Thermal Image Super-Resolution Challenge - PBVS 2021

Rafael E. Rivadeneira, Angel D. Sappa, Boris X. Vintimilla
Sabari Nathan, Priya Kansal, Armin Mehri, Parichehr Behjati Ardakani,
Anurag Dalal, Aparna Akula, Darshika Sharma, Shashwat Pandey, Basant Kumar,
Jiaxin Yao, Rongyuan Wu, Kai Feng, Ning Li, Yongqiang Zhao,
Heena Patel, Vishal Chudasama, Kalpesh Prajapati, Anjali Sarvaiya,
Kishor P. Upla, Kiran Raja, Raghavendra Ramachandra, Christoph Busch,
Feras Almasri, Thomas Vandamme, Olivier Debeir, Nolan B. Gutierrez, Quan H. Nguyen, William J. Beksi

Abstract

This paper presents results from the second Thermal Image Super-Resolution (TISR) challenge organized in the framework of the Perception Beyond the Visible Spectrum (PBVS) 2021 workshop. For this second edition, the same thermal image dataset considered during the first challenge has been used; only mid-resolution (MR) and high-resolution (HR) sets have been considered. The dataset consists of 951 training images and 50 testing images for each resolution. A set of 20 images for each resolution is kept aside for evaluation. The two evaluation methodologies proposed for the first challenge are also considered in this opportunity. The first evaluation task consists of measuring the PSNR and SSIM between the obtained SR image and the corresponding ground truth (i.e., the HR thermal image downsampled by four). The second evaluation also consists of measuring the PSNR and SSIM, but in this case, considers the ×2 SR obtained from the given MR thermal image; this evaluation is performed between the SR image with respect to the semi-registered HR image, which has been acquired with another camera. The results outperformed those from the first challenge, thus showing an improvement in both evaluation metrics.

1. Introduction

Single image super-resolution (SISR) is a classic ill-posed problem in computer vision that is still an active field of research. The goal of SISR is to recover an accurate high-resolution (HR) representation from a given low-resolution (LR) image. Different approaches have been proposed during the last two decades, but in recent years, novel deep learning-based techniques have shown significant improvements. Most learning-based techniques use a downsampled image from the HR image augmented with noise and blur as input. These poor-quality images, together with their corresponding HR images, are used to train the networks. Most of these approaches have been mainly used in the visible spectral domain. With regard to thermal images, their resolution is a common limitation that is related to the sensor technology used. The usage of these images is increasing for miscellaneous applications in several fields (i.e., medicine, security, industry, among others). Hence, the SR approaches initially developed for visible spectrum images have motivated the research community to look for solutions to this problem in the thermal image domain.

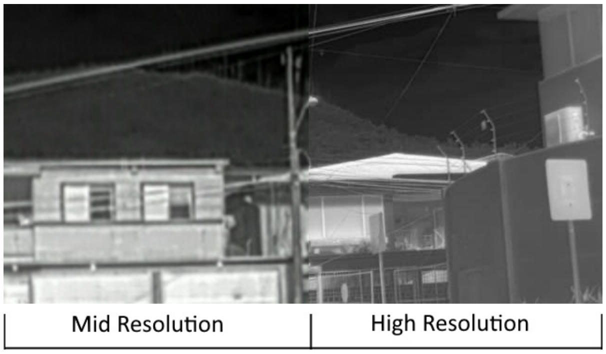


Figure 1: A mosaic illustrating two different resolution thermal images from the same camera viewpoint: $(left)$ a crop from a MR image; $(right)$ a crop from a HR image [13]

In order to define a standard benchmark for evaluating different contributions, the first challenge on TISR was pro-

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.

posed at the PBVS 2020 workshop. Due to the success of that first challenge and to keep improving the obtained results, a second challenge on TISR was proposed in the framework of the PBVS 2021 workshop. For this second challenge, only the MR and HR sets of images from the original thermal image dataset have been used (i.e., the LR set is not considered).

The TISR 2021 challenge[1] has two evaluation approaches. Evaluation 1 measures the ×4 SR result for images from the HR camera. In other words, each participant should downsample and add noise to the given ground-truth (GT) image and use that data to train their network. Evaluation 2 compares the ×2 SR results obtained by using input images from the MR camera (Axis Q2901-E). These ×2 SR results are evaluated with respect to the corresponding semi-registered images obtained from the HR camera (FLIR FC-632O). Hence, the proposed ×2 scale solution should be able to tackle both problems, i.e., generating the SR images acquired with the MR camera as well as mapping images from the MR to HR domain.

The remainder of this manuscript is organized as follows. Section 2 introduces the objectives of the challenge and presents the dataset and evaluation methodology. A summary of the results obtained by the different teams is presented in Section 3. In Section 4, a short description of the different approaches proposed by the teams is provided. Finally, a conclusion is given in Section 5 followed by an appendix with the team information. Due to space limitations, only the two winning architectures are depicted in this work. All of the other proposed architectures are provided in the supplementary material (SM) document and are referred to as *SM-Figure* in this paper.

2. TISR 2021 Challenge

The objectives of the TISR 2021 challenge are the following: (i) introduce state-of-the-art approaches for the thermal image SR problem; (ii) evaluate and compare different solutions using last year's benchmark; and (iii) promote a novel thermal image dataset to be used as a benchmark by the community.

2.1. Thermal Image Dataset

The dataset used in this challenge was presented in [13] and used in the first TISR challenge [14]. It consists of 1021 thermal images acquired with three thermal cameras, at varying resolutions, from indoor and outdoor scenes. The images were acquired under various lighting conditions (e.g., morning, afternoon, evening) and contain diverse objects (e.g., vegetation, people, cars, buildings, etc.). The cameras were mounted in a rig to minimize the baseline distance between the optical axis such that the acquired images are almost registered. Figure 1 presents a mosaic created with images from the MR and HR cameras.

2.2. Evaluation Methodology

The performance of the proposed contributions is evaluated by means of peak signal-to-noise (PSNR) ratio, and structural similarity (SSIM) measures between the obtained SR images with respect to the sequestered GT images that were not shared. As mentioned before, two kinds of evaluations are performed. The first one evaluates SR results from a set of 10 downsampled and noisy images obtained from an HR camera. A downsampling process with a scale factor of ×4 is applied. Additionally, Gaussian noise ($\sigma = 10\%$) is added. Figure 2 shows an illustration of this first evaluation process.

The second evaluation process analyzes a set of 10 SR images obtained by a ×2 scale factor from the given MR images. These 10 SR images are evaluated with respect to the corresponding HR GT images. The GT images are of the same resolution as the computed SR; however, they were acquired with a different camera. The SR images from the MR set are registered with the corresponding HR images by using a feature point-based approach. The evaluation (i.e., PSNR and SSIM) is performed just over 90% of the central cropped region of the image. Figure 3 illustrates this second evaluation process.

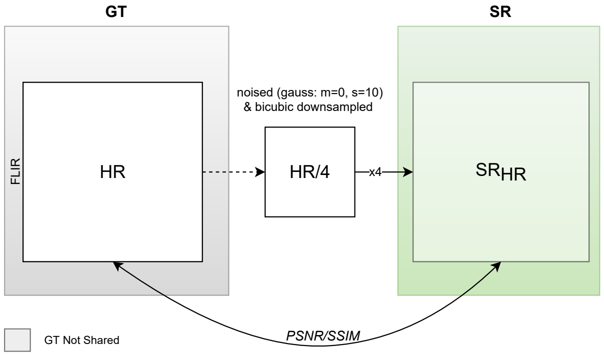


Figure 2: The first evaluation process.

3. Challenge Results

Nine teams submitted their results, together with 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. 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 provided in Appendix A. According to the results presented in Table 1, the winners of the second TISR 2021 challenge are the SVNIT_NTNU-1 and ULB-LISA teams who achieved the best results in

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

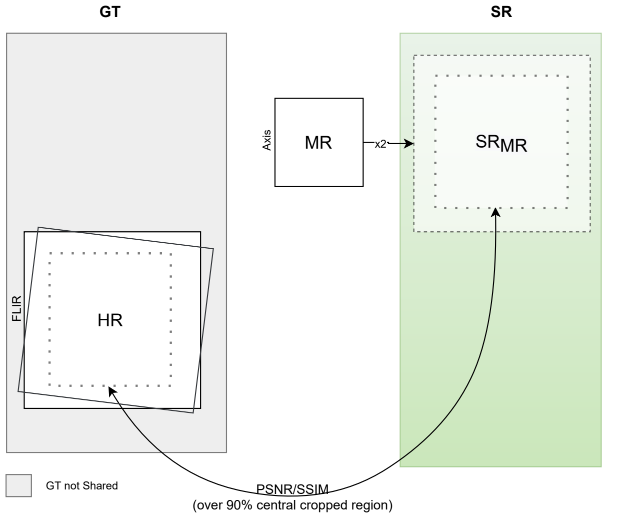


Figure 3: The second evaluation process from MR to HR.

Approach	**Evaluation1**		**Evaluation2**	
	×4		×2 **(MR to HR)**	
	PSNR	**SSIM**	**PSNR**	**SSIM**
COUGER AI	29.13	0.8469	20.03	0.7484
CVC	29.21	0.9032	20.03	0.7533
ISESL-CSIO	30.39	0.8992	16.15	0.5432
MNNIT	14.14	0.5226	20.03	0.7375
NPU-MPI-LAB	29.51	0.8937	21.96	0.7618
SVNIT_NTNU-1	**30.70**	**0.9290**	20.09	0.7510
SVNIT_NTNU-2	30.69	0.9288	21.44	0.7758
SVNIT_NTNU-3	30.59	0.9254	20.10	0.7485
ULB-LISA	24.86	0.8420	**22.32**	**0.7899**
UTA-RVL-1	25.95	0.8812	19.95	0.7389
UTA-RVL-2	29.40	0.8881	18.40	0.6712

Table 1: The 2021 TISR challenge results. Average results from the evaluations are detailed in Section 2.2. Bold and underline values correspond to the best- and second-best results, respectively.

Evaluation 1 and Evaluation 2, respectively.

4. Proposed Approaches and Teams

4.1. COURGER AI

The COURGER AI team's proposed approach is based on a neural network that utilizes the coordinate convolutional layer [11], and residual units [6], along with multi-level supervision and attention unit to map the information.

As shown in SM-Figure 1, the proposed architecture is a simple five-block stacked network. Each block consists of one convolutional layer followed by two Res2net blocks [4]. The output of the convolutional layer is upsampled by a factor of 2 for ×2 resolution and by a factor of 4 for ×4 resolution using a subpixel upscaling layer [16]. Each upscaled output is then fused for generating the final HR image as well as supervised to improve the pixel-wise resolution. However, to increase the dimensional feature information, the input image is first mapped to the Cartesian space using the coordinate convolutional layer [11] as inspired by [8, 7]. To supervise the model output, a combination of three losses is used: mean squared error (MSE), SSIM, and Sobel,

$$Total_{loss} = MSE + SSIM_{loss} + SOBEL_{loss}.$$

4.2. CVC

The CVC team proposes a lightweight overscaling network (L-OverNet) for thermal image SR. It consists of the following two main parts: a lightweight feature extractor and an overscaling module (OSM) for reconstruction. The feature extractor follows a novel recursive framework of skip and dense connections to reduce low-level feature degradation. The OSM is a new inductive bias that generates an accurate SR thermal image by internally constructing an over-scaled intermediate representation of the output features. The feature extractor computes useful representations of the LR patch in order to infer its HR version. Concretely, a recursive structure based on residual blocks (RBs) assembled into local dense groups (LDGs) and LDGs into global dense groups (GDGs) is proposed, see SM-Figure 2.

WDSR with wide low-rank convolutions is used. To make the network focus on more informative features, squeeze-and-excitation (SE) operations after these convolutions are used. The model learns a scalar multiplier λ to balance the amount of information that should be carried by the identity and activation operations within the RBs of the network. RBs are grouped into the LDGs. The input of an RB is concatenated with the output of all previous RBs in the group and merged with a 1×1 convolution. This recursion is repeated for all RBs within the LDG.

To increase the network capacity, a similar recursion is applied to the GDG, but this time incorporating skip connections between LDGs. This procedure is repeated while integrating the recursive concatenations through the LDGs into a single output. The output of each LDG is concatenated to the input of the next one. In order to facilitate access to local information, the final output of the network receives the concatenation of the outputs of all the LDGs. Therefore, the model incorporates features from multiple layers. This strategy makes information propagation efficient due to the multilevel representation and many shortcut connections.

To ensure that no information is lost before the reconstruction step, a long-range skip connection is incorporated to grant access to the original information. This encourages backpropagation of the gradients from the output of the feature extractor to the first 3×3 convolution layer. A global average pooling is also included, followed by a 1×1 convolution to fully capture channel-wise dependencies from the aggregated information.

As an example, consider the maximum scale factor N addressed by the network. First, an intermediate representation of the final image is generated. It consists of overscaled maps H^{OHR}, with an overscale factor $(N+1)$ times larger. Thus, given the **h** features extracted from I^{LR}, a 3×3 convolutional layer is used, followed by the strided subpixel convolution to upscale the features **h** to H^{OHR}. Finally, to obtain the output of the overscaling module, we include a second long-range skip connection from the original I^{LR} image. The final HR image is obtained by adjusting the overscaled maps and incorporating them into the native upscaling of the original LR image.

4.3. ISESL-CSIO

The ISESL-CSIO team uses an attention-based generative adversarial network (GAN) to compute SR images from LR thermal images. The generator model is inspired by SRGAN [10] and consists of five residual blocks with short and long skip connections. These are followed by upsampling blocks, which use pixel shuffle. An attention mechanism is introduced like RCAN [24] and is used in the upsampling blocks to adaptively select the predominant feature set. The discriminator is feature-based and learns to differentiate between SR and HR images.

The loss function is a crucial ingredient in training a network. Current advancements recommend the use of image quality assessment metrics such as PSNR, SSIM, MS-SSIM, etc., in comparison to L_2 for the SR task. Inspired by [22], we found that a combination of L_1 and SSIM loss performs the best. The loss function used for training the generator model is

$$\mathcal{L}^{Mix} = \alpha \frac{1}{n} \sum_{1}^{n} \mathcal{L}^{SSIM} + (1-\alpha) \frac{1}{n} \sum_{1}^{n} \mathcal{L}^{L_1},$$

where n is the batch size and α = 0.84.

As shown in SM-Figure 3, the network consists of a generator and discriminator model. The generator model creates the SR image from the LR or MR image, while the discriminator tries to distinguish if the image is from the generator (SR) or if it is a real-world sample (HR). As training progresses, the generator gets better at producing samples that closely resemble a HR image. Likewise, the discriminator also improves at discriminating HR and SR images. An equilibrium is reached when the discriminator takes a sample and tells with equal probability that the image is from the generator or the real world. When this equilibrium is reached the generator model is taken and the discriminator model is discarded. The generator is then used for the SR task.

4.4. MNNIT

The MNNIT team designed a lightweight model using SRCNN [3]. Lightweight models are computationally efficient, which makes them easier to deploy in real-world problems. The model, as shown in SM-Figure 4, is designed with three convolutional layers for feature extraction and a final fourth layer for the reconstruction of SR images. The first layer generates 128 feature maps using a kernel size of 9×9, followed by the SELU activation function. The second layer gives 64 features with a kernel size of 7×7 and is followed by a leaky ReLU activation function. The third layer gives out 32 features with a kernel size of 5×5, which is followed by the SELU activation function. The final layer regenerates high-quality SR images. Inspired by later models utilizing residual learning in SR, a residual connection is made from the input image to the final convolutional layer for improved training of the model to achieve better results.

4.5. NPU-MPI-LAB

The NPU-MPI-LAB team presents two approaches. For Evaluation 1, a simple network is proposed [9]. Initially, from an LR image, bicubic interpolation (×4) is performed. The network architecture is shown in SM-Figure 5. At the first layer, a convolution with a kernel size of 7 is applied to extract the low-frequency details and ten ResBlocks are used to extract the high-frequency details. The proposed approach is computationally efficient, needing only 1.44M parameters to obtain a four times thermal SR image. The L_2 loss is applied for this task.

For Evaluation 2 a network inspired by ESRGAN [18] is proposed to deal with SR and domain adaptation at the same time. The network architecture is shown in SM-Figure 6. The first convolution layer is set to extract shallow features and is followed by five multi-branch residual dense blocks (MRDB). MRDB aims to extract more levels of deep features through a variety of shortcut structures. Then, the shallow and deep features are sent to the ×2 upsampling through the shortcut structures and are decoded to get the generated image. Note that padding is set to 'same' and stride is set to 1 in all convolution layers to maintain the original size. Considering the serious misalignment of the image, a contextual bilateral loss (Cobi) [23] is selected to match between the generated image and GT. The discriminators are the same as ESRGAN and are called relativistic average discriminators. The total loss of the generator is composed of the adversarial loss, image-level Cobi loss, and perceptual-level Cobi loss (through VGG19),

$$L_G = \beta_1 L_G^{Ra} + \beta_2 L_{\text{Cobi}}^{img} + \beta_3 L_{\text{Cobi}}^{\text{vgg}}.$$

4.6. SVNIT_NTNU

The SVNIT_NTNU team presents three different approaches summarized below.

Approach 1: Figure 4(a) (SM-Figure 7(a)) shows the first proposed thermal image SR architecture for scaling factors of $\times 2$ and $\times 4$. The LR thermal image is used as input to the network, and it is passed through the low- and high-frequency feature extraction modules to extract salient features. The architecture uses an exponential linear unit (ELU) activation function [15] to improve learning performance at each layer in an efficient manner. A new and core element of the proposed architecture is the simple and effective design of ResBlock that preserves the high-frequency details of the SR image with a lesser number of parameters associated with the network and is shown in Figure 4(b) (SM-Figure 7(b)). The different kernel sizes (i.e., 1×1, 3×3, and 5×5) are adopted in the convolutional layer of ResBlock to recover details distributed at local and global regions. In addition, the proposed ResBlock involves depthwise convolution (DC) in the intermediate stage to make the network computationally cheaper [21]. The channel attention modules are further used to perform adaptive rescaling of features on a per-channel basis. A nearest-neighbor interpolation is used to upscale the feature maps to the desired scaling factor (i.e., $\times 2$ and $\times 4$). Finally, feature sharing with the reconstruction module is introduced that helps the model to focus on the features that are important.

Approach 2: The architecture of the proposed convolutional neural network (CNN) method for the upscaling factors $\times 4$ and $\times 2$ (Evaluation 1 and 2, respectively) is shown in SM-Figure 8(a). Additionally, SM-Figure 8(b) displays the UNet-based framework for upscaling $\times 2$ in Evaluation 2 LR images. For Evaluation 1, a CNN is trained using the L_1 loss between SR and HR images. In Evaluation 2, the network shown in SM-Figure 8(a) is trained using the L_1 loss on the dataset created by bicubic downsampling from the available Axis (MR) images [14] while the network in SM-Figure 8(b) is trained on the dataset created by the registration between FLIR (HR) and Axis (MR) datasets [14] using SURF features. However, this registration is not perfectly achieved leading them to employ a GAN framework [5] for semi-supervised learning. The UNet-based network uses the L_1, GAN, and SSIM losses for training. The final resulting SR image is generated using a combination of the SR images obtained from those two networks for the Evaluation 2 case. Moreover, a self-assemble technique [17] is also utilized to generate the SR results to enhance its quality in both evaluations.

A patch of 192×192 from the HR image and the corresponding size of its LR image is extracted randomly in the training process. The extracted patches are passed through a random horizontal flipping and (0^o or 90^o) rotation to perform augmentation. The Adam optimizer with default β values is used to train the network up to 2×10^5 iterations. The learning rate for the optimizer is set to 2×10^{-4} and is decayed by half at every 25% of the total iterations.

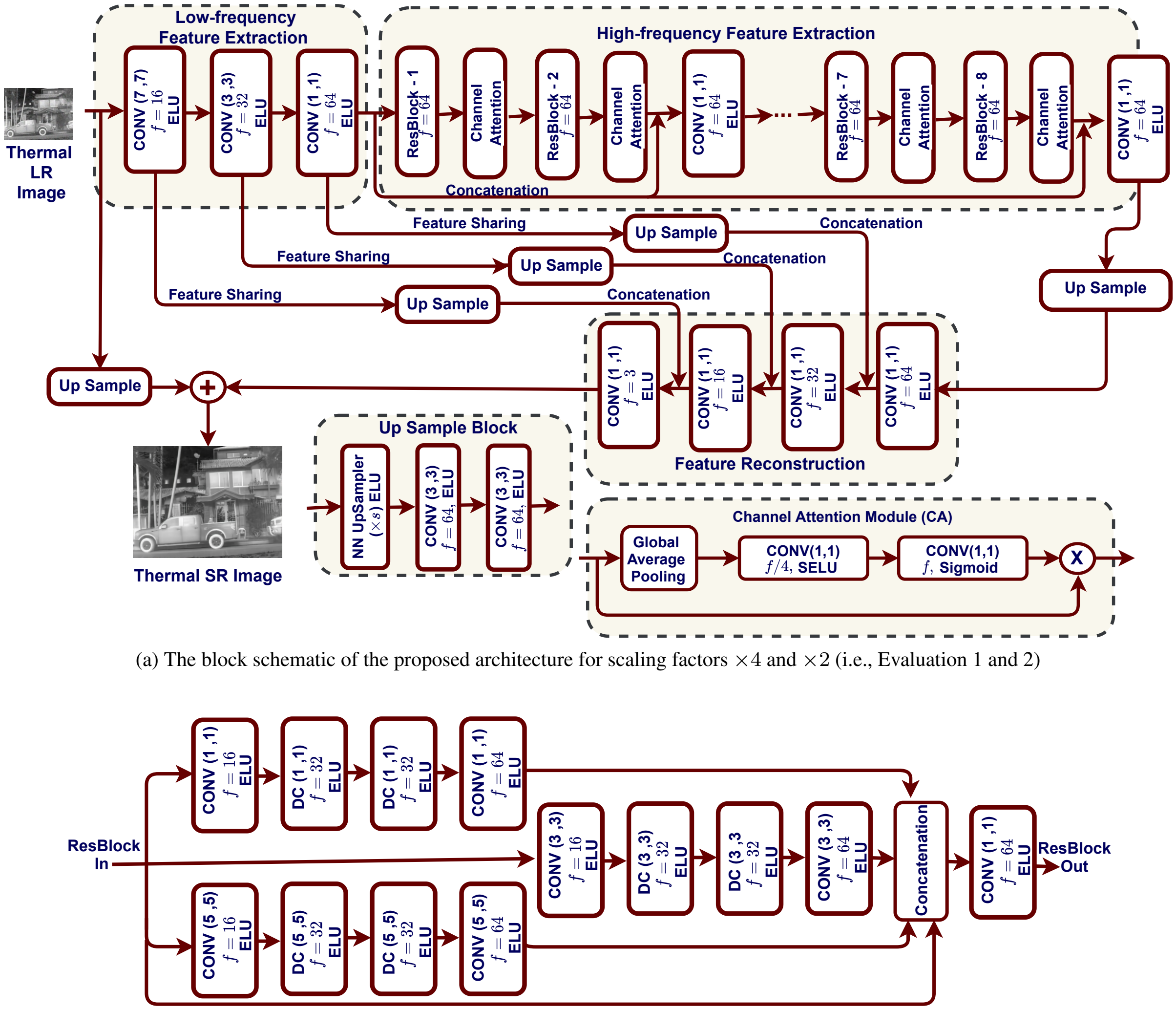
Approach 3: The design of the proposed approach to super resolve the thermal LR image is illustrated in SM-Figure 9(a), where it takes a thermal LR image as input and generates the corresponding SR image based on the chosen upscaling factors (i.e., $\times 2$, $\times 4$). The proposed approach consists of several residual groups and are considered as the core elements in the network to learn complex and rich features from the LR observation. The design of the proposed ResBlock is depicted in SM-Figure 9(b). Subpixel convolutions [16] in the upscaling block are used to upscale the feature maps to the desired resolution level. To mitigate the vanishing/exploding gradient problem, local and long skip connections are used where the higher-layer gradients are bypassed to lower layers, hence the learning ability of the proposed approach improves. Also, a global residual learning strategy is used.

4.7. ULB-LISA

The ULB-LISA team introduces a model referred to as the xcycles backprojection network (XCBP). It is comprised of the following two modules: cycle features correction (CFC) and residual features extraction (RFE). As shown in Figure 5(a) (SM-Figure 10(a)), the architecture of XCBP is unfolded with X number of CFCs, as each cycle contains one RFE module. The model uses an encoder (E) of one convolutional layer to extract the shallow features F from the LR input image I_{LR} and from its pre-upsampled version $\uparrow I_{LR}$. The pre-upsampling module is a bicubic operator. It uses a decoder (D) of one convolutional layer, which takes the final features $F_{SR,X}$ corrected by the CFC in cycle X to reconstruct the SR image I_{SR}.

The CFC serves as a feature correction mechanism, it is designed to supply the encoded features of the two parallel features' spaces to its RFE module for further feature extraction and corrects the encoded features one at a time. The output of this RFE module is backprojected by addition to one of the two parallel features spaces, while in the $F_{SR,x}$, the output of the RFE passes by the upsampler (U) to match the scale of the features. U is a resize-convolution consisting of a predefined nearest-neighbor interpolation operator of scale factor $\times 2$ and a convolution layer with a receptive field of size 5×5 pixels represented by two stacked 3×3 convolutions. In contrast to VCBP [14], the backprojection corrects the previous features in both encoded features manifold.

The RFE module takes both features and extracts new residual features for the next feature space correction procedure based on the similarity and dissimilarity between the previously corrected features' spaces. Illustrated in Figure 5(b) (SM-Figure 10(b)), the RFE has two identical sub-



(a) The block schematic of the proposed architecture for scaling factors ×4 and ×2 (i.e., Evaluation 1 and 2)

(b) The design of the ResBlock used in the proposed model.

Figure 4: SVNIT_NTNU's first proposed architecture (**winner at Evaluation 1**).

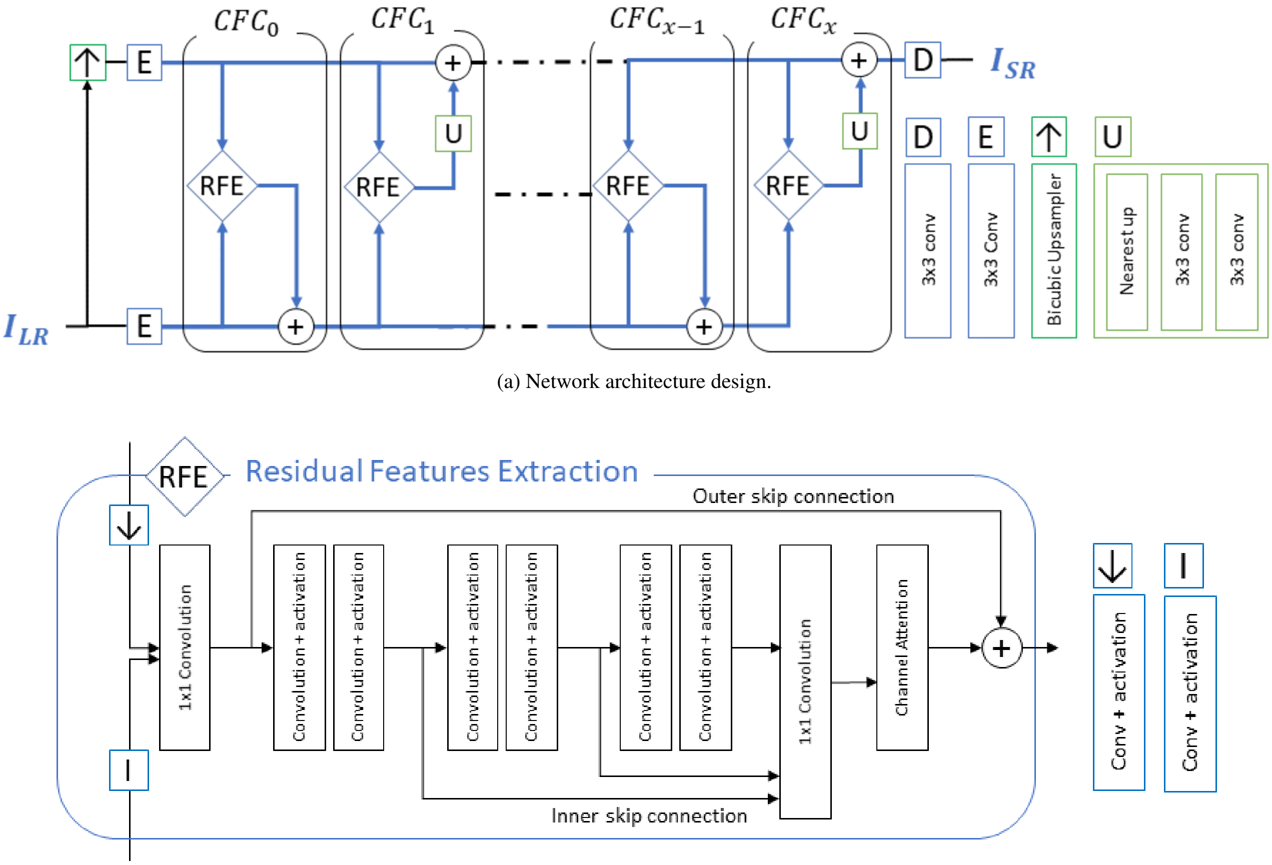
modules. An internal features encoder (I) is responsible for extracting deep features from the two parallel spaces of one convolutional layer, with strided convolution in the HR space to adapt the different resolution scales. The core of the RFE module consists of three levels (L) of double activated convolution layers connected sequentially. The output of L, defined as inner skip connections, are concatenated to create dense residual features. They are then fed to a channel attention module inspired by RCAN [24] to weight the residual channels priority before they are added to the outer skip connection.

For Evaluation 2, FLIR images are registered with the Axis images using SIFT and ORB. Only 60% of registered images with high NCC and SSIM values are selected for training. One model with ×2 scale factor was trained on the following three sets: FLIR ×2 downsampled, Axis ×2 downsampled, and FLIR registered with Axis images. Another model with ×2 scale factor is attached to the input of the first model trained for the noisy FLIR ×4 scale factor.

4.8. UTA-RVL-1

The UTA-RVL-1 team proposes a novel network for thermal image SR called the deep attention to varying smooth receptive fields network (AVSRN). The proposed architecture is displayed in SM-Figure 11. The network expands upon popularized attention mechanisms [24, 2], which have attempted to model local features of LR images throughout SR networks.

The inputs to AVSRN are normalized LR images, which are fed to a residual in residual (RIR) architecture [24]. The output of RIR, from the last residual group, is passed to an upscaling module with different parameters for each



(a) Network architecture design.

(b) Residual feature extraction module.

Figure 5: The architecture proposed by the ULB-LISA team (**winner at Evaluation 2**).

upscaling factor. The RIR is very similar to the original; thus, these contributions lie at the residual block (RB) level shown in the right half of SM-Figure 11. The left half of SM-Figure 11 shows its place within RIR. Inside the RIR is a residual group containing two RBs.

At the RB level, input from the previous RB is fed to a parallel structure of three smooth dilated convolutions (SDC). Dilated convolutions allow CNNs to have access to more contextual information at a lower cost. Each of the SDCs in the parallel structure contains a separable convolution and then an atrous convolution, in sequence, to lessen gridding issues associated with dilated convolutions [19]. The features from the SDCs are aggregated by concatenation and Conv2D before being passed to the second-order channel attention (SOCA) module. SOCA's output is multiplied by a residual connection from the aggregating Conv2D preceding addition with a residual connection from the previous RB.

4.9. UTA-RVL-2

The UTA-RVL-2 team, inspired by recent successes of implicit representations on 3D data [1, 12, 20], designed a compact neural network for the 2D counterpart. The network, shown in SM-Figure 12, learns local features from the inputs through a series of 2D convolutions resulting in multiple feature maps. Given a normalized continuous query location (x, y) in the image space, the network extracts the corresponding features of the location on each learned feature map. These collected features are then used to predict the intensity of a pixel corresponding to the location in the output image.

More specifically, the proposed network first learns local information from input image I through a series of three convolution blocks of increasing depth. Each block consists of two consecutive 2D convolution layers with ReLU activation followed by a batch normalization layer. Each convolution block produces a feature grid that is aligned with the image space. Next, the features that correspond to the input coordinates (x, y) are extracted from the feature grids. Since the location is continuous and the feature grids are discrete, the query features are calculated using bilinear interpolation on each feature grid. The extracted features from all feature grids are then concatenated and fed into a three-layer fully-connected block, which predicts the intensity value at location (x, y) in the output image.

In each iteration of training, the network takes a LR image and a set of uniformly generated query locations as inputs. The GT intensity of each query location is collected from the SR image of the same scene. Over time, the network is optimized by continuous query locations. At test time, given a LR image an output image of desired resolution is obtained by applying the network to an equally spaced 2D grid of locations where the number of points in the grid equals the number of pixels in the output image. The proposed network is size agnostic and can be applied to images of varying shapes.

5. Conclusion

This paper summarizes the contributions from the nine teams that reached the final validation phase in the Thermal Image Super-Resolution Challenge - PBVS 2021. This was the second edition of the challenge in thermal image SR. To conclude, we highlight that the number of teams reaching the final stage is higher than in the first year and the results obtained from the two quantitative evaluations outperformed last year's results. The results from this year will be used as a baseline for the next challenge. Lastly, this challenge has also been an opportunity to promote the testbed dataset used by the different teams.

Acknowledgements

This work has been partially supported by the ESPOL Polytechnic University; the Spanish Government under Project TIN2017-89723-P; and the "CERCA Programme / Generalitat de Catalunya". The first author has been supported by Ecuador government under a SENESCYT scholarship contract.

Appendix A. Teams Information

TISR 2021 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. COUGER AI:

Members: Sabari Nathan (sabari@couger.co.jp) and Priya Kansal.

Affiliation: Couger Inc, Japan.

A.2. CVC:

Members: Armin Mehri (amehri@cvc.uab.es) and Parichehr Behjati Ardakani.

Affiliation: Computer Vision Center, Campus UAB, 08193 Bellaterra, Barcelona, Spain.

A.3. ISESL-CSIO:

Members: Anurag Dalal[1,2] (anurag.dalal59@gmail.com) and Aparna Akula[1].

Affiliations: [1]Central Scientific Instruments Organization, Chandigarh, India; [2]Indian Institute of Engineering Science and Technology, Shibpur, India.

A.4. MNNIT:

Members: Darshika Sharma (darshika@mnnit.ac.in), Shashwat Pandey and Basant Kumar.

Affiliation: MNNIT Allahabad, Prayagraj, India.

A.5. NPU-MPI-LAB:

Members: Jiaxin Yao, Rongyuan Wu (rongyuanwu@mail.nwpu.edu.cn), Kai Feng, Ning Li and Yongqiang Zhao.

Affiliation: Northwestern Polytechnical University.

A.6. SVNIT_NTNU:

Members: Heena Patel[1] (hpatel1323@gmail.com), Vishal Chudasama[1], Kalpesh Prajapati[1], Anjali Sarvaiya[1], Kishor P. Upla[1], Kiran Raja[2], Raghavendra Ramachandra[2] and Christoph Busch[2].

Affiliations: [1]SVNIT, Surat, India; [2]NTNU, Gjøvik, Norway.

A.7. ULB-LISA:

Members: Feras Almasri (falmasri@ulb.ac.be), Thomas Vandamme and Olivier Debeir.

Affiliation: Université Libre de Bruxelles, Belgium.

A.8. UTA-RVL-1:

Members: Nolan B. Gutierrez (nolan.gutierrez@mavs.uta.edu) and William J. Beksi.

Affiliation: University of Texas at Arlington, United States.

A.9. UTA-RVL-2:

Members: Quan H. Nguyen (quan.nguyen4@mavs.uta.edu) and William J. Beksi.

Affiliation: University of Texas at Arlington, United States.

References

[1] Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5939–5948, 2019. 7

[2] Tao Dai, Jianrui Cai, Yongbing Zhang, Shu-Tao Xia, and Lei Zhang. Second-order attention network for single image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11065–11074, 2019. 6

[3] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *Transactions on Pattern Analysis and Machine Intelligence*, 38(2):295–307, 2015. 4

[4] Shanghua Gao, Ming-Ming Cheng, Kai Zhao, Xin-Yu Zhang, Ming-Hsuan Yang, and Philip HS Torr. Res2net: A new multi-scale backbone architecture. *Transactions on Pattern Analysis and Machine Intelligence*, 2019. 3

[5] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *arXiv preprint arXiv:1406.2661*, 2014. 5

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 3

[7] Priya Kansal and Sabarinathan Devanathan. Eyenet: Attention based convolutional encoder-decoder network for eye region segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 3688–3693, 2019. 3

[8] Priya Kansal and Sabari Nathan. A multi-level supervision model: A novel approach for thermal image super resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 94–95, 2020. 3

[9] 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. 4

[10] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4681–4690, 2017. 4

[11] Rosanne Liu, Joel Lehman, Piero Molino, Felipe Petroski Such, Eric Frank, Alex Sergeev, and Jason Yosinski. An intriguing failing of convolutional neural networks and the coordconv solution. *arXiv preprint arXiv:1807.03247*, 2018. 3

[12] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4460–4470, 2019. 7

[13] 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

[14] 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, 5

[15] Anish Shah, Eashan Kadam, Hena Shah, Sameer Shinde, and Sandip Shingade. Deep residual networks with exponential linear unit. In *Proceedings of the International Symposium on Computer Vision and the Internet*, pages 59–65, 2016. 5

[16] 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. 3, 5

[17] 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

[18] 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. 4

[19] Zhengyang Wang and Shuiwang Ji. Smoothed dilated convolutions for improved dense prediction. In *Proceedings of the International Conference on Knowledge Discovery & Data Mining*, pages 2486–2495, 2018. 7

[20] Qiangeng Xu, Weiyue Wang, Duygu Ceylan, Radomir Mech, and Ulrich Neumann. Disn: Deep implicit surface network for high-quality single-view 3d reconstruction. *arXiv preprint arXiv:1905.10711*, 2019. 7

[21] Rongtian Ye, Fangyu Liu, and Liqiang Zhang. 3d depthwise convolution: Reducing model parameters in 3d vision tasks. In *Proceedings of the Canadian Conference on Artificial Intelligence*, pages 186–199. Springer, 2019. 5

[22] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018. 4

[23] Xuaner Zhang, Qifeng Chen, Ren Ng, and Vladlen Koltun. Zoom to learn, learn to zoom. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3762–3770, 2019. 4

[24] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *Proc. of the European Conference on Computer Vision*, pages 286–301, 2018. 4, 6