LSwinSR: UAV Imagery Super-Resolution based on Linear Swin Transformer

Rui Li and Xiaowei Zhao

***Abstract*—Super-resolution, which aims to reconstruct high-resolution images from low-resolution images, has drawn considerable attention and has been intensively studied in computer vision and remote sensing communities. Super-resolution technology is especially beneficial for unmanned aerial vehicles (UAV), as the number and resolution of images captured by UAV are highly limited by physical constraints such as flight altitude and load capacity. In the wake of the successful application of deep learning methods in the super-resolution task, in recent years, a series of super-resolution algorithms have been developed. In this paper, for the super-resolution of UAV images, a novel network based on the state-of-the-art Swin Transformer is proposed with better efficiency and competitive accuracy. Meanwhile, as one of the essential applications of the UAV is land cover and land use monitoring, simple image quality assessments such as the Peak-Signal-to-Noise Ratio (PSNR) and the Structural Similarity Index Measure (SSIM) are not enough to comprehensively measure the performance of an algorithm. Therefore, we further investigate the effectiveness of super-resolution methods using the accuracy of semantic segmentation. The code is available at https://github.com/lironui/GeoSR.**

***Index Terms*—Super-resolution, Transformer, Semantic segmentation, Deep learning, UAV**

I. Introduction

WITH the continuous development in both hardware reliability and control strategy, Unmanned Aerial Vehicles (UAV) have been widely employed in more and more practical applications [1], such as atmosphere monitoring [2], tracking and surveillance [3] and land cover and land use monitoring [4]. However, limited by the physical characteristics of the UAV such as the flight altitude and load capacity, sometimes it is impractical to continuously obtain High-Resolution (HR) images through UAV, especially for those applications that need large-scale and long-duration UAV video data [5]. Super-resolution is a promising solution to alleviate this dilemma which can reconstruct high-resolution images from Low-Resolution (LR) observations [6]. Actually, the super-resolution technology particularly the deep-learning-based methods has already been widely applied for the processing of various kinds of images, such as natural images [7], [8], medical images [9], [10] and remote sensing images [11], [12], [13]. These works can not only enhance the image quality but also can further facilitate downstream applications [14], [15].

This work has received funding from the UK Engineering and Physical Sciences Research Council (grant number: EP/Y016297/1). The authors acknowledge the support of the Scientific Computing Research Technology Platform (SCRTP) at the University of Warwick for providing High-Performance Computing resources. (Corresponding author: Xiaowei Zhao.)

R. Li and X. Zhao are with Intelligent Control & Smart Energy (ICSE) Research Group, School of Engineering, University of Warwick, Coventry, CV4 7AL, UK. (e-mail: rui.li.8@warwick.ac.uk; xiaowei.zhao@warwick.ac.uk).

Among those revolutionary deep-learning-based super-resolution algorithms, most of them are constructed by the Convolutional Neural Network (CNN) [6], [16], [17]. However, as pointed out by [18], the utilization of CNN for super-resolution has two obvious drawbacks. First, as interactions between images and kernels are content-independent, using the same kernel to restore different image regions may not be the optimal solution. Second, as CNN is initially designed to focus on extracting local patterns, it lacks the capability for capturing long-range and non-local dependencies. By contrast, Transformer [19], a promising alternative to CNN, adopts the self-attention mechanism to capture global interactions between contexts, which has shown its great potential in boosting vision-related tasks [20], [21], [22], [23], [24].

For super-resolution, the interactions between images and self-attention blocks in the Vision Transformer (ViT) are content-adaptive as the attention weights are generated according to the relationship between contexts. Meanwhile, the shifted window mechanism embedded within the Transformer enables long-range dependency modelling. However, the utilization of the self-attention mechanism means significant memory and computational costs, which increases quadratically with the size of the input, i.e. $O(N^2)$ complexity [25]. Even though the shifted window attention operation [21], [23] can partly alleviate the massive memory and computational requirements by only applying the attention operation on a small local window, the quadratic complexity of the self-attention mechanism itself is still a problematic issue of concern, which is especially true when we need a large local window.

Meanwhile, for most super-resolution studies, the performance of the algorithm is only evaluated by the image quality assessment metrics such as the Peak-Signal-to-Noise Ratio (PSNR) and the Structural Similarity Index Measure (SSIM). The image quality assessment metrics may be enough and appropriate for entertaining applications such as enhancing the resolution of antique digital photos. However, when it comes to task-oriented applications such as UAV-based land cover and land use monitoring, those metrics which only reflect the image quality are far from the practical requirements.

In this paper, based on our previous work on linearizing the complexity of the self-attention mechanism [25], [26], [27], we propose a novel shifted window attention with linear complexity by employing the kernel attention mechanism [25], thereby designing the Linear Swin Transformer for UAV Super-Resolution (LSwinSR). The experimental results show that the inference speed of the proposed LSwinSR is faster than the Swin Transformer for Image Restoration (SwinIR) [28] but with competitive accuracy. Meanwhile, we compare

NOMENCLATURE

Abbreviations

Chat-GPT	Chat Generative Pre-trained Transformer
CNN	Convolutional Neural Network
EDSR	Enhanced Deep Super-Resolution Network
EDiffSR	Efficient Diffusion Probabilistic Model for Super-Resolution
FPS	Frames Per Second
GELU	Gaussian Error Linear Unit
HR	High-Resolution
LN	LayerNorm
LP-KPN	Laplacian Pyramid based Kernel Prediction Network
LR	Low-Resolution
LSTL	Linear Swin Transformer Layers
LSwinSR	Linear Swin Transformer for Super-Resolution
KA	Kernel Attention
MACs	Multiply-Accumulate Operations
MHKA	Multi-Head Kernel-Attention
MHSA	Multi-Head Self-Attention
NLP	Natural Language Processing
NLSA	Non-Local Sparse Attention
MLP	Multi-Layer Perceptron
PSNR	Peak-Signal-to-Noise Ratio
RLSTB	Residual Linear Swin Transformer Blocks
SR	Super-Resolution
SRCNN	Super-Resolution Convolutional Neural Network
SRGAN	Super-Resolution Generative Adversarial Network
SRResNet	Super-Resolution Residual Network
SSIM	Structural Similarity Index Measure
SwinIR	Swin Transformer for Image Restoration
TTST	Top-k Token Selective Transformer
UAV	Unmanned Aerial Vehicles
UNetFormer	UNet-like Transformer
VDSR	Very Deep Super Resolution network
ViT	Vision Transformer
W-MHSA	Window-based Multi-Head Self-Attention module

Symbols

$\boldsymbol{B}$	The relative position bias term
$\mathscr{C}_{DF}$	The convolutional layer in the deep feature extraction module
$\mathscr{C}_{RLSTB}$	The convolutional layer in the residual linear Swin Transformer block
$\mathscr{C}_{SF}$	The convolutional layer in the shallow feature extraction module
$\mathscr{D}$	The degradation mapping function
E	The expected loss
$\mathscr{F}$	The super-resolution model
F_{DF}	The deep features
F_{SF}	The shallow features
M	The side length of a local window
I	The number of image pairs
$\boldsymbol{I}_{LR}$	The low-resolution image
$\boldsymbol{I}_{HR}$	The high-resolution image
$\boldsymbol{I}_{SR}$	The super-resolution image
L	The loss function
$\mathscr{L}$	The linear Swin Transformer layer
$\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}$	The query, key and value matrices
$\mathscr{R}$	The residual linear Swin Transformer block
$\mathscr{R}e$	The reconstruction module
Φ	the regularization term
λ	the tradeoff parameter of $\Phi(\cdot)$
θ	the parameters of the super-resolution model
δ	the parameters of the degradation process
ρ	The normalization function

and evaluate the usability and dependability of the results generated by different super-resolution methods based on semantic segmentation accuracy, which further demonstrates the effectiveness of the proposed LSwinSR.

The remaining part of this paper is organized as follows: the related works are reviewed in Section II. Then, the methodology is described in Section III. Thereafter, the dataset, experimental setting and experimental result are reported and analyzed in Section IV. Finally, the conclusions are drawn in Section V.

II. RELATED WORK

A. Problem Definition

The super-resolution task aims at reconstructing the high-resolution images from the corresponding low-resolution inputs. Generally, the low-resolution image $\boldsymbol{I}_{LR}$ can be modelled as the output of the degradation process:

$$\boldsymbol{I}_{LR} = \mathscr{D}(\boldsymbol{I}_{HR}; \delta) \tag{1}$$

where $\mathscr{D}$ denotes a degradation mapping function, $\boldsymbol{I}_{HR}$ represents the corresponding high-resolution image and δ means the parameters of the degradation process. For real-world applications, the degradation process is normally unknown, while the target of the super-resolution model is to reverse the degradation thereby reconstructing the high-resolution image:

$$\boldsymbol{I}_{SR} = \mathscr{F}(\boldsymbol{I}_{LR}; \theta) \approx \mathscr{D}^{-1}(\boldsymbol{I}_{LR}; \theta) \tag{2}$$

where $\boldsymbol{I}_{SR}$ is the image generated by the super-resolution model $\mathscr{F}$ and θ indicates the parameters of the model $\mathscr{F}$. Hence, a super-resolution model $\mathscr{F}$ is trained to narrow the gap between the approximation $\boldsymbol{I}_{SR}$ and the high-resolution reference $\boldsymbol{I}_{HR}$ as closely as possible by optimizing the parameters θ:

$$\begin{aligned} \theta^* &= \arg\min_{\theta} E(\theta) \\ E(\theta) &= \sum\nolimits_{i=1}^{I} L(\boldsymbol{I}_{HR}, \boldsymbol{I}_{SR}) + \lambda\Phi(\theta) \\ L(\boldsymbol{I}_{HR}, \boldsymbol{I}_{SR}) &= L(\boldsymbol{I}_{HR}, \mathscr{F}(\boldsymbol{I}_{LR}; \theta)) \end{aligned} \tag{3}$$

where $E(\theta)$ means the expected loss, the loss function $L(\boldsymbol{I}_{HR}, \boldsymbol{I}_{SR})$ measures the disparity between the high-resolution references and the predicted results, $\Phi(\theta)$ is the regularization term weighted by the tradeoff parameter λ, and I represents the number of image pairs.

B. Super-Resolution

In recent years, benefitting from the significant advances of deep learning, a series of revolutionary methods have been proposed and verified for the super-resolution task. As a pioneering work, a Super-Resolution Convolutional Neural Network (SRCNN) was proposed by Dong et al. [6], demonstrating that the traditional sparse-coding-based super-resolution methods can be reformulated into a deep convolutional neural network. In the proposed pipeline, the low-resolution images were

upsampled to the same size as the high-resolution references using the Bicubic interpolation whereafter the upsampled images were taken as the input of the network. With only three convolutional layers, the performance of SRCNN outperformed the Bicubic interpolation method by a large margin. However, the over-simplified structure of SRCNN severely limited its full potential for more complicated scenarios. To address this problem, a Very Deep Super Resolution (VDSR) [29] structure was developed with 20 convolutional layers to learn the deep features of the images. Meanwhile, Dong et al. [30] also further optimized and accelerated their SRCNN by deepening the layer of the network, taking the original low-resolution image as input and adding the deconvolution layer at the end to enlarge the feature map. As the deconvolution operations were prone to checkerboard artefacts, EnhanceNet [31] alleviated this problem by replacing the deconvolution layer with the subpixel convolution layer. Besides, the residual block [32] was also introduced to design super-resolution networks with deeper or wider structures. For example, a Super-Resolution Generative Adversarial Network (SRGAN) was proposed in [33] based on the designed 16 blocks deep Super-resolution ResNet (SRResNet). The SRResNet was further improved by the Enhanced Deep Super-Resolution (EDSR) network [34], which removed the batch normalization layers as they would get rid of the range flexibility from the network. Moreover, several studies attempted to combine the advantages of both deep learning and traditional image processing technologies. For example, the Laplacian Pyramid based Kernel Prediction Network (LP-KPN) [35] applied the Laplacian pyramid technology to enlarge the receptive field and reduce the computational cost, while the Non-Local Sparse Attention (NLSA) proposed by [36] embraced the benefits of sparse representation and non-local operation. In recent years, the diffusion probabilistic model [37] has also been introduced to super-resolution and achieved promising progress, such as EDiffSR [38] and Resdiff [39].

C. Vision Transformer

Originally, the Transformer [19] was designed for Natural Language Processing (NLP) applications, which has achieved state-of-the-art performance and become the de-facto standard solution for many NLP tasks [20], such as the cutting-edge and high-profile Chat Generative Pre-trained Transformer (ChatGPT) [40]. Inspired by the great success of the Transformer in NLP, the Vision Transformer (ViT), a variant of the Transformer designed specifically for image processing, has recently gained much popularity in the computer vision community [18], [41]. Different from the CNN structure, the ViT converts 2-D images into 1-D sequences first and then applies the self-attention mechanism for feature extraction. With strong capabilities to capture long-range dependencies and non-local relationships, the self-attention mechanism can capture complex global interactions between different areas of the image [25].

In the state-of-the-art ViT backbone, i.e. Swin Transformer [21], [23], the Window-based Multi-Head Self-Attention module (W-MHSA) splits the input features into nonoverlapping windows before performing the standard Multi-Head Self-Attention (MHSA) in each local window. Specifically, giving vector $\boldsymbol{X}$ as the features within a local window, the self-attention mechanism for each head can be defined as:

$$Attention(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = SoftMax(cos(\boldsymbol{Q},\boldsymbol{K}^T)/\tau + \boldsymbol{B})\boldsymbol{V} \tag{4}$$

$$\begin{aligned} \boldsymbol{Q} &= \boldsymbol{X}\boldsymbol{W_q} \in \mathbb{R}^{M^2\times d} \\ \boldsymbol{K} &= \boldsymbol{X}\boldsymbol{W_k} \in \mathbb{R}^{M^2\times d} \\ \boldsymbol{V} &= \boldsymbol{X}\boldsymbol{W_v} \in \mathbb{R}^{M^2\times d} \end{aligned} \tag{5}$$

To calculate the attention map, the *query* matrix $\boldsymbol{Q}$, *key* matrix $\boldsymbol{K}$ and *value* matrix $\boldsymbol{V}$ need to be generated first by applying three projected matrices $\boldsymbol{W}_q$, $\boldsymbol{W}_k$ and $\boldsymbol{W}_v$ to the vector $\boldsymbol{X}$, respectively. The generated matrice are all in the shape of $M^2 \times d$, where M^2 means the number of pixels in a window and d indicates the *query/key* dimension. Thereafter, the SoftMax function is applied to each row of the similarity matrix $cos(\boldsymbol{Q},\boldsymbol{K}^T)/\tau + \boldsymbol{B}$. Here, the similarity matrix is in the shape of $M^2 \times M^2$, which models the relationship between each pair of pixels within the local window. For point (i, j) in the similarity matrix, the value is computed by:

$$cos(\boldsymbol{q}_i,\boldsymbol{k}_j) + \boldsymbol{B}_{i,j} = \left(\frac{\boldsymbol{q}_i^T}{\parallel \boldsymbol{q}_i^T \parallel_2}\right)\cdot\left(\frac{\boldsymbol{k}_j}{\parallel \boldsymbol{k}_j \parallel_2}\right) + \boldsymbol{B}_{i,j} \tag{6}$$

where $\boldsymbol{B}_{i,j}$ is the relative position bias between pixel i and j. τ is a learnable scalar, non-shared across heads and layers. Finally, the obtained similarity matrix is multiplied by the *value* matrix $\boldsymbol{V}$ to yield the attention map. Compared with the conventional multi-head self-attention module which calculates the similarity matrix among the whole input image, the complexity is significantly reduced as the self-attention operation in Swin Transformer is only conducted within the local window range.

Benefiting from this design, the ViT-based algorithms have demonstrated obvious superiority over CNNs and obtained numerous breakthroughs in fundamental vision tasks, such as image classification [20], [21], [23], semantic segmentation [24], [42], [43] and super-resolution [18], [44], [28], [45], [46].

III. Methodology

Although the complexity has been already decreased by the window-based multi-head self-attention module to a great extent, the computational and memory requirements for each self-attention operation are still quadratically related to the size of the window. Considering the large number of self-attention operations contained in the Swin Transformer, the optimization for the complexity of the self-attention operation still has considerable benefit in accelerating the inference speed and reducing the memory requirement, especially when we need a large window size to process images in large size.

A. Kernel Attention

Generally, the self-attention mechanism can be formalized as:

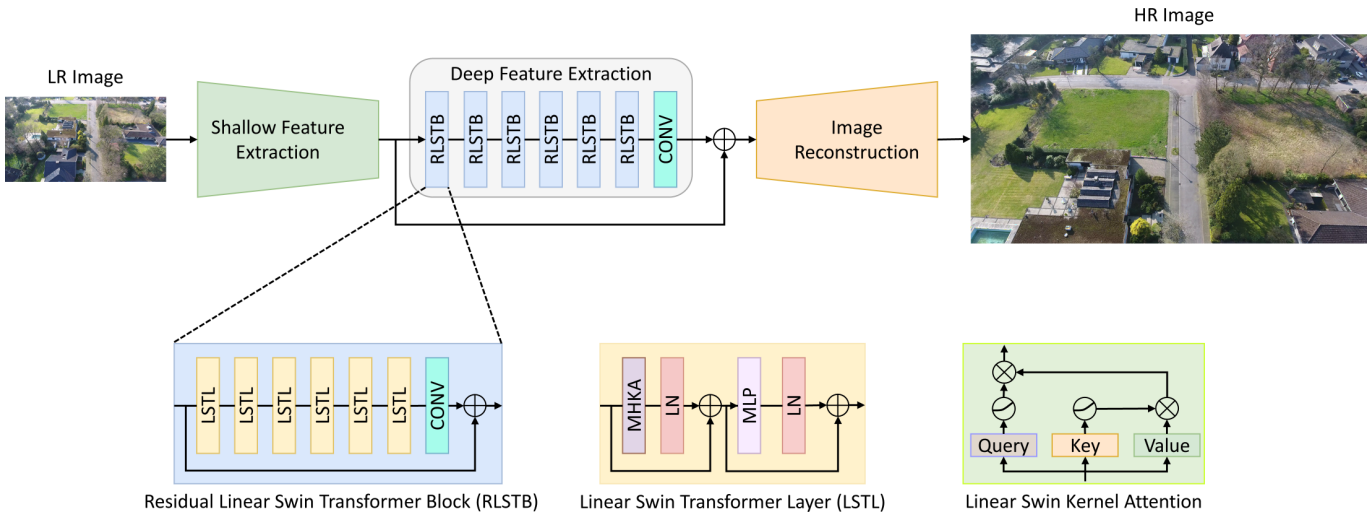

Fig. 1. The architecture of the proposed LSwinSR for UAV image super-resolution.

$$\begin{aligned} Attention(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) &= SoftMax(\boldsymbol{Q}\boldsymbol{K}^T)\boldsymbol{V} \\ &= \rho(\boldsymbol{Q}\boldsymbol{K}^T)\boldsymbol{V} \end{aligned} \tag{7}$$

where ρ represents the normalization function used to normalize the similarity matrix generated by the product of $\boldsymbol{Q}$ and $\boldsymbol{K}^T$. If $\rho(\cdot) = SoftMax(\cdot)$, the i-th row of the attention map can be written as:

$$Attention(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})_i = \frac{\sum_{j=1}^{M} e^{\boldsymbol{q}_i^T \cdot \boldsymbol{k}_j}\boldsymbol{v}_j}{\sum_{j=1}^{M} e^{\boldsymbol{q}_i^T \cdot \boldsymbol{k}_j}} \tag{8}$$

As can be drawn from Equation (8), the essence of the self-attention mechanism is to weigh the $\boldsymbol{v}_j$ by $e^{\boldsymbol{q}_i^T \cdot \boldsymbol{k}_j}$, where $sim(\boldsymbol{q}_i,\boldsymbol{k}_j) = e^{\boldsymbol{q}_i^T \cdot \boldsymbol{k}_j}$ measures the similarity between the $\boldsymbol{q}_i$ and $\boldsymbol{k}_j$. Therefore, Equation (8) can be rewritten as:

$$\begin{aligned} Attention(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})_i = \frac{\sum_{j=1}^{M} \mathrm{sim}(\boldsymbol{q}_i,\boldsymbol{k}_j)\boldsymbol{v}_j}{\sum_{j=1}^{M} \mathrm{sim}(\boldsymbol{q}_i,\boldsymbol{k}_j)}, where \\ \mathrm{sim}(\boldsymbol{q}_i,\boldsymbol{k}_j) \geq 0 \end{aligned} \tag{9}$$

Here, $\mathrm{sim}(\boldsymbol{q}_i,\boldsymbol{k}_j)$ can be expanded as $\mathrm{sim}(\boldsymbol{q}_i,\boldsymbol{k}_j) = \phi(\boldsymbol{q}_i)^T\varphi(\boldsymbol{k}_j)$, thereby reforming Equation (8) as:

$$Attention(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})_i = \frac{\sum_{j=1}^{M} \phi(\boldsymbol{q}_i)^T\varphi(\boldsymbol{k}_j)\boldsymbol{v}_j}{\sum_{j=1}^{M} \phi(\boldsymbol{q}_i)^T\varphi(\boldsymbol{k}_j)} \tag{10}$$

$$Attention(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})_i = \frac{\phi(\boldsymbol{q}_i)^T \sum_{j=1}^{M} \varphi(\boldsymbol{k}_j)\boldsymbol{v}_j^T}{\phi(\boldsymbol{q}_i)^T \sum_{j=1}^{M} \varphi(\boldsymbol{k}_j)} \tag{11}$$

In particular, Equation (10) is identical to Equation (8), when $\mathrm{sim}(\boldsymbol{q}_i,\boldsymbol{k}_j) = e^{\boldsymbol{q}_i^T \cdot \boldsymbol{k}_j}$. Further, Equation (11) can be represented as the vectorized form:

$$Attention(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \frac{\phi(\boldsymbol{Q})\varphi(\boldsymbol{K})^T\boldsymbol{V}}{\phi(\boldsymbol{Q})\sum_j \varphi(\boldsymbol{K})_j^T} \tag{12}$$

As the SoftMax function in Equation (12) is replaced by $\mathrm{sim}(\boldsymbol{q}_i,\boldsymbol{k}_j) = \phi(\boldsymbol{q}_i)^T\varphi(\boldsymbol{k}_j)$, the order of the operation can be altered. Specifically, we can compute the multiplication between $\varphi(\boldsymbol{K})^T$ and $\boldsymbol{V}$ first and then multiply the result and $\phi(\boldsymbol{Q})$, thereby avoiding the intensive computation and generation procedure of the matrix $\boldsymbol{Q}\boldsymbol{K}^T \in \mathbb{R}^{M^2\times M^2}$. Especially, $\phi(\cdot)$ and $\varphi(\cdot)$ in Equation (11) can be considered as kernel smoothers [47]. In our previous work [25], we have shown that by selecting $\mathrm{sim}(\boldsymbol{q}_i\boldsymbol{k}_j) = \mathrm{SoftPlus}(\boldsymbol{q}_i)^T\mathrm{SoftPlus}(\boldsymbol{k}_j)$ where $\mathrm{SoftPlus}(x) = \log(1+e^x)$, a revised self-attention mechanism (i.e. kernel attention [25]) with linear complexity can be achieved with competitive accuracy. For kernel attention, the Equation (11) and Equation (12) can be written as:

$$Attention(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})_i = \frac{\mathrm{SoftPlus}(\boldsymbol{q}_i)^T \sum_{j=1}^{M} \mathrm{SoftPlus}(\boldsymbol{k}_j)\boldsymbol{v}_j^T}{\mathrm{SoftPlus}(\boldsymbol{q}_i)^T \sum_{j=1}^{M} \mathrm{SoftPlus}(\boldsymbol{k}_j)} \tag{13}$$

$$Attention(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \frac{\mathrm{SoftPlus}(\boldsymbol{Q})\mathrm{SoftPlus}(\boldsymbol{K})^T\boldsymbol{V}}{\mathrm{SoftPlus}(\boldsymbol{Q})\sum_j \mathrm{SoftPlus}(\boldsymbol{K})_j^T} \tag{14}$$

B. LSwinSR

The proposed LSwinSR is based on the Swin Transformer [21] and SwinIR [28]. As can be seen from Fig. 1, similar to SwinIR [28], there are three components included in the LSwinSR, i.e. Shallow Feature Extraction, Deep Feature Extraction and Image Reconstruction.

1) Shallow Feature Extraction: Giving a low-resolution input image $\boldsymbol{I}_{LR} \in \mathbb{R}^{H\times W\times C}$, a 3×3 convolutional layer $\mathscr{C}_{SF}$ is first applied to extract the shallow feature $\boldsymbol{F}_{SF} \in \mathbb{R}^{H\times W\times D}$:

$$\boldsymbol{F}_{SF} = \mathscr{C}_{SF}(\boldsymbol{I}_{LR}) \tag{15}$$

where H, W, C and D are the image height, image width, input channel number and feature channel number, respectively. As pointed out by [48], the utilization of the convolution layer at the early visual processing is beneficial for stable optimization and better results. Besides, the convolution operation

Fig. 2. Visual comparison of SR ($\times 2$) images between different methods on training (top) and validation (bottom) sets for segmentation.

allows mapping the input image space to a higher dimensional feature space in a simple way.

2) Deep Feature Extraction: After obtaining the shallow feature, the deep features are extracted sequentially by the deep feature extraction module with H Residual Linear Swin Transformer Blocks (RLSTB) $\mathscr{R}$ and a 3×3 convolutional layer $\mathscr{C}_{DF}$:

$$\begin{aligned} \boldsymbol{F}_{DF} &= \mathscr{C}_{DF}(\boldsymbol{F}_H), where \\ \boldsymbol{F}_i &= \mathscr{R}_i(\boldsymbol{F}_{i-1}), i = 1, 2, \ldots, H \\ \boldsymbol{F}_0 &= \boldsymbol{F}_{SF} \end{aligned} \tag{16}$$

The convolutional layer at the end of the deep feature extraction module can provide the inductive bias to the Transformer-based network and pave a better way for the aggregation of shallow and deep features [28].

As can be seen from Fig. 1, a residual linear Swin Transformer block comprises K Linear Swin Transformer Layers (LSTL) $\mathscr{L}$ and a convolutional layer $\mathscr{C}_{RLSTB}$. Assuming the input feature maps of the i-th RLSTB is $\boldsymbol{F}_{i,0}$, then the output of the current RLSTB $\boldsymbol{F}_{i,out}$ can be represented as:

$$\begin{aligned} \boldsymbol{F}_{i,out} &= \mathscr{C}_{RLSTB}(\boldsymbol{F}_{i,K}) + (\boldsymbol{F}_{i,0}), where \\ \boldsymbol{F}_{i,j} &= \mathscr{L}_j(\boldsymbol{F}_{i,j-1}), j = 1, 2, \ldots, K \end{aligned} \tag{17}$$

Especially, the short identity-based connection provided by the residual connection enables the aggregation of different levels of features.

For each linear Swin Transformer layer, there exist two residual connections and four layers, i.e. a Multi-Head Kernel Attention (MHKA) operation, a Multi-Layer Perceptron (MLP) and two LayerNorm (LN) operations. The LSTL first reshapes the input from the size of $H \times W \times D$ to the $HW/M^2 \times M^2 \times D$ by partitioning the input into non-overlapping $M \times M$ local windows, where HW/M^2 equals the total number of windows. Then, the MHKA operation is applied separately for each window, while the technical details have been illustrated in Section II-C and Section III-A. Thereafter, a MLP with two fully connected layers and a Gaussian Error Linear Unit (GELU) between them is employed for further feature transformations. Meanwhile, the LN layer and the residual connection are adopted for both MHKA and MLP. The whole process of the LSTL can be formulated as:

$$\boldsymbol{X} = MHKA(LN(\boldsymbol{X})) + \boldsymbol{X} \tag{18}$$

$$\boldsymbol{X} = MLP(LN(\boldsymbol{X})) + \boldsymbol{X} \tag{19}$$

Meanwhile, to provide connections across local windows, the regular and shifted window partitioning [21] is used which shifts the feature by $(M/2, M/2)$ pixels before partitioning.

3) Image Reconstruction: To reconstruct the high-resolution images $\boldsymbol{I}_{SR}$, the shallow features $\boldsymbol{F}_{SF}$ and deep features $\boldsymbol{F}_{DF}$ are aggregated as:

$$\boldsymbol{I}_{SR} = \mathscr{R}e(\boldsymbol{F}_{SF} + \boldsymbol{F}_{DF}) \tag{20}$$

where $\mathscr{R}e$ is the reconstruction module. The reason why a long skip connection exists here is that deep features mainly focus on recovering high-frequencies while shallow features primarily centre on reconstructing low-frequencies [28]. With the long skip connection, the low-frequency information contained in shallow features can be directly fed into the reconstruction module, enabling the deep feature extraction module to mainly extract high-frequency information.

IV. Results and Discussions

A. Experimental Setting

1) Dataset: To compare the performance between different methods, we conduct a series of experiments based on a high-resolution UAV semantic segmentation dataset, i.e. the UAVid dataset [51]. The UAVid dataset focuses on urban street scenes with 4K resolutions (3840×2160 or 4096×2160) and eight classes. Both the super-resolution and segmentation of the UAVid dataset are extremely challenging because of the high resolution, heterogeneous spatial variation and generally complex scenes.

UAVid has 42 sequences with a total of 420 images in the dataset (10 images in each sequence), where 27 sequences are used for training and validation and 15 sequences are officially provided for testing. In our experiments, 15 sequences in the test set are remained for testing the performance of both the super-resolution and the segmentation. For the training and validation set, the images within a sequence are divided into four parts: images with ID: 000000, 000200, 000400, 000600 and the image with ID: 000800 are used for training and validating the super-resolution model while images with ID: 000100, 000300, 000500, 000700 and the image with ID: 000900 are used for training and validating the segmentation

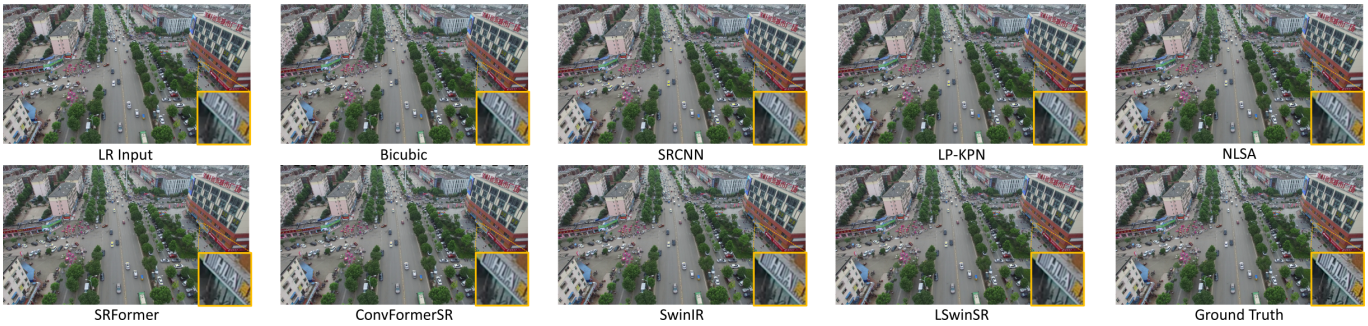

Fig. 3. Visual comparison of SR ($\times 2$) images between different methods on the test set.

TABLE I: Quantitative comparison including PSNR, SSIM (%) and MAE (%) with different methods on the training and validation sets for segmentation and the test set, where the best result is **highlighted** while the second best is underlined.

Models	Scale	Train			Validation			Test		
		PSNR ↑	SSIM ↑	MAE ↓	PSNR ↑	SSIM ↑	MAE ↓	PSNR ↑	SSIM ↑	MAE ↓
Bicubic	×2	28.760	90.878	2.273	28.817	90.863	2.329	27.933	85.956	2.437
SRCNN [6]	×2	32.994	93.195	1.986	32.907	93.079	2.045	30.516	83.791	2.859
LP-KPN [35]	×2	33.782	93.719	1.832	33.722	93.579	1.891	31.089	87.668	2.551
NLSA [36]	×2	34.033	93.969	1.779	33.950	93.815	1.839	31.412	88.571	2.469
SRFormer [49]	×2	34.175	94.150	1.770	34.049	93.976	1.838	31.991	88.641	2.313
ConvFormerSR [50]	×2	33.504	93.661	1.874	33.411	93.532	1.933	27.830	83.216	12.645
SwinIR [28]	×2	**34.354**	**94.240**	**1.720**	**34.253**	**94.073**	**1.782**	31.997	**89.957**	2.305
LSwinSR	×2	34.277	94.169	1.738	34.173	94.004	1.801	**32.040**	89.868	**2.303**
Bicubic	×4	22.915	66.157	4.862	22.904	66.022	4.991	23.967	69.091	3.815
SRCNN [6]	×4	25.296	67.972	4.250	25.275	67.671	4.388	24.938	54.851	8.662
LP-KPN [35]	×4	25.960	71.908	3.888	25.887	71.431	4.037	27.963	75.774	3.668
NLSA [36]	×4	26.117	72.182	3.914	26.013	71.628	4.069	27.524	73.330	4.212
SRFormer [49]	×4	26.107	72.885	3.894	26.001	72.311	4.052	27.851	67.786	4.528
ConvFormerSR [50]	×4	26.207	72.908	3.814	26.131	72.411	3.950	27.882	74.943	3.883
SwinIR [28]	×4	**26.338**	**73.563**	**3.731**	**26.253**	**73.040**	**3.872**	**28.150**	**76.505**	**3.628**
LSwinSR	×4	26.227	73.075	3.777	26.134	72.558	3.928	27.975	76.202	3.687
Bicubic	×8	19.923	46.763	6.930	19.920	46.740	7.094	21.044	50.527	5.562
SRCNN [6]	×8	21.581	48.912	6.175	21.614	48.865	6.320	20.000	34.340	23.075
LP-KPN [35]	×8	22.151	51.735	5.653	22.159	51.538	5.819	25.787	67.759	**5.124**
NLSA [36]	×8	**22.886**	**55.511**	**5.223**	**22.859**	**55.117**	**5.385**	18.453	37.480	26.185
SRFormer [49]	×8	22.458	53.487	5.451	22.464	53.231	5.669	22.359	37.027	16.639
ConvFormerSR [50]	×8	21.863	50.150	5.889	21.899	50.077	6.040	25.002	44.352	11.993
SwinIR [28]	×8	22.402	52.770	5.501	22.434	52.645	5.642	26.144	**67.954**	5.303
LSwinSR	×8	22.494	53.493	5.438	22.506	53.265	5.588	**26.256**	66.269	5.683

model, respectively. In such a setting, the datasets for super-resolution and for segmentation are totally separated, thereby ensuring an independent and objective evaluation.

2) Model Training: Based on the UAVid dataset, the super-resolution for three kinds of upsampling scales, i.e. $\times 2$, $\times 4$ and $\times 8$, are carried out. For those three scales, the high-resolution images are first cropped into 256×256, 512×512 or 1024×1024 patches, respectively. Then, based on the resize function defined in the OpenCV [52], the cropped patches are downsampled to $1/2$, $1/4$ or $1/8$ of the original sizes, i.e. the low-resolution images for three experiments are both in 128×128. Before the downsampling procedure of the test set, we apply the Gaussian Blur to raw high-resolution images first, in order to simulate the practical usage scenario where the degradation procedure of the test set is normally unknown and different from the experimental setting of the training set. Finally, the super-resolution models are trained to reconstruct the corresponding $\times 2$, $\times 4$ or $\times 8$ high-resolution images. For model training, the L_1 pixel loss is selected as the loss function:

$$L(\boldsymbol{I}_{HR}, \boldsymbol{I}_{SR}) = \| \boldsymbol{I}_{HR} - \boldsymbol{I}_{SR} \|_1 \tag{21}$$

The super-resolution models are trained for 35, 50 and 75 epochs for $\times 2$, $\times 4$ and $\times 8$ upsampling scales with a learning rate of 3×10^{-4} and batch size of 4. The segmentation model is trained 50 epochs with a learning rate of 6×10^{-4} and batch size of 8.

3) Model Evaluation: After training and validation, the super-resolution model is then employed to upsample the training, validation and test sets for segmentation. To evaluate the performance of the super-resolution results, three frequently used metrics are adopted including the Peak-Signal-to-Noise Ratio (PSNR), the Structural Similarity Index Measure (SSIM) and the Mean Absolute Error (MAE). Thereafter, we train, validate and test a segmentation model, i.e. the UNet-like transformer (UNetFormer) [24] based on the upsampled images generated from different super-resolution algorithms. By observing, comparing and analysing the segmentation accuracy, a more application-oriented evaluation for super-resolution can be achieved.

B. Super-resolution Performance

In this section, the performance of eight different super-resolution methods including Bicubic, SRCNN [6], LP-

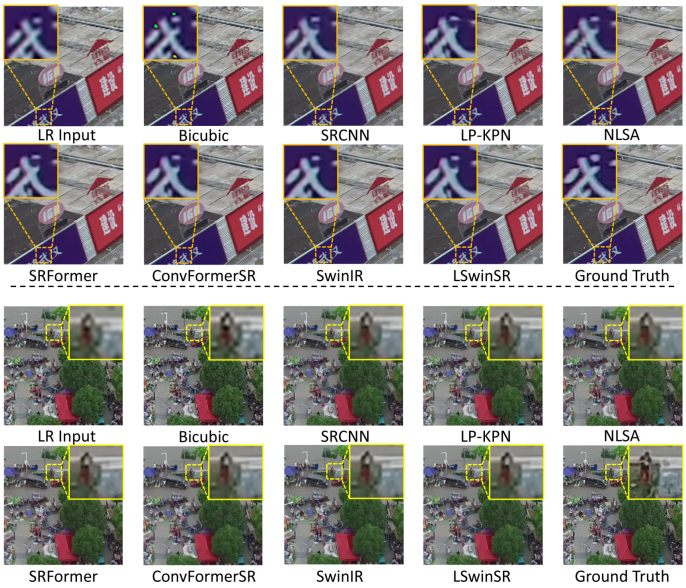

Fig. 4. Visual comparison of SR ($\times 4$) images between different methods on the training (top) and validation (bottom) sets for segmentation.

KPN [35], NLSA [36], SRFormer [49], ConvFormerSR [50], SwinIR [28] and the proposed LSwinSR is evaluated based on the image quality assessments, i.e. PSNR, SSIM and MAE. Among these eight methods, the SRCNN [6], LP-KPN [35] and NLSA [36] are CNN-based models, while the LP-KPN [35] and NLSA [36] enhance the model performance by combining the traditional image processing technologies. The ConvFormerSR [50] combines the CNN and Transformer to improve the performance. The SRFormer [49], SwinIR [28] and the proposed LSwinSR, by contrast, are constructed based on the state-of-the-art Swin Transformer [21]. The SwinIR [28] and the proposed LSwinSR in this section are both based on the lightweight version. Please note that the training and validation sets in this section refer to the training and validation sets for the semantic segmentation which is totally separated from the training and validation sets for super-resolution.

1) Quantitative Results: As explained in Section IV-A2, we add the Gaussian Blur to the test set before generating the low-resolution images. Therefore, the probability distributions between the training set and the test set are different, which means that untrained patterns exist when predicting the test set. This is part of the reason why the performance of SRCNN [6] is better than the Bicubic interpolation on the training and validation sets but worse on the test set. To be specific, the SRCNN [6] takes the Bicubic upsampled images as the input (i.e. pre-upsampling) and then processes the input based on only three convolutional layers. Thus, the errors that existed in the Bicubic upsampled images will be further enlarged by the post-processing procedures, which is especially true when considering that there are only three convolutional layers in the SRCNN [6] to process the upsampled images. By contrast, also as a pre-upsampling-based network, the enhanced post-processing procedures in the LP-KPN [35] guarantee a more stable and more robust performance on the untrained test set.

NLSA [36], SRFormer [49], ConvFormerSR [50], SwinIR [28] and the proposed LSwinSR are all based on the mainstream post-upsampling structure, which extracts the feature maps in the low-resolution input and upsamples the features at the end. A significant advantage of the post-upsampling structure is the much lower computational cost as the feature extraction procedure is only conducted on the low-resolution space. Thus, the more complex and advanced structure becomes acceptable and practical. For example, the NLSA [36] holds better performance than the pre-upsampling-based network most of the time. However, for the $\times 8$ scale, a severe over-fitting problem occurs in the NLSA [36] and SRFormer [49]. The accuracy of the NLSA [36] for $\times 8$ scale is the worst among all eight methods, although with the best performance on both train and validation sets.

As shown in Table I, the proposed LSwinSR holds a competitive performance with the SwinIR [28] for all three scales ($\times 2$, $\times 4$ and $\times 8$). Taking the $\times 2$ scale as an example, although the SwinIR [28] performs better on the training and validation sets, our LSwinSR delivers the best PSNR and MAE on the test set, demonstrating strong robustness and generalizability. More quantitative results for the super-resolution of the proposed LSwinSR can refer to Appendix A.

2) Qualitative Comparisons: In addition to quantitative results, the qualitative comparison of the reconstructed results with different methods is also provided from Fig. 2 to Fig. 7, where an enlarged region is presented on the corner within each corresponding image for convenient comparison.

Obviously, a better reconstruction can be achieved on the lower super-resolution scale. For example, in Fig. 2, the e-bike riders (top) and the building contours (bottom) generated by different methods are all recognizable. However, the noises of images generated by the Bicubic, SRCNN [6], LP-KPN [35] and NLSA [36] are much clearer and more than those by the SwinIR [28] and the LSwinSR, while some greenish points are wrongly added by the ConvFormerSR [50]. In the bottom part of Fig. 2, there exist dozens of obvious reddish and blackish noisy points on the white vertical ridge of the building in the predicted images (as enlarged). Particularly, such severe noisy points are much less frequent for images generated by SwinIR [28] and LSwinSR which contain only two and one such noisy points, respectively.

With the increase of the upsample scale, the available information for accurate reconstruction becomes more and more scarce, leading to great challenges for super-resolution algorithms. For example, in the top part of Fig. 4, the transformer-based methods provide smooth results for the text in the image when upsampling $\times 4$ scale. By comparison, the image generated by the Bicubic interpolation contains serious noises, while the image predicted by SRCNN [6] is more blurry than others. For LP-KPN [35] and NLSA [36], the problem is the unnecessary wrinkles in the reconstructed texts. When it comes to $\times 8$ super-resolution, the super-resolution algorithms can only rely on very limited input information. As shown in Fig. 7, the details of the building structure are nearly indistinguishable for all methods, but the SwinIR [28]

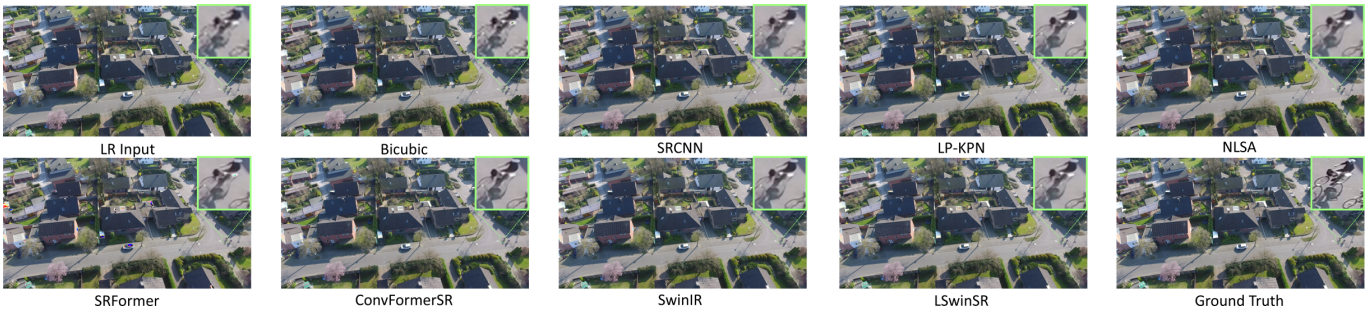

Fig. 5. Visual comparison of SR ($\times 4$) image between different methods on the test set.

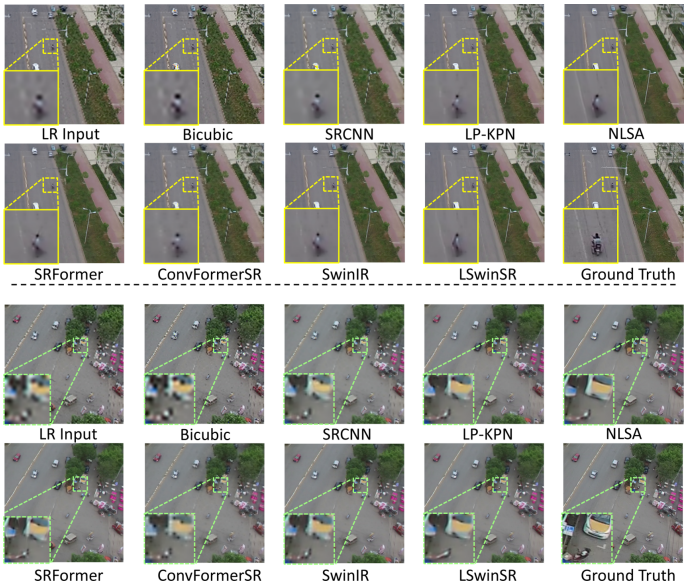

Fig. 6. Visual comparison of SR ($\times 8$) images between different methods on the training (top) and validation (bottom) sets for segmentation.

and LSwinSR still provide a clearer contour.

TABLE II: The comparison of complexity and inference speed between SwinIR [28] and LSwinSR under different window sizes. The complexity is measured by the Multiply-Accumulate Operations (MACs), while the inference speed, i.e. Frames Per Second (FPS), is evaluated based on the $4 \times 3 \times 128 \times 128$ input.

Window Size	Model	Complexity (G) ↓	Speed ↑
8	LSwinSR	**15.3**	**14.2**
	SwinIR	16.8	12.0
16	LSwinSR	**17.6**	**12.2**
	SwinIR	25.8	6.3
32	LSwinSR	**26.6**	**6.7**
	SwinIR	62.0	1.9

3) Inference Speed: As the proposed LSwinSR is an improved and simplified version of SwinIR [28], the complexity and the inference speed between the two models are compared. As can be seen in Table II, the complexity and the inference speed of our LSwinSR are slightly better than SwinIR [28] when the window size is eight. The reason why the gap is so small is two-fold. First, the other operations except for the self-attention mechanism such as patch embedding, patch merging and MLP also occupy a large part of the network structure. Second, the complexity of the traditional self-attention mechanism increases quadratically with the size of the input, while the high computational requirement is not obvious for a small window size. Therefore, the complexity and inference speed gaps between the LSwinSR and SwinIR [28] are dramatically widened with the increase in window sizes. When the window size reaches 32, the inference speed of the LSwinSR is at least three times faster than the SwinIR [28]. The ablation study about the impact of window size for super-resolution performance can refer to Appendix B.

TABLE III: The comparison of complexity and inference speed between different models. The complexity is measured by the Multiply-Accumulate Operations (MACs), while the inference speed, i.e. Frames Per Second (FPS), is evaluated based on the $4 \times 3 \times 128 \times 128$ input.

Models	Complexity (G) ↓	Speed ↑
SRCNN	18.2	**78.2**
LP-KPN	50.4	7.4
NLSA	213.2	6.2
SRFormer	63.3	6.3
ConvFormerSR	237.9	4.8
SwinIR	16.8	12.0
LSwinSR	**15.3**	14.2

Furthermore, we also compare the efficiency of the proposed LSwinSR and other methods in Table III. As can be seen, the proposed LSwinSR hold advantages in both complexity and inference speed. Although the SRCNN has the fastest speed, the low performance in both super-resolution and semantic segmentation highly limits its practical potential.

C. Semantic Segmentation Performance

As image quality assessments such as PSNR and SSIM cannot directly reflect the reliability of the predicted images for practical applications, we further evaluate the robustness of the reconstructed images from different methods for semantic segmentation.

For each super-resolution method, the UNetFormer [24] is trained and validated using the upsampled training and validation sets for segmentation, whereafter the trained UNetFormer [24] is then used to predict the segmentation maps of the upsampled test set, i.e. the training, validation and test sets are all upsampled by each corresponding super-resolution method. As a reference, we train the model based on the original high-resolution images and report the segmentation

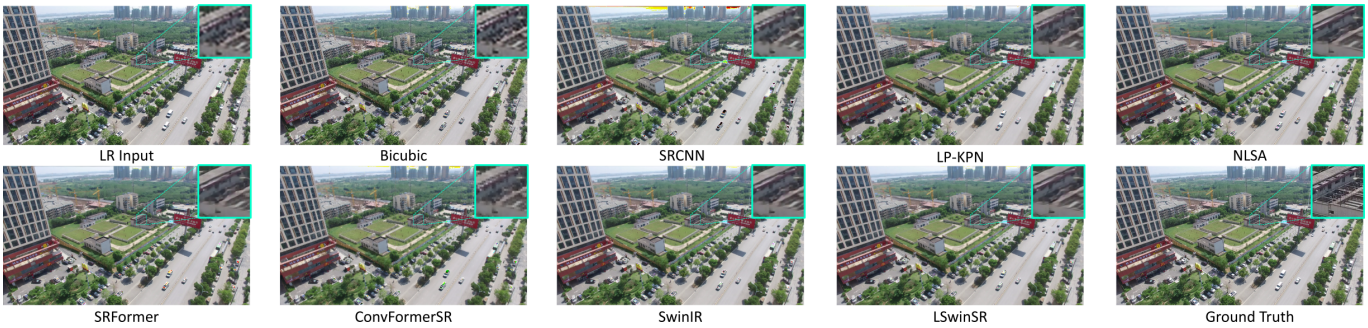

Fig. 7. Visual comparison of SR ($\times 8$) images between different methods on the test set.

TABLE IV: Quantitative comparison for segmentation based on super-resolution images generated by different methods on the test set. The accuracy of each class is measured by IoU scores. The results based on the original high-resolution images (HR) and the high-resolution image with Gaussian Blur (HR-Blur) are marked in light blue and light purple, respectively, while the best result is **highlighted** and the second best is underlined.

Model	Scale	Clutter	Building	Road	Tree	Vegetation	Static Car	Moving Car	Human	mIoU
Bicubic	$\times 2$	0.592	0.825	0.761	0.749	0.577	0.417	0.661	0.229	0.601
SRCNN [6]	$\times 2$	0.602	0.814	0.766	0.734	0.575	0.360	0.661	0.244	0.595
LP-KPN [35]	$\times 2$	0.596	0.815	0.747	0.756	0.559	0.414	0.658	0.246	0.599
NLSA [36]	$\times 2$	0.599	0.817	0.761	0.754	0.583	0.440	0.656	0.257	0.608
SRFormer [49]	$\times 2$	0.588	0.819	0.735	0.758	0.575	**0.480**	0.652	0.197	0.600
ConvFormerSR [50]	$\times 2$	0.598	0.824	0.758	0.758	0.583	0.437	0.656	0.049	0.583
SwinIR [28]	$\times 2$	**0.617**	**0.828**	**0.771**	0.757	0.599	0.441	0.683	0.265	0.620
LSwinSR	$\times 2$	0.605	0.821	0.756	**0.774**	**0.602**	0.469	**0.686**	**0.273**	**0.623**
HR-Blur	$\times 2$	0.603	0.835	0.745	0.705	0.562	0.476	0.630	0.232	0.599
HR	$\times 2$	0.635	0.850	0.775	0.777	0.593	0.509	0.676	0.273	0.636
Bicubic	$\times 4$	0.594	0.812	0.755	0.757	0.580	0.427	0.648	0.187	0.595
SRCNN [6]	$\times 4$	0.593	0.813	0.743	**0.759**	**0.587**	0.392	0.643	0.203	0.592
LP-KPN [35]	$\times 4$	0.603	0.827	0.765	**0.759**	0.585	0.436	0.653	0.206	0.604
NLSA [36]	$\times 4$	0.585	0.816	0.755	**0.759**	0.582	0.399	0.645	0.202	0.593
SRFormer [49]	$\times 4$	0.591	0.810	**0.776**	0.740	0.582	0.398	0.624	**0.213**	0.592
ConvFormerSR [50]	$\times 4$	0.599	**0.833**	0.764	0.761	**0.587**	0.528	0.636	0.000	0.588
SwinIR [28]	$\times 4$	**0.607**	0.829	0.768	0.756	0.579	0.483	0.655	0.202	0.610
LSwinSR	$\times 4$	0.603	0.824	0.761	0.757	0.586	**0.486**	**0.657**	**0.213**	**0.611**
HR-Blur	$\times 4$	0.611	0.841	0.758	0.698	0.557	0.532	0.609	0.225	0.604
HR	$\times 4$	0.642	0.850	0.785	0.782	0.601	0.530	0.680	0.270	0.643
Bicubic	$\times 8$	0.553	0.774	0.732	0.726	0.562	0.322	0.588	0.007	0.533
SRCNN [6]	$\times 8$	0.446	0.677	0.640	0.602	0.464	0.036	0.150	0.000	0.377
LP-KPN [35]	$\times 8$	0.565	0.795	0.743	0.735	0.558	0.338	0.611	0.149	0.562
NLSA [36]	$\times 8$	0.572	0.793	0.748	0.730	0.565	**0.376**	0.611	0.145	0.567
SRFormer [49]	$\times 8$	0.548	0.769	0.720	0.717	0.554	0.361	0.560	0.000	0.528
ConvFormerSR [50]	$\times 8$	0.543	0.769	0.731	0.730	0.556	0.337	0.595	0.000	0.532
SwinIR [28]	$\times 8$	0.573	0.791	0.751	0.744	**0.576**	0.330	0.637	**0.173**	0.572
LSwinSR	$\times 8$	**0.590**	**0.803**	**0.753**	**0.748**	0.572	0.321	**0.640**	0.159	**0.573**
HR-Blur	$\times 8$	0.628	0.849	0.776	0.713	0.567	0.535	0.675	0.229	0.621
HR	$\times 8$	0.654	0.856	0.800	0.787	0.615	0.546	0.712	0.292	0.658

accuracy based on the original test set (HR) and the test set after Gaussian Blur (HR-Blur). The visual comparisons between segmentation maps based on super-resolution images generated by different algorithms are provided from Fig. 8 to Fig. 10.

In Table IV, the performance of semantic labelling for different models is assessed based on the standard mean IoU metric. Due to the different probability distribution between the training set and test set, the accuracy of the test set with Gaussian Blur (HR-Blur) is obviously lower than the original test set (HR). Fortunately, this distributional difference can be offset by the super-resolution methods to a certain degree. For example, as shown in Table IV, the segmentation accuracy of the Bicubic interpolation for $\times 2$ scenario (mIoU: 0.601) is even better than the HR-Blur (mIoU: 0.599). We conjecture that the distributional difference gap between the training set and test set is narrowed after downsampling and upsampling operations, relieving the adverse impact caused by the Gaussian Blur. However, when it comes to the $\times 8$ scale, the accuracy of HR-Blur surpasses all super-resolution methods, even with a different probability distribution with the training set. The reason is that small objects such as humans and cars are hardly identifiable after downsampling to $1/8$ resolution, while such severe information loss cannot be totally recovered by the super-resolution methods. Therefore, the adverse impact caused by information loss overshadows the distributional difference, leading to the worse segmentation accuracy of super-resolution results.

Among seven deep-learning-based methods, only the SwinIR [28] and the proposed LSwinSR can always deliver better results than the simple Bicubic interpolation for all three scales, demonstrating the reliability and robustness of

the Transformer-based super-resolution methods on different scenarios. Meanwhile, what the segmentation accuracy can show is only the superiority rather than the full potential of the Transformer-based solution. For example, as shown in the enlarged region in Fig. 5, the cyclist can be clearly identified. However, only a small part of those pixels is correctly classified as Human in Fig. 9, even for the original high-resolution input. The reason is that the humans in the UAVid are very small and only occupy limited pixels in the high-resolution 4K images. Without enough training data, the segmentation model naturally lacks the ability to interpret those small objects. Actually, the best accuracy on the Human object for the UAVid dataset achieved by the state-of-the-art segmentation model is only 0.33 measured by mIoU. In other words, the capability of the segmentation model limits the full potential of the super-resolution results in segmentation performance. Thus, with the development of the segmentation model especially the optimization for the identification of small objects, we believe a much better segmentation accuracy can be achieved by the super-resolution results.

It is also noteworthy that the segmentation accuracy of NLSA [36] for $\times 8$ scale (mIoU: 0.567) is much higher than the SRCNN [6] (mIoU: 0.377), although the latter holds better image quality assessments, e.g. the PSNR of SRCNN is 20.000 while the one of the NLAS is 18.453. This phenomenon can illustrate two issues. First, the higher image quality metrics are not always connected with better reliability. Second, even though the over-fitting problem can cause an extremely low image quality assessment, a tolerable segmentation result can still be guaranteed so long as the model can narrow the probability distribution gap between the training set and test set. More results on different experimental settings are reported in Section C.

V. Conclusions

In this work, a novel Linear Swin Transformer for Super-Resolution (LSwinSR) was proposed. The LSwinSR has addressed the high memory and computational requirements of the original Swin Transformer caused by the quadratic complexity of the embedded self-attention mechanism by introducing the kernel attention mechanism. The super-resolution experiments conducted on the large-scale UAVid dataset demonstrated that the proposed LSwinSR could provide competitive performance compared to the SwinIR [28] but with better efficiency. Furthermore, the experiments for semantic segmentation demonstrated that the super-resolution technology could indeed enhance the segmentation accuracy, where only the SwinIR [28] and proposed LSwinSR could always deliver better performance than the simple Bicubic interpolation.

In the future, we will develop a framework equipping the super-resolution and semantic segmentation models with real-time capacity, thereby deploying it to the UAV platform for offshore renewable infrastructure monitoring.

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.

Appendix A
Super-Resolution Performance on AID

To further demonstrate the effectiveness of the proposed LSwinSR, we conduct the experiment on the publicly available AID [53] dataset, which has been used for verifying the super-resolution performance. The AID dataset contains 10000 high-resolution images in the shape of 600×600 with a 0.5m spatial resolution from 30 different types of remote sensing scenarios, such as airports, bridges and churches. For super-resolution, following the setting in [45], 7850 images are selected as the training set, 150 as the validation set and the remaining 2000 images are used as the test set. The experimental results are reported in Table A1. As shown in Table A1, in the large-scale and high-resolution remote sensing dataset, i.e. the AID [53], the proposed LSwinSR can still deliver competitive accuracy compared with the existing super-resolution algorithms. Please note that the LSwinSR-L means the lightweight version while the LSwinSR-C denotes the classical version.

TABLE A1: Quantitative comparison including PSNR and SSIM (%) with different methods on the AID dataset, where the best result is **highlighted** while the second best is underlined.

Model	Scale	PSNR ↑	SSIM ↑
Bicubic	×2	32.39	89.06
SRCNN [6]	×2	34.49	92.86
FSRCNN [30]	×2	34.73	93.30
VDSR [29]	×2	35.05	93.46
LGCNet [54]	×2	34.80	93.20
DCM [55]	×2	35.21	93.66
HSENet [56]	×2	35.24	93.68
TransENet [45]	×2	35.28	93.74
SwinIR [28]	×2	35.44	93.77
LSwinSR-L	×2	35.42	93.76
LSwinSR-C	×2	**35.52**	**94.01**
Bicubic	×3	29.08	78.63
SRCNN [6]	×3	30.55	83.72
FSRCNN [30]	×3	30.98	84.00
VDSR [29]	×3	31.15	85.22
LGCNet [54]	×3	30.73	84.17
DCM [55]	×3	31.31	85.61
HSENet [56]	×3	31.39	85.72
TransENet [45]	×3	31.45	85.95
SwinIR [28]	×3	31.52	85.94
LSwinSR-L	×3	31.47	85.85
LSwinSR-C	×3	**31.60**	**86.28**
Bicubic	×4	27.30	70.36
SRCNN [6]	×4	28.40	75.61
FSRCNN [30]	×4	28.77	77.20
VDSR [29]	×4	28.99	77.53
LGCNet [54]	×4	28.61	76.26
DCM [55]	×4	29.17	78.24
HSENet [56]	×4	29.21	78.50
TransENet [45]	×4	29.38	79.09
SwinIR [28]	×4	29.28	78.60
LSwinSR-L	×4	29.30	78.62
LSwinSR-C	×4	**29.44**	**79.25**

Appendix B
Ablation Study

In this section, the ablation study about the influence of the window size and the patch embedding dimension on super-resolution performance is carried out. Specifically, taking the $\times 4$ upsampling scale as an example, we investigate the super-resolution performance of the proposed LSwinSR with

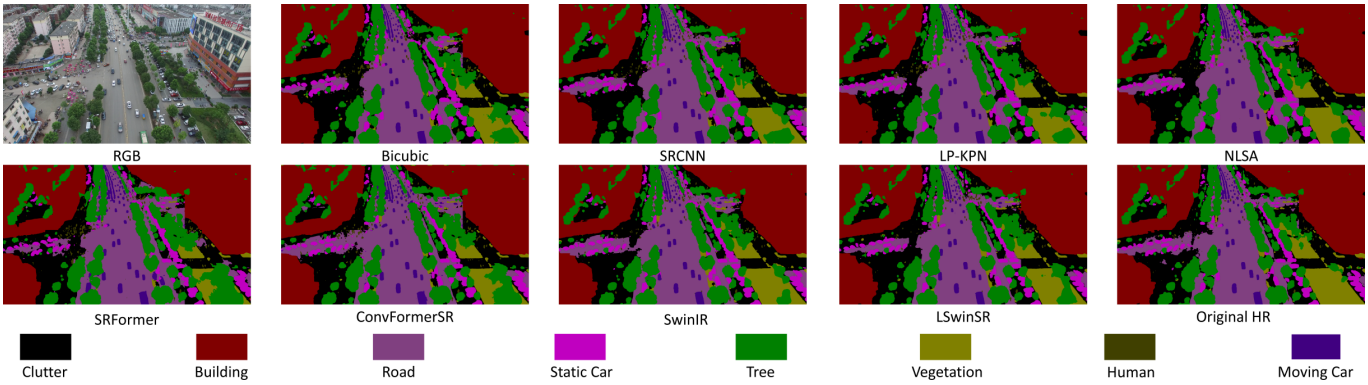

Fig. 8. The segmentation results on the test set based on $\times 2$ SR images generated by different methods.

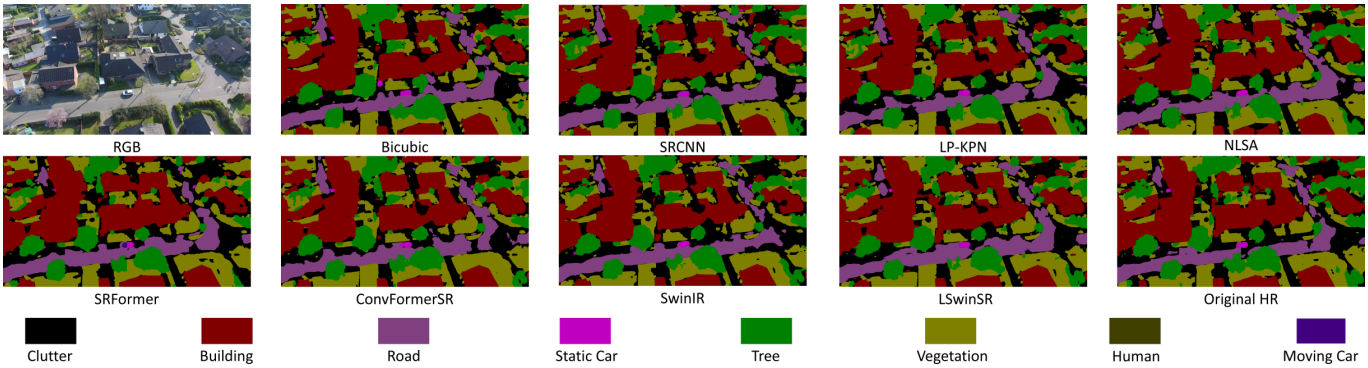

Fig. 9. The segmentation results on the test set based on $\times 4$ SR images generated by different methods.

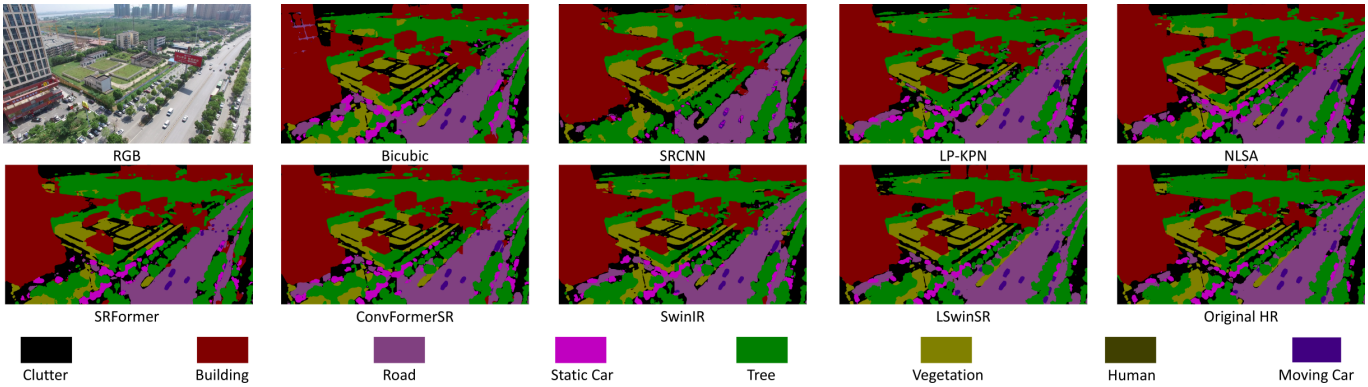

Fig. 10. The segmentation results on the test set based on $\times 8$ SR images generated by different methods.

different window sizes, i.e. 8, 16 and 32, and with different patch embedding dimensions, i.e. 48, 60, 72, while the results are reported in Table B1 and Table B2. Basically, as the larger window size normally can include more contextual information within the same window, the super-resolution therefore indeed improves with the increase of the window size. When it comes to patch embedding dimensions, the performance gets stead enhancement as the increase in embedding dimensions will introduce more parameters of the whole framework.

Appendix C
Experiments under different settings

In this section, to demonstrate the robustness and generalizability of different methods, we further conduct two extra experiments based on the UAVid dataset. As the Gaussian blur is added to the test set for the segmentation task, we then test and report the performance of different methods based on the original test set. Specifically, the raw test set is downsampled by a scale of 4 without any further processing, and then different super-resolution methods are used to upsample it. Then, based on the upsampling results and the trained UNetFormer structure, the segmentation performance is evaluated. As shown in Table C1, the mIoU of most methods improves in varying degrees. The proposed LSwinSR, in particular, increases from 0.611 to 0.627, further demonstrating its effectiveness.

Furthermore, we also conducted super-resolution experiments with Gaussian noise on the test set, as Gaussian noise

TABLE B1: Ablation study about the window size. Please note that LSwinSR-8, LSwinSR-16 and LSwinSR-32 represent the proposed LSwinSR with window size of 8, 16 and 32, respectively.

	Model	PSNR ↑	SSIM ↑	MAE ↓
Train	LSwinSR-8	26.227	73.075	3.777
	LSwinSR-16	26.397	73.717	3.710
	LSwinSR-32	**26.403**	**73.759**	**3.704**
Validation	LSwinSR-8	26.134	72.558	3.928
	LSwinSR-16	26.306	73.174	3.853
	LSwinSR-32	**26.310**	**73.179**	**3.850**
Test	LSwinSR-8	27.975	76.202	3.787
	LSwinSR-16	28.019	76.179	3.737
	LSwinSR-32	**28.182**	**76.219**	**3.720**

TABLE B2: Ablation study about the patch embedding dimension. Please note that LSwinSR-48, LSwinSR-60 and LSwinSR-72 represent the proposed LSwinSR with 48, 60 and 72 patch embedding dimensions, respectively.

	Model	PSNR ↑	SSIM ↑	MAE ↓
Train	LSwinSR-48	26.191	72.585	3.796
	LSwinSR-60	26.227	73.075	3.777
	LSwinSR-72	**26.351**	**73.642**	**3.723**
Validation	LSwinSR-48	26.097	72.504	3.944
	LSwinSR-60	26.134	72.558	3.928
	LSwinSR-72	**26.257**	**73.119**	**3.868**
Test	LSwinSR-48	27.364	75.679	3.796
	LSwinSR-60	27.975	76.202	3.787
	LSwinSR-72	**28.243**	**76.704**	**3.659**

is very common in images caused by sensors and electronic circuits. Specifically, we add the Gaussian noise with a mean of 0.1 and a sigma of 0.01 for case1 and a mean of 0.0 and a sigma of 0.03 for case2 to the original test set and then downsample the images with a 4 scale. Then, the trained super-resolution models are used to upsample the images. As shown in Table C2 and C3, all methods experience different levels of decrease under two scenarios. However, the proposed LSwinSR and SwinIR [28] still have the best performance compared with other methods.

REFERENCES

[1] Q. Zhang, S. Zheng, C. Zhang, X. Wang, R. Li, Efficient large-scale oblique image matching based on cascade hashing and match data scheduling, Pattern Recognition 138 (2023) 109442.

[2] H. Yuan, C. Xiao, Y. Wang, X. Peng, Y. Wen, Q. Li, Maritime vessel emission monitoring by an uav gas sensor system, Ocean Engineering 218 (2020) 108206.

[3] S. Hu, W. Ni, X. Wang, A. Jamalipour, D. Ta, Joint optimization of trajectory, propulsion, and thrust powers for covert uav-on-uav video tracking and surveillance, IEEE Transactions on Information Forensics and Security 16 (2020) 1959–1972.

[4] J. Xie, L. Fang, B. Zhang, J. Chanussot, S. Li, Super resolution guided deep network for land cover classification from remote sensing images, IEEE Transactions on Geoscience and Remote Sensing 60 (2021) 1–12.

[5] P. Mao, J. Ding, B. Jiang, L. Qin, G. Y. Qiu, How can uav bridge the gap between ground and satellite observations for quantifying the biomass of desert shrub community?, ISPRS Journal of Photogrammetry and Remote Sensing 192 (2022) 361–376.

[6] C. Dong, C. C. Loy, K. He, X. Tang, Image super-resolution using deep convolutional networks, IEEE transactions on pattern analysis and machine intelligence 38 (2) (2015) 295–307.

[7] Z. Lu, J. Li, H. Liu, C. Huang, L. Zhang, T. Zeng, Transformer for single image super-resolution, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 457–466.

[8] W. Zhang, Y. Liu, C. Dong, Y. Qiao, Ranksrgan: Super resolution generative adversarial networks with learning to rank, IEEE Transactions on Pattern Analysis and Machine Intelligence 44 (10) (2021) 7149–7166.

[9] Z. Chen, X. Guo, P. Y. Woo, Y. Yuan, Super-resolution enhanced medical image diagnosis with sample affinity interaction, IEEE Transactions on Medical Imaging 40 (5) (2021) 1377–1389.

[10] F. A. Dharejo, M. Zawish, F. Deeba, Y. Zhou, K. Dev, S. A. Khowaja, N. M. F. Qureshi, Multimodal-boost: Multimodal medical image super-resolution using multi-attention network with wavelet transform, IEEE/ACM Transactions on Computational Biology and Bioinformatics 20 (4) (2022) 2420–2433.

[11] M. T. Razzak, G. Mateo-García, G. Lecuyer, L. Gómez-Chova, Y. Gal, F. Kalaitzis, Multi-spectral multi-image super-resolution of sentinel-2 with radiometric consistency losses and its effect on building delineation, ISPRS Journal of Photogrammetry and Remote Sensing 195 (2023) 1–13.

[12] V. Vasilescu, M. Datcu, D. Faur, A cnn-based sentinel-2 image super-resolution method using multiobjective training, IEEE Transactions on Geoscience and Remote Sensing 61 (2023) 1–14.

[13] J. Wang, Z. Shao, X. Huang, T. Lu, R. Zhang, Y. Li, From artifact removal to super-resolution, IEEE Transactions on Geoscience and Remote Sensing 60 (2022) 1–15.

[14] C. Xiang, W. Wang, L. Deng, P. Shi, X. Kong, Crack detection algorithm for concrete structures based on super-resolution reconstruction and segmentation network, Automation in Construction 140 (2022) 104346.

[15] J. Jiang, J. Liu, J. Fu, W. Wang, H. Lu, Super-resolution semantic segmentation with relation calibrating network, Pattern Recognition 124 (2022) 108501.

[16] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, C. Change Loy, Esrgan: Enhanced super-resolution generative adversarial networks, in: Proceedings of the European conference on computer vision (ECCV) workshops, 2018, pp. 0–0.

[17] Z. Li, J. Yang, Z. Liu, X. Yang, G. Jeon, W. Wu, Feedback network for image super-resolution, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 3867–3876.

[18] M. V. Conde, U.-J. Choi, M. Burchi, R. Timofte, Swin2sr: Swinv2 transformer for compressed image super-resolution and restoration, in: Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part II, Springer, 2023, pp. 669–687.

[19] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems, Vol. 30, 2017.

[20] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, in: International Conference on Learning Representations, 2021.

[21] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: Hierarchical vision transformer using shifted windows, in: Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 10012–10022.

[22] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, J. Dai, Deformable detr: Deformable transformers for end-to-end object detection, in: International Conference on Learning Representations, 2021.

[23] Z. Liu, H. Hu, Y. Lin, Z. Yao, Z. Xie, Y. Wei, J. Ning, Y. Cao, Z. Zhang, L. Dong, et al., Swin transformer v2: Scaling up capacity and resolution, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 12009–12019.

[24] L. Wang, R. Li, C. Zhang, S. Fang, C. Duan, X. Meng, P. M. Atkinson, Unetformer: A unet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery, ISPRS Journal of Photogrammetry and Remote Sensing 190 (2022) 196–214.

[25] R. Li, S. Zheng, C. Zhang, C. Duan, J. Su, L. Wang, P. M. Atkinson, Multiattention network for semantic segmentation of fine-resolution remote sensing images, IEEE Transactions on Geoscience and Remote Sensing 60 (2021) 1–13.

[26] R. Li, S. Zheng, C. Duan, J. Su, C. Zhang, Multistage attention resunet for semantic segmentation of fine-resolution remote sensing images, IEEE Geoscience and Remote Sensing Letters 19 (2021) 1–5.

[27] R. Li, S. Zheng, C. Zhang, C. Duan, L. Wang, P. M. Atkinson, Abcnet: Attentive bilateral contextual network for efficient semantic segmentation of fine-resolution remotely sensed imagery, ISPRS Journal of Photogrammetry and Remote Sensing 181 (2021) 84–98.

[28] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, R. Timofte, Swinir: Image restoration using swin transformer, in: Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 1833–1844.

[29] J. Kim, J. K. Lee, K. M. Lee, Accurate image super-resolution using very deep convolutional networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 1646–1654.

[30] C. Dong, C. C. Loy, X. Tang, Accelerating the super-resolution convolutional neural network, in: Computer Vision–ECCV 2016: 14th

TABLE C1: Segmentation accuracy based on super-resolution images generated by different methods on the test set without Gaussian Blur by $\times 4$ upsampling scales. The results based on the original high-resolution images (HR) are marked in light blue, while the best result is **highlighted** and the second best is underlined.

Models	Clutter	Building	Road	Tree	Vegetation	Static Car	Moving Car	Human	mIoU
Bicubic	0.607	0.825	0.759	0.763	0.586	0.413	0.638	0.186	0.597
SRCNN [6]	0.614	0.824	**0.784**	0.765	0.590	0.377	0.680	0.220	0.607
LPKPNet [35]	0.564	0.792	0.748	0.759	0.582	0.395	0.651	0.208	0.587
NLSA [36]	0.554	0.792	0.747	0.759	0.572	0.380	0.636	0.197	0.579
SRFormer [49]	**0.625**	0.832	0.769	0.765	0.580	0.432	0.649	0.226	0.610
ConvFormerSR [50]	0.618	**0.838**	0.770	0.766	0.590	**0.512**	0.654	0.000	0.593
SwinIR [28]	0.614	0.825	0.781	**0.777**	0.604	0.465	**0.686**	0.251	0.625
LSwinSR	0.615	0.827	0.782	0.776	**0.605**	0.472	0.685	**0.252**	**0.627**
HR	0.642	0.850	0.785	0.782	0.601	0.530	0.680	0.270	0.643

TABLE C2: Quantitative comparison including PSNR, SSIM (%) and MAE (%) with different methods on the test set with Gaussian noise (mean of 0.1 and sigma of 0.01), where the best result is **highlighted** while the second best is underlined.

Models	PSNR ↑	SSIM ↑	MAE ↓
Bicubic	17.559	62.941	10.277
SRCNN [6]	19.554	53.899	14.062
LP-KPN [35]	20.808	74.868	10.115
NLSA [36]	20.823	72.777	10.288
SRFormer [49]	20.712	67.094	10.935
ConvFormerSR [50]	20.706	74.057	10.141
SwinIR [28]	20.989	**75.832**	**10.001**
LSwinSR	**20.998**	75.671	10.015

TABLE C3: Quantitative comparison including PSNR, SSIM (%) and MAE (%) with different methods on the test set with Gaussian noise (mean of 0.0 and sigma of 0.03), where the best result is **highlighted** while the second best is underlined.

Models	PSNR ↑	SSIM ↑	MAE ↓
Bicubic	22.083	60.555	4.970
SRCNN	24.280	53.729	9.078
LP-KPN	27.081	72.660	4.388
NLSA	26.941	71.496	4.537
SRFormer	27.046	66.203	4.971
ConvFormerSR	27.138	72.882	4.312
SwinIR	27.185	**73.583**	**4.208**
LSwinSR	**27.194**	73.208	4.268

European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14, Springer, 2016, pp. 391–407.

[31] M. S. Sajjadi, B. Scholkopf, M. Hirsch, Enhancenet: Single image super-resolution through automated texture synthesis, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 4491–4500.

[32] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[33] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, et al., Photo-realistic single image super-resolution using a generative adversarial network, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 4681–4690.

[34] B. Lim, S. Son, H. Kim, S. Nah, K. Mu Lee, Enhanced deep residual networks for single image super-resolution, in: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, 2017, pp. 136–144.

[35] J. Cai, H. Zeng, H. Yong, Z. Cao, L. Zhang, Toward real-world single image super-resolution: A new benchmark and a new model, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 3086–3095.

[36] Y. Mei, Y. Fan, Y. Zhou, Image super-resolution with non-local sparse attention, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 3517–3526.

[37] J. Ho, A. Jain, P. Abbeel, Denoising diffusion probabilistic models, Advances in neural information processing systems 33 (2020) 6840–6851.

[38] Y. Xiao, Q. Yuan, K. Jiang, J. He, X. Jin, L. Zhang, Ediffsr: An efficient diffusion probabilistic model for remote sensing image super-resolution, IEEE Transactions on Geoscience and Remote Sensing 62 (2024) 3341437.

[39] S. Shang, Z. Shan, G. Liu, L. Wang, X. Wang, Z. Zhang, J. Zhang, Resdiff: Combining cnn and diffusion model for image super-resolution, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 38, 2024, pp. 8975–8983.

[40] OpenAI, Gpt-4 technical report, 2023.

[41] L. Wang, S. Fang, X. Meng, R. Li, Building extraction with vision transformer, IEEE Transactions on Geoscience and Remote Sensing 60 (2022) 3186634.

[42] R. Strudel, R. Garcia, I. Laptev, C. Schmid, Segmenter: Transformer for semantic segmentation, in: Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 7262–7272.

[43] L. Wang, R. Li, C. Duan, C. Zhang, X. Meng, S. Fang, A novel transformer based semantic segmentation scheme for fine-resolution remote sensing images, IEEE Geoscience and Remote Sensing Letters 19 (2022) 1–5.

[44] F. Yang, H. Yang, J. Fu, H. Lu, B. Guo, Learning texture transformer network for image super-resolution, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 5791–5800.

[45] S. Lei, Z. Shi, W. Mo, Transformer-based multistage enhancement for remote sensing image super-resolution, IEEE Transactions on Geoscience and Remote Sensing 60 (2022) 3136190.

[46] Y. Xiao, Q. Yuan, K. Jiang, J. He, C.-W. Lin, L. Zhang, Ttst: A top-k token selective transformer for remote sensing image super-resolution, IEEE Transactions on Image Processing 33 (2024) 738–752.

[47] Y.-H. H. Tsai, S. Bai, M. Yamada, L.-P. Morency, R. Salakhutdinov, Transformer dissection: An unified understanding for transformer's attention via the lens of kernel, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2019.

[48] T. Xiao, M. Singh, E. Mintun, T. Darrell, P. Dollár, R. Girshick, Early convolutions help transformers see better, Advances in Neural Information Processing Systems 34 (2021) 30392–30400.

[49] Y. Zhou, Z. Li, C.-L. Guo, S. Bai, M.-M. Cheng, Q. Hou, Srformer: Permuted self-attention for single image super-resolution, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 12780–12791.

[50] J. Li, Y. Meng, C. Tao, Z. Zhang, X. Yang, Z. Wang, X. Wang, L. Li, W. Zhang, Convformersr: Fusing transformers and convolutional neural networks for cross-sensor remote sensing imagery super-resolution, IEEE Transactions on Geoscience and Remote Sensing 62 (2024) 3340043.

[51] Y. Lyu, G. Vosselman, G.-S. Xia, A. Yilmaz, M. Y. Yang, Uavid: A semantic segmentation dataset for uav imagery, ISPRS journal of photogrammetry and remote sensing 165 (2020) 108–119.

[52] G. Bradski, The OpenCV Library, Dr. Dobb's Journal of Software Tools.

[53] G.-S. Xia, J. Hu, F. Hu, B. Shi, X. Bai, Y. Zhong, L. Zhang, X. Lu, Aid: A benchmark data set for performance evaluation of aerial scene classification, IEEE Transactions on Geoscience and Remote Sensing 55 (7) (2017) 3965–3981.

[54] S. Lei, Z. Shi, Z. Zou, Super-resolution for remote sensing images via local–global combined network, IEEE Geoscience and Remote Sensing Letters 14 (8) (2017) 1243–1247.

[55] J. M. Haut, M. E. Paoletti, R. Fernández-Beltran, J. Plaza, A. Plaza, J. Li, Remote sensing single-image superresolution based on a deep compendium model, IEEE Geoscience and Remote Sensing Letters 16 (9) (2019) 1432–1436.

[56] S. Lei, Z. Shi, Hybrid-scale self-similarity exploitation for remote sensing image super-resolution, IEEE Transactions on Geoscience and Remote Sensing 60 (2022) 3069889.

Rui Li received a Bachelor's degree from the School of Automation Science and Engineering, South China University of Technology, Guangzhou, China, in 2019 and a Master's degree from the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, China, in 2021 and a Ph.D. degree with the Intelligent Control & Smart Energy (ICSE) Research Group, School of Engineering, University of Warwick, Coventry, UK, in 2024. He has authored more than 20 peer-reviewed articles in international scientific journals such as ISPRS P&RS, IEEE TGRS, PR, APEN, ECM and Energy. His research interests lie in trans-disciplinary applications of deep learning, especially for remote sensing, computer vision and renewable energy. He was honoured with the U.V. Helava Award (Best Paper 2022) from the International Society for Photogrammetry and Remote Sensing for his work on Vision-Transformer-based semantic segmentation published in ISPRS P&RS.

Xiaowei Zhao Xiaowei Zhao (Member, IEEE) received a Ph.D. degree in control theory from Imperial College London, London, U.K., in 2010. He was a Postdoctoral Researcher with the University of Oxford, Oxford, U.K., for three years before joining the University of Warwick, Coventry, U.K., in 2013. He is currently a Professor of Control Engineering and an Engineering and Physical Sciences Research Council Fellow with the School of Engineering, University of Warwick. His main research interests include control theory and machine learning with applications in offshore renewable energy systems, smart grids, and autonomous systems.