Thermal Image Super-Resolution using Deep Learning Techniques

Rafael Eduardo Rivadeneira Campodónico

Facultad de Ingeniería en Electricidad y Computación - FIEC
Escuela Superior Politécnica del Litoral - ESPOL

This dissertation is submitted for the degree of
Doctor of Applied Computer Science

Guayaquil, April 2023

Director	**PhD. Angel D. Sappa** CIDIS-ESPOL & CVC
Co-Director	**PhD. Boris X. Vintimilla** CIDIS-ESPOL
Thesis Committee	**PhD. Daniel Ochoa D.** ESPOL
	PhD. Miguel Realpe R. ESPOL
	PhD. Dennis Romero L. Hongo Aerospace Inc.
	PhD. Marco Flores C. ESPE

This document was typeset by the author using LaTeX2e.

The research described in this book was carried out at the Centro de Investigación, Desarrollo e Innovación de Sistemas Computacionales - CIDIS, Escuela Superior Politécnica del Litoral - ESPOL, Guayaquil, Ecuador.

"Remember to look up at the stars and not down at your feet. Try to make sense of what you see and wonder about what makes the universe exist. Be curious. And however difficult life may seem, there is always something you can do and succeed at."

Stephen Hawking

Acknowledgements

I want to thank my tutors, who have been a great help and support during my PhD. Thank you, Angel and Boris, for your patience and guidance. Your expertise and knowledge have been invaluable in helping me develop as a researcher; even though this process has not been easy, you have always been there for me. Without you, this would not have been possible. I would also like to extend thanks to my thesis committee: PhD. Daniel Ochoa, PhD. Miguel Realpe, PhD. Dennis Romero and Marco Flores, for their guidance and support.

During my years as a PhD. candidate, I met many people who have helped me differently. I want to take this opportunity to thank all of them, especially the PhD. Patricia Suárez, for all her advice and support. Thank you, my closed friends, colleague, and fellows, for always being there for me. Thank you, professors, for your guidance and for helping me develop my skills; an extended thanks to CIDIS staff for all support.

I am very grateful to the SENESCYT, an Ecuadorian government institution that supported me with the scholarship for the PhD program.

And finally, I would like to express my heartfelt gratitude to my beloved family, who have been my reason and impetus throughout this journey. They have been my unwavering support system, standing by me through thick and thin, and I owe my success to their love and encouragement. A special thank you to my wonderful parents, Rocío and Eduardo, my brothers Alex, Fernando, Eduardo, and Joaquín, and my beloved wife Karen and our precious sons, Paulina and little Eduardo.

Guayaquil, April 2023

Abstract

In recent years, there has been an increasing demand for high-resolution images, especially in the field of security and surveillance. Super-resolution is a technique that can be used to improve the resolution of an image. Most of these techniques are based on using a single image or a set of low-resolution images from the visible spectrum, where the high-resolution image is reconstructed by using a model that considers a degradation process. Nevertheless, images from the visible spectrum are limited by the atmospheric conditions and the availability of light.

While human visual perception is limited to the visual-optical spectrum, machine vision is not. This dissertation presents the use of images from the long-wavelength infrared spectral band, namely thermal images, for the purpose of super-resolving them. Thermal images are not affected by atmospheric conditions, and they can be acquired even in low-light conditions. In order to obtain a high-resolution image from a set of low-resolution thermal images, deep learning techniques are used, specifically convolutional neural networks. The results show that improving the thermal images' resolution is possible while preserving the scene's main features. Two main paths are tackled in the present work, the single and multi-image super-resolution, where a dataset with an extensive collection of images is collected to address this purpose. One of the main properties of this thesis is to show that thermal image super-resolution can be tackled by using the proposed architectures and validating them with the acquired public dataset used in several challenges.

Keywords: super-resolution, deep learning, convolutional neural networks, thermal images, single-image/multi-image super-resolution, supervised/unsupervised approaches

Table of contents

List of figures

List of tables

Chapter 1

Introduction

1.1 Motivation

The world has seen great software and hardware technology advancements in the past decade. This has allowed industrial sectors to make use of modern technology to create electronic devices such as computer systems, smartphones, tablets, and other devices at a relatively low cost. In addition, the manufacturing methods of camera sensors have been highly developed, resulting in high-quality digital cameras.

The electromagnetic spectrum is divided into several bands, such as X-rays, ultraviolet, visible, infrared, and radar, among others. The visual-optical (VIS) spectrum is the range of wavelengths of reflected light that humans can see. VIS-sensitive cameras can capture this light in either RGB color images or gray-value images. These cameras are used in computer vision to automatically process images and videos for various applications. The limitations of human perception in poor weather or low illumination conditions can be handled through technological advancements in thermographic imaging. All objects emit infrared radiation by themselves, independently of any external energy source, and depending on the temperature they emit a different wavelength in the long-wavelength infrared spectrum (i.e., thermal).

There exist different sub-division schemes for the infrared region in the different scientific fields, but the common scheme is shown in Fig. 1.1, where it has five regions, the near (NIR: Near-infrared), short (SWIR: Short-wavelength infrared), middle (MWIR: Mid-wavelength infrared), long (LWIR: Long-wavelength infrared) and far (FIR: Far-infrared) spectral bands, where the Long-wavelength infrared is also known as Thermal InfraRed (TIR) [27]. Fig. 1.2 presents images captured from different spectral band of distinct scenarios. The research work on this thesis is focused on the usage of TIR images, which is motivated by the facts mentioned below.

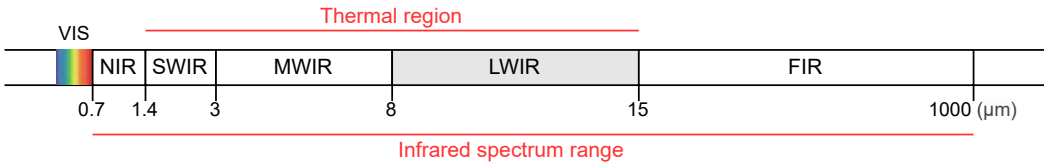
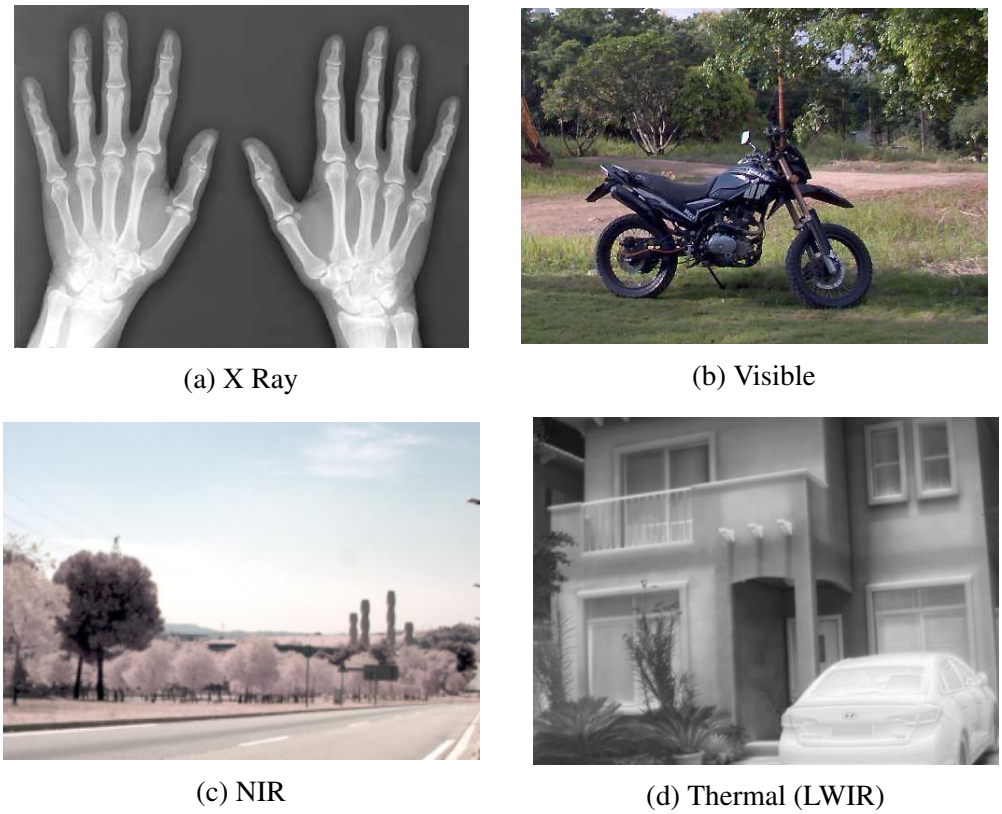


Fig. 1.1 Overview of the visible and infrared range in the electromagnetic spectrum.

Camera sensors can capture information of the electromagnetic spectrum. Visible cameras capture visible light and represent it as greyscale or RGB images. Infrared radiation is not visible to the human eye but can be detected by its effects on heat using an infrared-sensitive camera. This type of radiation is first discovered in the 1800s by astronomer Sir William Herschel. In 1929, British scientists leveraged this discovery to develop the first infrared-sensitive camera. This camera is designed for an anti-aircraft defense system. Passive sensors, like thermal cameras, can capture the information of the TIR spectral band; they capture the thermal-infrared radiation emitted by all objects with a temperature above absolute zero, based on object heat emission. No external illumination is required, such as natural or artificial lights. Thermal information can provide valuable extra information to the visible one (i.e., RGB camera), because it can detect objects that are not visible to the human eye. Furthermore, thermal images are not affected by the presence of artificial lights, such as streetlights, headlights, or flashlights. Meanwhile visible images can not capture nothing in total darkness, thermal cameras are not affected by this lack of illumination and do not depend on any external energy source. Fig. 1.3 shows a clear example of the pro and cons of thermal images; this figure shows two images from the same scene captured with a visible spectrum and a thermal camera, where the thermal image is represented as a grayscale image, with dark pixels for cold spots and the whites one for hot spots. In the figure mentioned above, a sitting person inside the garage is more distinguished in the thermal image, while in the visible spectrum, it is almost impossible to distinguish him, but the thermal image has a resolution lower than the visible image.

In recent years, the infrared imaging field has grown considerably; nowadays, there is a large set of infrared cameras available in the market (i.e., some of the most well-known brands are: FLIR[1], Axis[2], among others) with different technical specifications, lens, and costs. Innovative use of infrared imaging technology can, therefore, play an essential role in a wide range of various applications, such as medicine, military defense, surveillance and security, agriculture, building inspection, fire detection, among others, as well as detection, tracking and human recognition. However, thermal cameras have poor spatial resolutions compared

[1]https://www.flir.com
[2]https://www.axis.com



(a) X Ray (b) Visible

(c) NIR (d) Thermal (LWIR)

Fig. 1.2 Images captured from different spectral bands.

with RGB cameras, and depending on the thermal camera's specifications, the cost can vary between $200 till more than $20000 USD; the latter one is based on active technology with a cooled detector integrated using a cryocooler, providing better resolution and higher frame rate. On the contrary, cheap thermal cameras have smaller resolution than commercial RGB cameras, which can be 160×120 and 1280×1024, respectively. This low-resolution, at a moderate price, is a big limitation when thermal cameras need to be used for general-purpose solutions. In order to improve the overall quality of low-resolution images, Super-Resolution (SR) techniques are commonly used. Lasting research among Computer Vision community has been focused on SR techniques specifically for thermal images where there has been a major shift in recent years.

Image SR refers to the estimation of high-resolution (HR) image/video from a Low-Resolution (LR) of the same scene, in general with the use of digital image processing and Machine Learning (ML) techniques. SR has also important applications in a wide range of domains, such as surveillance and security (e.g., [136], [87], [102]), medical imaging (e.g., [78], [97],[46]), object detection in scene (e.g., [29]), among others. Actually, the possibility of obtaining a SR image has been largely exploited in the visible spectrum domain, where

Fig. 1.3 Visible and thermal images of the same scene at night. *(left)* captured with a RGB camera, and *(right)* captured with a thermal camera. Visible image has higher resolution than thermal.

different super-resolution approaches have been proposed from a conventional interpolation (e.g., [22], [52], [122]), but recently the development of deep learning techniques, have witnessed remarkable progress achieving the performance on various benchmarks of SR.

Learning-based super-resolution methods generally work by down-sampling and adding both noise and blur to the given images. These noisy and blurred poor quality images, together with the originally given images, which are considered as the ground truths, are used in the training process. The training process mentioned above has been chiefly used to tackle the super-resolution problem. However, there are few recent contributions where the training process is based on the usage of a pair of images (low and high-resolution images) obtained from different cameras (e.g., [49]); as mentioned above, the poor resolution of thermal cameras, due to physical limitations, could be improved by using new algorithms with learning-based super-resolution methods, allowing to increase image resolution.

Although the use of thermal imaging is not something new, and conventional super-resolution techniques (i.e., bicubic interpolation) have been used for many years in the visible spectrum, the use of deep learning techniques for super-resolution is something that has emerged in recent years, most of them are focused only on the visible spectrum. This thesis seeks to use deep learning techniques to obtain super-resolution representations in the thermal spectrum, allowing the use of high-resolution images in any of the applications mentioned above. In addition to the poor spatial resolution of thermal cameras, there is another limitation of this technology; this is related to the low frame rate. The 30 fps are generally used in the visible spectrum in a general blow. This low temporal resolution is not tackled in the research work of this thesis. It is worth to mention that recover the temperature measure of the image is also not tackled. In other words, this dissertation is

focused on the single and multi-image spatial super-resolution problems by proposing novel deep learning-based approaches.

1.2 Research objectives

This dissertation focuses on images from the long-wavelength infrared spectral band. Different methods for image super-resolution using deep learning techniques are proposed and compared to find the most suitable one for this particular subject. The following research objectives have been formulated to achieve the goals of this thesis:

1. Evaluate architectures of different convolutional networks used for image super-resolution in the visible spectrum.
2. Design and implement novel deep learning-based architectures to tackle the thermal image super-resolution problem.
3. Generate datasets for validating the implemented techniques.
4. Validate implemented models and compare them with the state-of-the-art approaches.

1.3 Contributions and outline

The outline and contributions of this dissertation are:

1. **Dataset (Chapter 3).** For training neural networks in super-resolution, it is crucial to have a large dataset with high-resolution images in diverse scenarios. Most of the images available for super-resolution are from the visible spectrum. This chapter, to go further and use Thermal Cameras, presents in detail the datasets collected in the long-wavelength infrared spectral band in different resolutions. This setting will provide deep learning-based models with sufficient data to train and evaluate their results.
2. **Single-Image Super-Resolution (Chapter 4).** Once the thermal cameras are used, restoring a high-resolution image from a low-resolution image is tackled from different approaches. This chapter describes the uses of supervised and unsupervised deep-learning techniques. The proposed models make thermal image super-resolution achieve outstanding results.

3. **Multi-Image Super-Resolution (Chapter 5).** In this chapter, multiple low-resolution images are used to restore a high-resolution image. The main idea is to take several images from the same scenario (with a little bit of shift, i.e., rotation and translation), grab the main features of each burst image, and recreate a high-resolution image. In contrast to single image from Chapter 4, multi-image can capture more critical information for image restoration. The model is trained with a simulated dataset.

The contributions mentioned above have been presented at conferences and a scientific journal. More details about these publications can be found in Chapter 6.

The thesis is organized as follows. Chapter 2 summarizes relevant works related to the proposed approaches into six sub-fields: *i*) Thermal image application areas, *ii*) Deep learning, *iii*) Datasets, *iv*) Evaluation metrics, *v*) Single image super-resolution and *vi*) multi-image super-resolution. Chapter 3 describes the acquired thermal dataset using different thermal cameras. Chapter 4 presents the single image super-resolution (SISR) with the proposed deep learning-based models using supervised and unsupervised methodologies. It also tackles multiple challenges in thermal image super-resolution. In Chapter 5, on the contrary to Chapter 4, a novel deep learning method based on the usage of multiple input images is proposed. Finally, in Chapter 6, the conclusions of the thesis are given and the future works presented.

Chapter 2

Literature Review

This chapter starts by reviewing some applications of thermal images. Then, it reviews the traditional methods used for upsampling. After that, a general overview of deep learning techniques is given. Then, common datasets and metrics used for training and testing are reviewed. Single and multi-image super-resolution techniques are then detailed. Finally, thermal image super-resolution challenge is tackled. Most reviewed approaches are intended for visible spectrum images while this dissertation's most relevant methods are for thermal image super-resolution.

2.1 Application areas

Human visual perception is limited to the visible spectrum. Being able to 'see' in other parts of the electromagnetic spectrum (infrared) can provide a lot of information that is currently hidden. Having access to the temperature of a given scene represents an attractive feature for many applications (e.g., detection, tracking, medicine, agriculture, among others) since it provides additional information that is not available in the visible spectrum. Applications that can benefit from using infrared imagery range from biometrics to surveillance and from robotics to automotive.

In the studies of warm-blood wild animals, [11] uses thermal imaging for the disease's diagnosis; where depending on blood circulation, temperature distributed over the animal body changes and can be detected. [130] uses thermal cameras to reveal inflammations in some part of the animal's body (e.g., the leg), or also to know the animal's health (e.g., [42]). In the food industry, [31] makes uses of thermal images because it is a non-invasive and non-contact method to measure the temperature of the fruit and vegetable, and get additional knowledge on information about the quality, such as damage and bruises. Thermal images in inanimate objects depend on the energy that generates heat on them. In [32], the authors

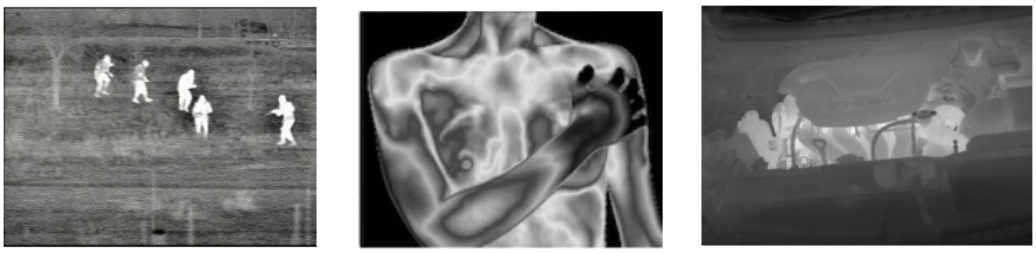


Fig. 2.1 Overview of some thermal image applications (security, medical, industrial).

propose to use of thermal images for building inspection, for the detection of heat loss. Meanwhile, [48] proposes it for industrial applications such as automatic issues detection in electrical installation or diagnosis for an object. Another application, proposed by [86] and [86], used for fire detection and military reasons (locating gunfire or hidden people), respectively. According to [139], thermal images can be also used for fault diagnosis methods on rotor-bearing systems.

Video surveillance is an application with great diversity. Thermal images are also widely used in detecting and tracking humans, which is the first step in many surveillance applications because it allows us to 'see' at night. In [124], authors propose an intruder detection system adjusting a thermal camera to detect objects in the human's temperature range and then classifying the objects based on their shape. [106] proposes a system of fall detection using just a low-quality thermal camera. Fig. 2.1 shows three examples of thermal applications (i.e., surveillance, medicine and driving assistance). It is important to mention that with higher spatial resolution of a thermal image, better application's performance.

However, many applications of computer vision still require high-resolution images, which often exceed the capabilities of HR digital cameras. Optical resolution is a measure of the ability of a camera system or its component to produce detailed images.

2.2 Upsampling methods

Before mentioning deep learning-based methods, it is worth to understand how traditional methods work (increasing the resolution of an image by adding more pixels). Over the last few decades, various interpolation methods have been proposed (e.g., [4, 52]). These methods have a relatively low complexity, but do not produce good results because they do not take into account the underlying structure of the image. Typically, it is done by upsampling the low-resolution image and then using a filter to smooth the appearance.

Interpolation-based methods refer to resizing an image by adding more pixels between existing pixels. This is done by using an interpolation function that estimates the value of

new pixels based on the importance of existing pixels. The most widely used interpolation methods are (see Fig. 2.2):

- Nearest neighbor interpolation is the simplest and fastest interpolation method, it requires the least processing time. This method just copies available values, not interpolate; has the effect of simply making each pixel bigger. Still, it does not produce good results and it should not be used in high resolution zooming.
- Bilinear interpolation is also simple interpolation method, and it works by performing two linear interpolations and also by using the weighted average of the 4 nearest pixels to estimate the value of a new pixel. This method makes images look much smoother than nearest neighbor interpolation. This technique reduces the visual distortion.
- Bicubic interpolation is a more sophisticated interpolation method, as compare to bilinear interpolation, which takes only 4 pixels (2×2) into account, this interpolation uses the weighted average of 16 nearest pixels to estimate the value of a new pixel. This method generates sharper images than the previous two methods, and is the ideal combination of processing time and output quality.

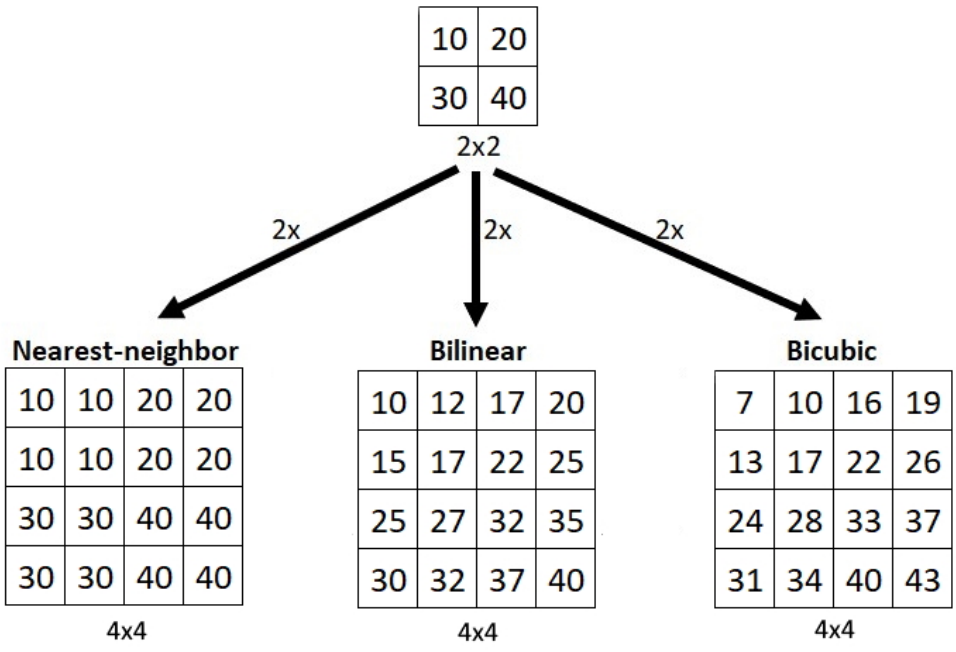


Fig. 2.2 Illustration of traditional interpolation-based methods.

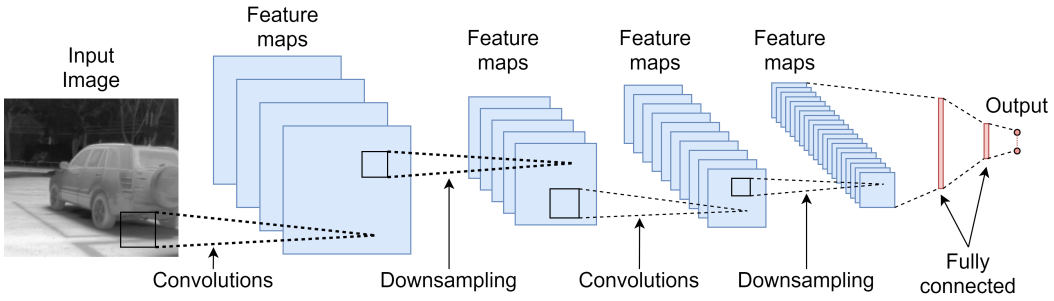


Fig. 2.3 The architecture of a typical convolutional neural network.

2.3 Deep learning

Deep Learning (DL) is a sub-class of machine learning that uses artificial neural network to learn complex relationships among data [14, 37]; its nature is on extracting the features from different types of data, supervised and unsupervised. DL has been widely used in computer vision, natural language processing, and many other fields [59, 39, 65].

Convolutional Neural Network (CNN) is a class of deep neural networks, mostly applied to analyze images. The convolutions layers are the core operations of a CNN; acting as a feature extractor and consist of a set of learnable kernels. Input images are two-dimensional data (matrix), where convolutions take advantage of doing a dot product with the kernels, producing a feature map with the most representative characteristic at some spatial position in the input. The illustration in Fig. 2.3 represents a typical CNN for classification. Transposed convolutional layers and sub-pixel convolutional layers are used in deep learning-based methods. Transposed convolutional layers are used to upsample an image by adding more pixels between existing pixels. This is done by using a convolutional layer with a stride of 2. Sub-pixel convolutional layers are used to upsample an image by adding more pixels between existing pixels. This is done by using a convolutional layer with a stride of 1 and a kernel size of 3. The down-sampling layer (a.k.a. Pooling Layer) has several non-linear functions that can be used, the most common is the *max pooling*; it divides the two-dimension input into several sub-areas, and for each sub-area, it takes the maximum value. These layers reduce the spatial dimension, by reducing the number of parameters being reflected in the computational cost.

Even though, CNN were proposed in the 1980s [26], it was not until 2012 that they became widely known after the success of the ImageNet Large Scale Visual Recognition Challenge (ILSVRC). The ILSVRC is an annual competition organized by the ImageNet project in which teams compete to correctly classify and detect objects in images. The competition is based on a subset of the ImageNet dataset [13], which contains 1.2 million images with 1000 object categories; where AlexNet [40] won the competition by a large

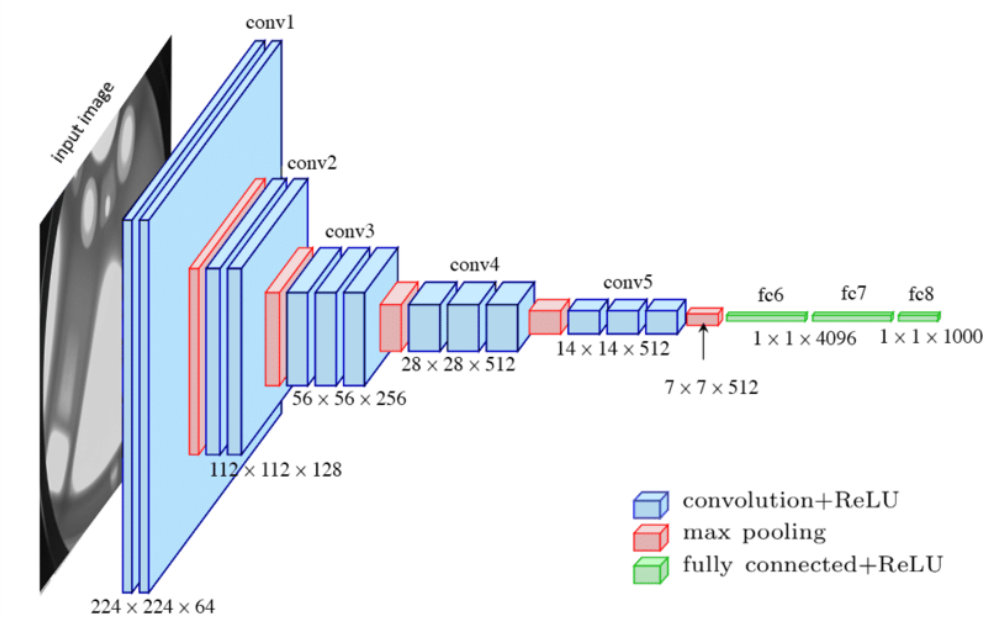


Fig. 2.4 The architecture of the VGG-16 network [105].

margin, using a deep CNN with 5 convolutional and three fully-connected layers. Since AlexNet, many CNN architectures have been proposed, such as VGG [105], ResNet [36], DenseNet [43], U-Net [98], MobileNet [41], among others for different computer vision applications such as image classification, object detection, edge detection, super-resolution, and semantic segmentation. Some of them are going to be mentioned below.

The Visual Geometry Group (VGG) proposes six CNN architectures, where VGG-16 and VGG-19 are the most used in different computer vision tasks. The VGG-16 is composed of 13 convolutional layers and 3 fully connected layers, while the VGG-19 has 16 convolutional layers and 3 fully connected layers. An illustration of the VGG-16 version can be appreciated in Fig. 2.4. VGG architecture scored the second place in the ILSVRC 2014 competition [100].

The Deep Residual Network for Image Recognition (ResNet) is proposed by Kaiming He et al. in 2015 [36]. As VGG, ResNet approach also has different versions according to the depth of the network. ResNet-50, ResNet-101, and ResNet-152 are the most used in different computer vision tasks. ResNet-50, shown in Fig. 2.5, has been frequently used in the literature. ResNet won the first place in the ILSVRC 2015 competition [100].

U-Net, proposed by Olaf Ronneberger et al. in 2015 [98], is a fully convolutional network for image segmentation. U-Net is composed with 23 convolutional layers, designed with an encoder and decoder path. The encoder path is a typical CNN that is used to capture

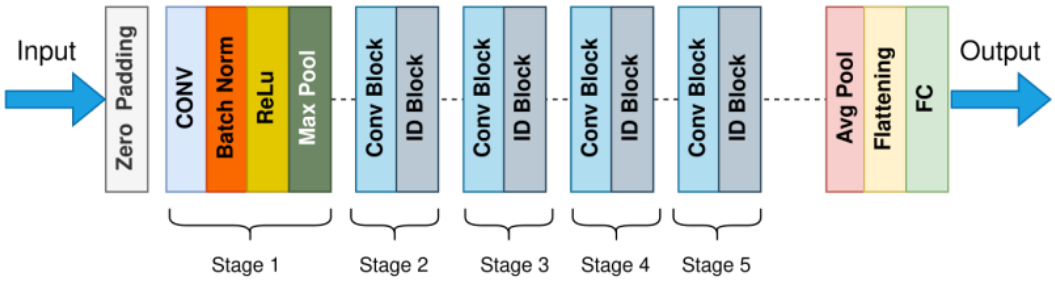


Fig. 2.5 The architecture of the ResNet-50 network [36].

the context in the image. The decoder consists of an upsampling process to enable precise localization. This architecture won the International Symposium of Biomedical Imaging (ISBI) cell tracking challenge. Google research proposes two CNN architectures that are widely used, Inception v3 [112] and Xception [10], with 23 and 22 million of parameters, respectively; both approaches use depthwise separable convolution instead of traditional convolutional layers.

Deep learning-based methods for image super-resolution (SR) have shown great promise in recent years. These methods are based on convolutional neural networks that are trained to map low-resolution images to high-resolution images. The main objective of deep learning is to develop mathematical models that enable artificial neural networks to learn by processing counterfeit information and knowledge in the human brain. The learning process on an artificial neural network is an iterative process that computes the loss function measuring the distance between the predicted value and the target value.

Convolutional Neural Networks are considered to answer the research questions. This decision has been made because CNN allows to get good representations of image features, which is the primary goal of computer vision. It has overcome the limitations of traditional upsampling methods. DL methods learn the mapping function from a large dataset of images; this allows them to capture the underlying structure of the image and produce better results.

2.4 Datasets

In order to train and evaluate proposed approaches in the visible spectrum, there are several public datasets available; the most widely and recently released dataset for visible image restoration is *DIV2K* [113], which contains high-quality (2K resolution) images split up into 800 images for training, 100 for testing, and 100 for validation. Another large-scale dataset is the Flickr2K dataset, which contains 2K images downloaded from Flickr. Most of the approaches in the literature use common benchmark datasets for evaluating their

Dataset	Avg. Resolution	Amount	Format	Content
Set5 [2]	313×336	5	PNG	baby, bird, butterfly, head, woman
Set14 [133]	492×446	14	PNG	humans, animals, insects, flowers...
BSD100 [71]	435×367	100	JPG	animal, building, food, landscape...
Urban100 [44]	984×797	100	PNG	architecture, city, structure, urban...
DIV2K [113]	1972×1437	1000	PNG	environment, flora, fauna, people...
General-100 [21]	435×381	100	BMP	animal, daily necessity, food, people...
L20 [114]	3843×2870	20	PNG	animal, building, landscape, people...
Manga109 [72]	826×1169	109	PNG	manga volume
OutdoorScene [117]	553×440	10624	PNG	animal, building, grass, mountain...
T91 [127]	264×204	91	PNG	car, flower, fruit, human face...

Table 2.1 List of some public visible image datasets commonly used for super-resolution training and benchmark.

performance, and all of them focus on the visible spectrum domain. Table 2.1 presents a list of the most commonly used visible spectrum dataset used by the SR community as a benchmark on super-resolution. These datasets provide HR images under different categories (e.g., animal, building, food, landscape, people, flora, fauna, car, etc.) with different images. Some even include low-resolution (LR) and high-resolution (HR) image pairs, while others only provide HR images. Some other datasets, intended for other vision task, have been also employed for SR, such as CelebA [67], VOC2012 [24], ImageNet [13], MS-COCO [66], BSDS300 [71], BSD500 [1], among others.

Regarding thermal images, in the last years some small-scale datasets have been proposed in the literature. In [12], a dataset with 284 thermal images is generated, with a resolution of 360×240. Images in this dataset have been acquired with a Raytheon 300D camera, images have been taken on their university campus at a walkway and street intersection, capturing images over several daytime and weather conditions. In [81], a 15224 thermal image dataset has been proposed, with a resolution of 164×129. This dataset is acquired with an Indigo Omega imager mounted on a vehicle, driving in outdoors urban scenarios. In [47] a FLIR-A35 is used to obtain more than 41500 thermal images, with a resolution of 320×256. In [9] a thermal dataset with 138 images is acquired, referred it as T-138, with a resolution of 640×480 using a FLIR T640 with a 41mm lens. A HR dataset is presented in [125]; the dataset contains seven different scenes, most of them collected with a FLIR SC8000, with a complete resolution of 1024×1024. The dataset consists of 63782 frames with thousands of recording objects; it is one of the enormous amounts of HR thermal images available. Most of the thermal image datasets mentioned above are usually designed for object detection and tracking; some others for applications in the biometric domain or medical applications are used (e.g., [8], [88]); and just a few of them are intended for super-resolution tasks. Also,

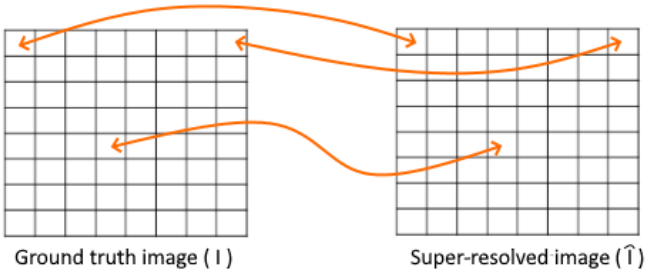


Fig. 2.6 Illustration of a pixel-wise comparison for metrics evaluation between GT and SR.

most of them contain low-resolution images, and others are from the same scenario, which gives poor variability.

Additionally, a dataset is proposed and acquired in [108], referred to as PROBA-V; which is composed of several images taken by the PROBA-V satellite [1]. The satellite provides images in two resolutions with different frequency bands; red visible (RED) and Near IR (NIR) spectral bands at 300-m (128×128 grayscale image) and 100-m (384×384 grayscale image) resolution. [75] presents SEN2VENµS, an open-data licensed dataset composed of 10 m and 20 m cloud-free surface reflectance patches, also acquired with a satellite. This dataset covers 29 locations on earth with a total of 132955 patches of 256×256. Furthermore, the use of generated noised downsampled images for multiple image super-resolution has been proposed (e.g., [3, 85]) to tackled the generation of a SR with the iterative process of multiple LR input images of the same scenario.

2.5 Evaluation metrics

Deep convolutional neural networks need evaluation metrics to learn the end-to-end mapping between an input and output image, in this case, the mapping of a reconstructed low-resolution to a high-resolution image or ground truth (Fig. 2.6). Even though there are several metrics, the most widely used quantitative metrics to evaluate the performance of the super-resolution methods in the single image super-resolution literature (e.g., [6, 56, 111, 127]) are: i) Peak Signal-to-Noise Ratio (PSNR) and ii) Structural Similarity Index Metric metrics (SSIM) [119].

Regarding PSNR metric, it is one of the basic and most popular denoising metrics and uses reconstruction quality measurement of lossy transformations and it has a higher correlation with the human perception. Eq. 2.1 show its formulation:

[1] https://esa.int/Applications/Observing the Earth/Proba-V

$$MSE = \frac{1}{N}\sum_{i=1}^{N}\left(I_i - \hat{I}_i\right)^2 \quad PSNR = 10 \cdot log_{10}\left(\frac{MAX^2}{MSE}\right), \tag{2.1}$$

where I refers to the given ground truth image with N pixels, and $\hat{\mathrm{I}}$ is the reconstructed image or the SR output image. MAX equals the highest pixel value of the image in general cases. MSE corresponds to Mean Square Error operation.

Respecting SSIM metric, as shown in Eq. 2.2, it is a perception-based metric that is based on the independent comparisons of luminance, contrast, and structure:

$$SSIM\left(I,\hat{I}\right) = \frac{\left(2\mu_I\mu_{\hat{I}} + C_1\right)\cdot\left(\sigma_{I\,\hat{I}} + C_2\right)}{\left(\mu_I^2 + \mu_{\hat{I}}^2 + C_1\right)\cdot\left(\sigma_I^2 + \sigma_{\hat{I}}^2 + C_2\right)}, \tag{2.2}$$

where μ_I and σ_I correspond to the luminance and the contrast, both are estimated as the mean and standard deviation of the image intensity; and C represents constant relaxation terms for avoiding instability. A higher score of PSNR or SSIM means better restoration fidelity.

Image quality assessments, focused on the perception of human viewers, are avoided in the current thesis due to these measures are not necessarily consistent in the case of thermal images. Furthermore, this is an expensive and time-consuming, and these methods are based on humans' perception, i.e., how realistic the image looks. Other methods, less popular SR metrics, for measuring image quality include the Multi-Scale Structural Similarity (MS-SSIM) [121] and the Feature Similarity Indexing Method (FSIM) [137]. These methods offer more flexibility than the single-scale SSIM in incorporating the variations of viewing conditions. Natural Image Quality Evaluator (NIQE) [76] and the Perceptual Quality Index (PQI) [80] are also proposed to assess the quality of images. NIQE makes use of measurable deviations from statistical regularities observed in natural images without exposure to distorted images. PQI uses a combination of human visual system models and machine learning to predict the perceptual quality of images.

Although there are different evaluation techniques to assess SR results [16], in general, PSNR and SSIM are considered quantitative metrics to evaluate the performance of different approaches [73]. Even though PSNR and SSIM are not the best metrics to measure the quality of the restored thermal images (that are acquired from the LWIR spectral band), these images are mostly represented in grayscale images so that these measure metrics can be used on them. Other types of evaluations could be performed to measure the improvements in the quality of the SR generated images (e.g., visual task-based evaluation metrics in medical imaging [51]) but these task-based metrics are not tackled in the present work.

2.6 Super-Resolution

Super-resolution is the process of recovering high-resolution images from low-resolution images, also known as image up-scaling or image enhancement. Super-resolution can be used to improve the quality of images or frame rate. Even though super-resolution is a well-studied problem in the field of image processing, it is still an active area of research due to the challenges involved in recovering high-resolution images from low-resolution images. In this process, there are mainly two types of super-resolution methods: single image super-resolution and multi-image super-resolution which are described in the following sections.

2.6.1 Single Image Super-Resolution

Although not intended for super-resolution, image enhancement approaches are related in the sense that these approaches tackle the process of restoring a given image to have a more suitable representation for the desired application. Generally, image enhancement covers many techniques to enhance the visual appearance of an original image for machine analysis or better appearance for humans. The main point of image enhancement is to restore the image with the same resolution where it has "noise", blurred focuses, and colorless, among others.

Like in image enhancement, the Single Image Super-resolution (SISR) aims to restore an image, not necessarily noisy or blurred images, but images with low-resolution being transformed into a high-resolution image. The SISR has been extensively studied during the last decades, and in the literature, there are many methods proposed to solve this problem; recently, different deep learning architectures have been proposed to solve this problems, obtaining better results than conventional methods.

Super-resolution approaches have different techniques, such as: Linear, Residual, Recursive, Progressive, Densely Connected, Multi-branch, Attention Based, Generative Adversarial Networks architecture, where there are mainly four model frameworks based on the employed upsampling operations for super-resolution:

1. **Pre-upsampling SR:** This framework first upsamples the low-resolution image using interpolation methods such as bicubic interpolation or bilinear interpolation. The upsampled image is then fed into a deep neural network for further processing (Fig. 2.7). This framework is simple and fast, but it does not produce good results.

2. **Post-upsampling SR:** In order to improve the computational efficiency, this framework first uses a deep neural network to map the low-resolution image to a high-resolution feature map. The feature map is then upsampled using interpolation methods such as bicubic

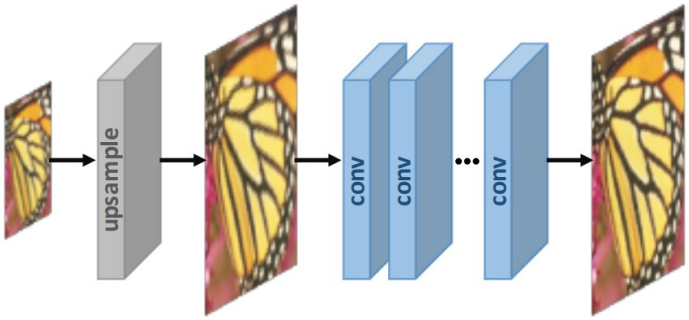


Fig. 2.7 Pre-upsampling SR illustration [120].

interpolation or bilinear interpolation (Fig. 2.8). This framework is more effective than the pre-upsampling SR framework because it uses a deep neural network.

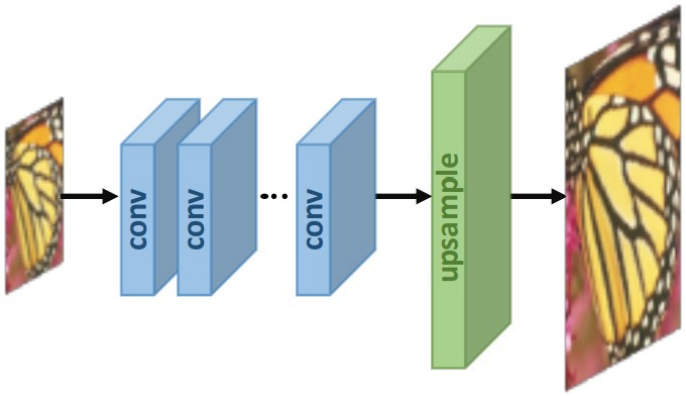


Fig. 2.8 Post-upsampling SR illustration [120].

3. **Progressive upsampling SR:** Even though post-upsampling reduced the computational cost, it has some overcoming. The upsampling is done in only one step, which increases the learning difficulty for large scaling factors. Also, it requires training an individual SR model for each scale, avoiding the multi-scale option. To overcome this drawbacks, a cascade of convolution layers and progressively reconstructed HR images is proposed (Fig. 2.9). This reduces the learning difficulty by decomposing it in several tasks.

4. **Iterative up-and-down sampling SR:** To capture the mutual dependency between LR-HR image pairs, an iterative back-projection is introduced (Fig. 2.10). This tries to iteratively apply back-projection refinement exploiting the up-and-down sampling layers, which connects upsampling and downsampling layers alternately and reconstructs the final HR result.

The use of convolutional neural networks has shown a great capability to increase the quality of SR results [129], where several CNN approaches have been implemented using different architectures. Dong et al. [18], [19] firstly propose a super-resolution using

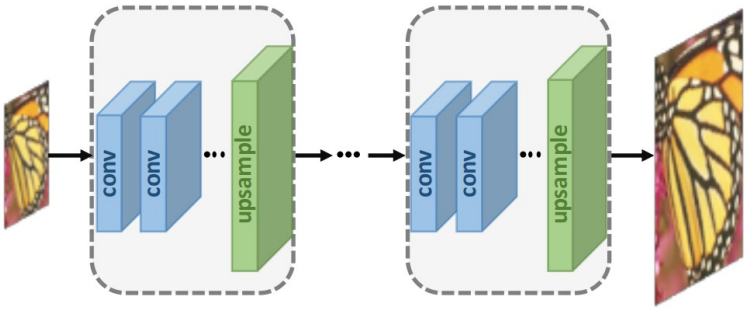


Fig. 2.9 Progressive upsampling SR illustration [120].

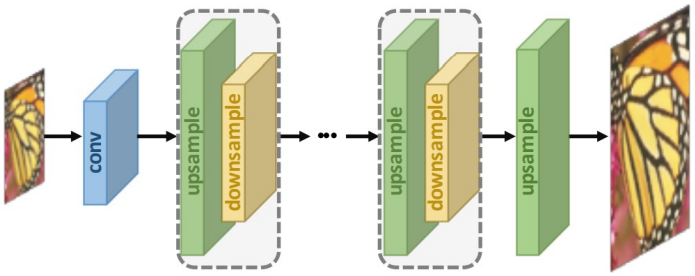


Fig. 2.10 Iterative up-and-down Sampling SR illustration [120].

convolutional neural networks (SRCNN) to learn an end-to-end mapping between the interpolated LR images and their HR counterparts. It uses traditional methods (e.g., bicubic interpolation) for having the same of HR output image, then deep CNNs are applied on these images to reconstruct high-quality details, reducing the learning difficulty; this first approach archives state-of-the-art performance and has become one of the most popular frameworks. However, using predefined upsampling, traditional methods introduce side effects like noise amplification and some blurring, and the cost of time and space is much higher than other frameworks. For better performance and increase the efficiency, a fast super-resolution convolutional neural network is proposed by [21], which performs a fast SRCNN extracting feature maps at the low-resolution image and upsample the image at the last layer; the authors introduce a deconvolution layer at the end of the network and adopt a smaller filter size but more mapping layers. Inspired by SRCNN, depth networks start to appear, and the first one is VDSR [54], which proposes a very deep network with 20 layers to extract features from the low-resolution image. The VDSR network is trained to minimize the MSE between the HR image and the reconstructed image. Same wise is proposed by [135] using a ResNet-based architecture to extract features from the low-resolution image and stacking more convolutional layers with residual learning [36]. A Deep Recursive Convolutional

Network (DRCN) is proposed by [55], where it applies the same convolution layers multiple times; the advantage of this method is that the number of parameters remains constant for more recursions. Another deep CNN is proposed by [70], referred to as RED, which uses a recurrent neural network to extract features from the low-resolution image. To speed up most of the training process, [63] proposes an architecture, referred to as EDSR, which removes the batch-normalization layer and takes advantage of residual learning. Yamanaka et al. [126] propose a CNN-based approach, where it uses a deep CNN with residual net, skip connections, and network-in-network and gets a computation complexity of at least ten times smaller than state-of-the-art (e.g., DRCN) reaching similar results.

Han et al. [35] propose a novel multi-scale spatial and spectral fusion architecture, taking advantage of the rich spatial context in HR-RGB and the spectral attribute in LR-HyperSpectral image, with a spatial structure reservation pathway for SR. Regarding thermal images, a CNN-based approach has been introduced by Choi et al. in [9], which is inspired by the proposal in [21]. The authors in [9] compare the accuracy of a network trained in different image spectral bands to find the best representation of thermal enhancement; concluding that a grayscale trained network provides better enhancement than the MWIR-based network for thermal image enhancement. On the other hand, Lee et al. [60] also propose a convolutional neural network based on image enhancement for thermal images. The authors evaluate four RGB-based domains, namely, gray, lightness, intensity, and V (from HSV color space) with a residual-learning technique. The approach improves the performance of enhancement and speed of convergence. The authors conclude that the V representation is the best for enhancing thermal images. In [64] the authors propose a parallelized 1x1 CNNs, named Network in Network, to perform image enhancement with a low computational cost for image reconstruction. In most previous approaches, thermal images are not considered during the training stage, although intended for thermal image enhancement; proposing to train a CNN-based approaches using images from the visible spectrum at different color space representations (e.g., grayscale, HSV). Yang et al. [128] propose a 3D convolution architecture for hyperspectral image super-resolution using wavelets prediction techniques, while [61] uses a 3D modify architecture, based for colorization, used in SR task.

Most of the aforementioned CNNs aim at minimizing the mean-square error between the SR and the ground truth (GT) image, tending to overthrow the high-frequency details in images. In other words, a supervised training process using a pair of images is followed. The main drawback of such approaches lies in the need of having pixel-wise registered SR and GT images to compute the MSE. As mentioned above, in most of the cases, the SR image is obtained from an image down-sampled from the GT. Since it is difficult to collect images of the same scene but with different resolutions and to overcome the pixel-wise

registration limitation between SR and GT images, unsupervised approaches have been proposed, where unpaired LR-HR images are provided for training. For instance, [104] proposes a single image super-resolution approach, referred to as SRGAN, which achieves impressive performance with respect to the state-of-the-art approaches. This approach is inspired by the seminal Generative Adversarial Network (GAN) presented in [30]; then, [118] proposes an improved version of SRGAN, denominated ESRGAN, with a better Residual-in-Residual Dense Block (RRBD), removing the batch-normalization and using a Relativistic average discriminator (RaGAN). In recent literature, different unsupervised training processes have been presented for applications such as transferring style [5], image colorization [74], image enhancement [7], feature estimation [109], among others. All these approaches are based on two-way GANs (CycleGAN) networks that can learn from unpaired datasets [140]. CycleGAN can be used to learn how to map images from one domain (source domain) into another domain (target domain); this functionality makes CycleGAN model appropriate for image SR estimation when there is not a pixel-wise registration.

The main challenge of SISR is to find the mapping between the LR and HR images, which is an ill-posed problem, where in the computer vision community this is still an active research field (e.g., [84], [34], [107], [123]). Several applications in different fields can benefit from SR representations, for instance security (e.g., [136], [102]), medical imaging (e.g., [78]), object detection (e.g., [29]), astronomical images (e.g., [68]), among others. In recent years, long-wavelength infrared (LWIR) images, a.k.a thermal images, have shown to be useful to efficiently solve problems from different domains (e.g., security monitor [99], medical imaging [38], car assistance [17], visual inspection [132], human detection [33], among others) because thermal images have the information of the radiation emitted by the surface of an object (temperature above zero [28]) captured by thermal cameras. As mentioned above, thermal cameras play an important role in different areas and have increased during the last two decades, in particular, thermal imagery, due to the cost reduction and availability of thermal cameras. Unfortunately, most affordable thermal cameras have poor resolution, and high-resolution ones are still expensive nowadays. Despite the continuous increase in the use of thermal cameras, there is still a limitation on image resolution. A possible way to overcome this limitation could be to develop a CNN-based architecture to generate a HR representation from a given LR image. Single thermal image super-resolution has become an active research topic in the computer vision community.

2.6.2 Multi-Image Super-Resolution

Most of the approaches mentioned in the previous section expose the use of deep learning techniques for single image super-resolution and include image restoration and noise

reduction since there are problems that are intrinsically related to the SR process. However, another research topic that will be considered in this thesis consists on using a set of LR images of the same scene with sub-pixel shifts to generate super-resolution images. Multi-Image Super-Resolution (MISR) involves extracting features from several LR images from the same scene to reconstruct a HR image [131]. MISR is also referred to as Multi-Frame Super-Resolution (MFSR). MFSR is typically regarded as a problem that lacks a unique and stable solution [25], as the HR image is degraded to the LR image by an unknown source. This unknown source can for instance, be modeled by considering a combination of atmosphere blur, camera blur, motion effect, and downsampling (Fig. 2.11). MISR approaches must often address some images inconsistencies that increase the complexity of their registration. Image registration is the process of aligning two or more images of the same scene. This is done by finding corresponding points between the images and then transforming them to be lineup. Image registration can be used for various purposes, such as motion estimation, image stitching, and super-resolution.

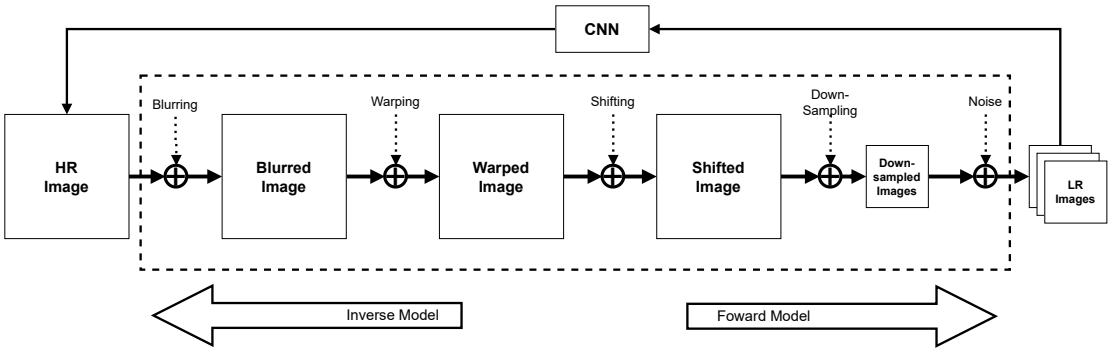


Fig. 2.11 Illustration model employed in most MISR techniques for multi LR generation.

SR image reconstruction approaches can be categorized into three classes: frequency-domain, interpolation-based, and regularization-based approaches [53].

- Frequency-domain approaches are based on the Fourier Transform (FT) of the LR images. The FT of the LR images is multiplied by a filter that is designed to enhance the high-frequency components. The Inverse Fourier Transform (IFT) is then applied to the filtered LR images to obtain the HR image.
- Interpolation-based approaches are based on the interpolation of the LR images. The interpolation is done by using a set of interpolation kernels that are applied to the LR images. The interpolated LR images are then combined to generate the HR image.
- Regularization-based approaches are based on the minimization of an objective function that is composed of a data fidelity term and a regularization term. The data

fidelity term is used to enforce the similarity between the HR image and the LR images. The regularization term is used to enforce the smoothness of the HR image.

The first work in MISR is proposed by Tsai and Huang [115], based on the frequency domain; it uses the combination of multiple images with sub-pixel displacements with frequency-domain techniques to upgrade the spatial resolution. Then, several spatial domain MISR techniques are considered [23] (e.g., projection onto convex sets, non-uniform interpolation, sparse coding, regularized methods). In [50], the authors take advantages of multiple image fusion by learning the low to high-resolution mapping using deep networks, as shown in Fig. 2.12; where each of LR input image is subject to single-image SR using ResNet, then these images undergoes to a registration process to determine sub-pixel shifts between images and finally employ a genetic algorithm to optimize the hyper-parameters and generate a SR image. Recently, a multi-image super-resolution algorithm applied to thermal imagery has been proposed by Mandanici et al. in [69]; this approach is tested and applied to terrestrial thermal imaging to overcome the limitation of the low resolution.

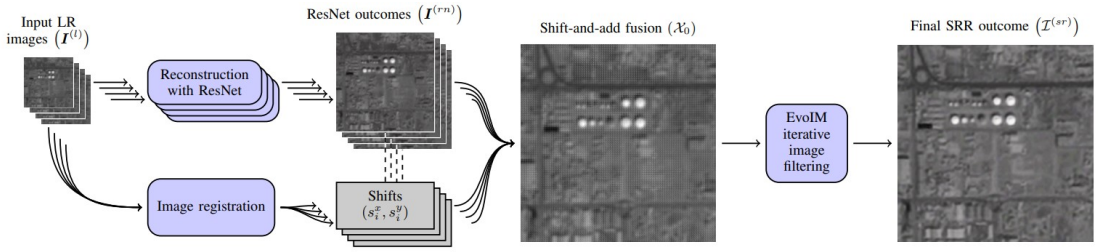
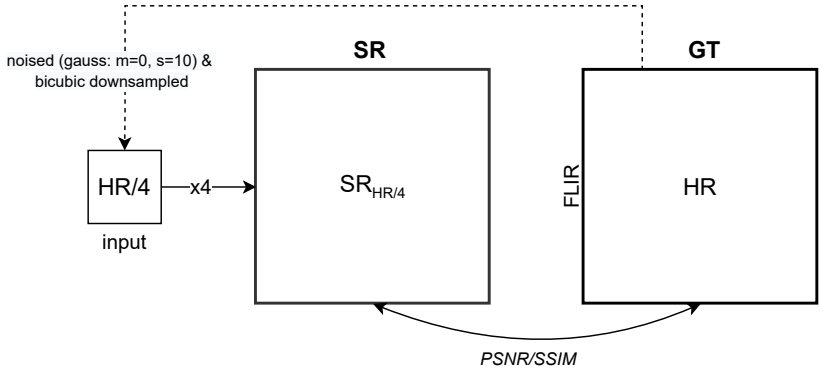
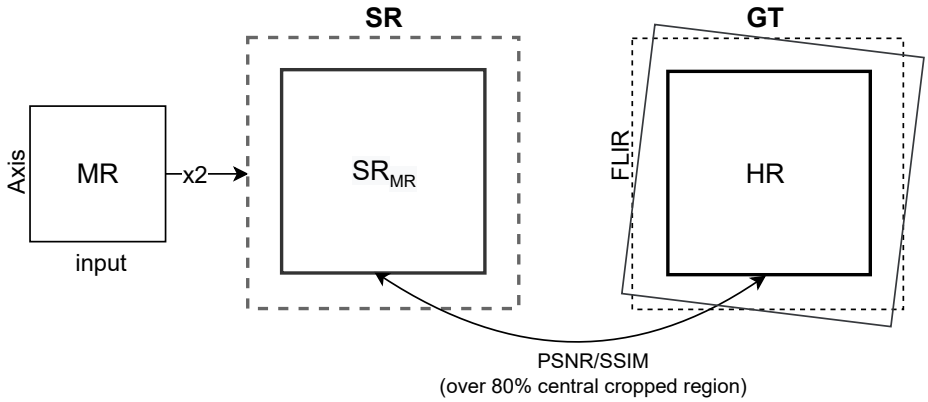


Fig. 2.12 Deep learning proposed for multiple-image super-resolution [50].

Recently, multi-image super-resolution of remotely sensed images using a novel Residual Feature Attention Deep Neural Networks has been proposed by Salvetti et al. [101]. It efficiently tackles MISR tasks, simultaneously exploiting spatial and temporal correlations to combine multiple images. European Space Agency set a recent challenge on MISR to tackle SR satellite images, where some novel architectures are proposed (e.g., [77] [15]). Multispectral satellite image datasets (e.g., Sentinel-2, MODIS) are generally used for tackling MISR; Lanaras et al. [58] infer all the spectral bands of multiresolution sensors in the highest available resolution of the sensor. Also, Muller et al. [79] propose a method to train state-of-the-art CNNs using pairs of LR multispectral and HR pan-sharpened image tiles to create SR representations. With more data available from the multiple observation of the scene is possible to achieve higher reconstruction accuracy than SISR approaches.



(a) First evaluation process ($\times 4$).



(b) Second evaluation process, from MR to HR ($\times 2$).

Fig. 2.13 Proposed evaluations processes for challenges [94].

2.7 Thermal Image Super-Resolution challenges

Three editions of the Thermal Image Super-Resolution (TISR) challenge have been organized in the Perception Beyond the Visible Spectrum (PBVS) workshop ([89, 95, 94]) of CVPR conference. The TISR challenge aims to introduce state-of-the-art approaches for the thermal image SR problem and evaluate and compare different solutions using the previous year results as benchmark. In computer vision, super-resolution takes a low-resolution image and turns it into a high-resolution image. Most techniques used for this purpose are deep learning-based and use a downsampled image from the high-resolution image as input. This image is then augmented with noise and blur to create a new image that is used to train the network. This challenge consists of two evaluations using the second acquired dataset (Th3D-1021) mentioned in Section 3.2.

In the framework of thermal image super-resolution, I have been in charge of these three challenges in the PBVS workshop. Although it does not represent a scientific contribution

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

Table 2.2 Average results for each evaluation metric of the PBVS-CVPR2022 challenge. Bold and underline values correspond to the best- and second-best results, respectively [94]. Results from just top teams are depicted.

of this thesis, results of the last challenge are presented here just to show the evolution of results and the architectures of the state-of-the-art.

2.7.1 Evaluations

Two kinds of evaluations are performed considering PSNR and SSIM as metrics. In Evaluation 1, a set of 10 noisy and downsampled images obtained from HR camera (FLIR FC-632O) of the Th3D-1021 dataset is evaluated. 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 2.13 (a) shows an illustration of this first evaluation process.

For Evaluation 2, the $\times 2$ SR results obtained from the input images of the MR camera (Axis Q2901-E) are evaluated with respect to the corresponding semi-registered images obtained from the HR camera (FLIR FC-632O). Somehow this evaluation tackles into account two problems, first generating the SR images acquired with the MR camera, and then mapping images from the MR domain (Axis camera) to the HR domain of another camera (FLIR camera). Figure 2.13 (b) shows an illustration of this second evaluation process.

There were more than 100 teams registered for the challenge on the last edition (PBVS-CVPR2022), and more than 50 submitted their final results. Top teams with higher evaluation results are selected; whom submit their corresponding extended abstracts. The quantitative average results (PSNR and SSIM) for each team in the two evaluations are shown in Table 2.2 (just top teams are depitected). More quantitative results can be found on

CodaLab Competition [83] webpage[2]. Next section presents details of the architectures of the teams with the best score in each evaluation metrics in last challenge (PBVS-CVPR2022).

2.7.2 Approaches

AIR team, winner of Evaluation 1 (on PSNR); it proposes a Convolution Attached Transformer Super-resolution Networks (CATS), as shown in Fig. 2.14, which is composed of convolutional blocks (CBs) and transformer blocks (TBs), allowing the model to take advantage of both the speed of the CNN-based approach and the restoration performance of the transformer-based approach. Detail-Fidelity Attention Module (DeFiAM) [45] is adopted instead of vanilla Residual Channel Attention Block to capture more accurate low-frequency structures and high-frequency details. Each DeFiAM extracts low-frequency structure and high-frequency details in a CNN-like manner. The data is then passed through a step consisting of one Swin TB [62] and a CB, which serves to transmit information to the subsequent TBs smoothly. The successive Swin TBs achieve high efficiency as it handles image information that has been refined before. 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.

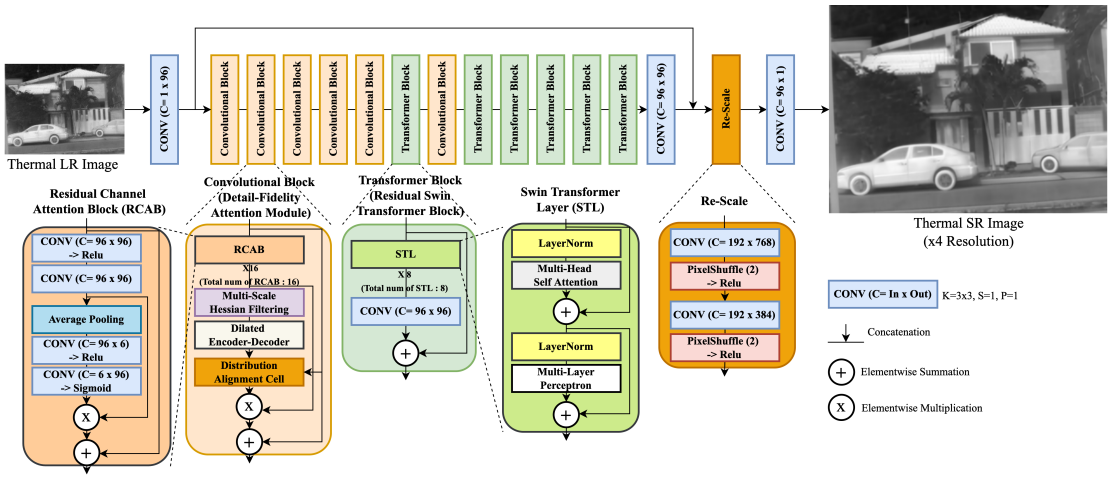


Fig. 2.14 Architecture proposed by AIR team (CATS) [94].

NJU team, winner of Evaluation 1 (on SSIM), proposes to use SwinIR [62] as base model, a pre-trained model for RGB images, for thermal image super-resolution; increasing the number of IR image channels and fed it into the SwinIR model. The output of the model is then averaged along the channels. For the training data, the high-resolution image is added

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

with Gaussian noise and then simulated the low-resolution image by JPEG compression. Random rotation and flip are added in training data generation. Finally, the three models are integrated using $L1$, $PSNR$, and MSE as the loss.

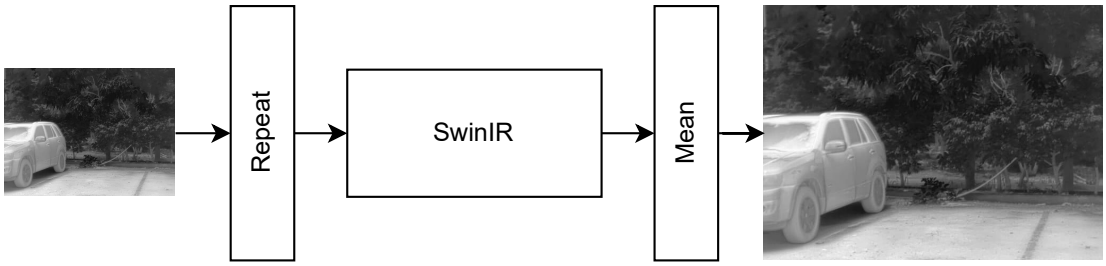


Fig. 2.15 Architecture proposed by NJU team [94].

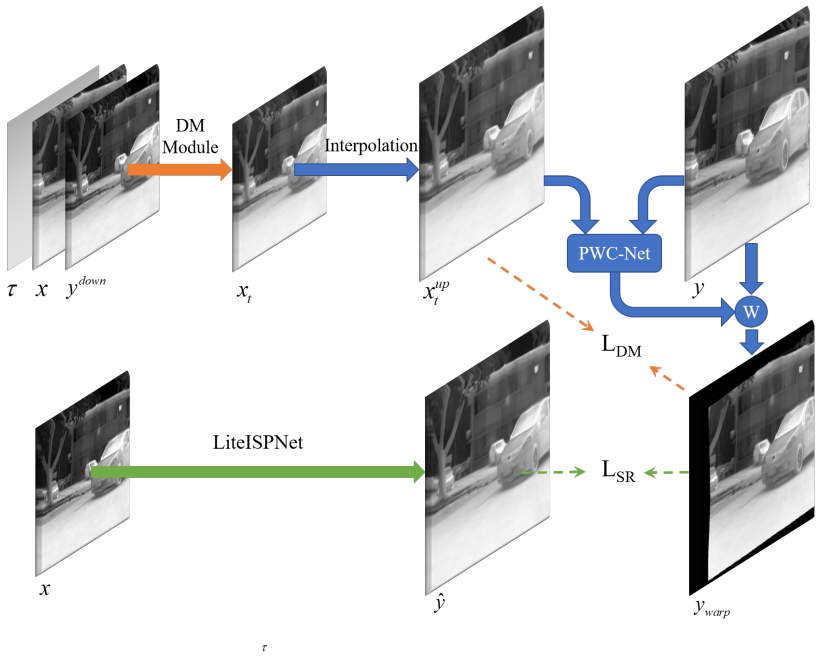


Fig. 2.16 Architecture proposed by NPU-LIFT-LAB [94].

NPU team, winner of Evaluation 2; it proposes a method to improve the performance of image super-resolution by leveraging domain adaptation, image alignment, and super-resolution. The technique is designed to solve the problem of domain shift, misalignment, and low resolution. The proposed network architecture, as shown in Fig. 2.16, is composed of a Domain Model module [138] (for generating a domain-adjusted image), a PWC-Net [110] (leveraged to estimate the optical flow) and a LiteISPNet [138] (to learn the mapping between the input and the updated label); all these modules can better

utilize the information from domain adaptation, image alignment, and super-resolution. The self-ensemble strategy [114] is adopted to increase the model's performance.

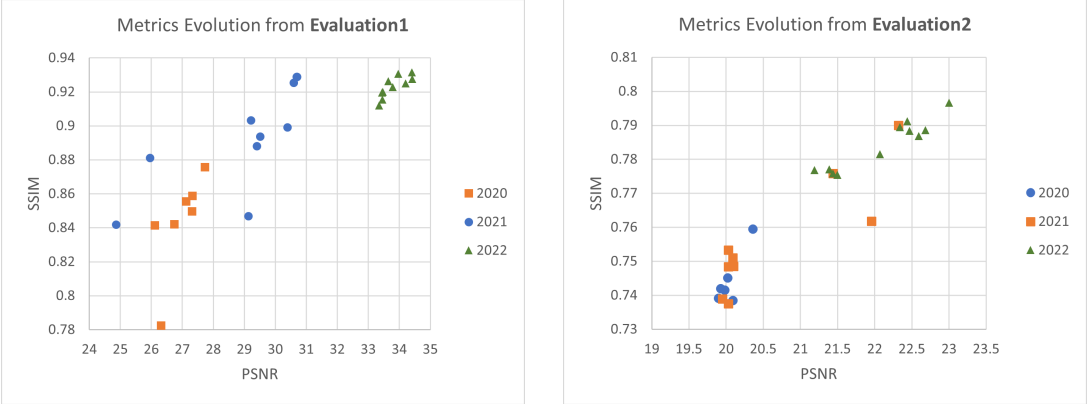


Fig. 2.17 Metrics evolution through all challenge editions [94].

Obtained results show the metrics' values are getting higher each year, as shown in Fig. 2.17. The presented approaches are all deep learning-based and use different CNN or transformer architectures, or combinations of them. The number of participants in the last edition is higher than in the previous years, and the results obtained from the quantitative evaluations outperformed the results of the last two years. These results can be used as a baseline for the following challenges editions.

Chapter 3

Dataset Acquisition

This chapter tackles the thermal camera sensor setup and the image acquisition task for collecting different datasets that will be used throughout the thesis. Also, it introduces the types of cameras available in the market and the ones used in this thesis.

3.1 Introduction

Although thermal cameras' price has decreased over the past years, they are still relatively expensive. Therefore, the setup of the thermal camera sensor is a crucial step in the image acquisition process. This camera sensor is a passive sensor that detects the infrared radiation emitted by objects. Infrared radiation is a type of electromagnetic radiation with a wavelength longer than visible light. The thermal camera sensor is a non-contact sensor that can estimate the infrared radiation emitted by objects without touching them. It can be used to measure the temperature of things cause is related to the amount of infrared radiation emitted by the objects.

Fig. 3.1 A FLIR A66xx cooled and a FLIR Quark 640 uncooled thermal cameras sensor.

There are many wavelengths thermal imaging camera sensor system choices in the marketplace. In general, they can be classified into two categories according to the type of

thermal imaging sensors (as shown in Fig. 3.1): cooled and uncooled. Cooled sensors are kept at an extremely low temperature, while uncooled sensors operate at ambient temperature. Cooled sensors are more sensitive and reliable, but expensive than uncooled ones.

The main advantages of uncooled thermal cameras sensor are that they are cheaper and easier to use than cooled thermal cameras sensor. However, the cooled ones have many advantages over uncooled cameras, including better image quality, faster speed, and the ability to see tiny temperature differences. For deciding which type of thermal camera sensor to use, it is essential to consider the specific needs of the application.

In this thesis, three different datasets have been acquired using uncooled thermal camera sensor due to the reasons mentioned above.

3.2 Dataset generation

To test the development of the thesis work and try to overcome the limitations mentioned in Section 2.4, three kinds of datasets have been acquired due to the lack of thermal image datasets, each one referred to as: ThD1-101, ThD2-200 and ThD3-1021.

Fig. 3.2 Example set of the first dataset acquired using a TAU2 camera (ThD1-101).

Image Description	Brand Camera	FOV	Focal Length	Native Resolution
Low-resolution (LR)	Axis Domo P1290	35.4	4mm	160×120
Mid-resolution (MR)	Axis Q2901-E	35	9mm	320×240
High-resolution (HR)	FC-632O FLIR	32	19mm	640×512*

Table 3.1 Thermal camera specifications (*the HR images have been center crop to 640×480) for ThD3-1021 dataset acquisition.

3.2.1 Single sensor

The first thermal dataset (ThD1-101) [96] has been acquired using a single TAU2[1] thermal camera with a 13mm lens (45° HFOV) with a native resolution of 640×512 pixels, in indoor and outdoor scenarios, in the morning, day and night time; containing multiple objects and people. Using the controller GUI software of the TAU2 camera with the default values, a set of 101 images (PNG format with a depth of 8 bits) has been acquired; Fig. 3.2 shows some images of this dataset. This dataset is used to train CNNs following the traditional method of down-sampling the images to have a registered image between SR and GT images. The idea of this first contribution is to provide to the thermal image dataset community, with a new dataset for the thermal image super-resolution problem.

3.2.2 Cross-spectral sensors

For cross-spectral image processing, where one spectral band guides or supplements another to solve a particular task, it is essential to have a dataset that contains images of the same scene acquired in different spectral bands. A second thermal dataset (ThD2-200) is acquired to tackle this problem by opening more robust solutions. It consists of visible and thermal pair images of the same scene caught in daylight conditions. They are all good quality images with clear edges with various scenarios, this makes it very suitable for the SR training. The main idea is to use a HR visible image as a guide for the LR thermal image to generate a SR thermal image.

This dataset has been acquired using Balser and TAU2 camera (Fig. 3.3) with a native resolution of 1080×1024 (using a 13mm lens) and 640×480 (using a 8mm lens) respectively. The dataset is composed of 200 pairs of images. The dataset is split up into training, validation, and testing sets. The training set contains 160 pairs of images, while the validation set contains 40 pairs of images and 20 pairs for testing. Examples of this dataset are depicted in Fig. 3.4.

The challenge of using thermal and visible images together is that it can be difficult to register the two images, especially if the overlap between the images is large. This registration

[1] https://www.flir.com/products/tau-2/

Fig. 3.3 TAU2 (*top*) and Balser (*bottom*) cameras mounted one over the other.

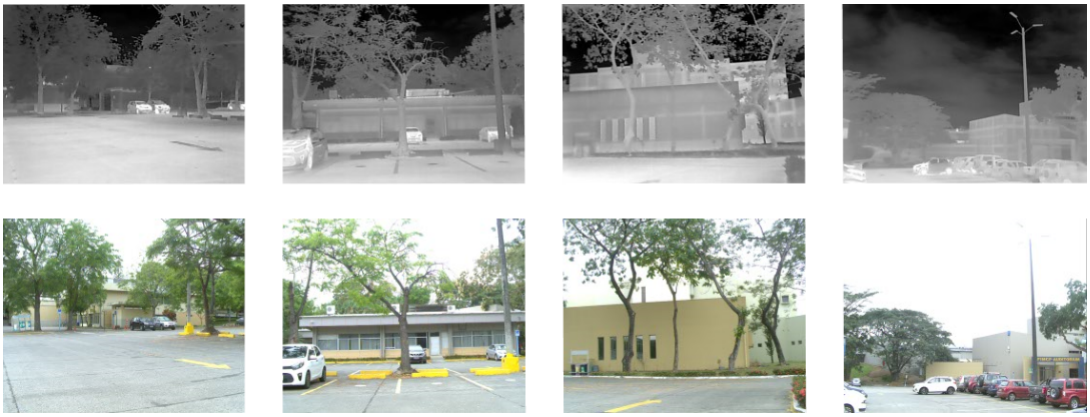


Fig. 3.4 Illustrations of ThD2-200 dataset, thermal and visible images of the same scenario.

problem is traditionally solved by matching descriptors, but this method depends on the quality of the representation. Ensuring that features are dense and uniformly distributed is not guaranteed. Recently, machine learning methods have addressed the issue of visible-to-visible matching, but just a few approaches address the multi-modality setting. The registration of this stereo images is not tackled in the present thesis.

3.2.3 Multiple sensors

Regarding the third and last thermal dataset (ThD3-1021) [90], unlike the first datasets, the main goal is to have three semi-registered thermal images, using three different thermal camera sensors of an outdoor scenario (that contains different objects such as buildings, cars, people, vegetation, among other) at different daytime. For this, the cameras are physically mounted on a panel, as shown in Fig. 3.5; this structure has been placed on a mobile platform to simultaneously capture different scenarios using a multi-thread developed script. Each

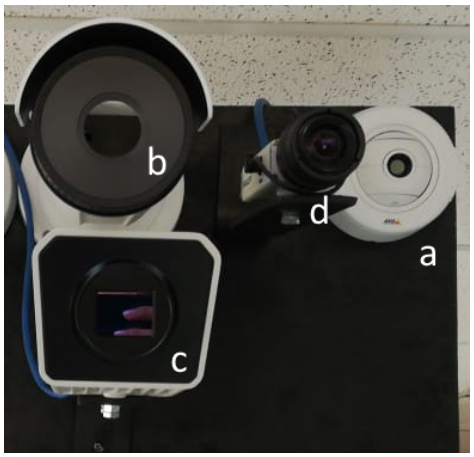


Fig. 3.5 Panel with the cameras for the second dataset: *a*) Axis Domo P1290 (LR); *b*) Axis Q2901-E (MR); *c*) FC-6320 FLIR (HR); *d*) Basler visible spectrum camera, which is not used in the current work.

camera has different resolution (low-, mid-, high- resolution) with 160×120, 320×240 and 640×480 pixels respectively, as shown in Table 3.1 and depicted in Fig. 3.6; Fig. 3.7 shows a mosaic build up with the three different resolutions. This dataset is acquired using an Axis Domo P1290, Axis Q2901-E, and FC-6320 FLIR thermal cameras, labeled as LR, MR, and HR, respectively. A set of 1021 images per camera is generated, split up into 951 images for training, 50 for validation, and 20 for testing.

Acquired images are saved in one channel, in PNG format with a depth of 8 bits and without compression. The idea of this dataset is to develop an architecture able to transform a real low-resolution image into a high-resolution image and with a semi-paired image to be able to have real GT images to be compared. Despite the effort during the camera setup, captured images have slightly different regions from the scene (Fig. 3.8); this is caused by the camera baseline between the optical center and the different intrinsic parameters of the cameras. Keeping in mind these limitations, a set of 10 images per each resolution (LR, MR, HR) is selected and registered to be used for testing any SR network approach, as shown in Fig. 3.9.

In all acquired datasets, thermal images are represented in grayscale, where hot and cold spots are represented with white and black pixels, respectively. Data augmentation process can be performed to have more variability and avoid the overfitting of any network implementation. This has been demonstrated to improve networks training results [103] [82].

Fig. 3.6 Examples of thermal images acquired by each camera of the second thermal dataset (ThD3-1021). $(left)$ LR image with 160×120 native resolution from Axis Domo P1290. $(middle)$ MR image with 320×240 native resolution from Axis Q2901-E. $(right)$ HR image with 640×480 resolution from FC-6320 FLIR (native resolution 640×512).

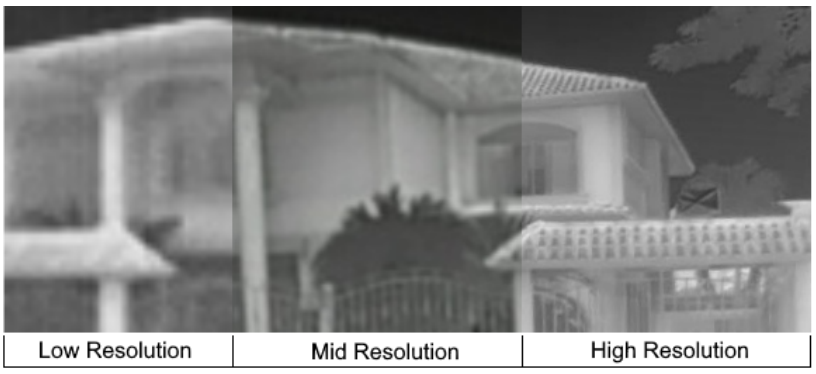


Fig. 3.7 Mosaic with three different resolution thermal images from ThD3-1021 dataset for a visual comparison: $(left)$ crop from a LR image; $(middle)$ crop from a MR image; $(right)$ crop from a HR image.

3.3 Summary

Thermal images offer great potential for many applications (e.g., security, medical, industrial applications) due to their ability to provide information about the temperature of an object in the scene. However, there is a lack of publicly available datasets that can be used to train and test algorithms for thermal image analysis and thermal image super-resolution. This chapter presents three new datasets that can be used for this purpose. The first dataset (ThD1-101) is a collection of over 101 thermal images of indoor and outdoor scenarios taken under various conditions. The second dataset (ThD2-200) that uses a visible camera and a thermal camera, contains 200 stereo images. And the third dataset

Fig. 3.8 ThD3-1021 dataset examples, thermal images of each camera and their slightly different regions of the scene. $(top-row)$ LR images from Axis Domo P1290. $(middle-row)$ MR images from Axis Q2901-E. $(bottom-row)$ HR images from FC-6320 FLIR.

Fig. 3.9 Image registration results. (top) From left to right: LR image; MR image; and image resulting from the registration of MR image with the results of SR_{LR} image. $(bottom)$ From left to right: MR image; HR image; and image resulting from the registration of HR image with the results of SR_{MR} image.

(ThD3-1021) uses three different thermal cameras to collect 1021 thermal images per camera of various scenarios during different daytime. Data augmentation process can be applied to these acquired datasets. All these acquired datasets are available and free to download at https://github.com/rafariva/ThermalDatasets.

Chapter 4

Single Image Super-Resolution

This chapter presents two kinds of techniques for single image super-resolution (SISR). The first approach follows a supervised scheme while the second one is an unsupervised method. Both approaches mainly consist of taking a single low-resolution (LR) image as input and generating a high-resolution (HR) one.

4.1 Introduction

As mentioned in Section 2.6, SISR is a classical and challenging ill-posed problem in computer vision that tries to infer a HR image from a single LR input image. This problem is ill-posed because there are an infinite number of possible HR images that could have generated the given LR image. The solution to this problem is usually not unique and it is often necessary to make assumptions or use some kind of prior knowledge in order to find a solution. A common approach for solving the SISR problem is to use a training set of HR/LR image pairs in order to learn the mapping between the LR and HR images. This mapping is then used to generate a HR image from a given LR image. This approach is known as example-based super-resolution, and it is effective in many cases. However, it requires a training set of HR/LR image pairs, which can be difficult to obtain.

An alternative approach is to use a generative model that can generate HR images from LR images. This approach is known as generative super-resolution, and it does not require a training set of HR/LR image pairs. Most of the proposed approaches presented in the state-of-the-art are focus on visible spectrum. One of the main contribution of this dissertation is to present several approaches to tackle thermal image SR within two techniques presented below. This chapter, focuses on two different generative super-resolution methods: a supervised method based on deep convolutional neural networks and an unsupervised method based on generative adversarial networks on thermal images.

4.2 Supervised method

One of the most relevant and first approach for supervised image enhancement has been presented in SRCNN [20]; the approach is based on a CNN, where the architecture is trained to learn how to get a high-resolution image from an image with a lower resolution. The authors explore the performance by using different color space representations; concluding that the best option is obtained by using the Y-channel from the YCbCr color space.

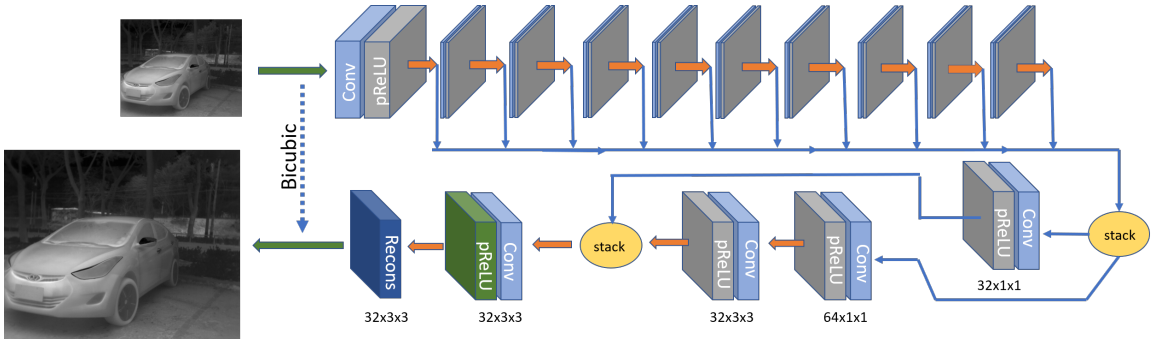


Fig. 4.1 Proposed supervised approach using CNN architecture (TISR-DCNN).

In most SR approaches, thermal images have not been considered during the training stage, although intended for thermal image enhancement. Most of them train the CNN-based techniques using images from the visible spectrum at different color space representations. On the contrary to all of them, in this chapter, thermal images are considered for training the proposed CNN architecture. This section presents a supervised CNN model designed for images of the thermal spectrum.

4.2.1 Proposed approach

This section presents a deep CNN architecture with a residual net and dense connections referred to as TISR-DCNN [96]. The architecture, presented in Fig. 4.1, has a part of the architecture dedicated to obtain the high-level features of the image, and another part, to perform the reconstitution of the image. All layers have dropouts and use parametric ReLU as an activator (preventing from learning a large negative bias term and getting better performance). Additionally, based on the work of [126], after concatenating all of the features, parallelized CNNs are used to reconstruct the image, and finally the generated image is estimated by typically adding these outputs with the bicubic interpolation to enhance the network's output.

For training this network, three datasets are considered. ThD1-101 [96] and T-138 [9] from thermal spectrum and BSD300 [71] from visible spectrum. Data augmentation process is performed to increase the number of training images.

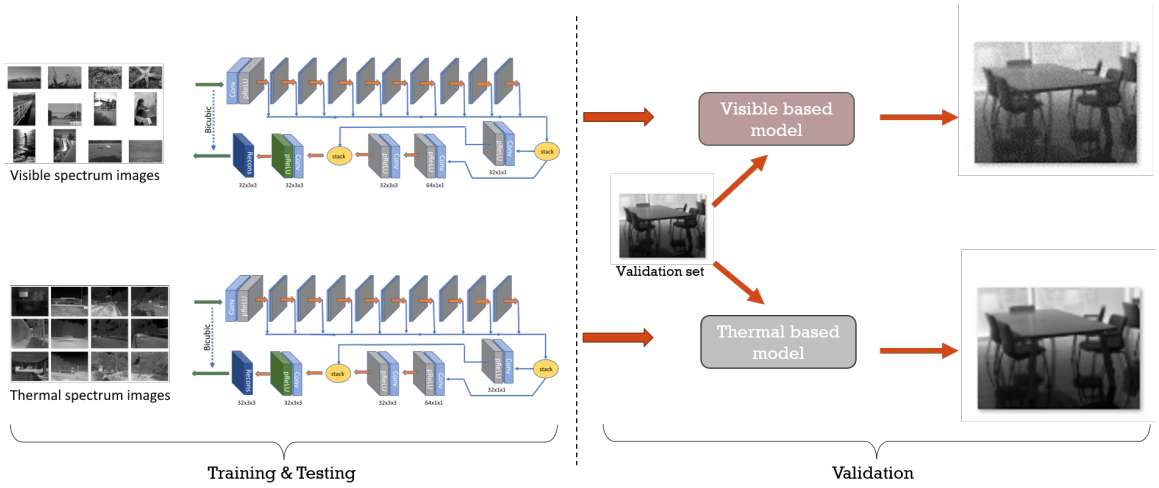


Fig. 4.2 Training process of TISR-DCNN network, which generate two strategies for the testing process.

This architecture is used to obtain the thermal image SR contrary to the state-of-the-art approaches, where most CNN-based architectures are trained with visible spectrum images and used with thermal images. This approach is trained with thermal images in order to obtain better results. The latter hypothesis is validated by training the networks twice, one with visible spectrum images and one with thermal spectrum images. This training process results in two models (see Fig. 4.2), which are finally validated with thermal images. More details are given in the following subsection.

4.2.2 Experimental result

The proposed architecture consists of a dense network, with the bicubic image interpolation generated to improve image details. For feature extraction, dropout and ReLU operations are used. A learning rate of 0.002 is applied to the model using MSE as a loss function to measure the difference between the ground truth and the output. The model uses Adam Optimizer, an adaptive learning rate method, which computes individual learning rates for different parameters. Its name is derived from adaptive moment estimation, and the reason it's called that is that Adam uses estimations of the first and second moments of the gradient to adapt the learning rate for each weight of the neural network. Each epoch is trained with a batch of 100000 patches for a total of 63 epochs.

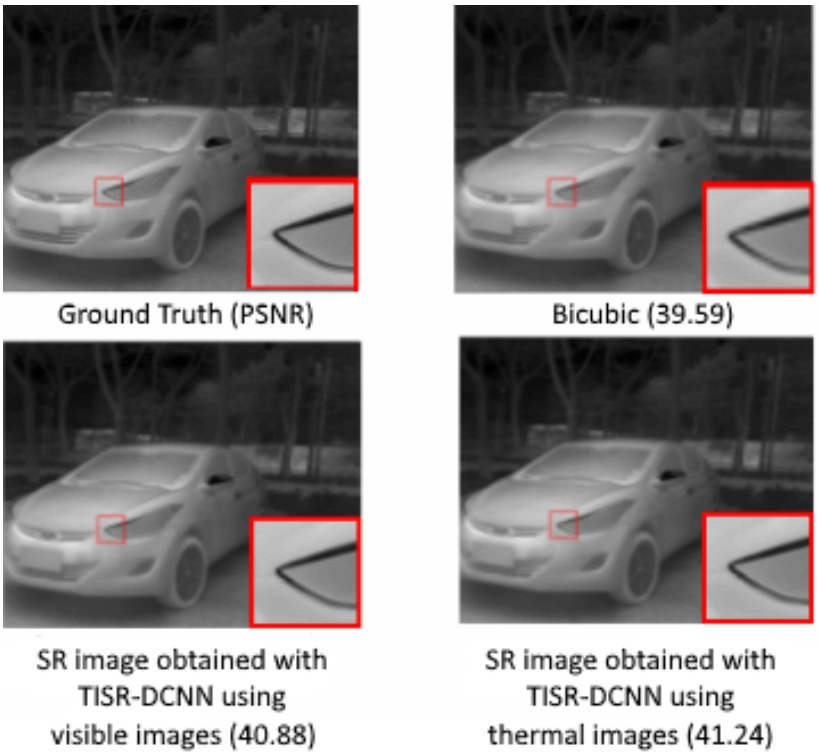


Fig. 4.3 Resulting images at $\times 2$ scale obtained with different strategies.

From ThD1-101 dataset, 77 images are for training, 18 images for validation and the remaining 6 thermal images for testing (referred to as Thermal6). In order to increase the number of training images, a data augmentation process is performed, rotating and flipping from top to bottom, from left to right all images. The quality and resolution of the images is maintained. The 77 training thermal images from Th-101 dataset and 138 from T-138 dataset, mixed with data augmentation, giving a total of 1720 thermal images. From the 300 visible images from BSD300 dataset, with data augmentation, gives a total of 2400 visible images.

As presented above, in order to evaluate the proposed approach, the same architecture is trained with the two different datasets; the 1720 thermal images are split up into 48×48 patches with 25 pixels overlapping adjacent patches, having a whole batch of 185760 patches. The 2400 visible images also are split up into 48×48 patches with 25 pixels overlapping, having a batch of 108000 patches (note that although there are more visible than thermal images, the number of thermal patches is more significant since thermal images have a larger resolution. The patches obtained above are used as ground truth, while the input patches are obtained by resizing them to half their original resolution. In this approach, there is no noise added to the input image.

The training is performed in Windows Server 2012, with a dual 2.50GHz CPU E5-2640, using one GPU K20m of 4GB. Each training consumes approximately 5GB of RAM and takes about 25 hours. This architecture is implemented using Tensorflow and Python. For a fair comparison, the two models are trained using the same infrastructure with the same number of batches per epoch, and hyperparameters are used.

Dataset	**Scale**	**Bicubic**	**Visible Model**	**Thermal Model**
Thermal6	×2	39.59	40.88	**41.24**
	×3	37.68	39.14	**39.62**
	×4	34.98	37.17	**37.85**
SET5	×2	33.64	**37.69**	37.46
	×3	30.37	**34.01**	33.74
	×4	28.41	**31.69**	31.25

Table 4.1 Average result of PSNR with proposed supervised architecture (TISR-DCNN).

As shown in Fig. 4.2, two strategies using the same network have been trained with different datasets; each trained network is validated with a set of six thermal images (Thermal6) and five RGB images (SET5), obtaining a visible and thermal model. Table 4.1 shows that with Thermal6, the trained thermal model shows a PSNR average value higher than the PSNR average value obtained with the visible trained model. Also, it indicates that SET5 got better PSNR values in the visible model than in the thermal model. A qualitative comparison can be appreciated in Fig. 4.3, where the SR images obtained with the two strategies and the images with the bicubic interpolation are depicted. Additionally, in this figure, the ground truth is presented (values in brackets correspond to the average PSNR presented in Table 4.1)

In this supervised approach, the usage of the proposed network has been considered to obtain thermal image SR. Two models have been obtained by training the same network with two different datasets to seek the best options when thermal images are considered. The experimental results indicate that the network model trained with the thermal image dataset is better than using the visible image dataset. The proposed method has the potential to increase the quality of thermal images, which is important for many applications such as surveillance, night vision, among others.

4.3 Unsupervised methods

Recently, unsupervised super-resolution approaches have been proposed to leverage unpaired images to overcome the limitation of having a pixel-wise registration without any assumption

on the degradation model. The Cycle Generative Adversarial Network (CycleGAN) [140], widely used for mapping features from one domain to another domain for image-to-image translation tasks in the absence of paired examples images, is used in the approaches proposed in this thesis. This framework is used to learn a mapping from the low-resolution to the high-resolution domain solving the SR problem. This is a recursive process where the mapping functions try to generate images with a similar distribution at each domain.

In the development of this thesis three different unsupervised approaches have been proposed, all mentioned in this section, which can be assumed as an evolution and an increment improvement from each other—published in [90, 91, 93].

4.3.1 Proposed approaches

The first unsupervised approach, referred to as TISR-US-1 [90], tackles the SR problem by using images from different cameras, which have been acquired at different resolutions, as mentioned in Chapter 3. The proposed approach is based on the usage of a CycleGAN, an unsupervised learning approach which is able to map information from one domain (low-resolution image) to another domain (high-resolution image). Figure 4.4 presents an illustration of the CycleGAN architecture proposed in this first approach. It consists of two generators (G_{L-H} and G_{H-L}) and two discriminators (D_H and D_L).

In the generators a ResNet with 6 residual blocks (ResNet-6), to avoid degradation of the optimization during the training process, are considered. Each residual block has (Conv -> InstaNorm -> Conv -> InstaNorm -> Relu), with skip connections. Regarding the discriminators, a PatchGAN-based architecture is considered. Each block in the discriminator has (Conv -> Conv -> InstaNorm -> Conv -> InstaNorm -> Conv -> LeakyReLU). The shape of the output after last layer is (batch_size, 30, 30, 1), each 30×30 of the output classifies a 70×70 portion of the input images. The discriminator receives two inputs, the target image (classified as a real image) and the generated image (which should be discriminated as a real or fake image by the discriminator).

The proposed architecture uses a combination of different loss functions: adversarial loss $\mathscr{L}_{Adversarial}$, cycle loss $\mathscr{L}_{Cycle}$, identity loss $\mathscr{L}_{Identity}$, and structural similarity loss $\mathscr{L}_{SSIM}$, which are detailed bellow.

The second approach [91], referred to as TISR-US-2, also uses a CycleGAN framework to tackle the SR problem. This time just by mapping information from the mid-resolution (MR) to the high-resolution (HR) domain (low-resolution domain is not considered). As shown in Fig. 4.5, the proposed approach consists of two generators (MR to HR and HR to MR), with their corresponding discriminators (DISC MR and DISC HR) that validate the generated images. As generators, a ResNet with 6 residual blocks (ResNet-6)

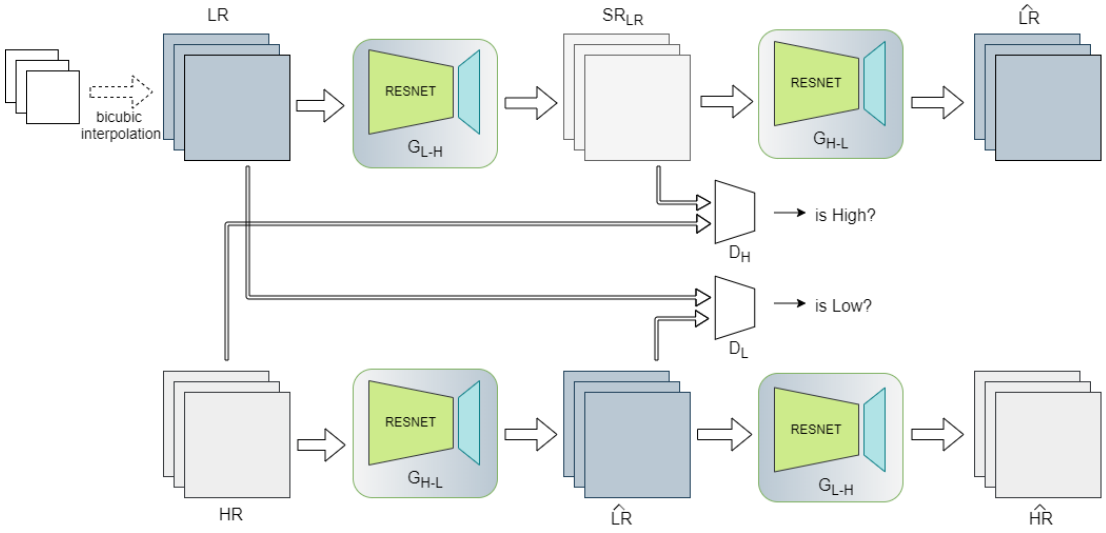


Fig. 4.4 First unsupervised proposed approach (TISR-US-1) illustration, based on a CycleGAN architecture; G_{L-H} and G_{H-L} represent generators from lower to higher and from higher to lower resolution, respectively. D_H and D_L represent the discriminator for each resolution.

is considered. It uses optimization to avoid degradation in the training phase. The residual blocks have convolutional layers, with instant normalization, ReLu, and skip connections. As discriminators, a patchGAN architecture is considered; the generated image and a non-paired GT image are used to validate if the output is real or not. In order to improve the results obtained with the first approach, a novel loss term based on edge information has been designed. It is based on the usage Sobel. This loss function helps in the training process to archive better results.

The same loss functions combination is used: i) adversarial loss $\mathscr{L}_{Adversarial}$, ii) cycle loss $\mathscr{L}_{Cycle}$, iii) identity loss $\mathscr{L}_{Identity}$, and iv) structural similarity loss $\mathscr{L}_{SSIM}$; additionally, a new loss function is proposed, Sobel loss $\mathscr{L}_{Sobel}$. This new loss consists in applying Sobel filter edge detector [57] to the input image, and the cycled generated image, getting the mean square difference between both images, helping to evaluate the contour consistency between the two images. Details on each of these loss terms are given below—Fig. 4.5 illustrates how these terms are computed.

The third and last unsupervised approach [93], referred to as TISR-US-3, accepted and published in Sensors Journal, is an extension of the first two proposed unsupervised approaches mentioned above. It uses the same Sobel cycle loss function. This time an Attention Module (AM) in the bottleneck in between the encoder and decoder of the generator in the GAN is proposed. Also, two datasets (Th3D-1021 and FLIR ADAS) are used, each

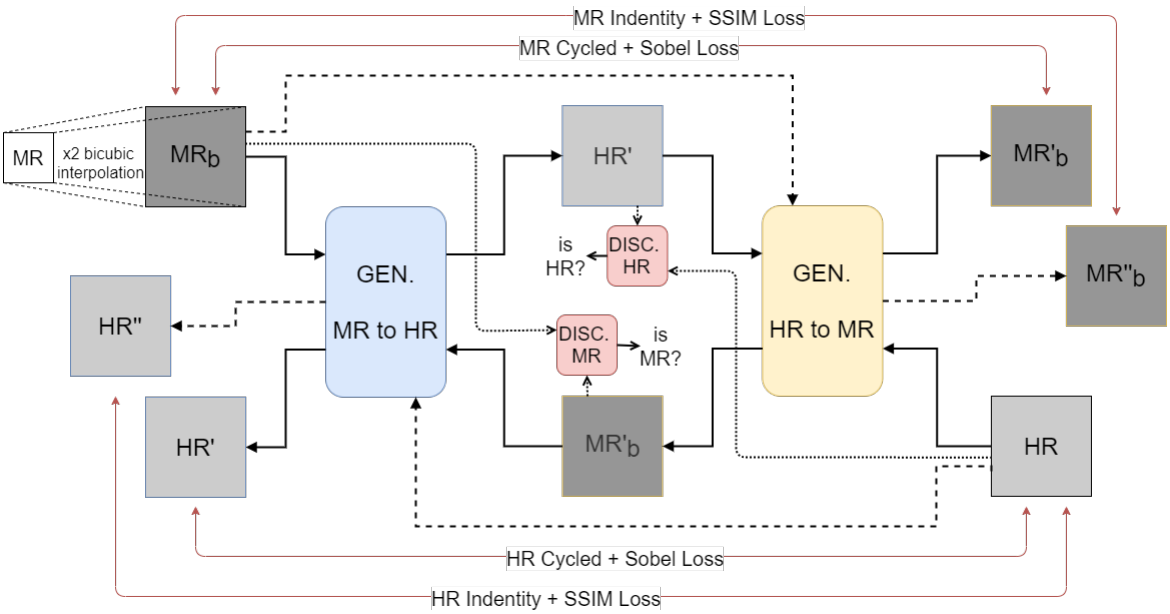


Fig. 4.5 Second unsupervised proposed approach (TISR-US-2) illustration, based on a CycleGAN architecture for MR to HR generator and for HR to MR; with cycled + sobel losses and identity + SSIM losses, and its respective discriminators.

one separately and also mixed together, for training the network. This unsupervised approach achieves results better than those obtained in the second evaluation of the PBVS-CVPR2021 challenge at the moment of the experimental results. This approach takes into consideration the gap between the generated and real HR images.

The proposed approach, shown in Fig. 4.6, consists of the same two generators, from the LR domain to the HR domain and vice versa. Each has its corresponding discriminator that validates the generated images. The generators are a ResNet with 6 residual blocks (ResNet-6). The residual blocks have convolutional layers, with instance normalization and ReLu activation with skip connections. Inspired in [134], an AM is added after the ResNet Encoder step (at the bottleneck of the generator), as shown in Fig. 4.7, which consists of the operation of three weight matrix obtained from a convolution operation of the last output layer in the encoder. A patchGAN architecture is also considered as a discriminator, and for validation, the same non-paired GT image and the generated image are used to validate if the output is real or not. The AM is a scaled dot-product as proposed in [116], and is computed as follows:

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \tag{4.1}$$

where Q, K, and V are the input matrices that contain the feature representation of the encoder, and d_k is a scaled-down factor. The scaling is done so that $softmax$ function's arguments do not become excessively large with a higher dimension.

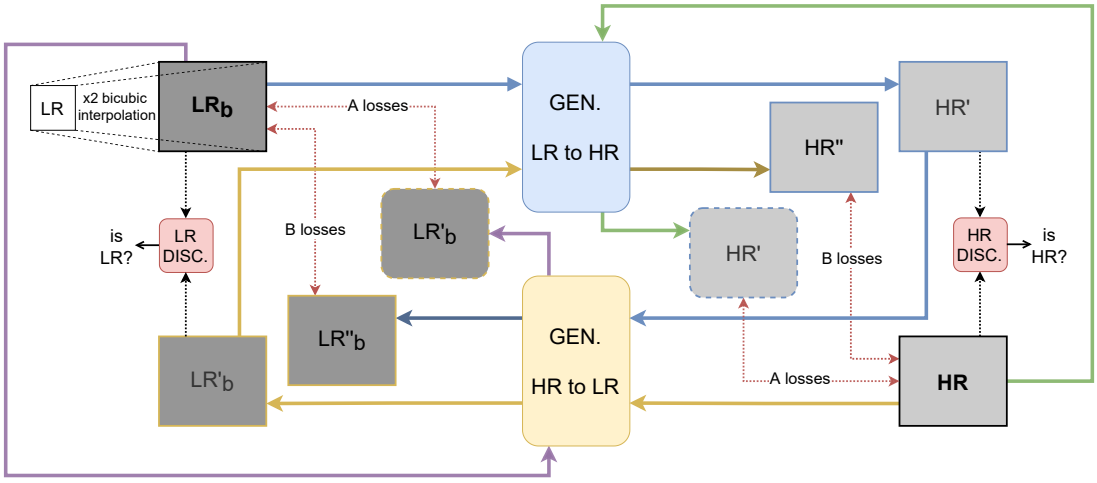


Fig. 4.6 Third unsupervised proposed approach (TISR-US-3) illustration, based on a CycleGAN architecture (for LR to HR and vice versa); *A losses* represents the Sobel and cycled losses, and *B losses* are SSIM and identity losses. Each cycle has its respective discriminators.

Following the first two unsupervised approaches mentioned above (TISR-US-1 and TISR-US-2), for this third approach (TISR-US-3) the combination of all different loss functions is used: *i*) adversarial loss $\mathscr{L}_{Adversarial}$, *ii*) cycle loss $\mathscr{L}_{Cycle}$, *iii*) identity loss $\mathscr{L}_{Identity}$, *iv*) structural similarity loss $\mathscr{L}_{SSIM}$; and Sobel loss $\mathscr{L}_{Sobel}$.

The loss functions mentioned above, which have been used for the three proposed approaches in this thesis are detailed next. The **adversarial loss** is designed to minimize the cross-entropy to improve the texture loss:

$$\mathscr{L}_{Adversarial} = -\sum_{i} logD(G_{L2H}(I_L), I_H), \tag{4.2}$$

where D is the discriminator, $G_{L2H}(I_L)$ is the generated image, I_L and I_H are the low and high-resolution images, respectively. The **cycled loss** ($\mathscr{L}_{Cycled}$) is used to determinate the consistency between input and cycled output; it is defined as:

$$\mathscr{L}_{Cycled} = \frac{1}{N}\sum_{i} ||G_{H2L}(G_{L2H}(I_L)) - I_L||, \tag{4.3}$$

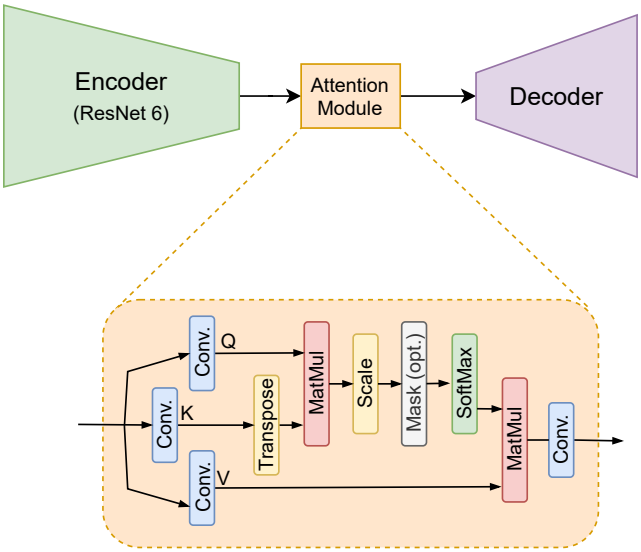


Fig. 4.7 Generator defined by: ResNet-6 as encoder, followed by the scaled dot-product attention module [116] and then the decoder.

where G_{L2H} and G_{H2L} are the generators that go from one domain to the other domain. The **identity loss** ($\mathscr{L}_{Identity}$) is used for maintaining the consistency between input and output; it is defined as:

$$\mathscr{L}_{Identity} = \frac{1}{N}\sum_{i} ||G_{H2L}(I_L) - I_L||, \tag{4.4}$$

where G is the generated image and I is the input image. The **structural similarity loss** ($\mathscr{L}_{SSIM}$) for a pixel P is defined as:

$$\mathscr{L}_{SSIM} = \frac{1}{NM}\sum_{p=1}^{P} 1 - SSIM(p), \tag{4.5}$$

where $SSIM(p)$ is the Structural Similarity Index (see [119] for more details) centered in pixel p of the patch (P). The **Sobel loss** ($\mathscr{L}_{Sobel}$), considered for the two last approaches, is used to determinate the edge consistency between input and cycled output; it is defined as:

$$\mathscr{L}_{Sobel} = \frac{1}{N}\sum_{i} ||Sobel(G_{H2L}(G_{L2H}(I_L))) - Sobel(I_L)||, \tag{4.6}$$

where G_{L2H} and G_{H2L} are the generators that go from one domain to the other domain, and Sobel gets the edges of each of the objects in the images. The **total loss function** ($\mathscr{L}_{total}$) used in this work is the weighted sum of the individual loss function terms:

$$\mathscr{L}_{total} = \lambda_1\mathscr{L}_{Adversarial} + \lambda_2\mathscr{L}_{Cycled} + \lambda_3\mathscr{L}_{Identity} + \lambda_4\mathscr{L}_{SSIM} + \lambda_5\mathscr{L}_{Sobel}, \tag{4.7}$$

where λ_i parameters for Adversarial, Cycled, and Identity losses are maintained as originally proposed CycleGAN. For SSIM and Sobel losses, λ_i are set empirically according to best results of the experiments; Cycled, and SSIM losses are sets with higher values.

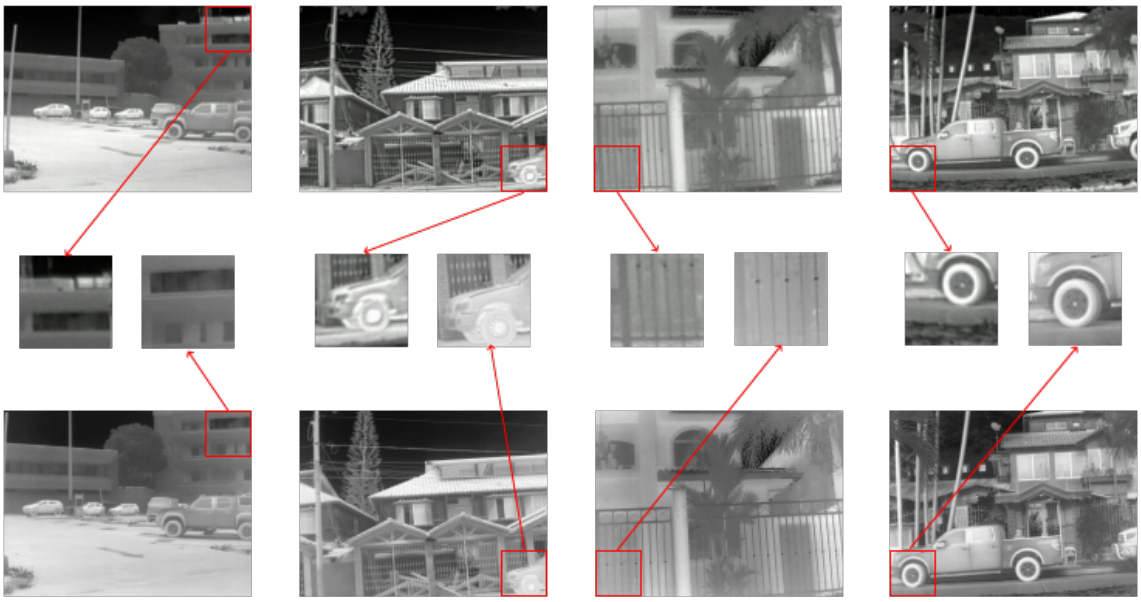


Fig. 4.8 Examples of thermal images acquired in [89]: (*top*) MR images from Axis Q2901-E (320×240), used in TISR-US-2 and TISR-US-3 approaches as LR images; (*bottom*) HR images from FC-6320 FLIR (640×480) [89]; (*middle*) enlargements to show the miss-registration between the images.

4.3.2 Experimental results

All these proposed unsupervised approaches have been trained with ThD3-1021 dataset mentioned in Chapter 3, which has images acquired with three different cameras at different resolutions; each resolution set has 951 images, and 50 images are left for validation. Just for TISR-US-2 and TISR-US-3 approaches, only mid-resolution images (referred to as LR inputs in these approaches) and high-resolution images are considered. It is worth noticing that the input images (LR and HR) are from different cameras and they are not pixel-wise registered, Figure 4.8 shows some illustrations of this dataset.

For TISR-US-2 and TISR-US-3 another thermal dataset is considered. It consists of a video sequence with 8862 thermal images from *FREE FLIR Thermal Dataset for Algorithm Training (FLIR-ADAS)*[1]; where just 985 images are selected (one out of nine images) to have a more variance scenario, Figure 4.9 shows some illustrations of this FLIR-ADAS

[1] https://www.flir.com/adasdataset/

Fig. 4.9 Examples of the FLIR-ADAS dataset.

dataset. Images from both datasets (ThD3-1021 and FLIR-ADAS) have a native resolution of 640×512. To exactly have ×2 size resolution on LR images, both datasets are centered cropped to 640×480 with the same format (8 bits in jpg format) but are acquired in different places and conditions.

The evaluation is the same as the PBVS-CVPR challenges (referred to as *Evaluation 2*), mentioned in Section 2.7.1; the quantitative evaluation of the presented approach is performed by means of the average PSNR and SSIM measures between the generated SR image and the semi-registered HR counterpart obtained from the other camera; this evaluation is illustrated in Fig. 2.13(b). Due to the camera baseline, the information in the images is not the same; hence, just a ROI of 80% of the image size, centered at each image, is considered.

The proposed architectures have been trained in a NVIDIA Titan X mounted in a workstation with 128GB of RAM. Python programming language and Tensorflow 2.0 with Keras library are used. Only ThD3-1021 and FLIR ADAS datasets are considered. No data-augmentation process has been applied to the given input data. CycleGAN transfer domain needs images at the same resolution; hence, the LR input images are up-sampled by bicubic interpolation and normalized in a [-1,1] range. The training process is done for 100 epochs without dropout (the model does not present overfitting). During the training, the input images are randomly selected in each epoch according to the batch size. The learning rate is set to 0.0002 for both generator and discriminator networks; epsilon=1e-05; exponential decay rate for the 1^{st} momentum, 0.5 for the discriminator, and 0.4 for the generator. The λ_i values that weigh each loss function used in all approaches are set as follows: $\mathscr{L}_{Cycled}$=10, $\mathscr{L}_{Identity}$=5, $\mathscr{L}_{SSIM}$=5 and $\mathscr{L}_{Sobel}$=10 in order to reach the best results, where the Cycled and Sobel losses have higher values for the importance in their corresponding loss functions.

Even though these images are from the thermal spectrum domain, they are represented like grayscale images, so PSNR and SSIM metrics are used for the evaluations and comparison

of the results. Quantitative results obtained with the first proposed CycleGAN architecture (TISR-US-1), for both evaluations, from LR to MR and from MR to HR, are shown in Table 4.2 and Table 4.3 respectively, together with the results obtained with the proposed supervised approach (TISR-DCNN) [96] and with the bicubic interpolation, which is used as a baseline. As can be appreciated, the proposed architecture achieves a better performance than both the previous publication and the bicubic interpolation.

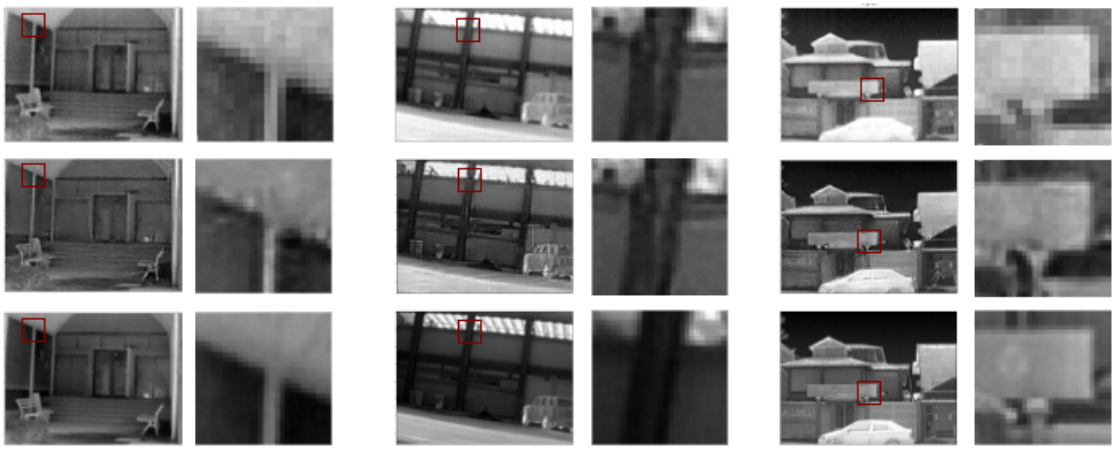


Fig. 4.10 SR results of TISR-US-1 approach on real-world LR images with a $\times 2$ scale factor—these illustrations correspond to the 80% centered area cropped from the images. $(top-row)$ Bicubic interpolation image, $(middle-row)$ Super-resolution results (SR_{LR}), $(bottom-row)$ Ground truth MR image.

Qualitative comparisons, for both evaluations, are depicted in Fig. 4.10 and Fig. 4.11. The results have shown that using this architecture it is possible to go from a lower resolution to a higher resolution representation, even though the network is trained with images from different cameras where there is not a perfect registration.

Approachs'	**PSNR**	**SSIM**
Bicubic Interpolation	16.46	0.6695
TISR-DCNN [96]	17.01	0.6704
TISR-US-1 [90]	**21.50**	**0.7218**

Table 4.2 Results from TISR-US-1 approach on LR set in a $\times 2$ scale factor, compared with its MR registered testing set.

For the second unsupervised approach (TISR-US-2), the quantitative results obtained for each training, and comparisons with previous work (TISR-US-1) and other approaches from the PBVS-CVPR2020 challenge are shown in Table 4.4 using PSNR and SSIM measures

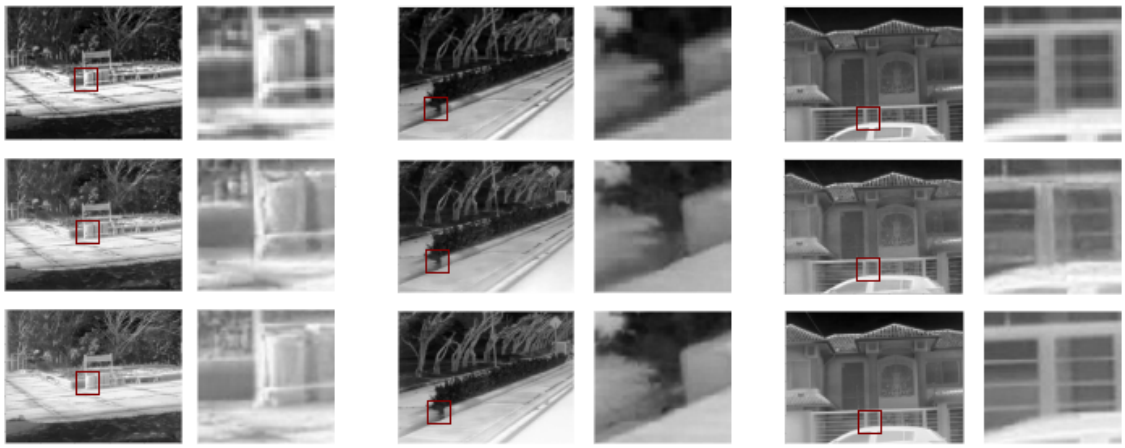


Fig. 4.11 SR results of TISR-US-1 on real-world MR images with a $\times 2$ scale factor—these illustrations correspond to the 80% centered area cropped from the images. $(top-row)$ Bicubic interpolation image, $(middle-row)$ Super-resolution results (SR_{MR}), $(bottom-row)$ Ground truth HR image.

Approachs'	PSNR	SSIM
Bicubic Interpolation	20.17	0.7553
TISR-DCNN [96]	20.24	0.7492
TISR-US-1 [90]	**22.42**	**0.7989**

Table 4.3 Results from TISR-US-1 approach on MR set in a $\times 2$ scale factor, compared with its HR registered testing set.

comparison. The best results are highlighted in bold, and the second-best result is underlined. As can be appreciated, this second approach achieves better results than other works. Using just one dataset (Th3D-1021) gets seven-tenths better PSNR results than using both datasets together (Th3D-1021 and FLIR ADAS). SSIM measure gets higher results using both datasets but just by one-thousandth. These results show that using just the ThD3-1021 dataset, the proposed approach archives better results, meaning that this dataset is varied enough to train a network and that it is possible to do a single thermal image super-resolution between two different domains using images acquired with different camera resolutions and without registration.

Regarding the quality of the obtained results, Fig. 4.12 shows the worst and best super-resolution results from the testing set. The worst resulting image gets 20.11/0.6464 PSNR and SSIM measures, respectively; it should be mentioned that although it is the worst resulting image from the whole validation set, it is considerably better than the results obtained with a bicubic interpolation: 17.36/0.6193 PSNR and SSIM, respectively. In the case

Approachs'	**PSNR**	**SSIM**
Bicubic Interpolation	20.24	0.7515
TISR-US-1 [90]	22.42	0.7989
MLVC-Lab* [89]	20.02	0.7452
COUGER AI* [89]	20.36	0.7595
TISR-US-2 (a) [91]	**22.98**	<u>0.8032</u>
TISR-US-2 (b) [91]	<u>22.27</u>	**0.8045**

Table 4.4 Results from TISR-US-2 approach. $(^*)$ Best approaches at the PBVS-CVPR2020 challenge. For TISR-US-2: (a) uses just ThD3-1021 dataset; (b) uses ThD3-1021 and FLIR-ADAS datasets. Bold and underline values correspond to the first and second best results, respectively.

of the best resulting image, 26.06/0.8651 PSNR and SSIM measures respectively are obtained; in this case, the bicubic interpolation reaches 19.41/0.8021 PSNR and SSIM, respectively. In conclusion, it could be stated that the most challenging scenarios are those with objects at different depths and complex textures. On the contrary, it can be appreciated that scenes with planar surfaces are more simple to obtain their corresponding super-resolution representation.

For the third approach (TISR-US-3), the proposed architecture has been trained four times, once with just ThD3-1021 dataset and once with ThD3-1021 and FLIR-ADAS datasets together, then once more for each but with and without the AM, referred to as TISR-US-3 (a, b, c, d) as an ablation study. FLIR-ADAS dataset are frame images from a video sequence, and for having more variability and the images of a balanced number than ThD3-1021 dataset, one out of nine frames are selected. As mentioned above, the testing is done with the same set of images used in the PBVS-CVPR2021 challenge [95] to compare this third approach with the most recent results in the state-of-the-art literature at the time.

Quantitative results of each training are compared with the best three approaches from the PBVS-CVPR2021 challenge. Table 4.5 depicts PSNR and SSIM measures for the comparisons. The best result is highlighted in bold, and the second-best result is underlined. Qualitative results are depicted in Fig. 4.13. As can be appreciated, the approach that reaches the best result in the PSNR metric (TISR-US-3 (c)) uses the AM; it gets the second-best result in SSIM. This approach has been trained with just ThD3-1021 dataset. The approach with attention module and trained with ThD3-1021 and FLIR-ADAS datasets (TISR-US-3 (d)) preserves the structural information (SSIM) better than other methods. The usage of just ThD3-1021 dataset shows a good performance; as TISR-US-2, this show that ThD3-1021

Fig. 4.12 Examples quality results from TISR-SR-2 approach. (top) from left to right, MR image (input), HR image, worst generated result. $(bottom)$ best generated result.

dataset has a large enough variability to train a network to perform a single thermal image super-resolution between two different domains.

Regarding the first approach (TISR-SR-1), the last approach (TISR-SR-3) shows betters results by adding and adjusting loss functions' variations and, better yet, with the attention module. The use of both datasets (without the attention module) increases the SSIM measure (reaching the best result in this measure), but with the attention module using just the ThD3-1021 dataset, overcoming the best result in the PSNR measure and better SSIM than previous work. With some changes from previous work, quantitative measures overcome previous results and the best approaches from PBVS-CVPR2021 challenge (Evaluation 2).

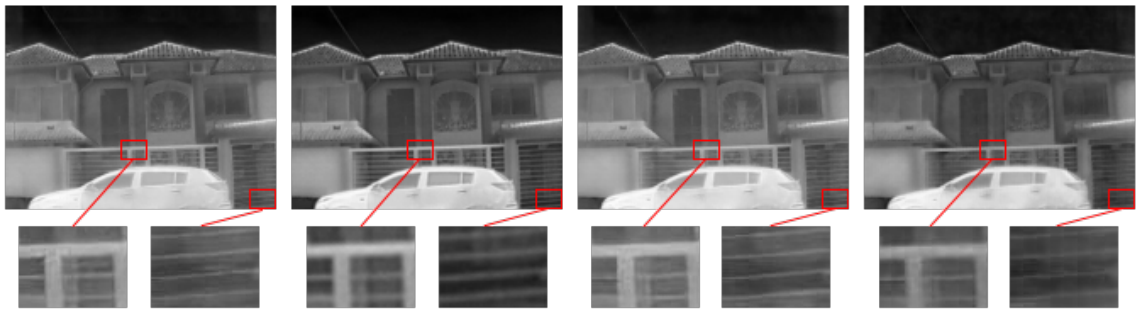


Fig. 4.13 Visual comparison of SR results obtained using TISR-US-3 (a, b, c and d) variations, respectively.

The main differences between the three unsupervised approaches mentioned above that make use of CycleGAN architecture are the usage of different datasets and loss functions

Approachs'	PSNR	SSIM
TISR-SR-1 [90]	22.42	0.7989
NPU-MPI-LAB* [95]	21.96	0.7618
SVNIT_NTNU-2* [95]	21.44	0.7758
ULB-LISA* [95]	22.32	0.7899
TISR-US-3 (a)	<u>22.98</u>	0.7991
TISR-US-3 (b)	21.93	<u>0.8117</u>
TISR-US-3 (c)	**23.19**	0.8023
TISR-US-3 (d)	21.23	**0.8167**

Table 4.5 Results from TISR-US-3 approach [93]. $(^*)$ Best approaches at the PBVS-CVPR2021 challenge (Evaluation 2). For TISR-US-3, (a) is trained just with ThD3-1021 dataset and without AM; (b) is trained using ThD3-1021 and FLIR-ADAS datasets without AM. (c) is trained with just ThD3-1021 dataset with AM. (d) is trained using ThD3-1021 and FLIR-ADAS datasets with AM. Bold and underline values correspond to the first and second best results, respectively.

for the two last approaches, and the last approach makes use of an attention module in the bottleneck of the generator. According to the results, the last approach (TISR-US-3) presents higher results an improvement concerning previous works.

4.4 Summary

Single image super-resolution is a well-studied problem in the computer vision community and has shown great success in the past few years, where most of them are focused on the visible spectrum. However, the problem remains largely unexplored in the thermal infrared (TIR) spectrum. This chapter reviews the supervised and unsupervised deep learning-based methods for single image super-resolution in the TIR spectrum. The methods are evaluated and compared using the two public datasets mentioned in Chapter 3. The results show that the proposed methods can increase the resolution of TIR images by a large margin concerning state-of-the-art methods.

Chapter 5

Multi-Image Super-Resolution

This chapter presents a deep learning-based method for multi-image super-resolution of thermal images [92]. This method takes a set of low-resolution images as input and generates a single high-resolution image as output. The main idea behind the proposed method is to use a deep convolutional neural network to learn the mapping between the low-resolution (LR) images and the HR image. This mapping is then used to generate the high-resolution (HR) image from the set of LR images. For this, a synthetic dataset is generates, shifting and downsampling the HR images to create the LR set of images.

5.1 Introduction

Image Super-resolution (SR) is the problem of reconstructing a HR image from one or more LR images of the same scene. HR images provide supplementary information that makes the problem widely studied with several practical applications. Most of the super-resolution community has focused on the single image super-resolution (SISR) problem, which estimates the HR image just from a single LR input.

On the contrary, multi-image super-resolution (MISR) reconstructs a HR image using multiple LR images of the same scenes. It is possible to apply this concept to thermal images, which would require several LR images from the same scene as input. In the current work, a dataset with synthesized images is generated due to the lack of a benchmark of multi-thermal image datasets. This dataset contains several LR images of a given scene by down-sampling, adding noise and blur, and randomly shifting (X and Y coordinates) to simulate being captured by a burst of input images. The main idea of this approach is to combine information from multiple frames to obtain a more detailed reconstruction of the HR image. As far as we know, a few approaches in the literature use a multi-image scheme to generate HR thermal images.

5.2 Proposed approach

The current approach, referred to as TISR-MI, takes as an input a sequence of multiple noisy, RAW, LR thermal images and combines their features to generate a SR image. As shown in Fig. 5.1, it consists of two main paths, a 2D Attention Path and a 3D Attention Path, where both paths use Residual Attention Blocks, which are the core of the model that focuses on the images' High-Frequency (HF) features. For SR generation, HF features have more valuable information. The up-sampling operation is done at the end of each path for better computational performance. In the end, the results from both paths are added to generate the SR image.

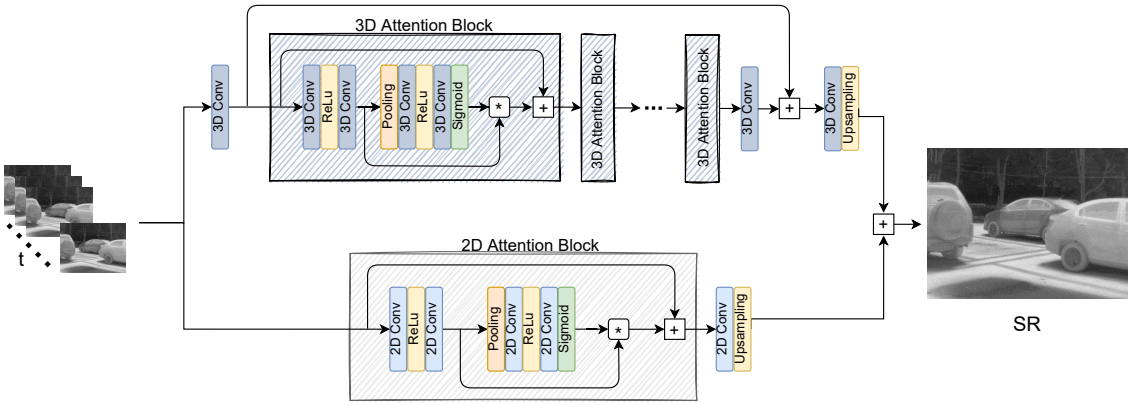


Fig. 5.1 Proposed multi-image thermal super-resolution architecture using a 2D and 3D attention blocks.

The 2D Attention Path allows the network to generate a simple super-resolution solution for up-sampling a set of multi-LR images. This attention path consists of: 2DConv -> ReLU -> 2DConv -> GlobalPoll -> 2DConv -> ReLU -> 2DConv -> Sigmoid, with respective skip connection, followed by 2DConv -> UpSampling.

The 3D Attention Path uses 3D convolutions residual-based blocks to extract spatial correlations from the pool of inputs LR images. This path is the main branch of the approach. First, a 3D convolution layer is applied to extract shallow features from the LR input images. After this, a cascade of N concatenations, 3D Attention Blocks are applied for higher extractions of features exploiting the spatial and local, and non-local correlations. A long skip connection is used for redundant low-frequency signals and several short skip connections inside each block. Finally, the up-sample operation is done. In summary, this attention path consists of 12 times: 3DConv -> ReLU -> 3DConv -> GlobalPoll -> 3DConv -> ReLU -> 3DConv -> Sigmoid, with respective skip connection, and a long skip connection, followed by 3DConv -> UpSampling.

The multi-image super-resolution approach can be summarized as follow:

$$SR = U\left(2D_{attB}(LR_t)\right) + U\left([3D_{attB}(LR_t)]^N\right) \tag{5.1}$$

where U represents the up-sampling operation, 2D and 3D are each attention block paths, and N is the number of times the 3D path repeats. LR_t represents the multi-image set from the same scene, and SR represents the generated super-resolution image.

5.3 Experimental results

The results of the MISR proposed approach, training it on a synthesized dataset and comparing its performance with state-of-the-art SISR algorithms are presented in this section.

SR reconstruction is highly dependent on the degradation model. Several factors such as relative motion (handshake), atmospheric variation (haze), optical blurring, and preprocessing are used to generate a simulated burst of multi-LR thermal images. The thermal dataset used to evaluate the proposed model is ThD3-1021; it consists of 1021 thermal images (951 for training, 50 validating, and 20 for testing). Assuming all thermal LR images are generated under the same condition, the degradation model can be formulated as:

$$Y_t = (X + G_t) * S_t * D_t; t = 1, 2, ..., T, \tag{5.2}$$

where X, Y_t represent the t^{th} HR image and LR image respectively. G_t is the additional Gauss noise to Y_t. S_t and D_t represents the random u and v shift and downsampled factor by 4, respectively; ten LR images (T = 10) per each given image are generated. When the random shift is done, reflect padding is performed to fill the gap of the shift. The degradation process is illustrated in Fig. 5.2. Random $Gauss - noise$ with a value of 2 std. Random $left - right$ shift ± 4, $up - down$ shift $\pm$ 3, and $bicubic - downsampled$ method. No rotation is applied in this degradation method.

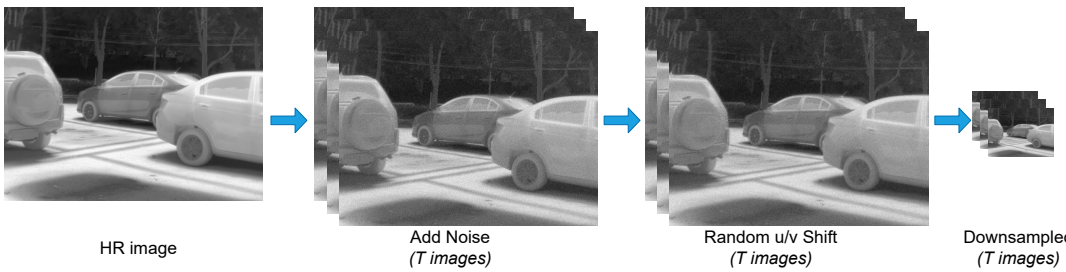


Fig. 5.2 Illustration of the model used as degradation prepossessing and synthesized data set generation.

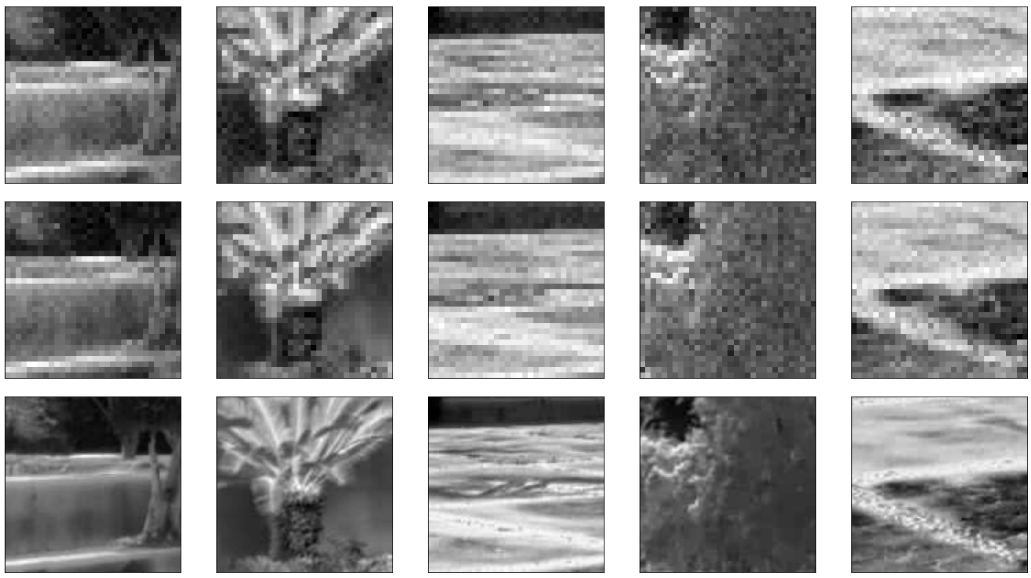


Fig. 5.3 Examples of the patch image registration process. $(top-rows)$ represent different generated LR patches—synthesized images. $(bottom-row)$ show the corresponding HR image patch.

To complete the synthesized multi-image, each T-generated image is registered using the first image as a reference, which has no shift. The registration is done using an efficient sub-pixel image translation by cross-correlation to have real simulated sub-pixel shifts with respect to each other, as it would be generated due to, e.g., camera motion, providing different LR samplings of the underlying scene, registered patch examples are depicted in Fig. 5.3. Reflect padding is used to complete the dismissed pixels. The synthesized dataset is saved in npy files to be loaded during the training process. The data format of the images is in uinit8, and each image is normalized between [-1,1] at the beginning, and after passing the network, they are denormalized. No extra data augmentation is used.

5.3.1 Training

In all convolutional layers, on both paths of the network, 32 filters with a kernel size of 3×3 are set. The reduction factor in the attention blocks is set to 8. The number of times that 3D Attention Block is repeated has been set to 12 (lower value causes a loss of performance, a higher value increases the number of parameters unnecessarily). In total, the network has less than 750K parameters.

For training, patches of 32×32 pixels, with an overlap of 22 pixels, are extracted from each LR image, giving more than 23K patches for training and 1.2K for validation. An initial learning rate of 0.0005 and Adam loss function optimization are used. To learn the end-to-end

Approachs'	PSNR	SSIM
SVNIT_NTNU-1	<u>30.70</u>	**0.9290**
SVNIT_NTNU-2	30.69	<u>0.9288</u>
SVNIT_NTNU-3	30.59	0.9254
ISESL-CSIO	30.39	0.8992
CVS	29.21	0.9032
TISR-MI	**32.99**	0.9236

Table 5.1 Results from TISR/MI approach, and state-of-the-art SISR approaches from PBVS 2021 challenge [95].

mapping process, L_1 and SSIM losses are considered by minimizing their values between the generated and the ground truth images. The proposed architecture has been trained in a NVIDIA Titan X mounted in a workstation with 128GB of RAM. Python programming language and Tensorflow 2.0 library are used. The model is trained for 50 epochs, taking less than 24 hours.

5.3.2 Results

The standard fidelity-based metrics PSNR and SSIM measures are used for testing and validating the proposed model, which consists in evaluating the SR generated from the multi-noisy down-sampled image with the corresponding HR image, as shown below:

$$R = \frac{1}{N}\sum_{1}^{N} eval\left(HR, SR(LR_t)\right), \tag{5.3}$$

where *eval* is PSNR and SSIM measures metrics separately calculated, SR is the super-resolution generated image from the t multi-image LR noise inputs, and HR represents the corresponding GT image. N is the number of validation images.

The metrics mentioned above to evaluate the results are: i) Peak Signal-to-Noise Ratio (PSNR), which is commonly used to measure the reconstruction quality of lossy transformations; and ii) Structural Similarity Index Metric (SSIM) [119], which is based on the independent comparisons of luminance, contrast, and structure. Due to thermal images being represented in grayscale, these metrics can also be used.

Quantitative results obtained with this proposed architecture are shown in Table 5.1, together with the SISR results of the state-of-the-art approaches from PBVS 2021 challenge [95]. As it can be appreciated, the proposed architecture achieves a better performance in PSNR with 32.99dB and is highly good on SSIM metrics (just 0.0054 below the best results, SVNIT_NTNU team achieves slightly better results). The SVNIT_NTNU-1 team uses an

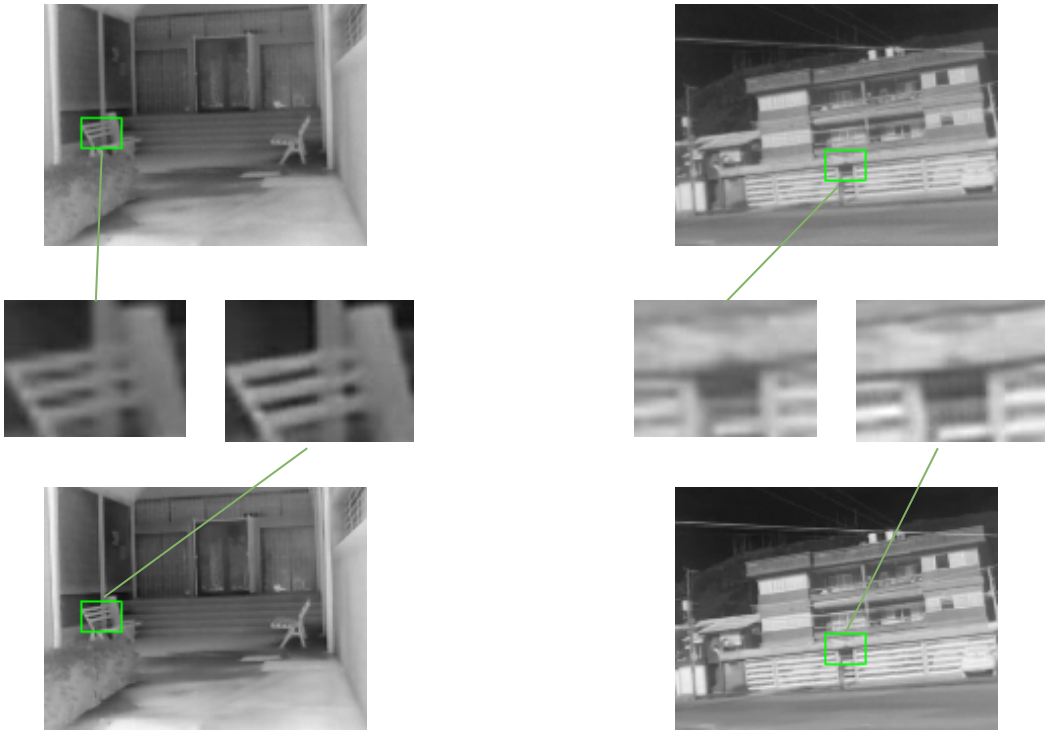


Fig. 5.4 SR results with a $\times 4$ scale factor: $(top-row)$ results from bicubic interpolation; $(bottom-row)$ results from the proposed approach.

effective design of ResBlock, that preserves the HF details with fewer parameters and uses channel attention modules; using an exponential linear unit (ELU) activation function to improve learning performance at each layer in an efficient manner. The SVNIT_NTNU-2 team uses a cascade of convolution with layer attention, including Residual Blocks, using a self-assemble technique to generate the SR result. Finally, the SVNIT_NTNU-3 team proposes several residual groups to learn complex and rich features from the LR observation, using subpixel convolutions in the up-sampling block with local and long skip connections.

Qualitative comparisons between bicubic interpolations and results from the proposed approach are depicted in Fig. 5.4. Enlarged patches are provided for a closed inspection showing that the obtained results are sharper and less noisy than bicubic interpolation. This comparison shows that using this architecture to go from a multi-LR to a HR image on the thermal spectrum is possible, even though the network is trained with synthesized images.

5.4 Summary

A multi-image super-resolution architecture for thermal images exploits recent deep learning advancements. In this case, two attention paths are proposed, a 2D and a 3D attentions

block mechanisms, that are used to train the network to perform SR at a $\times 4$ scale. To train the proposed architecture, synthesized RAW burst noisy LR images are generated. As loss functions, L_1 and SSIM are considered. At the moment of the experiments, results obtained with the proposed MISR approach reach the state-of-the-art SISR approaches presented in the PBVS-CVPR2021 challenge when SSIM is considered; on the contrary, when PSNR is considered, results from the proposed approach considerably overcome results from the state-of-the-art approaches.

Chapter 6

Conclusions and Future Work

This dissertation has made several significant contributions, summarized in this final chapter. Additionally, future work and possible research lines identified during the development of this dissertation are presented.

6.1 Conclusions

This dissertation presents works focused on the processing of low-resolution thermal images acquired by cheap thermal cameras. These cameras can capture the long-wavelength infrared band (temperature of an object) in a single shot but with a low spatial resolution. Most of the images acquired with these thermal cameras are in low-resolution because of the use of a low-cost camera sensor. To face this problem, the following contributions have been proposed.

One contribution is the creation of three datasets (ThD1-101, ThD2-200 and Th3D-1021) with high-resolution thermal images to be used for training and evaluation deep learning based approaches. These datasets are composed of images acquired in different scenarios and with different objects. Other contributions are the design of deep learning-based architectures for single image super-resolution (SISR) and multi-image super-resolution (MISR). For SISR, two methods are considered, supervised (TISR-DCNN) and unsupervised (TISR-US-1, TISR-US-2, TISR-US-3) proposed approaches. For MISR, the use of multi low-resolution input images is considered and TISR-MI network is proposed. All these architectures are based on the use of a deep convolutional neural network, using several features, such as skip connection, generation modules, attention modules, loss's function, among others. The proposed architectures are able to achieve state-of-the-art results in the super-resolution of thermal images. Finally, even it does not represent a scientific contribution to this thesis, thermal images super-resolution challenges are presented (PBVS-CVPR2020,

PBVS-CVPR2021, PBVS-CVPR2022), which I have been in charge of the organization, showing the evolution of the state-of-the-art.

List of Contributions:
This dissertation has led to the following publications, in chronological order:

- **Rivadeneira, R. E.**, Suárez, P. L., Sappa, A. D., & Vintimilla, B. X. (2019, August). Thermal image superresolution through deep convolutional neural network. In International conference on image analysis and recognition (pp. 417-426). Springer, Cham [96].

- **Rivadeneira, R. E.**, Sappa, A. D., & Vintimilla, B. X. (2020, February). Thermal Image Super-resolution: A Novel Architecture and Dataset. In VISIGRAPP (4: VISAPP) (pp. 111-119) [90].

- **Rivadeneira, R. E.**, Sappa, A. D., & Vintimilla, B. X. (2020, February). Thermal Image Super-Resolution: A Novel Unsupervised Approach. In International Joint Conference on Computer Vision, Imaging and Computer Graphics (pp. 495-506). Springer, Cham. **Chapter Book** [91].

- **Rivadeneira, R. E.**, Sappa, A. D., Vintimilla, B. X., Almasri, F., ... & Debeir, O et al. (2020, June). Thermal Image Super-Resolution Challenge - PBVS 2020. In 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (pp. 432-439) [89].

- **Rivadeneira, R. E.**, Sappa, A. D., Vintimilla, B. X., Nathan, S., Kansal, P., Mehri, A., ... & Beksi, W. J. (2021). Thermal image Super-Resolution Challenge - PBVS 2021. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 4359-4367) [95].

- **Rivadeneira, R. E.**, Sappa, A. D., Vintimilla, B. X., Kim, J., Kim, D., ... & Jiang, J. (2022). Thermal Image Super-Resolution Challenge Results - PBVS 2022. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 418-426) [94].

- **Rivadeneira, R. E.**, Sappa, A. D., & Vintimilla, B. X. (2022). Multi-Image Super-Resolution for Thermal Images. In VISIGRAPP (4: VISAPP) (pp. 635-642) [92].

- **Rivadeneira, R. E.**, Sappa, A. D., Vintimilla, B. X., & Hammoud, R. (2022). A Novel Domain Transfer-Based Approach for Unsupervised Thermal Image Super-Resolution. Sensors, 22 (6), 2254. **Journal** [93].

6.2 Future work

Regarding thermal image super-resolution, despite of the recent advances, there is still some work to be done. The most important one is to have more and more high-quality datasets to train and validate different models. The other one is to develop unsupervised methods for this problem. Several possible further research lines could be pursued in relation to the algorithms proposed in this dissertation in order to increase their performance. Some of the most direct areas of improvement are describe below:

- Improve the unsupervised thermal image super-resolution by designing new loss functions such as the content loss or the style loss, could also help to increase the performance of the proposed models.
- The use of GANs for data augmentation to increase the amount of data available for training the proposed models.
- Use of transfer learning of models pre-trained on other datasets could help to improve the performance of the proposed models, especially in the case of the unsupervised approach.
- Guided super-resolution to increase the performance of the unsupervised methods; the use of a guidance image (e.g. a HR visible spectrum image) could be explored.
- Implementation of new metrics for the evaluation of the obtained results.
- Evaluate the proposed methods in some specific application.

References

[1] Arbel, P., Maire, M., Fowlkes, C., and Malik, J. (2011). Contour detection and hierarchical image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell*, 33(5):898–916.

[2] Bevilacqua, M., Roumy, A., Guillemot, C., and Alberi-Morel, M. L. (2012). Low-complexity single-image super-resolution based on nonnegative neighbor embedding.

[3] Cascarano, P., Corsini, F., Gandolfi, S., Piccolomini, E. L., Mandanici, E., Tavasci, L., and Zama, F. (2020). Super-resolution of thermal images using an automatic total variation based method. *Remote Sensing*, 12(10):1642.

[4] Catmull, E. and Rom, R. (1974). A class of local interpolating splines. In *Computer aided geometric design*, pages 317–326. Elsevier.

[5] Chang, H., Lu, J., Yu, F., and Finkelstein, A. (2018). Pairedcyclegan: Asymmetric style transfer for applying and removing makeup. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 40–48.

[6] Chang, H., Yeung, D.-Y., and Xiong, Y. (2004). Super-resolution through neighbor embedding. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, volume 1, pages I–I. IEEE.

[7] Chen, Y.-S., Wang, Y.-C., Kao, M.-H., and Chuang, Y.-Y. (2018). Deep photo enhancer: Unpaired learning for image enhancement from photographs with gans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6306–6314.

[8] Cho, Y., Bianchi-Berthouze, N., Marquardt, N., and Julier, S. J. (2018). Deep thermal imaging: Proximate material type recognition in the wild through deep learning of spatial surface temperature patterns. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, page 2. ACM.

[9] Choi, Y., Kim, N., Hwang, S., and Kweon, I. S. (2016). Thermal image enhancement using convolutional neural network. In *Intelligent Robots and Systems (IROS), 2016 IEEE/RSJ International Conference on*, pages 223–230. IEEE.

[10] Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258.

[11] Cilulko, J., Janiszewski, P., Bogdaszewski, M., and Szczygielska, E. (2013). Infrared thermal imaging in studies of wild animals. *European Journal of Wildlife Research*, 59(1):17–23.

[12] Davis, J. W. and Keck, M. A. (2005). A two-stage template approach to person detection in thermal imagery. In *2005 Seventh IEEE Workshops on Applications of Computer Vision (WACV/MOTION'05)-Volume 1*, volume 1, pages 364–369. IEEE.

[13] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

[14] Deng, L., Yu, D., et al. (2014). Deep learning: methods and applications. *Foundations and trends® in signal processing*, 7(3–4):197–387.

[15] Deudon, M., Kalaitzis, A., Goytom, I., Arefin, M. R., Lin, Z., Sankaran, K., Michalski, V., Kahou, S. E., Cornebise, J., and Bengio, Y. (2020). Highres-net: Recursive fusion for multi-frame super-resolution of satellite imagery. *arXiv preprint arXiv:2002.06460*.

[16] Dijk, J., Schutte, K., van Eekeren, A. W., and Bijl, P. (2012). Measuring the performance of super-resolution reconstruction algorithms. In *Infrared Imaging Systems: Design, Analysis, Modeling, and Testing XXIII*, volume 8355, pages 389–400. SPIE.

[17] Ding, M., Zhang, X., Chen, W.-H., Wei, L., and Cao, Y.-F. (2019). Thermal infrared pedestrian tracking via fusion of features in driving assistance system of intelligent vehicles. *Proceedings of the Institution of Mechanical Engineers, Part G: Journal of Aerospace Engineering*, 233(16):6089–6103.

[18] Dong, C., Loy, C. C., He, K., and Tang, X. (2014). Learning a deep convolutional network for image super-resolution. In *European conference on computer vision*, pages 184–199. Springer.

[19] Dong, C., Loy, C. C., He, K., and Tang, X. (2015). Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(2):295–307.

[20] Dong, C., Loy, C. C., He, K., and Tang, X. (2016a). Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(2):295–307.

[21] Dong, C., Loy, C. C., and Tang, X. (2016b). Accelerating the super-resolution convolutional neural network. In *European conference on computer vision*, pages 391–407. Springer.

[22] Duchon, C. E. (1979). Lanczos filtering in one and two dimensions. *Journal of applied meteorology*, 18(8):1016–1022.

[23] Elad, M. and Hel-Or, Y. (2001). A fast super-resolution reconstruction algorithm for pure translational motion and common space-invariant blur. *IEEE Transactions on image Processing*, 10(8):1187–1193.

[24] Everingham, M., Eslami, S., Van Gool, L., Williams, C. K., Winn, J., and Zisserman, A. (2015). The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111(1):98–136.

[25] Farsiu, S., Robinson, M. D., Elad, M., and Milanfar, P. (2004). Fast and robust multiframe super resolution. *IEEE transactions on image processing*, 13(10):1327–1344.

[26] Fukushima, K. and Miyake, S. (1982). Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition. In *Competition and cooperation in neural nets*, pages 267–285. Springer.

[27] Gade, R. and Moeslund, T. B. (2014a). Thermal cameras and applications: A survey. *Machine Vision and Applications*, 81:89–96.

[28] Gade, R. and Moeslund, T. B. (2014b). Thermal cameras and applications: a survey. *Machine vision and applications*, 25(1):245–262.

[29] Girshick, R., Donahue, J., Darrell, T., and Malik, J. (2015). Region-based convolutional networks for accurate object detection and segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 38(1):142–158.

[30] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680.

[31] Gowen, A., Tiwari, B., Cullen, P., McDonnell, K., and O'Donnell, C. (2010). Applications of thermal imaging in food quality and safety assessment. *Trends in food science & technology*, 21(4):190–200.

[32] Grinzato, E., Bison, P., and Marinetti, S. (2002). Monitoring of ancient buildings by the thermal method. *Journal of Cultural Heritage*, 3(1):21–29.

[33] Haider, A., Shaukat, F., and Mir, J. (2021). Human detection in aerial thermal imaging using a fully convolutional regression network. *Infrared Physics & Technology*, 116:103796.

[34] Han, J., Yang, Y., Zhou, C., Xu, C., and Shi, B. (2021). Evintsr-net: Event guided multiple latent frames reconstruction and super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4882–4891.

[35] Han, X.-H., Zheng, Y., and Chen, Y.-W. (2019). Multi-level and multi-scale spatial and spectral fusion cnn for hyperspectral image super-resolution. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 0–0.

[36] He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

[37] Heaton, J. (2018). Ian goodfellow, yoshua bengio, and aaron courville: Deep learning.

[38] Herrmann, C., Ruf, M., and Beyerer, J. (2018). Cnn-based thermal infrared person detection by domain adaptation. In *Autonomous Systems: Sensors, Vehicles, Security, and the Internet of Everything*, volume 10643, page 1064308. International Society for Optics and Photonics.

[39] Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A.-r., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N., et al. (2012a). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal processing magazine*, 29(6):82–97.

[40] Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. R. (2012b). Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*.

[41] Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., and Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*.

[42] Howell, K., Dudek, K., and Soroko, M. (2020). Thermal camera performance and image analysis repeatability in equine thermography. *Infrared Physics & Technology*, 110:103447.

[43] Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708.

[44] Huang, J.-B., Singh, A., and Ahuja, N. (2015). Single image super-resolution from transformed self-exemplars. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5197–5206.

[45] Huang, Y., Li, J., et al. (2021). Interpretable Detail-Fidelity Attention Network for Single Image Super-Resolution. *IEEE Transactions on Image Processing*, 30:2325–2339.

[46] Huang, Y., Shao, L., and Frangi, A. F. (2019). Simultaneous super-resolution and cross-modality synthesis in magnetic resonance imaging. In *Deep Learning and Convolutional Neural Networks for Medical Imaging and Clinical Informatics*, pages 437–457. Springer.

[47] Hwang, S., Park, J., Kim, N., Choi, Y., and So Kweon, I. (2015). Multispectral pedestrian detection: Benchmark dataset and baseline. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1037–1045.

[48] Jadin, M. S., Ghazali, K. H., and Taib, S. (2013). Thermal condition monitoring of electrical installations based on infrared image analysis. In *2013 Saudi International Electronics, Communications and Photonics Conference*, pages 1–6. IEEE.

[49] Joze, H. R. V., Zharkov, I., Powell, K., Ringler, C., Liang, L., Roulston, A., Lutz, M., and Pradeep, V. (2020). Imagepairs: Realistic super resolution dataset via beam splitter camera rig. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 518–519.

[50] Kawulok, M., Benecki, P., Piechaczek, S., Hrynczenko, K., Kostrzewa, D., and Nalepa, J. (2019). Deep learning for multiple-image super-resolution. *IEEE Geoscience and Remote Sensing Letters*, 17(6):1062–1066.

[51] Kelkar, V. A., Zhang, X., Granstedt, J., Li, H., and Anastasio, M. A. (2021). Task-based evaluation of deep image super-resolution in medical imaging. In *Medical Imaging 2021: Image Perception, Observer Performance, and Technology Assessment*, volume 11599, pages 207–213. SPIE.

[52] Keys, R. (1981). Cubic convolution interpolation for digital image processing. *IEEE transactions on acoustics, speech, and signal processing*, 29(6):1153–1160.

[53] Khattab, M. M., Zeki, A. M., Alwan, A. A., Badawy, A. S., and Thota, L. S. (2018). Multi-frame super-resolution: A survey. In *2018 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)*, pages 1–8. IEEE.

[54] Kim, J., Kwon Lee, J., and Mu Lee, K. (2016a). Accurate image super-resolution using very deep convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1646–1654.

[55] Kim, J., Kwon Lee, J., and Mu Lee, K. (2016b). Deeply-recursive convolutional network for image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1637–1645.

[56] Kim, K. I. and Kwon, Y. (2010). Single-image super-resolution using sparse regression and natural image prior. *IEEE transactions on pattern analysis and machine intelligence*, 32(6):1127–1133.

[57] Kittler, J. (1983). On the accuracy of the sobel edge detector. *Image and Vision Computing*, 1(1):37–42.

[58] Lanaras, C., Bioucas-Dias, J., Baltsavias, E., and Schindler, K. (2017). Super-resolution of multispectral multiresolution images from a single sensor. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 20–28.

[59] LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *nature*, 521(7553):436–444.

[60] Lee, K., Lee, J., Lee, J., Hwang, S., and Lee, S. (2017). Brightness-based convolutional neural network for thermal image enhancement. *IEEE Access*, 5:26867–26879.

[61] Li, Y., Tsiminaki, V., Timofte, R., Pollefeys, M., and Gool, L. V. (2019). 3d appearance super-resolution with deep learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9671–9680.

[62] Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., and Timofte, R. (2021). Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1833–1844.

[63] Lim, B., Son, S., Kim, H., Nah, S., and Mu Lee, K. (2017). 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.

[64] Lin, M. (2014). Cq, and yan, s. network in network. In *International Conference on Learning Representations (ICLR)*.

[65] Lin, M. and Qiang Chen, S. (2012). Imagenet classification with deep convolutional neural networks.

[66] Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

[67] Liu, Z., Luo, P., Wang, X., and Tang, X. (2015). Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738.

[68] Lobanov, A. P. (2005). Resolution limits in astronomical images. *arXiv preprint astro-ph/0503225*.

[69] Mandanici, E., Tavasci, L., Corsini, F., and Gandolfi, S. (2019). A multi-image super-resolution algorithm applied to thermal imagery. *Applied Geomatics*, 11(3):215–228.

[70] Mao, X., Shen, C., and Yang, Y.-B. (2016). Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections. In *Advances in neural information processing systems*, pages 2802–2810.

[71] Martin, D., Fowlkes, C., Tal, D., Malik, J., et al. (2001). A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. Iccv Vancouver:.

[72] Matsui, Y., Ito, K., Aramaki, Y., Fujimoto, A., Ogawa, T., Yamasaki, T., and Aizawa, K. (2017). Sketch-based manga retrieval using manga109 dataset. *Multimedia Tools and Applications*, 76(20):21811–21838.

[73] Mehri, A., Ardakani, P. B., and Sappa, A. D. (2021). Mprnet: Multi-path residual network for lightweight image super resolution. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2704–2713.

[74] Mehri, A. and Sappa, A. D. (2019). Colorizing near infrared images through a cyclic adversarial approach of unpaired samples. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*.

[75] Michel, J., Vinasco-Salinas, J., Inglada, J., and Hagolle, O. (2022). Sen2venμs, a dataset for the training of sentinel-2 super-resolution algorithms.

[76] Mittal, A., Soundararajan, R., and Bovik, A. C. (2012). Making a "completely blind" image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212.

[77] Molini, A. B., Valsesia, D., Fracastoro, G., and Magli, E. (2019). Deepsum: Deep neural network for super-resolution of unregistered multitemporal images. *IEEE Transactions on Geoscience and Remote Sensing*, 58(5):3644–3656.

[78] Mudunuri, S. P. and Biswas, S. (2015). Low resolution face recognition across variations in pose and illumination. *IEEE transactions on pattern analysis and machine intelligence*, 38(5):1034–1040.

[79] Müller, M. U., Ekhtiari, N., Almeida, R. M., and Rieke, C. (2020). Super-resolution of multispectral satellite images using convolutional neural networks. *arXiv preprint arXiv:2002.00580*.

[80] Neelima, D. and Sultana, S. (2012). An efficient perceptual quality index metrics.

[81] Olmeda, D., Premebida, C., Nunes, U., Armingol, J. M., and de la Escalera, A. (2013). Pedestrian detection in far infrared images. *Integrated Computer-Aided Engineering*, 20(4):347–360.

[82] Pandey, S., Singh, P. R., and Tian, J. (2020). An image augmentation approach using two-stage generative adversarial network for nuclei image segmentation. *Biomedical Signal Processing and Control*, 57:101782.

[83] Pavao, A., Guyon, I., Letournel, A.-C., Baró, X., Escalante, H., Escalera, S., Thomas, T., and Xu, Z. (2022). Codalab competitions: An open source platform to organize scientific challenges. *Technical report*.

[84] Pesavento, M., Volino, M., and Hilton, A. (2021). Attention-based multi-reference learning for image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14697–14706.

[85] Pickup, L. C. (2007). *Machine learning in multi-frame image super-resolution*. PhD thesis, Oxford University, UK.

[86] Price, J., Maraviglia, C., Seisler, W., Williams, E., and Pauli, M. (2004). System capabilities, requirements and design of the gdl gunfire detection and location system. In *33rd Applied Imagery Pattern Recognition Workshop (AIPR'04)*, pages 257–262. IEEE.

[87] Rasti, P., Uiboupin, T., Escalera, S., and Anbarjafari, G. (2016). Convolutional neural network super resolution for face recognition in surveillance monitoring. In *International conference on articulated motion and deformable objects*, pages 175–184. Springer.

[88] Ring, E. and Ammer, K. (2012). Infrared thermal imaging in medicine. *Physiological measurement*, 33(3):R33.

[89] Rivadeneira, R., Sappa, A., Vintimilla, B., Guo, L., Hou, J., Mehri, A., Ardakani, P., Patel, H., Chudasama, V., Prajapati, K., et al. (2020a). Thermal image superresolution challenge-pbvs 2020. in 2020 ieee. In *CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 432–439.

[90] Rivadeneira, R. E., Sappa, A. D., and Vintimilla, B. X. (2020b). Thermal image super-resolution: A novel architecture and dataset. In *VISIGRAPP (4: VISAPP)*, pages 111–119.

[91] Rivadeneira, R. E., Sappa, A. D., and Vintimilla, B. X. (2020c). Thermal image super-resolution: A novel unsupervised approach. In *International Joint Conference on Computer Vision, Imaging and Computer Graphics*, pages 495–506. Springer.

[92] Rivadeneira, R. E., Sappa, A. D., and Vintimilla, B. X. (2022a). Multi-image super-resolution for thermal images. In *VISIGRAPP (4: VISAPP)*, pages 635–642.

[93] Rivadeneira, R. E., Sappa, A. D., Vintimilla, B. X., and Hammoud, R. (2022b). A novel domain transfer-based approach for unsupervised thermal image super-resolution. *Sensors*, 22(6):2254.

[94] Rivadeneira, R. E., Sappa, A. D., Vintimilla, B. X., Kim, J., Kim, D., Li, Z., Jian, Y., Yan, B., Cao, L., Qi, F., et al. (2022c). Thermal image super-resolution challenge results-pbvs 2022. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 418–426.

[95] Rivadeneira, R. E., Sappa, A. D., Vintimilla, B. X., Nathan, S., Kansal, P., Mehri, A., Ardakani, P. B., Dalal, A., Akula, A., Sharma, D., et al. (2021). Thermal image super-resolution challenge-pbvs 2021. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4359–4367.

[96] Rivadeneira, R. E., Suárez, P. L., Sappa, A. D., and Vintimilla, B. X. (2019). Thermal image superresolution through deep convolutional neural network. In *International Conference on Image Analysis and Recognition*, pages 417–426. Springer.

[97] Robinson, M. D., Chiu, S. J., Toth, C. A., Izatt, J. A., Lo, J. Y., and Farsiu, S. (2017). New applications of super-resolution in medical imaging. In *Super-Resolution Imaging*, pages 401–430. CRC Press.

[98] Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer.

[99] Rövid, A., Vámossy, Z., and Sergyán, S. (2016). Thermal image processing approaches for security monitoring applications. In *Critical Infrastructure Protection Research*, pages 163–175. Springer.

[100] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. (2015). Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252.

[101] Salvetti, F., Mazzia, V., Khaliq, A., and Chiaberge, M. (2020). Multi-image super resolution of remotely sensed images using residual attention deep neural networks. *Remote Sensing*, 12(14):2207.

[102] Shamsolmoali, P., Zareapoor, M., Jain, D. K., Jain, V. K., and Yang, J. (2019). Deep convolution network for surveillance records super-resolution. *Multimedia Tools and Applications*, 78(17):23815–23829.

[103] Shawky, O. A., Hagag, A., El-Dahshan, E.-S. A., and Ismail, M. A. (2020). Remote sensing image scene classification using cnn-mlp with data augmentation. *Optik*, 221:165356.

[104] Shi, W., Ledig, C., Wang, Z., Theis, L., and Huszar, F. (2018). Super resolution using a generative adversarial network. US Patent App. 15/706,428.

[105] Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

[106] Sixsmith, A. and Johnson, N. (2004). A smart sensor to detect the falls of the elderly. *IEEE Pervasive computing*, 3(2):42–47.

[107] Song, D., Wang, Y., Chen, H., Xu, C., Xu, C., and Tao, D. (2021). Addersr: Towards energy efficient image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15648–15657.

[108] Sterckx, S., Benhadj, I., Duhoux, G., Livens, S., Dierckx, W., Goor, E., Adriaensen, S., Heyns, W., Van Hoof, K., Strackx, G., et al. (2014). The proba-v mission: Image processing and calibration. *International journal of remote sensing*, 35(7):2565–2588.

[109] Suarez, P. L., Sappa, A. D., Vintimilla, B. X., and Hammoud, R. I. (2019). Image vegetation index through a cycle generative adversarial network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*.

[110] Sun, D., Yang, X., Liu, M.-Y., and Kautz, J. (2018). 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.

[111] Sun, J., Xu, Z., and Shum, H.-Y. (2008). Image super-resolution using gradient profile prior. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE.

[112] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826.

[113] Timofte, R., Agustsson, E., Van Gool, L., Yang, M.-H., and Zhang, L. (2017). Ntire 2017 challenge on single image super-resolution: Methods and results. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 114–125.

[114] Timofte, R., Rothe, R., and Van Gool, L. (2016). 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.

[115] Tsai, R. (1984). Multiframe image restoration and registration. *Advance Computer Visual and Image Processing*, 1:317–339.

[116] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

[117] Wang, X., Yu, K., Dong, C., and Loy, C. C. (2018a). Recovering realistic texture in image super-resolution by deep spatial feature transform. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 606–615.

[118] Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., Qiao, Y., and Change Loy, C. (2018b). Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0.

[119] Wang, Z., Bovik, A. C., Sheikh, H. R., Simoncelli, E. P., et al. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612.

[120] Wang, Z., Chen, J., and Hoi, S. C. (2020). Deep learning for image super-resolution: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3365–3387.

[121] Wang, Z., Simoncelli, E. P., and Bovik, A. C. (2003). Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, volume 2, pages 1398–1402. Ieee.

[122] Watson, D. and Philip, G. (1987). Neighborhood-based interpolation. *Geobyte*, 2(2):12–16.

[123] Wei, Y., Gu, S., Li, Y., Timofte, R., Jin, L., and Song, H. (2021). Unsupervised real-world image super resolution via domain-distance aware training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13385–13394.

[124] Wong, W. K., Tan, P. N., Loo, C. K., and Lim, W. S. (2009). An effective surveillance system using thermal camera. In *2009 International Conference on Signal Acquisition and Processing*, pages 13–17. IEEE.

[125] Wu, Z., Fuller, N., Theriault, D., and Betke, M. (2014). A thermal infrared video benchmark for visual analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 201–208.

[126] Yamanaka, J., Kuwashima, S., and Kurita, T. (2017). Fast and accurate image super resolution by deep CNN with skip connection and network in network. *CoRR*, abs/1707.05425.

[127] Yang, J., Wright, J., Huang, T. S., and Ma, Y. (2010). Image super-resolution via sparse representation. *IEEE transactions on image processing*, 19(11):2861–2873.

[128] Yang, J., Zhao, Y.-Q., Chan, J. C.-W., and Xiao, L. (2019a). A multi-scale wavelet 3d-cnn for hyperspectral image super-resolution. *Remote sensing*, 11(13):1557.

[129] Yang, W., Zhang, X., Tian, Y., Wang, W., Xue, J.-H., and Liao, Q. (2019b). Deep learning for single image super-resolution: A brief review. *IEEE Transactions on Multimedia*, 21(12):3106–3121.

[130] Yanmaz, L. E., Okumus, Z., and Dogan, E. (2007). Instrumentation of thermography and its applications in horses. *J Anim Vet Adv*, 6(7):858–62.

[131] Yue, L., Shen, H., Li, J., Yuan, Q., Zhang, H., and Zhang, L. (2016). Image super-resolution: The techniques, applications, and future. *Signal Processing*, 128:389–408.

[132] Zefri, Y., ElKettani, A., Sebari, I., and Ait Lamallam, S. (2018). Thermal infrared and visual inspection of photovoltaic installations by uav photogrammetry—application case: morocco. *Drones*, 2(4):41.

[133] Zeyde, R., Elad, M., and Protter, M. (2010). On single image scale-up using sparse-representations. In *International conference on curves and surfaces*, pages 711–730. Springer.

[134] Zhang, H., Goodfellow, I., Metaxas, D., and Odena, A. (2019). Self-attention generative adversarial networks. In *International conference on machine learning*, pages 7354–7363. PMLR.

[135] Zhang, K., Zuo, W., Gu, S., and Zhang, L. (2017). Learning deep cnn denoiser prior for image restoration. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3929–3938.

[136] Zhang, L., Zhang, H., Shen, H., and Li, P. (2010). A super-resolution reconstruction algorithm for surveillance images. *Signal Processing*, 90(3):848–859.

[137] Zhang, L., Zhang, L., Mou, X., and Zhang, D. (2011). Fsim: A feature similarity index for image quality assessment. *IEEE transactions on Image Processing*, 20(8):2378–2386.

[138] Zhang, Z., Wang, H., Liu, M., Wang, R., Zhang, J., and Zuo, W. (2021). Learning raw-to-srgb mappings with inaccurately aligned supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4348–4358.

[139] Zhiyi, H., Haidong, S., Xiang, Z., Yu, Y., and Junsheng, C. (2020). An intelligent fault diagnosis method for rotor-bearing system using small labeled infrared thermal images and enhanced cnn transferred from cae. *Advanced Engineering Informatics*, 46:101150.

[140] Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232.