Reviewer 2:

This manuscript presents a lightweight hybrid transformer architecture for single-image super resolution (SISR) of remote sensing images. The authors employed existing deep learning modules to improve the SISR performance. The reviewer's comments are as follows.

Comments 1: There are several previous methods reporting superior performance on the UCMerced and AID datasets, compared to the proposed method. It is required to highlight the benefit and main contributions of the proposed method.

Response 1: Thank you very much for your suggestion. We have re-summarized our main contributions and benefit in the following three main points:

1. We propose the Efficient Super-Resolution Hybrid Network (EHNet), a lightweight RSISR network that efficiently fuses CNN and swin transformer within a UNet-like structure. This hybrid model is capable of utilizing both the inductive bias of convolution and the long range modeling capability. On the other hand, the multi-scale capability of unet and skip connection can reconstruct images with richer details.
2. We design a lightweight and efficient convolutional block, as the fundamental unit for image feature extraction. The dual-branch design of CSP enables the integration of features from different stages, aiding the model in understanding and utilizing these varied stage features. In addition, We found that SELayer can also realize channel feature combination with much less computational cost than pointwise convolution.
3. In the decoder, we innovatively propose an upsampling method SUB based on sequence of tokens. Compared with convolution-based upsampling methods, our SUB is more suitable for transformer-based models and can improve image detail recovery capabilities by focusing on semantic information.

Comments 2: It is required to conduct more experiments at least with the scaling factor of 3, following previous SISR literatures on the UCMerced and AID datasets.

Response 2: Thank you very much for your suggestion. We have added comparison

results with the scaling factor of 3 on UCMerced and AID datasets. The new results tables are as follows:

Table 7. Comparative results for the UCMerced dataset.

Method	Scale	Params	FLOPs	PSNR	SSIM
Bicubic	×2	-	-	30.76	0.8789
SRCNN [10]	×2	69K	0.028G	32.84	0.9152
VDSR [11]	×2	671K	0.275G	33.47	0.9234
LGCNet [34]	×2	193K	0.195G	33.48	0.9235
DCM [35]	×2	1.84M	1.299G	33.65	0.9274
CTNet [36]	×2	349K	0.105G	33.59	0.9255
HSENet [37]	×2	5.29M	3.886G	34.22	**0.9327**
TransENet [38]	×2	37.3M	1.012G	34.03	0.9301
SwinIR [28]	×2	3.87M	1.693G	34.15	0.9307
HAT[33]	×2	5.12M	2.247G	34.17	0.9311
EHNet(ours)	×2	2.64M	0.785G	**34.29**	0.9320
Bicubic	×3	-	-	27.46	0.7631
SRCNN [10]	×3	69K	0.028G	28.97	0.8132
VDSR [11]	×3	671K	0.275G	29.75	0.8346
LGCNet [34]	×3	193K	0.195G	29.28	0.8238
DCM [35]	×3	1.84M	1.299G	29.86	0.8393
CTNet [36]	×3	349K	0.105G	29.44	0.8319
HSENet [37]	×3	5.29M	3.886G	30.04	0.8433
TransENet [38]	×3	37.3M	1.012G	29.90	0.8397
SwinIR [28]	×3	3.87M	1.693G	30.12	0.8487
HAT[33]	×3	5.12M	2.247G	**30.15**	**0.8489**
EHNet(ours)	×3	2.64M	0.785G	30.09	0.8465
Bicubic	×4	-	-	25.65	0.6725
SRCNN [10]	×4	69K	0.028G	26.78	0.7219
VDSR [11]	×4	671K	0.255G	27.54	0.7522
LGCNet [34]	×4	193K	0.195G	27.02	0.7333
DCM [35]	×4	2.16M	0.754G	27.22	0.7528
CTNet [36]	×4	360K	0.180G	27.41	0.7512
HSENet [37]	×4	5.43M	4.136G	27.73	0.7623
TransENet [38]	×4	37.46M	2.599G	27.78	0.7635
SwinIR [28]	×4	4.54M	2.801G	27.87	0.7659
HAT[33]	×4	5.29M	2.497G	27.86	0.7683
EHNet(ours)	×4	2.64M	1.205G	**28.02**	**0.7711**

Comparative results for the UCMerced dataset.

Table 8. Comparative results for the AID dataset.

Method	Scale	Params	FLOPs	PSNR	SSIM

Bicubic	×2	-	-	32.39	0.8906
SRCNN [10]	×2	69K	0.028G	34.49	0.9286
VDSR [11]	×2	671K	0.275G	35.11	0.9340
LGCNet [34]	×2	193K	0.195G	34.80	0.9320
DCM [35]	×2	1.84M	1.299G	35.21	0.9366
CTNet [36]	×2	349K	0.105G	35.13	0.9354
HSENet [37]	×2	5.29M	3.886G	35.50	0.9383
TransENet [38]	×2	37.3M	1.012G	35.40	0.9372
SwinIR [28]	×2	3.87M	1.693G	35.35	0.9370
HAT[33]	×2	5.12M	2.247G	35.49	0.9388
EHNet(ours)	×2	2.64M	0.785G	35.42	0.9390
Bicubic	×3	-	-	32.39	0.8906
SRCNN [10]	×3	69K	0.028G	30.55	0.8372
VDSR [11]	×3	671K	0.275G	31.17	0.8511
LGCNet [34]	×3	193K	0.195G	30.86	0.8498
DCM [35]	×3	1.84M	1.299G	31.31	0.8561
CTNet [36]	×3	349K	0.105G	31.16	0.8515
HSENet [37]	×3	5.29M	3.886G	31.49	0.8588
TransENet [38]	×3	37.3M	1.012G	31.50	0.8588
SwinIR [28]	×3	3.87M	1.693G	31.47	0.8600
HAT[33]	×3	5.12M	2.247G	31.53	0.8612
EHNet(ours)	×3	2.64M	0.785G	31.51	0.8609
Bicubic	×4	-	-	27.30	0.7036
SRCNN [10]	×4	69K	0.028G	28.40	0.7561
VDSR [11]	×4	671K	0.255G	28.99	0.7753
LGCNet [34]	×4	193K	0.195G	28.61	0.7626
DCM [35]	×4	2.16M	0.754G	29.17	0.7824
CTNet [36]	×4	360K	0.180G	29.00	0.7768
HSENet [37]	×4	5.43M	4.136G	29.32	0.7867
TransENet [38]	×4	37.46M	2.599G	29.44	0.7912
SwinIR [28]	×4	4.54M	2.801G	29.26	0.7863
HAT	×4	5.29M	2.497G	29.43	0.7921
EHNet(ours)	×4	2.64M	1.205G	29.44	0.7922

Comments 3: The quality of all figures should be improved. It is not able to reproduce the proposed model based on the current version of the figures.

Response 3: Thank you very much for your suggestion. We have redrawn some figures as required. These figures are as follows:

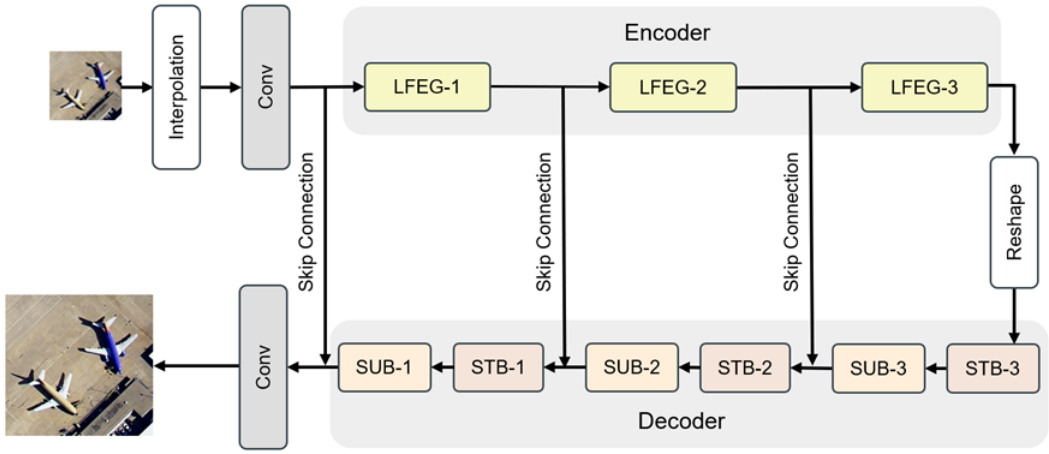


Figure 1. Architecture of the proposed EHNet.

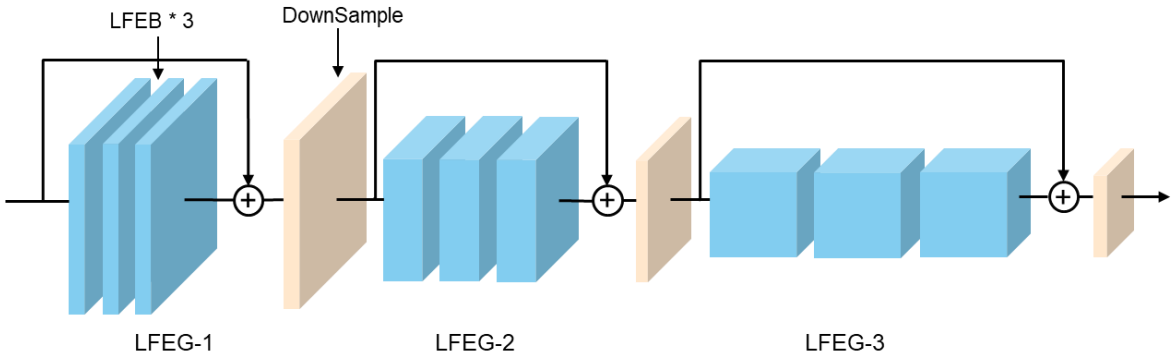


Figure 2. Architecture of encoder in EHNet.

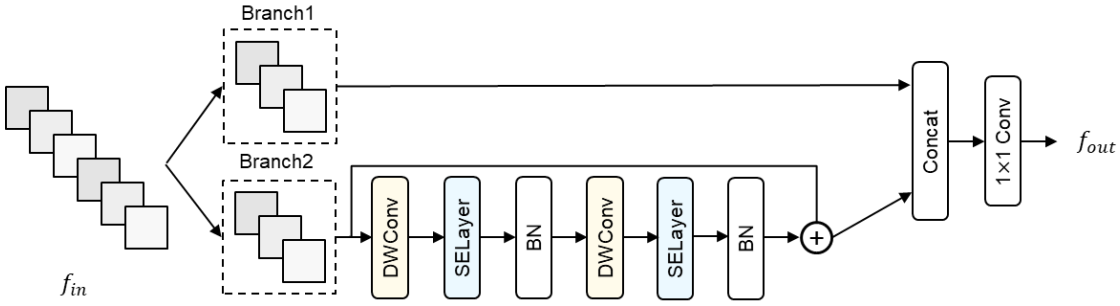


Figure 3. Lightweight Feature Extraction Block(LFEB).

Comments 4: It would be beneficial to explain the motivation of developing lightweight deep learning models for the remote sensing imagery application.

Response 4: Thank you very much for your suggestion. We have added the motivation of developing lightweight deep learning models as follows:

"Existing algorithms usually choose to increase the number of network layers in order

to more adequately extract features from low-resolution images and improve the quality of reconstructed images. However, too many layers may bring other negative effects for the performance of the network. For example, gradient vanishing, network degradation, and overfitting. In recent years, there are some lightweight super-resolution models such as feature enhancement network (FeNet) and Omnisr, all of which have less computational resource consumption but poor reconstruction quality. Larger network models generate higher quality super-resolution images, but consume a large amount of resources. Achieving a balance between reconstruction quality and model complexity is an important goal in image super-resolution research.”