remote sensing

MDPI

Article

AEFormer: Zoom Camera Enables Remote Sensing Super-Resolution via Aligned and Enhanced Attention

Ziming Tu [1,2], Xiubin Yang [1,*], Xingyu Tang [1,2], Tingting Xu [3], Xi He [1,2], Penglin Liu [4], Li Jiang [4] and Zongqiang Fu [1,2]

[1] Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun 130033, China; tuziming21@mails.ucas.ac.cn (Z.T.); tangxingyu22@mails.ucas.ac.cn (X.T.); hexi21@mails.ucas.ac.cn (X.H.); fuzongqiang20@mails.ucas.ac.cn (Z.F.)
[2] Daheng College, University of Chinese Academy of Sciences, Beijing 100039, China
[3] School of Computer Science and Technology, China University of Mining and Technology, Xuzhou 221116, China; tingting_xu@cumt.edu.cn
[4] Physics Department, Changchun University of Science and Technology, Changchun 130022, China; 2022100131@mails.cust.edu.cn (P.L.); jiangli@cust.edu.cn (L.J.)
* Correspondence: yangxiubin@ciomp.ac.cn

Abstract: Reference-based super-resolution (RefSR) has achieved remarkable progress and shows promising potential applications in the field of remote sensing. However, previous studies heavily rely on existing and high-resolution reference image (Ref), which is hard to obtain in remote sensing practice. To address this issue, a novel structure based on a zoom camera structure (ZCS) together with a novel RefSR network, namely AEFormer, is proposed. The proposed ZCS provides a more accessible way to obtain valid Ref than traditional fixed-length camera imaging or external datasets. The physics-enabled network, AEFormer, is proposed to super-resolve low-resolution images (LR). With reasonably aligned and enhanced attention, AEFormer alleviates the misalignment problem, which is challenging yet common in RefSR tasks. Herein, it contributes to maximizing the utilization of spatial information across the whole image and better fusion between Ref and LR. Extensive experimental results on benchmark dataset RRSSRD and real-world prototype data both verify the effectiveness of the proposed method. Hopefully, ZCS and AEFormer can enlighten a new model for future remote sensing imagery super-resolution.

Keywords: remote sensing imagery; reference-based super-resolution; attention

Citation: Tu, Z.; Yang, X.; Tang, X.; Xu, T.; He, X.; Liu, P.; Jiang, L.; Fu, Z. AEFormer: Zoom Camera Enables Remote Sensing Super-Resolution via Aligned and Enhanced Attention. *Remote Sens.* **2023**, *15*, 5409. https://doi.org/10.3390/rs15225409

Academic Editors: Yuxuan Liu, Oktay Karakus, Li Zhang, Paul Rosin and Zhihua Hu

Received: 8 September 2023
Revised: 2 November 2023
Accepted: 13 November 2023
Published: 18 November 2023

1. Introduction

Image super-resolution aims at reconstructing a super-resolution image (SR) from a low-resolution image (LR). Since the mapping between SR and LR is not bijective, it results in countless possibilities of SR reconstruction. Although image super-resolution is a long-standing topic with history of several decades [1,2], the effectiveness of SR has been benefiting from recent deep-learning (DL) neural networks due to the rap-id-evolving computers in the past few years [3,4]. Usually, in DL-based single image super-resolution (SISR), LR is reconstructed into SR result based on pretrained model [5–7]. Although great progress has been achieved by DL-based SISR methods, it's still challenging to reconstruct the fine textures and missing details across SR imagery [8–11]. Herein, SR tasks remain challenging despite decades of researches.

To overcome the shortcoming of SISR, previous studies attempted to introduce more details from a reference image (Ref) to enrich the reconstruction, the process of which is called reference-based super-resolution (RefSR) [12–14]. In RefSR, HR texture details are extracted, aligned, then transferred from a given Ref to LR. The core difficulty and key to high-quality RefSR progress lie in the alignment and fusion problem between Ref and LR. To alleviate this problem, previous RefSR methods adopted optical-flow or spatial alignment.

These methods usually introduced a Ref, which is imaged from an-other perspective, video frame, or at different times, to reconstruct corresponding LR image. However, they have limitations and shortcomings that prevent them from being directly applicable in the field of remote sensing.

First, adopting external data as Ref results in heavy temporal and spatial redundancy. For example, RRSGAN [11], the first RefSR method in the field of remote sensing, adopted images from Google Earth as Ref and degraded images from GF-X satellite as LR. Note that images from different satellites vary in both spatial content and spectral characteristics. Pioneering though it is, RRSGAN still restricts the potential of RefSR in the field of remote sensing. Second, it's extremely hard to obtain equivalent high-quality Ref image in remote sensing practice unless significantly enhancing imaging hardware. But it's impractical to enhance hardware significantly due to limited satellite assembly space, and more importantly, of high cost [15].

To address these problems, in this study proposes a feasible approach by establishing zoom camera structure (ZCS) [16,17]. It allows simultaneous imaging of region of interest (ROI) by shifting focal length of zoom camera, thereby reducing the temporal redundancy and mitigating content irrelevance caused by different imaging times or different satellite cameras disparities. Specifically, in ZCS, camera with short long length is equipped with n$\times$ times focal length than the other camera, where n is the magnification factor of SR task. The difference of focal length between two cameras aims at capturing n$\times$ magnified image as Ref (with a resolution of $4s \times 4s$) and original image as LR (with a resolution of $s \times s$). In this way, the 4$\times$ amplified LR image, denoted as LR$\uparrow$, share the same resolution as Ref, both of which can serve as inputs to the subsequent RefSR network.

Herein, two mentioned problems are alleviated. Based on ZCS, Ref and LR images are obtained subsequently through consistent camera structure, which can dramatically reduce temporal and spatial redundancy. Besides, the zoom camera enables us to capture high-quality Ref image without significantly increasing hardware costs.

Furthermore, to achieve better RefSR performance, this study proposes a vision transformer (ViT)-based network through aligned and enhanced attention, namely AEFormer. By replacing deformable convolution (DConv) [18] with attention mechanisms [19], more spatial information across the whole image is accessible. Through the proposed aligned and enhanced attention, features of LR and Ref are utilized, aligned and fused thoroughly, which contributes to a more valid and effective SR progress. It turns out that the proposed network, AEFormer, demonstrates remarkable performance in reconstructing high-quality SR image, surpassing existing SISR networks and RefSR networks.

The main contributions of this study are summarized as follows:

(1) This study proposes a novel network for super-resolving remote sensing imagery, namely AEFormer. To the best of our knowledge, AEFormer is one of the first ViT-based RefSR networks in the field of remote sensing. Compared with existing SR networks, especially CNN-based ones, AEFormer exhibits extraordinary performance both qualitatively and quantitatively;
(2) The core advantage of AEFormer lies in the proposed aligned and enhanced attention. Due to the strong representation capability of ViT, aligned and enhanced attention represents a significant improvement to existing RefSR frameworks;
(3) The proposed ZCS is capable of enhancing the efficiency and quality of remote sensing imagery in both temporal and spatial dimensions. To the best of our knowledge, ZCS is pioneering in the field of remote sensing, which may provide insights for future satellite camera design.

2. Related Works

2.1. Single Image Super Resolution (SISR)

Single image super-resolution (SISR) aims at reconstructing SR result from LR input based on the learned end-to-end mapping between LR and high-resolution (HR) training data. SRCNN is the first DL-based method adopting a three-layer convolution neural

network (CNN) to achieve SISR [3]. The groundbreaking ResNet [20] improved the relationship between convolution layers and network effectiveness, which leads to SRResNet and other achievements in the field of SISR [21,22]. While most CNN-based networks are optimized towards minimizing mean-square-error (MSE) or mean-absolute-error (MAE), previous studies have found it not sufficient or accurate for human vision [23]. To address with this problem, generative adversarial network (GAN) offers a reliable solution by generating more photo-realistic texture, in which SRGAN is the first GAN-based SR network [21,24]. Recently, SPSR [25] proposed dual-domain encoding by adding an additional gradient domain, which contributes to a more effective feature representation process. Liu et al. introduced detail complement (DMDC) into GAN-based SR to improve the recovery capability of detailed supplyment [26]. Diffusion model was closely related to and involved in recent SR practice due to its capability generate more high-frequency information [27,28].

2.2. Reference-Based Super Resolution (RefSR)

RefSR alleviates the shortcomings of SISR by transferring more relevant details from Ref to LR [29]. Apparently, the key of RefSR lie in the alignment between LR and Ref. There are two mainstream ways for achieving alignment between LR and Ref, which are image alignment [11,29–31] and patch match [13,14,32–34]. In RefSR studies, alignment process aims at bridging the gap between LR and Ref image, and in turn obtaining aligned Ref features, which would be transferred into LR feature space during SR reconstruction process [35]. It's noteworthy that some fusion-based methods [36–38], though aimed for different tasks, can also be attributed to the mentioned transfer process, which aims at bridging the gap between target image or information and corresponding reference label for improved network performance or effect. For example, Zhou et al. proposed a multiscale feature adaptive fusion module to effectively reduce the redundancy in low-level features and background noise in S2EPC [36]. Yuan et al. proposed an enhanced fusion module for deep features from both M and RGB images via Encoder and DenseNet fusion structures with receptive fields in MCRN, which contributed to a more valid fusion than single-modal encoding [37]. Besides, Yuan et al. proposed a multi-level fusion module for global and local information, which leverages complementarity between them to generate prominent visual representation in GaLR [38].

Two typical image-alignment-based methods in RefSR tasks are optical flow [39] and deformable convolution [18]. They tend to warp the aligned parts in a flexible and rapid way. However, they prove less effective in long-distance correspondence [15]. On the other hand, patch matching-based methods prove more stable but resume higher calculation resources. SRNTT is one of the first patch-matching-based RefSR network [13], which swaps feature patches and transfers swapped patches based on pretrained VGG [40]. However, SRNTT ignores the correspondence between LR and Ref, because pretrained VGG is not in-volved in the end-to-end training. To address this problem, recent studies introduce patch-based attention to enable a learnable framework, which proves valid for most scenes yet invalid for in-patch misalignment [14]. To solve this problem, this study proposes aligned-and-enhanced attention for a more thorough patch match during alignment. Different from fusion module in previous works, this study proposes a three-level transfer module, in which each level consists of a learning mask to ensure the complete fusion effect between Ref branch and LR branch.

2.3. Vision Transformer (ViT)

The success of transformer [19] has brought unparalleled rapid development to fields like computer vision (CV) and natural language processing (NLP). In the field of CV, transformer is introduced as ViT. SwinIR is the first ViT-based SR network [41], whose performance surpassing most of previous CNN-based methods. ESRT improves SwinIR by feasibly adjusting the size of the feature map and extracting deep features with a low computational cost [42]. In fact, ESRT is a combination of convolution and transformer [43], in which the former one aimed at recognizing low-level information while the latter aimed

at exploring deeper information. Recently, transformer-based SR networks are focusing on cross-window information iteration to release further potential of ViT [44,45]. Since the above ViT-based methods are towards SISR, currently, there are few studies on ViT-based RefSR [14,46]. In this study, we aim to propose the first transformer-based RefSR network in the field of remote sensing.

2.4. Dual Camera for Super Resolution

The first demand for super-resolution is driven by on-orbit remote sensing imagery when the satellite resolution is rather poor, in which researchers incorporated two satellite camera arrays for better visual results [1]. Wilburn et al. are one of the first to achieve higher imaging quality based on multiple camera arrays [47]. In the past, non-learning-based methods [48,49] focused on searching image similarity to achieve image registration, which tend to be low-efficiency and inaccurate. CameraSR tried to reverse the latent model which was regarded responsible for degradation of camera imagery due to intrinsic tradeoff between field of view and resolution [50]. Recently, Guo et al. proposed a dual camera system to achieve low-light color imaging, which consists of a high-resolution monochromatic camera and a low-resolution color camera [51]. However, it focused on fusing spectral dynamic range, while had limited effect on high-quality super-resolution. As a matter of fact, dual camera array is a favorable structure for implementing RefSR, because dual camera guarantees a reliable implementation platform for RefSR. However, there are few previous studies extended on this topic [17].

To the best of our knowledge, we're one of the first to explore the application feasibility and potential of RefSR via zoom camera in the field of remote sensing. Different from above dual camera structures, this study only uses single camera. Considering there's limited satellite assembly space, this study utilizes only one zoom camera to achieve the functions of dual cameras by changing its focal length. In this study, ZCS is the foundation to implementing the proposed AEFormer.

3. Methodology

The proposed method follows the process of 'imaging then super-resolving', which refers to the design of zoom camera structure (ZCS) and the proposed RefSR network, namely AEFormer.

As shown in Figure 1, a zoom camera is installed and imaged towards a common region of interest (ROI). To fully estimate the effect of super resolution, the LR and Ref image are cropped from either camera imaging as shown in Figure 2, aligned with each other, and then super-resolved according to the proposed network AEFormer as shown in Figure 3. Given cameras with different focal lengths, L_1 and L_2, where L_1 is equipped with the focal length of f while L_2 is equipped with $4f$, L_1 and L_2 are imaged towards the same target. LR is cropped from imagery by L_1, with a size of $n \times n$, while Ref is cropped from imagery by L_2, with a size of $4n \times 4n$.

ZCS works following illustration in Figure 2. Constrained by the contradiction between the camera's field of view (FOV) and imaging spatial resolution, for most cases, the zoom camera works as short-focus camera L1 to expand the effective imaging field, as shown in position 1,2,3. When encountering ROI, zoom camera changes focal length and works as long-focus camera L2 to obtain optical magnified image (Ref), as shown in position 4. With LR from Pos 1,2,3 and Ref from Pos 4 as inputs, corresponding SR im-ages can be obtained through the proposed AEFormer, as shown in right side of Figure 2.

Section 3 is arranged as follows. Section 3.1 introduces aligned and enhanced attention mechanism for feature alignment process. Section 3.2 presents feature transfer based on dynamic transfer module. Section 3.3 illustrates the loss function of the proposed AEFormer.

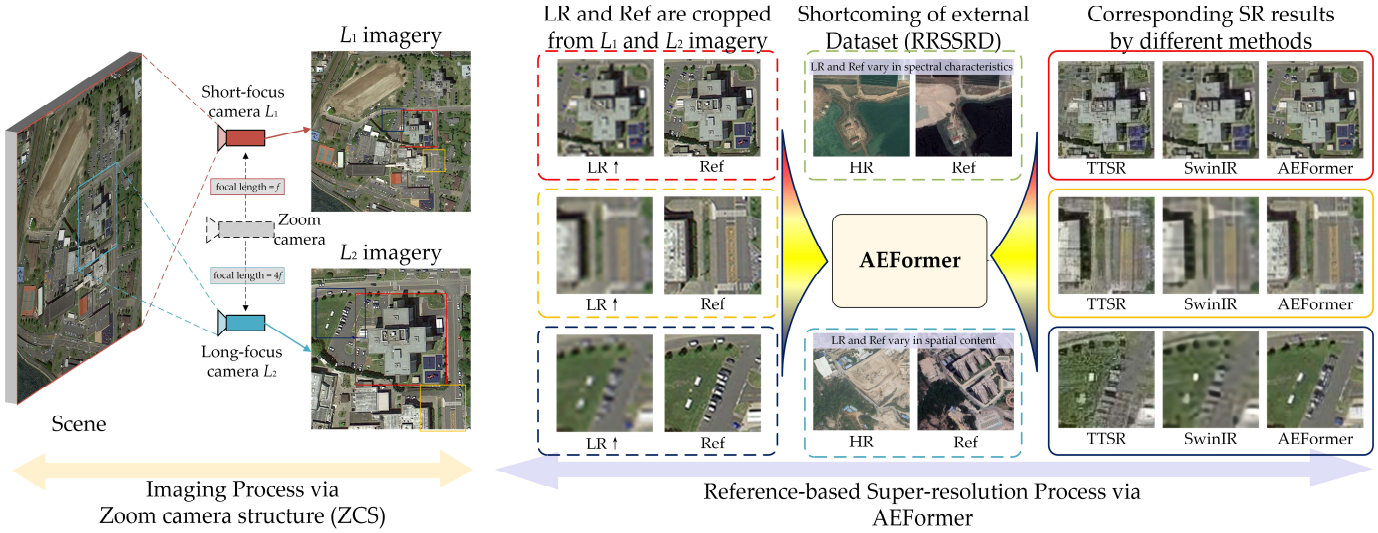

Figure 1. Overall framework of the proposed method follows the process of 'imaging then super resolution'. Imaging process is based on the proposed ZCS which is composed of a short focus camera L_1 and a long focus camera L_2. Super-resolution process is based on the proposed AEFormer which adopts LR and Ref as inputs and generates SR as an output. LR and Ref images are cropped from L_1 imagery and L_2 imagery, denoted in squares of different colors, respectively. ↑ denotes upscale ×4. SR results by TTSR [14], SwinIR [41], and our method are compared.

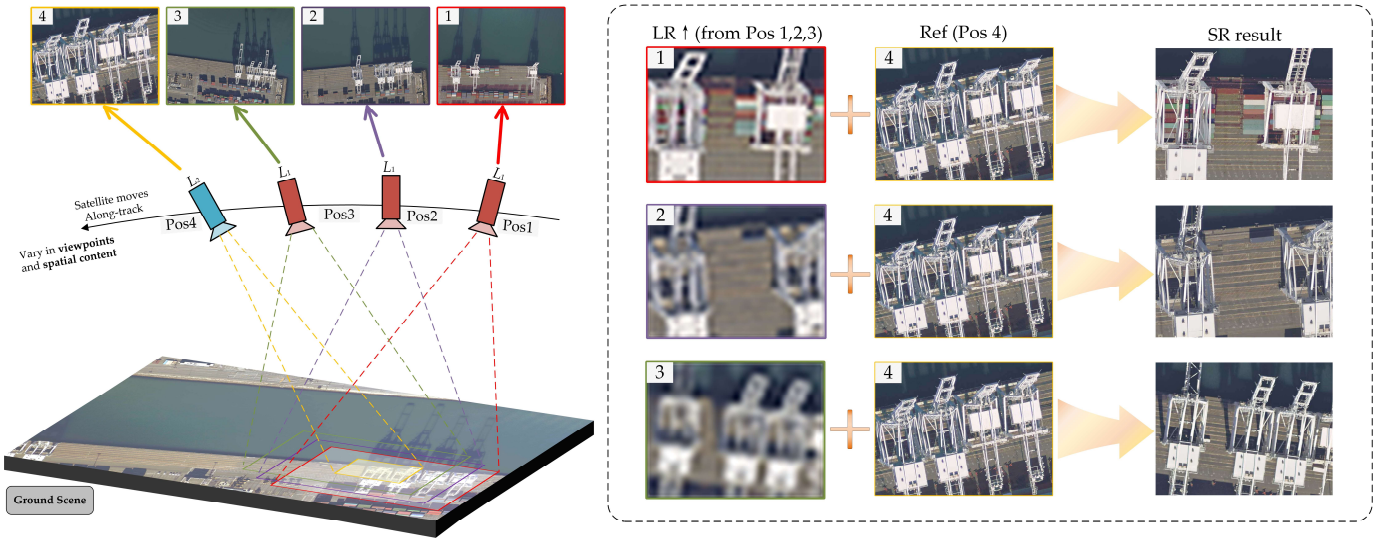

Figure 2. The working mechanism of ZCS in remote sensing practice. In most cases, to expand the imaging area, short-focus camera L_1, which is equipped with wide FOV, is used for wide-field imaging, as shown in position 1, 2, and 3. However, when it is in need for super-resolving ROI in the image captured by L_1, a zoom camera switches to long-focus camera L_2 which is equipped with narrow FOV, to capture Ref, as shown in position 4. ↑ denotes upscale ×4.

3.1. Feature Alignment Based on Aligned and Enhanced Attention

Considering the time intervals for changing the focal length of a zoom camera, there are subtle differences in the spatial content and viewpoint of images acquired by L_1 and L_2 which results in the misalignment problem between LR↑ and Ref. Addressing the misalignment is crucial for achieving a high-quality RefSR [11]. Since there is a certain similarity between patches across correlated images, the alignment strategy in the proposed network is based on the patch match [34].

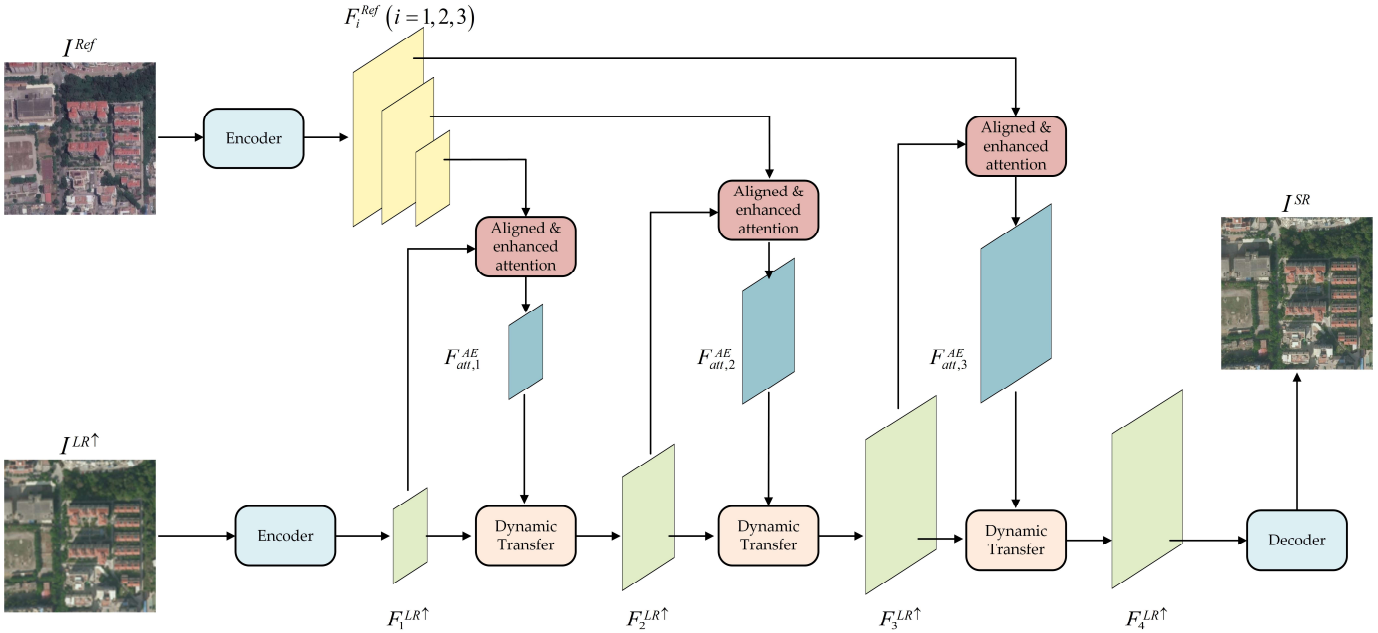

Figure 3. Overview of the proposed AEFormer. In this study, F denotes the feature space while I denotes the image space. Three levels of Ref feature F_i^{Ref} are aligned with LR via aligned and enhanced attention and then transferred to the LR feature space through dynamic transfer. For improved model generalization, LR and Ref images are selected from external dataset RRSSRD [11] which vary in spectral characteristics and spatial content. $\uparrow$ denotes upscale $\times 4$.

Different from previous patch-match-based methods where patches are swapped from non-learning process [13], the alignment strategy proposed in this study is an end-to-end ViT-based learnable process, namely of aligned and enhanced attention, as shown in Figure 4. For improved model generalization, Ref and LR, in this section, are selected from the external dataset which differ in spectral characteristics and spatial content despite being aimed for the same ROI. It corresponds to a normal RefSR verification process.

To bridge the gap between Ref and LR image, and obtain valid and effective aligned Ref features to be involved in SR reconstruction, it occurred to us that attention mechanism may be an alternative for deformable convolution [18] which is widely used in previous RefSR alignment practice [11], and the combination of feature swapping [13] and attention mechanism may lead to a reinforced aligned features acquisition. To obtain the mentioned aligned features, it takes several steps as follows.

First, considering there is a difference in content and viewpoint between LR$\uparrow$ and Ref, Ref feature map F^{Ref} is swapped according to feature swapping [13], denoted as Swap$(\cdot)$, for a rough processing. The swapped Ref feature map F_{tmp}^{Ref} is only temporary because it does not involve a learning process.

$$F_{tmp}^{Ref} = \text{Swap}\left(F^{Ref}\right) \tag{1}$$

where the role of feature swap aims at swapping features which searches over the entire I^{Ref} for locally similar textures of I^{LR} for enhanced SR reconstruction. The swapped LR and Ref patches, which may differ in color and illumination, are matched in neural feature space $\phi(\cdot)$ [13] to emphasize the structural and textural information. The similarity between both can be calculated as:

$$s_{i,j} = \left\langle P_i\left(\phi\left(I^{LR\uparrow}\right)\right), \frac{P_j\left(\phi\left(I^{Ref\downarrow\uparrow}\right)\right)}{\| P_j\left(\phi\left(I^{Ref\downarrow\uparrow}\right)\right) \|} \right\rangle \tag{2}$$

where P_i denotes sampling i-th patch from corresponding feature map. $s_{i,j}$ denotes the similarity between the i-th LR patch and the j-th Ref patch. The similarity computation S_j can be efficiently implemented as a set of convolution operations over all LR patches with each kernel corresponding to a Ref patch:

$$S_j = \phi\left(I^{LR\uparrow}\right) * \frac{P_j\left(\phi\left(I^{Ref\downarrow\uparrow}\right)\right)}{\| P_j\left(\phi\left(I^{Ref\downarrow\uparrow}\right)\right) \|} \tag{3}$$

where S_j denotes the similarity map for the j-th Ref patch, and * denotes the correlation operation. Use $S_j(x, y)$ to denote the similarity between the LR patch centered at location (x, y) and the j-th Ref patch. Based on the similarity score, a swapped feature map M can be constructed to represent texture-enhanced LR image. Each patch in M centered at (x, y) is defined as

$$P_{\omega(x,y)}(M) = P_{j^*}\left(\phi\left(I^{Ref}\right)\right), j^* = \underset{j}{\operatorname{argmax}} S_j(x, y) \tag{4}$$

where $\omega(x, y)$ maps patch center to patch index. As a result, swapped feature map M can be obtained from feature swap, as the basis for subsequent operations.

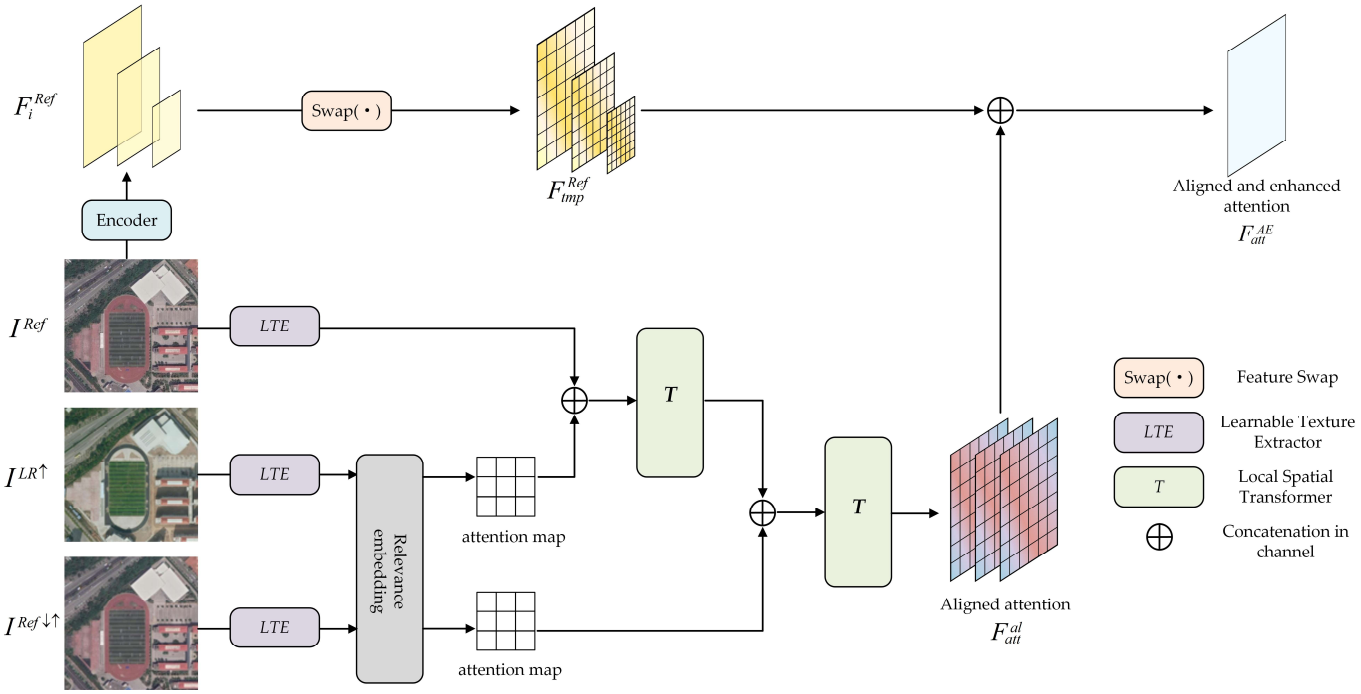

Figure 4. Aligned and enhanced attention. All image branches are transferred to feature space via the learnable texture extractor (LTE). Three branches of the feature map are aligned and concatenated into aligned attention F_{att}^{al} Finally, F_{att}^{al} is concatenated with the swapped feature map F_{tmp}^{Ref} to obtain the final aligned and enhanced attention F_{att}^{AE}. Considering the structure of the zoom camera, Ref↓↑ and LR↑ share a common image sampling frequency. In this way, the $I^{Ref\downarrow\uparrow}$ branch is added for the alignment process which is novel compared to previous alignment modules [52].

Second, for all the image branches, including Ref image I^{Ref}, LR bicubic image $I^{LR\uparrow}$, and Ref down sampled then up sampled image $I^{Ref\downarrow\uparrow}$, the learnable texture extractor (LTE) [14] is adopted as the deep feature extractor of each branch. Once feature maps of three branches are obtained from LTE, they need to be aligned. For an improved alignment effect, Ref↓↑ and LR↑, which share the same image sampling frequency, need to be aligned first, which is different from previous alignment process [52]. Specifically, the relevance between Ref↓↑ and LR↑ images are calculated by relevance embedding [14,19,32]. Then, the embedded images $I_{tmp}^{Ref\downarrow\uparrow}$ and $I_{tmp}^{LR\uparrow}$ are divided into patches p_i and q_i, in which

p_i ($i \in [1, H_{Ref\downarrow\uparrow} \times W_{Ref\downarrow\uparrow}]$) are from Ref$\downarrow\uparrow$ and q_i ($i \in [1, H_{LR\uparrow} \times W_{LR\uparrow}]$) are from LR$\uparrow$, respectively.

For each patch in $I_{tmp}^{Ref\downarrow\uparrow}$ and $I_{tmp}^{LR\uparrow}$, the relevance $r_{i,j}$ between two patches calculated by

$$\begin{aligned} r_{i,j} &= \left\langle \frac{p_i}{\|p_i\|}, \frac{q_j}{\|q_j\|} \right\rangle \\ R &= \|r_{i,j}\|^2 \end{aligned} \tag{5}$$

Third, through relevance map R obtained from Equation (5), the attention map can be achieved by incorporating the relevance map and feature map. Specifically, for Ref attention map, F_{tmp}^{Ref} and R^{Ref} are integrated to obtain Ref attention map F_{att}^{Ref}:

$$F_{att}^{Ref} = T\left(F_{tmp}^{Ref} \oplus R^{Ref}\right) \tag{6}$$

where T denotes the local transformer network [53] which aims at estimating patch-wise alignment parameters for all patches. $\oplus$ denotes concatenation in channel. Similarly, for the $LR\uparrow$ branch, F_{att}^{Ref} and $R^{LR\uparrow}$ are incorporated to obtain the aligned attention F_{att}^{al}:

$$F_{att}^{al} = T\left(F_{att}^{Ref} \oplus R^{LR\uparrow}\right) \tag{7}$$

Such aligned attention F_{att}^{al} can compensate for the weakness of a swapped patch match feature map F_{tmp}^{Ref} where nonlinear misalignment is usually hard to cope with. However, non-learning feature map F_{tmp}^{Ref} contains a basic high-similarity patch match which means combining the strength of F_{tmp}^{Ref} and F_{att}^{al} can hopefully enhance the aligned attention map F_{att}^{al}. In this way, to further enhance F_{att}^{al}, concatenate F_{tmp}^{Ref} and F_{att}^{al} to obtain the final aligned and enhanced attention F_{att}^{AE}.

$$F_{att}^{AE} = F_{tmp}^{Ref} \oplus F_{att}^{al} \tag{8}$$

Compared with DATSR [35] where feature maps are obtained through convolution and aligned attention mechanism, this study enhanced the aligned attention on the basis on swapped features.

Based on the above process, F_{att}^{AE} can be transferred to the LR feature space to obtain the SR result through the proposed dynamic transfer module (DTM) in Section 3.2.

3.2. Feature Transfer Based on the Dynamic Transfer Module

Transferring an aligned feature to the LR feature space has been a long challenging task [11,54]. Direct transfer, such as summation or concatenation, would lead to information loss or misalignment.

To address the challenging task of feature transfer, this study proposes a novel dynamic transfer module (DTM), shown as Figure 5. From what is presented in Figure 3, aligned and enhanced attention F_{att}^{AE} is transferred to the LR feature space in a multi-level way. Within each level, DTM adopts the LR feature map $F_{level}^{LR\uparrow}$ and corresponding $F_{att,level}^{AE}$ as inputs and generates $F_{level+1}^{LR\uparrow}$ as the output. The number of total levels corresponds to the number of feature encoding levels.

As shown in Figure 5, the transfer within each DTM can be divided into 3 stages. First, embed the concatenation between $F_{att,level}^{AE}$ and $F_{level}^{LR\uparrow}$ with a learnable convolutional layer which is denoted as $C_1(\cdot)$. The normalized attention map is then elementwise multiplied by the LR feature map. Second, obtain more information through another convolutional layer, which is denoted as C_2, then elementwise add (denoted as +) that with the LR feature map. Finally, to obtain the output of DTM, the above attention map goes through residual blocks to prevent degradation resulting from multiple convolutions [20].

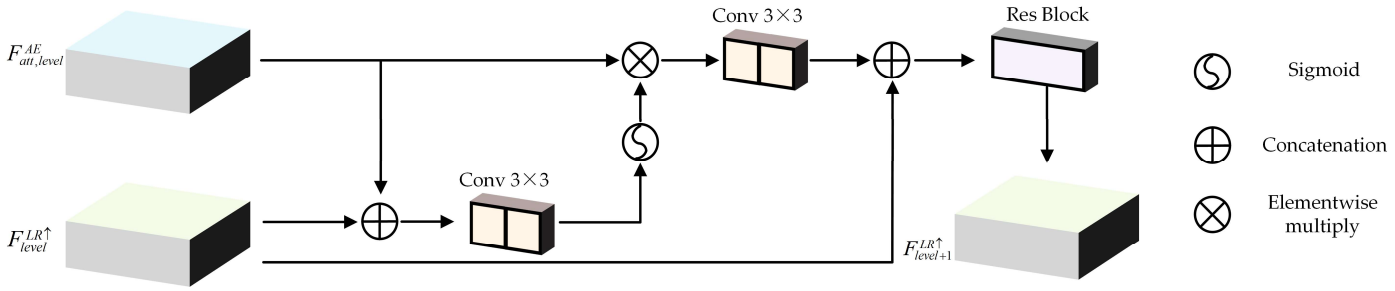

Figure 5. Dynamic transfer module (DTM). Aligned and enhanced attention F^{AE}_{att} is transferred into the LR feature space for a better transfer effect with less information loss. Given aligned and enhanced attention $F^{AE}_{att,level}$ and the LR↑ feature map $F^{LR\uparrow}_{level}$ of the same level, DTM generates the LR↑ feature map of the next level $F^{LR\uparrow}_{level+1}$ as the output.

The above feature transfer process can be represented as

$$\begin{aligned} F^{mul}_{level} &= \mathrm{Sig}\left(\mathrm{C}_1\left(F^{LR\uparrow}_{level} \oplus F^{AE}_{att,level}\right)\right) \otimes F^{LR\uparrow}_{level} \\ F^{sum}_{level} &= \mathrm{C}_2\left(F^{mul}_{level}\right) + F^{LR\uparrow}_{level} \\ F^{LR\uparrow}_{level+1} &= \mathrm{Res}\left(F^{sum}_{level}\right) \end{aligned} \tag{9}$$

where $\mathrm{C}_{1/2}(\cdot)$ denotes the learnable convolution layer with a kernel size of 3×3. $\mathrm{Sig}(\cdot)$ denotes the sigmoid operation. $\mathrm{Res}(\cdot)$ denotes residual blocks [20]. $\oplus$ denotes concatenation in the channel. $\otimes$ denotes elementwise multiplication. Level = 1, 2, and 3. Note that the output of DTM (level) equals the input of DTM (level + 1) which corresponds to Figure 3.

Furthermore, as can be seen from Figure 3 and Equation (9), $F^{LR\uparrow}_4$ can be obtained from the final level of feature transfer. With the decoder, the feature map $F^{LR\uparrow}_4$ can be transformed back to the image space, in other words, the SR result.

3.3. Loss Function

To achieve a better SR effect, the loss function in this study consists of four components which are reconstruction loss $\mathcal{L}_{rec}$, adversarial loss $\mathcal{L}_{adv}$, perceptual loss $\mathcal{L}_{per}$, and texture loss $\mathcal{L}_{txt}$. For improved clarity, the component and configuration of the loss function are different from previous studies [12–14,31].

$$\mathcal{L}_{total} = \lambda_{rec}\,\mathcal{L}_{rec} + \lambda_{adv}\mathcal{L}_{adv} + \lambda_{per}\,\mathcal{L}_{per} + \lambda_{txt}\mathcal{L}_{txt} \tag{10}$$

Reconstruction loss. To preserve the basic structure of LR-SR mapping, reconstruction loss aims at making SR infinitely approach HR (GT) during training. In this study, l_1 norm is adopted within $\mathcal{L}_{rec}$. Notably, $\mathcal{L}_{rec}$ is the most basic component of the SR training process.

$$\mathcal{L}_{rec} = \left\| I^{HR} - I^{SR} \right\|_1 \tag{11}$$

Adversarial loss. Since GAN [24] is capable of reconstructing a visually satisfactory image, adversarial loss is common yet effective in recent SR tasks [11,31]. Herein, $\mathcal{L}_{adv}$ proposed in WGAN-GP [55] is adopted in our loss function.

$$\begin{aligned} \mathcal{L}_D &= \mathbb{E}_{\tilde{x}\sim\mathbb{P}_g}[D(\tilde{x})] - \mathbb{E}_{x\sim\mathbb{P}_r}[D(x)] + \lambda\mathbb{E}_{\hat{x}\sim\mathbb{P}_{\hat{x}}}\left[(\| \nabla_{\hat{x}} D(\hat{x}) \|_2 - 1)^2\right] \\ \mathcal{L}_{adv} &= -\mathbb{E}_{\tilde{x}\sim\mathbb{P}_g}[D(\tilde{x})] \end{aligned} \tag{12}$$

where $\mathcal{L}_D$ denotes loss of the discriminator. $D(\cdot)$ denotes 1-Lipschitz functions [56]. $\mathbb{P}_r$ and $\mathbb{P}_g$ denote the distribution of the proposed model and actual situation, respectively [14].

Perceptual loss. Inspired by [13,14,57,58], perceptual loss aiming for better visual perception, in our study, is different from previous studies. Specifically, $\mathcal{L}_{per}$ consists of two parts in this study. The first part is consistent with traditional studies while the second part combines the perceptual effect of an aligned and enhanced attention map because it records certain information of certain stages during training.

$$\mathcal{L}_{per} = \frac{1}{C_i H_i W_i} \parallel \phi_i^{vgg}\left(I^{SR}\right) - \phi_i^{vgg}\left(I^{HR}\right) \parallel_2^2 + \frac{1}{C_j H_j W_j} \parallel \phi_j^{E}\left(I^{SR}\right) - F_{att}^{AE} \parallel_2^2 \quad (13)$$

where ϕ_i^{vgg} represents the feature map of VGG-19 of the i-th layer while (C_i, H_i, W_i) represents the shape of the feature map. I^{SR} denotes the super-resolved (generated) image of the corresponding iteration. ϕ_j^E denotes the feature map warped from the j-th layer of the rough alignment $E(\cdot)$ while F_{att}^{AE} denotes the aligned and enhanced attention of the corresponding iteration.

Texture loss. Following [13], texture loss, aimed at alleviating texture differences between I^{SR} and I^{Ref}, is also involved in our loss function.

$$\mathcal{L}_{txt} = \sum_l \lambda_l \parallel Gr\left(\phi_l\left(I^{SR}\right) \cdot S_l^*\right) - Gr(M_l \cdot S_l^*) \parallel_F \quad (14)$$

where $Gr(\cdot)$ computes the Gram matrix and λ_l is a normalization factor corresponding to the feature size of layer l. S_l^* denotes a weighting map for all LR patches.

4. Experiment

This section presents three aspects of our experiment. First, dataset and implementation details are presented as the basis of the experiment. Second, comparisons between state-of-the-art methods and ours are carried out for validating the effect of super-resolving. Finally, AEFormer's capability of super-resolving real-world imagery is verified based on the proposed ZCS.

4.1. Dataset and Implementation Details

Dataset. To verify the effectiveness of the proposed method and train the proposed AEFormer, benchmark dataset RRSSRD [11] is adopted in our experiment. Specifically, 4047 pairs of HR-Ref remote sensing images are used for training while four groups of images are used for testing. RRSSRD is constructed based on GF-X satellite open source, Google Earth Engine, and Microsoft Virtual Earth, which contains various remote sensing scenes, such as urban architecture, an airport, farmland, a parking lot, and more. The spatial resolution of images from RRSSRD is approximately 0.5 m.

HR and Ref within RRSSRD have the same image resolution of 480 × 480 pixels, while LR is downsampled from HR with resolution of 120 × 120 (unit in pixel). Figure 6 displays some examples of HR-Ref pairs within RRSSRD. Following standard protocol in SR tasks [4,11,21], all LR images are obtained by bicubic downsampling corresponding HR images to a $\frac{1}{4}$ size. LR↑ denotes LR upsampling four times. Ref↓↑ denotes Ref downsampling then upsampling to the original size, which aims at seeking better matching band frequency between images [11,14], as shown in Figure 4.

In Section 4.2, LR (down sampled from HR) and Ref from RRSSRD test sets are used for a classical SR verification. In Section 4.3, LR and Ref are obtained from camera L_1 and L_2 (within ZCS) for real-world SR verification.

Implementation details. For a valid verification, following some previous known works [8,11,14,29], scale factor is set to be large as 4 in this study. For improved clarity, state-of-the-art methods are compared with the proposed AE-Former, including CNN-based SISR method EDSR [22], GAN-based SISR method SPSR [25], CNN-based RefSR method CrossNet [12], GAN-based RefSR method SRNTT [13], ViT-based SISR method SwinIR [41], ViT-based RefSR method TTSR [14]. For fair comparison, all methods are configured via the default provided by their authors to achieve their best performance

and trained for a consistent or close iteration. AEFormer is trained for 200,000 iterations to achieve convergence, which took about 53 h on 2 × Nvidia RTX 4090. SRNTT was implemented on TensorFlow while others were implemented on PyTorch. Inspired by some previous known works [11,15,25,41,52,59,60], hyperparameter configuration in our study, which are ($\mathcal{L}_{rec}$, $\mathcal{L}_{per}$, $\mathcal{L}_{adv}$, $\mathcal{L}_{txt}$) in Equation (10), are set as 1, 1×10^{-3}, 1.5×10^{-7}, and 1×10^{-7}, respectively. Quantitative comparison between state-of-the-art methods and AEFormer in Section 4.2 are evaluated in terms of LPIPS [61], PSNR, SSIM, and FID [62–64], all of which compare SR with HR (GT) to calculate the corresponding evaluation metrics. For real-world experimentation in Section 4.3, SR quality is evaluated by NIQE [65] and PI [23], both of which are non-reference image quality evaluation metrics [66]. Besides, the metrics for evaluating the diversity of GAN capability, IS, is also adopted in the comparison study [64].

Figure 6. Some examples of HR-Ref pairs within the benchmark dataset RRSSRD [11]. HR images are selected from WorldView-2 (2015) and GF-2 (2018) with a spatial resolution of 0.5 m while Ref images are selected from Google Earth Engine (2019) with consistent or close spatial resolution.

$$\begin{aligned} \mathrm{LPIPS}(I_{SR}, I_0) &= \sum_l \frac{1}{H_l W_l} \sum_{h,w} \left\| w_l \odot \left(\hat{y}^l_{hw} - \hat{y}^l_{0_{hw}} \right) \right\|_2^2 \\ \mathrm{PI} &= \tfrac{1}{2}((10 - \mathrm{Ma}) + \mathrm{NIQE}) \\ \mathrm{FID}(\mathbb{P}_r, \mathbb{P}_g) &= \| \mu_r - \mu_g \| + \mathrm{Tr}\left(\mathbf{C}_r + \mathbf{C}_g - 2(\mathbf{C}_r \mathbf{C}_g)^{1/2} \right) \\ \mathrm{IS}(P_g) &= e^{E_{x \sim P_g}[KL(p_M(y|x) \| p_M(y))]} \end{aligned} \tag{15}$$

where H_l and W_l represent the height and width of l-th layer. $\hat{y}^l_{hw}$ and $\hat{y}^l_{0_{hw}}$ indicate the features at the specific location (h, w) of l-th layer from the generated image and the ground truth image. w_l is a supervised weight vector. $\odot$ denotes elementwise multiply. Tr($\cdot$) represents the sum of elements on the diagonal of a matrix. PI, as the non-reference metrics, can be calculated by incorporating the criteria of Ma [67] and NIQE [65]. It indicates the perception quality of the generated image. More details in the generated image can lead to a better PI result, which is lower in score. In FID equation, the distance between these two univariate Gaussian distributions is calculated using mean and variance, in which r denotes

the ground-truth image, while g denotes the generated image. A lower FID means that the two distributions are closer, which means that the quality and diversity of the generated images are higher. Lastly, in IS equation, the conditional probability $P(y \mid x)$ is expected to be highly predictable (low entropy) for GAN. For example, given an image, the object type should be known easily. In turn, an Inception network is used to classify the generated images and predict $P(y \mid x)$, where y is the label and x is the generated data. This reflects the quality of the images. Besides, if the generated images are diverse, the data distribution for y should be uniform (high entropy) [64]. Finally, to combine these two criteria (quality and diversity), their KL-divergence can be computed and the equation can be used to obtain IS score. Based on previous knowledge [63] and experimental results below, the biggest difference between FID and IS is that FID focuses more on image similarity, while IS focuses more on data diversity especially for GAN methods.

4.2. Comparison with State-of-the-Art Methods

In this section, both qualitative and quantitative comparisons are carried out to fully estimate the performance of the proposed AEFormer.

Qualitative comparison. To verify the visual quality of SR results, AEFormer is compared to state-of-the-art methods. As shown in Figure 7, AEFormer elicits the best visual quality on the dis-played test sets. It's also observable that, by enriching details from Ref, RefSR methods are more visually satisfactory than SISR methods. Although Ref and HR differ in viewpoints, spectral characteristics, and spatial content within RRSSRD, AEFormer successfully utilizes Ref and LR to reconstruct the best SR effect among selected methods. Specifically, in the first set in Figure 7, AEFormer retains the best roof details while suppressing noise across whole image (compared to TTSR). In the second set, only AEFormer recovers the horizontal structures. In the third set, SPSR and TTSR generate massive artifacts and blurriness on the scaffolding, while other methods even fail to distinguish between multiple scaffolding elements. Only AEFormer succeeds in distinguishing them.

It's observable that SwinIR, which is trained with reconstruction only, achieves the second best PSNR and SSIM scores, only second to AEFormer_rec (AEFormer trained from scratch with reconstruction loss only). Based on previous known studies [11,13,14], it's commonly known that SR network trained with reconstruction loss (e.g. l_1 loss) can lead to better PSNR and SSIM scores, because they're usually oriented towards MAE or MSE. However, higher PSNR and SSIM don't guarantee a better visual result because over-smooth texture can lead to higher PSNR, this is why we need different loss functions and additional metrics to evaluate its effectiveness.

Quantitative comparison. Table 1 shows the quantitative comparison on RRSSRD in terms of LPIPS, PI, NIQE, FID, IS, PSNR, and SSIM. Red bold score denotes the 1st best result, while blue bold score denotes the second best result. Considering the combination of different loss functions con-tribute to a more photo-realistic details with slightly worse PSNR and SSIM score [11,21], AEFormer is trained from scratch with all losses, denoted as AEFormer. For fair comparison, AEFormer is additionally trained with reconstruction loss only to achieve higher PSNR and SSIM score, with adversarial loss, perceptual loss and texture loss removed, which is denoted as AEFormer_rec.

It's observable that AEFormer outperforms the second best method, SwinIR [41], 53.28%, 1.98%, 2.14% in terms of average LPIPS, PNSR, and SSIM. It further reiterates the superiority of AEFormer. Considering there are many indicators in this section, some of which are lower, the better, whereas for others is the opposite.

For more intuitive com-parison, improvement percentage in this study is defined as

$$\text{Improvement Percentage} = \frac{|Indicator(Method) - Indicator(Bic)|}{Indicator(Bic)} \times 100\% \quad (16)$$

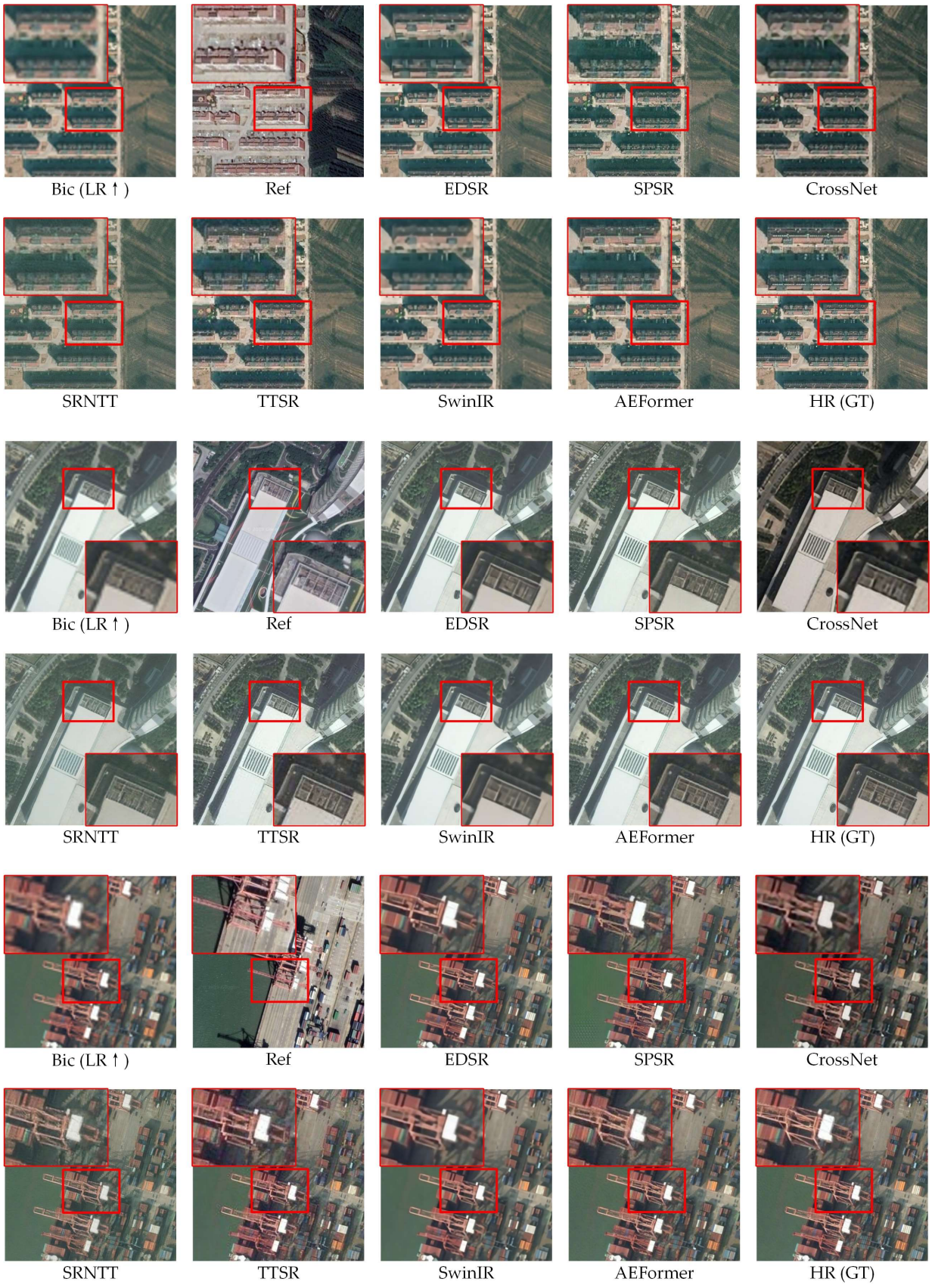

Figure 7. Visual comparison of SR images between the selected methods and ours. Zoom in for better visualization. Bic denotes the bicubic image of LR which is also denoted as LR↑ in the field of super resolution. HR denotes high-resolution imagery which is also known as ground truth (GT).

Table 1. Average LPIPS, PI, NIQE, FID, IS, PSNR, and SSIM scores of selected SR methods of ×4 factor on different test sets. For LPIPS, PI, NIQE and FID, a lower score indicates a better result whereas for IS, PSNR and SSIM, a higher score indicates a better result. The best results are **highlighted** in **red** (first best) and **blue** (second best).

Test Set	Metrics	Bicubic	EDSR [22]	SPSR [25]	CrossNet [12]	SRNTT [13]	TTSR [14]	TTSR_rec [14]	SwinIR [41]	AEFormer (Ours)	AEFormer _rec (Ours)
1	LPIPS	0.3667	0.1688	0.1723	0.2897	0.2004	0.1944	0.2323	0.2351	**0.1035**	**0.1639**
	PI	7.1020	5.1290	**3.2918**	5.3949	3.7471	3.4192	5.7505	6.2364	**3.2000**	5.5191
	NIQE	7.7933	5.6877	**4.2980**	6.0194	4.1205	4.1673	6.2667	6.9530	**4.2309**	6.0150
	FID	126.0413	89.9880	**75.6160**	90.0235	89.4061	89.2271	97.1016	123.0567	**39.1042**	79.7634
	IS	1.9299	**2.0919**	2.0071	1.9316	2.0796	1.9836	2.0115	2.0883	1.9882	**2.0982**
	PSNR	29.6840	32.2835	30.4216	31.0801	29.1221	30.9589	32.8617	**33.4122**	32.4519	**34.2224**
	SSIM	0.7914	0.8750	0.7908	0.8590	0.7977	0.7957	0.8558	**0.8758**	0.8470	**0.8915**
2	LPIPS	0.3920	0.2704	0.2105	0.3093	0.2348	0.2110	0.2400	0.2681	**0.1340**	**0.1960**
	PI	7.0139	4.9010	**3.0975**	5.2704	3.7498	3.1980	5.8804	6.2632	**3.1927**	5.6102
	NIQE	7.6505	5.5122	**4.0160**	5.9580	4.1104	4.0012	6.5053	6.9457	**4.1917**	6.2035
	FID	127.7582	104.1339	98.8441	113.6950	115.4572	99.1046	107.9276	125.6784	**47.9718**	**83.3458**
	IS	2.0822	2.0836	2.0543	1.9980	**2.1643**	2.0933	2.1141	**2.1376**	2.1200	2.1272
	PSNR	29.5621	31.1646	29.2977	31.0895	27.9063	29.9981	32.1126	**32.5679**	31.6659	**33.2675**
	SSIM	0.7638	0.8319	0.7446	0.8295	0.7727	0.7543	0.8275	**0.8406**	0.8119	**0.8615**
3	LPIPS	0.4748	0.2712	0.2040	0.3266	0.2426	**0.2212**	0.3210	0.3491	**0.1414**	0.2874
	PI	7.0493	5.0103	**2.8630**	5.1190	3.9161	3.0213	5.8568	6.5596	**2.8314**	5.7783
	NIQE	7.6804	5.4921	**3.9324**	5.1493	4.3254	3.9501	6.3419	7.3428	**3.9168**	6.2950
	FID	138.0656	101.5913	**73.8934**	106.9849	80.1084	75.8917	99.1774	127.0837	**43.2799**	98.4504
	IS	1.9127	1.9825	2.0001	1.9197	**2.1088**	2.0276	2.0170	1.9656	**2.0359**	1.9612
	PSNR	27.7658	29.4900	28.0084	28.9832	27.8890	28.7439	30.5491	**30.9032**	29.6163	**31.3204**
	SSIM	0.7275	0.8071	0.7082	0.7897	0.7472	0.7281	0.7997	**0.8117**	0.7646	**0.8282**
4	LPIPS	0.3749	0.2807	0.2427	0.3216	0.2574	0.2326	0.2470	0.2817	**0.1509**	**0.2143**
	PI	7.1793	5.3996	**3.3195**	5.4073	4.2947	3.4277	6.3508	6.5094	**3.5329**	6.1760
	NIQE	7.7796	5.6910	**4.1089**	5.8788	4.5599	4.0359	6.8673	7.2425	**4.2829**	6.6927
	FID	123.8509	98.8013	104.6603	112.0716	112.9005	99.0680	100.5237	124.4605	**57.2116**	**88.4482**
	IS	2.1570	2.1563	2.1753	2.0994	2.1225	**2.1860**	2.1172	2.1140	**2.1787**	2.1158
	PSNR	30.0465	31.5246	29.1996	30.1077	29.2247	30.1338	32.4450	**32.9487**	31.9228	**33.5955**
	SSIM	0.7593	0.8233	0.7231	0.7690	0.7687	0.7394	0.8184	**0.8310**	0.7962	**0.8496**
Average	LPIPS	0.4021	0.2478	**0.2074**	0.3118	0.2338	0.2148	0.2601	0.2835	**0.1325**	0.2154
	PI	7.0861	5.1100	**3.1430**	5.2979	3.9269	3.2666	5.9596	6.3922	**3.1893**	5.7709
	NIQE	7.7260	5.5958	**4.0888**	5.7514	4.2791	4.0386	6.4953	7.1210	**4.1556**	6.3016
	FID	128.9290	98.6286	88.2535	105.6938	99.4681	90.8229	101.1826	125.0698	**46.8919**	**87.5020**
	IS	2.0205	2.0786	2.0592	1.9872	**2.1188**	2.0726	2.0650	2.0764	**2.0807**	2.0756
	PSNR	29.2646	31.1157	29.2318	30.3151	28.5355	29.9587	31.9921	**32.4580**	31.4142	**33.1015**
	SSIM	0.7605	0.8343	0.7417	0.8118	0.7716	0.7544	0.8254	**0.8398**	0.8049	**0.8577**

For each indicator in Table 1, denote the improvement of each average indicator of the compared methods by calculating corresponding improvement percentage. The higher the percentage is, the more effective corresponding method is. The graphical result of improvement percentage is shown in Figure 8. It's observable that the pro-posed AEFormer (and AEFormer_rec) can elicit the best improvement percentage in most circumstances.

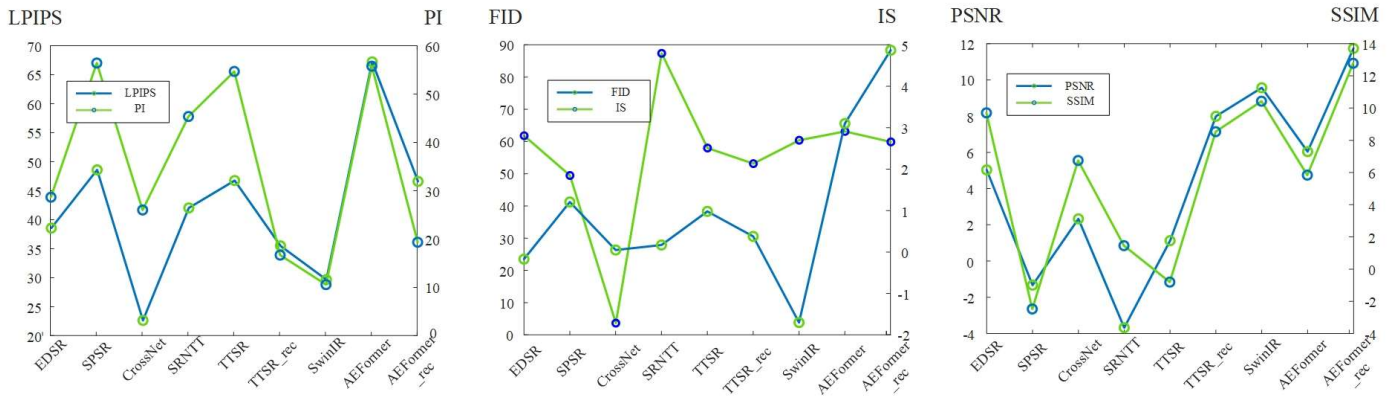

Figure 8. Improvement percentage (corresponding method compared to bicubic imagery) of each indicator in Table 1. The higher improvement percentage is, the better corresponding method is.

4.3. ZCS for Real-World Super Resolution

The above experiment adopts degraded images (LR down sampled from HR) as LR to verify the effectiveness of AEFormer. This typical super-resolution verification differs from super-resolving real-world data because real-world LR data are not obtained through down sampling. Since there is no equivalent corresponding HR (GT) of real-world LR data, the quality of SR results cannot be evaluated by previous metrics which are LPIPS, PSNR, and SSIM. Instead, they should be evaluated by PI [23] and NIQE [65], both of which are non-reference image quality evaluation metrics [66]. In this section, SR results of real-world Orbita satellite data [68] are verified.

The experimental process of super-resolving real-world data based on ZCS also follows the flowchart outlined in Figure 1. Evidently, as shown in Figure 9, short-focus camera L_1 is aimed for capturing the wide FOV image where LR images are cropped from. Long-focus camera L_2, with limited and narrow FOV, is aimed for obtaining Ref images, which share the same image resolution as LR↑. In this section, both LR↑ and Ref have the same image resolution as 600 × 600 (unit in pixel).

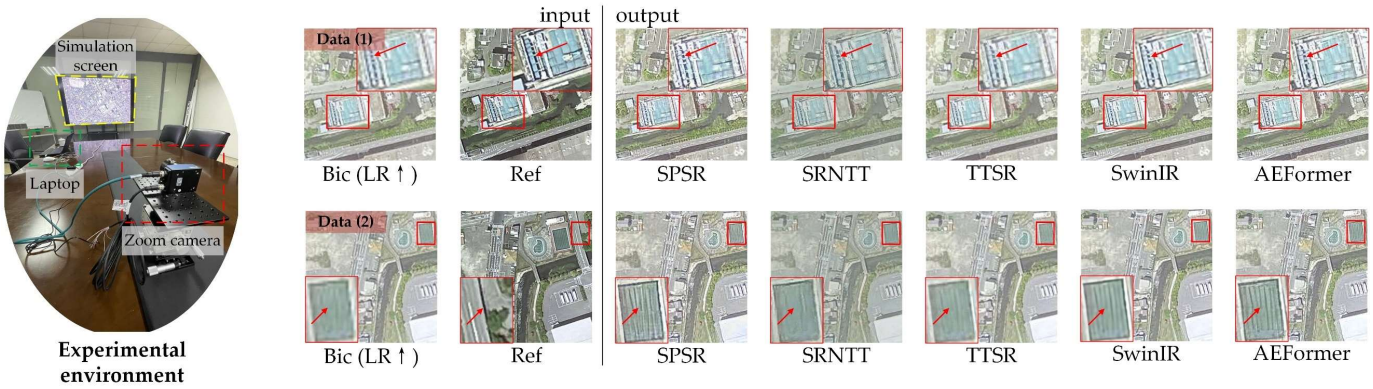

Figure 9. Real-world imagery super resolution based on ZCS. The arrow points to the main differences among compared methods. Zoom in for better visualization.

Although Orbita satellite data are not involved in our training, the SR result of AEFormer still elicits the best performance both qualitatively and quantitatively when compared with the selected state-of-the-art methods.

As shown in Figure 9, the SR image of SPSR is full of noise and artifacts. On one hand, the SR image of SRNTT lacks real image intensity despite achieving the second best scores. On the other hand, SwinIR, trained only on reconstruction loss, achieves low scores when facing real-world imagery super resolution, which is different from the previous section. It proves the necessity of a combination of different loss functions, especially perceptual loss and adversarial loss. Most importantly, the SR image of AEFormer shows the most detailed contents and achieves the best scores in terms of PI and NIQE (Table 2). It demonstrates the robustness and effectiveness of the proposed method.

Table 2. Evaluation of Real-world imagery SR. Both PI and NIQE are non-reference image quality evaluation metrics. For PI and NIQE, a lower score indicates better. The best results are **highlighted in bold**.

Methods	Data (1)		Data (2)	
	PI	NIQE	PI	NIQE
SPSR [25]	3.508	3.3052	3.0197	3.1995
SRNTT [13]	3.5142	3.9132	3.1793	3.8485
TTSR [14]	5.6652	5.8128	5.3023	5.5889
SwinIR [41]	6.4644	6.9157	6.4344	7.0353
AEFormer	**2.8890**	**3.5367**	**2.9647**	**3.6203**

5. Discussion

In this section, three key points of the proposed method, which are ZCS, aligned and enhanced attention, and dynamic transfer module, are discussed. Specifically, the implementation process of ZCS is discussed. Its current effectiveness and future development are detailed. Additionally, ablation studies are carried out on the aligned and enhanced attention and dynamic transfer module to demonstrate their effectiveness.

5.1. Effectiveness and Limitation of ZCS

As introduced in Section 2.2, the performance of RefSR greatly depends on the alignment between the LR and Ref image. In other words, it also indicates that the quality of the Ref image could possibly affect the performance of RefSR. In the proposed ZCS, zoom camera L_2 is equipped with four times the focal length of L_1 to obtain the magnified yet same resolution Ref image as LR↑. What would happen when L_2 is equipped with different focal lengths? Moreover, what would happen when an irrelevant image or external dataset is adopted as a Ref for the RefSR process?

To verify the necessity and effectiveness of ZCS, an ablation study is carried out on the quality of Ref which is photographed by zoom camera L_2. Specifically, to address the above questions, two types of experiments are conducted. First, adopt a different focal length to obtain a different Ref image, which captures different regions from LR↑ or the previous Ref. Second, adopt an irrelevant image or external dataset as the Ref. By comparing corresponding SR results based on the above Ref both qualitatively and quantitatively, a comprehensive conclusion about the effectiveness of the proposed ZCS can be arrived at.

As shown in Figure 10 and Table 3, different ZCS configurations result in obtaining different Ref images, in turn leading to different SR results. By evaluating corresponding SR results, the effectiveness of ZCS can be estimated. Specifically, given a fixed L_1 configuration (f), different ZCS configurations include L_2 imaging with focal lengths of f, $2f$, and $4f$. Moreover, irrelevant imagery and external datasets (from Google Earth Engine 2023) are also involved in ZCS configurations.

Table 3. Evaluation of SR results according to different ZCS configuration. For PI and NIQE, a lower score indicates better. The best results are **highlighted in bold**.

ZCS Configuration	Data (1)		Data (2)	
	PI	NIQE	PI	NIQE
Irrelevant image as Ref	2.9043	3.5912	2.9829	3.6536
Google Earth as Ref	2.9959	3.7749	**2.8806**	**3.5593**
Ref with focal length = f	3.1877	3.7340	2.9960	3.6807
Ref with focal length = $2f$	2.9189	3.5839	2.9810	3.6476
Ref with focal length = $4f$	**2.8890**	**3.5367**	2.9647	3.6203

It can be concluded from Figure 10 and Table 3 that adopting Google Earth data and L_2 $4f$ imaging as Ref can possibly lead to the best SR result while adopting L_2 f imaging or irrelevant imagery as the Ref leads to a poor SR result. This is possibly due to the high-frequency details in Ref because there are many high-frequency details in Ref from L_2 $4f$ imaging whereas there are few in Ref from L_2 f imaging. Since the external dataset may vary significantly from LR in spatial content or other issues, ZCS for capturing ahigh-quality Ref proves feasibly accessible and practically effective for contributing to a better RefSR performance. In remote sensing practice, determining which Ref, from 4f ZCS or Google Earth, to use depends on data accessibility.

In future remote sensing practice, common path system [69,70] can be introduced into satellite camera design for shortening the interval time of the changing focal length.

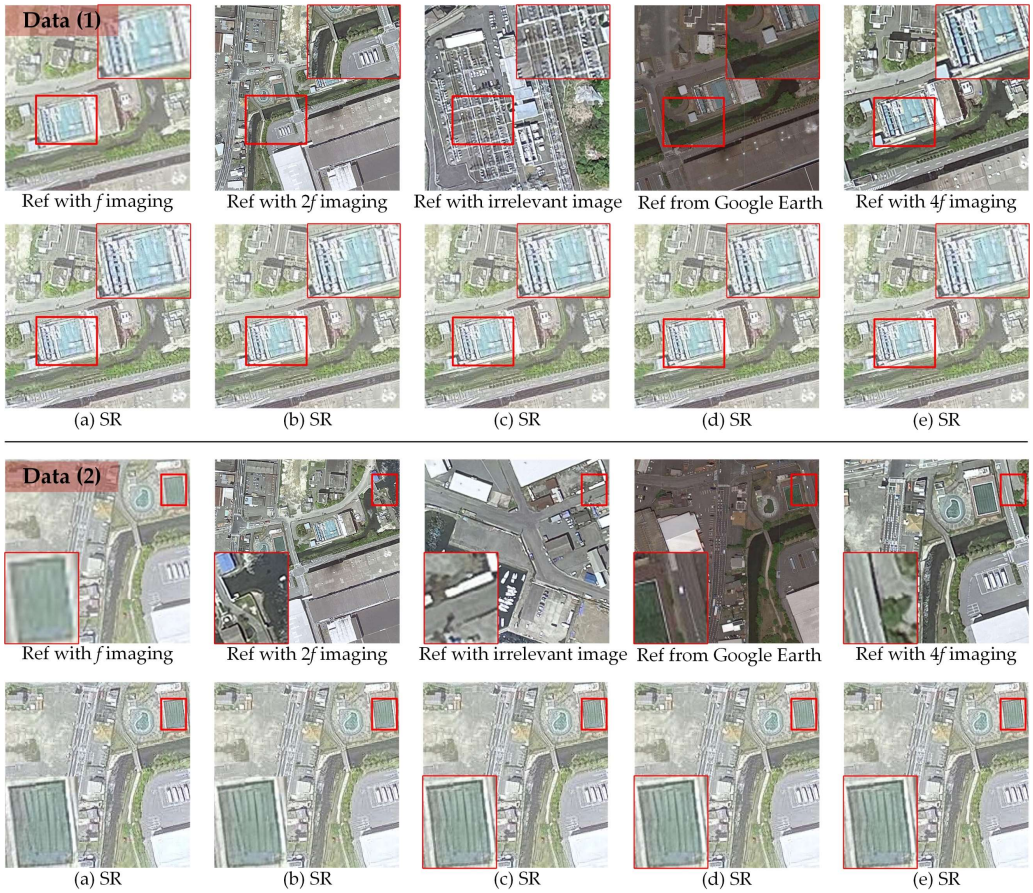

Figure 10. SR results according to different ZCS configuration. The first row displays different Ref images obtained by different ZCS configuration, while the second row displays corresponding SR results. Given fixed L1 configuration (L1 imaging with f), changes in ZCS include: (**a**) L2 imaging with f; (**b**) L2 imaging with 2f; (**c**) irrelevant imagery; (**d**) external data (from Google Earth Engine 2023); (**e**) Real ZCS configuration (L2 imaging with 4f).

5.2. Effectiveness of Aligned and Enhanced Attention

In this study, the remarkable performance of the proposed AEFormer is attributed to the configuration of Ref↓↑ branch, aligned attention, and aligned and enhanced attention, all of which are regarded as improvements from previous studies [14,25,31,41,71]. To verify the necessity and importance of these modules, ablation study is carried out on each mod-ule, as shown in Table 4.

Notably, 'Ref↓↑ branch', 'Aligned attention', and 'Enhanced attention' in Table 4 are not complementary. For 'Ref↓↑ branch', × denotes removing the Ref↓↑ branch in Figure 4 while √ denotes reserving this branch. 'Aligned attention' is the basis of 'Enhanced attention', which means the existence of enhanced attention depends on the existence of aligned attention in Figure 4.

It can be seen from the first row and second row within Table 4 that the existence of aligned attention contributes to an improvement among all scores significantly where LPIPS, PSNR, and SSIM are improved by 13.8%, 3.27%, and 2.70%, respectively. Moreover, given a considerable improvement among all metrics from the second and third row, it verifies the necessity of the Ref↓↑ branch. Furthermore, based on the third row and last row, the existence of aligned and enhanced attention leads to further improvements in LPIPS and SSIM despite a slight and acceptable decrease in PSNR. It indicates that the proposed aligned and enhanced attention fully exploit and utilize the feature space to release the potential for a more effective alignment process.

Table 4. Ablation study on the Ref↓↑ branch; aligned and enhanced attention. Ref↓↑ branch denotes the existence of Ref↓↑ in Figure 4 while aligned attention and enhanced attention denotes correspondence in Figure 4. Evaluation metrics are estimated on the first test set from RRSSRD. For LPIPS, a lower score indicates a better result whereas for PSNR and SSIM, a higher score indicates a better result. The best results are **highlighted in bold**.

Module			Evaluation Metrics		
Ref↓↑ Branch	**Aligned Attention**	**Enhanced Attention**	**LPIPS**	**PSNR**	**SSIM**
×	×	×	0.1580	30.9693	0.8171
×	√	×	0.1362	31.9833	0.8392
√	√	×	0.1139	**32.4702**	0.8396
√	√	√	**0.1035**	32.4519	**0.8470**

5.3. Effectiveness of Dynamic Transfer Module

To further verify the effectiveness of transfer process (in this study in DTM), ablation study is carried out on transfer module. There're many fusion-based and transfer-based methods currently [36–38]. Different from these studies, the proposed DTM consists of sigmoid module and convolution module, while DTM is used in each level of aligned Ref features transferring. In this way, ablation study is carried out on sigmoid part (for generating a learning mask) and convolution part (for extracting and rein-forcing features). It's observable from Table 5 that removing sigmoid part leads to approximately 13.5% decrease in terms of LPIPS, 4.4% decrease in PSNR, and 5.5% decrease in SSIM. It's also observable that removing convolution part leads to a slightly mild decrease. It verifies the necessity and effectiveness of the components within DTM.

Table 5. Ablation study of transfer module. The first row denotes features concatenation in channel without further operations. The last row denotes the proposed DTM. For LPIPS and PI, a lower score indicates better, whereas for PSNR and SSIM, a higher score indicates better. Evaluation metrics are estimated on the first test set from RRSSRD. The best results are **highlighted in bold**.

Module within Transfer Module		Evaluation Metrics			
Sigmoid Module	Convolution Module	LPIPS	PI	PSNR	SSIM
×	×	0.2495	6.2780	32.3798	0.8320
×	√	0.1896	6.1993	32.7164	0.8426
√	×	0.2098	6.3817	32.9354	0.8371
√	√	**0.1639**	**5.5191**	**34.2224**	**0.8915**

6. Conclusions

This study presents a novel method for achieving better RefSR performance in the field of remote sensing by exploring a novel imaging mode, namely ZCS, and a novel algorithm, namely AEFormer, in which ZCS serves as the structural basis for implementing AEFormer. On one hand, ZCS utilizes the magnification performance of the zoom camera to obtain a high-quality Ref image with the least temporal and spatial redundancy. On the other hand, AEFormer, with highlights in aligned and enhanced attention and dynamic transfer, achieves state-of-the-art performance among the selected SISR and RefSR methods. Aligned and enhanced attention prove superior to previous alignment modules which should be an enlightenment for future alignment module design. This study, with its blend of theoretical innovation and engineering applicability, proves potentially impactful for future remote sensing imaging.

In the future, efforts will be made towards optimizing the model, for example, introducing a semi-supervised mechanism into the loss function for improved learning effectiveness.

Author Contributions: Conceptualization, Z.T. and X.Y.; methodology, Z.T. and Z.F.; software, Z.T.; validation, Z.T., X.T., X.H. and T.X.; formal analysis, Z.T., Z.F. and X.T.; investigation, Z.T. and X.T.; resources, Z.T., P.L. and L.J.; data curation, Z.T. and Z.F.; writing—original draft preparation, Z.T.; writing—review and editing, Z.T., Z.F., T.X., X.H. and X.T.; visualization, Z.T.; supervision, Z.T. and X.Y.; project administration, Z.T.; funding acquisition, X.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by Key Research and Development Program of Jilin Province under Grant 20230201061GX, in part by Natural Science Foundation of Jilin Province under Grant 20210101099JC, in part by National Natural Science Foundation of China under Grant 62171430, in part by National Natural Science Foundation of China under Grant 62101071, in part by Entrepreneurship Team Project of Zhuhai City under Grant ZH0405190001PWC.

Data Availability Statement: The data of experimental images used to support the findings of this research are available from the corresponding author upon reasonable request.

Acknowledgments: The authors are sincerely grateful for the constructive comments and suggestions of the manuscript reviewers.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this paper:

Abbreviation	Full Name
SR	Super resolution
LR	Low resolution
Ref	Reference (image)
HR	High resolution
GT	Ground truth
ViT	Vision transformer
LTE	Learnable texture extractor
ZCS	Zoom camera structure
FOV	Field of view
SISR	Single-image super resolution
Ref-SR	Reference-based super resolution
CNN	Convolutional neural network
GAN	Generative adversarial network
AEFormer	Reference-based super-resolution network via aligned and enhanced attention

References

1. Tsai, R.Y.; Huang, T.S. Multiframe image restoration and registration. *Multiframe Image Restor. Regist.* **1984**, *1*, 317–339.
2. Zhang, H.; Yang, Z.; Zhang, L.; Shen, H. Super-Resolution Reconstruction for Multi-Angle Remote Sensing Images Considering Resolution Differences. *Remote Sens.* **2014**, *6*, 637–657. [CrossRef]
3. Dong, C.; Loy, C.C.G.; He, K.M.; Tang, X.O. Learning a Deep Convolutional Network for Image Super-Resolution. In Proceedings of the 13th European Conference on Computer Vision (ECCV), Zurich, Switzerland, 6–12 September 2014; pp. 184–199.
4. Dong, C.; Loy, C.C.; He, K.M.; Tang, X.O. Image Super-Resolution Using Deep Convolutional Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *38*, 295–307. [CrossRef] [PubMed]
5. Zhao, M.H.; Ning, J.W.; Hu, J.; Li, T.T. Hyperspectral Image Super-Resolution under the Guidance of Deep Gradient Information. *Remote Sens.* **2021**, *13*, 2382. [CrossRef]
6. Xu, Y.Y.; Luo, W.; Hu, A.N.; Xie, Z.; Xie, X.J.; Tao, L.F. TE-SAGAN: An Improved Generative Adversarial Network for Remote Sensing Super-Resolution Images. *Remote Sens.* **2022**, *14*, 2425. [CrossRef]
7. Guo, M.Q.; Zhang, Z.Y.; Liu, H.; Huang, Y. NDSRGAN: A Novel Dense Generative Adversarial Network for Real Aerial Imagery Super-Resolution Reconstruction. *Remote Sens.* **2022**, *14*, 1574. [CrossRef]
8. Wang, P.J.; Bayram, B.; Sertel, E. A comprehensive review on deep learning based remote sensing image super-resolution methods. *Earth-Sci. Rev.* **2022**, *232*, 25. [CrossRef]
9. Singla, K.; Pandey, R.; Ghanekar, U. A review on Single Image Super Resolution techniques using generative adversarial network. *Optik* **2022**, *266*, 31. [CrossRef]
10. Qiao, C.; Li, D.; Guo, Y.T.; Liu, C.; Jiang, T.; Dai, Q.H.; Li, D. Evaluation and development of deep neural networks for image super-resolution in optical microscopy. *Nat. Methods* **2021**, *18*, 194–202. [CrossRef]

11. Dong, R.; Zhang, L.; Fu, H. RRSGAN: Reference-Based Super-Resolution for Remote Sensing Image. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5601117. [CrossRef]
12. Zheng, H.T.; Ji, M.Q.; Wang, H.Q.; Liu, Y.B.; Fang, L. CrossNet: An End-to-End Reference-Based Super Resolution Network Using Cross-Scale Warping. In Proceedings of the 15th European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 87–104.
13. Zhang, Z.F.; Wang, Z.W.; Lin, Z.; Qi, H.R. Image Super-Resolution by Neural Texture Transfer. In Proceedings of the 32nd IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 7974–7983.
14. Yang, F.Z.; Yang, H.; Fu, J.L.; Lu, H.T.; Guo, B.N. Learning Texture Transformer Network for Image Super-Resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Electr Network, Seattle, WA, USA, 14–19 June 2020; pp. 5790–5799.
15. Wang, T.; Xie, J.; Sun, W.; Yan, Q.; Chen, Q. Dual-camera super-resolution with aligned attention modules. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 2001–2010.
16. Zhang, Y.F.; Li, T.R.; Zhang, Y.; Chen, P.R.; Qu, Y.F.; Wei, Z.Z. Computational Super-Resolution Imaging with a Sparse Rotational Camera Array. *IEEE Trans. Comput. Imaging* **2023**, *9*, 425–434. [CrossRef]
17. Liu, S.-B.; Xie, B.-K.; Yuan, R.-Y.; Zhang, M.-X.; Xu, J.-C.; Li, L.; Wang, Q.-H. Deep learning enables parallel camera with enhanced-resolution and computational zoom imaging. *PhotoniX* **2023**, *4*, 17. [CrossRef]
18. Zhu, X.Z.; Hu, H.; Lin, S.; Dai, J.F. Deformable ConvNets v2: More Deformable, Better Results. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 9300–9308.
19. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. In Proceedings of the 31st Annual Conference on Neural Information Processing Systems (NIPS), Long Beach, CA, USA, 4–9 December 2017.
20. He, K.M.; Zhang, X.Y.; Ren, S.Q.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 27–30 June 2016; pp. 770–778.
21. Ledig, C.; Theis, L.; Huszar, F.; Caballero, J.; Cunningham, A.; Acosta, A.; Aitken, A.; Tejani, A.; Totz, J.; Wang, Z.H.; et al. Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network. In Proceedings of the 30th IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 105–114.
22. Lim, B.; Son, S.; Kim, H.; Nah, S.; Lee, K.M. Enhanced Deep Residual Networks for Single Image Super-Resolution. In Proceedings of the 30th IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Honolulu, HI, USA, 21–26 July 2017; pp. 1132–1140.
23. Blau, Y.; Mechrez, R.; Timofte, R.; Michaeli, T.; Zelnik-Manor, L. The 2018 PIRM Challenge on Perceptual Image Super-Resolution. In Proceedings of the 15th European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 334–355.
24. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative Adversarial Networks. *Commun. ACM* **2020**, *63*, 139–144. [CrossRef]
25. Ma, C.; Rao, Y.M.; Lu, J.W.; Zhou, J. Structure-Preserving Image Super-Resolution. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 7898–7911. [CrossRef]
26. Liu, J.; Yuan, Z.; Pan, Z.; Fu, Y.; Liu, L.; Lu, B. Diffusion Model with Detail Complement for Super-Resolution of Remote Sensing. *Remote Sens.* **2022**, *14*, 4834. [CrossRef]
27. Han, L.; Zhao, Y.; Lv, H.; Zhang, Y.; Liu, H.; Bi, G.; Han, Q. Enhancing Remote Sensing Image Super-Resolution with Efficient Hybrid Conditional Diffusion Model. *Remote Sens.* **2023**, *15*, 3452. [CrossRef]
28. Yuan, Z.; Hao, C.; Zhou, R.; Chen, J.; Yu, M.; Zhang, W.; Wang, H.; Sun, X. Efficient and Controllable Remote Sensing Fake Sample Generation Based on Diffusion Model. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–12. [CrossRef]
29. Jiang, Y.M.; Chan, K.C.K.; Wang, X.T.; Loy, C.C.; Liu, Z.W. Robust Reference-based Super-Resolution via C-2-Matching. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Electr Network, Virtual, 19–25 June 2021; pp. 2103–2112.
30. Shim, G.; Park, J.; Kweon, I.S. Robust reference-based super-resolution with similarity-aware deformable convolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 8425–8434.
31. Zhang, J.Y.; Zhang, W.X.; Jiang, B.; Tong, X.D.; Chai, K.Y.; Yin, Y.C.; Wang, L.; Jia, J.H.; Chen, X.X. Reference-Based Super-Resolution Method for Remote Sensing Images with Feature Compression Module. *Remote Sens.* **2023**, *15*, 1103. [CrossRef]
32. Lu, L.Y.; Li, W.B.; Tao, X.; Lu, J.B.; Jia, J.Y. MASA-SR: Matching Acceleration and Spatial Adaptation for Reference-Based Image Super-Resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Electr Network, Virtual, 19–25 June 2021; pp. 6364–6373.
33. Chen, T.Q.; Schmidt, M. Fast patch-based style transfer of arbitrary style. *arXiv* **2016**, arXiv:1612.04337.
34. Barnes, C.; Shechtman, E.; Finkelstein, A.; Goldman, D.B. PatchMatch: A Randomized Correspondence Algorithm for Structural Image Editing. *ACM Trans. Graph.* **2009**, *28*, 11. [CrossRef]

35. Cao, J.Z.; Liang, J.Y.; Zhang, K.; Li, Y.W.; Zhang, Y.L.; Wang, W.G.; Van Gool, L. Reference-Based Image Super-Resolution with Deformable Attention Transformer. In Proceedings of the 17th European Conference on Computer Vision (ECCV), Tel Aviv, Israel, 23–27 October 2022; pp. 325–342.
36. Zhou, R.; Zhang, W.; Yuan, Z.; Rong, X.; Liu, W.; Fu, K.; Sun, X. Weakly Supervised Semantic Segmentation in Aerial Imagery via Explicit Pixel-Level Constraints. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–17. [CrossRef]
37. Yuan, Z.; Zhang, W.; Tian, C.; Mao, Y.; Zhou, R.; Wang, H.; Fu, K.; Sun, X. MCRN: A Multi-source Cross-modal Retrieval Network for remote sensing. *Int. J. Appl. Earth Obs. Geoinf.* **2022**, *115*, 103071. [CrossRef]
38. Yuan, Z.; Zhang, W.; Tian, C.; Rong, X.; Zhang, Z.; Wang, H.; Fu, K.; Sun, X. Remote Sensing Cross-Modal Text-Image Retrieval Based on Global and Local Information. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–16. [CrossRef]
39. Dosovitskiy, A.; Fischer, P.; Ilg, E.; Hausser, P.; Hazirbas, C.; Golkov, V.; van der Smagt, P.; Cremers, D.; Brox, T. FlowNet: Learning Optical Flow with Convolutional Networks. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 11–18 December 2015; pp. 2758–2766.
40. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2015**, arXiv:1409.1556.
41. Liang, J.Y.; Cao, J.Z.; Sun, G.L.; Zhang, K.; Van Gool, L.; Timofte, R. SwinIR: Image Restoration Using Swin Transformer. In Proceedings of the 18th IEEE/CVF International Conference on Computer Vision (ICCV), Electr Network, Montreal, BC, Canada, 11–17 October 2021; pp. 1833–1844.
42. Lu, Z.S.; Li, J.C.; Liu, H.; Huang, C.Y.; Zhang, L.L.; Zeng, T.Y. Transformer for Single Image Super-Resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 456–465.
43. Wu, H.P.; Xiao, B.; Codella, N.; Liu, M.C.; Dai, X.Y.; Yuan, L.; Zhang, L. CvT: Introducing Convolutions to Vision Transformers. In Proceedings of the 18th IEEE/CVF International Conference on Computer Vision (ICCV), Electr Network, Montreal, BC, Canada, 11–17 October 2021; pp. 22–31.
44. Chen, X.; Wang, X.; Zhou, J.; Dong, C. Activating More Pixels in Image Super-Resolution Transformer. *arXiv* **2022**, arXiv:2205.04437v3.
45. Grosche, S.; Regensky, A.; Seiler, J.; Kaup, A. Image Super-Resolution Using T-Tetromino Pixels. *arXiv* **2023**, arXiv:2111.09013.
46. Ma, J.Y.; Tang, L.F.; Fan, F.; Huang, J.; Mei, X.G.; Ma, Y. SwinFusion: Cross-domain Long-range Learning for General Image Fusion via Swin Transformer. *IEEE-CAA J. Autom. Sin.* **2022**, *9*, 1200–1217. [CrossRef]
47. Wilburn, B.; Joshi, N.; Vaish, V.; Talvala, E.V.; Antunez, E.; Barth, A.; Adams, A.; Horowitz, M.; Levoy, M. High performance imaging using large camera arrays. *ACM Trans. Graph.* **2005**, *24*, 765–776. [CrossRef]
48. Yu, S.; Moon, B.; Kim, D.; Kim, S.; Choe, W.; Lee, S.; Paik, J. Continuous digital zooming of asymmetric dual camera images using registration and variational image restoration. *Multidimens. Syst. Signal Process.* **2018**, *29*, 1959–1987. [CrossRef]
49. Manne, S.K.R.; Prasad, B.H.P.; Rosh, K.S.G. *Asymmetric Wide Tele Camera Fusion for High Fidelity Digital Zoom*; Springer: Singapore, 2020; pp. 39–50.
50. Chen, C.; Xiong, Z.W.; Tian, X.M.; Zha, Z.J.; Wu, F. Camera Lens Super-Resolution. In Proceedings of the 32nd IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 1652–1660.
51. Guo, P.Y.; Asif, M.S.; Ma, Z. Low-Light Color Imaging via Cross-Camera Synthesis. *IEEE J. Sel. Top. Signal Process.* **2022**, *16*, 828–842. [CrossRef]
52. Wang, X.T.; Chan, K.C.K.; Yu, K.; Dong, C.; Loy, C.C.G. EDVR: Video Restoration with Enhanced Deformable Convolutional Networks. In Proceedings of the 32nd IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 1954–1963.
53. Jaderberg, M.; Simonyan, K.; Zisserman, A. Spatial transformer networks. In *Advances in Neural Information Processing Systems*; The MIT Press: Cambridge, MA, USA, 2015; p. 28.
54. Zhang, S.; Yuan, Q.Q.; Li, J.; Sun, J.; Zhang, X.G. Scene-Adaptive Remote Sensing Image Super-Resolution Using a Multiscale Attention Network. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 4764–4779. [CrossRef]
55. Gulrajani, I.; Ahmed, F.; Arjovsky, M.; Dumoulin, V.; Courville, A. Improved Training of Wasserstein GANs. In Proceedings of the 31st Annual Conference on Neural Information Processing Systems (NIPS), Long Beach, CA, USA, 4–9 December 2017.
56. Bustince, H.; Montero, J.; Mesiar, R. Migrativity of aggregation functions. *Fuzzy Sets Syst.* **2009**, *160*, 766–777. [CrossRef]
57. Johnson, J.; Alahi, A.; Li, F.F. Perceptual Losses for Real-Time Style Transfer and Super-Resolution. In Proceedings of the 14th European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 8–16 October 2016; pp. 694–711.
58. Sajjadi, M.S.M.; Scholkopf, B.; Hirsch, M. EnhanceNet: Single Image Super-Resolution Through Automated Texture Synthesis. In Proceedings of the 16th IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 4501–4510.
59. Khan, S.; Naseer, M.; Hayat, M.; Zamir, S.W.; Khan, F.S.; Shah, M. Transformers in Vision: A Survey. *ACM Comput. Surv.* **2022**, *54*, 41. [CrossRef]
60. Wang, X.T.; Yu, K.; Wu, S.X.; Gu, J.J.; Liu, Y.H.; Dong, C.; Qiao, Y.; Loy, C.C. ESRGAN: Enhanced Super-Resolution Generative Adversarial Networks. In Proceedings of the 15th European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 63–79.

61. Zhang, R.; Isola, P.; Efros, A.A.; Shechtman, E.; Wang, O. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In Proceedings of the 31st IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 586–595.
62. Salimans, T.; Goodfellow, I.; Zaremba, W.; Cheung, V.; Radford, A.; Chen, X. Improved Techniques for Training GANs. In Proceedings of the Advances in Neural Information Processing Systems 29 (NIPS 2016), Barcelona, Spain, 5–10 December 2016.
63. Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; Hochreiter, S. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In Proceedings of the Advances in Neural Information Processing Systems 30 (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017.
64. Lucic, M.; Kurach, K.; Michalski, M.; Gelly, S.; Bousquet, O. Are GANs Created Equal? A Large-Scale Study. In Proceedings of the Advances in Neural Information Processing Systems 31 (NeurIPS 2018), Montréal, QC, Canada, 3–8 December 2018.
65. Mittal, A.; Soundararajan, R.; Bovik, A.C. Making a "Completely Blind" Image Quality Analyzer. *IEEE Signal Process. Lett.* **2013**, *20*, 209–212. [CrossRef]
66. Liu, L.X.; Liu, B.; Huang, H.; Bovik, A.C. No-reference image quality assessment based on spatial and spectral entropies. *Signal Process. Image Commun.* **2014**, *29*, 856–863. [CrossRef]
67. Ma, C.; Yang, C.Y.; Yang, X.K.; Yang, M.H. Learning a no-reference quality metric for single-image super-resolution. *Comput. Vis. Image Underst.* **2017**, *158*, 1–16. [CrossRef]
68. Tu, Z.; Yang, X.; Fu, Z.; Gao, S.; Yang, G.; Jiang, L.; Wu, M.; Wang, S. Concatenating wide-parallax satellite orthoimages for simplified regional mapping via utilizing line-point consistency. *Int. J. Remote Sens.* **2023**, *44*, 4857–4882. [CrossRef]
69. Wadduwage, D.N.; Singh, V.R.; Choi, H.; Yaqoob, Z.; Heemskerk, H.; Matsudaira, P.; So, P.T.C. Near-common-path interferometer for imaging Fourier-transform spectroscopy in wide-field microscopy. *Optica* **2017**, *4*, 546–556. [CrossRef]
70. Aleman-Castaneda, L.A.; Piccirillo, B.; Santamato, E.; Marrucci, L.; Alonso, M.A. Shearing interferometry via geometric phase. *Optica* **2019**, *6*, 396–399. [CrossRef]
71. Zeng, Y.; Fu, J.; Chao, H.; Guo, B. Learning pyramid-context encoder network for high-quality image inpainting. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 1486–1494.