Boosting Guided Super-Resolution Performance with Synthesized Images

Patricia L. Suárez[1]
[1]*ESPOL Polytechnic University*
Guayaquil, Ecuador
plsuarez@espol.edu.ec

Dario Carpio[1]
[1]*ESPOL Polytechnic University*
Guayaquil, Ecuador
dncarpio@espol.edu.ec

Angel Sappa[1,2]
[2]*Computer Vision Center*
08193-Bellaterra, Barcelona, Spain
sappa@ieee.org

***Abstract*—Guided image processing techniques are widely used for extracting information from a guiding image to aid in the processing of the guided one. These images may be sourced from different modalities, such as 2D and 3D, or different spectral bands, like visible and infrared. In the case of guided cross-spectral super-resolution, features from the two modal images are extracted and efficiently merged to migrate guidance information from one image, usually high-resolution (HR), toward the guided one, usually low-resolution (LR). Different approaches have been recently proposed focusing on the development of architectures for feature extraction and merging in the cross-spectral domains, but none of them care about the different nature of the given images. This paper focuses on the specific problem of guided thermal image super-resolution, where an LR thermal image is enhanced by an HR visible spectrum image. To improve existing guided super-resolution techniques, a novel scheme is proposed that maps the original guiding information to a thermal image-like representation that is similar to the output. Experimental results evaluating five different approaches demonstrate that the best results are achieved when the guiding and guided images share the same domain.**

***Index Terms*—Thermal-Like Image, Guided Super-resolution, HSV Image**

I. Introduction

In recent years the usage of thermal imaging has largely increased to tackle different applications (e.g., pedestrian tracking [1], firefighting [2], and many others). Unfortunately, the main limitation of this technology lies in the low resolution of thermal cameras; accessing HR thermal images is relatively difficult and expensive. Trying to overcome this limitation different single-image super-resolution (SR) approaches have been developed (e.g., [3], [4]). Hence, in order to encourage the development of new techniques in this area, a thermal image SR challenge has been organized since 2020 as a part of the Perception Beyond the Visible Spectrum workshop at the CVPR conference [5]–[7]. While promising results have been achieved for $\times 2$ and $\times 4$ SR, it remains challenging to recover the information when large-scale upsampling is required. Hence, the development of more effective and efficient super-resolution techniques remains an important research area in thermal imaging.

In order to reach acceptable results with higher upsampling resolutions, recent studies propose to use the information from an HR cheap visible spectrum camera as guidance (e.g., [8], [9], and [10]). These approaches extract guidance information to drive the SR process of LR thermal images. The need for guided image processing is justified by its potential to improve image quality, facilitate super-resolution, integrate data from multiple modalities, and enhance semantic understanding. However, several research gaps and challenges must be addressed to maximize its potential.

Guidance SR techniques have been studied in different super-resolution domains, such as depth-map SR [11], infrared SR [12], [13], thermal SR [14], hyperspectral SR [15], and some others. MSG-Net [11], employs CNNs to accomplish guidance super-resolution, which is the first CNN model that attempts to upsample depth images under multi-scale guidance from the corresponding HR visible images. In [13], a new unsupervised approach is proposed to improve the visual quality of infrared images using a generative adversarial framework without high-resolution ground truth. The method includes a dual discriminator module and a content constraint module to enhance image details and maintain basic content. The approach produces realistic super-resolved images and is simpler to generalize for higher scales compared to supervised algorithms. More recently, [16] presents a guided SR approach, where a very LR face image is super-resolved up to $\times 8$ by means of a CNN architecture—referred to as GWAInet. The approach leverages an HR face image of the same person to guide the super-resolution process. GWAInet is trained adversarially to produce high-quality perceptual results and uses a warper subnetwork and a feature fusion chain to align and extract features from the HR guiding image and the LR input image.

Guidance information can also help to reduce artifacts and noise in the output image, resulting in a more visually appealing and natural-looking image [17]. It can also be used to control certain aspects of the output image, such as preserving certain image features or enhancing specific image details. This can be especially important in applications where the output image is intended for further analysis or to be used in downstream tasks. Commonly, guidance information is a critical component of super-resolution algorithms and plays a significant role in determining the quality of the output image.

Most of the guided super-resolution approaches mentioned above focus on developing novel architectures to efficiently extract and integrate features from the HR guiding image toward the LR-guided image during the super-resolution process.

979-8-3503-7091-1/23/$31.00 ©2023 IEEE
DOI 10.1109/SITIS61268.2023.00036

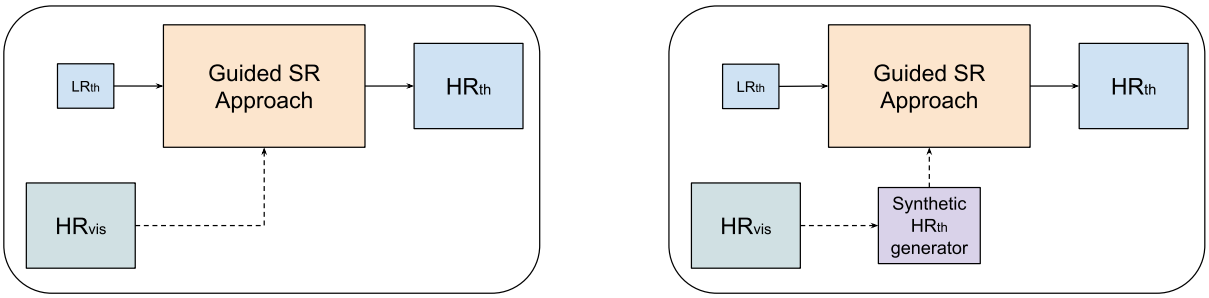

Fig. 1. Guided thermal image super-resolution approaches: $(left)$ State-of-the-art scheme; $(right)$ Proposed strategy.

On the contrary, the current work tackles the representation of the input images trying to map guiding to the guided domain. Our hypothesis is that as more similar the domain of the input images is as better the guidance process will be. This hypothesis is validated by testing different state-of-the-art approaches using as a guide an HR thermal-like image, instead of the given HR visible spectrum image. In all the tested approaches better results are obtained with the proposed strategy. The manuscript is organized as follows. Section II presents the strategy used for obtaining the guided thermal image super-resolution from a thermal-like representation. Experimental results and comparisons with different implementations are given in Section III. Finally, conclusions and future works are presented in Section IV.

II. Proposed Strategy

The coexistence of information from different modalities has inspired the research community to explore different strategies to combine this information. These strategies go from information fusion, where a new representation is obtained by merging the given inputs (e.g., [18], [19]), to guided approaches, where one of the images is used to guide the processing of the other (e.g., filtering [10], super-resolution [16]). In this paper, state-of-the-art guided super-resolution techniques are evaluated by updating the guiding image, in general, a high-resolution visible spectrum image, with a synthetic representation of it. Figure 1 depicts an illustration of the proposed strategy. This section first introduces the modifications proposed to the approach presented in [20], where a cyclic adversarial network with multiple loss functions is used to obtain a pseudo-thermal representation. One of the proposed changes is in the patch sampling process used for the contrastive loss, where the input data are shuffled so that the same patch IDs are not used for multiple images in the same batch when comparing the image regions. Another change lies at the optimizer level in order to increase the convergence of the model during the training process. In Section II-B, all the evaluated state-of-the-art guided super-resolution approaches are briefly described.

A. Thermal Image-Like Representation

The generation of synthetic representations has been widely explored in several machine vision applications, such as synthetic face representations [21], image colorization (e.g., [22], [23]), vegetation index estimation (e.g., [24], [25]), just to mention a few. Recently, in [20] a novel approach has been proposed to generate thermal image-like representations from low-cost visible images. The main idea is to generate pseudothermal synthetic images that can provide valuable information about the objects in the scene and can be used to improve the performance of other machine vision algorithms.

The architecture presented in [20] is a generative cycle GAN architecture that enables domain adaptation from the brightness channel of an HSV image to a thermal-like image. The model achieves convergence through the use of several loss functions, including relativistic, contrastive, cycle consistency, and identity losses. Herein, we provide a detailed explanation of the underlying principles guiding the use of these loss functions in the generation of synthetic thermal images. First, the use of relativistic adversarial loss, has demonstrated effectiveness in improving the stability and quality of GANs, especially when generating high-dimensional data. This loss function mandates that the generated samples resemble the actual samples more closely, mitigating model saturation and accelerating the training process. Given the benefits of this loss, we have integrated it into a cycled transformation model as proposed in [20], resulting in successful outcomes:

$$L_D^{RGAN} = \mathbb{E}_{(x_r,x_f)\sim(\mathbb{P},\mathbb{Q})}\left[f\left(C\left(x_r\right) - C\left(x_f\right)\right)\right], \quad (1)$$

$$L_G^{RGAN} = \mathbb{E}_{(x_r,x_f)\sim(\mathbb{P},\mathbb{Q})}\left[g\left(C\left(x_f\right) - C\left(x_r\right)\right)\right], \quad (2)$$

where f and g are functions mapping a scalar input to another scalar and x_r, x_f is the real and fake image respectively. According to [26], contrastive loss has also been implemented to minimize the distance between similar pairs of data points and maximize the distance between dissimilar pairs of data points in a given dataset. This contrastive loss can be defined as:

$$\mathcal{L}_{\text{cont}}(\hat{Y}, Y) = \sum_{l=1}^{L}\sum_{s=1}^{S_l} \ell_{\text{contr}}\left(\hat{v}_l^s, v_l^s, \bar{v}_l^s\right), \quad (3)$$

where the shape of the tensor $V_l \in \mathbb{R}^{S_l \times D_l}$ is dependant on model architecture, and S_l is the number of spatial locations of the tensor. Therefore, the tensor is indexed with the notation $v_l^s \in \mathbb{R}^{D_l}$, which is the D_l-dimensional feature vector at spatial location s^{th} and the notation $\hat{v}_l^s \in \mathbb{R}^{D_l}$, which is also

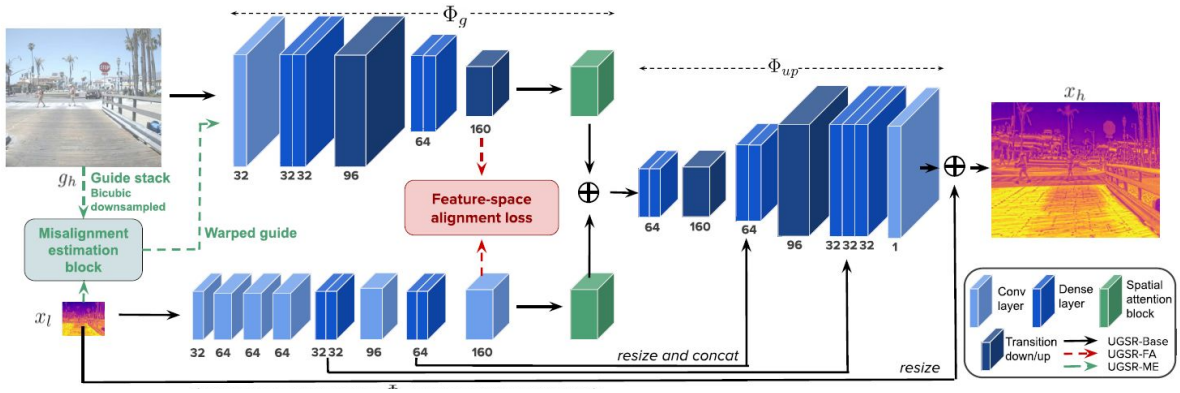

Fig. 2. UGSR architecture (illustration from [8]).

the D_l-dimensional predicted feature tensor at spatial location s^{th} . It has been denoted $\bar{v}_l^s \in \mathbb{R}^{(S_l-1)\times D_l}$ as the collection of feature vectors at all other spatial locations. Identity loss has been included, as proposed in [20]. It computes the difference between the features extracted from the real and generated images. By minimizing the intensity loss, the generator network can learn to produce outputs that not only look realistic but also have similar features and structures as the real images; this loss is defined as:

$$\begin{aligned}\mathcal{L}_{\text{ident}}\ (G,F) = \mathbb{E}_{c\sim P_{\text{data}}\,(c)}\left[\|F(c)-c\|\right] \\ + \mathbb{E}_{n\sim P_{\text{data}}\,(n)}\left[\|G(n)-n\|\right],\end{aligned} \tag{4}$$

where G and F correspond to mapping functions that generate the synthetic images $F(G(x))$ and $G(Y(z))$ respectively and c and n correspond to a real image from the source and target domains respectively.

Additionally, a cycle consistency loss is used in order to enhance the translation results and also, helps to ensure that the mapping between the domains is consistent and bijective. This means that if an image is translated from domain A to domain B and then back to domain A, it should be the same as the original image from domain A. It is defined as:

$$\begin{aligned}\mathcal{L}_{cycle}(G,F) = \mathbb{E}_{x}\sim_{p\ \text{data}_{(x)}}[\|F(G(x))-x\|_1] \\ +\mathbb{E}_{y}\sim_{p\ \text{data}_{(y)}}[\|G(F(y))-y\|_1],\end{aligned} \tag{5}$$

where G and F correspond to mapping functions that generate the reconstructed images $F(G(x))$ and $G(F(y))$ respectively and x,y correspond to real images. The cycle consistency loss encourages $F(G(x)) \approx x$ real and $G(F(y)) \approx y$ real. This loss allows for generating high-quality images that are both realistic and semantically meaningful. Finally, the multiple loss functions implemented in our model can be defined as:

$$\begin{aligned}\mathcal{L}_{\text{final}}\ = \lambda_1\mathcal{L}_{\text{RGAN}}(G,D,X,Y) + \lambda_2\mathcal{L}_{\text{cont}}\ (G,H,X) \\ +\lambda_3\mathcal{L}_{\text{cont}}(G,H,Y) + \lambda_4\mathcal{L}_{\text{ident}(G,F)} + \lambda_5\mathcal{L}_{\text{cycle}(G,F)},\end{aligned} \tag{6}$$

where λ_i are empirically defined.

B. Guided Super-Resolution Approaches

This section briefly details the guided SR approaches evaluated in the current work. Note that some of the approaches have been originally proposed for guiding SR of depth maps, while others are for thermal images. In order to evaluate all of them in a common framework just thermal images are considered as input for training them, in spite of the fact they were proposed for thermal or depth map images. All the approaches from the state-of-the-art included in this section have the corresponding code provided by the authors. All these approaches are used in the experimental results section of this work.

Toward Unaligned Guided Thermal Super-Resolution — UGSR: In [8] two models are proposed for guided super-resolution of unaligned thermal and visible images without pixel-to-pixel alignment. Figure 2 depicts the proposed UGSR architecture. The first model employs a correlation-based feature misalignment loss (UGSR-FA), while the second model includes a misalignment map estimation block to compensate for misalignment in an end-to-end manner (UGSR-ME). The USGR architecture is built with two branches of encoders, one for the low-resolution thermal image and the other for the high-resolution guide image. The encoders use dense blocks and a spatial attention module for feature extraction and rescaling. The features are merged and fed to a decoder for generating the final high-resolution thermal image. Bicubic upsampling of the input image is also used for residual learning and the spatial attention module is applied to remove texture edges or other information that is not common to both images and could cause artifacts. According to the authors, the correlation-based feature misalignment model (UGSR-FA) reaches the best results.

Guided Super-Resolution as Pixel-to-Pixel Transformation — PixTransform: In [9], instead of using a standard super-resolution approach, the authors propose a pixel-to-pixel transformation from the guiding image to the target image domain without changing the resolution (see illustration in Fig. 3). The authors use a multi-layer perceptron that takes the

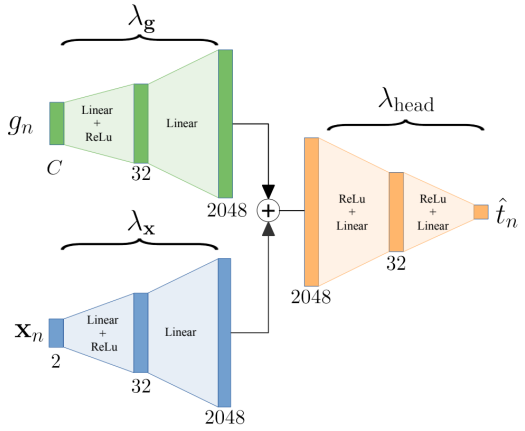

Fig. 3. Pixel-to-Pixel transformation architecture (illustration from [9]).

guiding image's pixel values, which are augmented with two extra channels to encode pixel location and passed through a convolutional network with 1×1 kernels. This setup allows for encoding spatial context relations implicitly, with a single set of transformation parameters that work for all pixels, making it unsupervised. The method fits a unique set of weights for each new image using its pixels as training data and consistency with the low-resolution source as supervision. The weights of the model are learned by minimizing the discrepancy between the source image and the downsampled target image.

Pyramidal Edge-Maps and Attention-based Guided Thermal Super-Resolution — PAGSR: This approach is intended for guided super-resolution of thermal images using visible images [14]; it relies on pyramidal edge maps to reduce artifacts in the resulting images. The method proposes to carry out the super-resolution using the pyramidal edges obtained from multiple hierarchical levels resulting from the merge of previously extracted edge features. The authors use attention mechanisms with this border information in the super-resolution network to integrate them. The key challenge is to extract high-frequency details from the guiding image and integrate them with the thermal image in an adaptive way so that the reconstructed image is both visually pleasing and free from artifacts. Applying this method improves the texture of the objects present in the images resulting from the guided super-resolution. Figure 4 shows the architecture illustrating the LR thermal image as well as the edges used as inputs.

Deformable Kernel Networks for Joint Image Filtering — DKN: The authors in [10] propose a new approach to joint image filtering called the deformable kernel network (DKN), which uses sparse and spatially-variant kernels instead of nonlinear activation of spatially-invariant kernels. DKN outputs sets of neighbors and corresponding weights for each pixel and the filtering result is computed as a weighted average. The authors also propose a fast version of DKN that runs 17 times faster for an image of size 640×480. The effectiveness and flexibility of DKN are demonstrated on various computer vision tasks such as depth map upsampling,

TABLE I
RESULTS OF THE GUIDED SUPER-RESOLUTION APPROACHES EVALUATED IN THE CURRENT WORK, A $\times8$ SCALE FACTOR IS CONSIDERED.

Methods	Visible guidance		Synthetic guidance	
	PSNR	SSIM	PSNR	SSIM
PixTransform [9]	23.360	0.627	23.831	0.742
PAGSR [14]	27.454	0.831	28.843	0.869
FDKN [10]	26.924	0.825	30.302	0.926
UGSR [8]	28.524	0.860	33.312	0.950
DCTNet [12]	28.910	0.835	32.215	0.922

saliency map upsampling, cross-modality image restoration, texture removal, and semantic segmentation. Figure 5 shows the DKN architecture of the guided filtering, which involves learning kernel weights and spatial sampling offsets from feature maps of guidance and target images to obtain a residual image. The model is fully convolutional and learned end-to-end, using element-wise multiplication and dot product operations. Reshaping and residual connections are also used.

Discrete Cosine Transform Network for Guided Depth Map Super-Resolution — DCTNet: A novel approach for Guided Depth Super-Resolution (GDSR) has been presented in [12]. The approach is based on the usage of the Discrete Cosine Transform Network (DCTNet) to improve the reconstruction of high-resolution depth maps from low-resolution ones. The proposed DCTNet method addresses the issues in GDSR by utilizing a Discrete Cosine Transform (DCT) to reconstruct high-resolution depth features, a semi-coupled feature extraction process with shared and individual convolutional kernels, and an attention mechanism to emphasize important edges in the image, see Fig. 6. The DCTNet consists of several modules that work together to perform super-resolution on a low-resolution depth image and HR RGB image. The first module, called SCFE, extracts both shared and private features from the two images. The GESA module then uses the RGB features to obtain edge attention weights that are helpful for super-resolution. These features and weights are then processed by the DCT module, which uses the DCT in each channel to obtain high-resolution depth features. The final step is performed by the reconstruction module, which outputs the super-resolved depth map.

C. Datasets

The first dataset, known as M3FD, was recently introduced for image fusion in Liu et al. [27]. This dataset was used to train the thermal image-like image generator described in Section II-A. Images were converted to HSV color space and the brightness channel was extracted to aid translation from the visible to the thermal domain. The M3FD data set contains 4500 pairs of visible and infrared images acquired using a binocular optical and infrared sensor. These images depict various scenes, such as highways, campuses, streets, forests, and more, captured during the day, night, and under cloudy skies. In addition to the M3FD dataset, we also use our own cross-spectral high-resolution dataset known as the Thermal Stereo dataset to evaluate the generalization capabilities of

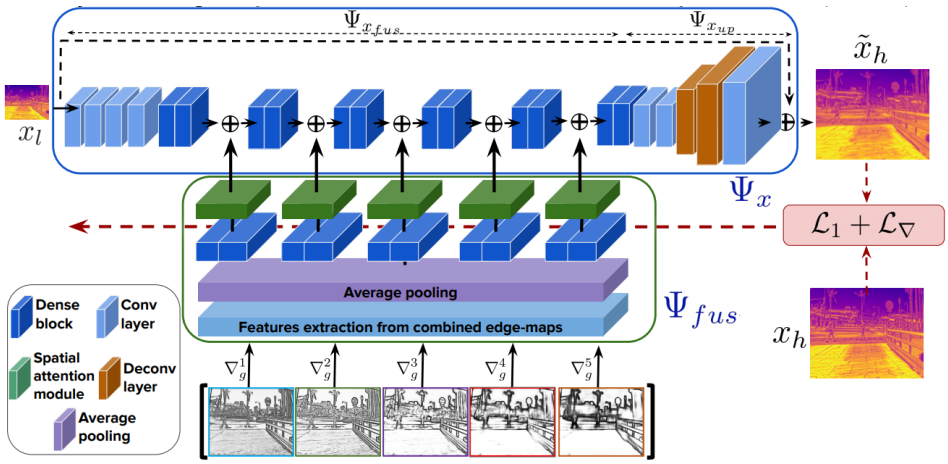

Fig. 4. Pyramidal edge-maps and attention-based architecture for guiding thermal image super-resolution (illustration from [14]).

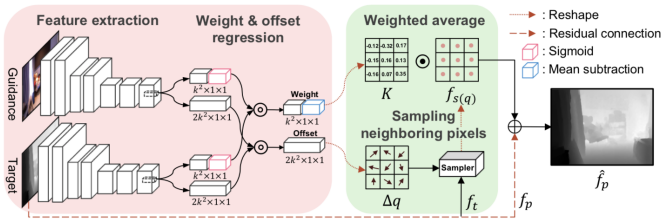

Fig. 5. DKN architecture (illustration from [10]).

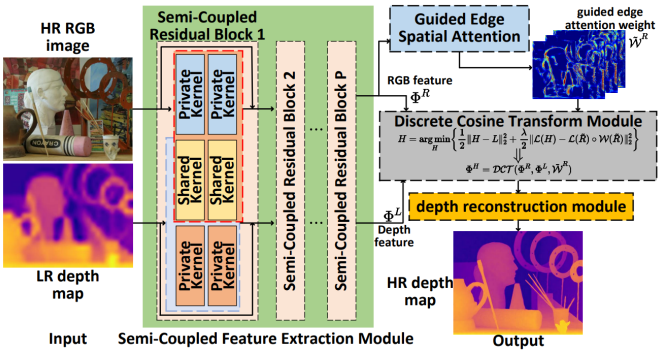

Fig. 6. Guided depth super-resolution architecture (illustration from [12]).

our synthetic thermal image generator. With these thermal-like generated images, it has been possible to train and evaluate each of the guided superresolution models already presented above, using synthetic images. This custom dataset was acquired using FLIR FC-6320 and TAU2 cameras and consists of 200 pairs of images, comprising both thermal images and their corresponding visible images. To ensure accurate alignment between the two modalities, the images were registered using the Elastix method [28], resulting in well-aligned pairs with a resolution of 640x480 pixels.

III. Experimental Results

This section presents the experimental results obtained with the state-of-the-art guided super-resolution techniques presented in Section II-B using two different strategies—i.e., guided with the visible spectrum images and guided with the synthesized thermal images generated in Section II-A. Also, this section presents the comprehensive evaluation of state-of-the-art guided super-resolution approaches discussed in Section II-B. Our evaluation encompasses two distinct strategies: i) thermal SR using HR RGB images as guidance and ii) thermal image SR considering HR thermal image-like as guidance. To conduct quantitative and qualitative evaluations, we use the Thermal Stereo dataset. Prior to evaluation, all images were pre-processed by resizing them to a uniform resolution of 512×512 pixels. For the generation of LR thermal images, we applied downsampling using bicubic interpolation on the HR thermal images. To train the guided super-resolution methods, we split the Thermal Stereo dataset up into three subsets: 160 image pairs for training, 30 image pairs for validation, and 10 image pairs for testing.

Our evaluation focused on the five state-of-the-art guided super-resolution methods presented in Section II-B. These methods were evaluated using the proposed strategy, which involves training on both visible and synthesized thermal images, with a scale factor of $\times 8$. To assess the performance of these methods, we employed widely used metrics such as SSIM (Structural Similarity Index) and PSNR (Peak Signal-to-Noise Ratio). Table I shows the results obtained by both guided strategies in all the evaluated approaches. In all cases, improvements can be seen in both metrics when using similar domain data images as a guide. Qualitatively, the super-resolved images with our strategy have greater contour detail than the images produced by the guided approaches with visible spectrum imaging. Fig. 7 shows the qualitative and quantitative results, in a sample of the test set, which was obtained with each superresolution method according to both guiding strategies. The SSIM and PSNR values are shown

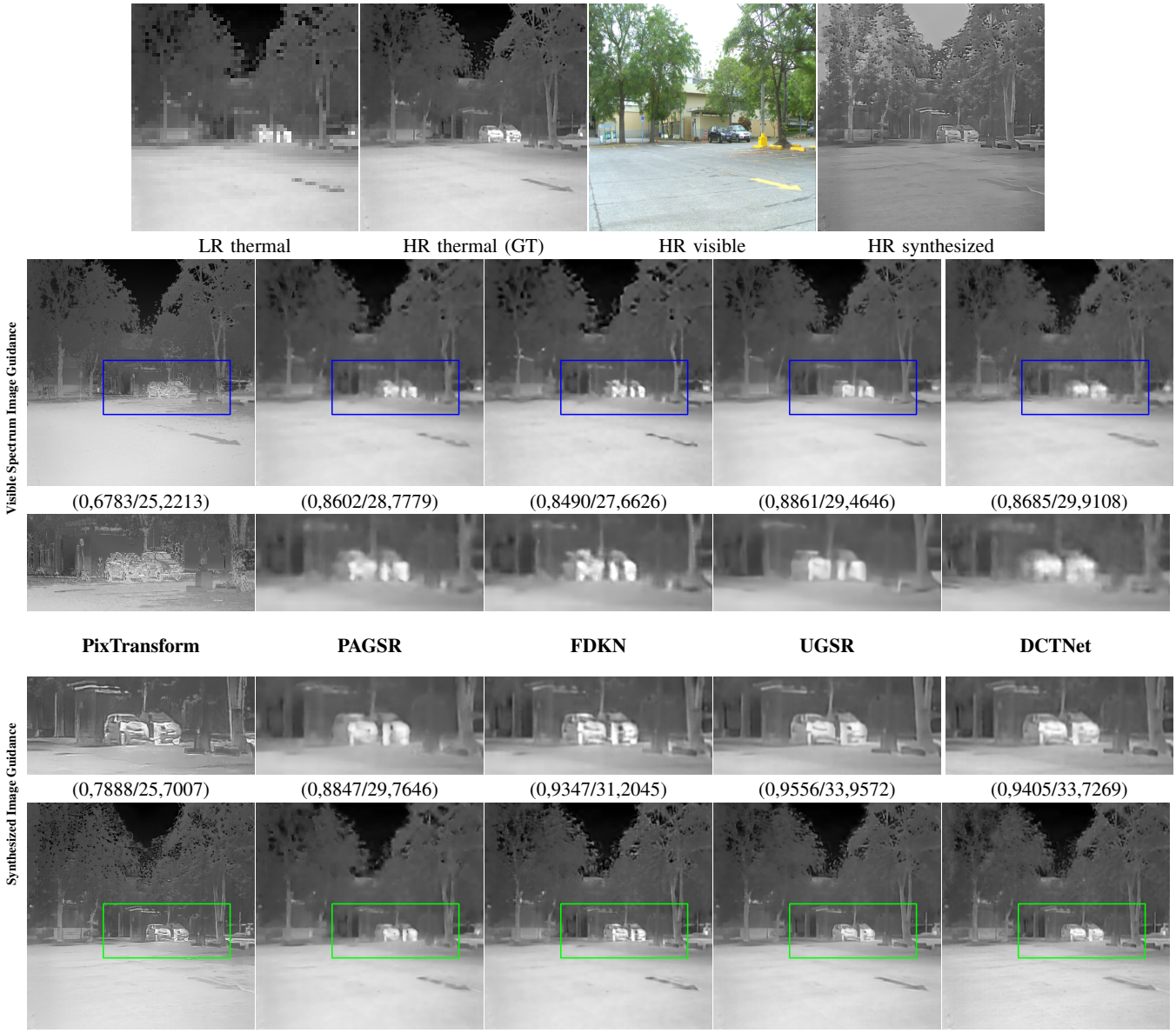

Fig. 7. (*top*) Illustration of LR thermal image, HR thermal image (ground truth), HR visible spectrum image, and the synthesized HR thermal image. (*middle*) Results of (×8) super-resolution guided by the HR visible spectrum image. (*bottom*) Results of (×8) super-resolution guided by the proposed synthesized HR thermal image — SSIM and PSNR values in parenthesis.

in parentheses. In all the evaluated state-of-the-art techniques the proposed strategy helps to improve results, in some cases, more than 18% in SSIM value, such is the case of PixTransform and more than 16% in PSNR value, such is the case of the UGSR architecture.

IV. Conclusions

This work shows that guided image processing approaches, in particular guided thermal image super-resolution, can be improved if guiding information overlaps the guided domain. Hence, the key factor is to have an efficient generator network able to synthesize the working domain. As more realistic the generated images as better the guidance process will be performed. The experiments involving the generation of synthesized thermal images are important because they demonstrate the ability to create high-quality thermal image-like representations that can be used as guidance in image processing approaches. By using synthesized images, the evaluated models can easily obtain the guidance information needed for super-resolution. As a future work, we will extend the present study to the guided depth super-resolution and guided denoising image processing. One of the limitations is the need to have high-quality guidance images available, which may not be feasible to acquire. Additionally, the number of images available may be limited, which could affect the performance of the model. Future work should explore strategies to handle scenarios that are not as complex or have lower resolution quality. Also the other limitation may be the minimal resource

requirements for our approach to support the scalability of images and computational resources, particularly when dealing with large-scale data sets or high-resolution images and the computational power and memory required for processing. which can become substantial as the complexity and size of the guidance and target data increase. Mitigating these challenges would involve optimizing the algorithms to ensure that our approach remains feasible and accessible in resource-constrained scenarios.

ACKNOWLEDGEMENTS

This material is based upon work supported by the Air Force Office of Scientific Research under award number FA9550-22-1-0261; and partially supported by the Grant PID2021-128945NB-I00 funded by MCIN/AEI/10.13039/501100011033 and by "ERDF A way of making Europe"; the "CERCA Programme / Generalitat de Catalunya"; and the ESPOL project CIDIS-12-2022.

REFERENCES

[1] M. Bertozzi, A. Broggi, C. Caraffi, M. Del Rose, M. Felisa, and G. Vezzoni, "Pedestrian detection by means of far-infrared stereo vision," *Computer Vision and Image Understanding*, vol. 106, no. 2-3, pp. 194–204, 2007.

[2] J. Baek, S. Hong, J. Kim, and E. Kim, "Efficient pedestrian detection at nighttime using a thermal camera," *Sensors*, vol. 17, no. 8, p. 1850, 2017.

[3] R. E. Rivadeneira, P. L. Suárez, A. D. Sappa, and B. X. Vintimilla, "Thermal image superresolution through deep convolutional neural network," in *Image Analysis and Recognition: 16th International Conference, ICIAR 2019, Waterloo, ON, Canada, August 27–29, 2019, Proceedings, Part II 16*, pp. 417–426, Springer, 2019.

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

[5] R. Rivadeneira, A. Sappa, B. Vintimilla, L. Guo, J. Hou, A. Mehri, P. Ardakani, H. Patel, V. Chudasama, K. Prajapati, *et al.*, "Thermal image superresolution challenge," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 432–439, 2020.

[6] R. E. Rivadeneira, A. D. Sappa, B. X. Vintimilla, S. Nathan, P. Kansal, A. Mehri, P. B. Ardakani, A. Dalal, A. Akula, D. Sharma, *et al.*, "Thermal image super-resolution challenge," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 4359–4367, 2021.

[7] R. E. Rivadeneira, A. D. Sappa, B. X. Vintimilla, J. Kim, D. Kim, Z. Li, Y. Jian, B. Yan, L. Cao, F. Qi, *et al.*, "Thermal image super-resolution challenge results," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 418–426, 2022.

[8] H. Gupta and K. Mitra, "Toward unaligned guided thermal super-resolution," *IEEE Transactions on Image Processing*, vol. 31, pp. 433–445, 2021.

[9] R. d. Lutio, S. D'aronco, J. D. Wegner, and K. Schindler, "Guided super-resolution as pixel-to-pixel transformation," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 8829–8837, 2019.

[10] B. Kim, J. Ponce, and B. Ham, "Deformable kernel networks for joint image filtering," *International Journal of Computer Vision*, vol. 129, no. 2, pp. 579–600, 2021.

[11] T.-W. Hui, C. C. Loy, and X. Tang, "Depth map super-resolution by deep multi-scale guidance," in *Proceedings of 14th European Conference on Computer Vision*, pp. 353–369, Springer, 2016.

[12] Z. Zhao, J. Zhang, S. Xu, Z. Lin, and H. Pfister, "Discrete cosine transform network for guided depth map super-resolution," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5697–5707, 2022.

[13] Y. Huang, W. Wang, and L. Wang, "Bidirectional recurrent convolutional networks for multi-frame super-resolution," *Advances in Neural Information Processing Systems*, vol. 28, 2015.

[14] H. Gupta and K. Mitra, "Pyramidal edge-maps and attention based guided thermal super-resolution," in *Proceedings of the European Conference on Computer Vision Workshops*, pp. 698–715, Springer, 2020.

[15] W. Dong, C. Zhou, F. Wu, J. Wu, G. Shi, and X. Li, "Model-guided deep hyperspectral image super-resolution," *IEEE Transactions on Image Processing*, vol. 30, pp. 5754–5768, 2021.

[16] B. Dogan, S. Gu, and R. Timofte, "Exemplar guided face image super-resolution without facial landmarks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, June 2019.

[17] K. He, J. Sun, and X. Tang, "Guided image filtering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 6, pp. 1397–1409, 2012.

[18] H. Xu, M. Gong, X. Tian, J. Huang, and J. Ma, "Cufd: An encoder–decoder network for visible and infrared image fusion based on common and unique feature decomposition," *Computer Vision and Image Understanding*, vol. 218, p. 103407, 2022.

[19] B. Meher, S. Agrawal, R. Panda, L. Dora, and A. Abraham, "Visible and infrared image fusion using an efficient adaptive transition region extraction technique," *Engineering Science and Technology, an International Journal*, vol. 29, p. 101037, 2022.

[20] P. L. Suárez and A. D. Sappa, "Toward a thermal image-like representation," in *Proceedings of the International joint Conference on Computer Vision*, 2023.

[21] J. C. Peterson, S. Uddenberg, T. L. Griffiths, A. Todorov, and J. W. Suchow, "Deep models of superficial face judgments," *Proceedings of the National Academy of Sciences*, vol. 119, no. 17, 2022.

[22] A. Mehri and A. D. Sappa, "Colorizing near infrared images through a cyclic adversarial approach of unpaired samples," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 0–0, 2019.

[23] P. L. Suárez, A. D. Sappa, and B. X. Vintimilla, "Learning to colorize infrared images," in *Trends in Cyber-Physical Multi-Agent Systems. The PAAMS Collection-15th International Conference, PAAMS 2017 15*, pp. 164–172, Springer, 2018.

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

[25] P. L. Suárez, A. D. Sappa, and B. X. Vintimilla, "Deep learning-based vegetation index estimation," in *Generative Adversarial Networks for Image-to-Image Translation*, pp. 205–234, Elsevier, 2021.

[26] A. Andonian, T. Park, B. Russell, P. Isola, J.-Y. Zhu, and R. Zhang, "Contrastive feature loss for image prediction," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1934–1943, 2021.

[27] J. Liu, X. Fan, Z. Huang, G. Wu, R. Liu, W. Zhong, and Z. Luo, "Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5802–5811, 2022.

[28] S. Klein, M. Staring, K. Murphy, M. A. Viergever, and J. P. W. Pluim, "elastix: A toolbox for intensity-based medical image registration," *IEEE Transactions on Medical Imaging*, vol. 29, no. 1, pp. 196–205, 2009.