IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 13 April 2024, accepted 29 April 2024, date of publication 2 May 2024, date of current version 9 May 2024.

*Digital Object Identifier 10.1109/ACCESS.2024.3396148*

RESEARCH ARTICLE

# Research on Super-Resolution Enhancement Technology Using Improved Transformer Network and 3D Reconstruction of Wheat Grains

**YIJUN TIAN[1,2,3], JINNING ZHANG[2], ZHONGJIE ZHANG[2], AND JIANJUN WU[1,3], (Member, IEEE)**

[1]College of Information Science and Engineering, Henan University of Technology, Zhengzhou 450001, China
[2]Academy of National Food and Strategic Reserves Administration, Beijing 100037, China
[3]Key Laboratory of Grain Information Processing and Control, Henan University of Technology, Ministry of Education, Zhengzhou 450001, China

Corresponding author: Jianjun Wu (t19138015212@gmail.com)

This work was supported in part by the "Self-Selected Topic" of the Academy of National Food and Strategic Reserves Administration's Scientific Research Institute "Study on Wheat Grain Moisture Detection Method Based on 3D Point Cloud Reconstruction" under Grant JY2202, and in part by the National Key Research and Development Program of China under Grant 2018YFD0401404.

**ABSTRACT** Three-dimensional reconstruction plays a crucial role in capturing plant phenotypes and expediting the process of agricultural informatization. However, the reconstruction of small objects such as plant specimens and grains often faces challenges like low two-dimensional image resolution and sparse textures. To enhance the three-dimensional reconstruction of plant specimens like wheat grains for comprehensive phenotypic characterization, this study proposes a novel super-resolution reconstruction network called T-transformer net. The network leverages the self-attention mechanism of Transformers to extract extensive global information from spatial sequences. By employing a hourglass block structure to construct spatial attention units and combining channel attention with window-based self-attention schemes, it effectively harnesses their complementary advantages. This encompasses utilizing global statistical data while capitalizing on potent local fitting capabilities. Evaluation of the model on publicly available datasets Set5, Set14, and Manga109 demonstrates superior overall performance of T-transformer net compared to mainstream super-resolution algorithms at upscaling factors of 2x, 3x, and 4x. In the context of super-resolution tasks involving wheat grain datasets, the peak signal-to-noise ratio reaches 42.89 dB, and the structural similarity index attains 0.9643. Subsequently, we subject the super-resolved wheat grain images to three-dimensional reconstruction. Through comprehensive extraction of high-level semantic information by neural networks, the reconstruction accuracy is improved by 38.96% compared with the unprocessed image, effectively mitigating challenges arising from sparse textures and repetitive patterns in wheat grain structures. This study contributes valuable methodology and insights to the realm of three-dimensional reconstruction in botany, holding significant implications for advancing agricultural informatization.

**INDEX TERMS** Three-dimensional reconstruction, super-resolution reconstruction, wheat grains, transformer, channel attention.

## I. INTRODUCTION

Plant phenotypic traits to some extent reflect the influence of genes and the environment on characteristics such as plant yield, quality, and stress resistance [1], [2]. Researchers have combined computer vision technology with agriculture [3] to explore new methods for overcoming the limitations of traditional crop breeding. High-throughput and rapid measurement of plant phenotypes can facilitate the study and cultivation of crop varieties [4], thereby enhancing breeding efficiency and advancing the development of phenomics.

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

Currently, the main techniques for obtaining plant phenotypes include 2D images and 3D models. Among these, methods based on 2D images have made progress in capturing one-dimensional and two-dimensional phenotypic features [5], [6]. However, in cases of severe occlusion and complex structures, the accuracy of the acquired phenotypic features is low [7]. Additionally, 2D images cannot fully capture the three-dimensional structure of crops, leading to the loss of 3D phenotypic information. In the field of agriculture, 3D reconstruction methods can be divided into active vision and passive vision. Active vision methods include laser scanning [8], [9], structured light [10], shadow methods [11], Time-of-Flight (TOF) technology [12], radar technology [13], Kinect technology [14], etc. Passive vision methods involve obtaining image sequences through visual sensors and then reconstructing the 3D structure. This approach first captures image sequences, extracts useful information, performs reverse modeling, and obtains the object's 3D structure. However, active vision methods are expensive, influenced by manual operations, time-consuming, and less widespread. In contrast, passive vision 3D reconstruction only requires capturing 2D images of plants with a camera and applying relevant algorithms to complete the 3D reconstruction. It offers advantages such as low cost, ease of use, and wide applicability [15]. Many researchers have reconstructed plant crops into 3D point clouds to extract 3D phenotypic information of plants. For instance, Chenxi et al. used a 3D scanner to acquire wheat plant phenotypic parameters [16]. Zhibin et al. employed a binocular vision system and SURF algorithm to obtain 3D point cloud information of crops like mustard and celery [17]. Paproki et al. measured the 3D model of cotton plants along with stems and leaves using multi-view stereo vision and segmentation techniques [18]. These studies indicate that rapid and high-precision reconstruction of plant 3D models is of great significance for accelerating the development of agricultural informatics.

In terms of wheat grains [19], [20], [21], image processing techniques have successfully extracted key phenotypic parameters such as spike number, spikelet number per row, and spikelet number per spike. However, challenges still exist in reconstructing the 3D model of grains and obtaining 3D information. Due to the small volume of wheat grains, similar surface textures, low pixel ratios, and a lack of sufficient feature information, accurate reconstruction of detailed features is difficult. Traditional passive methods perform well in the ideal Lambertian reflectance model but show lower accuracy when measuring the similarity in weak textures, repetitive textures, and non-Lambertian areas.

In recent years, a plethora of passive vision-based 3D reconstruction methods have emerged. Moulon et al. [22] proposed a motion recovery structure algorithm based on adaptive parameters. While achieving high reconstruction accuracy, this method is computationally intensive. Bao et al. [23] introduced a semantic motion recovery structure algorithm that enhances algorithm robustness through the recognition and estimation of high-level semantic information such as regions and objects in the 3D scene. Wu et al. [24] presented the VisualSFM algorithm, which employs preprocessed conjugate gradient descent to improve computational efficiency while maintaining accuracy. Schänberger et al. [25] introduced the COLMAP algorithm, enhancing key steps such as geometric calibration, viewpoint selection, and triangulation. It has made significant progress in both reconstruction accuracy and completeness, standing out as one of the best-performing algorithms among traditional passive methods. With deeper exploration of deep learning techniques, scholars have found that convolutional operations can introduce global semantic information, and supervised feature extraction can address stereo matching challenges under adverse conditions such as weak textures and non-Lambertian characteristics [26], [27]. Eigen et al. [28] conducted experiments on monocular depth estimation using convolutional neural networks, dividing the network into a global coarse estimation network and a local fine estimation network. This provided essential guidance for further research in multi-view 3D reconstruction. Yao et al. [29] introduced an end-to-end neural network model called MVSNet, pioneering multi-view depth learning for 3D reconstruction. Zhang et al. [30] proposed Vis-MVSNet, an end-to-end network structure that considers pixel visibility information, explicitly infers and integrates pixel-level occlusion information in the MVS network through uncertainty estimation and uses joint inference of pairwise uncertainty maps and depth maps as weighted guidance in multi-view cost volume fusion, effectively suppressing the adverse effects of occluded pixels.

3D reconstruction based on single-view images is an active research field in computer vision and graphics [31], which aims to reconstruct the 3D structure from a single image. Although 3D reconstruction based on single-view images has made some progress, single-view 3D reconstruction still faces many challenges because recovering three-dimensional information from only one two-dimensional image involves a high degree of uncertainty and ambiguity, including the consistency of reconstruction quality. And the ability to process complex scenes, as well as the generalization ability to different object types and different viewing angles. Considering that this research hopes to achieve high-precision acquisition of the three-dimensional point cloud model of wheat grains, 3D reconstruction based on single-view images is not used to achieve the goal of this article.

For low-resolution or blurry images, researchers have attempted to enhance image quality using super-resolution techniques [32]. Super-resolution technology aims to enhance the resolution and clarity of images by recovering detailed information from low-resolution images. This technology has a wide range of applications in various fields, especially in medical imaging, remote sensing, and facial imaging, where significant progress has been made. In the field of medical imaging [33], super-resolution technology can improve the resolution of imaging techniques such as magnetic resonance imaging (MRI) and computed

tomography (CT), providing doctors with more diagnostic information. In remote sensing [34], super-resolution technology can enhance the quality of satellite images and aerial images, providing more accurate data for environmental monitoring and urban planning. In the field of human-computer interaction [35], super-resolution technology can enhance the performance of video surveillance and facial recognition systems, improving the accuracy of image recognition. Traditional image super-resolution reconstruction methods primarily include three categories: interpolation-based algorithms such as nearest neighbor interpolation [36]; degradation model-based algorithms like iterative back-projection [37]; and learning-based algorithms including sparse coding methods [38]. Traditional super-resolution algorithms have achieved considerable success, but with increasing scale factors from 2x to 4x, 8x, the required information for super-resolution reconstruction becomes more demanding. Artificially defined prior knowledge cannot meet the requirements, making it difficult to achieve high-quality image reconstruction. With the significant success of deep learning in computer vision, Dong et al. [39] first introduced deep learning methods to image super-resolution tasks, achieving better results than traditional methods due to the powerful learning capacity of neural networks. Subsequent researchers proposed a series of continuously optimized algorithm models, from the earliest SRCNN (super-resolution convolutional neural network) model based on convolutional neural networks [40] used in the waifu2x project to the SRGAN (super-resolution generative adversarial network) [41] model based on generative adversarial networks. The field of deep learning-based image super-resolution reconstruction continues to make new breakthroughs. Jiang et al. [42] proposed a combination of association learning and Transformer architecture to solve the image denoising problem. Jiang et al. [43] proposed A dynamic association learning model that combines self-attention and convolution in the image restoration task to find an effective combination between self-attention mechanism and traditional convolution to take advantage of the advantages of both. Xiao et al. [44] proposed an improved model for the temporal dependence of the self-attention mechanism, and improved the performance of video super-resolution at local and global scales, Jiang et al. [45] proposed a hierarchical dense recursive network structure for image super-resolution. However, existing models do not consider the relationship between channels to better grasp the global relationship of image information.

Therefore, addressing the aforementioned issues in current research on 3D reconstruction of wheat grains, this study proposes a novel super-resolution network, namely the T-transformer net, based on the Transformer architecture. In the context of 3D reconstruction of wheat grains, an economical and high-precision reconstruction approach is investigated. The key findings and notable objectives of this study can be summarized as follows:

- We establish an economically viable data acquisition system by employing industrial-grade cameras for capturing RGB images. Simultaneously, a standardized dataset of wheat grain samples is curated, encompassing four distinct varieties, each represented across twelve moisture gradients.
- A Transformer-based super-resolution reconstruction network is introduced, incorporating channel attention modules to enhance feature utilization and information propagation, thereby extending the model's receptive field. Compared with existing super-resolution networks, the T-transformer net incorporates a channel attention mechanism, leading to significant improvements in image quality, as evidenced by PSNR and SSIM metrics. Furthermore, it specifically addresses the challenges associated with processing wheat grain images. This includes a demonstrated stronger ability to handle sparse textures and distinguish difficult repeated patterns.
- The proposed model is evaluated using the custom-designed dataset of wheat grain samples, and subsequently applied to 3D reconstruction using Vis-MVSNet. By adeptly extracting high-level semantic information from images, the reconstruction precision of wheat grains is notably enhanced under challenging conditions such as sparse or repetitive textures. Furthermore, the insights garnered herein serve as a reference for attaining high-precision 3D models of other agricultural products.

## II. MATERIALS AND METHODS

### A. MATERIALS

This study takes wheat as the research object and conducts experiments using four different varieties of wheat grains produced in North China in 2023, namely Huaimai 22, Vanke 189, Zimai 19 and Yangmai 15. Following the ASAE method, wheat samples were subjected to continuous drying in a 130°C oven for 19 hours. The difference in mass of the wheat samples before and after drying was measured to determine the moisture content of the wheat, resulting in an initial moisture content of 13.7% (w.b.). The wheat samples were grouped and labeled as samples 1 to 12. Each sample, weighing (200±1) grams, was sealed in self-sealing bags and stored in a constant-temperature chamber at 4°C. Prior to usage, the wheat samples were removed and gradually brought to room temperature.

The method outlined in literature [46] for adjusting grain moisture was employed to establish experimental samples with moisture contents ranging from 8% to 30% in twelve distinct gradients. The formula for calculating the change in moisture mass during the grain moisture adjustment process is provided below 1:

$$\triangle m = \begin{cases} \dfrac{(w_f - w_i)}{1 - w_f} m_i, w_f \geq w_i \\ \dfrac{(w_i - w_f)}{1 - w_f} m_i, w_f < w_i \end{cases} \quad (1)$$

In the equation provided: $m_i$ represents the initial mass of the grain to be conditioned; $\Delta m$ represents the change in mass of water during the conditioning process; $w_i$ represents the initial moisture content of the grain; $w_f$ represents the moisture content of the grain after conditioning. The specific steps for moisture adjustment are as follows:

#### 1) STEPS FOR INCREASING GRAIN MOISTURE CONTENT

1) Take a sample of 0.5 kg of grain, and calculate the required increase in water mass $\Delta m$ using Equation 1.
2) Add distilled water with a mass of $\Delta m$ to the grain sample and stir for 10 minutes to ensure uniform distribution of moisture.
3) Place the water-added grain into a sealed plastic bag, remove excess air, and store it in a refrigerator at 4°C for 72 hours. This allows the moisture to be thoroughly absorbed into the grain while the low temperature minimizes biological activity.
4) Transfer the grain sample to an open container at room temperature (25°C), and let it stand for 12 hours. During this period, stir every 2 hours. This process helps the grain sample return to room temperature and facilitates the evaporation of any unabsorbed surface moisture.

#### 2) STEPS FOR DECREASING GRAIN MOISTURE CONTENT

1) Take a sample of 0.5 kg of grain, and calculate the required decrease in water mass $\Delta m$ using Equation 1.
2) Place the grain sample in a 130°C drying oven. Remove the sample at regular intervals, weigh it, and thoroughly mix it to ensure uniformity.
3) Calculate the mass loss of the grain sample. If the mass loss is less than $\Delta m$, repeat step 2 to continue the drying process. If the mass loss exceeds $\Delta m$, the conditioning process is complete.

By employing the conditioning process, 12 samples of wheat with varying moisture content were obtained, as illustrated in Table 1.

**TABLE 1. Moisture Content Variation Among Different Varieties of Grain.**

| Moisture gradient | Huaimai 22 | Wanke 189 | Zimai 19 | Yangmai 15 |
|---|---|---|---|---|
| 8 | 8.691 | 8.179 | 7.899 | 7.924 |
| 10 | 10.487 | 10.453 | 9.900 | 9.645 |
| 12 | 12.239 | 12.148 | 11.160 | 11.452 |
| 14 | 14.257 | 14.838 | 13.553 | 13.832 |
| 16 | 16.593 | 16.260 | 15.411 | 15.709 |
| 18 | 17.757 | 18.533 | 17.976 | 17.916 |
| 20 | 20.289 | 20.754 | 19.535 | 19.559 |
| 22 | 22.734 | 22.527 | 21.623 | 22.305 |
| 24 | 24.453 | 24.708 | 24.313 | 24.003 |
| 26 | 25.863 | 25.679 | 25.499 | 25.550 |
| 28 | 28.575 | 28.446 | 27.447 | 28.707 |
| 30 | 30.264 | 29.781 | 29.885 | 29.978 |

### B. DATASET ACQUISITION

#### 1) IMAGE ACQUISITION

During the image acquisition phase, we employed the following method to capture images of wheat grains. Initially, wheat grains were positioned at the suction nozzle location, and utilizing the action of an air compressor, the wheat grains were affixed to the nozzle. Through the rotation of a stepper motor, the wheat grains adhered to the nozzle were set in rotational motion. Subsequent to each complete rotation executed by the stepper motor, the camera captured an image of the wheat grain. The imaging equipment used included a Canon EOS R5 C digital camera and a Canon EF 100mm f/2.8L IS USM macro lens, both sourced from Tokyo, Japan. The air compressor operated at a speed of 1450 revolutions per minute, with an airflow rate of 0.3 cubic meters per minute. The stepper motor, the Nanotec PD4-EX model, was set to rotate at 5 degrees per step, with a 5-second interval between captures. With these settings, we could capture a total of 72 images of wheat grains every 6 minutes, leading to a comprehensive image sequence over the week. In our study, we controlled image illumination using LED lights on the camera lens and conducted the capture in a darkroom to ensure consistent lighting and minimize external light interference, enhancing the experiment's validity.

The acquired images had a resolution of $8192 \times 5464$ pixels and were saved in JPG format. Macro lenses with exposure times ranging from 0.1 to 3 seconds were used during the capture process. The camera settings included an aperture of f/8, a focal length of 100 millimeters, and a lens-to-grain distance of 5 centimeters. To maintain the accuracy of experimental data and ensure model robustness, image acquisition was performed under these specific conditions for four different wheat varieties, with each variety comprising 50 grains per moisture level. A total of 600 grains per variety resulted in approximately 172,800 images. Furthermore, five grains from each moisture gradient were randomly chosen as the test set, ensuring a comprehensive representation within the dataset. This extensive collection of images is crucial for improving the model's generalization ability and robustness.

#### 2) IMAGE PREPROCESSING

This study identifies several distractions in the original images, including the suction device's nozzle and background noise points. These elements could lead to inaccuracies and disruptions in later stages, such as when training super-resolution models or aligning feature points for three-dimensional reconstruction. Thus, image preprocessing becomes a critical step.

Fig. 1 shows that, although the background appears black, the RGB values reveal subtle textures not actually pure black. These textures do not contribute to our analysis and can affect the accuracy of the super-resolution network by diminishing the precision of feature point extraction. This, in turn, could impact the precision of three-dimensional reconstruction and result in longer matching times. To mitigate this, we transformed the background to pure black using a brightness

**FIGURE 1.** Original wheat grain image.

threshold technique. This process involved converting the image to grayscale and then applying a threshold; pixels with values below the threshold were turned to pure black, while those above remained unchanged. For this study, we set the threshold at 40.

After eliminating background noise points, removing the nozzle area from the image becomes the next step. The color of the nozzle is quite similar to certain parts of the particles, this makes it challenging to depend only on color information for nozzle removal without risking damage to the main body. To resolve this, we incorporated spatial information. First, we sum the pixel values of each row and then subtract the pixel values between successive rows, resulting in an interpolated difference. Plotting these interpolated differences reveals a peak corresponding to the region with the most significant color variation, located above the particles. Using this peak as a reference, we crop the image, turning pixels outside this area to pure black to isolate the clean main body of the particle, as illustrated in Fig. 2.

**FIGURE 2.** The image after setting the background to pure black and removing the suction nozzle.

Furthermore, to mitigate the impact of rotational deviations introduced by the attachment system during image capture and to reduce the computational burden caused by excessive black regions during subsequent processing, we calculate the centroid of the colored information region—the area containing wheat particles. This centroid is then positioned as the center of the image, followed by cropping to achieve an image size of $1440 \times 2560$ pixels. Consequently, an image containing only the particles is obtained, as depicted in Fig. 3. These preprocessing steps are automated through our specially designed Python program, thereby enhancing the accuracy and efficiency of subsequent tasks.

**FIGURE 3.** Crop to remove the redundant background, and only keep the pixel part of the image of the wheat grain.

Post-image preprocessing, issues become observable within the dataset, as depicted in Fig. 4. These problems mainly stem from the inaccurate focus during capture, leading to blurred grain textures and unclear details. Given the relatively small diameter and volume of wheat grains, avoiding these issues during image acquisition is challenging. In response, one approach to address this challenge involves the increase of exposure time. However, to construct a three-dimensional model of the wheat grains, it is necessary to capture images from different angles. Merely extending exposure time would not only substantially increase the image capture time but also escalate workload and time costs.

**FIGURE 4.** Image of wheat grain with blurred parts.

Hence, in this study, we utilize super-resolution technology to tackle these problems. The core idea behind this approach is to improve image resolution, which simultaneously preserves clear texture information and compensates for the blurriness resulting from focus issues. Through this method, we are able to obtain clearer images without notably increasing image capture time and workload, thus providing improved input data for subsequent wheat grain three-dimensional model construction.

### C. METHODS

The network architecture is illustrated in Fig. 5, which comprises three main components: a shallow feature extraction module, a deep feature extraction module, and an image reconstruction module.

Given a low-quality image $I_{\mathrm{LQ}} \in R^{H\times W\times C_{in}}$ as input, with dimensions of height $H$, width $W$, and a channel count $C_{in}$ corresponding to the input image. The shallow feature extraction module is a $3 \times 3$ convolutional layer, expressed as $H_{\mathrm{SF}}(\cdot)$ is employed to extract shallow-level features, as indicated in (2), where $C$ represents the number of channels:

$$F_0 = H_{\mathrm{SF}}\left(I_{\mathrm{LQ}}\right), \tag{2}$$

Subsequently, the depth feature $F_{\mathrm{DF}} \in R^{H\times W\times C}$ is extracted from $F_0$:

$$F_{\mathrm{DF}} = H_{\mathrm{DF}}\left(F_0\right), \tag{3}$$

$H_{\mathrm{DF}}(\cdot)$ is a deep feature extraction module, which includes $K$ residual Fusion attention modules (RFAB) and a $3\times 3$ convolutional layer. The intermediate features $F_1, F_2, \ldots, F_K$ and the output deep features $F_{DF}$ are processed one by one as:

$$\begin{aligned} F_i &= H_{\mathrm{RSTB}_i}\left(F_{i-1}\right), i = 1, 2, \ldots, K, \\ F_{\mathrm{DF}} &= H_{\mathrm{CONV}}\left(F_K\right), \end{aligned} \tag{4}$$

where $H_{\mathrm{RFAB}_i}(\cdot)$ is the i-th RFAB module, and $H_{\mathrm{CONV}}$ is the last convolutional layer. As shown in Fig. 5, the residual Swin Transformer module is a residual module with $L$ Fusion attention layers (FAL) and a convolutional layer. Given the input feature $F_{i,0}$ of the i-th RFAB, we first extract intermediate features $F_{i,1}, F_{i,2}, \ldots, F_{i,L}$ by L Fusion attention layers as:

$$F_{i,j} = H_{\mathrm{FAL}_{i,J}}\left(F_{i,j-1}\right), j = 1, 2, \ldots, L, \tag{5}$$

$H_{\mathrm{FAL}_{i,j}}(\cdot)$ is the jth FAL layer in the ith RFAB. Subsequently, a convolutional layer is added before the residual connection. The output formula of the RFAB is:

$$F_{i,\mathrm{out}} = H_{\mathrm{CONV}_i}\left(F_{i,L}\right) + F_{i,0}, \tag{6}$$

$H_{\mathrm{CONV}_i}(\cdot)$ is the convolutional layer in the i-th RFAB. Many works have shown [47], [48] that convolution can help Transformer to obtain better visual representation or achieve easier optimization. Therefore, we incorporate an attentional convolution-based channel block into the standard Transformer block to enhance the representational power of the network. As shown in Fig. 5, after the first LayerNorm (LN) layer, a Channel Attention Block (CAB) is inserted. The CAB module is a network module that utilizes the channel attention mechanism. It is mainly used to enhance the deep learning network's attention to important features, thereby improving the performance of the network. The CAB module assigns different importance to the features of different channels, adaptively evaluates the importance of each channel (i.e., feature map), and automatically adjusts the weight distribution between channels according to task requirements, so that the network can pay more attention to the current task. A more helpful feature is that although the CAB module introduces additional computation, since its operations are mainly concentrated in the channel dimension, its computational overhead is smaller relative to spatial dimension operations. The CAB module is parallel to the Multi-headed Self-attention (MSA) layer to avoid conflicts between CAB and MSA in optimization and visual representation. If there is a conflict, multiply a small constant $\alpha$ by the output of the CAB. Among them, the MSA module

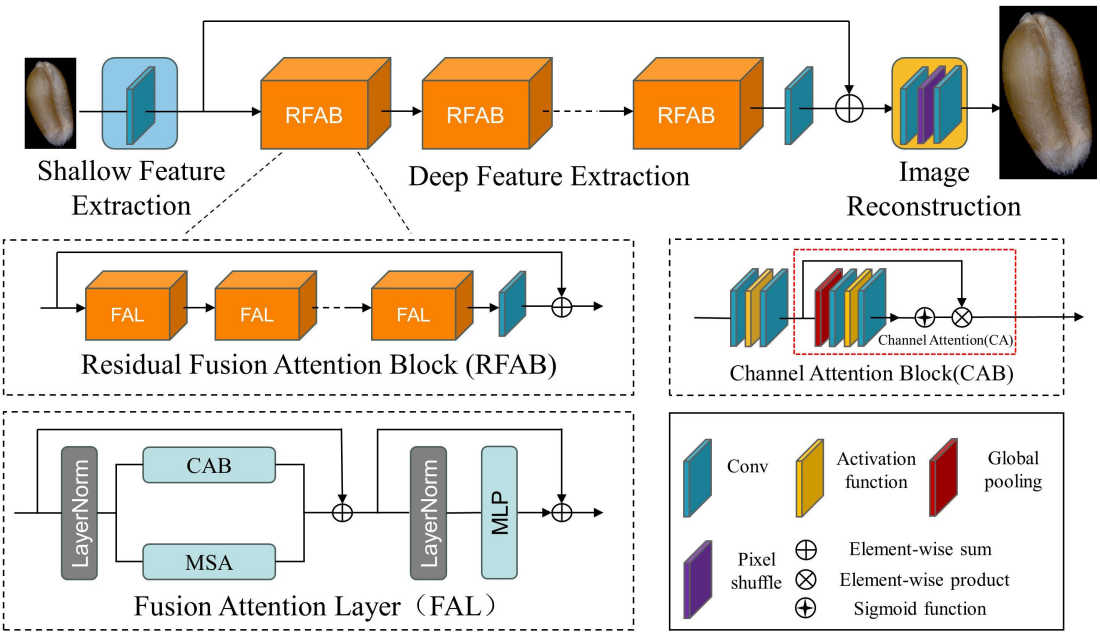



**FIGURE 5.** Network structure Image.

and the Self Attention layer are consistent with the standard Transformer structure [49]. For a given input feature $X$, the entire process of FAL is computed as:

$$\begin{aligned} X_N &= \mathrm{LN}(X), \\ X_M &= \mathrm{MSA}(X_N) + \alpha \mathrm{CAB}(X_N) + X, \\ Y &= \mathrm{MLP}(\mathrm{LN}(X_M)) + X_M, \end{aligned} \tag{7}$$

$X_N$ and $X_M$ represent intermediate features. $Y$ represents the output of the FAL. MLP is a multi-layer perceptron used to calculate the self-attention module. The channel attention module enhances the performance of deep learning networks by reinforcing the network's focus on crucial features and optimizing the utilization efficiency of these features. This module achieves performance improvement through adaptively assigning different weights to the features of each channel, thereby strengthening the focus on features that are more important for the current task. Traditional convolution operations are typically confined to local receptive fields, meaning that each unit in the output is only able to access information within its corresponding region, thus failing to capture contextual information beyond the region. The channel attention module, however, compresses the two-dimensional channel features into real-valued numbers with a global receptive field, effectively capturing the global distribution information of features in the channel dimension. This real-valued number reflects the overall importance of each channel feature. Another advantage of the channel attention mechanism is that, despite introducing additional computational steps, these operations primarily target the channel dimension. Compared to operations in the spatial dimension, the computational overhead is smaller, enabling the module to be efficiently integrated into various network architectures without imposing significant performance burdens. Given an input feature of size $H \times W \times C$ it is first divided into $\frac{HW}{M^2}$ local windows of size $M \times M$, and then self-attention is computed within each window. For a local window function $X_W \in R^{M^2 \times C}$, the query matrix, key matrix, and value matrix are computed by linear mappings such as $Q$, $K$, and $V$. Window-based self-attention is formulated as:

$$\mathrm{Attention}(Q, K, V) = \mathrm{SoftMax}\left(QK^T/\sqrt{d} + B\right)V, \tag{8}$$

Among them, $d$ represents the dimension of the key-value ratio. $B$ is the relative position code, and the calculation method is [49]. A CAB consists of two standard convolutional layers with a GELU activation [50] and a channel attention module, as shown in Fig. 5. Since the Transformer-based structure usually requires a large number of channels for token embedding, directly using constant-width convolutions will incur a large computational cost. Therefore, we compress the number of channels of the two convolutional layers with a constant $\beta$. For an input feature with $C$ channels, the number of output feature channels after the first convolutional layer is compressed to $C/\beta$, and then the feature is expanded to $C$ channels by the second layer. Next, the channel characteristics are adaptively scaled using a standard Channel Attention module [51].

The image reconstruction module aggregates shallow and deep features into features to reconstruct high-quality images $I_{\mathrm{RHQ}}$:

$$I_{\mathrm{RHQ}} = H_{\mathrm{REC}}(F_0 + F_{\mathrm{DF}}), \tag{9}$$

Among them, $H_{\mathrm{REC}}(\cdot)$ is the function of the reconstruction module. Shallow features mainly contain low frequencies,

while deep features focus on recovering lost high frequencies. T-transformer net can directly transmit low-frequency information to the reconstruction module through residual connection, helping the deep feature extraction module to focus on high-frequency information and stabilize training. For the implementation of the reconstruction module, we use sub-pixel convolutional layers [52] to upsample the features. In this task, the Loss function optimizes the parameters of T-transformer net by minimizing the L1 pixel loss:

$$\text{Loss} = \left\| I_{\text{RHQ}} - I_{\text{HQ}} \right\|_1, \tag{10}$$

$I_{\text{RHQ}}$ is obtained by taking $I_{\text{LQ}}$ as the input of T-transformer net, and $I_{\text{HQ}}$ is the corresponding real image. For classical and lightweight image SR, we only use the same L1 pixel loss as previous work, to demonstrate the effectiveness of the proposed network.

## III. RESULTS

### A. EXPERIMENTAL ENVIRONMENT AND CONFIGURATION

The primary experimental environment utilized Python 3.8.10, PyTorch 1.10.0, and CUDA 11.3. The specific host configurations are outlined in Table 2, while select experimental parameters can be found in Table 3.

**TABLE 2. Experimental environment.**

| Hardware | Specific Configuration |
|---|---|
| CPU | Intel(R) Xeon(R) Platinum 8255C |
| GPU | RTX 4090 |

**TABLE 3. Experimental parameters.**

| Experimental Parameters | Specific Parameters |
|---|---|
| Number of iterations | 100000 |
| batch size | 4 |
| Initial learning rate | $1 \times 10^{-4}$ |
| optimizer | adam |

### B. EVALUATION METRICS

In the experiment, the evaluation metrics of Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM) were utilized for quantitative assessment. PSNR is one of the most commonly used metrics for evaluating the reconstruction quality in lossy transformations, such as image compression and image inpainting. In the context of image super-resolution, PSNR is defined in terms of the maximum pixel value (L) and the Mean Squared Error (MSE) between images. Given a ground truth high-resolution image $h$ with $N$ pixels and a reconstructed image $s$, the PSNR between $h$ and $s$ is defined as follows:

$$\text{PSNR} = 10 \times log_{10}\left(\frac{L^2}{\frac{1}{N}\sum_{i=1}^{N}\left(I_h(i) - I_s(i)^2\right)}\right), \tag{11}$$

In the formula: $L = 255$ when the image pixel is represented by 8 bits. PSNR is only related to the mean square error at the pixel level and only concerns the difference between corresponding pixels. PSNR is currently the most widely used evaluation criterion in SR models. SSIM is used to measure differences in brightness, contrast, and structure between images. For a real high-resolution image $h$ and a reconstructed image $s$ with $N$ pixels, SSIM is defined as:

$$\text{SSIM} = \frac{(2\mu_s\mu_h + C_1)(2\sigma_s\sigma_h + C_2)(\omega_{sh} + C_3)}{(\mu_s^2 + \mu_h^2 + C_1)(\sigma_s^2 + \sigma_h^2 + C_2)(\sigma_s\sigma_h + C_3)}, \tag{12}$$

In the formula: $\mu_s$ represents the average value of image $s$, $\sigma_s$ represents the variance of image $s$, $\mu_h$ represents the average value of image $h$, $\sigma_h$ represents the variance of image $h$, $\omega_{sh}$ represents the covariance of image s and image $h$.

### C. EXPERIMENTAL RESULTS OF SUPER-RESOLUTION RECONSTRUCTION

#### 1) EFFECT OF DIFFERENT HYPERPARAMETERS

We initiated our study by training the network on the DIV2K dataset [53], followed by testing on the Manga109 dataset [54]. Evaluating at a magnification factor of 4, we employed PSNR as the benchmark metric to investigate the impact of varying channel numbers, RFAB quantities, and FAL quantities on performance. As depicted in Fig. 6, the influences of channel number, RFAB quantity, and FAL quantity on model performance are depicted. The observed outcomes indicated a positive correlation between PSNR and these three hyperparameters.

Regarding channel numbers, while performance demonstrated continuous enhancement with increasing channels, the total model parameters exhibited a quadratic growth trend. Striving for a balance between performance and model size, we opted for 120 channels for subsequent experiments. Regarding RFAB and FAL quantities, experimental findings revealed diminishing performance improvements with increasing module quantities. Thus, we chose 6 RFABs and 6 FALs to achieve a relatively compact model.

#### 2) ABLATION STUDY

We conducted experiments to demonstrate the effectiveness of the proposed CAB module. This article introduces an attention-based convolution channel block in the standard Transformer module. In order to verify that this choice is reasonable, we replaced the FAL module in the previous article. After removing the CAB module, the FAL module was replaced with the standard Transformer module as a comparison, and then train separately, obtain the weight file, and test it on the quantitative performance test set of the $\times 4$ SR Urban100 data set, and then compare it with the model effect after adding the CAB module.

As can be seen from the 4, compared with the baseline results, the performance gain of CAB is 0.1 dB. This result proves that after the introduction of the CAB module, thanks

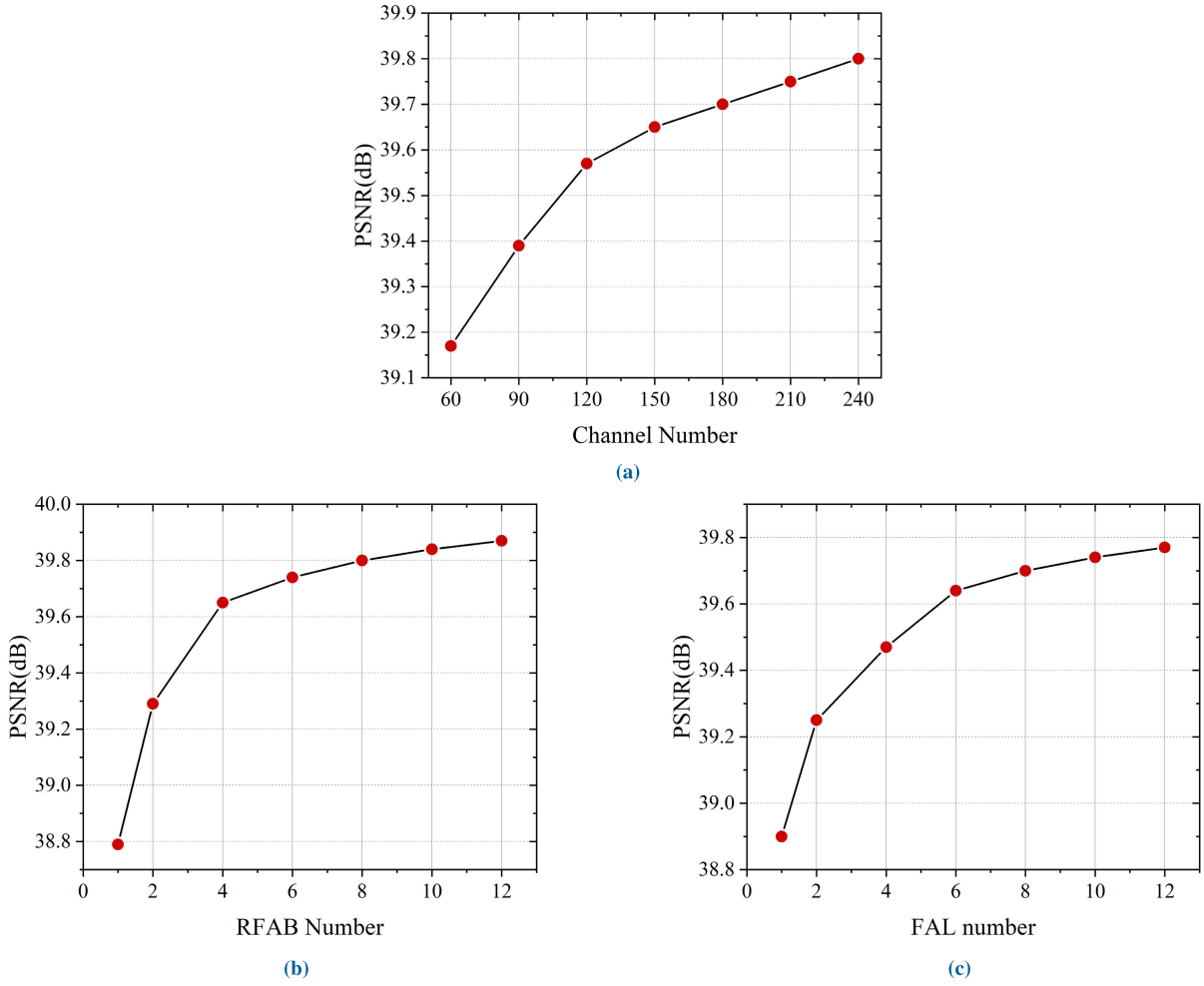



**FIGURE 6.** PSNR test results under different hyperparameters.

to the channel attention mechanism, the network can better capture the global information of the image for clearer and more accurate super-resolution reconstruction.

**TABLE 4.** Ablation study on the proposed CAB.

| | SSIM | PSNR |
|---|---|---|
| ✓ | 29.71 | 0.8362 |
| × | 29.61 | 0.8357 |

### 3) COMPARATIVE EXPERIMENT

To explore the reconstruction performance of the T-transformer net across different datasets, we compared it with contemporary high-performing reconstruction algorithms, namely EDSR [55], RCAN [51], SAN [56], and SwinIR [57]. For fair comparison and testing, we employed an image degradation mathematical model involving Gaussian blurring followed by downsampling on the original high-resolution images, yielding low-resolution images. Subsequently, we fed the initial high-resolution images into the model, enabling network learning and parameter optimization to generate the final high-resolution images. Throughout the experiments, we employed network models with the same iteration count for scaling factors of 2, 3, and 4 to ensure an objective and sound comparison. A summary of the experimental results is presented in Table 5.

The experimental results demonstrate the remarkable performance of the proposed model across various magnification levels and datasets. This better performance can be attributed to the utilization of a channel attention mechanism, which enables the network to better capture global information within images and consequently facilitate training and reconstruction based on these features. This mechanism was validated on datasets such as Set 5, Set 14, and Manga109, encompassing diverse image subjects and resolutions. The T-transformer net has a parameter size of 19.8M. Table 6 shows the performance and model size of other methods. Among these methods, SAN, RCAN, and SwinIR contain far fewer parameters, but at the expense of degraded performance. In contrast, our model has fewer parameters than EDSR but achieves higher performance, which means that our model can provide a good trade-off between performance and model complexity.

**TABLE 5.** The average PSNR/SSIM score of different models in wheat grain test set.

| Method | Scale | Set 5 [59] | | Set 14 [60] | | Manga109 [61] | |
|---|---|---|---|---|---|---|---|
| | | PSNR | SSIM | PSNR | SSIM | PSNR | SSIM |
| EDSR [55] | X2 | 38.11 | 0.9602 | 33.92 | 0.9195 | 39.10 | 0.9773 |
| RCAN [51] | X2 | 38.27 | 0.9614 | 34.12 | 0.9216 | 39.44 | 0.9786 |
| SAN [56] | X2 | 38.31 | 0.9620 | 34.07 | 0.9213 | 39.32 | 0.9792 |
| SwinIR [57] | X2 | 38.42 | 0.9623 | 34.46 | 0.9250 | 39.93 | 0.9800 |
| T-transformer net(ours) | X2 | **38.73** | **0.9637** | **35.13** | **0.9282** | **40.71** | **0.9819** |
| EDSR | X3 | 34.65 | 0.9280 | 30.52 | 0.8462 | 34.17 | 0.9476 |
| RCAN | X3 | 34.74 | 0.9299 | 30.65 | 0.8482 | 34.44 | 0.9499 |
| SAN | X3 | 34.75 | 0.9300 | 30.59 | 0.8476 | 34.30 | 0.9494 |
| SwinIR | X3 | 34.97 | 0.9318 | 30.93 | 0.8534 | 35.12 | 0.9537 |
| T-transformer net(ours) | X3 | **35.16** | **0.9335** | **31.33** | **0.8576** | **35.84** | **0.9567** |
| EDSR | X4 | 32.46 | 0.8968 | 28.80 | 0.7876 | 31.02 | 0.9148 |
| RCAN | X4 | 32.63 | 0.9002 | 28.87 | 0.7889 | 31.22 | 0.9173 |
| SAN | X4 | 32.64 | 0.9003 | 28.92 | 0.7888 | 31.18 | 0.9169 |
| SwinIR | X4 | 32.92 | 0.9044 | 29.09 | 0.7950 | 32.03 | 0.9260 |
| T-transformer net(ours) | X4 | **33.18** | **0.9073** | **29.38** | **0.8001** | **32.87** | **0.9319** |

**TABLE 6.** Computational and parameter comparison (2× Set5).

| Method | parameter | PSNR | FLOPs(G) |
|---|---|---|---|
| EDSR | 43M | 38.11 | 9387 |
| RCAN | 15.7M | 38.27 | 15445 |
| SAN | 16M | 38.31 | 15861 |
| SwinIR | 11.9M | 38.42 | 11752 |
| T-transformer net(ours) | 20.8M | 38.73 | 17584 |

Once the network's efficacy was established, the focus shifted to the primary objective of this study: three-dimensional reconstruction of wheat seeds. While the image data of wheat seeds involve a singular object, namely, the wheat seed itself, the high similarity in seed texture, coupled with potential focus-related blurring during image capture, could lead to partial image blurriness. Should the network fail to effectively learn the global information of the seeds, the reconstructed high-resolution images might suffer from inadequate texture details, thereby jeopardizing the success of the three-dimensional reconstruction. Thus, employing the model enhanced with the channel attention mechanism for the super-resolution reconstruction of original wheat seed images becomes inherently essential.

Initially, the image sets utilized for training and validation were determined. Seventy-two original images of the same wheat seed were downsampled by a factor of four, reducing their resolutions to 360 × 640. Simultaneously, to control parameter count during training, the original and downsampled images were segmented, dividing each image into 8 × 8 sub-images. Following segmentation, the pixel dimensions of training images and original images were 45 × 80 and 180 × 320, respectively. Ultimately, a dataset comprising 4608 training and validation images was compiled.

Subsequently, training and testing were conducted according to the parameters outlined in Table 3. The conclusive experimental outcomes are summarized in Table 7.

**TABLE 7.** The average PSNR/SSIM score of different models in wheat grain test set.

| Method | Scale | PSNR | SSIM |
|---|---|---|---|
| EDSR | X4 | 36.64 | 0.9075 |
| RCAN | X4 | 36.70 | 0.9077 |
| SAN | X4 | 36.72 | 0.9078 |
| SwinIR | X4 | 37.10 | 0.9232 |
| T-transformer net(ours) | X4 | 42.89 | 0.96431 |

It is evident that our model has achieved remarkable results in the task of super-resolution reconstruction of wheat grains. The trained model demonstrates the capability to efficiently learn and reconstruct the textural features of wheat grains from a global perspective. As depicted in Fig. 7, the outcomes of various models applied to the same image are showcased. These models exhibit variations in their ability to restore the texture of wheat grains. EDSR models typically divide the image into fixed-size blocks and operate on each block, leading to blocking artifacts after reconstruction. Although RCAN does not directly operate on pixel blocks, the model selectively focuses on different parts of the image based on the attention mechanism. However, it struggles to capture global information effectively, resulting in suboptimal reconstruction quality. SwinIR succeeds in producing relatively sharp images, but it may alter the original texture, potentially causing adverse effects in subsequent processing stages.

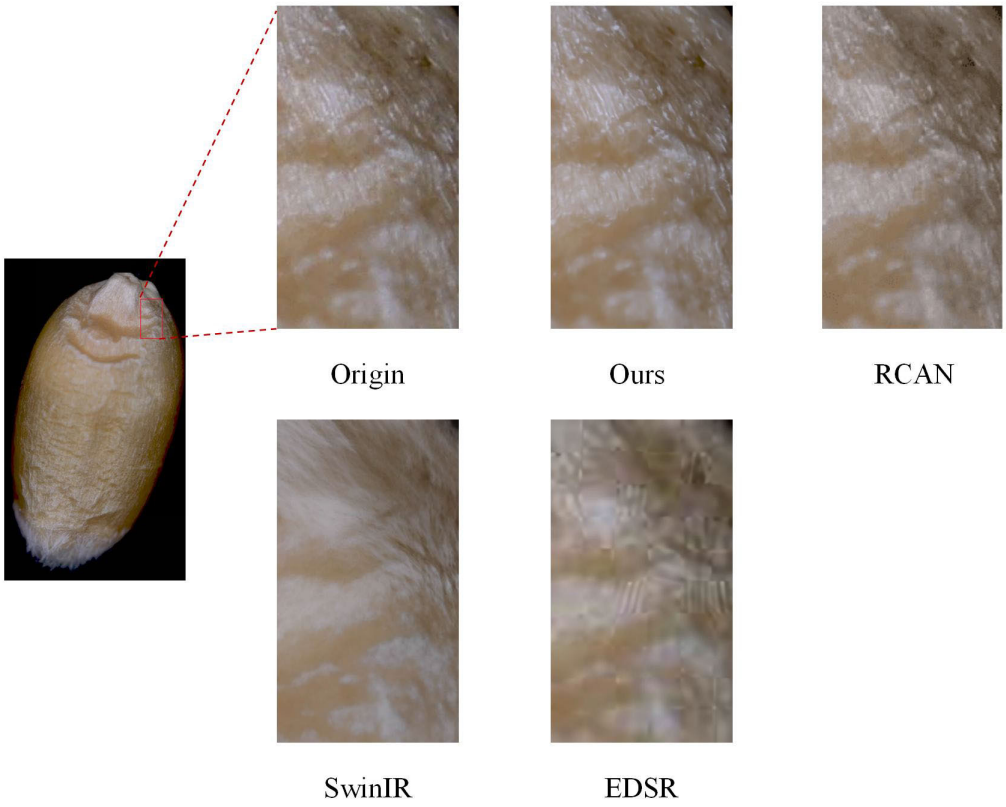



**FIGURE 7.** Super-resolution reconstruction results of wheat grain.

Thanks to the channel attention mechanism, our network model excels in restoring the intrinsic texture structure of wheat grains.

Subsequently, we performed an 8 × 8 segmentation on the original image, followed by super-resolution processing while maintaining a consistent 4-fold magnification factor. The segmented and processed images were then reassembled to reconstruct the complete image, now possessing a resolution four times higher than that of the original. This enhanced image was employed as the input for the subsequent stages of our three-dimensional reconstruction process.

**TABLE 8.** Quantitative result of the point cloud on the intermediate set of Tanks and Temples.

| Method | F-Score |
|---|---|
| COLMAP [63] | 42.14 |
| MVSNet [29] | 43.48 |
| Point-MVSNet [64] | 48.27 |
| CVP-MVENet [65] | 54.03 |
| UCSNet [66] | 54.83 |
| CasMVSNet [67] | 56.84 |
| ACMM [68] | 57.27 |
| Vis-MVSNet [30] | 60.03 |

### *D. THREE-DIMENSIONAL RECONSTRUCTION*

First, we assess the current mainstream 3D reconstruction methods on the intermediate set of the Tanks and Temples dataset [60]. They are trained according to the standard training methods provided by each model, and then images with a resolution of 1920×1080 are used as input for reconstruction on the Tanks and Temples dataset. In the benchmark test, we employ the F-score [61] as the evaluation index. As shown in Table 8, Vis-MVSNet [30] ranks first among all methods, outperforming both classic MVS methods [62] and other deep learning-based methods. Therefore, in the three-dimensional reconstruction task, this study adopts Vis-MVSNet as the reconstruction network.

The reconstruction process framework is shown in Fig. 8. Structure from Motion (SfM) [25] techniques take a set of images as input and produce two pieces of information: the camera parameters of each image and a set of three-dimensional points visible in the image, which are usually encoded as trajectories. A trajectory is defined as a list of the 3D coordinates of the reconstructed 3D points and the corresponding 2D coordinates in a subset of the input image. First, the camera pose and parameters are estimated using SfM techniques. This is followed by a dense reconstruction phase to generate dense point clouds. Finally, post-processing steps such as filtering, mesh generation, and smoothing are executed.

For model training, we utilized the publicly available DTU dataset [69]. This dataset comprises 128 scenes captured within controlled laboratory environments using a structured

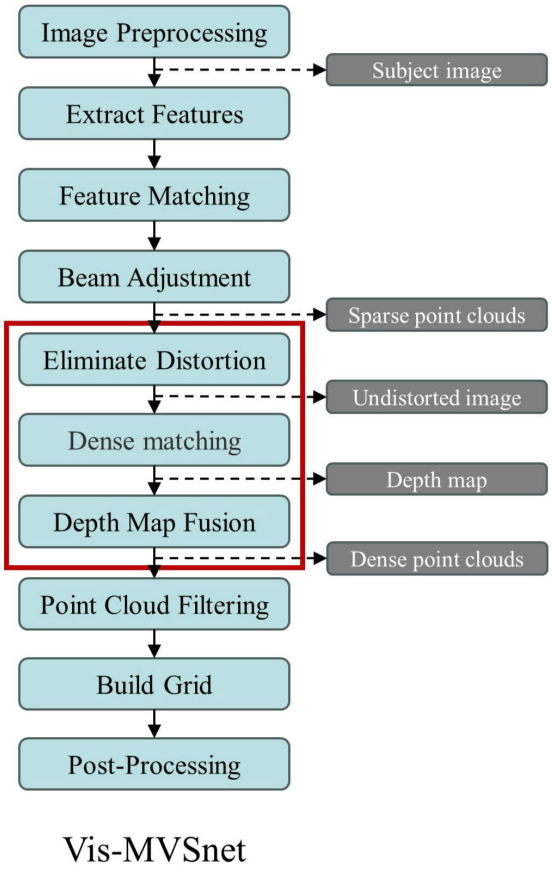



**FIGURE 8.** Frame diagram of Vis-MVSnet reconstruction process.

light scanner. Given the dataset's diversity encompassing various objects and materials, it serves as a suitable benchmark for training and evaluating deep learning-based Multi-View Stereo methods under real-world conditions. Although pre-trained models are made available by open-source methods, their applicability to our research dataset is limited due to differing task requirements. To address this, we employed transfer learning based on a pre-trained model. This approach facilitated fine-tuning of the model's weights specifically for our dataset, aiming to achieve enhanced performance.

After training the three-dimensional reconstruction network, we conducted three-dimensional reconstructions on both the raw images that had not undergone super-resolution processing and the images that had undergone super-resolution reconstruction. Ten wheat grains were randomly selected from four different varieties, and four distinct viewpoints were chosen for each grain. Subsequently, the reconstruction results of these viewpoints are compared with their corresponding original images, as illustrated in Fig. 9. The four viewpoints of the front, back, and sides are used as the reference viewpoints. Due to limitations in the image acquisition process, this analysis cannot be further completed. Since the top and bottom views of the wheat grain were not obtained, the top and bottom views cannot be used as indicators to measure the reconstruction results. The SSIM is used as a quantitative metric to evaluate the comparison between the reconstruction results, ensuring a direct comparison between the perspective of the model and that of the image. The average of these results is listed in Table 9.

**TABLE 9.** Effect of super-resolution reconstructed image and original image on SSIM score of 3D reconstruction.

| Image type | SSIM |
|---|---|
| Original image | 0.4514 |
| Super-resolution reconstructed image | 0.6273 |

Compared to the outcomes based solely on the original images, the results achieved with super-resolution reconstruction demonstrated an improvement in SSIM from 0.4514 to 0.6273, marking a significant enhancement of 38.96%. These outcomes suggest that employing super-resolution techniques in the task of three-dimensional reconstruction can notably increase the SSIM of the results, thus improving the overall quality of the final three-dimensional model.

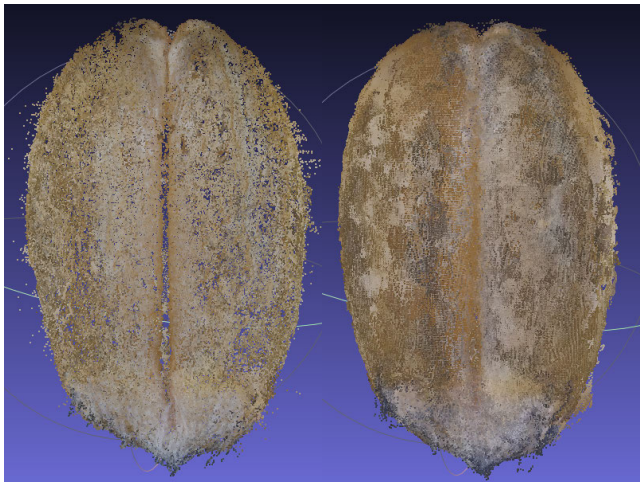
As illustrated in Fig. 10, the model obtained after super-resolution reconstruction presents a more detailed three-dimensional structure than the model created using only the original images. This improvement is due to addressing the focus issues encountered during the image acquisition of the original dataset, which resulted in blurred textures. These blurred textures impede accurate feature matching during the reconstruction process, leading to structural incompleteness.

Through the utilization of super-resolution networks for reconstruction, the network learns the process of grain texture restoration, transitioning from low-resolution to high-resolution image reconstruction. This acquired capability allows the post-super-resolution reconstructed images to effectively mitigate issues such as texture sparsity, repetitive patterns, and faint textural details within the images. The outcomes of the super-resolution reconstruction authentically capture the three-dimensional characteristics of wheat kernels.

Through the collaborative synergy of the super-resolution neural network and the three-dimensional reconstruction neural network, we are convinced that the network demonstrates proficiency in learning the textural and spatial attributes of wheat kernels. Subsequently, this acquired knowledge is then applied to our reconstruction task. Experimental data substantiate that employing images undergoing super-resolution reconstruction markedly improves the restoration of textural information pertaining to wheat kernels, which in turn facilitates the matching and reconstruction processes conducted by the three-dimensional reconstruction network.

In the process of sparse reconstruction and dense reconstruction, it has been observed that even with preprocessed images, the generated point clouds still contain a significant amount of noise related to the background. This complication renders direct filtering of the dense point cloud challenging, as the filtering procedure might inadvertently eliminate not


**FIGURE 9.** Comparison of original image and 3D reconstructed model from the same perspective.



**(a)** based on original image **(b)** based on super-resolution reconstructed image

**FIGURE 10.** Comparison of 3D reconstruction results based on original images and super-resolution reconstructed images.

only noise and errors but also essential three-dimensional points within sparse regions, thereby impeding subsequent mesh generation. Given this scenario, a novel approach of mesh generation followed by smoothing has been devised.

As depicted in Fig. 11(a), the presence of noisy points within the point cloud can lead to a considerable impact on the surface of the mesh model constructed based on the point cloud. To enhance the model's coherence, post-processing is imperative, with a key step involving the application of a smoothing procedure. By leveraging the mesh construction information and normal vectors, the model undergoes a smoothing process, as delineated in Fig. 11(b) and Fig. 11(c). Through multiple iterative smoothing iterations, remarkable enhancements in model quality are achieved.

After subjecting the grid models to a smoothing process, we employed Meshlab to efficiently convert them into point cloud models in bulk. This transformation facilitated the acquisition of uniformly distributed and pore-free point cloud models, rendering them suitable for subsequent research purposes. As depicted in Fig. 12 and Fig. 13, a representation of the diverse varieties of wheat grains, processed using the methodology outlined in this study, is displayed. Despite the minor differences in surface texture patterns and dimensions among the various wheat grain varieties, our reconstruction approach facilitated the successful generation of clear and structurally sound three-dimensional models for all strains. These outcomes underscore the robust generalizability of our

(a) original model (b) after single smoothing (c) after multiple smoothing

**FIGURE 11.** Model smoothing comparison.

(a) Huaimai 22 (b) Wanke 189 (c) Zimai 19 (d) Yangmai 15

**FIGURE 12.** The front results of 3D reconstruction of wheat grains of different varieties.

proposed refinement model, allowing for its broad applicability in the procurement of three-dimensional models across different wheat varieties. Simultaneously, this establishes a solid foundation for extending the model's utility in practical real-world scenarios.

Through this study, we have demonstrated the successful implementation of three-dimensional reconstruction using our method on wheat grains of different varieties. The obtained 3D point cloud model is more accurate than those models processed without the T-transformer net, which verifies the effectiveness and versatility of our method and provides important insights for future applications in the agricultural field.

In summary, this study achieved a more accurate and high-throughput three-dimensional reconstruction of wheat grains by combining super-resolution processing with neural network methods. This lays the foundation for subsequent wheat phenotypic measurements and the exploration of the relationship between the three-dimensional structure, physicochemical properties, and grain quality.

With our methodology, we not only attain high-quality three-dimensional models of wheat grains but also gain a

(a) Huaimai 22 (b) Wanke 189 (c) Zimai 19 (d) Yangmai 15

**FIGURE 13.** The back results of 3D reconstruction of wheat grains of different varieties.

deeper understanding of their characteristics in terms of texture, structure, and more. This offers a novel perspective and toolset for further studying aspects like wheat growth, development, and quality. Furthermore, the high-throughput nature of our approach enables swift processing of numerous wheat grain samples, providing robust technical support for large-scale experiments.

## IV. CONCLUSION

This study introduces a wheat grain reconstruction method based on the T-transformer net. The method utilizes SwinIR as the backbone network, combined with a channel attention mechanism, thereby achieving a significant improvement in accuracy. Through evaluation using standard datasets, both the PSNR and the SSIM of the images have improved by over one percentage point. When evaluated on a custom-built dataset, the PSNR and SSIM of the reconstructed images reached 42.89 and 0.96431, respectively. This represents an improvement of 15.60% and 4.45%, respectively, compared to the original network, achieving the highest scores compared to existing networks. Thanks to the channel attention mechanism, our network model effectively restores the original texture structure of wheat grains, outperforming previous approaches.

Subsequently, we trained the Vis-MVSnet network and conducted three-dimensional reconstruction on the super-resolved images, achieving satisfactory reconstruction accuracy. This method successfully addresses the challenges encountered when capturing two-dimensional images of wheat grains, such as the difficulty in obtaining high-resolution, clear, and accurately textured RGB images. Additionally, due to the symmetrical geometric characteristics of wheat grains, multiple groups of pixel blocks are highly similar, making it difficult to differentiate them and leading to stereo matching errors. Ultimately, we successfully reconstructed a batch of high-quality three-dimensional models of wheat grains, laying a solid foundation for subsequent work in phenomics, plant breeding, and research on the relationship between the three-dimensional structure and the physicochemical properties of grains. This effort bridges a gap in the field of real three-dimensional models of wheat grains.

However, this research method still has some limitations. Three-dimensional models obtained through passive methods lack real-scale information, which may limit the model's accuracy in certain applications. Additionally, highly accurate models often come with a large number of parameters, leading to slower reconstruction speeds. Therefore, further research is needed in the future to address these limitations and continuously optimize and develop this method. This could include exploring more efficient methods to obtain scale information, as well as using techniques such as model pruning and acceleration to improve reconstruction speed. The selection of wheat grain varieties did not consider the specific differences between different varieties. We only used the 2023 Chinese varieties of wheat grains available for this experiment as samples to create a dataset and a test set. Whether specific varieties of wheat grains exhibit the same high quality effects still requires further experimentation. These efforts will further drive the development of wheat grain three-dimensional reconstruction methods, bringing greater impact to the fields of agriculture and plant science.

## REFERENCES

[1] M. Tester and P. Langridge, "Breeding technologies to increase crop production in a changing world," *Science*, vol. 327, no. 5967, pp. 818–822, Feb. 2010.

[2] J.-M. Ribaut, M. de Vicente, and X. Delannay, "Molecular breeding in developing countries: Challenges and perspectives," *Current Opinion Plant Biol.*, vol. 13, no. 2, pp. 213–218, Apr. 2010.

[3] L. Huang, S. Shao, X. Lu, X. Guo, and J. Fan, "Segmentation and registration of lettuce multispectral image based on convolutional neural network," *Trans. Chin. Soc. Agricult. Machinery*, vol. 52, no. 9, pp. 186–194, 2021.

[4] J. Zhou, F. Tardieu, T. Pridmore, J. Doonan, D. Reynolds, N. Hall, S. Griffiths, T. Cheng, Y. Zhu, D. Jiang, and Y. Ding, "Plant phenomics: History, present status and challenges," *J. Nanjing Agricult. Univ.*, vol. 41, no. 4, pp. 580–588, 2018.

[5] S. Paulus, "Measuring crops in 3D: Using geometry for plant phenotyping," *Plant Methods*, vol. 15, no. 1, p. 103, 2019.

[6] S. Das Choudhury, S. Bashyam, Y. Qiu, A. Samal, and T. Awada, "Holistic and component plant phenotyping using temporal image sequence," *Plant Methods*, vol. 14, no. 1, p. 35, 2018.

[7] C. Lao, H. Yang, P. Li, and Y. Feng, "3D reconstruction of maize plants based on consumer depth camera," *Trans. Chin. Soc. Agricult. Machinery*, vol. 50, no. 7, pp. 222–228, 2019.

[8] K. Kraus and N. Pfeifer, "Determination of terrain models in wooded areas with airborne laser scanner data," *ISPRS J. Photogramm. Remote Sens.*, vol. 53, no. 4, pp. 193–203, Aug. 1998.

[9] W. Göbel, B. M. Kampa, and F. Helmchen, "Imaging cellular network dynamics in three dimensions using fast 3D laser scanning," *Nature Methods*, vol. 4, no. 1, pp. 73–79, Jan. 2007.

[10] C. Rocchini, P. Cignoni, C. Montani, P. Pingi, and R. Scopigno, "A low cost 3D scanner based on structured light," *Comput. Graph. Forum*, vol. 20, no. 3, pp. 299–308, Sep. 2001.

[11] N. Al-Najdawi, H. E. Bez, J. Singhai, and E. A. Edirisinghe, "A survey of cast shadow detection algorithms," *Pattern Recognit. Lett.*, vol. 33, no. 6, pp. 752–764, Apr. 2012.

[12] J. Park, H. Kim, Y.-W. Tai, M. S. Brown, and I. Kweon, "High quality depth map upsampling for 3D-TOF cameras," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 1623–1630.

[13] B. Schwarz, "Mapping the world in 3D," *Nature Photon.*, vol. 4, no. 7, pp. 429–430, Jul. 2010.

[14] K. Khoshelham and S. O. Elberink, "Accuracy and resolution of Kinect depth data for indoor mapping applications," *Sensors*, vol. 12, no. 2, pp. 1437–1454, Feb. 2012.

[15] Z. Taixiong, H. Shuai, L. Yongfu, and F. Ming-Chi, "Key techniques for vision based 3D reconstruction: A review," *Acta Automatica Sinica*, vol. 46, no. 4, pp. 631–652, 2020.

[16] Z. Chenxi, W. Weiliang, L. Xianju, G. Xinyu, and Z. Chunjiang, "Phenotypic traits extraction of wheat plants using 3D digitization," *Smart Agricult.*, vol. 4, no. 2, p. 150, 2022.

[17] Z. Zhang, S. Zhao, X. Luo, and F. Wei, "Matching method of green crops based on SURF feature extraction," *Trans. Chin. Soc. Agricult. Eng.*, vol. 31, no. 14, pp. 172–178, 2015.

[18] A. Paproki, J. Fripp, O. Salvado, X. Sirault, S. Berry, and R. Furbank, "Automated 3D segmentation and analysis of cotton plants," in *Proc. Int. Conf. Digit. Image Comput., Techn. Appl.*, Dec. 2011, pp. 555–560.

[19] Y. Li, M. Yao, L. Li, Q. Ding, and R. He, "Counting method of grain number based on wheatear spikelet image segmentation," *J. Nanjing Agricult. Univ.*, vol. 41, no. 4, pp. 742–751, 2018.

[20] N. Wang, B. Kong, C. Wang, W. Li, and H. Xu, "Counting grains per wheat spike base in fractal segmentation of image," *Comput. Syst. Appl.*, vol. 10, pp. 219–224, Jan. 2017.

[21] T. Liu, C. Sun, L. Wang, X. Zhong, X. Zhu, and W. Guo, "In-field wheatear counting based on image processing technology," *Trans. Chin. Soc. Agricult. Machinery*, vol. 45, no. 2, pp. 282–290, 2014.

[22] P. Moulon, P. Monasse, and R. Marlet, "Adaptive structure from motion with a contrario model estimation," in *Computer Vision—ACCV 2012: 11th Asian Conference on Computer Vision, Daejeon, Korea, November 5–9, 2012, Revised Selected Papers, Part IV 11*. Berlin, Germany: Springer, 2013, pp. 257–270.

[23] S. Y. Bao, M. Bagra, Y.-W. Chao, and S. Savarese, "Semantic structure from motion with points, regions, and objects," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2703–2710.

[24] C. Wu, "Towards linear-time incremental structure from motion," in *Proc. Int. Conf. 3D Vis. (3DV)*, Jun. 2013, pp. 127–134.

[25] J. L. Schönberger and J.-M. Frahm, "Structure-from-motion revisited," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4104–4113.

[26] Q. Zhu, C. Min, Z. Wei, Y. Chen, and G. Wang, "Deep learning for multi-view stereo via plane sweep: A survey," 2021, *arXiv:2106.15328*.

[27] X. Wang, C. Wang, B. Liu, X. Zhou, L. Zhang, J. Zheng, and X. Bai, "Multi-view stereo in the deep learning era: A comprehensive review," *Displays*, vol. 70, Dec. 2021, Art. no. 102102.

[28] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27, 2014, pp. 1–9.

[29] Y. Yao, Z. Luo, S. Li, T. Fang, and L. Quan, "MVSNet: Depth inference for unstructured multi-view stereo," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 767–783.

[30] J. Zhang, S. Li, Z. Luo, T. Fang, and Y. Yao, "Vis-MVSNet: Visibility-aware multi-view stereo network," *Int. J. Comput. Vis.*, vol. 131, no. 1, pp. 199–214, Jan. 2023.

[31] G. Fahim, K. Amin, and S. Zarif, "Single-view 3D reconstruction: A survey of deep learning methods," *Comput. Graph.*, vol. 94, pp. 164–190, Feb. 2021.

[32] M. Yu, J. Shi, C. Xue, X. Hao, and G. Yan, "A review of single image super-resolution reconstruction based on deep learning," *Multimedia Tools Appl.*, pp. 1–42, 2023.

[33] J. Li, W. Gao, and Y. Wu, "High-quality 3D reconstruction with depth super-resolution and completion," *IEEE Access*, vol. 7, pp. 19370–19381, 2019.

[34] R. Fernandez-Beltran, P. Latorre-Carmona, and F. Pla, "Single-frame super-resolution in remote sensing: A practical overview," *Int. J. Remote Sens.*, vol. 38, no. 1, pp. 314–354, Jan. 2017.

[35] K. Jiang, Z. Wang, P. Yi, G. Wang, K. Gu, and J. Jiang, "ATMFN: Adaptive-threshold-based multi-model fusion network for compressed face hallucination," *IEEE Trans. Multimedia*, vol. 22, no. 10, pp. 2734–2747, Oct. 2020.

[36] J. A. Parker, R. V. Kenyon, and D. E. Troxel, "Comparison of interpolating methods for image resampling," *IEEE Trans. Med. Imag.*, vol. MI-2, no. 1, pp. 31–39, Mar. 1983.

[37] M. Jiang and G. Wang, "Development of iterative algorithms for image reconstruction," *J. X-Ray Sci. Technol.*, vol. 10, nos. 1–2, pp. 77–86, 2002.

[38] Z. Wang, D. Liu, J. Yang, W. Han, and T. Huang, "Deep networks for image super-resolution with sparse prior," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 370–378.

[39] C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," in *Proc. 13th Eur. Conf. Comput. Vis. (ECCV)*. Zurich, Switzerland: Springer, 2014, pp. 184–199.

[40] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017.

[41] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4681–4690.

[42] K. Jiang, Z. Wang, C. Chen, Z. Wang, L. Cui, and C.-W. Lin, "Magic ELF: Image deraining meets association learning and transformer," 2022, *arXiv:2207.10455*.

[43] K. Jiang, X. Jia, W. Huang, W. Wang, Z. Wang, and J. Jiang, "Dynamic association learning of self-attention and convolution in image restoration," 2023, *arXiv:2311.05147*.

[44] Y. Xiao et al., "Local-global temporal difference learning for satellite video super-resolution," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 4, pp. 2789–2802, Apr. 2024, doi: 10.1109/TCSVT.2023.3312321.

[45] K. Jiang, Z. Wang, P. Yi, and J. Jiang, "Hierarchical dense recursive network for image super-resolution," *Pattern Recognit.*, vol. 107, Nov. 2020, Art. no. 107475.

[46] Z. Chi, "Research on grain moisture and temperature detection technology integrating microwave and sound wave," Ph.D. dissertation, Beijing Univ. Posts Telecommun., Beijing, China, 2020.

[47] T. Xiao, M. Singh, E. Mintun, T. Darrell, P. Dollár, and R. Girshick, "Early convolutions help transformers see better," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 30392–30400.

[48] K. Yuan, S. Guo, Z. Liu, A. Zhou, F. Yu, and W. Wu, "Incorporating convolution designs into visual transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 579–588.

[49] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.

[50] D. Hendrycks and K. Gimpel, "Gaussian error linear units (GELUs)," 2016, *arXiv:1606.08415*.

[51] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 286–301.

[52] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1874–1883.
[53] E. Agustsson and R. Timofte, "NTIRE 2017 challenge on single image super-resolution: Dataset and study," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jul. 2017, pp. 126–135.
[54] Y. Matsui, K. Ito, Y. Aramaki, A. Fujimoto, T. Ogawa, T. Yamasaki, and K. Aizawa, "Sketch-based Manga retrieval using manga109 dataset," *Multimedia Tools Appl.*, vol. 76, no. 20, pp. 21811–21838, Oct. 2017.
[55] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, "Enhanced deep residual networks for single image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 136–144.
[56] T. Dai, J. Cai, Y. Zhang, S.-T. Xia, and L. Zhang, "Second-order attention network for single image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 11065–11074.
[57] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, "SwinIR: Image restoration using Swin transformer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 1833–1844.
[58] M. Bevilacqua, A. Roumy, C. Guillemot, and M. L. Alberi-Morel, "Low-complexity single-image super-resolution based on nonnegative neighbor embedding," Tech. Rep., 2012.
[59] R. Zeyde, M. Elad, and M. Protter, "On single image scale-up using sparse-representations," in *Curves and Surfaces: 7th International Conference, Avignon, France, June 24–30, 2010, Revised Selected Papers 7*. Berlin, Germany: Springer, 2012, pp. 711–730.
[60] A. Knapitsch, J. Park, Q.-Y. Zhou, and V. Koltun, "Tanks and temples: Benchmarking large-scale scene reconstruction," *ACM Trans. Graph.*, vol. 36, no. 4, pp. 1–13, 2017.
[61] C. Goutte and E. Gaussier, "A probabilistic interpretation of precision, recall and F-score, with implication for evaluation," in *Proc. Eur. Conf. Inf. Retr.* Berlin, Germany: Springer, Mar. 2005, pp. 345–359.
[62] Y. Furukawa and C. Hernández, "Multi-view stereo: A tutorial," *Found. Trends Comput. Graph. Vis.*, vol. 9, nos. 1–2, pp. 1–148, 2015.
[63] J. L. Schönberger, E. Zheng, J.-M. Frahm, and M. Pollefeys, "Pixelwise view selection for unstructured multi-view stereo," in *Proc. 14th Eur. Conf. Comput. Visiona (ECCV)* Amsterdam, The Netherlands: Springer, 2016, pp. 501–518.
[64] R. Chen, S. Han, J. Xu, and H. Su, "Point-based multi-view stereo network," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1538–1547.
[65] J. Yang, W. Mao, J. M. Alvarez, and M. Liu, "Cost volume pyramid based depth inference for multi-view stereo," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4877–4886.
[66] S. Cheng, Z. Xu, S. Zhu, Z. Li, L. E. Li, R. Ramamoorthi, and H. Su, "Deep stereo using adaptive thin volume representation with uncertainty awareness," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2524–2534.
[67] X. Gu, Z. Fan, S. Zhu, Z. Dai, F. Tan, and P. Tan, "Cascade cost volume for high-resolution multi-view stereo and stereo matching," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2495–2504.
[68] Q. Xu and W. Tao, "Multi-scale geometric consistency guided multi-view stereo," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5483–5492.
[69] H. Aanæs, R. R. Jensen, G. Vogiatzis, E. Tola, and A. B. Dahl, "Large-scale data for multiple-view stereopsis," *Int. J. Comput. Vis.*, vol. 120, no. 2, pp. 153–168, Nov. 2016.

**YIJUN TIAN** was born in 1998. He is currently pursuing the master's degree in engineering with Henan University of Technology. His research interests include 3D reconstruction, computer vision, and deep learning.

**JINNING ZHANG** was born in 1995. She received the master's degree in agricultural engineering and information technology from Shanxi Agricultural University. Her main research interests include drying process quality data analysis and modeling, and intelligent software development of drying control systems.

**ZHONGJIE ZHANG** received the Ph.D. degree in agricultural product processing and storage engineering from China Agricultural University. He is currently the Director of the Grain Storage and Transportation Research Institute of the Academy of National Food and Strategic Reserves Administration. He is mainly engaged in theoretical research, technology development and application engineering in the fields of grain and oil storage, ventilation and drying, and environmental engineering.

**JIANJUN WU** (Member, IEEE) was born in November 1976. He received the B.E. degree in computer and applications from the Seventh Department, Information Engineering College, People's Liberation Army, in 1999, the M.E. degree in computer applications from the Department of Computer Science, Zhengzhou University, in 2004, and the Ph.D. degree from the School of Computer Science, Wuhan University of Technology, in 2014. He is currently an Associate Professor with the Information Science and Engineering Department, Henan University of Technology.