Heritage Science

RESEARCH Open Access

Reconstructing the local structures of Chinese ancient architecture using unsupervised depth estimation

Xiaoling Yao[1†], Lihua Hu[1*†] and Jifu Zhang[1†]

Abstract

Digitalization of ancient architectures is one of the effective means for the preservation of heritage structures, with 3D reconstruction based on computer vision being a key component of such digitalization techniques. However, Chinese ancient architectures are located in mountainous areas, and existing 3D reconstruction methods fall short in restoring the local structures of these architectures. This paper proposes a self-attention-guided unsupervised single image-based depth estimation method, providing innovative technical support for the reconstruction of local structures in Chinese ancient architectures. First, an attention module is constructed based on features extracted from architectural images learned by the encoder, and then embedded into the encoder-decoder to capture the interdependencies across local features. Second, a disparity map is generated using the loss constraint network, including reconstruction matching, smoothness of the disparity, and left-right disparity consistency. Third, an unsupervised architecture based on binocular image pairs is constructed to remove any potential adverse effects due to unknown scale or estimated pose errors. Finally, with the known baseline distance and camera focal length, the disparity map is converted into the depth map to perform the end-to-end depth estimation from a single image. Experiments on the our architecture dataset validates our method, and it performs well also well on KITTI.

Keywords Chinese ancient architecture, Reconstruction the local structures, Unsupervised depth estimation, Self-attention

Introduction

Chinese ancient architectures, such as Buddhist and Taoist temples, are wooden structures, and vulnerable to natural and manmade damages [1]. The 3D digitalization of such architectures is urgently needed for archiving and protection. Due to large scale and often time located on mountainous sites of ancient architectures, it is inevitable that some local structures are missed when attempting to digitize at once using either laser scanner or image-based technology [2, 3]. Therefore, it is necessary to find some easy methods to fill the missing parts afterwards.

Multi-view geometry methods rely on manual feature extraction and matching from images, but with limited feature points and high mismatch rates, the resulting sparse point cloud often fails to fully reconstruct local structures, particularly in scenes of ancient architecture with complex structures and repetitive textures. In contrast, single image depth estimation (SIDE) can estimate the depth of each pixel from a single image, enabling a more detailed and complete reconstruction of complex structures. Geometrically speaking, SIDE methods are an ill-posed problem because an infinitely large number of space points can project to the same image point [4]. However, thanks to the tremendous image representational ability of the convolutional

[†]Xiaoling Yao, Lihua Hu and Jifu Zhang have contributed equally to this work.

*Correspondence:
Lihua Hu
hlh@tyust.edu.cn
[1] School of Computer Science and Technology, Taiyuan University of Science and Technology, Waliu Road 66, Taiyuan 030024, China

neural network (CNN), SIDE methods have achieved exceptionally good depth estimation accuracy in either indoor or outdoor datasets, under either a supervised or unsupervised framework. However, we observed that current SIDE methods do not perform as well for ancient Chinese architectural images as for open indoor or outdoor datasets, such as NYU or KITTI, due to the peculiar nature of such images. We find that in addition to the shape complexity, those images have rich repeated textural and structural patterns, as shown in Fig. 1, which significantly complicates the SIDE because they make it difficult to distinguish depth differences between regions with similar textures.

To tackle the above interference problem of repeated textural and structural patterns in ancient Chinese architectural images, and considering that currently some kind of benchmark dataset is not available in the field, we proposed a self-attention-guided unsupervised SIDE method that is trained with calibrated stereo image pairs to remove any potential adverse effects in the estimated depth by unknown scale or relative pose errors estimated on-the-fly by such a method as PoseNet. Our main contributions include:

1. We observed that due to the pattern interference of repeated textural and structural patterns, which are abundant in ancient Chinese architectural images, the current state-of-the-art SIDE methods do not perform well on such images;
2. We propose an unsupervised binocular training-based depth estimation method for application scenarios of repeated textural and structural patterns.

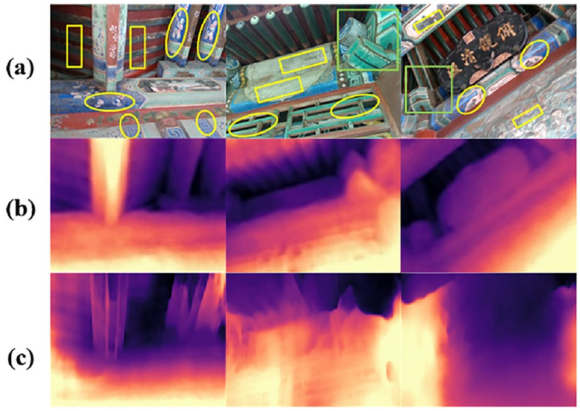


Fig. 1 The current state-of-art SIDE models do not perform well for ancient Chinese architectural images due to the presence of abundant repeated textural and structural patterns **a** The input images, the yellow circle or box represent regions of repeated textural and structural patterns, and the green boxes represent regions of complex shapes. **b** Experimental results of Mondepth2 [16] **c** Experimental results of Lite-mono [15]

3. We construct an ancient Chinese architectural image dataset which contains 14 typical local geometric structures.
4. Experimental results on the architecture and benchmark datasets validates the effectiveness of our method.

Related work

In the field of cultural heritage preservation and restoration, depth estimation techniques play a crucial role. They provide powerful tools for accurately capturing and reconstructing the 3D structures of ancient buildings and artworks, thereby supporting our understanding, preservation, and exhibition of these invaluable treasures. Within this context, unsupervised SIDE trained with stereo image pairs have shown particular promise. The methods leverage binocular vision principles to generate accurate depth maps without relying on external annotated data. Their adaptability makes them especially useful in the cultural heritage domain, where obtaining large amounts of labeled data is often impractical. Meanwhile, the attention mechanism enhance accuracy and efficiency by focusing on key regions within images. This is particularly important for complex and richly detailed ancient architecture. Garg et al. [5] proposed an approach to predict disparity map based on photometric errors. Godardet al. [6] followed the same line by introducing additional left-right disparity checking. Repala et al. [7] proposed a method based on dual CNN for cross image reconstruction. Tosi et al. [8] employed Huber loss to enhance the network's robustness to outliers. Ling et al. [9] used the first disparity estimation map to perform three image reconstructions as a new reconstruction loss function, which improved the accuracy and robustness of depth estimation. One of the disadvantages of stereo image pairs based SIDE is that it cannot exploit more contextual information embedded in the neighboring video frames during the training phase.

Wang et al. [10] proposed a non-local operation to capture long-range dependencies. More recently, non-local operation architecture has been introduced as the self-attention module in the SIDE problem. Johnston et al. [11] employed the self-attention module and discrete disparity volume in the network, which generate more robust and clearer depth estimation. Ji et al. [12] proposed a self-attention-guided scale regression network to estimate the global scale factor. Jiang et al. [13] proposed a monocular depth estimation method based on a dual attention mechanism, which enhances the expression ability of features and improves the accuracy and robustness of depth prediction by comprehensively considering spatial-channel attention. Lee et al. [14] proposed a module to extract structural information by segmenting the input image into blocks, and used an enhanced attention mechanism to fuse structural features with original features to enhance the performance of depth estimation. Zhang et al. [15] proposed a local–global feature interaction module that can use the self-attention mechanism to encode long-distance global information into features, thereby improving the performance and robustness of depth estimation. Godard et al. [16] employed multi-scale estimation together with a novel minimum re-projection loss for occlusion management to further reduce the performance difference between monocular and stereo-trained self-supervision.

Unsupervised SIDE trained with stereo image pairs were originally developed for tasks in the field of computer vision, such as autonomous driving and 3D modeling. However, the potential applications of these technologies in the preservation of cultural heritage are also significant. They offer a new pathway to reconstruct and preserve ancient cultural heritages with greater accuracy and efficiency. Future research will continue to explore the application of these technologies in the digitalization and virtual reconstruction of cultural heritage, allowing us to leave behind a richer and more vivid historical legacy.

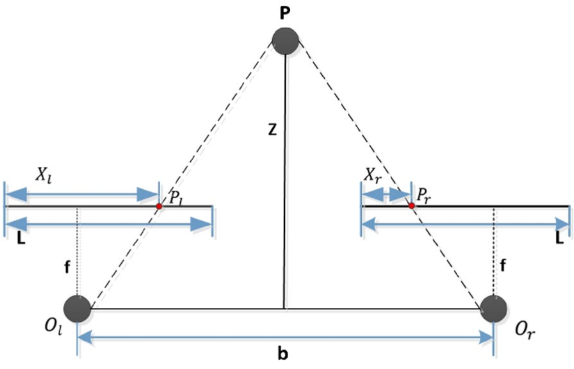


Fig. 2 Principle of binocular stereo vision system

Preliminaries

Depth and disparity

Depth and disparity are two closely connected concepts. Disparity is defined as the image position difference of a corresponding image pair [17, 18]. As shown in Fig. 2, P_l and P_r are the image points of a space point $\mathbf{P}$ on the left-right cameras, f denotes the focal length, and O_l and O_r are the optical centers of left-right cameras, and the baseline is b. Then the distance z from $\mathbf{P}$ to the optical center plane is the depth, and the disparity d is:

$$d = |X_l - X_r| \tag{1}$$

And for the standard stereo setting in Fig. 2, the depth is related to disparity by:

$$z = bf/d \tag{2}$$

Left-right disparity consistency

Figure 3 outlines the principle of the left-right disparity consistency, a well-known concept in stereo community. The left-right disparity consistency refers to the constraints: $x_l = x_r + d_l$ and $x_r = x_l + d_r$, where $(x_l \longleftrightarrow x_r)$ is a pair of corresponding image points, and d_l, d_r are associated left and right disparity. In ref. [6], the left-right disparity consistency was used as a separate loss to handle occlusion problem. Nowadays, it is generally established that SIDE with occlusion/disocclusion mechanism performs better than its counterpart without the consistency, and in this work, we also employ this consistency to alleviate the repeated pattern interference problem.

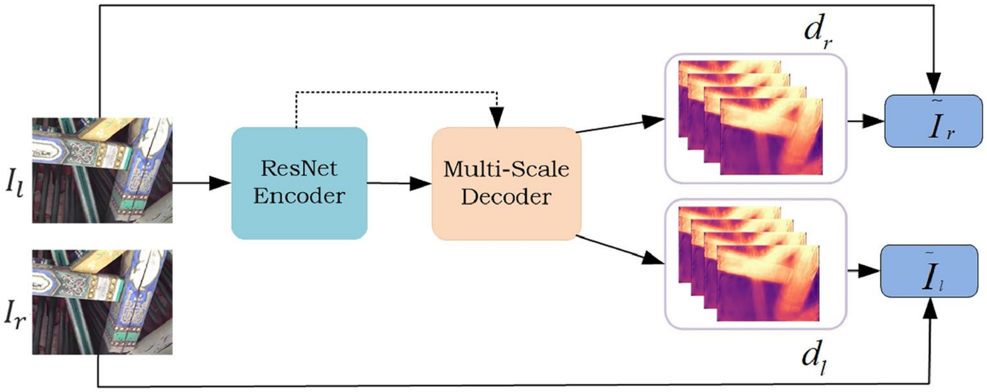


Fig. 3 Principle of left-right consistency

Method

In this section, the problem to be solved is introduced, and an overview of the proposed method is presented. Next, the network architecture of the model is elaborated on. Finally, we present the proposed learning strategy.

Problem formulation

We observe that the current SIDE methods do not perform well for ancient Chinese architectural images, some estimated depth maps are shown in Fig. 1 for Mondepth2 [16] and Lite-mono [15]. where repeated textural and structural patterns significantly complicate the depth estimation. We thought the major adverse factors of our ancient Chinese architectural images include:

(1) The convolutional neural network is inherently translation invariant which is an advantage for image categorization. However, depth estimation is a pixel wise dense estimation, the multi-layer convolutions could potentially aggravate the pattern interference by enlarging receptive fields;
(2) The severe interference of repeated textural and structures patterns in the Ancient Chinese architectural images;
(3) The influence of the unknown scale and potential pose error by the pose estimation module seem more severe for architectural images.

Hence in this work, we use the calibrated stereo image pairs for model training to eliminate scale and pose errors, and adopt the self-attention mechanism to alleviate the pattern interference. Like others, we also use the left-right disparity constraint to enforce geometry consistence. By such, we found our proposed method significantly improve the depth estimation performance for the ancient Chinese architectural images.

Network architecture

The proposed method is depicted in Fig. 4. Here, we describe our depth prediction network that takes a pair of stereo images, where only the left image is processed by the network and the right image provides supervision. We first review the key ideas behind the network structure for depth estimation, including feature extraction, self-attention, feature fusion and up-sampling, and then describe the loss function required for model optimization.

(1) *Feature extraction*

As shown in Fig. 5, the encoder employs residual network (ResNet) as the feature extractor, which leverages a residual network architecture to enable deeper networks for richer feature extraction. The architecture includes six modules: E_1(convolution module, Conv), E_2 (maximum pooling module, Maxpool), and four residual modules (Resblock). Each residual module features skip connections that facilitate multi-scale feature extraction by passing residual information across layers. This multi-scale approach enhances the network's ability to capture detailed and abstract features at various levels, improving learning performance and efficiency. Additionally, it helps address issues such as overfitting and network deepening by providing better feature representation and faster convergence. This process generates multi-channel feature maps, which are then fed into the self-attention module to learn contextual relationships between different regions of the image.

(2) *Self-attention*

Unlike the non-local operation in ref. [10], which calculates correlations between all pixel points, our model addresses the issue of capturing complex feature dependencies within local regions through a self-attention module. As shown in Fig. 4 and detailed in Fig. 6, the self-attention block captures the contextual relationships between different features within a local region. It allows the model to understand how features interact with each other, even when they share similar

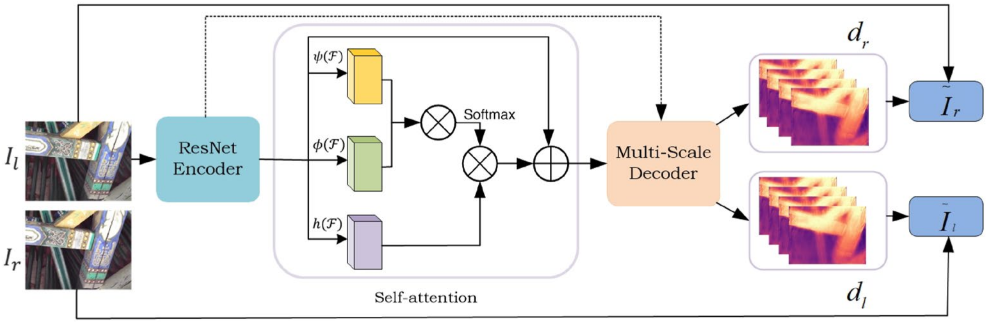


Fig. 4 Principle of our proposed method

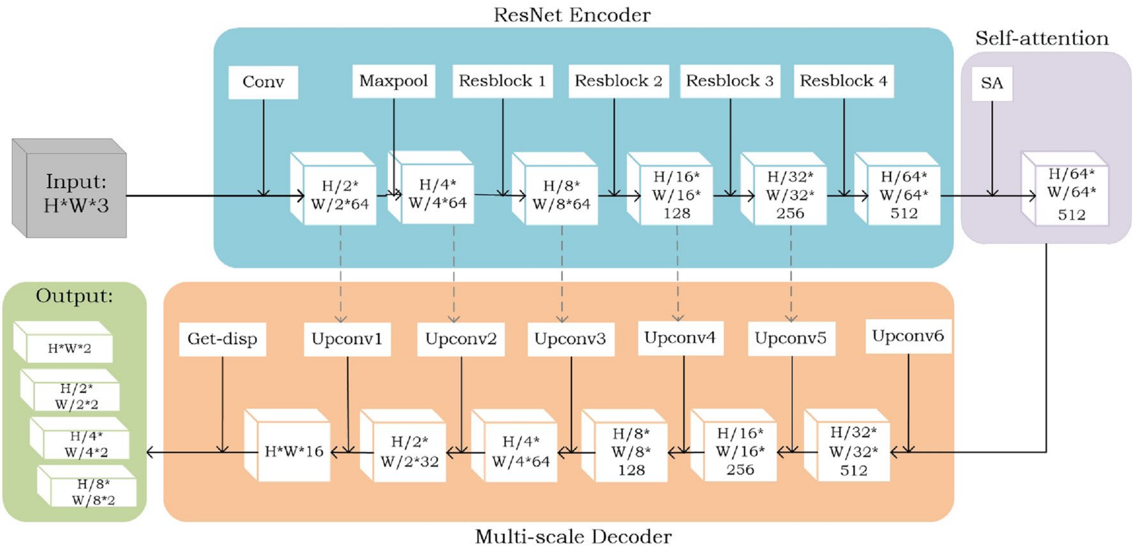


Fig. 5 Network architecture of our method. In the model, we use a cube to represent a feature map of size height×width× features. The input size of the model is H×W×3. The four scales' output sizes of the model are H×W×2, H/2×W/2×2, H/4×W/4×2, and H/8×H/8×2

textures. This helps the network differentiate between regions that might have similar surface appearances but different depths.

The self-attention block is used to accept the feature representations $\mathcal{F}$ that we have learnt from the input picture as input. From there, we build the query, the key, and the value output by:

$$\begin{aligned} Q(\psi(\mathcal{F})) &= W_{\psi}\mathcal{F} \\ K(\phi(\mathcal{F})) &= W_{\phi}\mathcal{F} \\ V(h(\mathcal{F})) &= W_{h}\mathcal{F} \end{aligned} \tag{3}$$

where W_{ψ}, W_{ϕ}, W_h are the parameters to be learned. The query and key values are then combined to compute the correlation between all features, and the self-attention matrix A is attained:

$$A\left(a_{i,j}\right) = \left(q^{i}\right)^{T} k^{j} = Q^{T}K \tag{4}$$

where each value in A records the correlation a between the corresponding two input features. The softmax operation is performed on the self-attention matrix A to obtain A', the resulting A' and V are used to calculate the output feature O obtained by the input feature through the self-attention module:

$$O = V(h(\mathcal{F}))A' \tag{5}$$

Finally, the self-attention block refines the feature representation by integrating information from different parts of the local region. The refined feature map O is combined with the original feature map $\mathcal{F}$ to produce $S_{\mathcal{F}}$, which is then input to the decoder.

$$S_{\mathcal{F}} = W_{S_{\mathcal{F}}}O + \mathcal{F} \tag{6}$$

(3) *Feature fusion and up-sampling*

The decoder, which is responsible for feature fusion and up-sampling, consists of six deconvolution modules (Upconv) and a disparity synthesis module (Get-disp). Each deconvolution module handles the restoration of feature maps to their original size by combining new feature maps from the previous stage with the original feature maps. This fusion process is crucial for effective up-sampling. The Get-disp module then generates the predicted disparity map. The network produces disparity maps d_j at four scales, each containing left-right disparity maps $(1 \leq j \leq 4)$.

Loss function

In ref. [6], the input left-right images are used for mutual supervision, which enables the depth estimation network to generate more accurate left-right disparity maps, and curb the discontinuity of image depth. The loss function is defined on four scales: $L = \sum_{S=1}^{4} L_S$, where s represents different output scales. The loss at each scale consists of three parts: reconstruction matching loss L_m, disparity smoothness loss L_{d_s}, and left-right disparity consistent loss L_{l_r}. L_S can be expressed as:

$$L_s = \alpha_m\left(L_m^l + L_m^r\right) + \alpha_{ds}\left(L_{ds}^l + L_{ds}^r\right) + \alpha_{lr}\left(L_{lr}^l + L_{lr}^r\right) \tag{7}$$

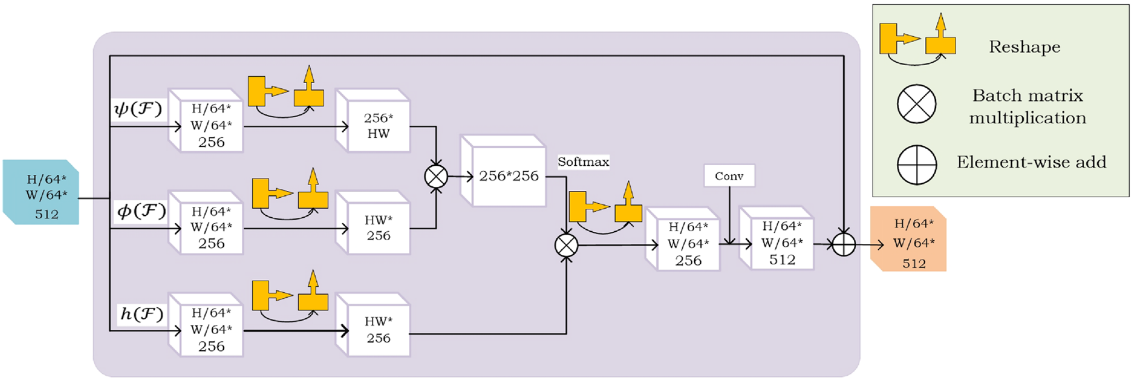


Fig. 6 Network architecture of self-attention

where $\alpha_m, \alpha_{ds}, \alpha_{lr}$ are the weights of three losses. The network uses the left image as input and outputs both left-right disparity maps, so each loss has both left-right versions. The three losses are described below using the left image version as an example:

Reconstruction matching loss L_m: The network attempts to estimate the disparity map for a single view, and then samples pixels from the opposing stereo picture to create an image based on this disparity. Consequently, the degree of similarity between the original and reconstructed images is determined by the correctness of the disparity map. This introduces the structural similarity indicator SSIM [19]. According to this theory, the reconstruction matching loss has the following definition:

$$L_m^l = \frac{1}{N}\sum_{i,j}\alpha\frac{1-\mathrm{SSIM}\left(I_{ij}^l,\tilde{I}_{ij}^l\right)}{2} + (1-\alpha)\left\|I_{ij}^l-\tilde{I}_{ij}^l\right\| \tag{8}$$

Disparity smoothness loss L_{d_s}: To avoid the discontinuity of depth estimation, the gradient information of the original image in the x and y directions is used to constrain the gradient of the disparity map. The disparity map of the smooth area in the original image should be smoother, and the boundary area with large gradient changes in the original image also guides the disparity map to attain a clearer boundary.

$$L_{ds}^l = \frac{1}{N}\sum_{i,j}\left|\partial_x d_{ij}^l\right| e^{-\left\|\partial_x d_{ij}^l\right\|} + \sum_{i,j}\left|\partial_y d_{ij}^l\right| e^{-\left\|\partial_y d_{ij}^l\right\|} \tag{9}$$

Left-right disparity consistency loss L_{l_r}: The consistency loss needs to be consistent in order to guarantee that the left-right disparity maps that are produced are valid. By combining the estimated disparity map d_l with the right disparity map d_r, we may create a new version of the left disparity map, d_l'. It is also possible to construct the new disparity map d_r' matched to the correct image using the same method. The left-right disparity consistency loss is then described as follows:

$$L_{lr}^l = \frac{1}{N}\sum_{i,j}\left|d_{ij}^l - d^r_{ij+d_{ij}^l}\right| \tag{10}$$

Results and discussion

Due to the peculiar characteristics of the ancient Chinese architectural images, in particular, the abundance of repeated textural and structural patterns, we observed that the current state-of-art single-image based depth estimation methods do not work well on such images even if they have achieved impressive results on public datasets such as NYUv2 or KITTI, as shown in Fig. 1. At present, there is no public dataset on architectural images, we construct an ancient Chinese architectural images dataset, denoted as ARCHITECT for our method training and testing. In addition, we also test our model on KITTI dataset. As shown in the next sections, our method performs well on both ARCHITECT and KITTI. In the following, dataset construction, evaluation metrics, and experimental results will be reported.

ARCHITECT dataset

The LenaCV CAM-OV9714-6 binocular camera from Lena Machine Vision was used to capture stereo image pairs. Binocular calibration and rectification were performed by MATLAB's Stereo Camera Calibrator app [20]. The original images are of 1400 × 1400 pixels, we first cropped the central part to 1280 × 960, then resized it to 640 × 480 using the OpenCV resize function. The calibrated baseline is 60 mm. Since the stereo rig is fixed, the system calibration is done before hand. The calibrated cameras' intrinsic and extrinsic parameters are saved as known parameters in training.

Ancient Chinese architectures have specific meaningful parts. After consultation with the experts in the field, 14 typical subscenes are selected from Summer Palace. A total of 4074 image pairs, comprising 8148 individual images, are collected as the training dataset.

The test dataset is constructed using Intel RealSense D435i depth camera whose nominal depth accuracy is less than 2% at 2 m. To ensure the accuracy of the captured "ground-truth depth" is less than 1.0 cm, we deliberately keep the depth camera within the range of less than 50 cm. The captured image with pixelwise depth value is of 640 × 480, and a total of 669 images across the 14 subscenes are used as the test dataset.

Evaluation metrics

Following the convention [21], the following 5 metrics are used for the model evaluation:

$$AbsRel : \frac{1}{|T|}\sum_{d\in T}\left|d-d^*\right|/d^* \tag{11}$$

$$\text{Sq}\ Rel : \frac{1}{|T|}\sum_{d\in T}\left\|d-d^*\right\|^2/d^* \tag{12}$$

$$RMSE : \sqrt{\frac{1}{|T|}\sum_{d\in T}\left\|d-d*\right\|^2} \tag{13}$$

$$RMSE\,\ln : \sqrt{\frac{1}{|T|}\sum_{d\in T}\left\|\ln d-\ln d*\right\|^2|} \tag{14}$$

$$\%ofd_i \quad s.t.max\left(\frac{d_i}{d_i^*},\frac{d_i^*}{d_i}\right)=\delta<thr \tag{15}$$

where d represents the predicted depth of pixel, d^* represents the ground truth depth of pixel, and T is the total number of pixels that can be obtained in the ground truth depth map.

Experimental environment

Our model was implemented using PyTorch [22] and trained on a single Nvidia GeForce RTX 2080 Ti. We set the input size to 640 × 480 and the number of epochs to 100. With $\beta_1 = 0.9, \beta_2 = 0.999, \varepsilon = 10^{-8}$, and an initial learning rate of 10^{-4}, Adam was employed as the optimizer. On the spot data augmentation was done. With a 50% chance, we flipped the input photos horizontally, making sure to additionally swap both images so they were in the proper alignment with respect to one another. Additionally, we included 50% chance color augmentations, in which we sampled from uniform distributions in the ranges [0.8, 1.2] for gamma, [0.5, 2.0] for brightness, and [0.8, 1.2] for each color channel independently to conduct random shifts in gamma, brightness, and color.

Comparison methods

Our comparison methods include supervised methods [21, 23], unsupervised methods [5, 6, 9, 11, 15, 16, 24–30]. The results were visualized in the form of heat maps.

Experiments on ARICHTECT dataset

Table 1, Figs. 7, and 8 are some of the estimated depth maps on ARCHITECT dataset, visualizing the interference effects of similar local image patterns for the different methods. Our method is evaluated under 3 different work architectures respectively: ResNet18, 50, and 101. We trained all the three architectures and used the post-processing method in ref. [6].

As can be seen in Table 1, our method is superior to most of the methods in terms of accuracy and error value. In addition, the third threshold accuracy is only slightly inferior to that of Monodepth2 [16] and Johnston [11] methods in binocular mode when the network architecture deepens. We observed a clear trend: transitioning from ResNet18 to ResNet50 and ResNet101 does not consistently improve performance. In fact, ResNet18 yields the best results, indicating that in complex scenes such as the reconstruction of ancient architectural details, deeper network architectures may introduce overfitting or learning difficulties, leading to suboptimal performance. Therefore, we chose ResNet18 as the optimal architecture, offering the best balance between accuracy, error minimization, and computational.

The visualization results in Fig. 7 show that, compared with other methods, the resulting depth map of our method has richer details and clearer edges, with fewer artifacts. The white boxes in the image represent areas with different positions but the same texture structure, and their depths are not the same. However, all comparative methods tend to regard the depth of these areas as consistent. The main reasons are as follows:

Garg [5], Godard [6], Monodepth2 [16], and Struct-Depth [26] are classic frameworks in the field of unsupervised single-image depth estimation. The current state-of-the-art unsupervised single-image depth estimation methods include Swindepth [27], Mono-FiVI [28], Xiong [29], and Lite-mono [15]. Among these, Garg, Godard, Monodepth2, Swindepth, and Xiong focus primarily on loss function design, such as reconstruction loss, disparity consistency, photometric consistency, and scale consistency. However, these methods often

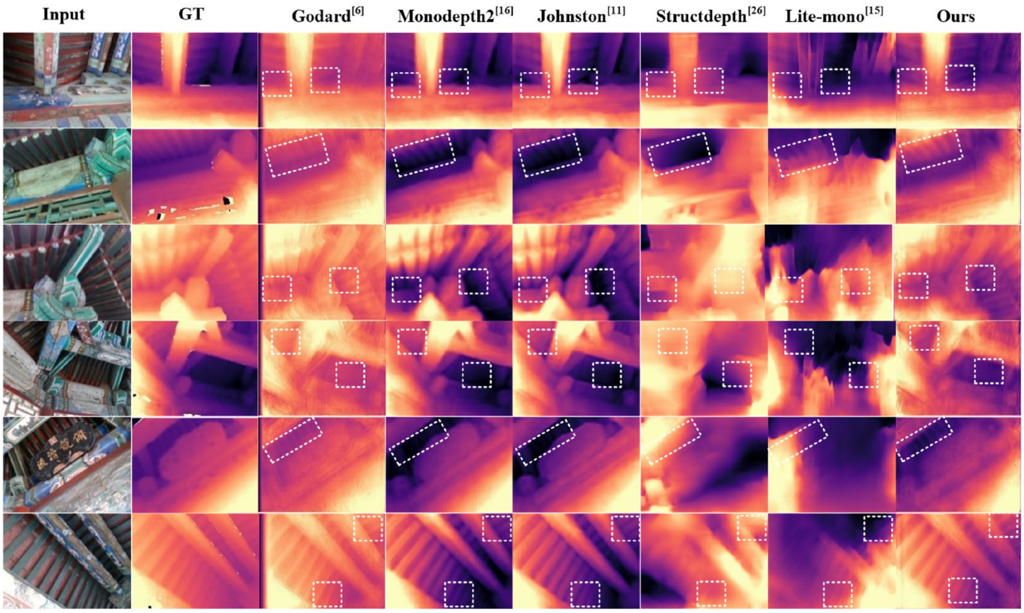


Fig. 7 Visualization of estimated depth map on the ancient Chinese architecture dataset

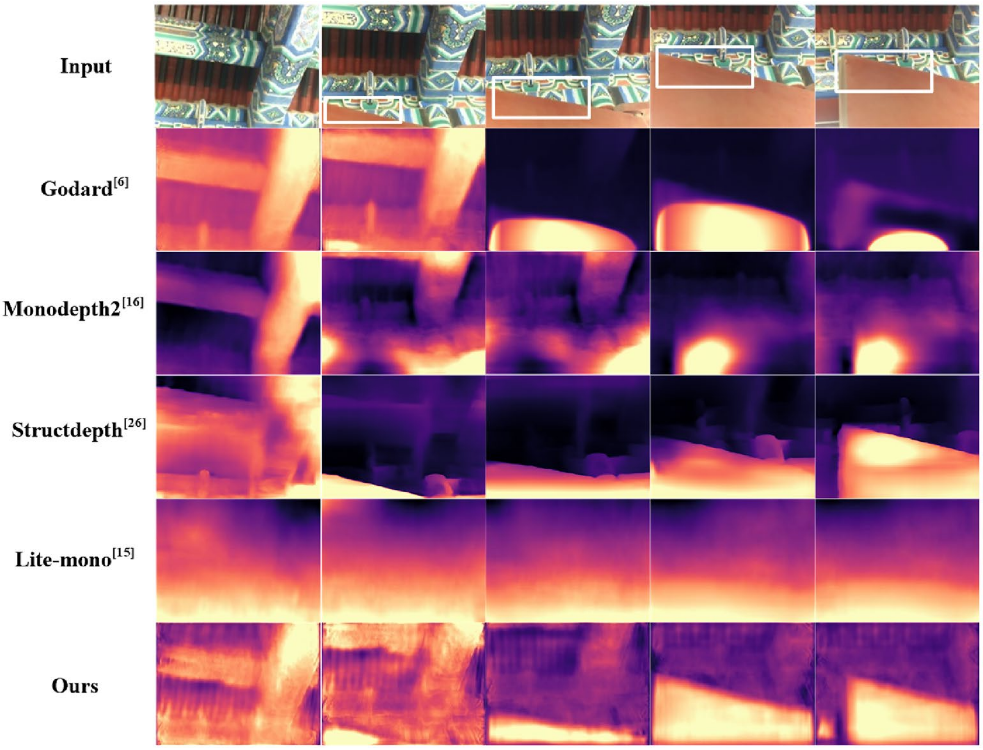


Fig. 8 "Pattern interference effect" for different models on the ancient architecture dataset

overlook the interference caused by repetitive texture patterns during feature extraction, which can affect the accuracy of depth estimation. Our method addresses this issue by capturing contextual relationships between different image regions.

Dac-CNN [30] also focuses on enhancing the feature learning process through an innovative cumulative convolution layer strategy, which integrates feature information from various directions to improve depth estimation accuracy. However, the cumulative convolution layers relies on pre-defined accumulation directions, which may lack sufficient adaptability and flexibility when dealing with complex and variable textures and structures. Lite-mono focuses mainly on reducing the computational complexity and parameter count of the model, but this may compromise performance in intricate scenes.

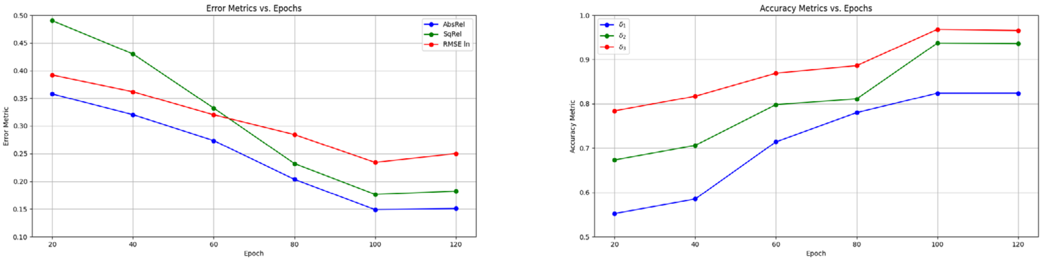
The improved methods based on classic frameworks include Multi-Warp [9], Johnston [11], and Mono-FiVI [28]. Multi-Warp enhances Godard's method by incorporating purposeless multi-reconstruction loss. Johnston builds upon Monodepth2 by integrating discrete disparity volume functionalities. Mono-FiVI introduces standard-view depth consistency and two-scale-aware depth consistency losses for regularization and distillation, further improving depth estimation accuracy. However, these methods do not offer specific strategies for handling complex structures and repetitive textures in local scenes of ancient architecture, which may limit their effectiveness in this area.

In summary, our method uses a self-attention module to capture correlations between local image features and considers the positional relationships in the image. Therefore, depth estimation in repetitive regions of texture structures is significantly superior to that of most other methods. Additionally, deepening the network architecture does not significantly improve depth estimation performance, while Monodepth2 and Johnston only improve accuracy at the third threshold. A possible reason for this is that the vanishing gradient phenomenon becomes more pronounced with deeper network architectures, leading to deterioration in depth estimation model performance. However, it may also be due to overfitting.

As shown in Fig. 8, when the "pattern interference" in the box gradually increases, the depth estimation of other methods for the entire image is seriously affected. However, our method is the least affected one.

In summary, our method trained with calibrated stereo image pairs and uses the self-attention module could effectively alleviate the pattern interference effects by the repeated textural and structural image patterns. We observed that the deepening of the network architecture cannot significantly improve the depth estimation performance, as Monodepth2 [16] and Johnston[11] only improve the third threshold accuracy. A possible reason could be that the scale effect of pattern interference has some "intrinsic scale" as advocated in multi-scale space theory [31], and further increasing the receptive field by deepening the network architecture could "saturate" the improvement. Other possible reasons could be that our dataset is relatively small, the capacity of deeper networks is not fully exploited. For the models with the PoseNet, relative pose errors estimated on-the-fly must also be a possible error source.

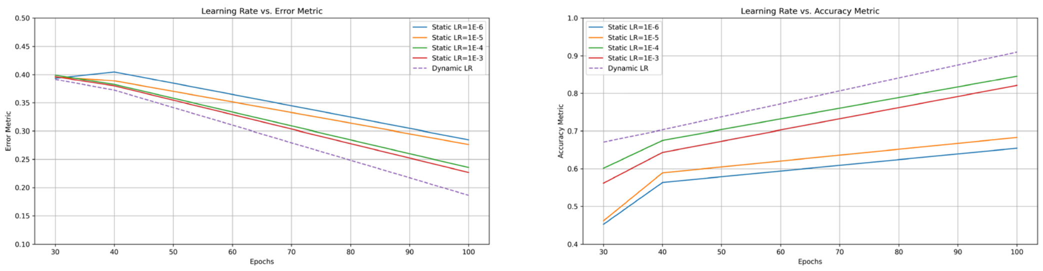
The impact of epochs on model performance: The Fig. 9 that when the training period (epoch) is too short, the model shows higher error and lower accuracy, indicating that it hasn't sufficiently learned the features of the ancient architecture. As the epoch increases, the model's performance improves significantly, reaching its best at 100 epochs. However, beyond 100 epochs, the changes in error and accuracy metrics stabilize, and there's even a trend of overfitting. Due to the complex textures and structures in the ancient architecture dataset, a longer training period is necessary to fully capture these features, making 100 epochs the optimal choice. In summary, increasing the training period according to the dataset's characteristics can significantly enhance model accuracy and stability, but exceeding a certain threshold may not bring additional performance gains and could even harm model performance due to overfitting.



(a) Error Metrics vs. Epochs

(b) Accuracy Metrics vs. Epochs

Fig. 9 The impact of epochs on model performance



(a) Error Metrics vs. Epochs (b) Accuracy Metrics vs. Epochs

Fig. 10 The impact of learning rates on model performance

Table 1 Comparison of model accuracy and error on ARCHITECT dataset

Method	Network	AbsRel	SqRel	RMSE	RMSE ln	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Godard [6] (CVPR'17)	ResNet18	0.1787	0.2619	0.839	0.289	0.801	0.914	0.946
Monodepth2 [16] (ICCV'19)	ResNet18	0.2853	0.3002	0.954	0.392	0.341	0.634	0.967
StructDepth [26] (ICCV'21)	ResNet18	0.1683	0.1762	0.807	0.234	0.856	0.930	0.952
multi-warp [9] (TMM'22)	ResNet18	0.2513	0.2142	0.833	0.286	0.785	0.885	0.906
Swindepth [27] (ICRA'23)	ResNet18	0.3450	0.285	0.276	0.363	0.527	0.814	0.933
Lite-mono [15] (CVPR'23)	ResNet18	0.2973	0.3125	0.978	0.397	0.333	0.562	0.690
Mono-ViFI [28] (ECCV'24)	ResNet18	0.3296	0.4818	1.182	0.522	0.339	0.587	0.774
Ours	ResNet18	**0.1488**	**0.1763**	**0.711**	**0.234**	**0.812**	**0.937**	**0.968**
Garg et al. [5] (ECCV'16)	ResNet50	0.3730	0.5775	1.257	0.540	0.338	0.590	0.775
Godardet al. [6] (CVPR'17)	ResNet50	0.1906	0.2161	0.770	0.271	0.738	0.912	0.950
Monodepth2 [16] (ICCV'19)	ResNet50	0.2854	0.2910	0.935	0.385	0.341	0.629	**0.963**
multi-warp [9] (TMM'22)	ResNet50	0.2379	0.2078	0.806	0.274	0.782	0.885	0.906
Dac-CNN [30] (ICCV'23)	ResNet50	0.1443	0.1910	0.723	0.242	0.824	0.933	0.961
Xiong et al. [29] (EAAI'24)	ResNet50	0.1673	0.1978	0.722	0.288	0.812	0.928	0.957
Ours	ResNet50	**0.1434**	**0.1837**	**0.711**	**0.238**	**0.824**	0.934	0.961
Johnston et al. [11] (CVPR'20)	ResNet101	0.2857	0.3001	0.950	0.390	0.341	0.631	**0.961**
multi-warp [9] (TMM'22)	ResNet101	0.2470	0.2014	0.847	0.276	0.793	0.888	0.908
Ours	ResNet101	**0.2035**	**0.2322**	**0.831**	**0.304**	**0.694**	**0.896**	0.948

Bold values indicate the optimal results for the respective metrics within each specific network structure

*The impact of learning rates on model performance:*The Fig. 10 illustrates the impact of learning rates on model performance, with error metrics representing the average of three error measures (AbsRel, SqRel, RMSE ln) and accuracy metrics (δ_1, δ_2, δ_3) representing the average of three accuracy measures. In this context, the static learning rate refers to maintaining a constant learning rate throughout the training period (e.g., epoch = 100). In contrast, the dynamic learning rate strategy involves starting with an initial learning rate of 1e-4, and adjusting it based on the epoch: 1e−4 for epochs < 30, (1e−4)/2 for epochs between 30 and 40, and (1e−4)/4 for epochs > 40.

The Fig. 10 shows that the dynamic learning rates results in lower training errors and higher accuracy across various training epochs, especially with longer training periods (100 epochs), where its effectiveness is most pronounced. This is because the dynamic learning rate can adapt to the training process: a higher learning rate at the beginning helps the model converge faster, while a lower learning rate in later epochs allows for finer adjustments and avoids overshooting minima in the loss landscape. This adaptability enhances the model's ability to capture complex patterns and improves its overall performance.

In contrast, the static learning rate exhibits relatively consistent performance across different training stages, which may not achieve the best results in the later stages of training due to its inability to adapt to the changing learning needs of the model. Overall, the dynamic

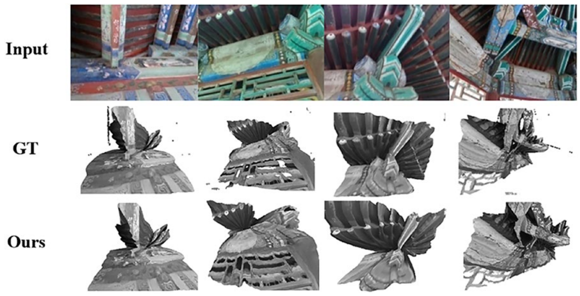


Fig. 11 The reconstructed point cloud quality evaluation by Chamber distance on our ARCHITECT dataset

learning rate strategy outperforms the static learning rate by effectively reducing training errors and enhancing model accuracy and stability through its flexible adjustment of the learning rate.

Point cloud reconstruction The Chamfer Distance (CD) is used to measure the quality of the reconstructed 3D point cloud with the ground truth [32]. With the estimated depth and the calibrated camera intrinsic parameters, the corresponding 3D point cloud in the camera coordinate system is computed, and this point cloud is compared with the point cloud obtained by the ground-truth depth. Given two point sets S1 and S2, the CD is defined as in (16). The smaller the distance, the better the reconstructed point cloud to the ground-truth.

$$d_{CD}(S_1, S_2) = \frac{1}{S_1}\sum_{x\in S_1}\min_{y\in S_2}\|x-y\|_2^2 + \frac{1}{S_2}\sum_{x\in S_2}\min_{y\in S_1}\|x-y\|_2^2 \tag{16}$$

The comparative results on our ARCHITECT dataset are listed in Table 2, and Fig. 11 shows some reconstructed point clouds. From Table 2 and Fig. 11, it can be seen that although our method has a smaller CD compared with the peers, some evident errors still exist, and more efforts are needed to tackle the pattern interference problem.

Experiments on KITTI data

As KITTI dataset is also captured with a stereo rig albeit under driving scenario, our method is also evaluated on this widely used dataset. Following the Eigen-split [21], 22,600 images were used for training, 888 for verification, and 697 for testing.

To comprehensively evaluate the performance and robustness of our method, we conducted additional tests on the KITTI dataset. The comparative experimental results on KITTI dataset are shown in Fig. 12. From Fig. 12, we can see that compared with other methods, the depth image boundary estimated by our method is clearer and smoother in the region where the depth remains unchanged.

Conclusion

The research successfully reveals the challenges faced by existing SIDE methods when applied to images of Chinese ancient architectures, which are characterized by rich repetitive textures and structural patterns. To achieve precise reconstruction of the local structures of Chinese ancient architectures, we propose an unsupervised depth estimation method using a self-attention mechanism. Trained with stereo image pairs,

Table 2 Chamfer distance for the ancient Chinese architectural dataset (Units: mm)

Method	Godard [6]	Johnston [11]	StructDepth [26]	MultiWarp [9]	Lite-mono [15]	Ours
CD	33.15	47.05	52.10	37.14	58.34	**27.79**

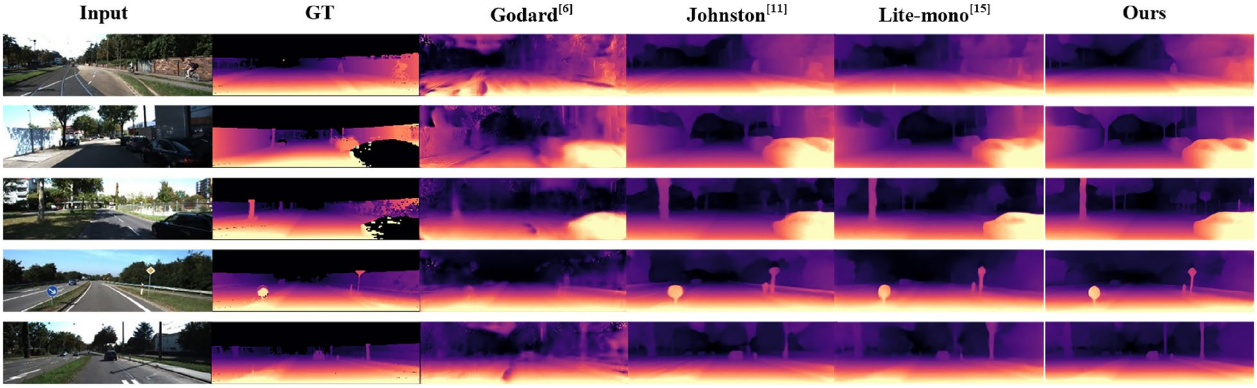


Fig. 12 Visualized depths of testing images on the KITTI dataset

the method mitigates pattern interference and demonstrates significant potential in maintaining detail and edge clarity.

Future work will explore the potential of multi-frame depth estimation methods and adapt the principles of positional encoding in transformers to further refine our method. Additionally, utilizing the known shapes and structures of architectural elements to provide more accurate information for the digital reconstruction of cultural heritage is also a promising direction for research.

In summary, the research is not only a preliminary attempt to address the issue of pattern interference in the digital preservation of Chinese ancient architecture but also lays a solid foundation for more targeted and effective solutions in the future, providing direction for the high-precision three-dimensional reconstruction of local structures of Chinese ancient architectures.

Abbreviations

SIDE	Single image-based depth estimation
3D	Three dimensions
CNN	Convolutional neural network
Conv	Convolution module
Resblock	Residual module
ReLU	Rectified linearunit
BN	Batch normalizatio
Upconv	Deconvolution module
Get-disp	Disparity module
SSIM	Structural Similarity Index Measure
ARCHITECT	Ancient Chinese architectures
NYU	New York University Depth Dataset V2
KITTI	Karlsruhe Institute of Technology and Toyota Technological Institute
RTX	Ray Tracing Texel eXtreme
CD	Chamfer distance

Acknowledgements
Not applicable.

Author contributions
XY wrote the main manuscript text; LH and JZ revised it critically for important intellectual content; XY conducted the experiment and collected the data. All authors reviewed the manuscript.

Funding
This work is supported by the National Natural Science Foundation of China (NSFC) under Grant No. 62273248.

Availability of data and materials
The datasets used and/or analysed during the current research are available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate
Not applicable.

Competing interests
The authors declared that they have no competing interests to this work.

Received: 10 June 2024 Accepted: 22 August 2024
Published online: 30 August 2024

References

1. Liu X, Liu Y, Wang K, Zhang Y, Lei Y, An H, Wang M, Chen Y. A color prediction model for mending materials of the Yuquan Iron Pagoda in China based on machine learning. Herit Sci. 2024;12(1):183.
2. Ming Y, Meng X, Fan C, Yu H. Deep learning for monocular depth estimation: a review. Neurocomputing. 2021;438:14–33.
3. Yan L, Yu F, Dong C. EMTNet: efficient mobile transformer network for real-time monocular depth estimation. Pattern Anal Appl. 2023;26(4):1833–46.
4. Li S, Shi J, Song W, Hao A, Qin H. Hierarchical object relationship constrained monocular depth estimation. Pattern Recognit. 2021;120: 108116.
5. Garg R, Bg VK, Carneiro G, Reid I. Unsupervised cnn for single view depth estimation: geometry to the rescue. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VIII 14. Springer; 2016; p. 740–56.
6. Godard C, Mac Aodha O, Brostow GJ. Unsupervised monocular depth estimation with left-right consistency. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017; p. 270–9.
7. Repala VK, Dubey SR. Dual cnn models for unsupervised monocular depth estimation. In: Pattern Recognition and Machine Intelligence: 8th International Conference, PReMI 2019, Tezpur, India, December 17–20, 2019, Proceedings, Part I, Springer; 2019; p. 209–17 .
8. Tosi F, Aleotti F, Poggi M, Mattoccia S. Learning monocular depth estimation infusing traditional stereo knowledge. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019; p. 9799–809.
9. Ling C, Zhang X, Chen H. Unsupervised monocular depth estimation using attention and multi-warp reconstruction. IEEE Trans Multimed. 2022;24:2938–49.
10. Wang X, Girshick R, Gupta A, He K. Non-local neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018; p. 7794–803.
11. Johnston A, Carneiro G. Self-supervised monocular trained depth estimation using self-attention and discrete disparity volume. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020; p. 4756–65.
12. Ji P, Li R, Bhanu B, Xu Y. Monoindoor: Towards good practice of self-supervised monocular depth estimation for indoor environments. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021; p. 12787–96.
13. Yan J, Zhao H, Bu P, Jin Y. Channel-wise attention-based network for self-supervised monocular depth estimation. In: 2021 International Conference on 3D Vision (3DV), IEEE; 2021. p. 464–73 .
14. Lee M, Hwang S, Park C, Lee S. Edgeconv with attention module for monocular depth estimation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2022; p. 2858–67.
15. Zhang N, Nex F, Vosselman G, Kerle N. Lite-mono: A lightweight cnn and transformer architecture for self-supervised monocular depth estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023; p. 18537–46.
16. Godard C, Mac Aodha O, Firman M, Brostow GJ. Digging into self-supervised monocular depth estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019; p. 3828–38.
17. Dhond UR, Aggarwal JK. Structure from stereo-a review. IEEE Trans Syst Man Cybern. 1989;19(6):1489–510.
18. Zhang R, Tsai P-S, Cryer JE, Shah M. Shape-from-shading: a survey. IEEE Trans Pattern Anal Mach Intell. 1999;21(8):690–706.
19. Wang Z, Bovik AC, Sheikh HR, Simoncelli EP. Image quality assessment: from error visibility to structural similarity. IEEE Trans Image Process. 2004;13(4):600–12.
20. Zhang Z. A flexible new technique for camera calibration. IEEE Trans Pattern Anal Mach Intell. 2000;22(11):1330–4.
21. Eigen D, Puhrsch C, Fergus R. Depth map prediction from a single image using a multi-scale deep network. Adv Neural Inform Process Syst. 2014;27:2366–74.
22. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L, et al. Pytorch: An imperative style, high-performance deep learning library. Adv Neural Inform Process Syst. 2019;32:8024–35.

23. Liu F, Shen C, Lin G, Reid I. Learning depth from single monocular images using deep convolutional neural fields. IEEE Trans Pattern Anal Mach Intell. 2015;38(10):2024–39.
24. Zhou T, Brown M, Snavely N, Lowe DG. Unsupervised learning of depth and ego-motion from video. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017; p. 1851–8.
25. Casser V, Pirk S, Mahjourian R, Angelova A. Depth prediction without the sensors: leveraging structure for unsupervised learning from monocular videos. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, 2019; p. 8001–8.
26. Li B, Huang Y, Liu Z, Zou D, Yu W. Structdepth: leveraging the structural regularities for self-supervised indoor depth estimation. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2021; p. 12643–53.
27. Shim D, Kim HJ. Swindepth: unsupervised depth estimation using monocular sequences via swin transformer and densely cascaded network. In: 2023 IEEE International Conference on Robotics and Automation (ICRA), IEEE; 2023. p. 4983–90 .
28. Liu J, Kong L, Li B, Wang Z, Gu H, Chen J. Mono-ViFI: a unified learning framework for self-supervised single- and multi-frame monocular depth estimation. arXiv:https://arxiv.org/abs/2407.14126. 2024.
29. Xiong M, Zhang Z, Liu J, Zhang T, Xiong H. Monocular depth estimation using self-supervised learning with more effective geometric constraints. Eng Appl Artif Intell. 2024;128: 107489.
30. Han W, Yin J, Shen J. Self-supervised monocular depth estimation by direction-aware cumulative convolution network. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023; p. 8613–23.
31. Lindeberg T. Scale-space theory in computer vision—introduction and overview, vol. 1994; p. 1–28. https://doi.org/10.1007/978-1-4757-6465-9.
32. Sun X, Wu J, Zhang X, Zhang Z, Zhang C, Xue T, Tenenbaum JB, Freeman WT. Pix3d: dataset and methods for single-image 3d shape modeling. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018; p. 2974–83.

Publisher's Note