "Depth and surface normal estimation from monocular images using regression on deep features and hierarchical CRFs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1119–1127.

[9] X. Qi, R. Liao, Z. Liu, R. Urtasun, and J. Jia, "GeoNet: Geometric neural network for joint depth and surface normal estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 283–291.

[10] R. Garg, B. Kumar, G. Carneiro, and I. Reid, "Unsupervised CNN for single view depth estimation," in *Proc. 14th Eur. Conf. Comput. Vis.*, Amsterdam, The Netherlands, 2106, pp. 740–756, doi: 10.1007/978-3-319-46484-8_45.

[11] C. Godard, O. M. Aodha, and G. J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 6602–6611, doi: 10.1109/CVPR.2017.699.

[12] J. Watson, M. Firman, G. Brostow, and D. Turmukhambetov, "Self-supervised monocular depth hints," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Seoul, South Korea, Oct. 2019, pp. 2162–2171, doi: 10.1109/ICCV.2019.00225.

[13] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015*. Cham, Switzerland: Springer, 2015, pp. 234–241.

[14] H. Xu, S. Lai, X. Li, and Y. Yang, "Cross-domain car detection model with integrated convolutional block attention mechanism," 2023, *arXiv:2305.20055*.

[15] M. S. A. Shawkat, M. M. Adnan, R. D. Febbo, J. J. Murray, and G. S. Rose, "A single chip SPAD based vision sensing system with integrated memristive spiking neuromorphic processing," *IEEE Access*, vol. 11, pp. 19441–19457, 2023.

[16] P. Melzi, C. Rathgeb, R. Tolosana, R. Vera-Rodriguez, D. Lawatsch, F. Domin, and M. Schaubert, "GANDiffFace: Controllable generation of synthetic datasets for face recognition with realistic variations," 2023, *arXiv:2305.19962*.

[17] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," 2017, *arXiv:1709.01507*.

[18] Z. Yang, L. Zhu, Y. Wu, and Y. Yang, "Gated channel transformation for visual recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11791–11800.

[19] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-Net: Efficient channel attention for deep convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11531–11539.
[20] Z. Wen, Y. Zhang, X. Chen, and J. Wang, "TOFG: A unified and fine-grained environment representation in autonomous driving," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2023, pp. 1565–1571.
[21] N. Alam, S. Kolawole, S. Sethi, N. Bansali, and K. Nguyen, "Vision transformers for mobile applications: A short survey," 2023, *arXiv:2305.19365*.
[22] J. Zhou, Y. Wu, W. Song, Z. Cao, and J. Zhang, "Towards omni-generalizable neural methods for vehicle routing problems," 2023, *arXiv:2305.19587*.
[23] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.
[24] F. Lin, Y. Ma, and S. Tian, "Exploring vision transformer layer choosing for semantic segmentation," in *Proc. ICASSP - IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2023, pp. 1–5.
[25] G. Rizzoli, D. Shenaj, and P. Zanuttigh, "Source-free domain adaptation for RGB-D semantic segmentation with vision transformers," 2023, *arXiv:2305.14269*.
[26] Y. Deng, S. Hui, S. Zhou, D. Meng, and J. Wang, "T-former: An efficient transformer for image inpainting," 2023, *arXiv:2305.07239*.
[27] A. Rehman Khan and A. Khan, "MaxViT-UNet: Multi-axis attention for medical image segmentation," 2023, *arXiv:2305.08396*.
[28] L. Shi, T. Gao, Z. Zhang, and J. Zhang, "STM-UNet: An efficient U-shaped architecture based on Swin transformer and multi-scale MLP for medical image segmentation," 2023, *arXiv:2304.12615*.
[29] Y. Yang, S. Dasmahapatra, and S. Mahmoodi, "ADS_UNet: A nested UNet for histopathology image segmentation," 2023, *arXiv:2304.04567*.
[30] P. Liang, W. Lin, G. Luo, and C. Zhang, "Research of hand–eye system with 3D vision towards flexible assembly application," *Electronics*, vol. 11, no. 3, p. 354, Jan. 2022, doi: 10.3390/electronics11030354.
[31] X.-Z. Cui, Q. Feng, S.-Z. Wang, and J.-H. Zhang, "Monocular depth estimation with self-supervised learning for vineyard unmanned agricultural vehicle," *Sensors*, vol. 22, no. 3, p. 721, Jan. 2022, doi: 10.3390/s22030721.
[32] H.-T. Duong, H.-M. Chen, and C.-C. Chang, "URNet: An UNet-based model with residual mechanism for monocular depth estimation," *Electronics*, vol. 12, no. 6, p. 1450, Mar. 2023, doi: 10.3390/electronics12061450.
[33] C. Godard, O. M. Aodha, M. Firman, and G. Brostow, "Digging into self-supervised monocular depth estimation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Seoul, South Korea, Oct. 2019, pp. 3827–3837, doi: 10.1109/ICCV.2019.00393.
[34] C. Q. Ma, X. H. Li, and L. J. Zhang, "Anti occlusion monocular depth estimation algorithm," *Comput. Eng. Appl.*, vol. 57, no. 2, pp. 217–222, 2021, doi: 10.3778/j.issn.1002-8331.1911-0346.
[35] M. Song, S. Lim, and W. Kim, "Monocular depth estimation using Laplacian pyramid-based depth residuals," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 11, pp. 4381–4393, Nov. 2021, doi: 10.1109/TCSVT.2021.3049869.
[36] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 3354–3361.
[37] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Montreal, QC, Canada, Oct. 2021, pp. 9992–10002, doi: 10.1109/ICCV48922.2021.00986.
[38] Z. Zhang, "Improved Adam optimizer for deep neural networks," in *Proc. IEEE/ACM 26th Int. Symp. Quality Service (IWQoS)*, Banff, AB, Canada, Jun. 2018, pp. 1–2, doi: 10.1109/IWQoS.2018.8624183.
[39] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and Li Fei-Fei, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.
[40] F. Aleotti, M. Poggi, and S. Mattoccia, "Learning optical flow from still images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 15196–15206.
[41] F. Liu, C. Shen, G. Lin, and I. Reid, "Learning depth from single monocular images using deep convolutional neural fields," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 10, pp. 2024–2039, Oct. 2016, doi: 10.1109/TPAMI.2015.2505283.
[42] M. Klodt and A. Vedaldi, "Supervising the new with the old: Learning SFM from SFM," in *Proc. ECCV*, 2018, pp. 698–713.
[43] X. Guo, H. Li, S. Yi, J. Ren, and X. Wang, "Learning monocular depth by distilling cross-domain stereo networks," in *Proc. ECCV*, 2018, pp. 484–500.
[44] H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao, "Deep ordinal regression network for monocular depth estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 2002–2011, doi: 10.1109/CVPR.2018.00214.
[45] C. Luo, Z. Yang, P. Wang, Y. Wang, W. Xu, R. Nevatia, and A. Yuille, "Every pixel counts++: Joint learning of geometry and motion with 3D holistic understanding," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 10, pp. 2624–2641, Oct. 2020, doi: 10.1109/TPAMI.2019.2930258.
[46] V. Guizilini, R. Ambrus, S. Pillai, A. Raventos, and A. Gaidon, "3D packing for self-supervised monocular depth estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, Jun. 2020, pp. 2482–2491, doi: 10.1109/CVPR42600.2020.00256.

WEI ZHAO received the Ph.D. degree in engineering from Chang'an University, in 2008. He is currently an Associate Professor with the Henan University of Science and Technology. His current research interests include machine vision and self-driving car technology.

YUNQING SONG is currently pursuing the M.Eng. degree in vehicle engineering with the Henan University of Science and Technology. His current research interests include computer vision and self-driving technology.

TINGTING WANG is currently pursuing the M.Eng. degree in transportation engineering with the Henan University of Science and Technology. Her current research interests include target detection and sensor fusion technology.