ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-6327-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSP49357.2023.10096174

PFT-SSR: PARALLAX FUSION TRANSFORMER FOR STEREO IMAGE SUPER-RESOLUTION

Hansheng Guo[1] *Juncheng Li*[2,3*] *Guangwei Gao*[4] *Zhi Li*[5] *Tieyong Zeng*[1*]

[1] The Chinese University of Hong Kong, Hong Kong, China; [2]Shanghai University, Shanghai, China
[3]Jiangsu Key Laboratory of Image and Video Understanding for Social Safety, Nanjing, China
[4]Nanjing University of Posts and Telecommunications, Nanjing, China
[5]East China Normal University, Shanghai, China

ABSTRACT

Stereo image super-resolution aims to boost the performance of image super-resolution by exploiting the supplementary information provided by binocular systems. Although previous methods have achieved promising results, they did not fully utilize the information of cross-view and intra-view. To further unleash the potential of binocular images, in this letter, we propose a novel Transformer-based parallax fusion module called Parallax Fusion Transformer (PFT). PFT employs a Cross-view Fusion Transformer (CVFT) to utilize cross-view information and an Intra-view Refinement Transformer (IVRT) for intra-view feature refinement. Meanwhile, we adopted the Swin Transformer as the backbone for feature extraction and SR reconstruction to form a pure Transformer architecture called PFT-SSR. Extensive experiments and ablation studies show that PFT-SSR achieves competitive results and outperforms most SOTA methods. Source code is available at `https://github.com/MIVRC/PFT-PyTorch`.

***Index Terms*—** Stereo Image Super-Resolution, Parallax Fusion Transformer, Stereo Cross Attention, SSR.

1. INTRODUCTION

Binocular cameras have been widely employed to improve the perception capabilities of vision systems in devices such as self-driving vehicles and smartphones. With the rapid development of binocular cameras, stereo image super-resolution (SSR) is becoming increasingly popular in academia and industry. Specifically, SSR attempts to reconstruct a high-resolution (HR) image from a pair of low-resolution (LR) images. With the help of additional information from a pair of binocular images at the same physical location, making full use of the information from both two images is crucial for stereo image super-resolution (SSR).

The easiest way to implement stereo image SR is to perform single image SR (SISR) methods [1–6] on stereo image pairs, respectively. These approaches, however, neglected the cross-view information between the pair of images and are incapable of reconstructing high-quality images. To address this problem, current strategies have focused on building novel cross-view feature aggregation modules, loss functions, and so on, to improve the efficiency with which image pair interaction features are used. For example, [7] first combined depth estimation and image resolution tasks with multiple image inputs. After that, StereoSR [8] took the lead in introducing CNN into Stereo SR. iPASSR [9] suggested a symmetric bi-directional parallax attention module (biPAM) and an inline occlusion handling scheme as its cross view interaction module to exploit symmetry cues for stereo image SR. Recently, several more advanced strategies for improving Stereo SR performance have been introduced. For instance, NAFSSR [10] designed a new CNN-based backbone NAFNet [11] and proposed a novel Stereo Cross Attention Module (SCAM) as parallax fusion block. These network topologies typically included a CNN backbone for obtaining intra-view information and a parallax fusion module for combining cross-view attention. Since the existence of parallax, we discovered that it is also highly crucial for cross-picture features and intra-picture features to promote each other in the process of binocular feature fusion. However, these two processes in existing works are often relatively independent, which is not conducive to the full use of image features. Meanwhile, the quality of the input features is vital for image fusion efficiency. However, existing works never consider the degree of match between the backbone networks and parallax fusion blocks. Therefore, the combination of these two pieces will be sub-optimal.

In this work, we address the aforementioned problems by introducing the Transformer to stereo image SR. Recently, Transformer demonstrated strong performance in various low-level tasks [12–14], which can learn global information

Corresponding author: Juncheng Li, Tieyong Zeng

This work was supported in part by the National Key Research and Development Program of China under Project no. 2021YFE0203700, 2018AAA0100102, and 2018AAA0100100, the National Natural Science Foundation of China under Grant no. 61972212, and the YangFan Project of Shanghai under Grant no. 23YF1412800.

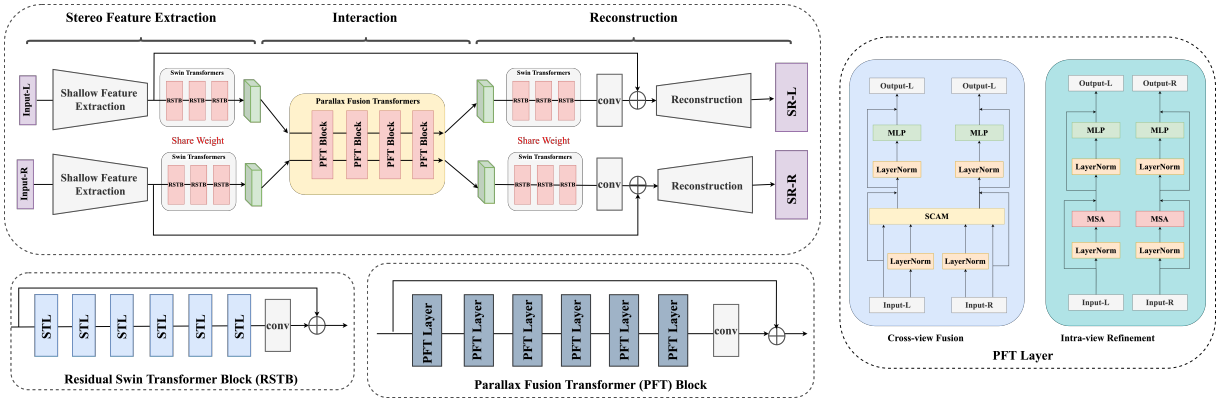

Fig. 1: The complete architecture of the proposed Parallax Fusion Transformer for Stereo Image Super-Resolution (PFT-SSR). This is a dual-stream network and interacts through an interaction module. **Due to page limit, please zoom in to see details.**

of images to further improve model performance. However, directly merging current CNN-based parallax fusion modules (PFM) and Transformer will not result in outstanding performance. This is because the CNN-based parallax fusion modules and Transformer have different properties, leading to PFM that cannot fully utilize the features from the Transformer backbone. To address this issue, we designed a new parallax fusion module, named Parallax Fusion Transformer (PFT). PFT contains a Stereo Cross Attention Module (SCAM) and a Feature Refining Module (FRM). Among them, the SCAM gets the cross-view attention and FRM will fuse the cross-view feature with the local window features. The cross-view features and intra-view features (local window features) will enhance each other to get a better representation for image super-resolution task. With the help of PFT, the proposed model can well-adapt the deep features with the parallax feature fusion blocks to fully utilize the representational potential of the Transformer.

The contributions of this letter can be summarized as follows: 1) We propose a novel Parallax Fusion Transformer (PFT) layer with a Cross-view Fusion Transformer (CVFT) and an Intra-view Refinement Transformer (IVRT). 2) Based on the proposed PFT, we design a pure Transformer network (named PFT-SSR) to further improve the feature extraction ability of Transformer-based networks. 3) Extensive experiments have illustrated the effectiveness of PFT-SSR.

2. METHODOLOGY

In this paper, we propose a Parallax Fusion Transformer for Stereo Image Super-Resolution, called PFT-SSR. As shown in Fig. 1, the proposed PFT-SSR consists of three parts: stereo feature extraction, feature interaction, and SR image reconstruction. For Stereo SR, the model takes two images $x_{LR}^L, x_{LR}^R \in R^{B\times C_{in}\times H\times W}$ as inputs and then outputs x_{HR}^L, $x_{HR}^R \in R^{B\times C_{out}\times S*H\times S*W}$. Among them, B, C_{in}, C_{out}, H, and W are the input batch size, the number of channels, height, and weight, respectively. Meanwhile, S is the upscaling factor, which is used to control the size of the output images. Specifically, we first use two convolutional layers to extract shallow features of the input images respectively. After that, we further extract the deeper feature representations with SwinIR [13] backbone, which contains three consequent Residual Swin Transformer Blocks (RSTBs)

$$I_d^L = f_{ex}(f_s(x_{LR}^L)), \quad I_d^R = f_{ex}(f_s(x_{LR}^R)). \tag{1}$$

Then, the extracted features are fed into the proposed Parallax Fusion Transformers (PFT) for cross-view interaction and intra-view refinement

$$I_f^L, I_f^R = f_{PFT}(I_d^R, I_d^L). \tag{2}$$

With fused features, we apply RSTBs again to obtain the refined features, with a residual connection from the shallow image feature (ignored in formula for simplicity).

$$I_r^L = f_{cov}(f_{re}(I_f^L)), \quad I_r^R = f_{cov}(f_{re}(I_f^R)). \tag{3}$$

Finally, a Reconstruction module that contains a single convolutional layer and a PixelShuffle layer is used to reconstruct the final SR images.

2.1. Swin Transformer Backbone

In this work, we use Swin Transformer Blocks [15] to build the backbone of our network. Specifically, a Swin Transformer Layer firstly reshapes the input feature map I_{in} to $\frac{HW}{M^2} \times M^2 \times C$ and performs standard self-attention locally on each window. For each of $\frac{HW}{M^2}$ feature maps, let input be $X \in R^{M^2\times C}$, then query, key, and value should be

$$Q = XP_Q, \quad K = XP_K, \quad V = XP_V, \tag{4}$$

where P_Q, P_K, and P_V are linear projection matrices. Then, the attention matrix is calculated within the local windows

$$Attention(Q, K, V) = SoftMax(QK^T/\sqrt{d} + B)V, \tag{5}$$

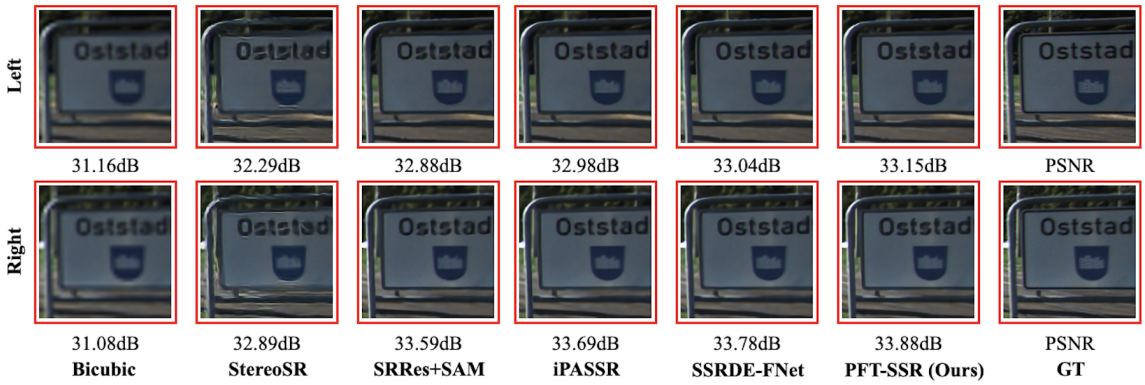

Fig. 2: Visual results (x4) achieved by different methods on Flickr1024.

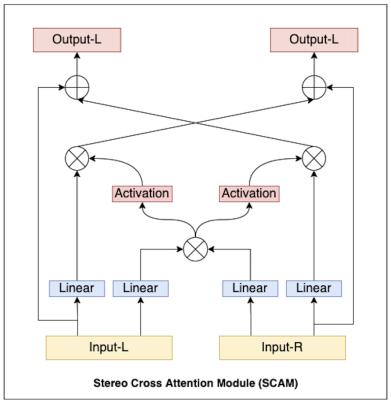

Fig. 3: The architecture of stereo cross attention module.

where B is the positional encoding for Transformer. The model also apply an MLP with two fully connected layers and GELU non-linearity on the attention matrix for feature transformations. Meanwhile, the LayerNorm [16] layer is added before both Attention Block and MLP with residual connection. Though local attention can greatly reduce the amount of computation, there is no connection across local windows. To solve this problem, Swin Transformer proposed a shifted window mechanism to shifts the feature map by $(\lfloor \frac{M}{2} \rfloor, \lfloor \frac{M}{2} \rfloor)$ pixels before partitioning. The process can be expressed as

$$X = MSA(LN(X)) + X, X = MLP(LN(X)) + X \quad (6)$$

where regular partitioning and shift partitioning are used alternately before each MSA. With the help of this backbone, our model can extract sufficient useful image features.

2.2. Parallax Fusion Transformer

In order to make full use of the features of the left and right images, we propose a Parallax Fusion Transformer (PFT). As shown in Fig. 1, PFT contains 4 PFT blocks, and each PFT block consists of 6 PFT layers and a convolutional layer. Meanwhile, each PFT layer has two different Transformer blocks, i.e. Cross-view Fusion Transformer (CVFT) and Intra-view Refinement Transformer (IVRT). Among them, CVFT adopts stereo cross-attention module (SCAM [10]) to learn the features of another view and IVRT takes the local-window Transformer to better merge features from the other view to its feature map. Specifically, we first apply CVFT to achieve cross-view attention via SCAM. However, using single-head SCAM to get the cross-view information cannot adapt to different parallax. Therefore, we further use IVRT to make cross-view information from the other branch better interact with intra-view features. With this 'Attention-Refine' paradigm, our PFT-SSR shows a compelling effect on cross-view attention.

Cross-view Fusion Transformer (CVFT): The core component of CVFT is SCAM, and the whole process of SCAM is shown in Fig. 3. Given input image features $X_L, X_R \in R^{H \times W \times C}$, we first perform layer normalization to get scaled features. Due to the nature of stereo images, we use the same Q and K for representing intra-view features. Then, we get cross-view attention both from right to left and from left to right by

$$\begin{aligned} F_{R \to L} &= Attention(T_1^L \overline{X_L}, T_1^R \overline{X_R}, T_2^R \overline{X_R}), \\ F_{L \to R} &= Attention(T_1^R \overline{X_R}, T_1^L \overline{X_L}, T_2^L \overline{X_L}), \end{aligned} \quad (7)$$

where $Attention$ is defined same as Eq. (5). Besides, T_1^L, T_1^R, T_2^L, and T_2^R are linear projection matrices. After getting the cross-view attention feature, we use a weighted residual connection to merge it to the corresponding image feature, which are formulated as

$$Y_L = \alpha_L F_{R \to L} + X_L, Y_R = \alpha_R F_{L \to R} + X_R, \quad (8)$$

where α_L and α_R are learnable scalars. After observing the corrected features, we apply MLP and LayerNorm to get the final outputs and the whole process can be expressed as

$$X = SCAM(X) + X, X = MLP(LN(X)) + X. \quad (9)$$

Intra-view Refinement Transformer (IVRT): One key difficulty of SSR is the different parallax brought by various

Table 1: Quantitative comparison on different datasets. PSNR/SSIM values achieved on both the left images (i.e., *Left*) and a pair of stereo images (i.e., $(Left + Right)/2$) are reported. Among them, the best results are **highlighted**.

Method	Scale	*Left*			$(Left + Right)/2$			
		KITTI 2012	KITTI 2015	Middlebury	KITTI 2012	KITTI 2015	Middlebury	Flickr1024
EDSR [3]	×2	30.83/0.9199	29.94/0.9231	34.84/0.9489	30.96/0.9228	30.73/0.9335	34.95/0.9492	28.66/0.9087
RCAN [17]	×2	30.88/0.9202	29.97/0.9231	34.80/0.9482	31.02/0.9232	30.77/0.9336	34.90/0.9486	28.63/0.9082
StereoSR [18]	×2	29.42/0.9040	28.53/0.9038	33.15/0.9343	29.51/0.9073	29.33/0.9168	33.23/0.9348	25.96/0.8599
PASSRnet [19]	×2	30.68/0.9159	29.81/0.9191	34.13/0.9421	30.81/0.9190	30.60/0.9300	34.23/0.9422	28.38/0.9038
iPASSR [9]	×2	30.97/0.9210	30.01/0.9234	34.41/0.9454	31.11/0.9240	30.81/0.9340	34.51/0.9454	28.60/0.9097
SSRDE-FNet [20]	×2	31.08/**0.9224**	30.10/**0.9245**	35.02/0.9508	31.23/**0.9254**	30.90/**0.9352**	35.09/0.9511	28.85/**0.9132**
PFT-SSR (Ous)	×2	**31.15**/0.9166	**30.16**/0.9187	**35.08**/**0.9516**	**31.29**/0.9195	**30.96**/0.9306	**35.21**/**0.9520**	**29.05**/0.9049
EDSR [3]	×4	26.26/0.7954	25.38/0.7811	29.15/0.8383	26.35/0.8015	26.04/0.8039	29.23/0.8397	23.46/0.7285
RCAN [17]	×4	26.36/0.7968	25.53/0.7836	29.20/0.8381	26.44/0.8029	26.22/0.8068	29.30/0.8397	23.48/0.7286
StereoSR [18]	×4	24.49/0.7502	23.67/0.7273	27.70/0.8036	24.53/0.7555	24.21/0.7511	27.64/0.8022	21.70/0.6460
PASSRnet [19]	×4	26.26/0.7919	25.41/0.7772	28.61/0.8232	26.34/0.7981	26.08/0.8002	28.72/0.8236	23.31/0.7195
SRRes+SAM	×4	26.35/0.7957	25.55/0.7825	28.76/0.8287	26.44/0.8018	26.22/0.8054	28.83/0.8290	23.27/0.7233
iPASSR [9]	×4	26.47/0.7993	25.61/0.7850	29.07/0.8363	26.56/0.8053	26.32/0.8084	29.16/0.8367	23.44/0.7287
SSRDE-FNet [20]	×4	26.61/**0.8028**	25.74/**0.7884**	29.29/0.8407	26.70/**0.8082**	26.45/**0.8118**	29.38/0.8411	23.59/**0.7352**
PFT-SSR (Ours)	×4	**26.64**/0.7913	**25.76**/0.7775	**29.58**/**0.8418**	**26.77**/0.7998	**26.54**/0.8083	**29.74**/**0.8426**	**23.89**/0.7277

stereo systems. Although SCAM shows great cross-view attention ability, it cannot adapt various parallax. After observing this, we used a Transformer with local-window attention for feature refinement. Regular partitioning is adopted before the MSA so that the features after the interaction of the two views can be further fused and enhanced, which is helpful for the final SR image reconstruction.

3. EXPERIMENT

3.1. Experimental Settings

800 images from Flickr1024 [21] and 60 images from Middlebury [22] are chosen for training. To make the Middlebury dataset matches the spatial resolution of the Flickr1024 dataset, we perform bicubic downsampling by a factor of 2 on each image. And then, we use bicubic downsampling to these GT images by the factors of 2 and 4 to get the input images. We follow previous works [9, 10, 20] on this setting to make comparison fair. During training, we use the L1 loss function for supervision, PSNR and SSIM as quantitative metrics to make easy comparison with previous methods. These metrics are calculated on RGB color space with a pair of stereo images. To evaluate SR results, we use KITTI 2012 [23], KITTI 2015 [24], Middlebury [22], and Flickr1024 [21] for test.

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

We compare our proposed PFT-SSR with several state-of-the-art methods, including SISR methods (e.g., EDSR [3], RCAN [17]) and stereo image SR methods (e.g., StereoSR [18], PASSRnet [19], iPASSR [9], and SSRDE-FNett [20]). According to TABLE 1, we can clearly observe that our PFT-SSR achieves outstanding results and outperforms most other SOTA methods, especially on Flickr102. Meanwhile, we also show the qualitative comparisons in Figs. 2. Obviously, our PFT-SSR can reconstruct more accurate SR images with more accurate edges and texture details. This fully demonstrates the effectiveness of the proposed PFT-SSR.

Table 2: Ablation study on PFT under Flickr1024.

Backbone	**Module**	**PSNR (x4)**	**SSIM (x4)**
Swin Transformer	None	23.54	0.7120
Swin Transformer	RSTB (SwinIR)	23.65	0.7164
Swin Transformer	BiPAM	23.42	0.7068
Swin Transformer	PFT (Ours)	**23.83**	**0.7268**

3.3. Ablation Study

Cross-view interaction is the key part in Stereo SR. In this part, we do ablation on the choice of this technology to show the strong stereo image fusion ability of the proposed PFT. We use Swin Transformer [15] Blocks as backbones and take the same number of Swin Transformer, biPAM [9], and our proposed PFT as the cross-view interaction module in this part. According to TABLE 2, it is obviously that the proposed PFT can improve the model performance more effectively, which fully illustrates the effectiveness of PFT.

4. CONCLUSION

In this paper, we proposed a PFT-SSR for stereo image super-resolution, which contains a well-designed Parallax Fusion Transformer (PFT). PFT consists of a Cross-view Fusion Transformer (CVFT) and an Intra-view Refinement Transformer (IVRT), specially designed for cross-view interaction. It is worth mentioning that PFT can better merge different parallaxes to utilize the features of the left and right images fully. Meanwhile, PFT can also better adapt to the current popular Transformer-based backbone. Extensive experiments show that PFT-SSR outperforms most current models and achieves promising outcomes.

5. REFERENCES

[1] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang, "Learning a deep convolutional network for image super-resolution," in *ECCV*, 2014, pp. 184–199.

[2] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee, "Accurate image super-resolution using very deep convolutional networks," in *CVPR*, 2016, pp. 1646–1654.

[3] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee, "Enhanced deep residual networks for single image super-resolution," in *CVPR Workshops*, 2017, pp. 136–144.

[4] Tao Dai, Jianrui Cai, Yongbing Zhang, Shu-Tao Xia, and Lei Zhang, "Second-order attention network for single image super-resolution," in *CVPR*, 2019, pp. 11065–11074.

[5] Yukai Shi, Haoyu Zhong, Zhijing Yang, Xiaojun Yang, and Liang Lin, "Ddet: Dual-path dynamic enhancement network for real-world image super-resolution," *IEEE Signal Processing Letters*, vol. 27, pp. 481–485, 2020.

[6] Yongsong Huang, Zetao Jiang, Rushi Lan, Shaoqin Zhang, and Kui Pi, "Infrared image super-resolution via transfer learning and psrgan," *IEEE Signal Processing Letters*, vol. 28, pp. 982–986, 2021.

[7] Arnav V Bhavsar and AN Rajagopalan, "Resolution enhancement in multi-image stereo," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1721–1728, 2010.

[8] Longguang Wang, Yulan Guo, Yingqian Wang, Juncheng Li, Shuhang Gu, Radu Timofte, Liangyu Chen, Xiaojie Chu, Wenqing Yu, Kai Jin, et al., "NTIRE 2022 challenge on stereo image super-resolution: Methods and results," in *CVPR Workshop*, 2022, pp. 906–919.

[9] Yingqian Wang, Xinyi Ying, Longguang Wang, Jungang Yang, Wei An, and Yulan Guo, "Symmetric parallax attention for stereo image super-resolution," in *CVPR Workshop*, 2021, pp. 766–775.

[10] Xiaojie Chu, Liangyu Chen, and Wenqing Yu, "Nafssr: Stereo image super-resolution using nafnet," in *CVPR*, 2022, pp. 1239–1248.

[11] Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun, "Simple baselines for image restoration," *arXiv preprint arXiv:2204.04676*, 2022.

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al., "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[13] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte, "Swinir: Image restoration using swin transformer," in *ICCV*, 2021, pp. 1833–1844.

[14] Zhisheng Lu, Juncheng Li, Hong Liu, Chaoyan Huang, Linlin Zhang, and Tieyong Zeng, "Transformer for single image super-resolution," in *CVPR*, 2022, pp. 457–466.

[15] 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," in *ICCV*, 2021, pp. 10012–10022.

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

[17] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu, "Image super-resolution using very deep residual channel attention networks," in *ECCV*, 2018.

[18] Daniel S Jeon, Seung-Hwan Baek, Inchang Choi, and Min H Kim, "Enhancing the spatial resolution of stereo images using a parallax prior," in *CVPR*, 2018, pp. 1721–1730.

[19] Longguang Wang, Yingqian Wang, Zhengfa Liang, Zaiping Lin, Jungang Yang, Wei An, and Yulan Guo, "Learning parallax attention for stereo image super-resolution," in *CVPR*, 2019, pp. 12250–12259.

[20] Qinyan Dai, Juncheng Li, Qiaosi Yi, Faming Fang, and Guixu Zhang, "Feedback network for mutually boosted stereo image super-resolution and disparity estimation," in *ACMMM*, 2021, pp. 1985–1993.

[21] Yingqian Wang, Longguang Wang, Jungang Yang, Wei An, and Yulan Guo, "Flickr1024: A large-scale dataset for stereo image super-resolution," in *ICCV Workshops*, 2019.

[22] Daniel Scharstein, Heiko Hirschmüller, York Kitajima, Greg Krathwohl, Nera Nešić, Xi Wang, and Porter Westling, "High-resolution stereo datasets with subpixel-accurate ground truth," in *GCPR*, 2014, pp. 31–42.

[23] Andreas Geiger, Philip Lenz, and Raquel Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *CVPR*, 2012, pp. 3354–3361.

[24] Moritz Menze and Andreas Geiger, "Object scene flow for autonomous vehicles," in *CVPR*, 2015, pp. 3061–3070.