This CVPR workshop paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Thermal Image Super-Resolution Challenge Results - PBVS 2022

Rafael E. Rivadeneira, Angel D. Sappa, Boris X. Vintimilla,
Jin Kim, Dogun Kim, Zhihao Li, Yingchun Jian,
Bo Yan, Leilei Cao, Fengliang Qi, Hongbin Wang
Rongyuan Wu, Lingchen Sun, Yongqiang Zhao, Lin Li,
Kai Wang, Yicheng Wang, Xuanming Zhang, Huiyuan Wei,
Chonghua Lv, Qigong Sun, Xiaolin Tian, Zhuang Jia, Jiakui Hu,
Chenyang Wang, Zhiwei Zhong, Xianming Liu and Junjun Jiang

Abstract

This paper presents results from the third Thermal Image Super-Resolution (TISR) challenge organized in the Perception Beyond the Visible Spectrum (PBVS) 2022 workshop. The challenge uses the same thermal image dataset as the first two challenges, with 951 training images and 50 validation images at each resolution. A set of 20 images was kept aside for testing. The evaluation tasks were to measure the PSNR and SSIM between the SR image and the ground truth (HR thermal noisy image downsampled by four), and also to measure the PSNR and SSIM between the SR image and the semi-registered HR image (acquired with another camera). The results outperformed those from last year's challenge, improving both evaluation metrics. This year, almost 100 teams participants registered for the challenge, showing the community's interest in this hot topic.

1. Introduction

The goal of super-resolution in computer vision is to take a low-resolution image and turn it into a high-resolution image; most techniques used for this purpose are deep learning-based. These methods typically use a downsampled image from the high-resolution image as input, which is then augmented with noise and blur. The resulting image is then used to train the network. Most of these approaches have been used primarily in the visible spectrum, but with the increasing usage of thermal images for various applications, there is a need for methods that can operate in the thermal image domain.

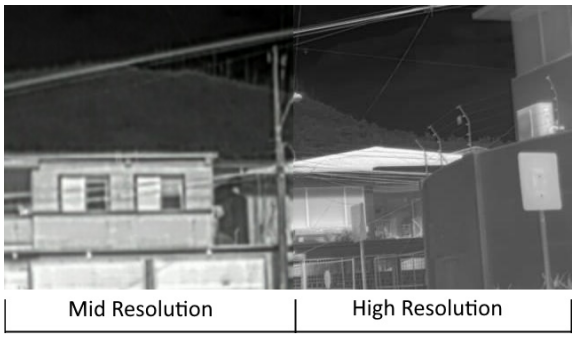


Figure 1: A mosaic illustrating two different resolution thermal images from the same camera viewpoint: $(left)$ a crop from the MR image; $(right)$ a crop from the HR image [11]

A standard benchmark for evaluating different contributions was first proposed at the PBVS 2020 workshop through a challenge on TISR. The success of the two first challenges led to the third challenge on TISR being proposed in the framework of the PBVS 2022 workshop. This third challenge also uses the MR and HR sets of images from the original thermal image dataset.

This TISR 2022 challenge[1] also has the two same evaluation approaches from last challenge [13]. Evaluation 1 measures the $\times 4$ SR result for images from the HR camera. This means that each participant must add noise and downsample the given ground-truth image and use it to train their network. Evaluation 2 compares the $\times 2$ SR results obtained by using input images from the MR camera (Axis Q2901-E). These $\times 2$ SR results are evaluated concerning the corresponding semi-registered images obtained from the HR camera (FLIR FC-632O). So the proposed $\times 2$ scale solu-

Rafael E. Rivadeneira* (rrivaden@espol.edu.ec), Angel D. Sappa*+ and Boris X. Vintimilla* are the TISR Challenge - PBVS 2020 organizers, while the other authors participated in the challenge.
*Escuela Superior Politécnica del Litoral, ESPOL, Guayaquil, Ecuador.
+Computer Vision Center, Campus UAB, 08193 Bellaterra, Barcelona, Spain.
Appendix A contains the authors' teams and affiliations.

[1] https://pbvs-workshop.github.io/challenge.html

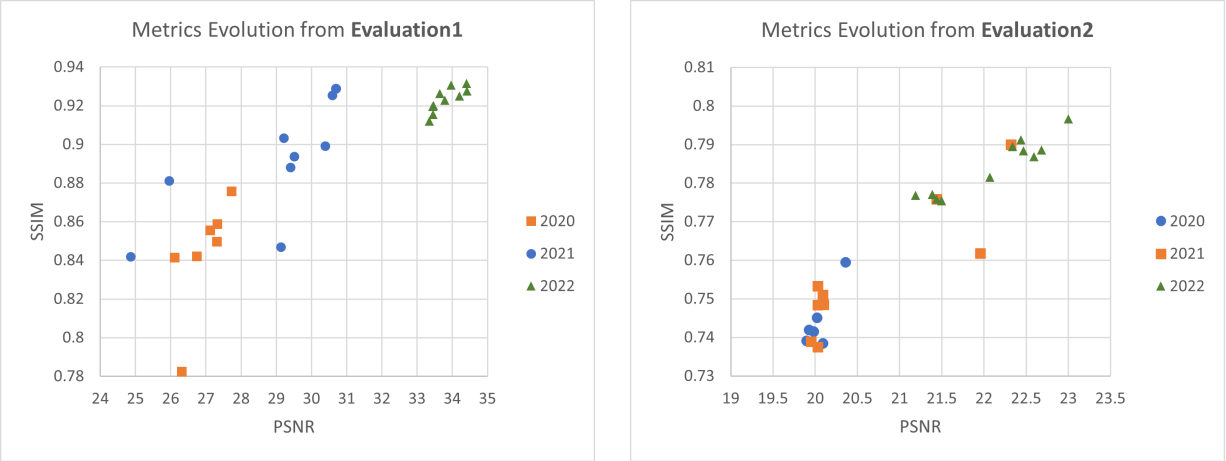


Figure 2: Metrics evolution through the last three years challenges.

tion must be able to tackle both problems, in other words, generating the SR images acquired with the MR camera as well as mapping images from the MR domain to the HR domain. Obtained results show the interest of the community, and each year metrics' values are getting higher as shown in Fig. 2.

The following sections are dedicated to an introduction of the challenge objectives and dataset in Section 2, a summary of the results obtained by different teams in Section 3, and a short description of the different approaches proposed by the teams in Section 4. Section 5 presents the conclusion followed by an appendix with the team information.

2. TISR 2022 Challenge

The TISR 2022 challenge aims to introduce state-of-the-art approaches for the thermal image SR problem, evaluate and compare different solutions using last year's benchmark, promote a novel thermal image dataset to be used as a benchmark by the community, and encourage future research in this area.

2.1. Thermal Image Dataset

The present challenge is based on the dataset introduced in [11], which was used in the first and second TISR challenge ([12], [13]). This dataset consists of 1021 thermal images acquired with three different thermal cameras under various lighting conditions (i.e., day, night, indoor, outdoor), resolutions (120×160, 240×320, 480×640), and object types (i.e., cars, people, vegetation, buildings). The cameras were mounted on a rig in order to minimize the baseline distance between the optical axis, such that the images acquired would be almost registered. Figure 1 presents a mosaic created with images from the MR and HR cameras.

2.2. Evaluation Methodology

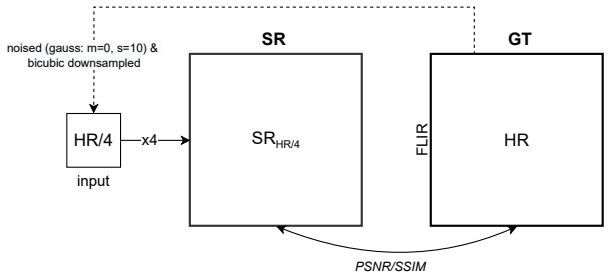
The evaluation methodology is the same as the last challenge. All contributions are evaluated using the mean of the obtained peak signal-to-noise (PSNR) ratio and structural similarity (SSIM) measures. As mentioned, two kinds of evaluations are performed. A set of 10 noisy and downsampled images obtained from an HR camera are evaluated in the first process. Gaussian noise ($\sigma = 10\%$) is added, and then the downsampling process is applied by a scale factor of $\times 4$ to the HR image. Figure 3 (a) shows an illustration of this first evaluation process.

Another set of 10 SR images obtained by a $\times 2$ scale factor from the given MR images is evaluated in the second process. These 10 SR images are evaluated with respect to the corresponding HR GT images (acquired from a different camera with the same resolution as the computed SR). Feature point-based registration is used to align the images. The evaluation on PSNR and SSIM is performed on 80% of the central cropped region of the image. Figure 3 (b) illustrates this second evaluation process.

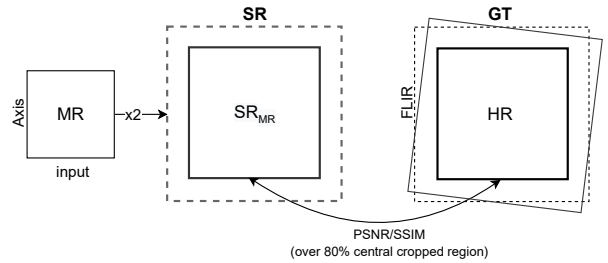
3. Challenge Results

From the initial 100 registered teams, more than 50 teams submitted their results. The top teams with higher results from each evaluation were selected, then submitted their corresponding extended abstracts and reached the final phase. The quantitative average results (PSNR and SSIM) for each team in the two evaluations are shown in Table 1. More quantitative results can be found on CodaLab Competition webpage[2]. Section 4 presents a brief description of the approach proposed by each team to perform SR. Information about the team members and their affiliations is

[2] https://codalab.lisn.upsaclay.fr/competitions/1990



(a) First evaluation process (×4).



(b) Second evaluation process, from MR to HR (×2).

Figure 3: Evaluations processes.

Team Approach	Evaluation 1		Evaluation 2	
	×4		×2 (MR to HR)	
	PSNR	SSIM	PSNR	SSIM
AIR	**34.42**	0.9275	20.63	0.7657
ANT GROUP	33.64	0.9263	21.08	0.7803
NJU	34.41	**0.9316**	20.23	0.7506
NPU-LIFT-LAB	30.19	0.9040	**23.00**	**0.7966**
SENSEXDU	33.57	0.9201	22.68	0.7886
SISYPHUS A.	31.95	0.9165	22.34	0.7896
WZ	33.79	0.9228	22.44	0.7912
XDU-JK	34.20	0.9249	21.50	0.7754

Table 1: Average results for each evaluation of the 2022 TISR challenge (see Section 2.2 for more details). Bold and underline values correspond to the best- and second-best results, respectively.

provided in Appendix A.

4. Proposed Approaches and Teams

4.1. AIR

The present team proposed Convolution Attached Transformer Super-resolution Networks(CATS), as shown in Fig. 4. CATS is composed of *convolutional block(CB)*s and *transformer block(TB)*s. This allows the model to take on both the speed of the CNN-based approach and the restoration performance of the Transformer-based approach. Furthermore, instead of using vanilla Residual Channel Attention Block (RCAB), Detail-Fidelity Attention Module (DeFiAM) [4] was adopted to capture more accurate low-frequency structures and high-frequency details.

Initially, the thermal image passes through five successive *CB*s. Each *CB* is a DeFiAM [4] that extracts low-frequency structure and high-frequency details in a CNN-like manner. Then, it goes through a step consisting of one Swin *TB* [6] and a *CB*, which serves to smoothly transmit information to the subsequent *TB*s. The successive 5 Swin *TB*s can achieve high efficiency as they handle the image information that has been refined once before. At the end of the blocks, the data before passing through each block is concatenated with the processed data through a long skip-connection. Finally, the reconstruction is completed by increasing the spatial size of the data through the *Re-Scale* module and passing it through a convolutional layer that maps the data from the latent space to the image space. Given a pair of reconstructed images extracted from CATS and high-quality (HQ) images, pixel-wise *MSE* loss function is calculated.

All reported implementations were based on PyTorch framework, while the proposed approaches were conducted with 16-Core CPU, 4 × V100 GPU, 32Gib RAM for about three days. Quantitative results shows that this team achieves **34.42** PSNR & 0.9275 SSIM on Evaluation 1, and 20.63 PSNR & 0.7657 SSIM on Evaluation 2.

4.2. ANT GROUP

Inspired in Channel Splitting Network(CSN) [19, 10] and Swin Transformer [8], this team proposed a Bilateral Network with Channel Splitting Network and Transformer (BN-CSNT). This network, as shown in Fig. 5, is designed to tackle the TISR problem. The context branch obtains sufficient context information. The spatial branch, with a shallow transformer, can preserve spatial information. And the attention refinement and feature fusion modules are designed to fuse features.

For the context branch, the input feature maps are passed through the convolution layer with a kernel size of 5×5 and 2N channels, then split N feature-maps input to Swin-Based Block and another N feature-maps input to Attention Refinement Module (ARM). Each Swin-Based Block unit takes N channel features and outputs 2N number of channels. The input N feature-maps to Swin-Based Block are passed through the Swin Basic Layers and output N feature-maps, then N input and output feature-maps are concatenated and passed through the convolution layer with a kernel size of 1×1 and 2N channels. For the output 2N feature maps of Swin-Based Block, split N feature maps to next Swin-Based Block, and another N feature maps to ARM. For the spatial branch, the input feature maps are passed through the convolution layer with a kernel size of 5×5 and N channels and then passed through shallow Swin Ba-

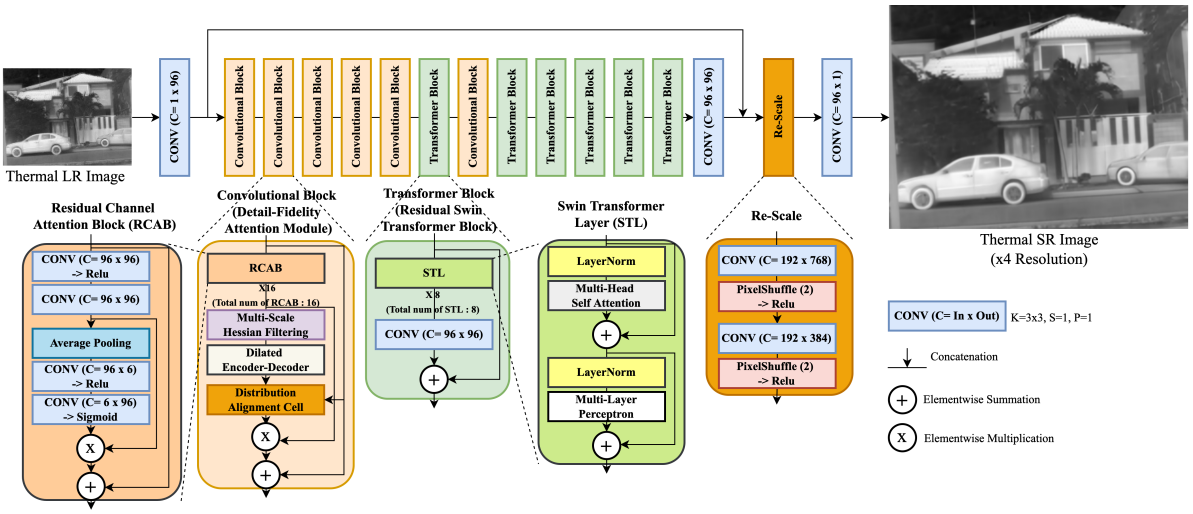


Figure 4: Architecture proposed by AIR team (CATS).

sic Layers to preserve the spatial information. The features obtained from the context branch, ARM, and spatial branch are concatenated and fused by Feature Fusion Module (FFM). Finally, a pixel-shuffle operator to increase the spatial resolution of the feature [14] is employed.

For Evaluation 1, the BN-CSNT network is trained with an upscaling factor of ×4 and $L1$ loss. The inputs are downsampled with a factor ×4, and the outputs are high-resolution FLIR images. For Evaluation 2, the BN-CSNT network is first trained with an upscaling factor of $\times 2$ and $L1$ loss. The inputs are down-sampled with a factor $\times 2$, and the outputs are medium resolution Axis images. Secondly, the BN-CSNT network is trained with an upscaling factor of $\times 2$, Least Squared GAN (LSGAN) loss [9], and $SSIM$ loss. The inputs are semi-matched medium resolution Axis images, and the outputs are high-resolution FLIR images. The average outputs of these two models are the final result for Evaluation 2.

Quantitative results shows that this team achieves 33.64 PSNR & 0.9263 SSIM on Evaluation 1, and 21.08 PSNR & 0.7803 SSIM on Evaluation 2.

4.3. NJU

Considering that thermal image super-resolution is not fundamentally different from RGB image super-resolution, the following team uses swinIR [6] as a based model, as shown in Fig. 6. So to take advantage of the model pretrain on RGB, first triple the number of IR image channels and then feed it into the model. The output of the model was average along the channels. For the training data, the high-resolution image was added with Gaussian noise, used cubic resize to 224 × 224 size, and then simulated the low-resolution image by JPEG compression with a quality factor of 95. Random rotation and flip are added in training data generation. Patch size of 120 was used for SwinIR [6] and finetune 10 epochs using the model pretrained on DF2K [1]. Finally, the three models were integrated using $L1$, $PSNR$, and MSE as the loss.

Quantitative results shows that this team achieves 34.41 PSNR & **0.9316** SSIM on Evaluation 1, and 20.23 PSNR & 0.7506 SSIM on Evaluation 2.

4.4. NPU-LIFT-LAB

The following team uses directly from swinIR [6] for classic $\times 4$ super-resolution (SR) for Evaluation 1. For Evaluation 2, there are three main problems to deal with, which are domain inconsistency, image misalignment and super-resolution respectively. The utilized network architecture is shown in Fig. 7. The method is to generate the more effective labels to guide the learning process. The labels are expected to have the same domain and alignment with the input images.

For convenience, the input low-resolution image acquired with the camera Axis Q2901-E is denoted as x, and the output high-resolution image acquired with the camera Axis Q2901-E is denoted as y. In order to transform domain, the domain mapping (DM) module [18] is used to generate a domain-adjusted image x_t from x and a 2D coordinate map containing the coordinate of the pixels τ, the process of which is guided by the image y^{down} downsampled from y. Note that the location of pixels in x_t is unchanged due to the pixel-wise mapping from x in DM module. In order to align x_t with y, there are three steps. Firstly, x_t is interpolated to get x_t^{up} that has the same size as y. Then PWC-Net [15] is leveraged to estimate the optical flow from x_t^{up} and y. Finally, the estimated optical flow is used to warp y to obtain the updated label y_{warp}. LiteISPNet [18] is used to learn the mapping between the

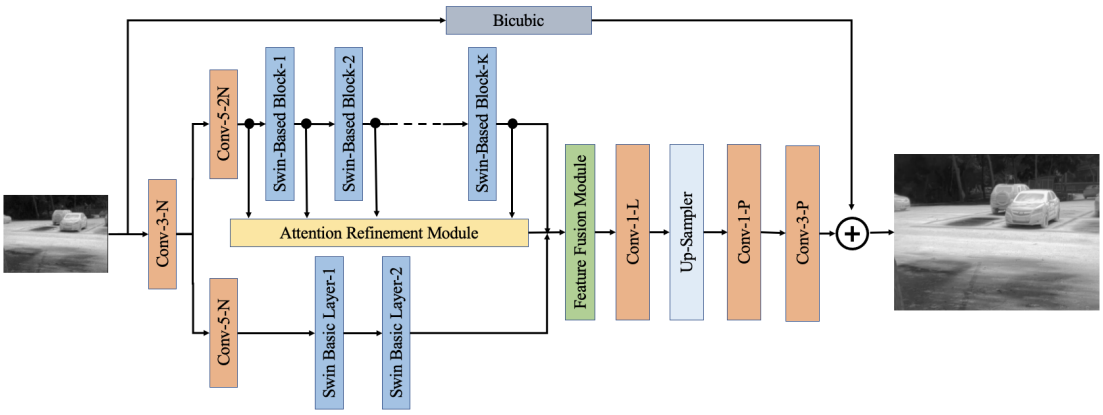


Figure 5: Architecture proposed by ANT GROUP team (BN-CSNT).

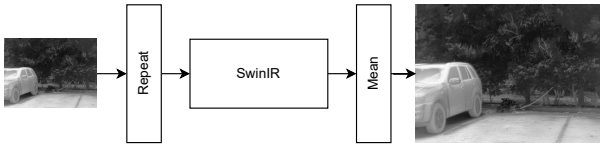


Figure 6: Architecture proposed by NJU team.

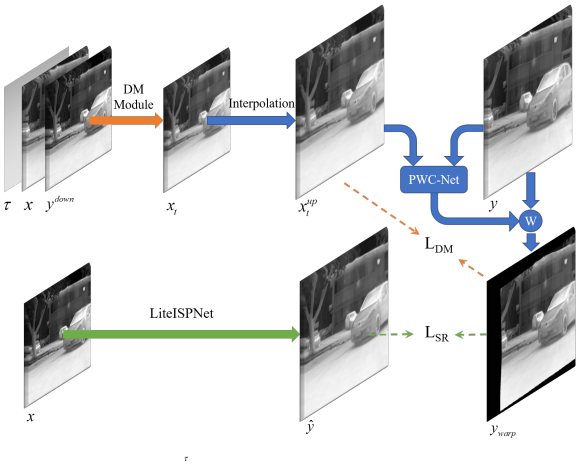


Figure 7: Architecture proposed by NPU-LIFT-LAB.

input x and the updated label y_{warp} to accomplish the final super-resolution task. In conclusion, the designed network architecture can solve the above three main problems.

The joint training is choose to better utilize the information from domain adaption, image alignment and super-resolution. In order to maximize the potential performance of the model, the self-ensemble strategy similarly to [16] was adopted.

Quantitative results shows that this team achieves 30.19 PSNR & 0.9040 SSIM on Evaluation 1, and **23.00** PSNR & **0.7966** SSIM on Evaluation 2.

4.5. SENSEXDU

This team presents two approaches. For Evaluation 1, a noise-squeezing super-resolution network, as shown in Fig. 8 (a), is composed of four modules: the feature extraction (FE) module, the feature refinement (FR) module, the noise squeezing (NS) module, and the up-sampling reconstruction (USR) module. The FE module consists of convolutional layers with multi-level long-skip connections; the FR module consists of multi-level channel attention. The NS module is only used in the training stage; it allows squeezing suppression of the noisy parts of the feature map coming from the FR module and enhancing the noise-free features. The USR module consists of sub-pixel convolutional layers, and the enhanced noise-free features enter the USR module to get high-quality super-resolution results.

For Evaluation 2, a network named Camera Internal Parameters Perceptual Super-Resolution Network (CIPPSR-Net) to deal with SR and misalignment. It is made of three major components, as illustrated in Fig. 8 (b): the Camera Internal Parameter Representation Network (CIPRNet), the U-shaped perceptual network, and the Cycle perceptual network. As the backbone of CIPRNet, ResNet18 gathers the internal information of different cameras through supervised learning. A U-shaped network is created to encode (i.e., DownSamplingBlock, DSB) and decode (i.e., UpSamplingBlock, USB) the thermal image's features and use the Perceptual Module (PM) to sense the camera internal information for low-resolution (LR) cameras. Finally, a reconstruction block is used to reconstruct the super-resolution thermal image. The used optimized are L_1, $SSIM$, $cyclic$ and $contrastive$ losses.

The proposed network has been trained using 4 x NVIDIA TELSA V100 graphics cards. Quantitative results

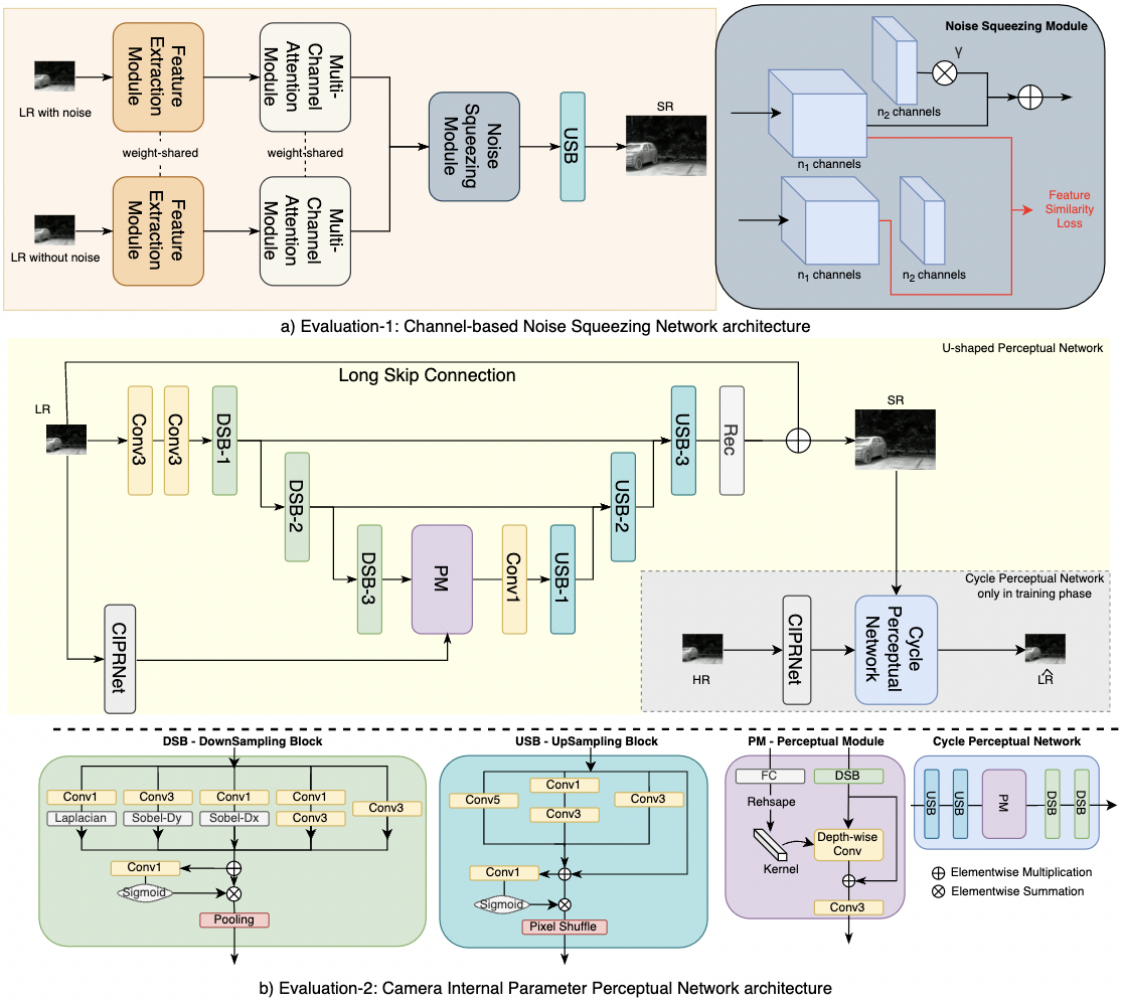


Figure 8: Architecture proposed by SENSEXDU team.

shows that this team achieves 33.47 PSNR & 0.9201 SSIM on Evaluation 1, and 22.68 PSNR & 0.7886 SSIM on Evaluation 2.

4.6. SISYPHUS AKADEMIA

The architecture of the network presented by this team is shown in Fig. 9. A convolutional layer is first used to extract the feature maps from the input noisy LR image; then, the feature maps go through two branches for feature refinement. The first branch on the top is composed of several IMDB blocks [5], which applies recursive channel split and convolutions for only part of channels to achieve information distillation. IMDB block can effectively learn the feature maps for low-level vision tasks with a relatively small number of parameters by exploiting feature map channels' variation. The outputs of each IMDB are concatenated and fused by a 1×1 conv layer and a 3×3 conv layer.

The other branch on the bottom is a series of residual blocks, which are without BN layers according to [7]. This branch serves as an alternative for feature extraction and makes the training more stable. The original features from the first convolutional layer are added to each output from two branches by skip connection. Then the two outputs are concatenated and up-sampled by convolution and pixel-shuffle operation. In order to reduce the learning difficulty, the bicubic up-sample LR image is added to the whole network output.

The parameter settings for Evaluation 1 is as follows: numbers of IMDBs and ResBlocks are 16 and 32 respectively, and their channel numbers are 128 and 64.

Data augmentation (flip, rotation) is used and cropped to 128×128 (HR patch size), AdamW optimizer with learning rate 2e-4 is applied for training. The network is trained for 100k iteration with batch size 128, with a learning rate halved in 10k, 30k, and 70k iterations. In addition, test time augmentation is applied by taking the average outputs from

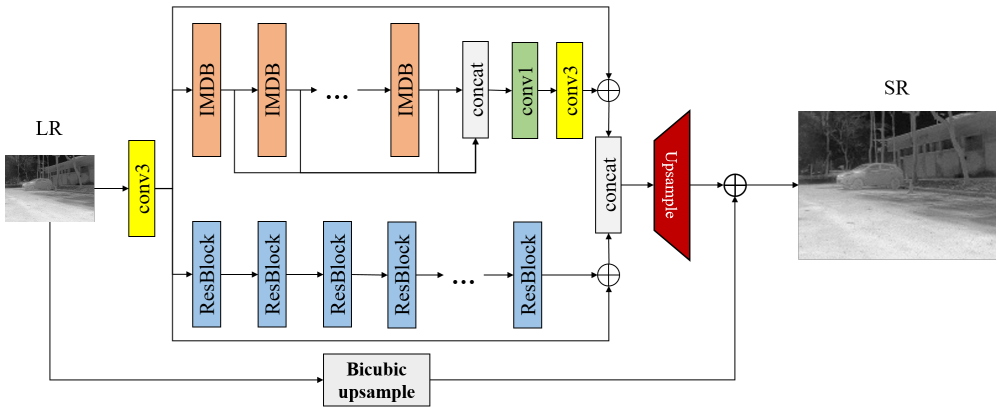


Figure 9: Architecture proposed by SISYPHUS AKADEMIA Team.

h-fliped, v-fliped, and original LR images.

For Evaluation 2, the network structure and parameter settings stay the same, except the up-sample scale is 2 instead of 4. Firstly the ECC Maximization method [3] is used to align the MR to downsampled HR images. The network is trained on the aligned dataset for 50k iterations with batch size 128. AdamW with learning rate 1e-4 and multi-step scheduler, which halves learning rate in 10k and 20k are applied in this experiment.

Python 3.7 and Pytorch 1.10.0 with CUDA 11.3 were used in the experiments. All trainings are done on 1 Nvidia A100 GPU card (40G) of Ubuntu 18.04 server. Quantitative results shows that this team achieves 31.95 PSNR & 0.9165 SSIM on Evaluation 1, and 22.34 PSNR & 0.7896 SSIM on Evaluation2.

4.7. WZ

This Team develops a thermal image super-resolution model using lightweight image super-resolution model IMDN [5] due to the limitation of computational resources. The network is comprised of several cascading in- formation multi-distillation blocks (IMDB) as shown in Fig. 10.

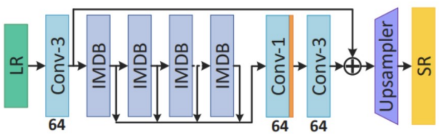


Figure 10: Architecture used by WZ Team from [5].

For ×4 super-resolution, the given dataset is used directly to train the model. For $\times 2$ super-resolution, use the given MR-HR pair and downsampled HR and MR by $\times 2$ to get additional train pair. The model is optimized by ADAM with $\beta 1$ = 0.9, $\beta 2$ = 0.99 and $\epsilon = 1e$ - 8. The learning rate is set as $1e$ - 4 in the training phase, and the mini-batch size of the model is set as 8.

Experiments were conducted on NVIDIA 3090 GPU with popular Pytorch. Quantitative results shows that this team achieves 33.79 PSNR & 0.9228 SSIM on Evaluation 1, and 22.44 PSNR & 0.7922 SSIM on Evaluation 2.

4.8. XDU-JK

The following team uses the same network in both evaluations, which can be used to enlarge the image by four times. Inspired in SwinIR [6], RRDBs in ESRGAN [17] as shallow feature extractor is used and replace the RSTB in SwinIR for deep feature extraction with the designed LKB by the team.

For perceptual field improvements, the features were analyzed by using a large kernel, like 13×13. To compensate for possible drawbacks of large convolution kernels, such as the inability to capture details, it use several 1×1 convolutions to connect residuals before and after each large convolution. The detailed structure of the whole model can be seen in Fig. 11. From the test results, it can be concluded that large convolutional energy can better restore content information, and its PSNR metric is better than most methods, but its ability to reconstruct details is not ideal. The training strategy is WGAN [2], and the discriminator is VGG19. At the same time, two up-sampling options are provided to deal with different situations: pixel-shuffle and nearest before convolution, the former for artificial down-sampled images for Evaluation1, and the latter for super-resolution tasks of real images for Evaluation2.

For Evaluation 1, loss function consisting of $L1Loss$, $SSIMLoss$ and $TVLoss$ were used to make our model converge. For Evaluation 2, the MR images were downsampled twice as much as the LR production method, using CUBIC. They are subsequently super-resolved using the model trained in the Evaluation 1.

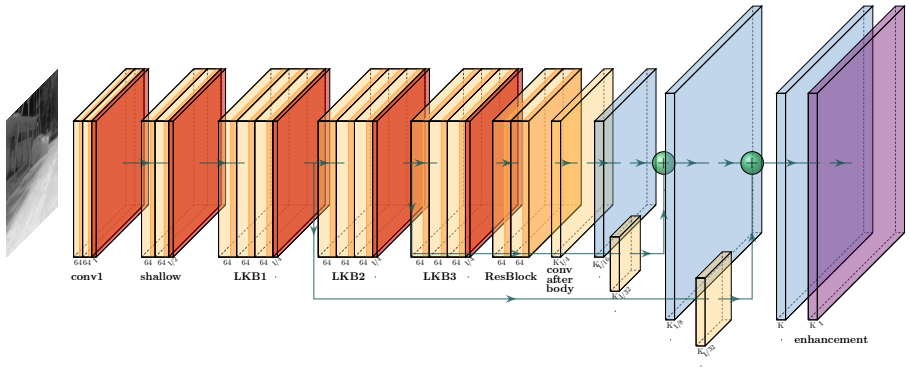


Figure 11: Architecture proposed by XDU-JK team.

Quantitative results shows that this team achieves 34.20 PSNR & 0.9249 SSIM on Evaluation1, and 21.50 PSNR & 0.7754 SSIM on Evaluation 2.

5. Conclusion

This paper summarizes the approach proposed by the teams that reached the final validation phase and submitted their contributions in the Thermal Image Super-Resolution Challenge - PBVS 2022. The different approaches are all deep learning based and use different CNN or transformer architectures, or combinations of them. This is the third edition of the challenge in thermal image SR. The number of participants is higher than in the two previous years, and the obtained results from the quantitative evaluations outperformed the last two years' results. These new results will be used as a baseline for the following challenges.

Acknowledgements

Appendix A. Teams Information

TISR 2022 organization team:

Members: Rafael E. Rivadeneira[1] (rrivaden@espol.edu.ec), Angel D. Sappa[1,2] and Boris X. Vintimilla[1]

Affiliations:

[1]Escuela Superior Politécnica del Litoral, ESPOL, Facultad de Ingeniería en Electricidad y Computación, CIDIS, Campus Gustavo Galindo Km. 30.5 Vía Perimetral, P.O. Box 09-01-5863, Guayaquil, Ecuador.

[2]Computer Vision Center, Campus UAB, 08193 Bellaterra, Barcelona, Spain.

A.1. AIR:

Members: Jin Kim[1,3] (jinmerong@konkuk.ac.kr) and Dogun Kim[2,3]

Affiliation: [1]Konkuk University, Seoul, Republic of Korea; [2]Sungkyunkwan University, Seoul, Republic of Korea; [3]Hanwha Systems, Seoul, Republic of Korea.

A.2. ANT GROUP:

Members: Bo Yan (lengyu.yb@antgroup.com), Leilei Cao, Fengliang Qi and Hongbin Wang

Affiliation: Ant Group, China.

A.3. NJU:

Members: Zhihao Li (lizhihao6@outlook.com) and Yingchun Jian

Affiliation: Nanjing University, Jiangsu, China.

A.4. NPU-LIFT-LAB:

Members: Rongyuan Wu (rongyuanwu@mail.nwpu.edu.cn), Lingchen Sun, Yongqiang Zhao and Lin Li

Affiliations: Northwestern Polytechnical University, Shaanxi, China.

A.5. SENSEXDU:

Members: Kai Wang[1,2] (kiwi0128@qq.com), Yicheng Wang[1], Xuanming Zhang[2], Huiyuan Wei[2], Chonghua Lv[2], Qigong Sun[1] and Xiaolin Tian[2]

Affiliation: [1]SenseTime Research; [2]Xidian University, Shaanxi, China.

A.6. SISYPHUS AKADEMIA:

Members: Zhuang Jia (jiazhuang@xiaomi.com)

Affiliation: Xiaomi Inc.

A.7. WZ:

Members: Chenyang Wang (wangchy02@hit.edu.cn), Zhiwei Zhong, Xianming Liu and Junjun Jiang

Affiliations: School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China.

A.8. XDU-JK:

Members: Jiakui Hu (jiakuihu29@gmail.com)

Affiliation: School of Physics and Optoelectronic Engineering, Xidian University, Xi'an 710126, P. R. China Belgium.

References

[1] Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, July 2017. 4

[2] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017. 7

[3] Georgios D Evangelidis and Emmanouil Z Psarakis. Parametric image alignment using enhanced correlation coefficient maximization. *IEEE transactions on pattern analysis and machine intelligence*, 30(10):1858–1865, 2008. 7

[4] Yuanfei Huang, Jie Li, et al. Interpretable Detail-Fidelity Attention Network for Single Image Super-Resolution. *IEEE Transactions on Image Processing*, 30:2325–2339, 2021. 3

[5] Zheng Hui, Xinbo Gao, Yunchu Yang, and Xiumei Wang. Lightweight image super-resolution with information multi-distillation network. In *Proceedings of the 27th acm international conference on multimedia*, pages 2024–2032, 2019. 6, 7

[6] 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. 3, 4, 7

[7] 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. 6

[8] Ze Liu et al. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 3

[9] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2794–2802, 2017. 4

[10] Kalpesh Prajapati, Vishal Chudasama, Heena Patel, Anjali Sarvaiya, Kishor P Upla, Kiran Raja, Raghavendra Ramachandra, and Christoph Busch. Channel split convolutional neural network (chasnet) for thermal image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4368–4377, 2021. 3

[11] Rafael E Rivadeneira, Angel D Sappa, and Boris X Vintimilla. Thermal image super-resolution: A novel architecture and dataset. In *Proc. of the International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, pages 111–119, 2020. 1, 2

[12] Rafael E Rivadeneira, Angel D Sappa, Boris X Vintimilla, Lin Guo, Jiankun Hou, Armin Mehri, Parichehr Behjati Ardakani, Heena Patel, Vishal Chudasama, Kalpesh Prajapati, Kishor P Upla, Raghavendra Ramachandra, Kiran Raja, Christoph Busch, Feras Almasri, Olivier Debeir, Sabari Nathan, Priya Kansal, Nolan Gutierrez, Bardia Mojra, and William J Beksi. Thermal image super-resolution challenge-PBVS 2020. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 96–97, 2020. 2

[13] Rafael E Rivadeneira, Angel D Sappa, Boris X Vintimilla, Sabari Nathan, Priya Kansal, Armin Mehri, Parichehr Behjati Ardakani, Anurag Dalal, Aparna Akula, Darshika Sharma, et al. Thermal image super-resolution challenge-pbvs 2021. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4359–4367, 2021. 1, 2

[14] 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. 4

[15] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8934–8943, 2018. 4

[16] Radu Timofte, Rasmus Rothe, and Luc Van Gool. Seven ways to improve example-based single image super resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1865–1873, 2016. 5

[17] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European Conference on Computer Vision Workshops*, 2018. 7

[18] Zhilu Zhang, Haolin Wang, Ming Liu, Ruohao Wang, Jiawei Zhang, and Wangmeng Zuo. Learning raw-to-srgb mappings with inaccurately aligned supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4348–4358, 2021. 4

[19] Xiaole Zhao, Yulun Zhang, Tao Zhang, and Xueming Zou. Channel splitting network for single mr image super-resolution. *IEEE Transactions on Image Processing*, 28(11):5649–5662, 2019. 3