symmetry

MDPI

Article

Transformer with Hybrid Attention Mechanism for Stereo Endoscopic Video Super Resolution

Tianyi Zhang [1,2] and Jie Yang [1,2,*]

[1] Department of Automation, Shanghai Jiao Tong University, Shanghai 200240, China; autozty@sjtu.edu.cn
[2] Institute of Medical Robotics, Shanghai Jiao Tong University, Shanghai 200240, China
* Correspondence: jieyang@sjtu.edu.cn

Abstract: With stereo cameras becoming widely used in invasive surgery systems, stereo endoscopic images provide important depth information for delicate surgical tasks. However, the small size of sensors and their limited lighting conditions lead to low-quality and low-resolution endoscopic images and videos. In this paper, we propose a stereo endoscopic video super-resolution method using transformer with a hybrid attention mechanism named HA-VSR. Stereo video SR aims to reconstruct high-resolution (HR) images from corresponding low-resolution (LR) videos. In our method, the stereo correspondence and temporal correspondence are incorporated into the HA-VSR model. Specifically, the Swin transformer architecture is utilized in proposed framework with hybrid attention mechanisms. The parallel attention mechanism is utilized by using the symmetry and consistency of left and right images, and the temporal attention mechanism is utilized by using the consistency of consecutive frames. Detailed quantitative evaluation and experiments on two datasets show the proposed model achieves advanced SR reconstruction performance, showing that the proposed stereo VSR framework outperforms alternative approaches.

Keywords: stereo video super-resolution; endoscopic image; transformer

check for updates

Citation: Zhang, T.; Yang, J. Transformer with Hybrid Attention Mechanism for Stereo Endoscopic Video Super Resolution. *Symmetry* **2023**, *15*, 1947. https://doi.org/10.3390/sym15101947

Academic Editor: Zhixun Su

Received: 24 September 2023
Revised: 11 October 2023
Accepted: 18 October 2023
Published: 20 October 2023

Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

For a long time now, endoscopy has been widely used for surgical navigation and operation in minimally invasive surgery [1]. However, an endoscopic image with a single camera can provide limited depth information and field-of-view. For this reason, stereo cameras are increasingly used in endoscopy for operations that require delicate manipulation, particularly in robot-assisted procedures [2]. Compared with a single endoscopic image, stereo endoscopic images recorded by two different views provide depth cues and more sub-pixel information [3].

However, the image quality and resolution of endoscopy suffer from narrow surgical space and a limited endoscopic view. In order to record different scales of tubular cavities and lumens, the size of optical sensors need to be enough small. In addition, unstable illumination conditions also lead to image degradation and information loss of stereo endoscopic images, which may result in negative influence in down-stream procedures, such as image classification, segmentation, and reconstruction [4,5]. Therefore, it is beneficial to enhance the resolution and quality of stereo endoscopic images and video frames.

To deal with the aforementioned problem, the image super-resolution (SR) method provides a typical way to reconstruct a high-resolution (HR) image from the corresponding low-resolution (LR) image. Traditional SR methods are based on interpolations and sparse representation [6,7]. Recently, there have been various single image SR (SISR) methods based on deep learning networks [8–12]. Among these methods, convolutional neural networks (CNN) are most commonly used with increasingly more layers. In the SISR method, image feature extraction is the key point of the model for better performance. For stereo images, a simple method involves respectively applying the SISR networks to

both the left and right views. However, this method may damage the correspondence between the left and right views. Recently, CNN-based methods have also shown advanced performance for improving stereo image SR [13–16]. These methods jointly consider the SR problems from both views by using spatial shift patches and the parallel attention mechanism. Knowing how to incorporate the consistency between two views into the framework becomes the key issue of stereo image SR. The video SR methods are expected to extract complementary information from neighbouring video frames, which is beneficial for providing more details to solve the SR problem. Most of the previous video SR methods extract frames and perform reconstruction through single SR methods, while current video SR methods typically align the LR frames with the implicit motion compensation and optical flow [17–22].

For stereo endoscopic video SR, consecutive frames provide additional temporal consistency. In the traditional video SR method, the network receives several consecutive images as input and then extracts and synthesizes image features to reconstruct the HR output. Similarly, applying the 2D video SR method to stereo video frames may also lose the correspondence of left and right views. The recently proposed method [23], which integrated the spatial and temporal information from the stereo views, enhanced the video SR performance by using the optical flow-based feed-forward layer in its model. Therefore, the relationship between different views and adjacent frames should be considered to reconstruct the output image.

Although the visual performance of the CNN-based SR method is more advanced than conventional SR methods, limited by the size of convolution kernel and the locality of field-of-view regions, CNN models lack the utility of long-range dependency. Recently, the transformer network that incorporates self attention mechanisms provides a promising solution on different visual tasks [24–26] . In transformer-based methods, the input image and frames are divided into small patches as sequential token inputs, and the image feature is extracted by self-attention using global relationships of the tokens. The Swin transformer [27] integrates the advantages of both CNN and transformer via parallel computing and the shifted window technique and constructs a hierarchical feature representation by starting from small patches and gradually merging adjacent patches in deeper transformer layers. With the multi-scale feature maps, the Swin transformer model can efficiently and conveniently leverage advanced techniques for dense prediction and image reconstruction [28].

In order to make full use of stereo image pairs and adjacent frames in stereo endoscopic video SR, we propose a deep neural network (HA-VSR) based on Swin transformer with hybrid attention mechanisms. In detail, the Swin transformer blocks are utilized as the backbone of the network, along with parallel attention modules and temporal attention modules. The HA-VSR network consists of three modules: the image feature extractor, hybrid attention modules, and the HR image reconstruction module. The feature extractor consists of cascaded residual Swin transformer blocks (RSTB), which are utilized for deep feature extraction, with a shallow convolution layer ahead. Also, a global skip connection linked to the HR reconstruction module is utilized to preserve the low-frequency information. The input of the model contains left and right adjacent frames, and the weight of the feature extractor is shared by the input images. In addition, the parallel attention module (PAM) and temporal attention module (TAM) are applied to utilize the relationship and correspondence between the left and right views, and the consecutive frames. In the PAM module, the stereo correspondence along the epipolar line is utilized between according to the parallel attention mechanism, by calculating the cross correlation between the pixels on corresponding epipolar lines. In the TAM module, the cross transformer unit is utilized to generate refined image feature according to the cross attention mechanism. Finally, the refined features are concatenated in the reconstruction module to obtain the output image.

The main contributions of this paper are two-folds:

(1) A stereo endoscopic video super-resolution method using transformer with the hybrid attention mechanism (HA-VSR) is proposed for stereo endoscopic video SR. In order

to improve the efficiency, the symmetric Swin transformer blocks are utilized to extract image features in the network.

(2) To preserve the consistency of correspondence among different frames and different left and right views, we propose a hybrid attention block that combines the parallel attention module and temporal attention module in our framework. Extensive experiments show that the proposed HA-VSR outperforms previous SR methods on stereo endoscopic video datasets.

2. Related Works

This section is a brief review of the SR methods related to our work, including the single image SR, the stereo image SR, and the video SR methods.

2.1. Single Image SR

Conventional single image SR methods are based on interpolation. Unknown pixels are calculated by a mapping or interpolation function using known pixel values. Yang et al. [7] proposed a sparse representation-based image SR method by training coupled dictionaries. Patch-based SR methods were also commonly proposed to preserve the structure of the reconstructed HR images.

In recent years, convolutional neural networks are widely used in visual tasks and have shown promising performance on image SR methods. Dong et al. [8] proposed the first SR convolutional neural network (SRCNN) to generate HR images by using a 3-layer end-to-end network. Kim et al. [9] incorporated the residual connection and proposed a very deep CNN network for SR (VDSR), which consists of 20 convolutional layers. Based on these previous SR methods, Tai et al. [10] introduced the recursion into the deep recursive residual network (DRRN) to efficiently generate features in SR frameworks. Zhang et al. [11] proposed a very deep residual channel attention network (RCAN) for SR by using channel attention in feature extraction. The CNN based methods typically improve the amount of the layers in network to achieve better performance. Later, the transformer architecture presented its excellent capability on non-local feature extraction. Liang et al. [28] proposed an image reconstruction model named SwinIR with Swin transformer structure.

2.2. Stereo Image SR

Recently, stereo image SR has received increasing attention, and there have been some notable stereo SR works that utilize stereo information. In order to improve the result, kniwing how to effectively apply the corresponding information between two views to the SR network becomes the critical challenge for enhancing stereo images. Bhavsar et al. [13] proposed an integrated framework to estimate the image depth map and the SR image from multiple LR images. They produced a joint energy function while minimizing it by updating the SR image and disparity map iteratively. There are also several CNN-based stereo SR approaches using disparity and parallex attention. Jeon et al. [14] proposed a stereo enhancement super-resolution model (StereoSR) using a single image and a stack of auxiliary shifted images to generate SR results with more details. However, this technique was inapplicable for stereo images with variant disparities since the maximum parallax was fixed. To incorporate the stereo correspondence into the SR technique, Wang et al. [15] proposed a parallax-attention stereo super-resolution network (PASSRnet). The proposed parallax attention module (PAM) effectively exploits information from both views from a global receptive field along the epipolar line for correspondence matching. Ying also introduced several PAMs to various stages of the pre-trained SISR networks to improve the performance.

The above methods all extract the image feature by the CNN-like backbone, in which the global self-attention is a challenge. Moreover, only using pixels on the epipolar line in the parallel attention mechanism is insufficient since the stereo pair may not be accurately calibrated. In our proposed network, a Swin transformer-based backbone is introduced with both parallel attention and temporal attention between stereo video frames.

2.3. Video SR

Unlike SISR, the video SR problem is more challenging because of the mismatched information among adjacent video frames. It is similar to the stereo SR task, in which the implicit correspondence can be extracted by cross-image attention and recurrent methods. Recent video SR methods utilize more indirect approaches, such as deformable convolutions and alignment functions, to achieve better performance [29]. Chan et al. [17] redesigned the BasicVSR model by proposing second-order grid propagation and flow-guided deformable alignment. Their framework with enhanced propagation and alignment could exploit types of spatial-temporal information across misaligned video frames more effectively. Wang et al. [18] proposed a practical compression-aware video super-resolution model named CAVSR. In their model, a compression encoder was designed to model compression levels of input frames, which is beneficial for adapting the video enhancement process to the estimated compression level. Lu et al. [19] proposed a novel framework to achieve video SR at random scales. Their unified framework incorporated the spatial-temporal interpolation of events by learning implicit neural representations from queried spatial-temporal coordinates and features. Li et al. [20] proposed a method for high-quality and efficient video SR by leveraging the spatial-temporal information and the number of the divided video chunks. In order to compress the video SR models, Xia et al. [21] developed a structured pruning scheme for several key components in video SR frameworks.

Recently, Imani et al [23] proposed the Trans-SVSR model, which integrated the spatio-temporal information from the stereo views while maintaining consistency. In Trans-SVSR, an optical flow-based feed-forward layer in the transformer model spatially aligns input features by considering the correlations between all frames.

3. Methods

3.1. Network Architecture

As shown in Figure 1, we propose a stereo endoscopic video SR network, which consists of the feature extractor, hybrid attention modules, and the reconstruction module. Given the left LR frames $I_{n-k}^{L,LR}, \ldots, I_{n}^{L,LR}, \ldots, I_{n+k}^{L,LR}$ and right LR frames $I_{n-k}^{R,LR}, \ldots, I_{n}^{R,LR}, \ldots, I_{n+k}^{R,LR}$, the reconstructed SR result $I_{n}^{L,SR}$ and $I_{n}^{R,SR}$ can be generated by the model $\mathcal{F}(\cdot,\cdot,\theta)$ with model parameters θ, as Equation (1) shows.

$$I_{n}^{L,SR}, I_{n}^{R,SR} = \mathcal{F}(\{I_{n-k}^{L,LR}, \ldots, I_{n}^{L,LR}, \ldots, I_{n+k}^{L,LR}\}, \{I_{n-k}^{R,LR}, \ldots, I_{n}^{R,LR}, \ldots, I_{n+k}^{R,LR}\}; \theta) \quad (1)$$

In this work, the superscripts denote the attributes of the tensor, including the left and right views, the LR and SR resolutions, and the processed status by certain modules. The subscripts of tensors denote the temporal information, namely, the number of frames, and the subscripts of modules denote the order.

The superscripts L and R denote the left view and right view, respectively, and the subscript n represents the ordinal number of the input frame. As Equation (1) shows, the number of left view frames is $2k+1$. Inspired by [28], a 3×3 convolutional layer $\mathcal{F}_S$ is used to extract low-level features of input frames. Since the feature extractor is shared by all the input frames, the subscript can be omitted in the following equation:

$$F^{L,S} = \mathcal{F}_S(I^{L,LR}), \qquad F^{R,S} = \mathcal{F}_S(I^{R,LR}) \quad (2)$$

where F^S represents the shallow feature extracted by the $\mathcal{F}_S$. This shallow convolution layer is beneficial as pre-processing for image reconstruction, resulting in more stable optimization of the model and better performance.

Next, the high-level deep image feature F^D is produced by the deep feature extractor $\mathcal{F}_D$,

$$F^{L,D} = \mathcal{F}_D(F^{L,S}), \qquad F^{R,D} = \mathcal{F}_D(F^{R,S}) \quad (3)$$

For convenience, the superscripts L and R that denote left and right are omitted in the following equations since the operations on the left and right image features are consistent during feature extraction.

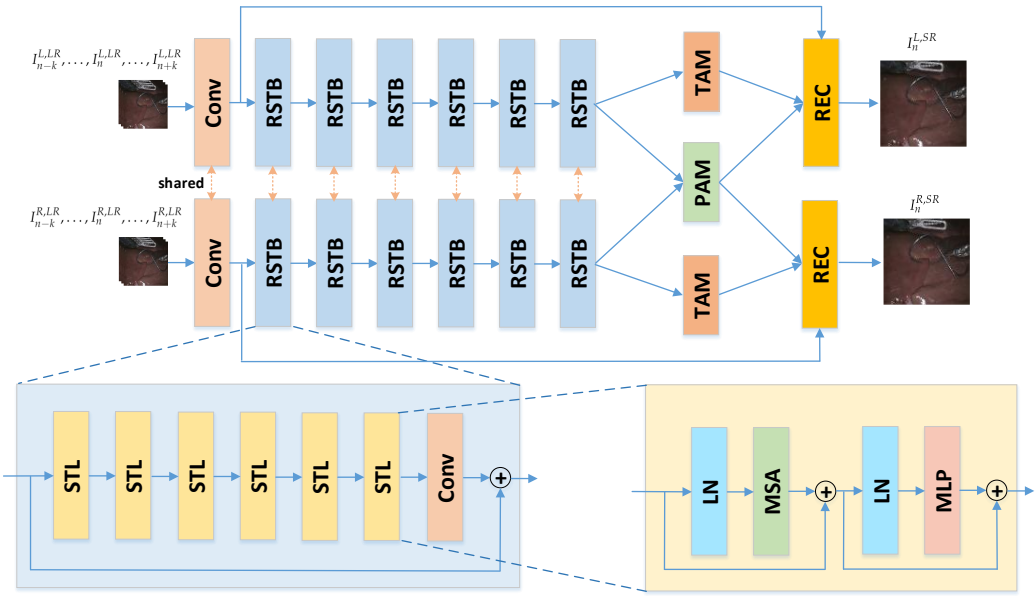


Figure 1. An overview architecture of the proposed HA-VSR network. The I^{LR} on the left side represents the input low-resolution video frames. And the I^{SR} on the right side represents the output reconstructed video frame. The RSTB represents the residual Swin transformer block, and REC denotes the reconstruction module. PAM and TAM denote the parallel attention and temporal attention modules.

Following the SwinIR network [28], the $\mathcal{F}_D$ consists of six sequential residual Swin transformer blocks (RSTB) and a 3×3 convolutional layer. The Swin transformer layers for local attention and cross-window interaction are utilized in RSTB. In detail, the intermediate features $F^{D1}, F^{D2}, \ldots, F^{D6}$ and the output deep feature F^D are extracted as

$$\begin{aligned} F^{D_1} &= \mathcal{F}_{RSTB_1}(F^S) \\ F^{D_i} &= \mathcal{F}_{RSTB_i}(F^{D_{i-1}}) \qquad i = 2,3,4,5,6 \\ F^D &= \mathcal{F}_{conv}(F^{D_6}) \end{aligned} \tag{4}$$

where $\mathcal{F}_{RSTB_i}$ represents the i-th RSTB and $\mathcal{F}_{conv}$ is the convolutional layer, which is beneficial for aggregation of different features.

After the feature extractor, the left and right deep image features will be fed into the parallel attention module (PAM):

$$F_n^{L,P}, F_n^{R,P} = \mathcal{F}_{PAM}(F_n^{L,D}, F_n^{R,D}) \tag{5}$$

And the left frame features are served as the input of the temporal attention module (TAM):

$$F_n^{L,T} = \mathcal{F}_{TAM}(F_{n-k}^{L,D}, \ldots, F_n^{L,D}, \ldots, F_{n+k}^{L,D}) \tag{6}$$

and same to the right frame features. Finally, the SR result image is reconstructed by aggregating the above features as

$$I_n^{L,SR} = \mathcal{F}_{REC}(F_n^{L,S}, F_n^{L,D}, F_n^{L,P}, F_n^{L,T}) \tag{7}$$

where $\mathcal{F}_{REC}$ denotes the concatenation operation and a pixelshuffle layer with a convolutional layer.

3.2. Residual Swin Transformer Block

As shown in Figure 1, the residual Swin transformer block (RSTB) consists of a residual block with Swin transformer layers (STL) and a convolutional layer. Given the input feature $F^{D_{i,0}}$ of the i-th RSTB, the intermediate features $F^{D_{i,1}}, F^{D_{i,2}}, \ldots, F^{D_{i,6}}$ are first extracted by 6 Swin transformer layers as

$$F^{D_{i,k}} = \mathcal{F}_{STL_{i,k}}(F^{D_{i,k-1}}) \qquad k = 1, 2, 3, 4, 5, 6 \tag{8}$$

where $\mathcal{F}_{STL_{i,k}}$ is the k-th Swin transformer layer in the i-th RSTB. Then, an additional convolutional layer is attached. The output of RSTB is formulated as

$$F^{D_i} = \mathcal{F}_{conv_i}(F^{D_{i,6}}) + (F^{D_{i,0}}) \tag{9}$$

where $\mathcal{F}_{conv_i}$ is the convolutional layer in the i-th RSTB. The skip connection in RSTB allows for the fusion of different stages of extracted features.

Based on the original multi-head self-attention of the original transformer layer, the Swin transformer layer (STL) [27] introduces local attention and the shifted window mechanism. Given an input tensor of size $H \times W \times C$, the STL first reshapes the input to a $HW/M^2 \times M^2 \times C$ feature by dividing the input into $M \times M$ local windows with no overlapped region, in which HW/M^2 is the number of divided windows. For each window, the local mechanism computes the standard self-attention independently. For a local window feature $F \in \mathbb{R}^{M^2 \times C}$, the query, key, and value matrices (Q, K, and V, respectively) are computed as

$$Q = FW^Q, \qquad K = FW^K, \qquad V = FW^V \tag{10}$$

where W^Q, W^K, and W^V are weight matrices that are shared across different windows. For $Q, K, V \in \mathbb{R}^{M^2 \times d}$, the attention matrix is generated by the self-attention mechanism in a local window as

$$F_{attn}(Q, K, V) = \text{softmax}(QK^T/\sqrt{d} + B)V \tag{11}$$

where B is the variable relative positional encoding. And $\sqrt{d}$ needs to be divided in order to prevent excessive magnitude of QK^T. To effectively utilize the transformer layer, the multihead self-attention (MAS) mechanism has been incorporated into the model by performing the attention function for n_h times in parallel and concatenating the results.

Then, a multi-layer perceptron (MLP) consisting of two fully-connected (FC) layers with the GELU function between them is utilized for further feature transformations. The LayerNorm (LN) is added before both MSA and MLP, and the residual connection is applied for both modules. The whole process is formulated as

$$\begin{aligned} F_{im} &= \mathcal{F}_{MSA}(\text{LN}(F)) + F \quad \text{or} \\ F_{im} &= \mathcal{F}_{SWMSA}(\text{LN}(F)) + F \\ F_{out} &= \mathcal{F}_{MLP}(\text{LN}(F_{im})) + F \end{aligned} \tag{12}$$

where F_{im} denotes the immediate feature and F_{out} represents the output feature of the block.

In image restoration tasks, the window partition is consistent across all the stages in the feature extractor. In order to conduct the cross-window connection, the MSA and Shifted Window MSA (SWMSA) are settled alternately. In SWMSA, the input feature is shifted by $\left(\lfloor \frac{M}{2} \rfloor, \lfloor \frac{M}{2} \rfloor\right)$ pixels before partitioning.

3.3. Parallel Attention Module

In order to utilize the stereo correspondence along the epipolar line between left and right views, the parallel attention module is introduced into our model. The original PAM [15] calculates the correlation of pixels on corresponding epipolar lines. In the PAM, features along the epipolar line are utilized to build sparse attention maps by the reshape operation and geometry-aware matrix multiplication. We adopted an atrous convolution layer to extract multiple-line stereo image features, as Figure 2 shows.

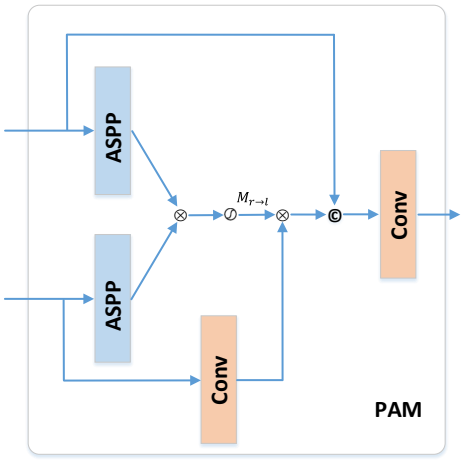


Figure 2. The architecture of PAM. Only the procedure of left PAM feature generation is shown for convenience. The left and right image features are utilized to calculate the attention map $M^{R\to L}$.

Given the left and right feature maps F_l^D, F_r^D extracted from the deep feature extractor, the features are first fed to a residual atrous spatial pyramid pooling (ASPP) block.

Then, the resulting features are utilized to generate the output feature and the parallax-attention map $M^{R\to L}$ based on PAM. Additionaly, stereo features are exchanged to repeat the progress. Finally, the concatenation of the output feature and the identity of input feature is fed to a convolutional layer for feature fusion.

$$\begin{aligned} F^{L,Q} &= \mathcal{F}^{L,ASPP}(F^{L,D}) \\ F^{R,Q} &= \mathcal{F}^{L,ASPP}(F^{R,D}) \\ M^{R\to L} &= \text{softmax}(F^{L,Q} \otimes F^{R,Q}) \\ F_l^P &= \mathcal{F}_{conv}(\text{concat}(M^{R\to L} \otimes \mathcal{F}_{conv}(F^{R,D}), F^{L,D}) \end{aligned} \tag{13}$$

3.4. Temporal Attention Module

In transformer blocks, the self-attention module is used to exploit the intra-image feature, and the PAM module above is designed to effectively extract the stereo image feature by using the relationship of pixels on epipolar lines between the left and right images. In this section, we introduce the temporal attention module into our model. As shown in Figure 3, the TAM takes two frame features as the input, which are passed into a cross transformer unit. Using $Q_1, K_1, V_1, Q_2, K_2, V_2$ to denote the query, key, and value of two frame features, respectively, the outputs of MSA are formed with the following equations:

$$\begin{aligned} F_{attn,1}(Q_1, K_2, V_2) &= \text{softmax}(Q_1 K_2' / \sqrt{d} + B_1) V_2 \\ F_{attn,2}(Q_2, K_1, V_1) &= \text{softmax}(Q_2 K_1' / \sqrt{d} + B_2) V_1 \end{aligned} \tag{14}$$

where B_1 and B_2 are the variable relative positional encodings and K' denotes the transposition of K. The generated feature of TAM is the following item:

$$F_n^{TAM} = \mathcal{F}_{MLP}(\mathrm{LN}(F_{attn,n})) + F_{attn,n} \tag{15}$$

The output of TAM is calculated by:

$$F_n^T = \mathcal{F}_{conv}(\mathrm{concat}(F_{n-k}^{TAM}, \ldots, F_n^{TAM}, \ldots, F_{n+k}^{TAM}) \tag{16}$$

Hence, the output of the TAM in this model takes cross-relationships of adjacent frames into consideration.

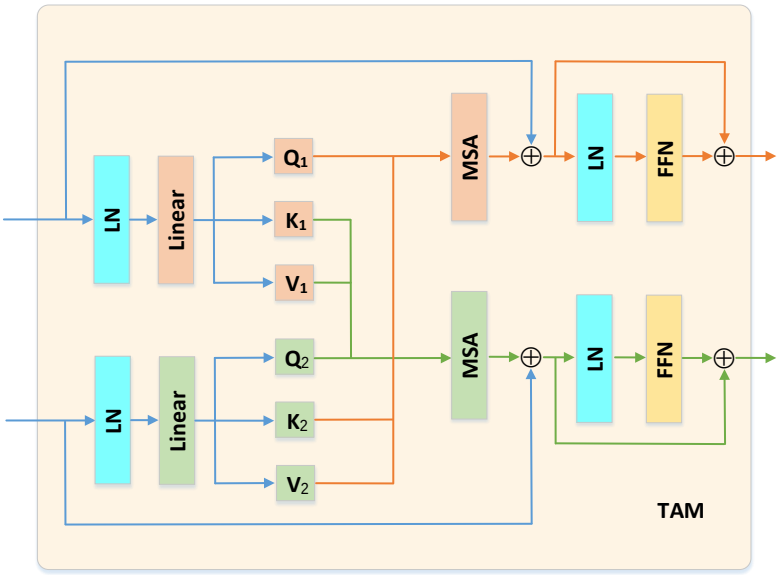


Figure 3. The architecture of TAM. This module takes stereo image features as the input, with performed result, which is calculated by cross-attention.

3.5. Loss Function

We introduce the SR loss $\mathcal{L}_{SR}$ and PAM loss $\mathcal{L}_{PAM}$ to train our network. The overall loss function is defined as

$$\mathcal{L} = \mathcal{L}_{SR} + \lambda \mathcal{L}_{PAM} \tag{17}$$

where λ is set to be 0.02 in this work.

The SR loss is utilized to minimize the difference between SR images and ground truth HR images as

$$\mathcal{L}_{SR} = \|I^{L,SR} - I^{L,HR}\|_1 + \|I^{R,SR} - I^{R,HR}\|_1 \tag{18}$$

where I^{HR} is the corresponding ground truth HR image, and we used the L_1 loss function as the measurement of SR loss.

Consisting of the photometric loss, the cycle loss, and the smoothness loss, the PAM loss function is used to regularize the parallax attention maps:

$$\mathcal{L}_{PAM} = \mathcal{L}_{photo} + \mathcal{L}_{cycle} + \mathcal{L}_{smooth} \tag{19}$$

Since the parallax attention map is available in PAM, the right view image can be calculated by multiplying the attention map to the left view image matrix. The photometric loss can be formulated as

$$\mathcal{L}_{photo} = \|I^{L,LR} - (M^{R\to L} \otimes I^{R,LR})\|_1 + \|I^{R,LR} - (M^{L\to R} \otimes I^{L,LR})\|_1 \tag{20}$$

where $\otimes$ denotes batch-wise matrix multiplication.

Moreover, a cycle loss is used to keep the cycle consistency of the parallax attention map, which can be represented as

$$\mathcal{L}_{cycle} = \|M^{L\to R} \otimes M^{R\to L} - \mathbf{I}\|_1 + \|M^{R\to L} \otimes M^{L\to R} - \mathbf{I}\|_1 \tag{21}$$

where $\mathbf{I} \in \mathbb{R}^{H\times W\times W}$ denotes a stack of H identity matrices.

The smoothness loss function is introduced to generate a smoother and more consistent parallax attention map:

$$\mathcal{L}_{smooth} = \sum_{d\in\{L\to R, R\to L\}} \sum_{i,j,k} (\|M^d(i,j,k) - M^d(i+1,j,k)\|_1 + \|M^d(i,j,k) - M^d(i,j+1,k+1)\|_1) \tag{22}$$

4. Experiments

In this section, we first introduce the datasets and experimental settings. Then, we compare the proposed HA-VSR to several image SR and video SR methods. Finally, the ablation studies are conducted to validate the components of our proposed method.

4.1. Experimental Settings

For the model training, we adopted 240 pairs of stereo video frames from the da Vinci dataset [30] as the training dataset. To produce LR images, the HR images were downscaled on particular scaling factors by using the bicubic operation. The vertically flipping was used for data augmentation.

For testing, we adopted two stereo endoscopic video datasets: the da Vinci dataset with the testset consists of 80 pairs of stereo endoscopic video frames collected by stereo cameras of the da Vinci system, and the SCARED dataset [31] consists of 120 stereo video frames.

The network was implemented using PyTorch and trained on NVIDIA Titan Xp GPU. All the models were optimized using Adam with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The batch size was set to 8, and the initial learning rate was set to 1×10^{-4}. In this work, the k was set to 1, which means three adjacent frames were used as data input.

4.2. Evaluation Results

The peak signal-to-noise ratio (PSNR) is a broadly utilized quantitative measurement in image SR, which measures the proximity between HR and SR images. The structure similarity index measure (SSIM) is also quantified as a perceptual metric of image similarity. We compare our algorithm with several SR methods. These metrics were calculated on RGB color space, and the PSNR and SSIM scores are the average on the left and right image pairs among the frames (i.e., (Left + Right)/2). As Table 1 shows, the proposed HA-VSR network achieves remarkable PSNR and SSIM scores on test sets for $\times 2$ and $\times 4$ SR tasks. Specifically, the PSNR values of our method on the two test sets are higher than other single, stereo, and VSR methods. For the $\times 2$ stereo SR task, our HA-VSR performs better SSIM values on the majority of the datasets. The quantitative evaluation results demonstrate the ability of our HA-VSR in utilizing the temporal cross attention and parallel attention to reconstruct the HR image.

Figures 4 and 5 show the qualitative performance comparison of different methods for $\times 4$ SR on da Vinci dataset. Details could be observed in zoomed-in regions. According to the qualitative evaluation, stereo SR methods perform better in detail than SISR approaches in general. And transformer-based methods produce clearer images than most of the CNN-based methods. Moreover, compared to other methods, our HA-VSR incorporates parallel attention and temporal attention in stereo image pairs to improve the SR performance in edge and texture details.

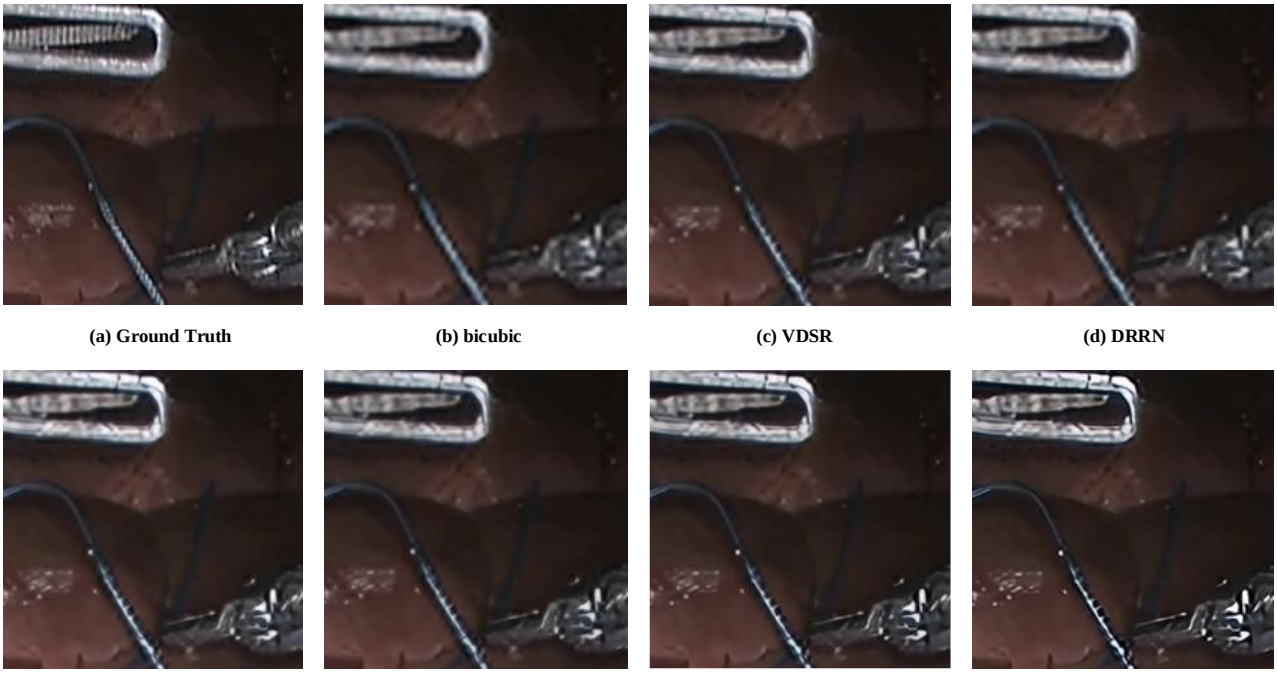


Figure 4. SR endoscopic images (da Vinci dataset) recovered by different methods, and corresponding HR image with scale factor $\times$ 4.

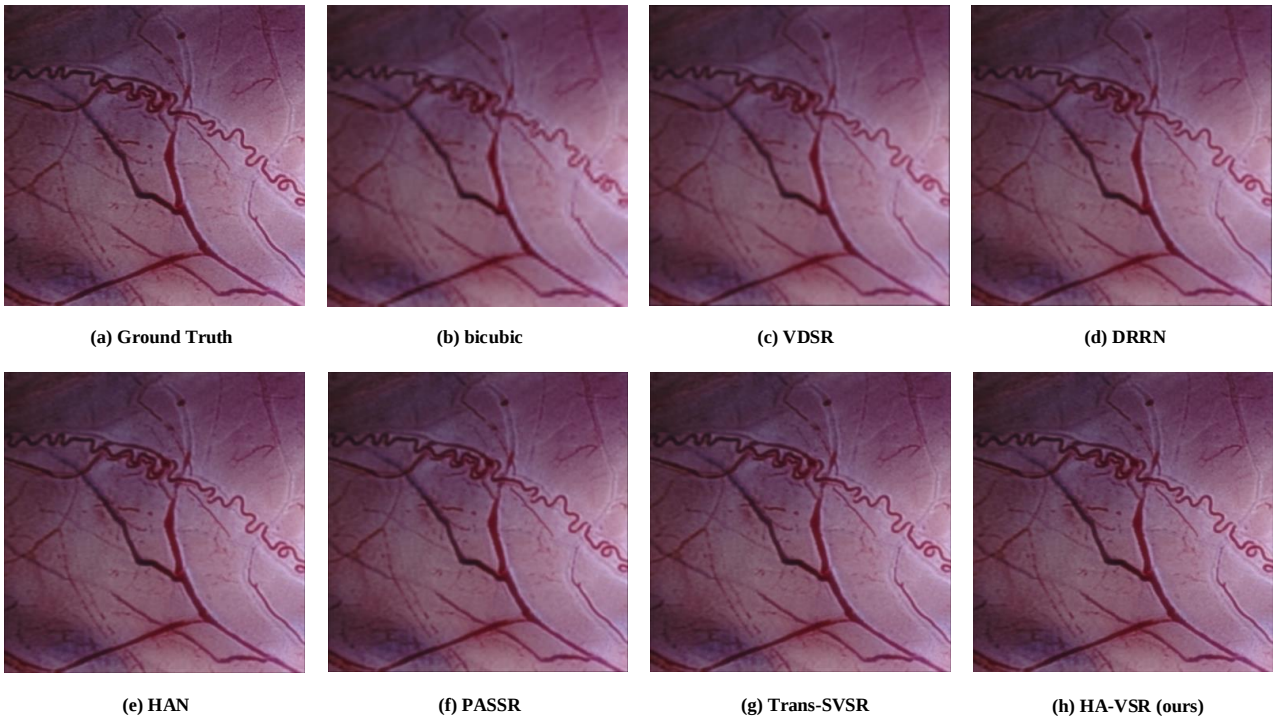


Figure 5. SR endoscopic images (SCARED dataset) reconstructed by different methods, and corresponding HR image with scale factor $\times$ 4.

Table 1. Performance comparison (PSNR/SSIM) between our proposed method and other methods on da Vinci and SCARED datasets for ×2 and ×4 SR.

Method	Scale	da Vinci	SCARED
bicubic	×2	35.6629/0.9645	38.6021/0.9792
VDSR	×2	37.1054/0.9681	39.5793/0.9824
DRRN	×2	37.9829/0.9733	40.1844/0.9858
HAN	×2	38.2513/0.9765	40.6208/0.9869
PASSR	×2	37.6501/0.9714	40.3617/0.9860
Trans-SVSR	×2	38.2165/0.9767	40.7365/0.9875
HA-VSR (proposed)	×2	38.3702/0.9771	40.8019/0.9870
bicubic	×4	30.0670/0.9358	32.8524/0.9480
VDSR	×4	31.1425/0.9410	33.3527/0.9516
DRRN	×4	31.6728/0.9428	34.0189/0.9558
HAN	×4	31.7459/0.9433	34.5003/0.9569
PASSR	×4	31.4637/0.9415	34.1275/0.9547
Trans-SVSR	×4	31.8952/0.9469	34.6870/0.9573
HA-VSR (proposed)	×4	32.0321/0.9477	34.7903/0.9576

4.3. Ablation Analysis

In this section, we present the ablation experiments. Table 2 shows the average performance under different configurations on da Vinci and SCARED datasets.

- Network without residual connection: The residual connection provides the SR result with low-frequency information. As Table 2 shows, the network without residual connections suffers an average decrease of ∼0.15 dB in PSNR.
- Network without the PAM module: The PAM module makes use of the stereo relationship on epipolar lines. The performance of the network without PAM module suffers an average decrease of ∼0.09 dB in PSNR on a da Vinci dataset and ∼0.12 dB on a SCARED dataset.
- Network without the TAM module: From the comparative results shown in Table 2, the SR performance benefits from the TAM module on both datasets.
- The network without both PAM and TAM: If both the PAM and TAM modules are removed from the network, the PSNR value suffers an average decrease of ∼0.2 dB on the da Vinci dataset.

Table 2. Mean PSNR(dB)/SSIM values calculated with different ablations on da Vinci and SCARED datasets for ×2 and ×4 SR.

Method	Scale	da Vinci	SCARED
HA-VSR	×2	38.3702/0.9771	40.8019/0.9870
w/o res connection	×2	38.0164/0.9748	40.5362/0.9851
w/o PAM	×2	38.2835/0.9764	40.6838/0.9859
w/o TAM	×2	38.2459/0.9769	40.7129/0.9862
w/o PAM and TAM	×2	38.1221/0.9764	40.6540/0.9856
HA-VSR	×4	32.0321/0.9477	34.7903/0.9576
w/o res connection	×4	31.8971/0.9452	34.5219/0.9566
w/o PAM	×4	31.9805/0.9473	34.6961/0.9573
w/o TAM	×4	31.9738/0.9474	34.7259/0.9575
w/o PAM and TAM	×4	31.9457/0.9473	34.6472/0.9570

5. Conclusions

In this paper, we propose a deep neural network HA-VSR based on the Swin transformer with parallel attention and temporal cross attention. In order to improve the stereo SR performance, we propose the hybrid attention modules to utilize stereo and adjacent correspondence information between the left and right views.

We demonstrate that our proposed Swin transformer-based SR network has good performance by qualitatively and quantitatively comparing it with other models in the field of stereo SR through experiments. The effectiveness of our parallel attention module and temporal attention module is also demonstrated through experiments for quantitative comparisons.

Author Contributions: Methodology, T.Z.; Writing—original draft, T.Z.; Supervision, J.Y.; Funding acquisition, J.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research was partly supported by Ministry of Science and Technology, China (No. 2019YFB1311503).

Data Availability Statement: Publicly available datasets were analyzed in this study. This data can be found here: https://endovissub2019-scared.grand-challenge.org and https://github.com/hgfe/DCSSR (accessed on 23 September 2023).

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

References

1. Peters, B.S.; Armijo, P.R.; Krause, C.; Choudhury, S.A.; Oleynikov, D. Review of emerging surgical robotic technology. *Surg. Endosc.* **2018**, *32*, 1636–1655. [CrossRef] [PubMed]
2. Mueller-Richter, U.D.A.; Limberger, A.; Weber, P.; Ruprecht, K.W.; Spitzer, W.; Schilling, M. Possibilities and limitations of current stereo-endoscopy. *Surg. Endosc.* **2004**, *18*, 942–947. [CrossRef] [PubMed]
3. Park, J.; Hwang, Y.; Yoon, J.H.; Park, M.G.; Kim, J.; Lim, Y.J.; Chun, H.J. Recent development of computer vision technology to improve capsule endoscopy. *Clin. Endosc.* **2019**, *52*, 328–333. [CrossRef] [PubMed]
4. Wang, C.C.; Chiu, Y.C.; Chen, W.L.; Yang, T.W.; Tsai, M.C.; Tseng, M.H. A deep learning model for classification of endoscopic gastroesophageal reflux disease. *Int. J. Environ. Res. Public Health* **2021**, *18*, 2428. [CrossRef] [PubMed]
5. Ali, S.; Dmitrieva, M.; Ghatwary, N.; Bano, S.; Polat, G.; Temizel, A.; Krenzer, A.; Hekalo, A.; Guo, Y.B.; Matuszewski, B.; et al. Deep learning for detection and segmentation of artefact and disease instances in gastrointestinal endoscopy. *Med. Image Anal.* **2021**, *70*, 102002. [CrossRef] [PubMed]
6. Zhou, F.; Yang, W.; Liao, Q. Interpolation-based image super-resolution using multisurface fitting. *IEEE Trans. Image Process.* **2012**, *21*, 3312–3318. [CrossRef] [PubMed]
7. Yang, J.; Wright, J.; Huang, T.S.; Ma, Y. Image super-resolution via sparse representation. *IEEE Trans. Image Process.* **2010**, *19*, 2861–2873. [CrossRef] [PubMed]
8. Dong, C.; Loy, C.C.; Tang, X. Accelerating the super-resolution convolutional neural network. In Proceedings of the Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; pp. 391–407.
9. Kim, J.; Lee, J.K.; Lee, K.M. Accurate image super-resolution using very deep convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 1646–1654.
10. Tai, Y.; Yang, J.; Liu, X. Image super-resolution via deep recursive residual network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 3147–3155.
11. Zhang, Y.; Li, K.; Li, K.; Wang, L.; Zhong, B.; Fu, Y. Image super-resolution using very deep residual channel attention networks. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 286–301.
12. Niu, B.; Wen, W.; Ren, W.; Zhang, X.; Yang, L.; Wang, S.; Zhang, X.; Cao, K.; Shen, H. Single image super-resolution via a holistic attention network. In Proceedings of the Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Part XII 16, pp. 191–207.
13. Bhavsar, A.V.; Rajagopalan, A.N. Resolution enhancement in multi-image stereo. *IEee Trans. Pattern Anal. Mach. Intell.* **2010**, *32*, 1721–1728. [CrossRef] [PubMed]
14. Jeon, D.S.; Baek, S.H.; Choi, I.; Kim, M.H. Enhancing the spatial resolution of stereo images using a parallax prior. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 1721–1730.
15. Wang, L.; Wang, Y.; Liang, Z.; Lin, Z.; Yang, J.; An, W.; Guo, Y. Learning parallax attention for stereo image super-resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 12250–12259.
16. Ying, X.; Wang, Y.; Wang, L.; Sheng, W.; An, W.; Guo, Y. A stereo attention module for stereo image super-resolution. *IEEE Signal Process. Lett.* **2020**, *27*, 496–500. [CrossRef]
17. Chan, K.C.; Zhou, S.; Xu, X.; Loy, C.C. BasicVSR++: Improving video super-resolution with enhanced propagation and alignment. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 5972–5981.
18. Wang, Y.; Isobe, T.; Jia, X.; Tao, X.; Lu, H.; Tai, Y.W. Compression-Aware Video Super-Resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 2012–2021.

19. Lu, Y.; Wang, Z.; Liu, M.; Wang, H.; Wang, L. Learning Spatial-Temporal Implicit Neural Representations for Event-Guided Video Super-Resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 1557–1567.
20. Li, G.; Ji, J.; Qin, M.; Niu, W.; Ren, B.; Afghah, F.; Guo, L.; Ma, X. Towards High-Quality and Efficient Video Super-Resolution via Spatial-Temporal Data Overfitting. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023.
21. Xia, B.; He, J.; Zhang, Y.; Wang, Y.; Tian, Y.; Yang, W.; Van Gool, L. Structured sparsity learning for efficient video super-resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 22638–22647.
22. Tu, Z.; Li, H.; Xie, W.; Liu, Y.; Zhang, S.; Li, B.; Yuan, J. Optical flow for video super-resolution: A survey. *Artif. Intell. Rev.* **2022**, *55*, 6505–6546. [CrossRef]
23. Imani, H.; Islam, M.B.; Wong, L.K. A new dataset and transformer for stereoscopic video super-resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 706-715.
24. Han, K.; Xiao, A.; Wu, E.; Guo, J.; Xu, C.; Wang, Y. transformer in transformer. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 15908–15919.
25. Xu, Y.; Wei, H.; Lin, M.; Deng, Y.; Sheng, K.; Zhang, M.; Tang, F.; Dong, W.; Huang, F.; Xu, C. transformers in computational visual media: A survey. *Comput. Vis. Media* **2022**, *8*, 33–62. [CrossRef]
26. Liu, Y.; Zhang, Y.; Wang, Y.; Hou, F.; Yuan, J.; Tian, J.; Zhang, Y.; Shi, Z.; Fan, J.; He, Z. A survey of visual transformers. *IEEE Trans. Neural Netw. Learn. Syst.* **2023**. [CrossRef] [PubMed]
27. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 11–17 October 2021; pp. 10012–10022.
28. Liang, J.; Cao, J.; Sun, G.; Zhang, K.; Van Gool, L.; Timofte, R. Swinir: Image restoration using swin transformer. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 11–17 October 2021; pp. 1833–1844.
29. Ying, X.; Wang, L.; Wang, Y.; Sheng, W.; An, W.; Guo, Y. Deformable 3D convolution for video super-resolution. *IEEE Signal Process. Lett.* **2020**, *27*, 1500–1504. [CrossRef]
30. Zhang, T.; Gu, Y.; Huang, X.; Yang, J.; Yang, G.Z. Disparity-constrained stereo endoscopic image super-resolution. *Int. J. Comput. Assist. Radiol. Surg.* **2022**, *17*, 867–875. [CrossRef] [PubMed]
31. Allan, M.; Mcleod, J.; Wang, C.; Rosenthal, J.C.; Hu, Z.; Gard, N.; Eisert, P.; Fu, K.X.; Zeffiro, T.; Xia, W.; et al. Stereo correspondence and reconstruction of endoscopic data challenge. *arXiv* **2021**, arXiv:2101.01133.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.