SwiniPASSR: Swin Transformer based Parallax Attention Network for Stereo Image Super-Resolution

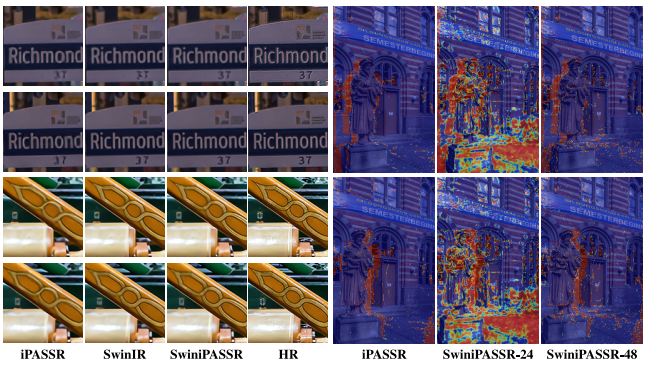
Kai Jin[1] Zeqiang Wei[2] Angulia Yang[1] Sha Guo[3]
Mingzhi Gao[1] Xiuzhuang Zhou[2*†] Guodong Guo[4]

Bigo Technology Pte. Ltd.[1]
Smart Medical Innovation Lab, Beijing University of Posts and Telecommunications[2]
Institute of Digital Media, Peking University[3] Institute of Deep Learning, Baidu Research[4]

{jinkai, yangying.angulia, gaomingzhi}@bigo.sg
{weizeqiang, xiuzhuang.zhou}@bupt.edu.cn
sandyKwok@pku.edu.cn Guodong.Guo@mail.wvu.edu

Abstract

*With binocular cameras being widely accepted, the study of stereo image super resolution (Stereo SR) has received increasing attention. Different from single image super resolution (SISR) setting, it is more challenging for utilizing both intra-view and cross-view information. Although prior convolution-based works have achieved admirable progress, few attempts have explored the possibility of the Transformer-based architecture for stereo image SR, which has demonstrated promising performance in several visual tasks. In this paper, we propose a novel approach namely SwiniPASSR, which adopts Swin Transformer as the backbone, meanwhile incorporating it with the Bi-directional Parallax Attention Module (biPAM) to maximize auxiliary information given by the binocular mechanism. Even Transformer and parallax attention mechanism (PAM) have been separately proved usefulness by prior studies, we find that simply integrating convolution-based PAM with Transformer or directly optimizing for stereo SR problem was may not achieve desirable result. We therefore introduced a conversion layer to resolve integration and adopted progressive training strategy to learn disparity correspondence through progressively enlarged receptive fields. Both extensive experiments and ablation studies demonstrate the effectiveness of our proposed SwiniPASSR. In particular, in the NTIRE 2022: Stereo Image Super-Resolution Challenge, we report **23.71dB PSNR** and 0.7295 SSIM performance which ranked **2nd place** on the leaderboard. Source*



(a) reconstruction comparison (b) valid mask comparison

Figure 1. Visualization of solid mask heatmaps and reconstruction outputs for different approaches: (a) by visualizes details restored by different approaches, for either multi-view information issue or structural reconstruction issue, we could see from the marked regions that SwiniPASSR achieves most satisfactory fine-grained detail restoration. (b) by comparing the valid mask heatmaps between iPASSR and SwiniPASSR, we observe that SwiniPASSR with patch size 24 fails to precisely recognize precise occlusions and boundaries (the salient areas) but patch size 48 is able to capture boundaries well.

code is available at https://github.com/SMI-Lab/SwinIPASSR.

1. Introduction

Super Resolution (SR) is one of the long-lasting low-level vision tasks, whose objective is to reconstruct the fine-grained high-resolution (HR) image from degraded low-resolution (LR) ones. For a long time, extensive convolution-based methods [6, 25, 45] have been proposed to address low-level vision problems and achieved excel-

* Corresponding author
† This work was supported by the National Natural Science Foundation of China under grants 61972046, and in part by the Beijing Natural Science Foundation under grants 4202051.

lent performance. Recently, Transformer-based methods have also shown impressive capability to model global interactions via self-attention mechanism and present promising results on various low-level vision tasks [16, 42], while few attempts has been made on stereo image SR problem. For another, the photographic imaging also embraces upgrading, dual cameras have been widely adopted by a large amount of devices, such as mobile phones, drones, and autonomous vehicles, by which people capture stereo images more easily. SR will benefit from extra information on stereo image pairs, therefore the stereo SR has drawn increasing attention from both industries and academics.

After witnessing success of deep-based single image super resolution works, researchers started to discover the Stereo image Super Resolution task [9, 15, 23, 39, 47]. However it is challenging to incorporate superb stereo correspondence due to varied parallax among objects, disparity at different depths, cross-view information incorporation or potential occlusions. Several prior works made their attempts to address the above-mentioned issues: Wang *et al*. [33] proposed an attention module for parallax learning, while further extended such scheme by tactfully using symmetry among stereo images in iPASSR [36]; Ying *et al*. [43] introduced a generic stereo attention module (SAM) to interact cross-view information; and Yan *et al*. [41] designed a Feature Modulation Dense Block (FMDB) adaptively inserted into the network with disparity prior.

Previous studies worked closely on various attention mechanisms to address the disparity incorporation and obtained splendid quantitative results, but modeling long-range stereo correspondence dependency has not been well studied, which leads restoration deficiency in global textural and detailed reference. As show in Fig. 1, even existing state-of-the-art iPASSR [36] fails to restore textures and sharped lines on the vein effectively. Although Transformer-based architectures offer an possibility and recent methods have achieved impressive performance on other low-level vision tasks [16, 42], it is still unknown how to combine the merit of both Transformer and parallax attention mechanism.

Thus in this paper, we propose a unified framework called SwiniPASSR for the stereo SR problem. Inspired by prior work SwinIR [16] for image restoration, we design the siamese neural network without any downsampling opeartion instead composed of residual Swin Transformer blocks (RSTB) [19], where we incorporate parallax attention mechanism by using the Bi-directional Parallax Attention Module (biPAM) [36] that has proved success on learning stereo correspondence. However, simply integrating the biPAM with RSTB is unsatisfactory, the performance could even be worse than directly using SwinIR for the stereo SR problem. Experiments find that the placement of parallax attention module in the network hierarchy is critical and it also requires a feature translation between convolution-based biPAM and patch-based RSTB. To solve this issue, we introduce layer conversion and place it in the middle of RSTB for using well-representation features to compute stereo correspondence and benefiting following feature fusion as shown in Fig. 2. While directly optimizing SwiniPASSR end up with sub-optimal result due to the difficulty of jointly learning low-level signal restoration and disparity estimation, therefore we propose a progressive training strategy by progressively enlarging training receptive fields in multi-stage to simmer the network and achieve a promising improvement. In summary, our proposed method contributes as follows:

- We propose a Swin Transformer based parallax attention network for the stereo SR problem that can well model global texture and local image details while yielding accurate stereo correspondence.
- We conduct a progressive training strategy to address the joint optimization problem of image restoration and disparity estimation by using gradually expanding patch size during multiple training stages.
- Extensive experiments demonstrate the effectiveness of proposed method, and it obtains state-of-the-art performance with 24.13 dB PSNR and 0.7579 SSIM[1] on validation set of Flickr1024 [35]. Meanwhile, in the NTIRE 2022 Stereo Image Super-Resolution Challenge [32], SwiniPASSR achieves **23.71dB PSNR** and 0.7295 SSIM, ranks **2nd place** in the leaderboard.

2. Related Work

2.1. Stereo Image Super-Resolution

Different from the single image super resolution (SISR) task which utilizes only one LR image to reconstruct HR image, the stereo image super resolution task provides image pairs with stereo viewpoint as auxiliary information for super resolution task, while it upgrades challenges by requiring detail consistency in each high resolution outputs. Considering that the stereo SR task and depth estimation problem are intertwined, [2] first propose an integrated approach to jointly predict the HR depth and the SR image from multiple LR stereo observations, and following traditional methods [26, 27] extend the pipeline but with high computational cost. Recently, due to strong representation ability for vision signals, deep learning based super resolution evolve rapidly. [11] propose a two-stage network by learning parallax prior which enhances spatial resolution remarkably, then multiple works springs up: [14] explores correlative information among stereos images and propose interaction module, DFAM [7], SSRDEFNET [5] exploits

[1]The result uses self-ensemble strategy [17] in inferencing.

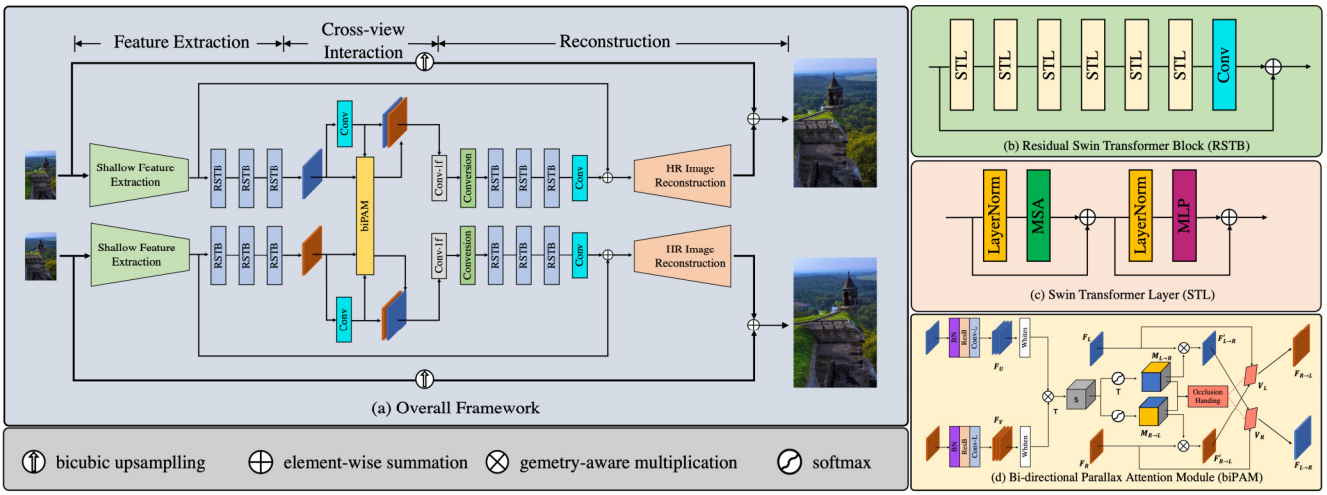


Figure 2. Network Architecture of SwiniPASSR and its core modules. (a) **Overall Framework.** SwiniPASSR can be divided into three part: Feature Extraction, Cross-view Interaction, Reconstruction. Given a pair of LR images, SwiniPASSR extract features for each input from two respective yet parameter sharing branches: a CNN layer acts as shallow feature extractor, followed by 3 Residual Swin Transformer Blocks (RSTB) sequentially, within Feature Extraction part, SwiniPASSR is capable of capturing sufficient features for each viewing angle. As a matter of course, next we utilize Bi-directional Parallax Attention Module (biPAM) to exchange complementary features from each branch, by such Cross-view Interaction our network is able to acquire better stereo correspondence. In the later Reconstruction part, similar consecutive blocks of RSTB remains, equipped with element-wise summation between different feature maps, SwiniPASSR outcome the HR pair. (b) **Residual Swin Transformer Block(RSTB).** a computation block which consists of several Swin Transformer Layer (STL) and element-wise sum operation. (c)**Swin Transformer Layer(STL).** Basic computing unit of RSTB, it could provide commendable feature extraction. (d) **Bi-directional Parallax Attention Module(biPAM).** A flexible attention mechanism that we utilize to capture better stereo correspondence and global context existed in LR features.

disparity features, and PSSR [22] manages to improve perceptual performance and construct a StereoSRQA database. Meanwhile a batch of prior works continuously explore the attention mechanism: PASSRnet [33] propose a parallax-attention module, which boosts the performance by capturing correspondence between stereo images, iPASSR [36] extends the core concept and further propose symmetric bi-directional parallax attention module into their iPASSR. Beyond this, prior works such as SAM [43], CPASSRnet [3], BSSRnet [40], and SPAMnet [29] also further explore different attention modules respectively. Different from prior work iPASSR [36], we focus more on modelling long-range dependency of global textural description based on the characteristic of Swin Transformer.

2.2. Vision Transformer

Not only exemplary performance the Transformer models have achieved on a broad range of natural language processing tasks, but Transformer also reveals its adaptive transferability upon vision. Vision Transformer (ViT) [8] and DeiT [30] as the representative works showcase how Transformer can replace standard convolutions in deep networks, then the following works such as Pyramid ViT [34], Convolutional vision Transformer (CvT) [37], Twins [4], and etc. iteratively update their architecture and proves solid performance in vision recognition task, until Swin Transformer [19], which constructs hierarchical architecture by shifted windows approach, can fulfill promising performance on high-level vision tasks such as recognition, object detection and segmentation. In addition to that, Transformer has been applied on low-level vision task such as super resolution: Yang et al. [42] propose Transformer network (TTSR) for super-resolution, in which it proposes learnable texture extractor and relevance embedding achieve fair performance, considering the prohibitively high computation cost, efficient Transformer based super resolution models SwinIR [16], ESRT [21] come up as well. SwinIR builds upon Swin Transformer, extract features through Swin Transformer blocks (RSTB), conduct state-of-the-art results on multiple low-level vision tasks.

3. Method

In this section, we present our proposed SwiniPASSR as follows: we firstly elaborate the network architecture and the impact of the conversion layer, then we explain the progressive training strategy and learning objects we built.

3.1. Network Architecture

As shown in Fig. 2, the input of the SwiniPASSR network is a pair of low-resolution RGB stereo images I_L^{LR} and I_R^{LR}, then through network computation flow, it outputs the corresponding generated high-resolution RGB stereo im-

ages I_L^{SR} and I_R^{SR}. Referring to Fig. 2(a), SwiniPASSR consists of three modules: feature extraction, cross-view interaction, and reconstruction. And we build two identical branches to symmetrically process I_L^{LR} and I_R^{LR}, in which they share the same parameters and utilize complementary information in each other.

Feature Extraction. Given the stereo image pair $I_L^{LR}, I_R^{LR} \in R^{H\times W\times 3}$, at first we feed them to a 3×3 convolutional layer to extract shallow features $F_L^0, F_R^0 \in R^{H\times W\times C}$, where C is the number feature channels. Then the shallow features with the preliminary perception of images are fed to consecutive $\lfloor K/2 \rfloor$ cascade Residual Swin Transformer Blocks (RSTB) [16], to extract features $F_L, F_R \in R^{H\times W\times C}$. Where K is the number of RSTB blocks in the overall architecture, $\lfloor\ \rfloor$ is rounded down. The RSTB block is one kind residual block with Swin Transformer layer (STL) and convolutional layers as shown in Fig. 2(b), in RSTB we first extract intermediate features through L STL layers, where L is the number of STL. Then we reuse another convolutional layer before residual skip connection, since additional convolutional layer could enhance the translational equivalence of the network due to its spatially invariant filter properties. Besides, the design of residual skip connections provides reconstruction modules with identity-based connections from different RSTB blocks, by which strengthens the aggregation of features at different levels. STL is a variant of the standard multi-head self-attention in Transformer [31], the key differences reside in local attention and the shifted window mechanism, referring details in Fig. 2(c).

Cross-view Interaction. As it is well-known the importance of cross-view information fusion in Stereo SR, we here interact with cross-view information of stereo features F_L, F_R through the bi-directional parallax attention module (biPAM), where we refer to the detailed module structure from [36], the detailed workflow shown in Fig. 2(d). It is worth noting that even though existing works have indicated that hierarchical features are beneficial to stereo correspondence learning, which does not make it to be applicable to Transformer networks. Compared with single-layer features, concatenating the hierarchical features as inputs does not bring significant gain, so as shown in Fig. 2(a), we only use the features after $\lfloor K/2 \rfloor$ RSTB as the input of the biPAM module. In the final, through the module, we can obtain left and right attention maps $\{M_{R\to L}, M_{L\to R}\}$, valid masks $\{V_L, V_R\}$, and cross-view interaction features $\{F_{R\to L}, F_{L\to R}\}$. With the simple yet appropriate adjustment, the cross-view interaction module will achieve the best feature interaction effect at the position.

Reconstruction. As we've mentioned multiple times, HR reconstructions including complete information and refine details are essential in the whole process. Here, we firstly concatenate conversion layer generated features with their correpsonding cross-view interaction features $\{F_L, F_{R\to L}\}, \{F_R, F_{L\to R}\}$ in each branch. Then subsequent feature integration is performed by a 3×3 convolution and restores the number of feature channels to C. By using the conversion layer, the regularization transformation is performed on the integration features. Then, we symetrically use $\lceil K/2 \rceil$ RSTB exactly as feature extraction module, acting as the base blocks of the reconstruction module, here $\lceil\ \rceil$ represents rounded up. Finally, the super-resolution images I_L^{SR}, I_R^{SR} are generated by a 1×1 convolution and sub-pixel convolution layers [28].

3.2. Conversion layer

Because the bi-PAM and the subsequent fusion modules are convolutional-based operations, features are inherent in consistent distribution among mini-batch via using batch normalization [10], which is different from Transformer-based feature that is more inclined to establish intra-sample dependencies and maintain independence between samples. To this end, we introduce a conversion layer by a learnable layer normalization [1] for strengthening the internal connection of samples. Experiments demonstrate that directly feeding the fused feature into subsequent RSTB hardly converges, which also proves the necessity of conversion layer. This conversion design usually is used in Transformer and Convolutional fusion works, [18, 38] also exist similar design module.

3.3. Progressive training strategy

Since learning parallax attention requires accurate structured information, we propose a multi-stage training procedure namely progressive training strategy, which could guide the SwiniPASSR to learn better stereo correspondence by enlarging the training patch size gradually. Different from HERN [24] that adds resolution without any parameter changing or ProGAN [12] that grows resolution and trainable layers, the proposed method simultaneously changes the network structure, patch size and training strategy. We will explain the procedure in the following.

SwinIR-S/M. In the first stage, the stereo SR task is regarded as a SISR task by training a SwinIR with L1 loss and 24×24 patch size. The trained network could construct well-represented low-level features with precise textural and structural information. As shown in SwinIR-M visualized outputs Fig. 5, those fine lines and boundaries are clearer and more sharp than iPASSR.

SwiniPASSR-S1/M1. In the second stage, on one side two branches process LR input pairs simultaneously, share weights of the shallow, deep extractor, and keep low-level feature representation for reconstruction. On the other side, the biPAM module is introduced to connect them for constructing better stereo correspondence and learning to handle both occlusions and fuzzy boundaries. As quantitative

results, SwiniPASSR-M1/S1 achieves higher PSNR refer to Tab. 3, but it fails to handle occlusions or restore clear boundaries as shown in Fig. 4.

SwiniPASSR-S2/M2. In the last stage, to better utilize cross-view information and facilitate occlusions handling ability, we enlarge the training patch size to as large as 48×48 because training patch size larger than 48×48 is unable to fit into GPU card for limited 11G GPU memory. In addition, we use smaller learning rate and adopt exponential moving average (EMA) technique to stabilize training process in this stage. As shown in Fig. 4, compared with the second stage results of SwiniPASSR-S1/M1, the SwiniPASSR-S2/M2 model in the last stage presents a well-defined stereo correspondence.

3.4. Objectives

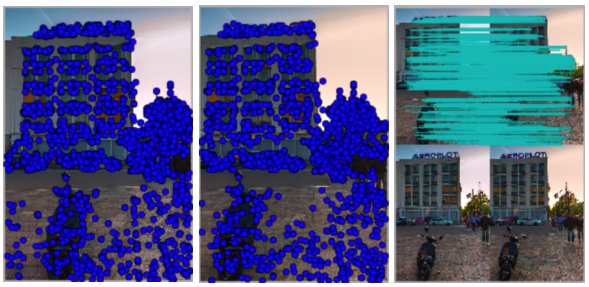
To construct parallax attention mechanism (PAM), we introduce biPAM module and its related loss function, which includes L1 pixel loss for super resolution reconstruction, photometric loss for illumination robustness, cycle loss for consistency, smooth loss for stereo correspondence, and consistency loss for super-resolved stereo consistency. Thus our ultimate optimization objective is defined as $\mathcal{L} = \mathcal{L}_{rec} + \lambda_{photo} * \mathcal{L}_{photo} + \lambda_{cycle} * \mathcal{L}_{cycle} + \lambda_{smooth} * \mathcal{L}_{smooth} + \lambda_{cons} * \mathcal{L}_{cons}$, where λ_{photo}, λ_{cycle}, λ_{smooth}, and λ_{cons} denotes the corresponding weight of each loss item, respectively. On general in this part we greatly refer to prior work [36], more detailed explanation could be found there.

4. Experiments

In this section, We firstly illustrate the stereo SR benchmark dataset used for the contest and our experimental settings in Sec. 4.1. Then in Sec. 4.3, we justify the efficiency of extra augmentation and characteristics of the parallax attention map under different patch sizes. Finally, in Sec. 4.2 and Sec. 4.4, we thoroughly compare our proposed method with both previous state-of-the-art and winning solutions from other teams.

4.1. Implementation Details

Dataset. Flickr1024 [35] as a commonly used benchmark dataset, provides 800 training pairs, 112 validation pairs and 112 test pairs. NTIRE 2022 Stereo Image Super-Resolution Challenge followed the same training and validation settings, but changed the testing set to 100 pairs. Flickr1024 includes plenty of images from various real-life scenarios and even synthesis scenarios, in which the image sizes range from minimum 123×198 to maximum 675×525. Considering variance appeared in image contents, we preliminarily compute SIFT features [20] on the images, in order to search for matching parts and obtain



(a) (b) (c)

Figure 3. Parallax illustration on training dataset. (a) and (b) are corresponding to SIFT computed feature of a stereo image pair. (c) is the matching points and warped image between two views.

disparity values. Visualized outcomes of SIFT are illustrated in Fig. 3a and Fig. 3b, the matching points are more concentrated on edges and corners, yet less on the texture, while Fig. 3c indicates that paired points under the same depth tend to have similar disparity values. Besides, based on SIFT analysis, we also find that there is merely vertical disparity in the dataset, instead there is a 2.5 pixel average disparity in the horizontal direction for LR training pairs.

Evaluation Metrics. For evaluation, Peak Signal to Noise Ratio (PSNR) and Structural Similarity Index (SSIM) indices are used to estimate the restoration level between SR images and HR images in the RGB domain. Furthermore, for NTIRE 2022 Stereo Image Super-Resolution Challenge, the final leaderboard is ranked by PSNR calculated in the test dataset.

Model Setting. Our proposed SwiniPASSR uses suffix **-S** denotes small-sized network and suffix **-M** denotes medium-sized network. For the two different model-sized networks, the RSTB number, STL number, window size, channel number and attention head number are set to 8/12, 6/6, 12/12, 180/180 and 6/10, respectively.

Training Details. Refer to Sec. 3.3, progressive training requires three stages training with different hyper-parameters. During the first training stage, we use HR training patches size of 96×96, and the sizes of the corresponding LR patches are 24×24 for 4× SISR. Also, training patches are augmented with clockwise rotating of 90, 180, and 270 degrees and flipping horizontally, then randy selected. In addition to above, extra RGB channel shuffling augmentation is also applied for photometric consideration. For optimization, the Adam optimizer is used and set to $\beta_1 = 0.9$, $\beta_2 = 0.999$, with initial learning rate as 1e-4. During training, total 16 samples in a batch are equally distributed to 4 parts scattered on 4 Nvidia RTX 2080Ti GPUs. At 250k, 375k, and 450k iterations, we decay the learning rate to half of the former value, and stop training this stage when it completes total 500k iterations.

During the second training stage, we continue to use a pair of LR patches with 24×24 and HR patches with

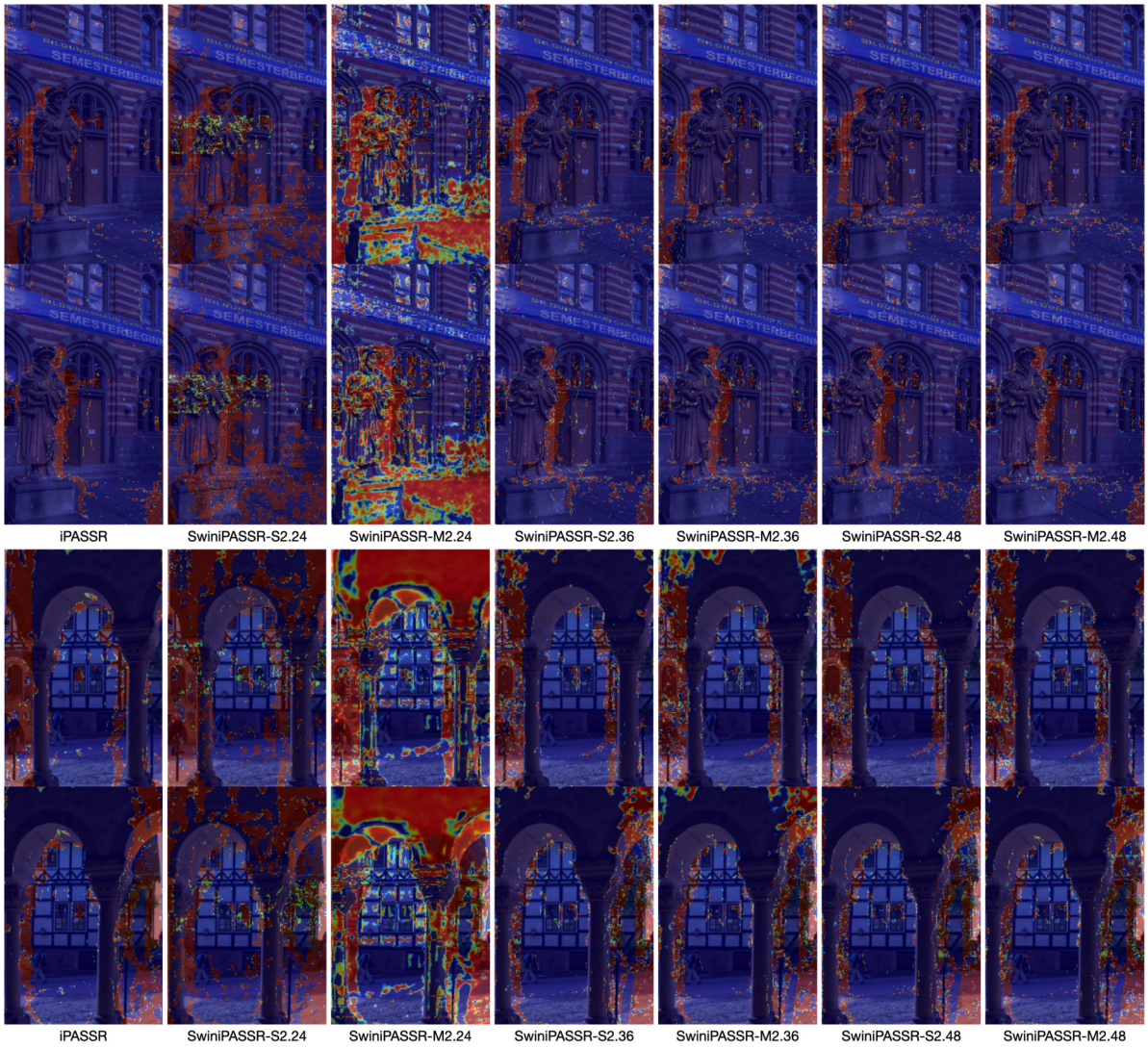


Figure 4. Valid mask heatmaps across different approaches over '0026' and '0080' validation images. iPASSR uses 30x90 patch size as training, and variants of SwiniPASSR employ 24, 36 and 48 patch sizes for training. Distinct red-colored area means high possibility of occlusion and boundaries. It is obvious that larger training patch size visualize more accurate valid mask heatmaps based on occlusion handling scheme from iPASSR.

96×96, but LR and HR patches will have exactly the same coordinate positions and apparent disparity values. In this stage, we reduce augmentations to only clockwise of 180 degrees and corresponding flipping vertically to keep parallax attention maps' consistency. Similarly, RGB channel shuffling scheme is used. All training hyper-parameters settings remain the same as the first training stage. Except that, we will use parallax losses in this stage, hence the hyperparameters mentioned in Sec. 3.4, λ_{SR}, λ_{photo}, λ_{cycle}, λ_{smooth}, and λ_{cons} coefficients are set to 1.0, 0.1, 0.1, 0.01, and 0.1, respectively.

During the last training stage, we expand the LR patches to 36×36 and 48×48 and corresponding HR patches to 144×144 and 192×192 for learning better parallax relationships. Also, the EMA technique is used for stabilizing training process and improving model robustness. Note that due to the limitation of GPU memory, we reduce the batch size to 8 for SwiniPASSR-M2 model with 48×48 LR patch size.

4.2. Comparison to state-of-the-art methods

Tab. 1 shows the quantitative results among SwiniPASSR and state-of-the-arts. Here we includes VDSR [13], EDSR [17], RDN [46], RCAN [45], and SwinIR [19]. We retrain two SwinIR models with different model sizes on Flickr1024 dataset, training details refer to Sec. 4.1. Four stereo SR, approaches such as PASSRNet [33], SR-Res+SAM [44], iPASSR [36] are included. Note that our implementation of SwinIR and SwiniPASSR use all the

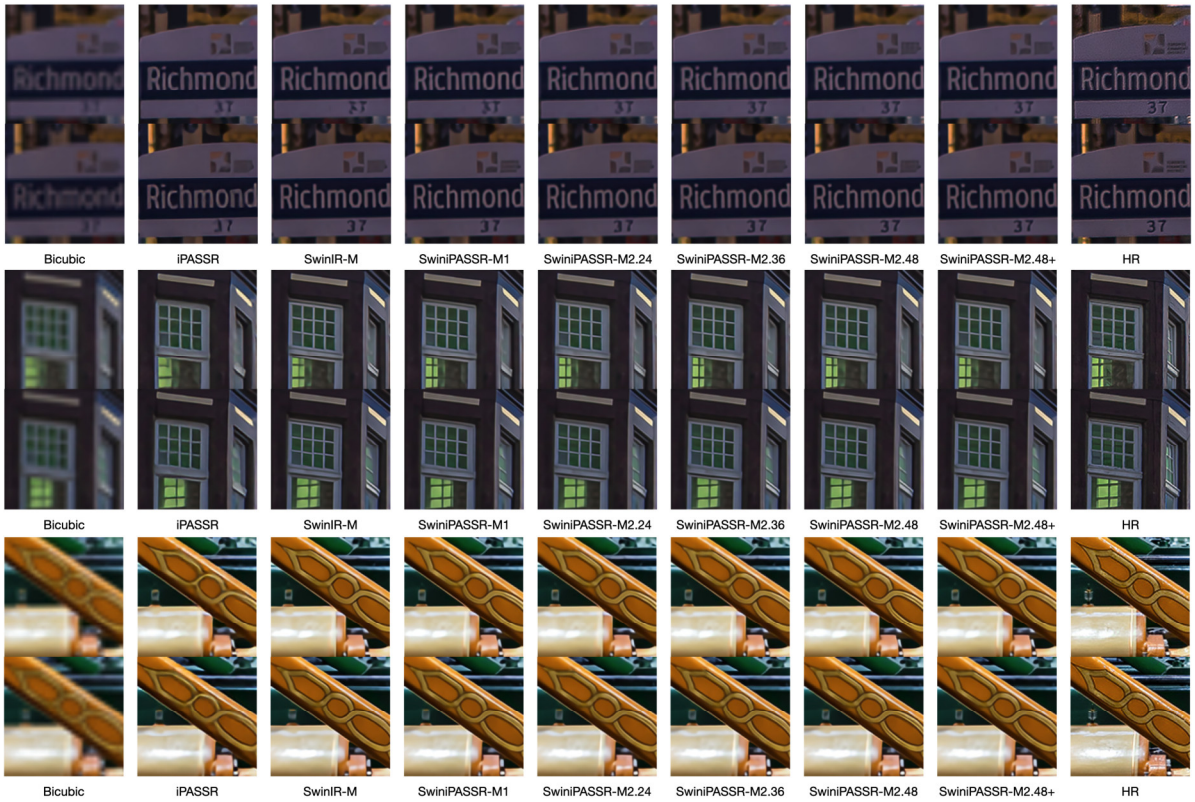


Figure 5. Qualitative validation samples visualization. iPASSR and SwinIR methods sometimes fail to restore edges or fine-grained texture of objects. Our proposed SwiniPASSR-M2 with patch size 48 model enhances image details and alleviate the blur condition.

Table 1. Quantitative results across different super resolution approaches. Params stands for the volumer of parameters. PSNR and SSIM are computed by RGB image and reported in terms of a pair of stereo images. Methods with * represent using Flickr1024 and Middlebury mixed dataset for training.

Method	#Params.	PSNR	SSIM
*Bicubic	/	21.82	0.6293
*VDSR	0.66M	22.46	0.6718
*EDSR	38.9M	23.46	0.7285
*RDN	22.0M	23.47	0.7295
*RCAN	15.4M	23.48	0.7286
SwinIR-S	14.95M	23.81	0.7444
SwinIR-M	21.20M	23.84	0.7450
*PASSRnet	1.42M	23.31	0.7195
*SRRes+SAM	1.73M	23.27	0.7233
*iPASSR	1.42M	23.44	0.7297
SwiniPASSR-S2	16.55M	24.00	0.7549
SwiniPASSR-M2	22.81M	24.05	0.7560
SwiniPASSR-M2†	22.81M	**24.13**	**0.7579**

mentioned augmentation techniques while other compared works only use randomly horizontal and vertical flipping. Compared with SwinIR-S/M, SwiniPASSR-S2/M2 has extra 1.6M parameters but the performance gains 0.19 dB and 0.21 dB improvements as well, which proves the efficiency of our approach in addressing stereo super resolution issue. Also, by further using self-ensemble strategy refer to [17] during inference, SwiniPASSR-M2† achieves highest 24.13 dB on PSNR and 0.7579 on SSIM.

Table 2. Ablation studies about the position of biPAM in RSTBs. Bottom, middle and top denote placing the biPAM at either the first, third, or last RSTB block respectively. Listed methods do not use rotation, RGB shuffling, or dropout augmentation.

Method	Training	Position	PSNR
SwiniPASSR-S1	directly	bottom	21.70
SwiniPASSR-S1	directly	top	23.67
SwiniPASSR-S1	directly	middle	23.75
SwiniPASSR-S1	progressively	middle	23.82

4.3. Ablation Study

Incorporation. As described in Tab. 2, we explore the network performance under three fusion strategies of the bottom layer, the middle layer and the top layer. It turns out that top-level fusion works the worst since the fused information has little chance to change the well-represented reconstructed features. The bottom-level fusion has a certain

Table 3. Quantitative results from different model sizes and training stages on validation dataset. Patch size stands for size of the square-croped patches from images.

Method	Patch Size	PSNR	SSIM
SwinIR-S	24	23.81	0.7444
SwiniPASSR-S1	24	23.92	0.7483
SwiniPASSR-S2	24	23.97	0.7524
SwiniPASSR-S2	36	23.99	0.7542
SwiniPASSR-S2	48	24.00	0.7549
SwinIR-M	24	23.84	0.7450
SwiniPASSR-M1	24	23.95	0.7499
SwiniPASSR-M2	24	23.98	0.7504
SwiniPASSR-M2	36	24.04	0.7556
SwiniPASSR-M2	48	24.05	0.7560

Table 4. Quantitative test results of Top 10 Teams for NTIRE 2022 Challenge on Stereo Image Super-Resolution Challenge.

Team Name	PSNR
The Fat, The Thin and The Young	23.7873
BigoSR	**23.7126**
NUDT-CV&CPLab	23.6007
BUPT-PRIV	23.5983
NKU_caroline	23.5770
BUAA-MC2	23.5733
No War	23.5664
GDUT_506	23.5601
DSSR	23.5533
xiaozhazha	23.5490

effect, while the features entering the biPAM are rudimentary and lack sufficient structural information. Mid-level fusion utilizes well-represented features to estimate disparity correspondences, providing better reference information for subsequent feature learning.

Model Size. As shown in Tab. 3, experimental results demonstrate larger model size tends to have a general improvement across different training stage. For SwinIR-based models, medium sized network SwinIR-S has 0.03 dB higher PSNR than small one SwinIR-M. When it progressively enlarges patch size to 48, the final staged SwinIR-M2 can achieve 24.05 dB PSNR, remarkably higher than SwinIR-S2 by 0.05 dB, which proves a better representative ability of larger model size.

Patch Size. As shown in Fig. 4, even SwiniPASSR-S2/SwiniPASSR-M2 with patch size 24 present higher quantitative results as 23.97 dB and 23.98 dB, they still fail to finely reconstruct the valid mask, meanwhile iPASSR with patch size 30×90 has clearer and more accurate valid mask heatmap. For the first and second rows in Fig. 4, SwiniPASSR-S2/SwiniPASSR-M2.24 give inaccurate valid mask over low-frequency texture area and scattered responses around edges and corners. By enlarging the training patch size from 24 to 36 or even 48, SwiniPASSR-S2 and SwiniPASSR-M2 gain a remarkable improvement from 23.97 and 23.98 dB (24) to 24.00 and 24.05 dB (48), respectively.

Progressive Training. Proposed SwiniPASSR is supposed to address parallax issue in an unsupervised learning alike fashion, which leads difficulty to simultaneously restore low-level image signals and model stereo correspondence from cross-view information. As shown in Tab. 3, the progressive training method enhance performance of the small network from 23.81 dB to 24.00 dB on PSNR and helps medium network from 23.84 dB to 24.05 dB similar visualization results from Fig. 5, compared to intermediate model results, restored super resolution images by SwiniPASSR-M2.48 reflect the more identifiable text of '37' for left view and it presents more sharp black lines according to the last row in Fig. 5.

4.4. NTIRE 2022 Stereo SR Challenge Results

The top-10 results selected by NTIRE 2022 committee are shown in Tab. 4. Our method finally ranked 2nd place and obtained 23.7126 dB on Flickr1024 test dataset. In the challenge, all participants are required not to use any external model and data, including pre-trained backbone and optical flow network, hence our results are acquired from Flickr1024 dataset solely. As for final reporting metrics, we utilize the average ensemble method to combine multiple SwiniPASSR outputs which are produced by different models trained on various patch sizes or model sizes. It is also worth noting that identical ensemble technique is implemented over validation dataset and acquires 24.1557 dB on PSNR and 0.7574 on SSIM, which is merely 0.025 dB higher than our best single model.

5. Conclusion

In this paper, we propose a unified framework namely SwiniPASSR to better fulfill the stereo SR task. By introducing the conversion layer with biPAM and placing it into RSTBs carefully, the siamese-like architecture could effectively model global textures, refined details, and accurate stereo correspondences. Due to the difficulty of simultaneously optimizing image reconstruction and disparity estimation, we propose a progressive training strategy to learn stereo correspondences from progressively enlarged receptive fields. Extensive ablation studies demonstrate the effectiveness of proposed method. In NTIRE 2022: Stereo Image Super-Resolution Challenge, SwiniPASSR achieves 23.71dB PSNR and 0.7295 SSIM and ranks 2nd. In the future, we will focus more on Transformer-based parallax attention mechanism and related optimization strategies.

References

[1] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[2] Arnav V Bhavsar and AN Rajagopalan. Resolution enhancement in multi-image stereo. *IEEE transactions on pattern analysis and machine intelligence*, 32(9):1721–1728, 2010.

[3] Canqiang Chen, Chunmei Qing, Xiangmin Xu, and Patrick Dickinson. Cross parallax attention network for stereo image super-resolution. *IEEE Transactions on Multimedia*, PP:1–1, 01 2021.

[4] Xiangxiang Chu, Zhi Tian, Yuqing Wang, Bo Zhang, Haibing Ren, Xiaolin Wei, Huaxia Xia, and Chunhua Shen. Twins: Revisiting spatial attention design in vision transformers, 04 2021.

[5] Qinyan Dai, Juncheng Li, Qiaosi Yi, Faming Fang, and Guixu Zhang. Feedback network for mutually boosted stereo image super-resolution and disparity estimation. pages 1985–1993, 10 2021.

[6] Tao Dai, Jianrui Cai, Yongbing Zhang, Shu-Tao Xia, and Lei Zhang. Second-order attention network for single image super-resolution. pages 11057–11066, 06 2019.

[7] Jiawang Dan, Zhaowei Qu, Xiaoru Wang, and Jiahang Gu. A disparity feature alignment module for stereo image super-resolution. *IEEE Signal Processing Letters*, PP:1–1, 06 2021.

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 10 2020.

[9] Chenyang Duan and Nanfeng Xiao. Parallax-based second-order mixed attention for stereo image super-resolution. *IET Computer Vision*, 16(1):26–37, 2022.

[10] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. PMLR, 2015.

[11] Daniel Jeon, Seung-Hwan Baek, Inchang Choi, and Min Kim. Enhancing the spatial resolution of stereo images using a parallax prior. pages 1721–1730, 06 2018.

[12] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

[13] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1646–1654, 2016.

[14] Jianjun Lei, Zhe Zhang, Xiaoting Fan, Bolan Yang, Xinxin Li, Ying Chen, and Qingming Huang. Deep stereoscopic image super-resolution via interaction module. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(8):3051–3061, 2021.

[15] Huiling Li, Qiong Liu, and You Yang. Sa-gnn: Stereo attention and graph neural network for stereo image super-resolution. In *International Conference on Image and Graphics*, pages 400–411. Springer, 2021.

[16] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1833–1844, 2021.

[17] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 136–144, 2017.

[18] Yun Liu, Guolei Sun, Yu Qiu, Le Zhang, Ajad Chhatkuli, and Luc Van Gool. Transformer in convolutional neural networks. *arXiv preprint arXiv:2106.03180*, 2021.

[19] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. pages 9992–10002, 10 2021.

[20] David Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60:91–, 11 2004.

[21] Zhisheng Lu, Hong Liu, Juncheng Li, and Linlin Zhang. Efficient transformer for single image super-resolution, 08 2021.

[22] Chenxi Ma, Bo Yan, Weimin Tan, and Jiang Xuhao. Perception-oriented stereo image super-resolution. pages 2420–2428, 10 2021.

[23] Li Ma and Sumei Li. Enhanced back projection network based stereo image super-resolution considering parallax attention. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 1834–1838. IEEE, 2021.

[24] Kangfu Mei, Juncheng Li, Jiajie Zhang, Haoyu Wu, Jie Li, and Rui Huang. Higher-resolution network for image demosaicing and enhancing. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 3441–3448. IEEE, 2019.

[25] Ben Niu, Weilei Wen, Wenqi Ren, Xiangde Zhang, Lianping Yang, Shuzhen Wang, Kaihao Zhang, Xiaochun Cao, and Haifeng Shen. *Single Image Super-Resolution via a Holistic Attention Network*, pages 191–207. 10 2020.

[26] Haesol Park, Kyoung Mu Lee, and Sang Uk Lee. Combining multi-view stereo and super resolution in a unified framework. In *Proceedings of The 2012 Asia Pacific Signal and Information Processing Association Annual Summit and Conference*, pages 1–4. IEEE, 2012.

[27] Hee Seok Lee and Kuoung Mu Lee. Simultaneous super-resolution of depth and images using a single camera. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 281–288, 2013.

[28] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1874–1883, 2016.

[29] Wonil Song, Sungil Choi, Somi Jeong, and Kwanghoon Sohn. Stereoscopic image super-resolution with stereo consistent feature. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34:12031–12038, 04 2020.

[30] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers distillation through attention, 12 2020.

[31] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[32] Longguang Wang, Yulan Guo, Yingqian Wang, Juncheng Li, Shuhang Gu, and Radu Timofte. Ntire 2022 challenge on stereo image super-resolution: Methods and results. In *CVPRW*, 2022.

[33] Longguang Wang, Yingqian Wang, Zhengfa Liang, Zaiping Lin, Jungang Yang, Wei An, and Yulan Guo. Learning parallax attention for stereo image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12250–12259, 2019.

[34] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. pages 548–558, 10 2021.

[35] Yingqian Wang, Longguang Wang, Jungang Yang, Wei An, and Yulan Guo. Flickr1024: A large-scale dataset for stereo image super-resolution. In *International Conference on Computer Vision Workshops*, pages 3852–3857, Oct 2019.

[36] Yingqian Wang, Xinyi Ying, Longguang Wang, Jungang Yang, Wei An, and Yulan Guo. Symmetric parallax attention for stereo image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 766–775, 2021.

[37] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. Cvt: Introducing convolutions to vision transformers. pages 22–31, 10 2021.

[38] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. Cvt: Introducing convolutions to vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22–31, 2021.

[39] Wangduo Xie, Jian Zhang, Zhisheng Lu, Meng Cao, and Yong Zhao. Non-local nested residual attention network for stereo image super-resolution. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2643–2647. IEEE, 2020.

[40] Qingyu Xu, Longguang Wang, Yingqian Wang, Weidong Sheng, and Xinpu Deng. Deep bilateral learning for stereo image super-resolution. *IEEE Signal Processing Letters*, PP:1–1, 03 2021.

[41] Bo Yan, Chenxi Ma, Bahetiyaer Bare, Weimin Tan, and Steven Hoi. Disparity-aware domain adaptation in stereo image restoration. pages 13176–13184, 06 2020.

[42] Fuzhi Yang, Huan Yang, Jianlong Fu, Hongtao Lu, and Baining Guo. Learning texture transformer network for image super-resolution. pages 5790–5799, 06 2020.

[43] Xinyi Ying, Yingqian Wang, Longguang Wang, Weidong Sheng, Wei An, and Yulan Guo. A stereo attention module for stereo image super-resolution. *IEEE Signal Processing Letters*, PP:1–1, 02 2020.

[44] Xinyi Ying, Yingqian Wang, Longguang Wang, Weidong Sheng, Wei An, and Yulan Guo. A stereo attention module for stereo image super-resolution. *IEEE Signal Processing Letters*, 27:496–500, 2020.

[45] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European conference on computer vision (ECCV)*, pages 286–301, 2018.

[46] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2472–2481, 2018.

[47] Xiangyuan Zhu, Kehua Guo, Hui Fang, Liang Chen, Sheng Ren, and Bin Hu. Cross view capture for stereo image super-resolution. *IEEE Transactions on Multimedia*, 2021.