ISPRS Journal of Photogrammetry and Remote Sensing 218 (2024) 50–68

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

journal homepage: www.elsevier.com/locate/isprsjprs

Effective variance attention-enhanced diffusion model for crop field aerial image super resolution

Xiangyu Lu [a], Jianlin Zhang [a], Rui Yang [a], Qina Yang [a], Mengyuan Chen [a], Hongxing Xu [b,*], Pinjun Wan [c], Jiawen Guo [b], Fei Liu [a,*]

[a] College of Biosystems Engineering and Food Science, Zhejiang University, Hangzhou 310058, China
[b] State Key Laboratory for Managing Biotic and Chemical Threats to the Quality and Safety of Agro-Products, Institute of Plant Protection and Microbiology, Zhejiang Academy of Agricultural Sciences, Hangzhou 310021, China
[c] State Key Laboratory of Rice Biology and Breeding, China National Rice Research Institute, Hangzhou 310006, China

ARTICLE INFO

Keywords:
Super-resolution
Diffusion model
Variance attention
Aerial imagery
Super-resolution relative fidelity index

ABSTRACT

Image super-resolution (SR) can significantly improve the resolution and quality of aerial imagery. Emerging diffusion models (DM) have shown superior image generation capabilities through multistep refinement. To explore their effectiveness on high-resolution cropland aerial imagery SR, we first built the CropSR dataset, which includes 321,992 samples for self-supervised SR training and two real-matched SR datasets from high-low altitude orthomosaics and fixed-point photography (CropSR-OR/FP) for testing. Inspired by the observed trend of decreasing image variance with higher flight altitude, we developed the Variance-Average-Spatial Attention (VASA). The VASA demonstrated effectiveness across various types of SR models, and we further developed the Efficient VASA-enhanced Diffusion Model (EVADM). To comprehensively and consistently evaluate the quality of SR models, we introduced the Super-resolution Relative Fidelity Index (SRFI), which considers both structural and perceptual similarity. On the × 2 and × 4 real SR datasets, EVADM reduced Fréchet-Inception-Distance (FID) by 14.6 and 8.0, respectively, along with SRFI gains of 27 % and 6 % compared to the baselines. The superior generalization ability of EVADM was further validated using the open Agriculture-Vision dataset. Extensive downstream case studies have demonstrated the high practicality of our SR method, indicating a promising avenue for realistic aerial imagery enhancement and effective downstream applications. The code and dataset for testing are available at https://github.com/HobbitArmy/EVADM.

1. Introduction

Low-altitude remote sensing offers high-resolution imagery that greatly benefits a wide range of fields, including vegetation monitoring (Sagan et al., 2019), road monitoring (Inzerillo et al., 2022), and forestry investigation (Kong et al., 2023). High flight altitudes are often required for large-area aerial photography, resulting in a coarse Ground Sampling Distance (GSD) and compromised image quality lacking fine-scale details, as illustrated in Fig. 1. Recent studies have shown that the flight height of Unmanned Aerial Vehicle (UAV) has a significant impact on the measurement of structural features of vegetation (Mao et al., 2023). For optical remote sensing, the trade-off between photography efficiency and image quality has constrained the ability to conduct precise observation over large areas (Mao et al., 2022). Table A.1 presents the statistical efficiency and imagery GSD across more altitudes. Another challenge is that downstream task models built under specific conditions struggle to adapt well to other scale domains. This issue is mainly due to altitude differences and sensor variations, resulting in GSD and feature shifting (Hu et al., 2019). Downgrading high-resolution data through down-sampling is straightforward. In contrast, the reverse is not true. Conventional up-sampling methods enlarge the image size but do not increase effective detailed features.

With the development of deep-learning (DL) techniques and computing power, super-resolution (SR) methods can reconstruct fine details during up-sampling. In remote sensing, SR is promising to transform high-altitude, low-quality images into high-resolution counterparts, making it a popular research area. As an upstream image optimization method, SR can be achieved from multiple images (MISR,

* Corresponding authors.
E-mail addresses: luxyzju@zju.edu.cn (X. Lu), jianlin.zhang@zju.edu.cn (J. Zhang), ryang@zju.edu.cn (R. Yang), qnyang@zju.edu.cn (Q. Yang), mychen_1998@zju.edu.cn (M. Chen), xuhongxing@zaas.ac.cn (H. Xu), wanpinjun@caas.cn (P. Wan), guojw@zaas.ac.cn (J. Guo), fliu@zju.edu.cn (F. Liu).

https://doi.org/10.1016/j.isprsjprs.2024.08.017
Received 22 December 2023; Received in revised form 28 August 2024; Accepted 29 August 2024

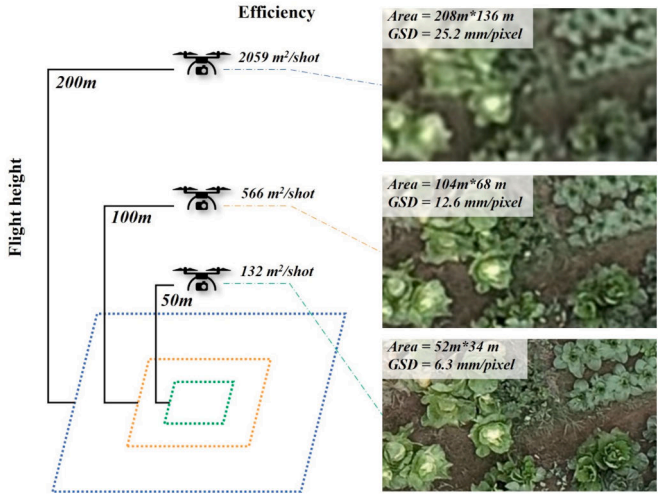

Fig. 1. Trade-off between aerial photography efficiency (flight altitude) and image quality (ground sampling distance, GSD). High-altitude sensing provides a large field of view (FOV) but with coarse detail.

from multi-sensor or multi-temporal) or single image (SISR) (Razzak et al., 2023). In (He et al., 2023), a two-stage self-supervised spectral image pan-sharpening method was proposed to fuse the high-resolution panchromatic (PAN) and low-resolution multispectral (MS) images for high-resolution spectral images. Arefin et al. (2020) reconstructed hidden high-resolution details using multi-temporal low-resolution observations of the same scene, and the method was further employed by Razzak et al. (2023) to enhance images from multiple observations. For SISR of remote sensing, the real SR image pairs were constructed from multi-sensors co-mounted on one carrier (Aslahishahri et al., 2021), multiple platforms (Wang et al., 2021a), or simulated from degradation of high-resolution image (Qiu et al., 2023). Zhang et al. (2022) utilized the SR technique to improve drone imagery resolution, which effectively enhanced the cabbage plant segmentation performance. A generic network was introduced by Feng et al. (2022) to conduct multiple tasks, including image colorization, image SR, and simultaneous SR colorization, demonstrating the effect of multitask integration. For small and low-resolution objects in remote sensing, SR can enhance the images and assist downstream object detection tasks (Courtrai et al., 2020; Zhou et al., 2019). Besides, the SR is widely utilized and significantly improves the performance of various downstream tasks, including individual building extraction (Chen et al., 2023), crop segmentation (Yun et al., 2023), structure-from-motion point cloud densification (Pashaei et al., 2020) and land cover classification (Li et al., 2021).

The development of DL-based SR can be divided into three crucial stages: it originated from network regression, evolved with the Generative-Adversarial-Network (GAN), and is now thriving from the diffusion-based models. Most of the early works on SR were regression-based and trained with (1-norm distance) L1 loss or mean squared error (MSE) L2 loss. The first SISR network, SRCNN, built by Dong et al. (2016), showed the effectiveness of DL in SR. The subsequent introduction of techniques including PixelShuffle (Shi et al., 2016), the elimination of batch normalization (Lim et al., 2017), the closed-loop regression scheme (Guo et al., 2020), and more comprehensive loss objectives (Sajjadi et al., 2017; Zhong et al., 2023), laid the foundation for DL-based SR. Despite effectively estimating the posterior mean (Saharia et al., 2022c), regression-based methods often produce blurry and unsatisfying images, lacking high-frequency details. To address the smooth and unreal SR issues, the generative-adversarial training scheme was proposed in SRGAN (Ledig et al., 2017), using a discriminator network to facilitate the training of the generator. ESRGAN was improved upon SRGAN by optimizing network blocks and employing a comprehensive loss function (Wang et al., 2018). Besides, as the real degradation schemes are unknown, complex and various degradations are vital for model generalizability for blind SISR self-supervised training. To cover or close to the real domain, researchers proposed to estimate blur kernels (Bell-Kligler et al., 2019; Cai et al., 2019) and real noise distributions (Ji et al., 2020), and constructed high-order degradation (Wang et al., 2021b; Zhang et al., 2021a) from repetitive blurring, down-sampling, noising, and compression processes. Although GAN-based models sometimes suffer from mode collapse (Thanh-Tung and Tran, 2020), recent research has developed techniques to mitigate this issue (Shi et al., 2022). Overall, the GAN-based method paves the way for SR and remains a strong baseline.

Diffusion model (DM) has recently been widely used in text-to-image generation (Saharia et al., 2022b), image inpainting (Wang et al., 2023), extrapolating (Saharia et al., 2022a), and many other image-generation applications. DM was inspired by non-equilibrium thermodynamics, including the forward (nosing) and the reverse (denoising) processes. In the forward process, noise is gradually added to the data, while in the reverse process, a learned noise prediction model estimates the noise and gradually removes it to recover the original features (Ho et al., 2020). As the diffusion-based SISR pioneer, SRDiff generated diverse and realistic SR results with rich details and was easier to train with a small model size (Li et al., 2022). The potential of DM for SR was further investigated through model cascading (Ho et al., 2022), enabling the generation of high-fidelity images even at a large $\times$ 8 upscaling factor in SR3 (Saharia et al., 2022c). By incorporating high-order degradation as Real-ESRGAN (Wang et al., 2021b), the improved SR3 + achieved realistic results on the real-scene SR dataset (Sahak et al., 2023). However, the most prominent shortcomings lie in the numerous refinement steps, where hundreds of iterations are required to generate satisfactory images (Salimans and Ho, 2021). To address the issue, Denoising Diffusion Implicit Models (DDIM) proposed to adopt a deterministic forward process (Song et al., 2021), and a smaller step sequence could be sampled during denoising inference for acceleration. In addition, the Latent Diffusion Model (LDM) further reduces the computation and steps by encoding the images into compressed representation using a pre-trained VQGAN (Esser et al., 2021). And the iterative refinement is conducted in the latent space for efficiency (Rombach et al., 2022). The DMs show great promise for photorealistic SR generation, yet there is little reported research exploring their magic on remote sensing imagery.

Besides, to evaluate the quality of remote sensing imagery SR reconstruction results, reference-based metrics, including peak signal-to-noise ratio (PSNR) and structural similarity (SSIM), are commonly used (Wang et al., 2004). Although they efficiently measure consistency, they over-penalize non-matching details and favor smooth images (Saharia et al., 2022c). Perceptual evaluation was later introduced by measuring the distance in the feature space of well-trained networks, which aligns well with human assessment. As a widely used metric in SR (Saharia et al., 2022c), the Fréchet Inception Distance (FID) utilizes the feature maps from the Inception V3 network as the feature representation (Heusel et al., 2017). For practical remote sensing SR data pairs, such as multi-temporal or cross-sensor ones, HR and LR images are not always perfectly aligned due to temporal variations, cloud contamination, or environmental changes (Razzak et al., 2023), causing the intrinsic disparities of matched SR pairs and difficulties across SR datasets.

In summary, high-resolution remote sensing imagery is valuable but costly. SR techniques provide an efficient solution, while there is limited research dedicated to high-resolution aerial images (Aslahishahri et al., 2021; Inzerillo et al., 2022). The development and optimization of DM have provided powerful new approaches for image SR. Besides, one single metric is insufficient for comprehensive model assessment (Wells et al., 2023) and can lead to inconsistency across datasets. Faced with these issues, the contributions of this study are as follows:

1). A super high-resolution (1 cm/pixel GSD) Cropland aerial image SR dataset (CropSR) consisting of 321,992 images was built, and two

real-matched SR datasets from multi-altitude orthomosaics and fixed-point photography (CropSR-OR/FP with 5,392 matched SR pairs).

2). By observing the cropland aerial image variance trends to flight height, the variance-average-spatial attention (VASA) was proposed, which significantly enhanced various SR models.

3). The Efficient VASA-enhanced Diffusion Model (EVADM) was constructed for agricultural aerial imagery SR, which effectively facilitated manifold downstream tasks.

4). A cross-scale Super-resolution Relative Fidelity Index (SRFI) considering structural similarity and perceptual distance was designed, demonstrating robust evaluation across multiple datasets.

The remaining part of the article is organized as follows. **Section 2** introduces the construction process of CropSR, CropSR-OR/FP datasets, and the open Agriculture-Vision dataset. **Section 3** illustrates the variance trends, the VASA-enhanced SR networks, the proposed EVADM model, and the SRFI index. All experimental results and the downstream case studies are reported in **Section 4**. Moreover, **Section 5** discusses the effectiveness of the proposed SR method, the relationship between variance and the SR process, and potential future research in detail.

2. Datasets

In this part, we first collected and built a high-resolution aerial image dataset, namely CropSR, with more than three hundred thousand samples. In **Section 2.2**, two real SR datasets, CropSR-OR and CropSR-FP, were constructed from matched orthomosaic mapping and fixed-point photographs, with more than 5,000 pairs in total for testing. The last section introduced the open dataset, Agriculture-Vision, which is used for generalization studies.

2.1. CropSR dataset

To experiment on the proposed SR methods, in this study, three kinds of datasets were built for training or testing: 1. CropSR, which is the main SR dataset with 321,992 high-resolution aerial cropland photographs for self-supervised degradation SR training; 2. CropSR-Ortho (OR), where the low–high resolution pairs are cut from high-low altitude geo-referenced orthomosaic and finely matched; 3. CropSR-Fixed-Point (FP), which is built from fixed-point photography at different heights. All training data is from the main CropSR at the experiments, and the CropSR-OR/FP are used for SR model testing only.

For the CropSR dataset, the images are from multiple equipment or sensors at different heights and various rural scenes, ensuring data generality and diversity. Table 1 lists the five sensors used to collect aerial images and their photogrammetry parameters. The sensors include P1, H20T, and L1 (only the wide-angle color camera used here) cameras mounted on the DJI Matrice 300 platform, DJI Mavic 3E/3M (carrying the same wide-angle cameras), and the DJI Mavic 2. Due to variations in film size, pixel size, lens types, and focal lengths, imaging quality and ground sampling distance (GSD) can vary significantly.

There are dozens of flight missions involved with multifarious ground cover and scenes, as detailed in Table A.2. Because most of the images were acquired under 25 to 50 m in height, the overall GSD of the CropSR is finer than 1 cm/pixel. The objects mainly consist of crops and vegetation, including rice at various growth stages, various types of vegetables, maize, oil-seed rape, wheat, monarda, shrubs, and trees. Additionally, a small portion of the samples includes bare soil, greenhouses, roads, ponds, and buildings, which appear distinct from the plants. As a high overlap ratio (70 %) is always set in flight missions, to reduce duplicate regions, we selected images from the original with a sparse sampling technique (Lu et al., 2023). Then, all images were segmented into 512x512 pixel patches using a sliding window approach with a step size of 512 pixels and a 10 % image margin ratio.

A total of 321,992 samples were generated, with 90 % randomly selected for training (289,792), 1 % for validation (3,220), and the remaining 9 % for testing (28,980). Due to the large number of test samples, to ensure efficient evaluation, a subset of 10 % from the original 9 % test samples (2,898) was randomly chosen as the Test_s set, which is labeled as CropSR-Test. These HR patches constitute the CropSR dataset, as the process depicted at the top of Fig. 2, and typical samples are displayed in the lower left.

2.2. Real super-resolution dataset

The construction process of CropSR-OR is depicted in the middle of Fig. 2, where the UAVs perform area mapping missions at two different heights to produce high-low altitude orthomosaics (OMs). In this instance, a region with diverse ground cover was scanned by P1 from altitudes of 25 and 50 m. Mission details can be found in Table A.3.

As real-time kinematic (RTK) was used during each flight, overlaps of the two OMs were roughly matched. Initially, we used a 5-meter square sliding window on both OMs to extract roughly matched image pairs. Then, low-resolution (LR) and high-resolution (HR) samples from each pair were finely aligned using SURF feature alignment. All LR samples were then resized to half the size of HR using bicubic interpolation to achieve a standardized $\times$ 2 SR ratio. Subsequently, multiple 512–256-sized image patches are cropped from each aligned pair using a sliding window with sizes of 512 and 256 pixels. To address color inconsistencies between the LR and HR images, the LR patch is adjusted in the HSV space according to the HR patch for calibration, which is crucial for addressing spectral disparity (Sagan et al., 2022). Ultimately, we assessed the quality of the finely matched patch pairs using reference-based metrics. Only pairs with down-sampled HR and LR achieving $SSIM > 0.7$ *and* $PSNR > 20$ were retained for further analysis.

The workflow of constructing CropSR-FP is similar, with data sourced from an M3E-w camera using fixed-point photography at 25 and 50 m, as detailed in Table A.3. The high- and low-altitude images with the same index capture distinct field of view (FOV). Therefore, each pair is first aligned using SURF, and the intersection areas are cropped accordingly. These coarsely matched overlapped pairs are further processed in the same manner as shown in the middle of Fig. 2.

Through this approach, the $\times$ 2-ratio CropSR-OR/FP datasets were constructed. Additionally, bicubic down-sampling of the LR images generated the $\times$ 4 and $\times$ 8-ratio datasets. The CropSR-OR/FP datasets are used exclusively for testing, and all pairs from these datasets were excluded from training.

2.3. Agriculture-vision dataset

To evaluate the generalization ability of SR models, we incorporate a

Table 1
Data acquisition equipment and photogrammetry parameters.

Camera	Film size (inch)	Film width (mm)	Image width (pixel)	Pixel size (um)	FL (mm)	35 mm EQ FL (mm)	FOV (°)	GSD-100 m (cm)
P1	Full	35.9	8192	4.4	35	35	63.5	1.26
M3-w	4/3	18	5280	3.4	12.29	24	84	2.77
H20T-w	1/2.3	6.4	4056	1.6	4.5	24	82.9	3.56
L1-c	1	12.7	5472	2.3	8.8	24	84	2.61
M2	1	12.7	5472	2.3	10.26	28	77	2.24

The FL denotes the focal length, and EQ FL means the equivalent focal length. FOV and GSD are field of view and ground sampling distance, respectively.

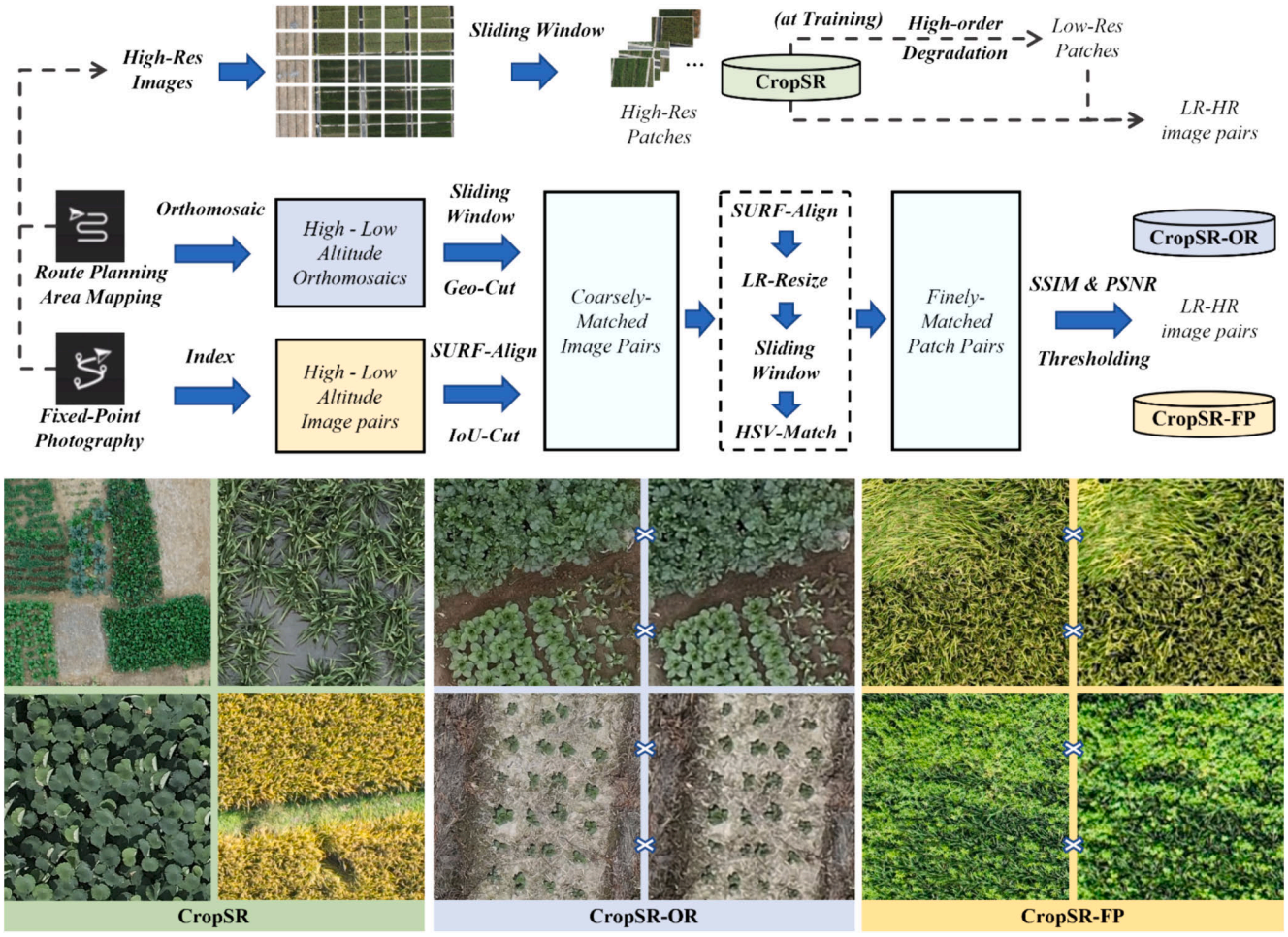

Fig. 2. Construction workflows and samples of the CropSR and CropSR-OR/FP super-resolution (SR) datasets. The CropSR is built through down-sampling and degradation for self-supervised training, while CropSR-OR/FP comprises pair-matched aerial images from different heights for real-world testing.

portion of the Agriculture-Vision (AgV) dataset in our experiments (Chiu et al., 2020). The AgV dataset was built for semantic segmentation of field anomaly patterns. The images, primarily from farmland, were captured by cameras on aerial vehicles with a GSD of approximately 10 cm per pixel. The AgV dataset contains a total of 94,986 images, with 19,708 samples allocated for testing.

Given that the GSD of AgV is ten times coarser than that of the CropSR dataset, the images are fed into the SR model trained on CropSR, and the super-resolved outputs are compared for generalization study. Due to the large size of the dataset, only 1/10 of the AgV test-set images are adopted. A total of 1,971 samples are randomly selected, each with an original size of 512 pixels. Details of data processing for the generalization test will be further discussed in **Section 4.4**.

3. Methods

This part first described the changes in aerial image variance and average statistics relative to flight height. Based on these variance and average trends, we constructed the VASA for three types of SR models: regression-based EDSR (Lim et al., 2017), GAN-based Real-ESRGAN (Wang et al., 2021b), and LDM (Rombach et al., 2022). We further improved the structure of the LDM with VASA for efficiency. Additionally, for SR quality evaluation, a cross-scale SRFI metric considering both structural similarity and perceptual distance was introduced **in Section 3.5**.

3.1. The variance trends and variance-average-spatial attention

For cropland, grassland, and many other wilderness areas, the homogenous groups observed from the air reveal a consistent and uniform morphology as the view zooms out. The spatial variance of imagery is promising in measuring this consistency. Thus, we calculated the variance and average statistics from 29 thousand randomly sampled aerial images (512 × 512 size) at different heights in rural areas. As shown in Fig. 3, as the image GSD increases with coarser spatial resolution, the variance tends to decrease because of blurry details, resulting in minor variations of each channel. Meanwhile, for one scenario, as altitude changes, the average values of each channel barely change. We reckon the statistics of image spatial variance may help SR models assess the scale of detail presented in one image, and the average may facilitate feature identification of different scenarios.

To validate the effectiveness of image spatial variance and average, we further developed variance-attention (Var-Attn), average-attention (Avg-Attn), and the combined Variance-Average Attention (VA-Attn), as displayed in Fig. 4**a**. As expressed in Equation (1), the spatial variance values of each channel, or channel-wise variance $v_X^{cw} \in \mathbb{R}^{1\times1\times C}$ are first calculated. The standardized variance v_{std}^{cw} is then computed by subtracting the mean of v_X^{cw} and dividing by the sqrt root of the variance of v_X^{cw}. The $v_{std}^{cw} \in \mathbb{R}^{1\times1\times C}$ is modified by a 2-layer fully connected (FC) learnable network with a residual connection. Finally, the Var-Attn is obtained through a sigmoid function:

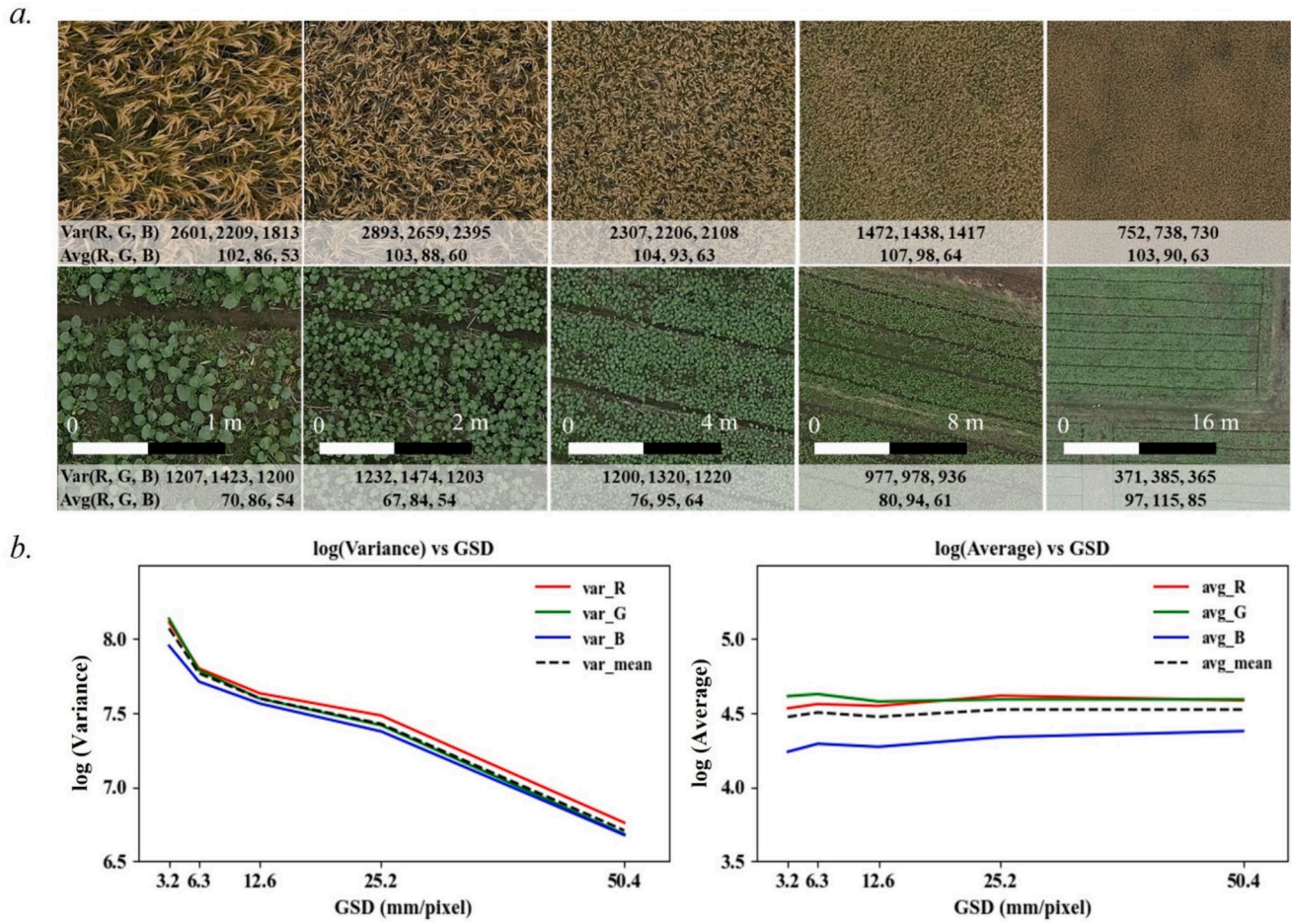

Fig. 3. Higher altitudes reduce variance significantly while the average value remains constant: a. Samples of cropland images collected at 25, 50, 100, 200, and 400 m altitudes, and b. changing trends of overall variance and average value vs. GSD across multiple altitudes on 29,000 aerial images.

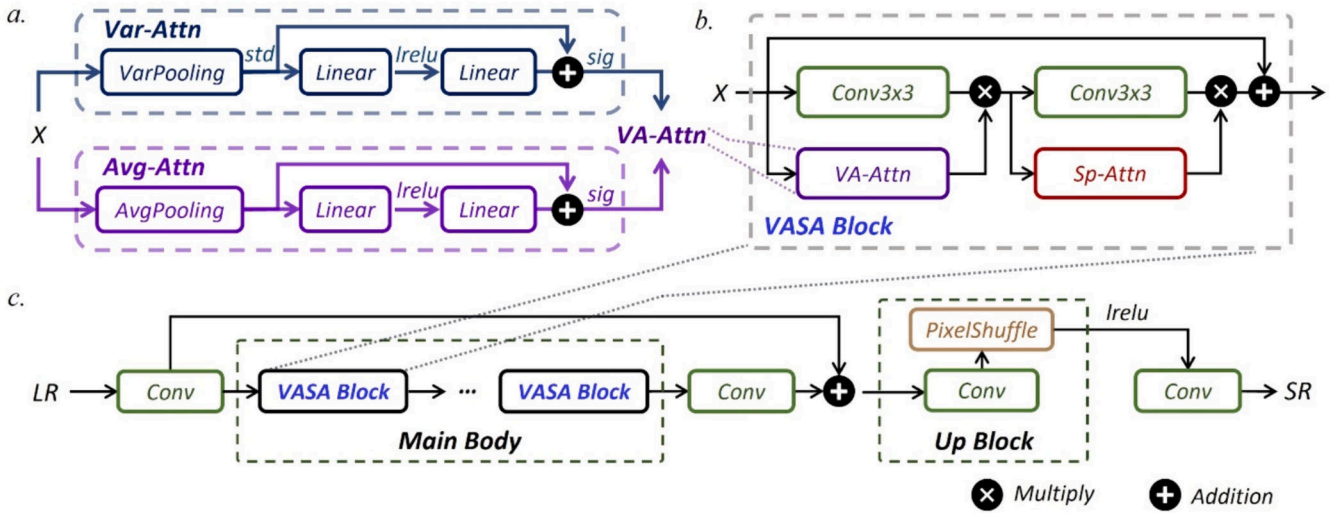

Fig. 4. Architecture of a. Variance-Average Attention (VA-Attn), b. The Variance-Average-Spatial Attention (VASA) Block, and c. the basic VASA SR (VASR) model for × 2 scale. VA-Attn is implemented to the first convolution layer of the VASA Block, and the main body of VASR is formed by stacking VASA Blocks.

$$v_X^{cw} = var_{cw}(X),\ v_{std}^{cw} = \frac{v_X^{cw} - mean(v_X^{cw})}{\sqrt{var(v_X^{cw})}} \tag{1}$$

$$Attn_{var} = sig\left(FC2\left(v_{std}^{cw}\right) + v_{std}^{cw}\right) \tag{2}$$

where $X \in \mathbb{R}^{W\times H\times C}$ represents the input features with W width, H height, and C channels. $var_{cw}()$ and $mean_{cw}()$ denote the variance and average statistics of each channel, while $var()$ and $mean()$ denote the variance and average statistics of all elements. $sig()$ means the sigmoid function, and $FC2()$ is a 2-layer learnable network with $\frac{C}{8}$ hidden layer channels activated by a leaky-ReLU ($lrelu$) function. Besides, the $Attn_{var} \in \mathbb{R}^{1\times 1\times C}$ is the Var-Attn.

The Avg-Attn is calculated from the channel-wise mean value $m_X^{cw} \in \mathbb{R}^{1\times1\times C}$, in a similar process while without standardized:

$$m_X^{cw} = mean_{cw}(X), Attn_{avg} = sig\left(FC2\left(m_X^{cw}\right) + m_X^{cw}\right) \quad (3)$$

It is worth noting that standardization is only applied to variance to mitigate the impact of extreme values and convert data of different magnitudes to the same level. In the case of Avg-Attn, standardization is usually not employed (Zhang et al., 2021b). As the average value implies the attribute and category of one feature map, the standardization may shift the mean and harm the model optimization.

The ultimate VA-Attn, $Attn_{VA} \in \mathbb{R}^{1\times1\times C}$, is acquired by element-wise addition. The VA-Attn is inserted into convolution layer by side, as shown in Fig. 4**b**.

$$Attn_{VA} = Attn_{var} + Attn_{avg} \quad (4)$$

For SR models, the feature maps from different channels store various feature attributes, and the importance of each channel varies spatially (Kim et al., 2020). Therefore, channel-specific spatial attention (Woo et al., 2018) is implemented in the second convolution layer of one block, as shown in Fig. 4**b**. The spatial attention (*Sp-Attn*), $Attn_{Sp} \in \mathbb{R}^{W\times H\times C}$, is achieved using a depthwise separable convolution *DWConv*() and a sigmoid function:

$$Attn_{Sp} = sig(DWConv(X)) \quad (5)$$

where the kernel size and number of *DWConv*() are 3×3 and match the number of input channels (Howard et al., 2017), the 2-layer residual convolution block, injected with VASA through element-wise multiplication, is illustrated in Fig. 4**b**. To state clear, the attention or features are denoted by capitalized words, while all model names are presented in uppercase.

3.2. VASA super resolution model

Among regression-based SR models, EDSR is one of the most classic and effective architectures (Lim et al., 2017). Compared to the earlier SRResNet of SRGAN (Ledig et al., 2017), EDSR removes batch normalization layers, which can be detrimental to range flexibility in the residual blocks of the main body. Therefore, the effectiveness of VASA on regression-based models was evaluated using EDSR in this part.

We replace the original residual blocks of EDSR with the proposed VASA block in the main body to construct a VASA Super-Resolution (VASR) model. Most of the components and details are kept unchanged for a fair comparison. As shown in Fig. 4**c**, the VASR network includes a main body and an up-sampling (Up) block. By default, the input low-resolution sample $LR \in \mathbb{R}^{W\times H\times3}$ is expanded to 64 channels first, and 16 VASA blocks contained in the main body further process the features. After the residual addition, the feature spatial size would be upscaled by the Up block, which consists of a stacked convolution layer and pixel-shuffle layers. The upsampled feature maps are activated through a *lrelu* function and aggregated to $HR \in \mathbb{R}^{rW\times rH\times3}$ by a convolution layer in the end.

Three down-sampling operations (including bilinear, area, and bicubic) are randomly sampled for self-supervised training to construct low-resolution pairs. The L1-norm between *SR* and *HR* is utilized as the training loss for better convergence:

$$\mathbb{L}_1 = \|\mathbb{F}(LR) - HR\|_1 \quad (6)$$

where $\mathbb{F}()$ denotes the SR model, $||\ ||_1$ is the 1-norm distance. For further details, please refer to Lim et al. (2017) and the model implementation specifics provided in **Section 4.1**.

3.3. VASA residual dense GAN model

GAN-based methods incorporate an additional discriminator and composite loss to enhance training, producing more realistic images than regression-based models. The Residual-in-Residual Dense Block (RRDB) from ESRGAN (Wang et al., 2018) has been widely adopted as a generator structure in remote sensing SR and other applications (Dong et al., 2022). Consequently, RRDB was selected as the basic GAN-based model for this study.

To construct a VASA-enhanced GAN model, we modified the RRDBs by adding the VA-Attn and Sp-Attn alternately to the first four convolution layers, as shown in Fig. 5**a**. The VASA Residual Dense (VARD) block contains five convolution layers, where each layer is fed with the concatenated features from all previous outputs in the block. And the residual scaling (× 0.2) is implemented to the output features of the last convolution layer before adding them to the main path for stability (Wang et al., 2018).

The overall architecture of the VASA Residual Dense GAN (VARDGAN) is illustrated in Fig. 5**b**. Similar to that of VASR, the input low-resolution sample $LR \in \mathbb{R}^{W\times H\times3}$ is first expanded by a convolution layer. The features are then processed by the main body, which consists of stacked VARD blocks. Unlike VASR, the up-sampling progress is achieved by interpolation to maintain consistency with the original ESRGAN.

For practical aerial image SR tasks, the actual degradations (from HR to LR) are unknown and complex. To achieve satisfactory reconstruction in real-world scenarios, we adopted a high-order degradation (HOD) process from Real-ESRGAN (Wang et al., 2021b) to synthesize a variety of LR samples for self-supervised training. The HOD involves randomly sampled degradation processes, repeated twice, including blurring, down-sampling, noising, and JPEG compression. This combination of multiple processes forms a large set, simulating and covering the real degradation conditions of low- and high-altitude aerial images.

The U-Net discriminator from Real-ESRGAN is also kept for per-pixel feedback during training. The total loss target of the VARDGAN generator is a hybrid of perceptual, adversarial, and consent loss:

$$\mathbb{L}_G = \mathbb{L}_{percep} + \alpha\mathbb{L}_{UDisc} + \beta\mathbb{L}_1 \quad (7)$$

where $\mathbb{L}_{percep}$ is the perceptual loss based on pre-trained VGG19 networks, which measures the distance of *SR* and *HR* in feature space; $\mathbb{L}_{UDisc}$ denotes the adversarial loss from the U-Net discriminator with spectral normalization as defined to represent per-pixel realness and α, β are weights used to control the loss importance. For details, please refer to (Wang et al., 2021b) and implementation details in **Section 4.1**.

3.4. Efficient VASA-enhanced diffusion model

The diffusion model includes a forward diffusion process that gradually adds noise to the data and a reverse process that removes the noise step-by-step. DMs can capture the complex statistics of the visual features and demonstrate promising performance in high-quality image super-resolution (Li et al., 2022). To efficiently process iterative denoising, we followed the LDM framework (Rombach et al., 2022).

LDM first compresses the input data into a compact latent space representation. During the diffusion process, Gaussian noise is sampled and injected into this latent space data, progressively degrading it towards pure noise. In the reverse denoising process, a learnable network is trained to predict the noise of each diffusion step, using the low-resolution data as guidance for SR. By iteratively removing the predicted noise, the original features could be restored.

However, the denoising UNet is computationally heavy and requires dozens of iterative forwarding steps to generate satisfactory results (Li et al., 2023). To focus on this bottleneck, we designed a time-conditional VASA residual (TVR) block, as shown in Fig. 6**a**. Its structure is similar to that of a basic VASA block with two convolution layers. The TVR additionally received a sinusoidal time-embedding diffusion step integrated into the median features after a learnable linear layer. We also modified the block by scaling the skip connections by $1/\sqrt{2}$ to facilitate

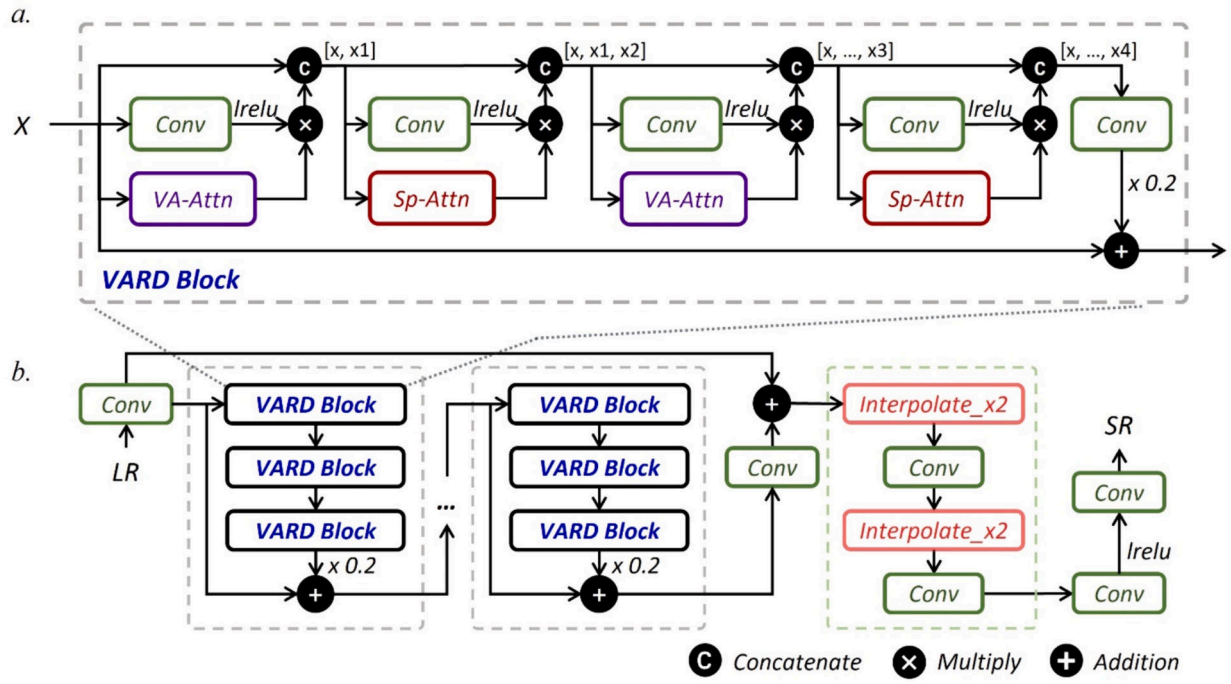

Fig. 5. Structure of a. VASA Residual Dense (VARD) Block and b. the derived VASA Residual Dense GAN (VARDGAN) SR generation model. Each block contains five layers of convolution, and every three stacked blocks of the main branch have a residual skip connection.

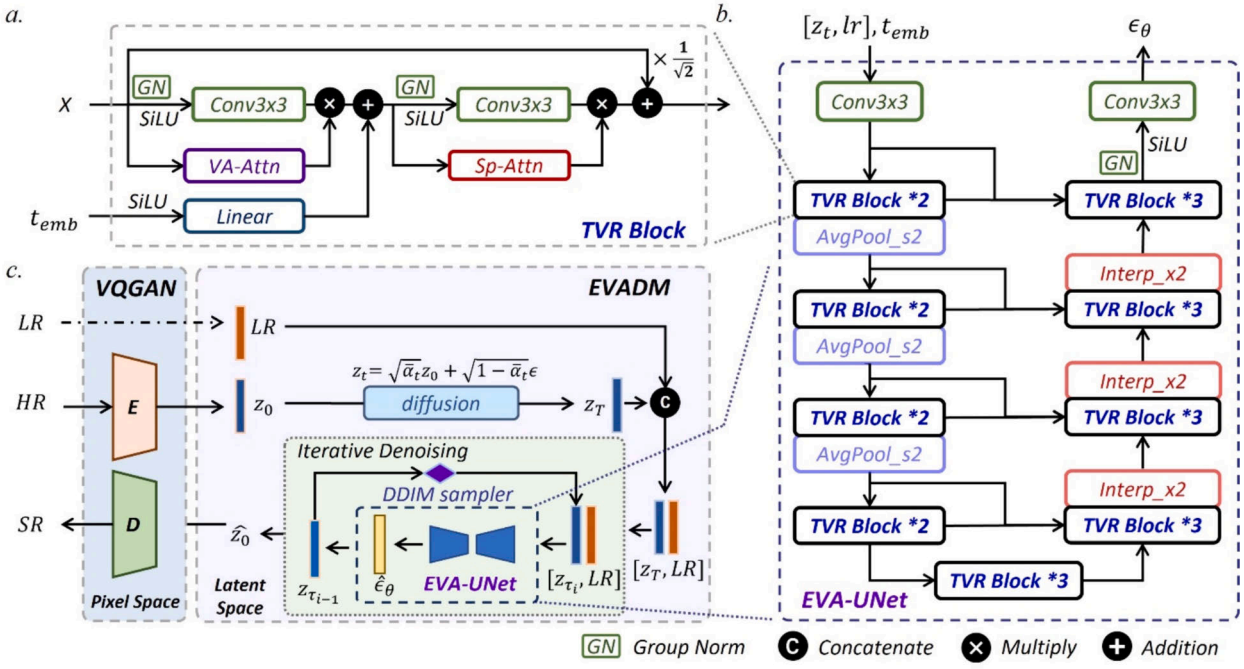

Fig. 6. Structure of a. Time-conditional VASA Residual block (TVR), b. the derived Efficient-VASA UNet (EVA-UNet) model for noise prediction and c. The workflow of the Efficient VASA Diffusion Model (EVADM). TVR contains a time embedding branch for injecting the step information, and EVA-UNet predicts noise at each diffusion step for iterative denoising (feature reconstruction). Notably, all diffusion and denoising processes are conducted in the VQGAN-compressed latent space.

convergence (Saharia et al., 2022b). The group normalization is conducted before the SiLU activation, and the normalized features are further fed into the convolution layer, which all line up with that of LDM.

We further optimized the overall structure of UNet for efficiency, drawing from recent studies such as SR3+ (Sahak et al., 2023) and SnapFusion (Li et al., 2023). Cross-attention layers, which limit generalization to different resolutions and have quadratic computation complexity relative to feature size (Li et al., 2023), were removed to create a fully convolutional structure. A comparative experiment was conducted to assess the effect of CA layers in **Section 5.1**. Due to the expanded channel dimensions, the parameters of UNet are concentrated in the middle stages (Li et al., 2023). Therefore, as shown in Table 2, for the standard EVADM_x4, we restrict the channel dimensions (channel

multipliers) while adding one more TVR block for the low-resolution stage to maintain the model capacity. The ultimate efficient VASA UNet (EVA-UNet) structure is shown in Fig. 6**b**, with the standard x4 model architecture breakdown detailed in Table A.4.

All components and overall workflow of the EVADM are illustrated in Fig. 6**c**. During training, the high-resolution ground truth ($HR \in \mathbb{R}^{rW\times rH\times 3}$) in pixel space is first compressed by a VQGAN encoder (E_{VQ}) into the latent space (Rombach et al., 2022):

$$z_0 = E_{VQ}(HR) \tag{8}$$

where $r \in 2^N$ represents the SR ratio, $z_0 \in \mathbb{R}^{W\times H\times c}$ is the latent presentation, and c denotes the number of output channels. By default, c is set to 3, but for $r \geq 8$, to ensure a decent reconstruction, c is set to 4. As shown in Table 2, the input channel number of EVADM_x8 is 7 (3 of which are the *LR* channels).

In the forward diffusion process, Gaussian noise q is sampled and gradually added to the latent space variable z_0, until the original signal is completely destroyed, eventually resulting in $z_T \sim \mathcal{N}(0, \mathrm{I})$. The sampled noise q is controlled by a fixed linear variance schedule $\beta_1, \cdots, \beta_T$:

$$q(z_t|z_{t-1}) = \mathcal{N}\left(z_t; \sqrt{1-\beta_t}z_{t-1}, \beta_t \mathrm{I}\right) \tag{9}$$

For an arbitrary timestep, $t \sim \{0, \cdots, T\}$, the median z_t can be reparameterized as:

$$z_t = \sqrt{\overline{\alpha}_t}z_0 + \sqrt{1-\overline{\alpha}_t}\epsilon \tag{10}$$

where $\overline{\alpha}_t = \prod_{s=1}^{t}(1-\beta_s)$, and $\epsilon \sim \mathcal{N}(0, \mathbf{I})$. For the derivation of this formula, please refer to Ho et al. (2020).

During training, we sample t and ϵ, then calculated z_t, which is concatenated with *LR* input and fed into the EVA-UNet, denoted as $\widehat{\epsilon}_\theta()$, which is parameterized by θ. The L2 distance between the estimated noise $\widehat{\epsilon}$ and the actual noise ϵ is used as MSE loss for optimizing the noise predictor:

$$\mathbb{L}(\theta) = \mathbb{E}_{LR,t,\epsilon}\|\widehat{\epsilon}_\theta(LR, z_t, t) - \epsilon\|_2^2 \tag{11}$$

To accelerate the generative processes, the DDIM sampler is used during the inference process (Song et al., 2021). Unlike the original Denoising Diffusion Probabilistic Models (DDPM), which sampled t sequence from $L = [T, T-1, \cdots, 1]$ (Ho et al., 2020), a subsequence $\tau = [\tau_s, \tau_{s-1}, ..., 1]$ is constructed from L. This skipping process is feasible because the DDPM sampling sequence used during training includes this subsequence.

When $\tau = 1$, the denoising process is completed, resulting in an estimated $\widehat{z_0}$. This super-resolved sample is then converted from the latent space to the pixel space by the VQGAN decoder.

A high-order degradation method similar to BSRGAN was adopted for all diffusion models to construct SR data pairs in self-supervised training (Zhang et al., 2021a). This method combines multiple degradation techniques, including blurring, down-sampling, noise, and JPEG compression, with a random shuffle mechanism similar to the high-order degradation discussed in **Section 3.3**.

3.5. Super-resolution relative fidelity index

In SR tasks, reference-based metrics like PSNR and SSIM (Wang et al., 2004) are commonly used. However, these metrics often over-penalize the reconstructed high-frequency details that are not perfectly aligned with the target image, leading to discrepancies with subjective evaluations, especially for large SR magnifications (Saharia et al., 2022c).

To address this issue, perceptual metrics such as FID ((Heusel et al., 2017)) and LPIPS (Zhang et al., 2018) have been reported in recent studies. They use pre-trained networks to extract feature maps and measure similarity in feature space, providing evaluations that correlate better with human observations.

The difficulty of reconstructing different datasets or scenarios varies in image SR tasks. For example, animated cartoon images are much easier to recover than real-world photographs. The imaging quality of UAV can be affected by the instability of flight environments and illumination conditions (Sidike et al., 2018). Additionally, for real-matched remote sensing SR pairs, intrinsic disparities across real SR image pairs arise due to temporal variations, cloud contamination, or environmental changes. Therefore, we developed a relative FID (RFID) metric to measure how the synthetic SR images compare to the upsampled LR images:

$$RFID = \frac{FID_{up} - FID_{sr}}{FID_{up}} = 1 - \frac{FID_{sr}}{FID_{up}}, RFID < 1 \tag{12}$$

where FID_{sr} is the FID score of the SR images, and FID_{up} is the FID score of the upsampled images (with bilinear interpolation adopted).

RFID is the relative FID score of the SR compared to the interpolated ones, and a large value (close to 1) suggests better performance. The baseline distance FID_{up} below suggests the disparity between the LR and HR data, where a larger value indicates a higher level of difficulty (lower data reliability) in reconstruction and a more significant gap in distance.

The RFID score can evaluate the results in a relative and perceptual view at a fixed SR scale. For different scales, because the FID_{up} increases dramatically with the SR scale rising, high-scale SR results in a higher *RFID*. This evaluation is inconsistent with the perceptual intuition that low-scale SR is easier and can yield better samples.

Besides, alignment and consistency remain important for SR tasks, especially for low-scale SR and regions with fewer high-frequency details (Sahak et al., 2023). A comprehensive assessment of SR quality can be achieved by using multifaceted evaluation metrics (Wells et al., 2023). Therefore, to evaluate multi-scale SR models both visually photorealistic and statistically close to the HR, we combined both the perceptual measure RFID and structural similarity SSIM and designed a cross-scale SRFI:

$$SRFI = 10^{(SSIM+\alpha\bullet RFID)/r}, 0 < SRFI < 10^{\frac{\alpha+1}{r}} \tag{13}$$

where *SSIM* is the structural similarity index of the SR, r is the SR ratio, and α controls the weight of perceptual measure (by default, $\alpha = 2$).

SRFI is a hybrid SR index that considers both image structural similarity, perceptual distance, and the SR ratio, where a large value suggests better SR quality.

4. Results and case studies

In this section, the experimental setup and implementation details were first outlined. Subsequently, in **Sections 4.2 and 4.3**, the experiment results on the holdout CropSR-Test dataset and the real CropSR-OR/FP dataset were presented. **Section 4.4** compared the models' generalization capabilities on the AgV dataset. Finally, The last section

Table 2
EVA-UNet architecture parameters.

Model	r	input shape	output shape	channel dim	channel multipliers	blocks number
LDM_x4	4	6x64x64	3 × 256^2	160	{1, 2, 2, 4}	2
EVADM_x2	2	6x128x128	3 × 256 × 256	128	{1, 2, 2}	2
EVADM_x4	4	6x64x64	3 × 256^2	160	{1, 2, 2, 2}	2
EVADM_x8	8	7x64x64	3x512x512	160	{1, 2, 2, 4}	2

The suffix_xr and r indicate the super-resolution ratio. The capitalized model name indicates an altered structure.

validated the models' effectiveness through practical downstream task facilitation.

4.1. Implementation details

In this study, all models were trained on the CropSR dataset using the previously mentioned degradation methods to synthesize LR samples. Random transposition and flip data augmentation techniques were employed to improve the model's robustness and avoid overfitting (Khan et al., 2023; Lu et al., 2022). For the × 2 and × 4 CropSR datasets, the intrinsically paired HR-LR pairs (with 256–512 and 128–512 size pairs) were simultaneously cropped to a random 128–256 and 64–256 size pair. For × 8, a full 64–512 size pair was adopted to incorporate sufficient prior contextual information for LR. All training processes were initiated from scratch with Kaiming initialization.

For regression- and GAN-based models, the Adam optimizer with $\beta_1, \beta_2 = 0.9, 0.99$, and an initial learning rate of 1×10^{-4} was used. Guided by the prior fitting test, the mini-batch size and total iteration steps were set to 16 and 5×10^4. For diffusion-based models, the AdamW optimizer was adopted with $\beta_1, \beta_2 = 0.9, 0.999$, and an initial learning rate of 1×10^{-6}. The batch size was set to 32 with accumulated gradients per 2 batches, and a total of 1×10^5 training iterations were used for abundant fitting. The training weights were saved and evaluated every 5,000 iterations. The best-trained weights on the evaluation dataset were kept for testing. For diffusion models, 16 refinement steps were utilized for × 2, and 32 steps for × 4 and × 8 models (optimized based on preliminary studies in **Section 5.4**).

All models were implemented using Python and PyTorch framework and trained on an NVIDIA RTX 4090 GPU. The diffusion model was based on the LDM (Rombach et al., 2022). The regression- and GAN-based models were derived from MMEDIT (MMEditing, 2022).

4.2. Results on CropSR dataset

This section focuses on the performance of the baseline and proposed models on the simulated CropSR-Test after training with different SR ratios (r). The representative baseline methods include the regression-based EDSR (Lim et al., 2017), GAN-based Real-ESRGAN (Real-ESRGAN) (Wang et al., 2021b), and LDM (Rombach et al., 2022), which serve as the base models for the experiments.

Table 3 reports the quantitative results in terms of reference-based metrics PSNR and SSIM, perceptual metrics FID (as the key performance indicator), and SRFI as introduced in **Section 3.5**, with the best performance of each SR ratio (r) highlighted in bold. The model computational costs (floating point operations, FLOPs) and the number of weight parameters (Parms) are also listed.

On the simulated CropSR-Test, it is observed that no single type of baseline model consistently performs well across different magnification ratios. For the × 2 ratio, the regression-based EDSR_x2 scored highly in all metrics. The GAN-based RealESRGAN_x2 performed poorly in terms of FID, indicating it failed to reconstruct details perceptually. This might be due to the mode collapse of GANs (Thanh-Tung and Tran, 2020), where we found that the colorful flowers in LR samples faded in the reconstructed SR samples. For the × 4 ratio, the LDM_x4 achieved a higher FID and SRFI than both EDSR_x4 and RealESRGAN_x4. We also expanded the original EDSR and RealESRGAN to × 8 ratio by adding additional up-sampling layers, but the results were unsatisfactory. Although EDSR_x4 and EDSR_x8 acquired high values from the reference-based metrics, the reconstructed images were over-smoothed and lacked details (Li et al., 2022; Wang et al., 2018). This is also the reason why EDSR achieved worse FID and SRFI scores than RealESRGAN in the × 4 SR task.

The results comparison of baseline models and the VASA-enhanced counterparts are shown in Table 3, including EDSR_x2 vs. VASR_fc1_x2 and RealESRGAN_x4 vs. VARDGAN_x4. The injected VASA attention proved to be effective, requiring only a small fraction of computations while significantly enhancing the results. For the × 2 SR ratio, our VASR_fc1_x2 achieved the highest performance across all metrics, primarily due to the high structural similarity. The suffix_fc1 denotes that the variance-attention and average-attention share one fully connected layer, whereas, by default, the two branches are separate. This fc1 trick is beneficial for regression models, and we also conducted further ablation studies in **Section 5.1**. For the × 4 SR ratio, the proposed EVADM_x4 attained the best FID and overall SRFI evaluation results, with a decrease of 5.7 in FID and a 7 % increase in SRFI compared to the LDM_x4 baseline. Furthermore, the number of parameters was nearly halved, with a reduction in computation by over 10 %.

The visual comparisons of crucial models on the CropSR-Test are presented at the top of Fig. 7. For the × 2 SR ratio, the sample from VASR_fc1_x2 reveals better details with brilliant color and sharp edges around the red and yellow flowers. Although EVADM_x2 reconstructed a more realistic crop canopy, it did not process well with the rare flowers and tended to erase them. For the more challenging × 4 SR task, the images generated by GANs show significant distortion and discoloration. In contrast, the reconstructed features from the diffusion model may not perfectly align with the ground truth (GT), but a similar pattern is often reconstructed.

Direct × 8 ratio augmentation SR experiments were also conducted. Due to significant information loss, the model could only infer and hallucinate similar features in many details. However, in contrast to regression- and GAN-based methods, the SR results generated by EVADM_x8 exhibit clear edges and distinguishable details of the objects, with considerably better numerical indicators, reducing FID by 67.1. This validates the feasibility of diffusion models for high-ratio image SR. Due to limited space, the visual comparisons of all x8 models are provided in Appendix B.

Table 3
Model performance comparison on CropSR-Test.

Model	Type	r	PSNR↑	SSIM↑	FID↓	RFID*↑	SRFI↑	FLOPs/G↓	Parms/M↓
EDSR_x2	Reg	2	26.18	0.90	6.07	0.70	14.17	90.18	1.37
VASR_fc1_x2	Reg	**2**	**26.63**	**0.92**	**4.74**	**0.77**	**16.74**	90.87	1.45
RealESRGAN_x2	GAN	2	21.75	0.78	27.13	−0.34	1.11	294.24	37.78
EVADM_x2	Diff	2	22.69	0.83	10.01	0.50	8.29	22.97	29.14
EDSR_x4	Reg	4	**19.18**	**0.49**	51.78	0.53	2.44	130.27	1.52
RealESRGAN_x4	GAN	4	17.27	0.44	40.38	0.63	2.67	1176.61	37.77
VARDGAN_x4	GAN	4	17.43	0.45	31.73	0.71	2.95	1188.19	40.71
LDM_x4	Diff	4	16.37	0.35	31.36	0.71	2.79	40.22	113.62
EVADM_ca_x4	Diff	4	16.21	0.34	25.98	0.76	2.93	35.96	64.03
EVADM_x4	Diff	**4**	16.49	0.36	**25.66**	**0.77**	**2.97**	35.91	63.20
EDSR_x8	Reg	8	**16.79**	**0.20**	131.24	0.41	1.34	290.60	1.67
RealESRGAN_x8	GAN	8	14.84	0.13	280.12	−0.27	0.89	1453.58	37.77
EVADM_x8	Diff	**8**	14.30	0.14	**64.14**	**0.71**	**1.57**	35.92	63.21

The best result of each scale is in **bold**. * FIDs of 2x, 4x, and 8x Bilinear Interpolation are 20.189, 109.829, and 220.712. The suffix_xr and r indicate the super-resolution ratio. The_fc1 means VA and AA share one FC layer. The_ca denotes the cross-attention.

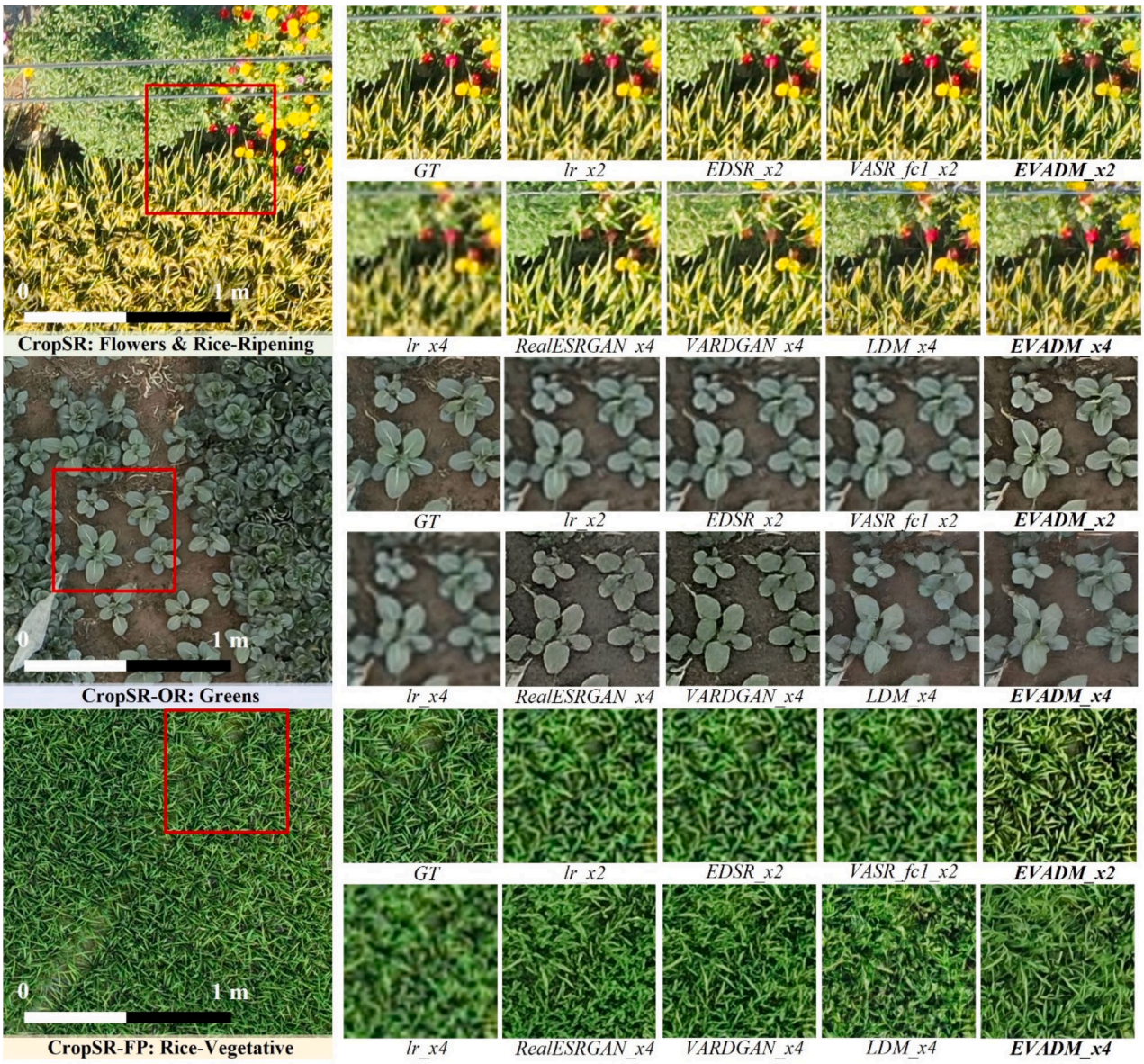

Fig. 7. Visual comparison of SR results on CropSR-Test, CropSR-OR, and CropSR-FP dataset. The left side provides an overview of different cropland scenes, while the model names are displayed below the small blocks on the right.

4.3. Results on CropSR-OR/FP dataset

Although similar objects exist in the CropSR dataset, the real-world degradation and imperfect matching of data in CropSR-OR and CropSR-FP present challenges to the SR models. We tested the primary models on CropSR-OR/FP, as shown in Table 4. Although the regression-based EDSR_x2 and VASR_fc1_x2 still achieved high SSIM values, their SR results are blurred and not much better than the interpolation, as presented in the middle (CropSR-OR) and bottom (CropSR-FP) of Fig. 7. For both the × 2 and × 4 ratios, the proposed EVADM models achieved

Table 4
Model Performance Comparison on CropSR-OR/FP.

Model	r	CropSR-OR				CropSR-FP			
		SSIM↑	FID↓	RFID*↑	SRFI↑	SSIM↑	FID↓	RFID*↑	SRFI↑
EDSR_x2	2	0.55	66.60	0.13	2.54	0.50	78.71	0.14	2.46
VASR_fc1_x2	2	0.55	65.96	0.14	2.60	0.50	78.14	0.15	2.49
EVADM_x2	**2**	0.42	**52.30**	0.32	**3.36**	0.35	**62.50**	0.32	**3.11**
RealESRGAN_x4	4	0.29	124.75	0.00	1.17	0.35	141.89	0.11	1.39
VARDGAN_x4	4	0.30	116.09	0.07	1.28	0.34	126.24	0.21	1.55
LDM_x4	4	0.28	61.75	0.50	2.10	0.35	92.47	0.42	1.98
EVADM_x4	**4**	0.28	**52.47**	0.58	**2.28**	0.33	**85.78**	0.46	**2.06**
EVADM_x8	8	0.15	101.72	0.57	1.45	0.21	151.09	0.39	1.33

The best result of each scale is in **bold**. The r indicates the super-resolution ratio. * FIDs of 2x, 4x, and 8x Bilinear Interpolation for −O and −FP are 76.706, 124.323, 239.271 and 91.601, 159.668, 246.671. The_fc1 means VA and AA share one FC layer.

the best FID and highest comprehensive SRFI index. On the two datasets, EVADM_x2 achieved an average reduction of 14.6 in the FID compared with the best baseline model. Similarly, EVADM_x4 obtained an impressive FID distance decrease of 8.0. Regarding the SRFI index, the EVADM_x2 and x4 showed average gains of 27.1 % and 6.3 %, respectively.

The GAN-based models exhibit a near-zero RFID score, suggesting their inability to generate perceptual perceptually satisfactory SR results. By observing the VARDGAN_x4 image of the second sample, the detailed features on the leaves are omitted, while EVADM_x4 depicts vivid leaf veins in the middle of Fig. 7. For × 4 SR of dense crop canopy in the third row, the GAN models are unable to reconstruct the narrow blades. By comparing the last two images generated by the x4 diffusion models, it can be concluded that our VASA attention mechanism significantly improved the ability of SR models to reconstruct fine details.

In contrast to the performance of regression-based models on simulated CropSR-Test datasets, the decrease in numerical metrics and visual quality of the reconstruction results on real SR datasets implies their limited generalization ability. By comparison, the diffusion-based models are more robust to domain shifts.

A single metric is insufficient for comprehensive model evaluation and can lead to inconsistent assessments across datasets. In Table 4, on the CropSR-OR dataset, EVADM_x4 and LDM_x4 have identical SSIM scores, and FID scores for EVADM_x2 and EVADM_x4 are nearly the same. Perceptually, EVADM_x2 is superior to EVADM_x4, which outperforms LDM_x4. This performances are difficult to assess with one metric. In addition, FID favors LDM_x4 over VASR_fc1_x2 on the OR dataset, but inconsistently, the evaluation reverses on the FP dataset, with VASR_fc1_x2 performing better. In practice, the x2 task has more initial information, and VASR_fc1_x2 is perceptually perceived as superior. The proposed SRFI provides consistent evaluation across the two datasets, aligning with the actual perception.

To further illustrate the capability of the SRFI in evaluating SR models across different datasets and ratios, we visualized the differences between FID and SRFI metrics in Fig. 8. By comparing the absolute difference (abs(Dif), lower in black); it is evident that the SRFI performs robustly across different scenarios, while the FID evaluation presents a larger disparity between the two datasets.

4.4. Generalization studies on agriculture-vision dataset

To investigate the generalization ability of the SR models across different data sources, this part compares the super-resolved images of the AgV test-set, as introduced in **Section 2.3**. All models were trained solely on CropSR and were not tuned. Considering there is no ground truth, the outputs are evaluated using no-reference image quality assessment metrics, including the Natural Image Quality Evaluator (NIQE) (Mittal et al., 2013) and the Perception-based Image Quality Evaluator (PI) (Blau et al., 2019). These two metrics are aligned with human ratings and are commonly used in remote sensing SR studies (Dong et al., 2022).

Table 5
Model Generalization Comparison on a part of the Agriculture-Vision dataset.

Model	Type	r	NIQE↓	PI↓
original	–	1	7.8035	5.7889
up_x2	Intp	2	8.4615	7.5576
EDSR_x2	Reg	2	8.0355	7.3005
RealESRGAN_x2	GAN	2	6.4506	4.9310
EVADM_x2	Diff	2	**4.2562**	**3.5697**
up_x4	Intp	4	9.6495	8.4720
RealESRGAN_x4	GAN	4	6.2094	4.8678
LDM_x4	Diff	4	3.4117	3.3903
EVADM_x4	Diff	4	**3.1454**	**3.2394**

The best result of each scale is in **bold**. The r indicates the super-resolution ratio.

The evaluation results on the 1,971 samples of the AgV dataset are listed in Table 5. For NIQE and PI, lower values indicate better SR image quality. The proposed EVADM models achieved the best scores for both the × 2 and × 4 SR ratios. Among existing methods, the GAN-based model outperformed the regression-based ones. By comparing the LDM and EVADM in the × 4 task, the effectiveness of VASA attention on the enhancement of model generalization ability is revealed.

The visual comparison of super-resolved images from AgV is shown in Fig. 9. The regression-based EDSR models failed to reconstruct details, representing blurred images similar to interpolation. This is due to the limited generalization ability of regression-based models, as stated in the previous section. The GAN-based RealESRGAN models produced scaly and poor features, especially for bare land and high-scale tasks. This issue becomes more pronounced with the increase of the SR scale (check the x8 results in Appendix B). This could be due to the mode collapse of GANs (Thanh-Tung and Tran, 2020).

In contrast, diffusion-based models provide greater realism and detailed representation. Moreover, the proposed EVADM surpasses other models by delivering more convincing results with enhanced textures and realistic details, such as the leaf blades above the crops and fine twigs atop the tree canopy, as shown in the last column of Fig. 9.

4.5. Downstream case studies

In this section, the utilities and effectiveness of the SR models are demonstrated by comparing the results of downstream tasks on super-resolved samples from different SR models. Samples exhibiting perceptually consistent detection boxes or segmentation masks are considered superior. Considering the efficiency and differences in SR, the existing

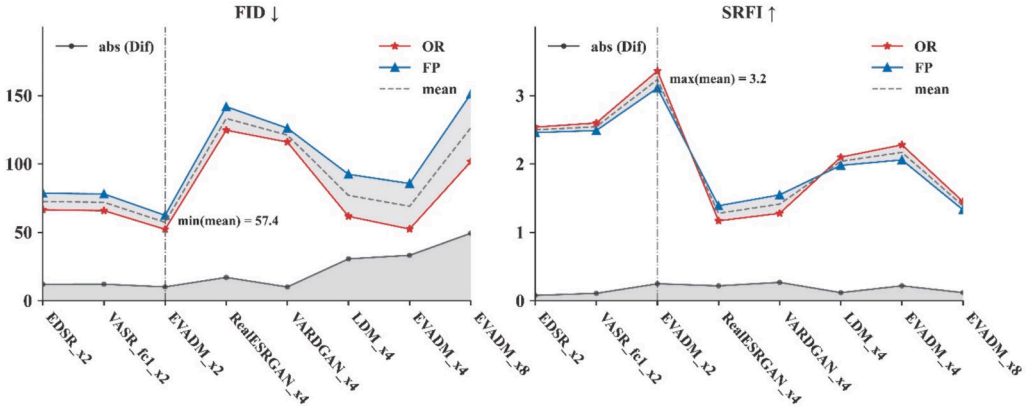

Fig. 8. Differences between FID and SRFI matrics for evaluating the same SR model on different real datasets (CropSR-OR/FP). The proposed SRFI provides more consistent evaluations across different datasets for the same model, with a smaller absolute difference (abs(Dif)).

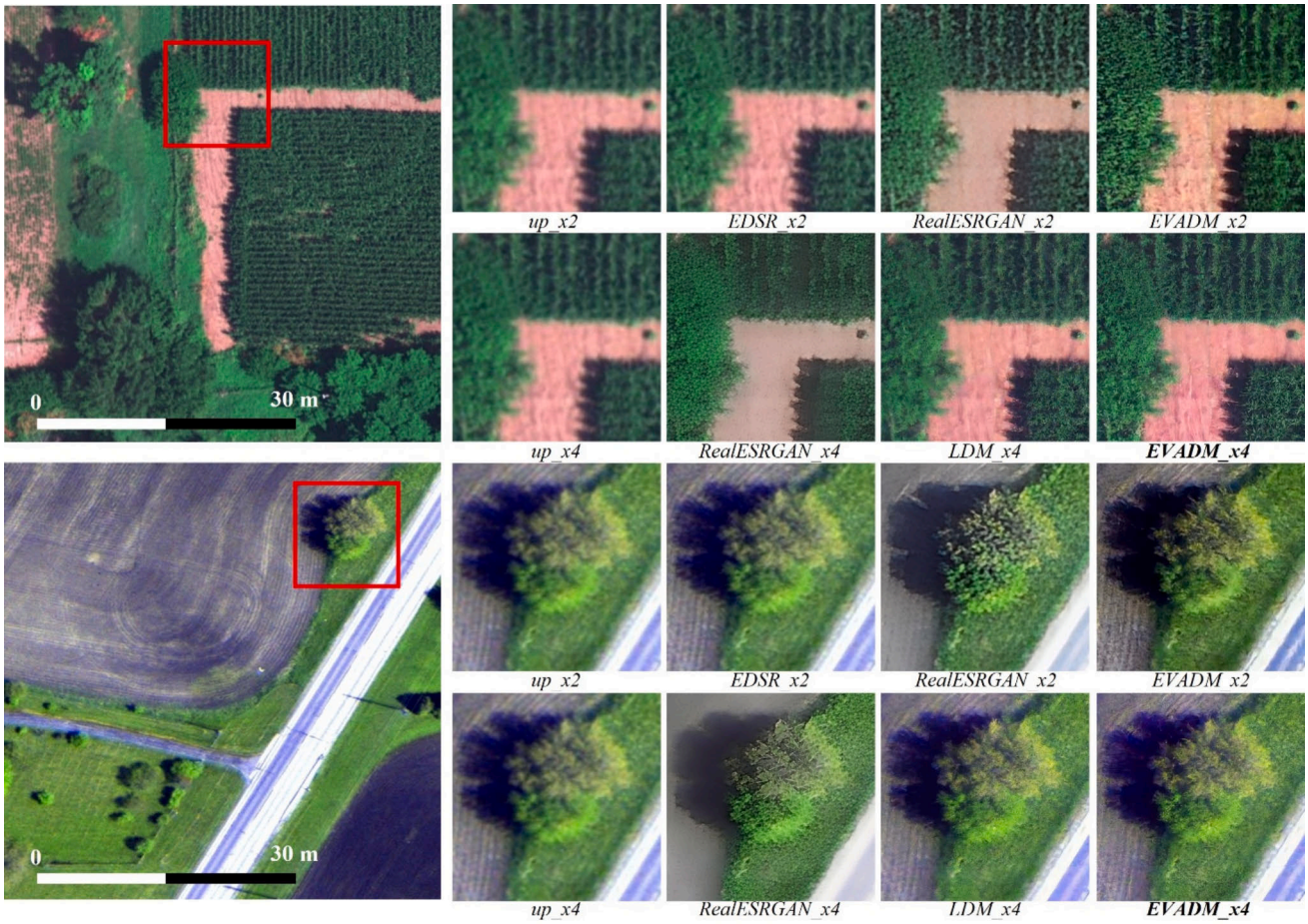

Fig. 9. Visual comparison of SR generalization studies on the Agriculture-Vision (AgV) dataset. The left side shows the FOV of AgV data samples, while the right side presents SR results, where EVADM produces more photorealistic views and finer details.

and proposed models were compared at a relatively high × 4 ratio.

Four kinds of downstream tasks were tested, including paddy field segmentation (PaddySeg), vegetable detection (VegDet), tree crown segmentation (CrownSeg), and tree detection (TreeDet). Images of the first two tasks were acquired from a P1 camera at a height of 100 m with a GSD of around 1.26 cm, while images of the other two were sourced from the AgV dataset with a GSD of 10 cm. For ease of presentation, the input pixel size of VegDet is 128x128, and for the other tasks, it is 512x512. As shown in Fig. 10, three typical existing SR models and the proposed EVADM were incorporated in the experiments. All models were trained only on the CropSR dataset, as detailed in **Section 4.1**, without fine-tuning.

For the first task, the downstream PaddySeg model proposed by Lu et al. (2023) was trained on 25 m P1 images for paddy segmentation and growth stage identification. As shown in the first input of Fig. 10, the left grid with sparse plants depicts the seedling stage (in lime green) of rice, and the right area with a dense crop canopy indicates the jointing stage (in yellow), according to agronomy observation. By comparing the predicted paddy masks, it is noticed that the fields are barely recognized from the EDSR_x4 sample. More adorable results are obtained from the RealESRGAN_x4, and the seedling stage is correctly recognized with LDM_x4 SR. Furthermore, the EVADM_x4 appears to produce the best perceptual results. This downstream comparison suggests that, compared to regression or GAN-based SR methods, the diffusion-based model generated images closer to the original high-resolution field (from high altitude, 100 m to low altitude, 25 m). Additionally, the proposed VASA further enhanced the diffusion SR model.

Apart from PaddySeg, the other three case studies are based on the interactive demo of T-Rex, a generic model that detects all objects with a similar pattern to the given visual prompts, such as points or boxes (Jiang et al., 2023). For VegDet and TreeDet tasks, four fixed typical targets (vegetables or trees) inputs are given for each SR sample. As depicted in the second and last column, the predicted instance boxes (in random color) in the lowest row are recalled more evenly and thoroughly compared with other cases. The CrownSeg task further employs the segmentation function of T-Rex, where the pixels of each tree crown are segmented separately. Many instances are omitted in the EDSR_x4 and RealESRGAN_x4 examples due to poor reconstruction quality. In the LDM_x4 case, almost all trees were detected, while a certain quantity of shadows was incorrectly segmented as individual crowns, such as the green and blue masks in the middle. The image from EVADM_x4 avoided this misleading trap in downstream tasks with a clearer output.

Overall, the images generated by the proposed SR method exhibit better detection boxes and segmentation masks across multiple downstream tasks. The reconstructed results are both visually appealing and practically effective. Although SR images cannot fully replace HR images, they are closer to the original HR domain and facilitate downstream models' performance.

5. Discussion

In this part, we first conducted ablation studies of VASA and other components of SR models. Then, in **Section 5.2**, we described the variance changes during degradation and SR processes. **Section 5.3** interpreted the relationship between the variance-average statistic values and network attention. In **Section 5.4**, we investigated the

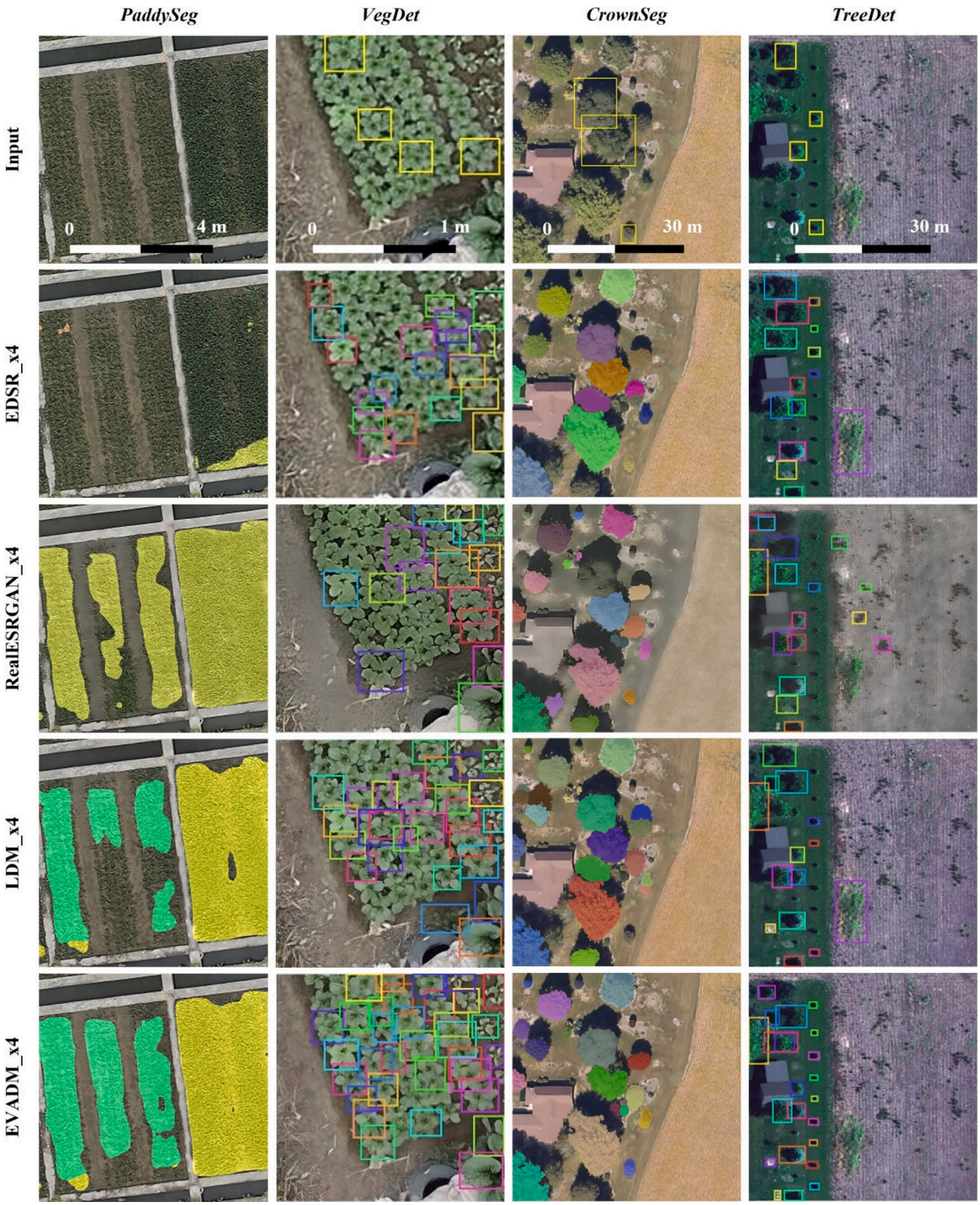

Fig. 10. Visual comparison of downstream tasks studies on × 4 SR models. PaddySeg and VegDet data are from the CropSR-OR/FP, while CrownSeg and TreeDet use samples from the AgV dataset. Only PaddySeg uses a pre-trained semantic segmentation model to predict paddy boundaries and stages; the others use the interactive detection tool T-Rex. The first row shows input visual prompts (boxes). Visually, SR data from EVADM produce more uniform and desirable results in downstream tasks.

influence of refinement steps on image quality in diffusion models. The final section discussed the potential future research directions in remote sensing SR.

5.1. Ablation studies

To demonstrate the effectiveness of the proposed VASA and other vital model components, we conducted a set of ablation experiments, as

shown in Table 6. The best results are highlighted in black for each type, including regression (Reg), GAN, Diffusion (Diff), and Effective Diff (DiffE).

For both the regression- and GAN-based models, those with only variance-spatial-attention (VA) (VSSR_x2, VSGAN_x4) provided considerable improvement compared to models with only average-spatial-attention (AA) (ASSR_x2, ASGAN_x4). The VASA combination exhibited the best overall performance in the SRFI evaluation. Interestingly, models that solely employ VA achieved the highest PSNR in both categories, suggesting that VA is beneficial for ensuring feature consistency in the SR process. Additionally, the fc1 trick mentioned in **Section 4.2** (where variance-attention and average-attention share one FC layer) demonstrated an improvement for the VASR_x2 model. However, in the case of VARDGAN_x4 and LDM_x4, the shared FC weights did not work and even caused a slight decline.

We also conducted comprehensive studies on DMs. A similar pattern was observed, where the combined VASA performed best, and the VA significantly facilitated the reference-based measure. By comparing the LDM_x4 and LDM-na_x4 models without cross-attention (CA), it was found that CA produced little effect despite occupying 10 % more parameters (LDM_x4: 113.62 M vs. LDM-na_x4: 103.77 M). Additionally, the performance of EVADM was better without the inclusion of CA (EVADM_ca_x4). Thus, CA was not adopted in the final version.

5.2. Variance change in pixel & feature space

To better understand the variance of images during the degradation and SR processes, we calculated the statistics of HR, degraded-upsampled LR, and VASR_fc1_x2 SR images, as illustrated in Fig. 11. It is observed that the degradation process results in the loss of detailed information and a decrease in variance across all channels. Higher degradation rates result in more significant information loss and variance reduction. In contrast, the SR process raises the variance to a level that closely approximates that of the HR image. This is due to the reconstructed high-frequency information, including fine details and sharp edges, which increase local deviations from the mean.

However, when evaluating the average value of each channel, it remains stable regardless of degradation or SR process. This suggested that the mean statistic helps determine the characteristics of an object, while the variance statistic assists in measuring the feature scale and image definition.

In addition, as shown in Fig. 12, we extracted the feature maps before the residual addition from each block of the VASR_fc1_x2 network. It was observed that the variance trend mentioned above also exists within the feature space of SR models, where the variance values of feature maps increase gradually, especially in shallow layers. The mean variance initially rises and then stabilizes around 0.7 after the eighth block, indicating subtle detail adjustments within the feature maps.

In contrast, as the depth increases, the average values of different channel feature maps gradually deviate from zero. The disparity of features expressed by different channels is enlarged, and these diverse features are merged into the SR images at the final output layer. Besides, it is observed from Fig. 12**a** that the shallow layers mainly focus on the differences in basic color information (B1/C16). As the depth increases, there is more emphasis on semantic information (B12/C64) and edge details (B16/C16).

5.3. How VASA attention works

To compare variance and average-attention, this section investigates the differences between statistics and their derived attention value across the depth of SR networks. The channel average and variance statistics represent different types of information. As described in **Section 3.1**, The variance-attention (Var-Attn) is obtained from the channel variance (Var) through standardization, FC layer modulation, and sigmoid processing successively. The average-attention (Avg-Attn) is derived from the channel average (Avg) value in a similar pattern without standardization. Wherein the FC layer can adjust the statistics adaptively after training (Kim et al., 2020). The sigmoid function weakens exceeded features and limits the distribution of the values between 0 and 1 while retaining the salient features. The statistic values and the derived attention values from the input of each block are graphed in Fig. 13.

As depicted in the first row of Fig. 13, the mean variances (in blue dashed line) of all channels are zero due to the standardization in the Var-Attn branch. Conversely, in the second row, the mean averages (in red dashed line) vary significantly across different blocks. The dispersion of average values from diverse channels exhibits a trend of increasing and then decreasing. The median diversity is advantageous for deep networks to focus on various features. As the network approaches output, the representations of each layer become more similar to ensure a robust feature aggregation for the final SR result.

The upper part of each sub-figure in Fig. 13 compares VA-Attn and Var/Avg-Attn (in black), where VA-Attn is acquired by adding Var-Attn and Avg-Attn. It can be seen that the patterns of VA-Attn across different channels are generally consistent with the Var-Attn, while the modulated Avg-Attn values are relatively small and have minimal impact. Therefore, it is inferred that Var-Attn plays a leading role in overall channel attention. In addition, the upper three graphs show that the original Var (in blue) and the derived Var-Attn are generally consistent

Table 6
Ablation study results on CropSR-test.

Model	Type	r	VA	AA	fc2	CA	PSNR↑	SSIM↑	FID↓	RFID*↑	SRFI↑
EDSR_x2	Reg	2	×	×	–	–	26.18	0.904	6.07	0.70	14.17
VSSR_x2	Reg	2	✓	×	–	–	**26.71**	**0.920**	5.04	0.75	16.22
ASSR_x2	Reg	2	×	✓	–	–	26.51	0.912	5.36	0.73	15.51
VASR_fc1_x2	Reg	**2**	✓	✓	×	–	26.63	0.917	**4.74**	**0.77**	**16.74**
VASR_x2	Reg	2	✓	✓	✓	–	26.45	0.913	5.73	0.72	14.89
RealESRGAN_x4	GAN	4	×	×	–	–	17.27	0.440	40.38	0.63	2.67
VSGAN_x4	GAN	4	✓	×	–	–	**17.46**	0.451	33.07	0.70	2.90
ASGAN_x4	GAN	4	×	✓	–	–	17.38	0.446	32.75	0.70	2.90
VARDGAN_fc1_x4	GAN	4	✓	✓	×	–	17.35	0.448	31.76	0.71	2.93
VARDGAN_x4	GAN	4	✓	✓	✓	–	17.43	**0.455**	**31.73**	**0.71**	**2.95**
LDM_x4	Diff	4	×	×	–	✓	16.37	0.353	31.36	0.71	2.79
LDM-na_x4	Diff	4	×	×	–	×	16.37	0.350	32.60	0.70	2.75
VSLDM_x4	Diff	4	✓	×	–	×	**16.45**	**0.353**	28.39	0.74	2.88
VASCLDM_fc1_x4	Diff	4	✓	✓	×	✓	16.20	0.339	28.80	0.74	2.84
VASCLDM_x4	Diff	4	✓	✓	✓	✓	16.33	0.344	**27.70**	**0.75**	**2.88**
EVADM_x4	DiffE	4	✓	✓	✓	×	**16.49**	**0.359**	**25.66**	**0.77**	**2.97**
EVADM_ca_x4	DiffE	4	✓	✓	✓	✓	16.21	0.342	25.98	0.76	2.93

The best result of each type is in **bold**. The r is the super-resolution ratio, and the VA, AA, and CA indicate variance, average, and cross-attention. The_fc1 means VA and AA share one FC layer. The suffix_ca denotes having CA, while –na means without CA.

	dsx4-upx4	dsx2-upx2	dsx2-SR	HR
Var(R, G, B)	822, 887, 698	903, 996, 769	1155, 1278, 1007	1224, 1335, 1067
mean(Var)	**802**	**889**	**1147**	**1209**
Avg(R, G, B)	42, 51, 37	42, 51, 37	42, 51, 37	42, 51, 36
mean(Avg)	43	43	43	43

Fig. 11. Variance changes from degradation and SR in pixel space. Degradation reduces image variance, while SR restores variance to the original HR level. The average value remains unaffected by these operations.

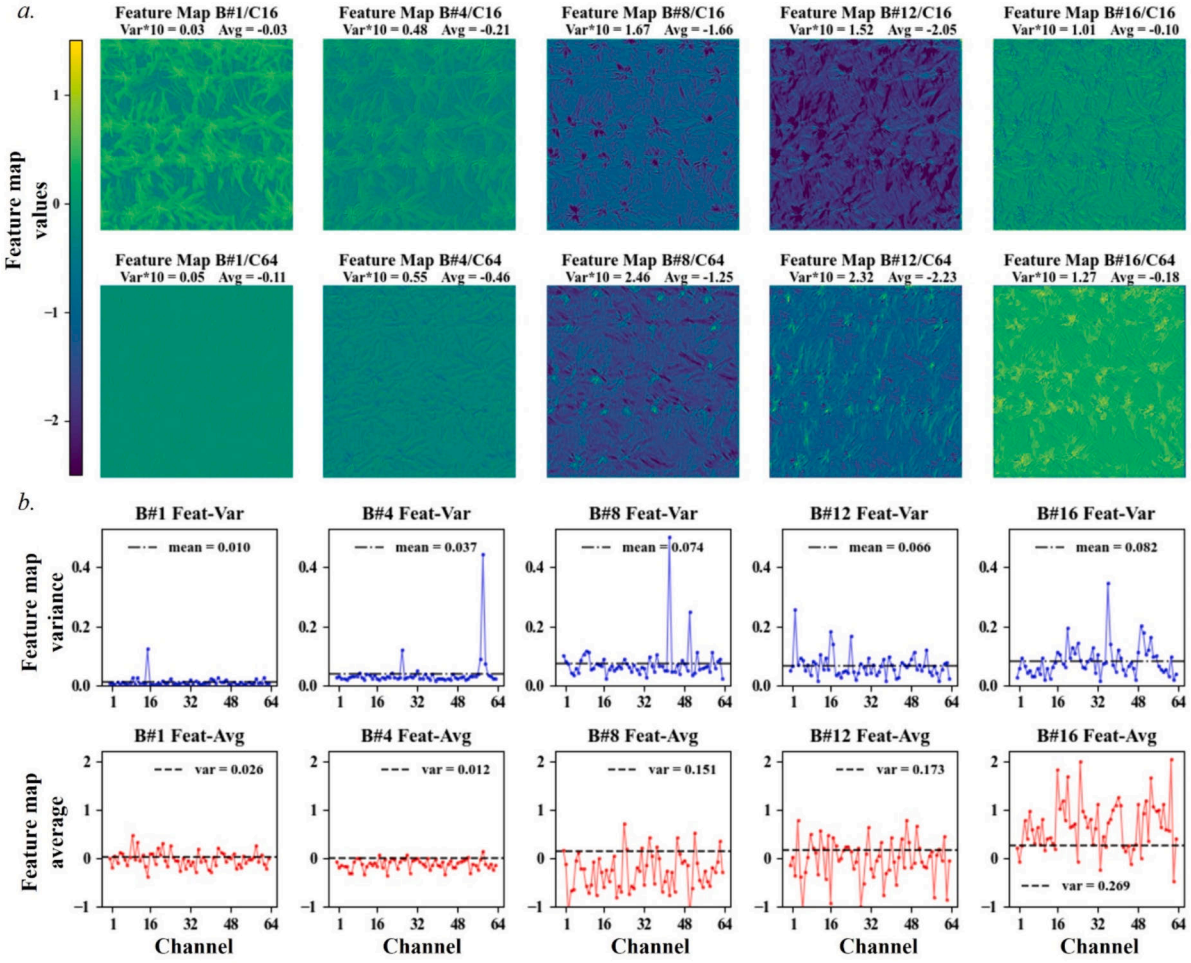

Fig. 12. Variance trends in feature space within networks. a. The feature maps of SR network blocks at different depths. b. Changes in variance and average values across different block depths, with overall feature variance gradually rising to a stable level.

across different channels. The modulated Var-fc (in cyan) values either closely follow the Var or remain at low levels near zero. In contrast, for the lower Avg2Attn graphs of Fig. 13, the learned Avg-fc (in orange) strongly modulates the Avg value, mainly by weakening with negative values, particularly in the middle block like B#12. This suggests that Var is more effective than Avg in SR networks' channel attention.

Furthermore, to understand how spatial attention works, Fig. 14 shows the spatial attention maps of various channels and depths of blocks. It can be observed that different channels exhibit varying spatial attention patterns. For instance, in B#1-C16, the background soil area is significantly activated, while in B#1-C32, attention is focused on the leaves and ears of the corn. Additionally, the features attended at

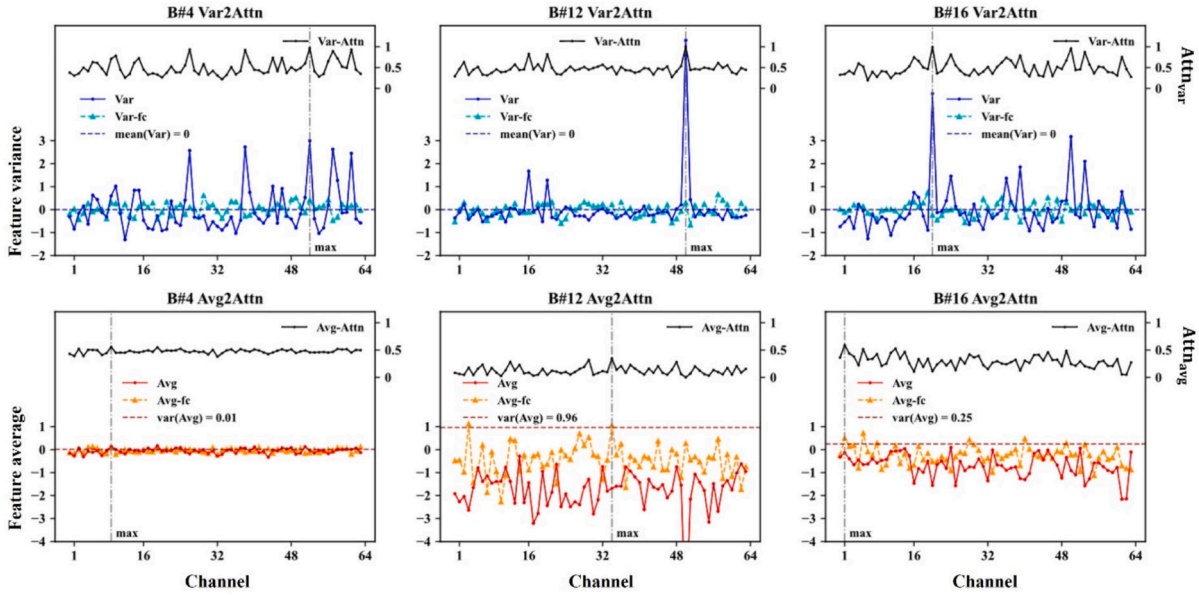

Fig. 13. From Var/Avg statistics to the corresponding attention weights. The top row shows the consistency between input variance (blue) and Var-Attn (black) across different channels at various block depths. While the bottom row shows no significant correlation between Avg and Avg-Attn. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

different depths are different.

In the shallow layers, the attention maps exhibit similarities to primary image features like HSV color extraction (B#1-C16) or edge contours (B#4-C16), indicating that the shallow layers sketch the basic properties and outlines of the objects. As the layer deepens, the spatial attention maps do not simply capture edge features. Instead, ghosting patterns emerge around specific regions or targets, such as the local features of the corn ears in B#8-C16, leaves in B#12-C32, and phantom contours in B#16–32. These attention maps within the network reflect the learned focus on specific details and memories of texture.

5.4. Refinement steps analysis of diffusion model

To clarify the effect of refinement steps on SR image quality for diffusion models, Fig. 15**a** represents the reconstruction progress, and Fig. 15**b** visualizes the overall numerical indicators changing with the exponentially increasing steps. The iterative refinement process of the diffusion model is evident during the initial stages, progressing from noise to a general outline, and then to fine details. However, comparing steps 16 and 250 reveals that excessive adjustments can adversely affect the results, causing the model to generate hallucinations that deviate from reality.

When testing the quality of results from different refinement steps on the CropSR-Test dataset in Fig. 15**b**, it was observed that the x2 model reached its peak with fewer steps compared to the x4 and x8 models. This is because large-scale reconstructions are more complex than smaller ones. Therefore, it is necessary to conduct premium step testing for diffusion models of different scales. Moreover, the SRFI, which considers perceptual and structural variations, effectively captures the degradation and helps identify the optimal step, as signified by the peak of EVADM_x2.

5.5. Potential future research

Future research in remote sensing image SR presents several promising avenues for exploration. Firstly, a comprehensive analysis of the variance trends across more general scenes and satellite data is necessary. This would extend the variance-based attention to a wider range of cross-scale remote sensing tasks and validate its generality. Additionally, investigating other kinds of variance-based attention could further enhance the network structure of SR models (Behjati et al., 2023).

Diffusion models excel at generating fine details, even at large

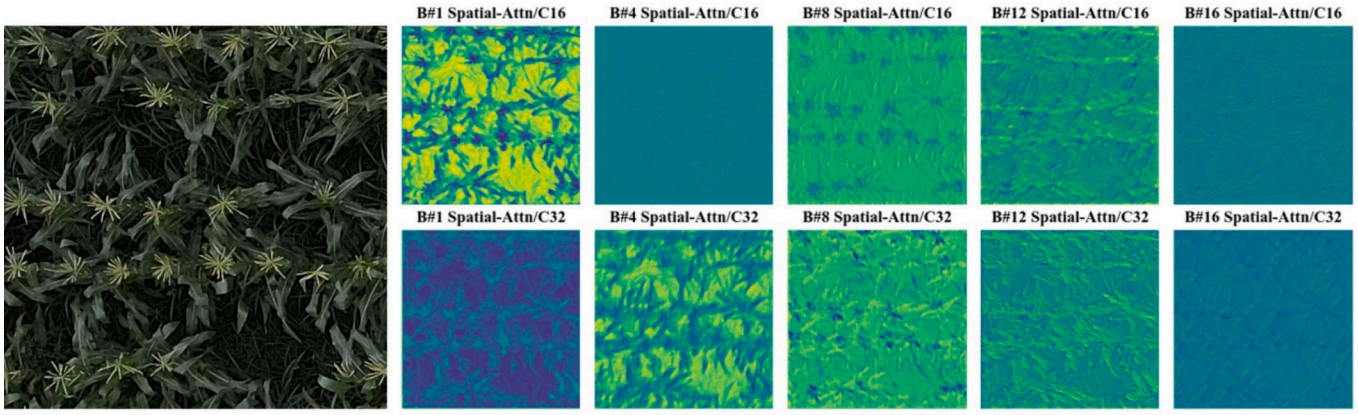

Fig. 14. Spatial attention maps at different channels and depths of VASA blocks. As depth increases, the spatial attention maps change from coarse edge features to detailed features, with ghosting patterns around contours.

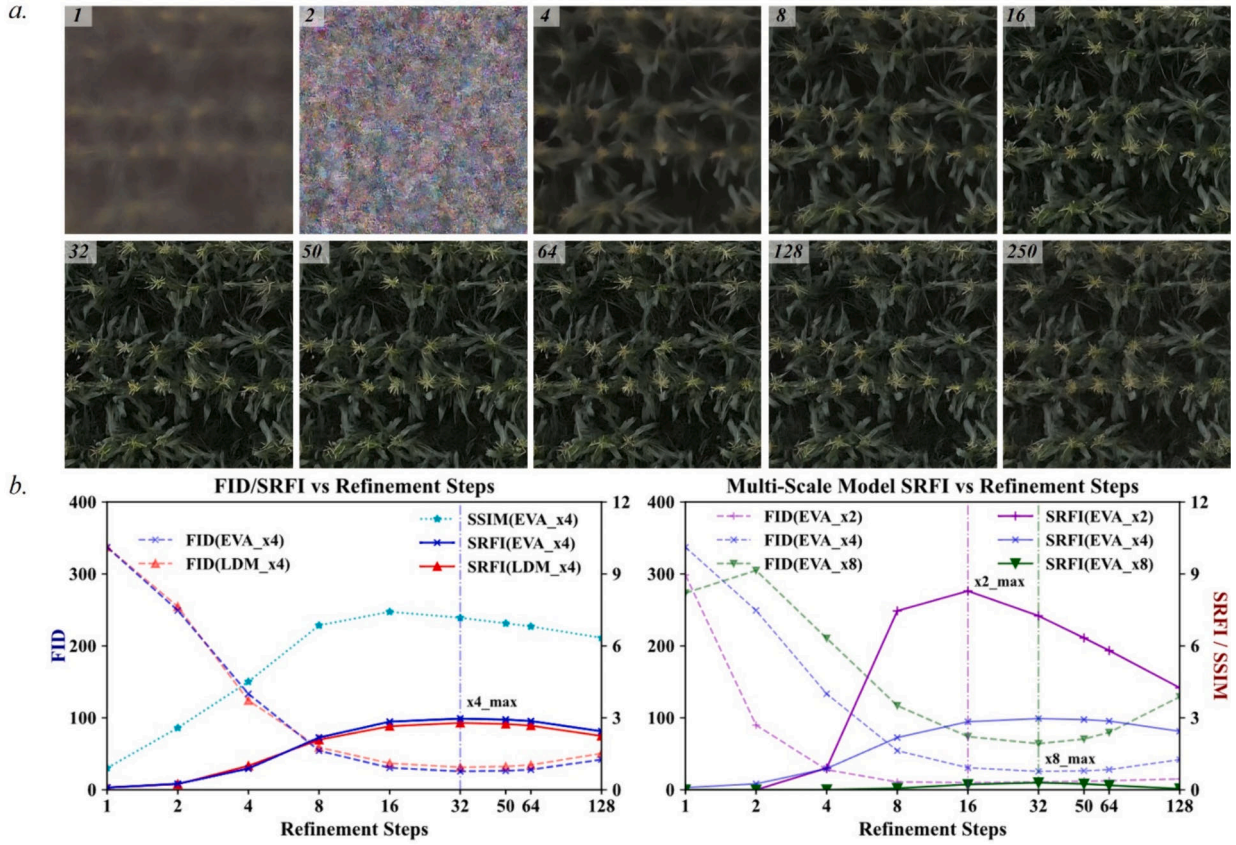

Fig. 15. Impact of refinement steps on the SR quality of diffusion models. a. SR results from EVADM_x4 at different refinement steps. b. Quality of results vs. refinement steps on the CropSR-Test dataset. (EVADM is abbreviated as EVA in the graph legend).

magnification factors, surpassing previous regression- and GAN-based models. However, in some cases, particularly at high magnification scales, the generated content may visually resemble but be inconsistent with the ground truth, such as reconstructing tree features as crops. This limitation stems from the inherently ill-posed nature of the SR problem, where the model estimates the most likely scenario under given constraints. It may be alleviated by incorporating spatial feature conditioning control (Zhang et al., 2023), using multimodal information, or integrating cognitive and semantic priors (Sun et al., 2023).

Despite the high-quality results from DM, the multistep refinement increases the computational time during inference (Cao et al., 2024). The acceleration of DM has garnered widespread attention and research. Apart from the perceptual compression in latent space encoding (LDM), reducing the refinement steps during the inference process is also crucial for accelerating DMs. Recent research introduced consistency models that decrease the number of refinement steps (Song et al., 2023). In addition, network distillation and light-weighting of the first-stage VQGAN could further enhance the efficiency and speed.

Beyond agriculture, the effectiveness of SR on aerial imagery could extend to other scenarios, such as urban land use (Hartling et al., 2021), tree breeding (D'Odorico et al., 2020), and ecology sensing (Wong et al., 2022). It is also worth exploring how SR techniques could bridge the gap between the large-area coverage of satellite imagery and the fine-grained details of aerial imagery. This would provide a clearer view of earth observation, aiding various downstream applications like disaster response, environmental monitoring, and urban planning.

6. Conclusion

In this work, we focus on enhancing the resolution and quality of aerial images of croplands using the image SR method. By observing the decreasing trends of image variance with flight height rising, we designed the Variance-Average Spatial Attention. This attention mechanism significantly boosted regression, GAN, and diffusion-based SR models. The novel cross-scale metric, SRFI, integrating structural and perceptual similarity, provided robust evaluation even on real-world SR datasets with imperfect matching. Across synthetic and real SR datasets, our EVADM model exhibited superior performance, achieving a 14.6 and 8.0 reduction in FID distance and 21 % and 6 % SRFI gains for $\times$ 2 and $\times$ 4 SR ratios, respectively. Generalization studies and SR downstream case studies also demonstrated the efficacy of EVADM compared to existing models.

Through extensive ablation studies, we found that variance attention within SR networks surpasses average attention in effectiveness. Unlike image degradation, SR processes elevate variance, especially in shallow layers. Intermediate feature and attention maps revealed that early blocks sketch rough contours while deeper blocks capture finer local details. Additionally, our research confirmed that diffusion models require more refinement steps for high-ratio SR tasks compared to low-ratio ones. Future research in aerial imagery SR should continue advancing and speeding up diffusion models while exploring broader applications of variance-based attention in diverse scenarios.

CRediT authorship contribution statement

Xiangyu Lu: Writing – original draft, Visualization, Software, Methodology, Formal analysis, Data curation, Conceptualization. **Jianlin Zhang:** Validation, Methodology, Data curation. **Rui Yang:** Validation, Software, Investigation, Data curation. **Qina Yang:** Writing – review & editing, Validation, Investigation, Data curation. **Mengyuan Chen:** Visualization, Validation, Software, Investigation. **Hongxing Xu:** Supervision, Resources, Project administration, Data curation. **Pinjun**

Wan: Validation, Resources, Project administration, Funding acquisition. **Jiawen Guo:** Writing – review & editing, Resources, Project administration, Data curation. **Fei Liu:** Writing – review & editing, Supervision, Resources, Project administration, Methodology, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This study was supported by the National Key Research and Development Program of China (2023YFD2000203), Science and Technology Department of Zhejiang Province (2022C02034), and Science and Technology Department of Shenzhen (CJGJZD20210408092401004).

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.isprsjprs.2024.08.017.

References

Aslahishahri, M., Stanley, K.G., Duddu, H., Shirtliffe, S., Vail, S., Stavness, I., 2021. Spatial super resolution of real-world aerial images for image-based plant phenotyping. Remote Sens. (Basel) 13, 2308. https://doi.org/10.3390/rs13122308.

Behjati, P., Rodriguez, P., Fernández, C., Hupont, I., Mehri, A., Gonzàlez, J., 2023. Single image super-resolution based on directional variance attention network. Pattern Recogn. 133, 108997 https://doi.org/10.1016/j.patcog.2022.108997.

Bell-Kligler, S., Shocher, A., Irani, M., 2019. Blind Super-Resolution Kernel Estimation using an Internal-GAN, in: Advances in Neural Information Processing Systems. Curran Associates, Inc. doi: 10.48550/arXiv.1909.06581.

Blau, Y., Mechrez, R., Timofte, R., Michaeli, T., Zelnik-Manor, L., 2019. The 2018 PIRM Challenge on Perceptual Image Super-Resolution, in: Leal-Taixé, L., Roth, S. (Eds.), Computer Vision – ECCV 2018 Workshops, Lecture Notes in Computer Science. Springer International Publishing, Cham, pp. 334–355. doi: 10.1007/978-3-030-11021-5_21.

Cai, J., Zeng, H., Yong, H., Cao, Z., Zhang, L., 2019. Toward Real-World Single Image Super-Resolution: A New Benchmark and a New Model, in: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE, Seoul, Korea (South), pp. 3086–3095. doi: 10.1109/ICCV.2019.00318

Cao, H., Tan, C., Gao, Z., Xu, Y., Chen, G., Heng, P.-A., Li, S.Z., 2024. A survey on generative diffusion models. IEEE Trans. Knowl. Data Eng. 36, 2814–2830. https://doi.org/10.1109/TKDE.2024.3361474.

Chen, S., Ogawa, Y., Zhao, C., Sekimoto, Y., 2023. Large-scale individual building extraction from open-source satellite imagery via super-resolution-based instance segmentation approach. ISPRS J. Photogramm. Remote Sens. 195, 129–152. https://doi.org/10.1016/j.isprsjprs.2022.11.006.

Chiu, M.T., Xu, X., Wei, Y., Huang, Z., Schwing, A.G., Brunner, R., Khachatrian, H., Karapetyan, H., Dozier, I., Rose, G., Wilson, D., Tudor, A., Hovakimyan, N., Huang, T.S., Shi, H., 2020. Agriculture-Vision: A Large Aerial Image Database for Agricultural Pattern Analysis. Presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2828–2838.

Courtrai, L., Pham, M.-T., Lefèvre, S., 2020. Small object detection in remote sensing images based on super-resolution with auxiliary generative adversarial networks. Remote Sens. (Basel) 12, 3152. https://doi.org/10.3390/rs12193152.

D'Odorico, P., Besik, A., Wong, C.Y.S., Isabel, N., Ensminger, I., 2020. High-throughput drone-based remote sensing reliably tracks phenology in thousands of conifer seedlings. New Phytol. 226, 1667–1681. https://doi.org/10.1111/nph.16488.

Dong, C., Loy, C.C., He, K., Tang, X., 2016. Image super-resolution using deep convolutional networks. IEEE Trans. Pattern Anal. Mach. Intell. 38, 295–307. https://doi.org/10.1109/TPAMI.2015.2439281.

Dong, R., Mou, L., Zhang, L., Fu, H., Zhu, X.X., 2022. Real-world remote sensing image super-resolution via a practical degradation model and a kernel-aware network. ISPRS J. Photogramm. Remote Sens. 191, 155–170. https://doi.org/10.1016/j.isprsjprs.2022.07.010.

Esser, P., Rombach, R., Ommer, B., 2021. Taming Transformers for High-Resolution Image Synthesis, in: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, Nashville, TN, USA, pp. 12868–12878. doi: 10.1109/CVPR46437.2021.01268

Feng, J., Jiang, Q., Tseng, C.-H., Jin, X., Liu, L., Zhou, W., Yao, S., 2022. A deep multitask convolutional neural network for remote sensing image super-resolution and colorization. IEEE Trans. Geosci. Remote Sens. 60, 1–15. https://doi.org/10.1109/TGRS.2022.3154435.

Guo, Y., Chen, J., Wang, J., Chen, Q., Cao, J., Deng, Z., Xu, Y., Tan, M., 2020. Closed-Loop Matters: Dual Regression Networks for Single Image Super-Resolution, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5406–5415. doi: 10.1109/CVPR42600.2020.00545.

Hartling, S., Sagan, V., Maimaitijiang, M., 2021. Urban tree species classification using UAV-based multi-sensor data fusion and machine learning. Gisci. Remote Sensing 58, 1250–1275. https://doi.org/10.1080/15481603.2021.1974275.

He, J., Yuan, Q., Li, J., Xiao, Y., Zhang, L., 2023. A self-supervised remote sensing image fusion framework with dual-stage self-learning and spectral super-resolution injection. ISPRS J. Photogramm. Remote Sens. 204, 131–144. https://doi.org/10.1016/j.isprsjprs.2023.09.003.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S., 2017. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium, in: Advances in Neural Information Processing Systems. Curran Associates, Inc. doi: 10.48550/arXiv.1706.08500.

Ho, J., Jain, A., Abbeel, P., 2020. Denoising Diffusion Probabilistic Models, in: Advances in Neural Information Processing Systems. Curran Associates, Inc., pp. 6840–6851. doi: 10.48550/arXiv.2006.11239.

Ho, J., Saharia, C., Chan, W., Fleet, D.J., Norouzi, M., Salimans, T., 2022. Cascaded diffusion models for high fidelity image generation. J. Mach. Learn. Res. 23, 1–33.

Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H., 2017. MobileNets: efficient convolutional neural networks for mobile vision applications. CoRR abs/1704.04861.

Hu, P., Guo, W., Chapman, S.C., Guo, Y., Zheng, B., 2019. Pixel size of aerial imagery constrains the applications of unmanned aerial vehicle in crop breeding. ISPRS J. Photogramm. Remote Sens. 154, 1–9. https://doi.org/10.1016/j.isprsjprs.2019.05.008.

Inzerillo, L., Acuto, F., Di Mino, G., Uddin, M.Z., 2022. Super-resolution images methodology applied to UAV datasets to road pavement monitoring. Drones 6, 171. https://doi.org/10.3390/drones6070171.

Ji, X., Cao, Y., Tai, Y., Wang, C., Li, J., Huang, F., 2020. Real-World Super-Resolution via Kernel Estimation and Noise Injection, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 1914–1923. doi: 10.1109/CVPRW50498.2020.00241.

Jiang, Q., Li, F., Ren, T., Liu, S., Zeng, Z., Yu, K., Zhang, L., 2023. T-Rex: Counting by Visual Prompting. arXiv preprint. doi: 10.48550/arXiv.2311.13596.

Khan, M.A., Menouar, H., Hamila, R., 2023. Revisiting crowd counting: state-of-the-art, trends, and future perspectives. Image Vis. Comput. 129, 104597 https://doi.org/10.1016/j.imavis.2022.104597.

Kim, J.-H., Choi, J.-H., Cheon, M., Lee, J.-S., 2020. MAMNet: Multi-path adaptive modulation network for image super-resolution. Neurocomputing 402, 38–49. https://doi.org/10.1016/j.neucom.2020.03.069.

Kong, J., Ryu, Y., Jeong, S., Zhong, Z., Choi, W., Kim, J., Lee, K., Lim, J., Jang, K., Chun, J., Kim, K.-M., Houborg, R., 2023. Super resolution of historic Landsat imagery using a dual generative adversarial network (GAN) model with CubeSat constellation imagery for spatially enhanced long-term vegetation monitoring. ISPRS J. Photogramm. Remote Sens. 200, 1–23. https://doi.org/10.1016/j.isprsjprs.2023.04.013.

Ledig, C., Theis, L., Huszar, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., Shi, W., 2017. Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 105–114. doi: 10.1109/CVPR.2017.19

Li, Y., Du, Z., Wu, S., Wang, Y., Wang, Z., Zhao, X., Zhang, F., 2021. Progressive split-merge super resolution for hyperspectral imagery with group attention and gradient guidance. ISPRS J. Photogramm. Remote Sens. 182, 14–36. https://doi.org/10.1016/j.isprsjprs.2021.09.023.

Li, Y., Wang, H., Jin, Q., Hu, J., Chemerys, P., Fu, Y., Wang, Y., Tulyakov, S., Ren, J., 2023. SnapFusion: Text-to-Image Diffusion Model on Mobile Devices within Two Seconds. doi: 10.48550/arXiv.2306.00980.

Li, H., Yang, Y., Chang, M., Chen, S., Feng, H., Xu, Z., Li, Q., Chen, Y., 2022. SRDiff: Single image super-resolution with diffusion probabilistic models. Neurocomputing 479, 47–59. https://doi.org/10.1016/j.neucom.2022.01.029.

Lim, B., Son, S., Kim, H., Nah, S., Lee, K.M., 2017. Enhanced deep residual networks for single image super-resolution. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). IEEE, pp. 1132–1140. https://doi.org/10.1109/CVPRW.2017.151.

Lu, X., Yang, R., Zhou, J., Jiao, J., Liu, F., Liu, Y., Su, B., Gu, P., 2022. A hybrid model of ghost-convolution enlightened transformer for effective diagnosis of grape leaf disease and pest. Journal of King Saud University - Computer and Information Sciences 34, 1755–1767. https://doi.org/10.1016/j.jksuci.2022.03.006.

Lu, X., Zhou, J., Yang, R., Yan, Z., Lin, Y., Jiao, J., Liu, F., 2023. Automated rice phenology stage mapping using UAV images and deep learning. Drones 7, 83. https://doi.org/10.3390/drones7020083.

Mao, P., Ding, J., Jiang, B., Qin, L., Qiu, G.Y., 2022. How can UAV bridge the gap between ground and satellite observations for quantifying the biomass of desert shrub community? ISPRS J. Photogramm. Remote Sens. 192, 361–376. https://doi.org/10.1016/j.isprsjprs.2022.08.021.

Mao, P., Jiang, B., Shi, Z., He, Y., Shen, T., Qiu, G.Y., 2023. Effects of UAV flight height on biomass estimation of desert shrub communities. Ecol. Ind. 154, 110698 https://doi.org/10.1016/j.ecolind.2023.110698.

Mittal, A., Soundararajan, R., Bovik, A.C., 2013. Making a "Completely Blind" Image Quality Analyzer. IEEE Signal Process Lett. 20, 209–212. https://doi.org/10.1109/LSP.2012.2227726.

MMEditing, 2022. MMEditing: OpenMMLab Image and Video Editing Toolbox [WWW Document]. URL https://github.com/open-mmlab/mmagic/tree/0.x (accessed 11.30.23).

Pashaei, M., Starek, M.J., Kamangir, H., Berryhill, J., 2020. Deep learning-based single image super-resolution: an investigation for dense scene reconstruction with UAS photogrammetry. Remote Sens. (Basel) 12, 1757. https://doi.org/10.3390/rs12111757.

Qiu, Z., Shen, H., Yue, L., Zheng, G., 2023. Cross-sensor remote sensing imagery super-resolution via an edge-guided attention-based network. ISPRS J. Photogram. Remote Sens. 199, 226–241. https://doi.org/10.1016/j.isprsjprs.2023.04.016.

Razzak, M.T., Mateo-García, G., Lecuyer, G., Gómez-Chova, L., Gal, Y., Kalaitzis, F., 2023. Multi-spectral multi-image super-resolution of Sentinel-2 with radiometric consistency losses and its effect on building delineation. ISPRS J. Photogramm. Remote Sens. 195, 1–13. https://doi.org/10.1016/j.isprsjprs.2022.10.019.

Arefin, R.M., Michalski, V., St-Charles, P.-L., Kalaitzis, A., Kim, S., Kahou, S.E., Bengio, Y., 2020. Multi-Image Super-Resolution for Remote Sensing using Deep Recurrent Networks, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). IEEE, Seattle, WA, USA, pp. 816–825. doi: 10.1109/CVPRW50498.2020.00111.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B., 2022. High-Resolution Image Synthesis with Latent Diffusion Models, in: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, New Orleans, LA, USA, pp. 10674–10685. doi: 10.1109/CVPR52688.2022.01042.

Sagan, V., Maimaitijiang, M., Sidike, P., Eblimit, K., Peterson, K.T., Hartling, S., Esposito, F., Khanal, K., Newcomb, M., Pauli, D., Ward, R., Fritschi, F., Shakoor, N., Mockler, T., 2019. UAV-based high resolution thermal imaging for vegetation monitoring, and plant phenotyping using ICI 8640 P, FLIR Vue Pro R 640, and thermoMap cameras. Remote Sens. (Basel) 11, 330. https://doi.org/10.3390/rs11030330.

Sagan, V., Maimaitijiang, M., Paheding, S., Bhadra, S., Gosselin, N., Burnette, M., Demieville, J., Hartling, S., LeBauer, D., Newcomb, M., Pauli, D., Peterson, K.T., Shakoor, N., Stylianou, A., Zender, C.S., Mockler, T.C., 2022. Data-driven artificial intelligence for calibration of hyperspectral big data. IEEE Trans. Geosci. Remote Sens. 60, 1–20. https://doi.org/10.1109/TGRS.2021.3091409.

Sahak, H., Watson, D., Saharia, C., Fleet, D., 2023. Denoising Diffusion Probabilistic Models for Robust Image Super-Resolution in the Wild. doi: 10.48550/arXiv.2302.07864.

Saharia, C., Chan, W., Chang, H., Lee, C.A., Ho, J., Salimans, T., Fleet, D.J., Norouzi, M., 2022a. Palette: Image-to-Image Diffusion Models, in: ACM SIGGRAPH 2022 Conference Proceedings. Presented at the SIGGRAPH '22, Association for Computing Machinery, New York, NY, USA, p. 10. doi: 10.48550/arXiv.2111.05826.

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S.K.S., Ayan, B.K., Mahdavi, S.S., Lopes, R.G., Salimans, T., Ho, J., Fleet, D.J., Norouzi, M., 2022b. photorealistic text-to-image diffusion models with deep language understanding, in: Advances in Neural Information Processing Systems. NeurIPS, 2022, pp. 36479–36494. doi: 10.48550/arXiv.2205.11487.

Saharia, C., Ho, J., Chan, W., Salimans, T., Fleet, D.J., Norouzi, M., 2022c. Image super-resolution via iterative refinement. IEEE Trans. Pattern Anal. Mach. Intell. 45, 4713–4726. https://doi.org/10.1109/TPAMI.2022.3204461.

Sajjadi, M.S.M., Scholkopf, B., Hirsch, M., 2017. EnhanceNet: Single Image Super-Resolution Through Automated Texture Synthesis. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4491–4500. https://doi.org/10.1109/ICCV.2017.481.

Salimans, T., Ho, J., 2021. Progressive Distillation for Fast Sampling of Diffusion Models. In: International Conference on Learning Representations. https://doi.org/10.48550/arXiv.2202.00512.

Shi, W., Caballero, J., Huszar, F., Totz, J., Aitken, A.P., Bishop, R., Rueckert, D., Wang, Z., 2016. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, Las Vegas, pp. 1874–1883. https://doi.org/10.1109/CVPR.2016.207.

Shi, Y., Han, L., Han, L., Chang, S., Hu, T., Dancey, D., 2022. A latent encoder coupled generative adversarial network (LE-GAN) for efficient hyperspectral image super-resolution. IEEE Trans. Geosci. Remote Sens. 60, 1–19. https://doi.org/10.1109/TGRS.2022.3193441.

Sidike, P., Sagan, V., Qumsiyeh, M., Maimaitijiang, M., Essa, A., Asari, V., 2018. Adaptive trigonometric transformation function with image contrast and color enhancement: application to unmanned aerial system imagery. IEEE Geosci. Remote Sens. Lett. 15, 404–408. https://doi.org/10.1109/LGRS.2018.2790899.

Song, Y., Dhariwal, P., Chen, M., Sutskever, I., 2023. Consistency Models. In: Proceedings of the 40th International Conference on Machine Learning, ICML'23. Presented at the International Conference on Machine Learning, PMLR, Honolulu, Hawaii, USA. p. 32211–32252. doi: 10.48550/arXiv.2303.01469.

Song, J., Meng, C., Ermon, S., 2021. Denoising Diffusion Implicit Models. In: International Conference on Learning Representations. https://doi.org/10.48550/arXiv.2010.02502.

Sun, H., Li, W., Liu, J., Chen, H., Pei, R., Zou, X., Yan, Y., Yang, Y., 2023. CoSeR: Bridging Image and Language for Cognitive Super-Resolution. arXiv preprint. doi: 10.48550/arXiv.2311.16512.

Thanh-Tung, H., Tran, T., 2020. Catastrophic forgetting and mode collapse in GANs. In: 2020 International Joint Conference on Neural Networks (IJCNN). Presented at the 2020 International Joint Conference on Neural Networks (IJCNN). pp. 1–10. doi: 10.1109/IJCNN48605.2020.9207181.

Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P., 2004. Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. 13, 600–612. https://doi.org/10.1109/TIP.2003.819861.

Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., Qiao, Y., Change Loy, C., 2018. ESRGAN: Enhanced Super-Resolution Generative Adversarial Networks, in: Proceedings of the European Conference on Computer Vision (ECCV) Workshops. pp. 0–0. doi: 10.48550/arXiv.1809.00219.

Wang, X., Xie, L., Dong, C., Shan, Y., 2021b. Real-ESRGAN: Training real-world blind super-resolution with pure synthetic data. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1905–1914. doi: 10.1109/ICCVW54120.2021.00217.

Wang, S., Saharia, C., Montgomery, C., Pont-Tuset, J., Noy, S., Pellegrini, S., Onoe, Y., Laszlo, S., Fleet, D.J., Soricut, R., Baldridge, J., Norouzi, M., Anderson, P., Chan, W., 2023. Imagen Editor and EditBench: Advancing and Evaluating Text-Guided Image Inpainting. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, Vancouver, BC, Canad. pp. 18359–18369. doi: 10.1109/CVPR52729.2023.01761.

Wang, J., Gao, K., Zhang, Z., Ni, C., Hu, Z., Chen, D., Wu, Q., 2021a. Multisensor Remote sensing imagery super-resolution with conditional GAN. J. Remote Sensing 2021. https://doi.org/10.34133/2021/9829706.

Wells, K., Lopes, F.A., Sagan, V., Esposito, F., 2023. A multifaceted benchmarking of GAN architectures on generating synthetic satellite imagery. In: 2023 IEEE Applied Imagery Pattern Recognition Workshop (AIPR). Presented at the 2023 IEEE Applied Imagery Pattern Recognition Workshop (AIPR), IEEE, St. Louis, MO, USA, pp. 1–7. doi: 10.1109/AIPR60534.2023.10440718.

Wong, C.Y.S., Mercado, L.M., Arain, M.A., Ensminger, I., 2022. Remotely sensed carotenoid dynamics improve modelling photosynthetic phenology in conifer and deciduous forests. Agric. For. Meteorol. 321, 108977 https://doi.org/10.1016/j.agrformet.2022.108977.

Woo, S., Park, J., Lee, J.-Y., Kweon, I.S., 2018. CBAM: Convolutional Block Attention Module. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 3–19. doi: 10.1007/978-3-030-01234-2_1.

Yun, C., Kim, Y.H., Lee, S.J., Im, S.J., Park, K.R., 2023. WRA-net: wide receptive field attention network for motion deblurring in crop and weed image. Plant Phenomics 5, 0031. https://doi.org/10.34133/plantphenomics.0031.

Zhang, S., Cheng, Y., Luo, D., He, J., Wong, A.K.Y., Hung, K., 2021b. Channel attention convolutional neural network for Chinese Baijiu detection with E-nose. IEEE Sens. J. 21, 16170–16182. https://doi.org/10.1109/JSEN.2021.3075703.

Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O., 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 586–595. doi: 10.1109/CVPR.2018.00068.

Zhang, K., Liang, J., Van Gool, L., Timofte, R., 2021. Designing a Practical Degradation Model for Deep Blind Image Super-Resolution, in: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4791–4800. doi: 10.1109/ICCV48922.2021.00475.

Zhang, L., Rao, A., Agrawala, M., 2023. Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. arXiv, pp. 3836–3847. doi: 10.48550/ARXIV.2302.05543.

Zhang, J., Wang, X., Liu, J., Zhang, D., Lu, Y., Zhou, Y., Sun, L., Hou, S., Fan, X., Shen, S., Zhao, J., 2022. Multispectral drone imagery and SRGAN for rapid phenotypic mapping of individual Chinese cabbage plants. Plant Phenomics 2022, 0007. https://doi.org/10.34133/plantphenomics.0007.

Zhong, Z., Zhu, J., Dai, Y., Zheng, C., Huo, Y., Chen, G., Bao, H., Wang, R., 2023. FuseSR: Super resolution for real-time rendering through efficient multi-resolution fusion, in: SIGGRAPH Asia 2023 Conference Papers. ACM. doi: 10.1145/3610548.3618209.

Zhou, J., Vong, C.-M., Liu, Q., Wang, Z., 2019. Scale adaptive image cropping for UAV object detection. Neurocomputing 366, 305–313. https://doi.org/10.1016/j.neucom.2019.07.073.