

SwinFuSR: an image fusion-inspired model for RGB-guided thermal image super-resolution

Cyprien Arnold
Polytechnique Montréal
Montréal, Canada

cyprien.arnold@polymtl.ca

Philippe Jouvet
CHU Sainte Justine
Montréal, Canada

philippe.jouvet.med@ssss.gouv.qc.ca

Lama Seoud
Polytechnique Montréal
Montréal, Canada

lama.seoud@polymtl.ca

Abstract

Thermal imaging plays a crucial role in various applications, but the inherent low resolution of commonly available infrared (IR) cameras limits its effectiveness. Conventional super-resolution (SR) methods often struggle with thermal images due to their lack of high-frequency details. Guided SR leverages information from a high-resolution image, typically in the visible spectrum, to enhance the reconstruction of a high-res IR image from the low-res input. Inspired by SwinFusion, we propose SwinFuSR, a guided SR architecture based on Swin transformers. In real world scenarios, however, the guiding modality (e.g. RGB image) may be missing, so we propose a training method that improves the robustness of the model in this case. Our method has few parameters and outperforms state of the art models in terms of Peak Signal to Noise Ratio (PSNR) and Structural SIMilarity (SSIM). In Track 2 of the PBVS 2024 Thermal Image Super-Resolution Challenge, it achieves 3rd place in the PSNR metric. Our code and pretrained weights are available at <https://github.com/VisionICLab/SwinFuSR>.

1. Introduction

Improving the quality of digital images is crucial in numerous fields, from mobile photography [19] and healthcare [14, 34, 36] to law enforcement [33]. Super-resolution (SR) has emerged as a promising technique to achieve this goal, allowing the reconstruction of high-resolution (HR) images from their low-resolution (LR) counterparts. In the realm of RGB images, SR has witnessed significant advancements in recent years. Diverse techniques, ranging from traditional methods to deep learning, have been developed to

exploit information within LR images and generate realistic detailed HR images.

Infrared (IR) images, capturing the heat emitted by objects, enable night vision and the ability to detect features invisible to the naked eye. The IR modality is used for continuous and contactless monitoring of patients' vital signs in the intensive care units (ICU) [4, 41] and to integrate this information into clinical decision support systems [52]. To achieve this, IR acquisitions are often combined with RGB images, and even 3D images [3]. High definition IR sensors with spatial resolution of up to $1,024 \times 768$ pixels are commercially available, but can cost tens of thousands of dollars. Hence, lower resolution IR sensors tend to be used instead in ICU rooms.

Thermal image super-resolution (TISR) tackles this challenge by increasing image resolution and revealing details obscured in the LR image. This topic is increasingly studied because of its many applications [18] including in medical science [40, 46], agricultural management [5, 35] or even space studies [15, 53]. Several challenges remain in fully realizing the potential of IR super-resolution. One key challenge lies in the inherent differences between IR and RGB images. IR images exhibit higher noise and poorer texture information [18], making HR reconstruction more complex.

Guided thermal image super resolution (GTISR) presents itself as a particularly promising approach for IR image reconstruction. By relying on an HR reference image as input, such as a corresponding visible spectrum image, guided SR can improve the accuracy and consistency of the reconstruction. In effect, HR RGB images are cost-effective to obtain and have higher frequencies than IR images. To encourage researchers to innovate in this little-explored field, the 19th IEEE Workshop on Perception Be-

yond the Visible Spectrum introduced a challenge track [38] in 2023 to generate x8 super-resolution thermal images by using visible HR images as guidance. Candidates are ranked according to Peak Signal to Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM) between images produced from the test set and the (non-public) ground truth HR IR images.

In this paper, we draw inspiration from multimodal image fusion based on Swin Transformers to propose Swin-FuSR, a novel method for RGB-guided thermal image super-resolution. Our contributions are two-fold:

- a lightweight transformer-based model that outperforms other state of art GTISR methods.
- a modified training strategy that improves the robustness of GTISR in the absence of the guiding modality.

2. Related Works

2.1. Visible image super-resolution

The first approaches to super resolution employed so-called “traditional” approaches [18]. These methods either focused on the frequency domain, trying to model the relationship between the HR and LR images using mathematical models [1, 37, 42], or used dictionaries methods to map LR patches to HR patches [49, 54, 57].

2015 saw the emergence of deep learning-based methods using convolutional neural networks (CNNs) such as SRCNN [9], FSRCNN [10] and ESPCN [43], which introduced subpixel convolutional layers, a new upsampling operation. The advent of residual networks [16] (to solve the vanishing gradient problem in particular) led to new architectures like VDSR [21], RED [32] and EDSR [26], the latter proposing a new residual connection and winning the NTIRE2017 Super-Resolution Challenge [45].

In 2017, SRGAN [23] achieved remarkable results by applying a generative adversarial network (GAN) to the SR task. One year later appeared ESRGAN[50], an enhanced version of SRGAN.

Since transformers [47] have been adapted to the field of computer vision with the Vision Transformer ViT [11], the Swin transformer [29] resolved the computational complexity problem of ViT by using shifted windows. With this mechanism, SwinIR [24] applied the Swin architecture to the image reconstruction task and outperformed the best existing architectures. SwinIR’s main strength lies in its Residual Swin Transformer Blocks, which extract highly relevant features. More recently, the HAT [7] and SwinFIR [56] architectures have proposed improvements to SwinIR and represent the current state of the art in SR.

2.2. Thermal image super-resolution

Compared to RGB images, IR images are single channel, have low gradients and “overlapping information between

high and low frequencies” [18]. To manage these characteristics, specific architectures have been proposed for IR images.

Before deep learning, frequency domain-based solutions like [48] or dictionary-based methods [8] were proposed. Then, inspired by the methods used in the visible spectrum, [12] and [61] exploited CNNs and residual networks. Other architectures have come up with the idea of using visible information (more abundant data) to reconstruct the IR image. For example, [55] used visible information in the loss function, while PSRGAN [17] used a GAN framework and transfer learning from RGB images to train their SR algorithm.

More recently, approaches using transformers have appeared, namely DASR [25] that exploits spatial and channel attention. In the same spirit, [58] dynamically reweights the output of attention and non-attention branches to improve the resolution and restore high-frequency details, offering a lightweight structure suitable for edge device deployment.

2.3. Guided thermal super-resolution

Unlike the methods presented above, guided methods take two paired images as input: an LR thermal image (or target image) and a higher-resolution guide image to help with the SR task. One of the first GTISR works was introduced in [60]; it used a dual-path residual network to merge features from the visible and IR domains. More recently, CoRefusion [20] has been proposed. Its architecture is composed of two U-Nets [39] with residual connections to fuse both modalities. A contrastive term is added to the loss function and yields improved performance.

CoreFusion was part of the first GTISR track in the 2023 PBVS competition [38]. However, the winner of that competition was GuidedSR [38]; this latter approach concatenates RGB and IR features from the shallow feature extraction layers and uses Non-linear Activation Free (NAF) blocks [6] to fuse RGB and IR information.

More recently, the authors of [44] described several SR guided methods applied to thermal images and how transforming the RGB guide image into a “thermal-like image” improves performance. They show that this substitution boosts performance by a few percentage points in different super-resolution guided architectures.

2.4. Multimodal image fusion

Multimodal image fusion aims at combining relevant information from images acquired with different sensors into a single image. DL-based fusion methods can be divided into three categories: early fusion, late fusion and hybrid fusion[2]. The first one merges features before task related layers, the second one uses task related layers on each modality before aggregating the information, while the last one combines the first two approaches.

SwinFusion [30] proposed to fuse images from two modalities using Attention-guided Cross-domain Fusion (ACF). Inspired by the Swin Transform block, this model merges information from the two modality branches via alternating modules of “self-attention-based intra-domain fusion” and “cross-attention-based inter-domain fusion” units.

The work in [27] proposed a Target-aware Dual Adversarial Learning for object detection. The idea is to exploit structural information in the IR image and textural details from the visible image to improve object detection. This is made possible by means of a generator and two discriminators that seek to retain relevant information from the two modalities.

2.5. Robustness to missing imaging modality

GTISR is subject to degraded performance when one of the inputs, e.g. the guiding RGB image, is missing at inference time. Little work in the literature addresses this issue directly for the thermal image SR task, but some studies have examined it in other application areas.

In [51], the authors evaluated the impact of the type of architecture, data augmentation and image fusion technique on action recognition performance in the case of a missing modality. They concluded that transformer-based fusion is more robust in this situation than feature summation or concatenation. Meanwhile, [31] studied the impact of a missing modality (text, audio, or image) in training or testing a GAN or autoencoder model. They proposed a Bayesian meta-learning framework to better manage missing modalities.

3. Method

In this paper, we propose a novel architecture, named SwinFuSR, as a contender for the PBVS 2024 TISR Track 2 challenge. The aim of this competition is to obtain a high-resolution (x8) infrared image from a low-resolution IR image and a medium-resolution RGB image.

3.1. Proposed architecture

As in many other super resolution transformer architectures [7, 24, 38], our own, illustrated in Fig. 1, is composed of three modules.

The first module extracts shallow features using convolutional layers followed by N Swin Transformer (STL) layers. The second module focuses on deep feature extraction. Its role is to extract characteristics that are useful to reconstruct the image by combining IR and RGB features. L Attention-guided Cross-domain Fusion (ACF) blocks are used to extract useful information from RGB and IR features. Then, concatenation and convolution are performed to merge the two branches. The third module carries out

deep feature reconstruction. It is composed of P Swin Transformer layers to refine the merged features and three convolution layers to return to image space.

In the first two modules, the architecture is divided into two branches, similarly to SwinFusion [30]: one dedicated to the RGB image and the other to the IR image. A bicubic interpolation is performed on the IR image so that its dimensions (height (h) and width (w)) match those of its paired RGB image. Inspired by [24, 50, 61], a skip connection from the interpolated IR image to the reconstructed image is introduced for faster convergence and better performance. This gives the network an initial solution to improve upon.

3.2. Loss function

As a loss function, we use a combination of two differentiable pixel losses commonly used to measure the similarity between two images:

- An L_1 loss (or MAE) allows for relatively stable convergence and avoids gradient explosion [59]:

$$L_1 = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (1)$$

with n the number of pixels, y_i the value of the i^{th} pixel in the ground-truth (GT) image and \hat{y}_i the value of the i^{th} pixel in the reconstruction.

- An L_2 loss (or MSE) is more sensitive to higher reconstruction errors but can make the reconstruction smoother at the expense of valuable high-frequency details:

$$L_2 = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (2)$$

The lower these two metrics are, the closer the reconstruction is to the GT.

We use the loss strategy proposed by GuidedSR, the winning solution in the 2023 PBVS challenge, described in [38].

$$Loss = \begin{cases} L_1 & \text{for the first } T \text{ epochs} \\ L_2 & \text{after} \end{cases} \quad (3)$$

This strategy allows us to obtain good convergence properties with an L_1 loss, then refine the optimization with an L_2 loss.

3.3. Training strategy

Specific training strategies can help build missing modality robustness into the model. The literature proposes two main ways to handle this. The first one is to remove the entire portion of the network dealing with the missing information; in that case, the modalities must be processed independently

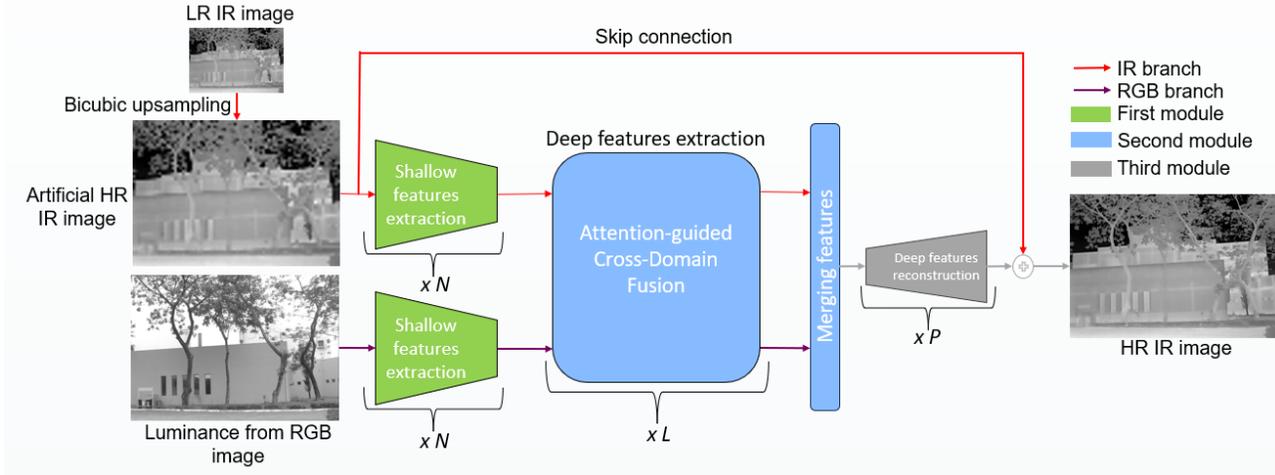


Figure 1. Architecture of the proposed SwinFuSR model.

as in CoRefusion[20]. The second and simpler method is to arbitrarily set the corresponding input values to the network to a fixed value such as zero.

To reduce performance degradation in the case of a missing modality, we propose a new model training regime that consists in randomly removing the training RGB images. More formally, at each training iteration, the input I to the network is given by:

$$I = \begin{cases} (I_{h,w}^{IR}, I_{h,w}^{RGB}) & \text{if } p < p_{th} \\ (I_{h,w}^{IR}, O_{h,w}) & \text{otherwise} \end{cases} \quad (4)$$

with $I_{h,w}^{RGB}$ the RGB image, $I_{h,w}^{IR}$ its corresponding IR image after bicubic interpolation, $O_{h,w}$ an all-zero image, p a random probability following a uniform distribution $\mathcal{U}(0, 1)$ and p_{th} a fixed threshold between 0 and 1.

4. Experiments

4.1. Implementation details

To train our model, we used the dataset provided for the second track of the PBVS 2024 TISR Challenge. It is composed of 700 training samples and 200 validation samples, each sample being a 640x448 IR image, along with its downsampled version by a factor of 8 and its paired 640x448 RGB image. The 100 testing samples are provided without the HR ground truths. These registered images were acquired by Balser (for RGB) and TAU2 (for IR) cameras and represent images of outdoor urban scenes. We evaluated our model’s performance on the training and validation sets using the PSNR and SSIM metrics.

Following common practice for training transformer architectures [24, 25, 30, 56], we used patches rather than the entire image as input. The patch size used was 128x128 and batch size was 16. The input patches were augmented with

random horizontal and vertical flips and random rotations. Pixel values were normalized between 0 and 1.

The number of heads, the window size and the embedding dimensions were set to 6, 9 and 60 respectively. We set the network module depths to $N = 2$, $L = 3$ and $P = 3$, according to the study detailed in Section 4.2.1 below.

For the training, the learning rate was set to 4×10^{-4} until $T = 3300$, then reduced to 1×10^{-4} for the remainder. We used the Adam optimizer. The run lasted 72 hours (4300 epochs) on two Tesla V100 GPUs with 32.0 GB of VRAM each.

4.2. Ablation study

4.2.1 Effect of the number of modules

To study the effect of the number of STL blocks (N), ACF blocks (L) and STL blocks (P) in the extraction, fusion and reconstruction modules respectively, we set as a baseline $N = 1, L = 2, P = 1$ as in the original SwinFusion paper [30]. Then, we increased for each module separately these values by 1 and by 2 and observed the effect on performance (PSNR and SSIM) (see Figure 2).

We can see that the increase in performance is most visible in the reconstruction module, suggesting that the latter is the network bottleneck. Increasing the number of modules in the extraction and fusion modules by 1 each also improves performance, but to a lesser degree. Based on these results, we set the numbers of modules to $N = 2, L = 3$ and $P = 3$ for the experiments in Section 5.1 below. For the remaining experiments, we set them to $N = 1, L = 2$ and $P = 1$ to limit required resources.

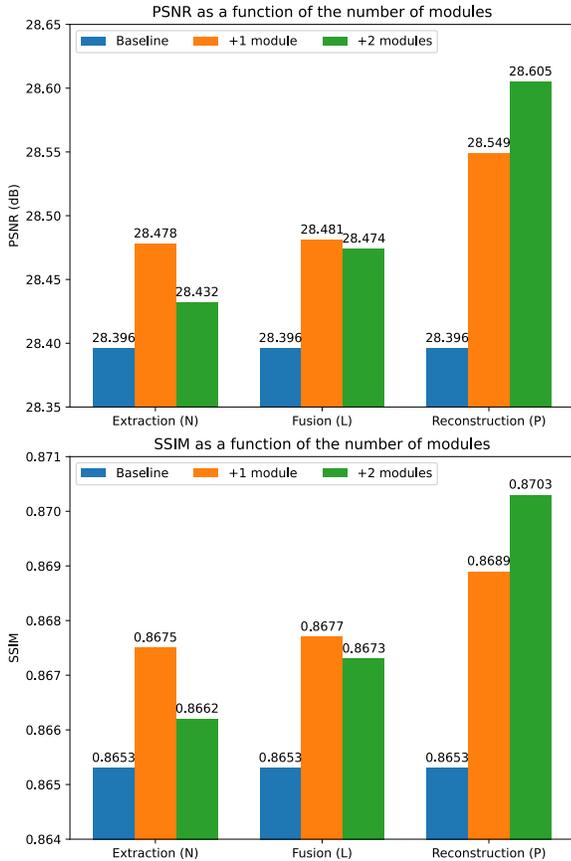


Figure 2. Effect of module depth on overall performance.

4.2.2 Effects of skip connection

In SR, it is common to use skip connections between an artificial upsampling or early feature extraction layer and the end of the network. We trained our model with and without the skip connection in our SwinFuSR model (Figure 1). Figure 3 shows the difference in performance.

The results demonstrate that using a skip connection improves the convergence speed of the model and improves final performance by 0.3%. It is important to note that this performance enhancement does not come at the cost of additional parameters in the model.

5. Results and discussion

5.1. RGB guided thermal image super-resolution

For a fair comparison between our solution and the existing methods GuidedSR and CoRefusion, we retrained the latter two models on the PBVS24 Track 2 dataset, using the same training setup as originally described in their respective papers [20, 38] (no pre-trained weights were available). Quantitative results are provided in Table 1.

Figure 4 provides some qualitative results for guided SR

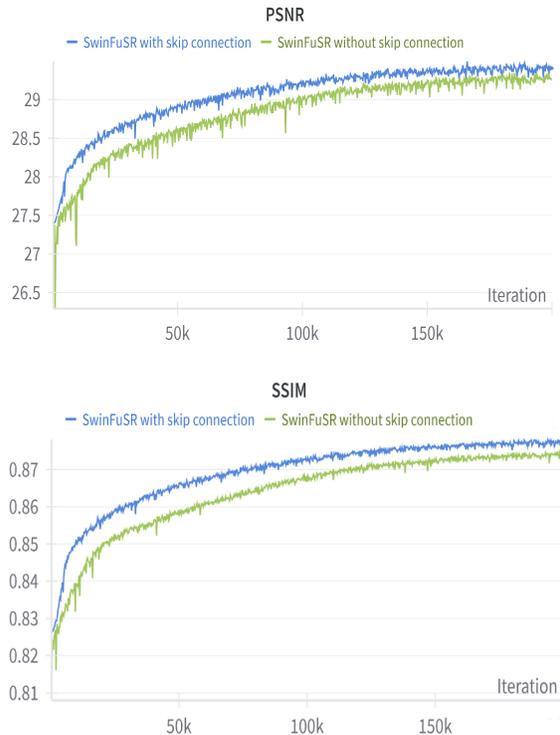


Figure 3. Performance with (blue) and without skip connection (green).

Method	PSNR	SSIM	#parameters
Bicubic	25.17	0.774	\emptyset
CoReFusion	27.27	0.835	46.31M
GuidedSR	27.22	0.834	116.35M
SwinFuSR (ours)	28.96	0.878	3.30M

Table 1. PSNR and SSIM on validation set.

on an image from the PBVS2024 challenge dataset. We can notice that SwinFuSR offers the closest output to the GT and seems clearer than the other 2 reconstructions.

To test our solution on different kinds of images and to verify generalization capabilities, we applied guided SR on images from the Simultaneously-collected multimodal Lying Pose (SLP) dataset [28]. This dataset is composed of low-resolution (120x160) infrared and RGB image pairs of adult subjects lying down in a hospital bed. Figure 5 shows the results of the x8 guided SR of an image from this dataset. Unfortunately no GT IR images of higher resolution are provided in SLP. Thus, we used the available images as is but could not compare the SR results to reference HR images. Qualitatively, all three SR solutions enhance the very low-quality original image. Nevertheless, the details of the hand generated by SwinFuSR seem to be the most accurate, even if the shape of the hand seems unrealistic.

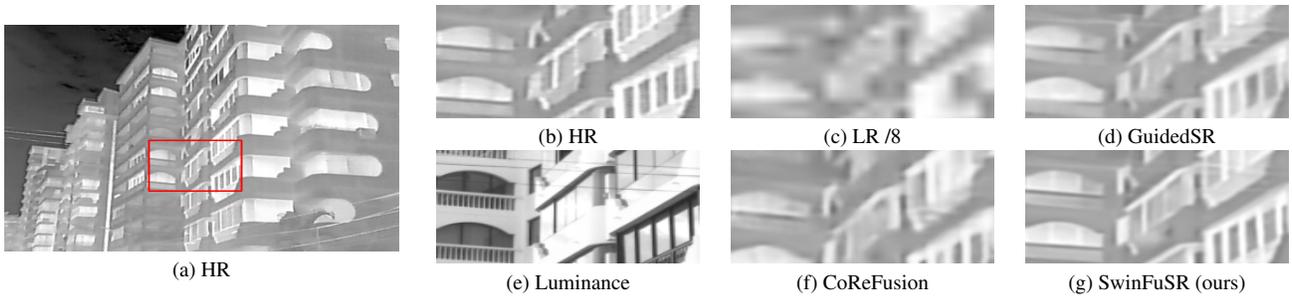


Figure 4. GTISR on image 292.01_D4 from PBVS 2024 Track- dataset.

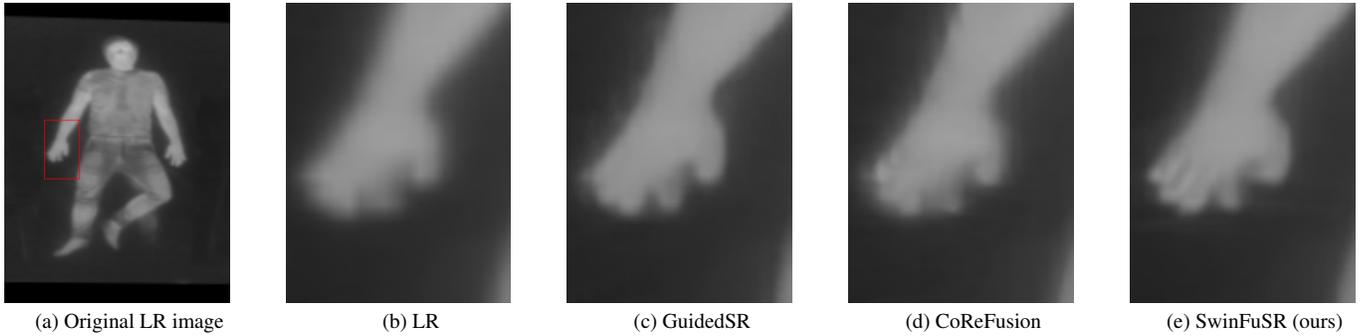


Figure 5. GTISR on sample image from SLP dataset [28].

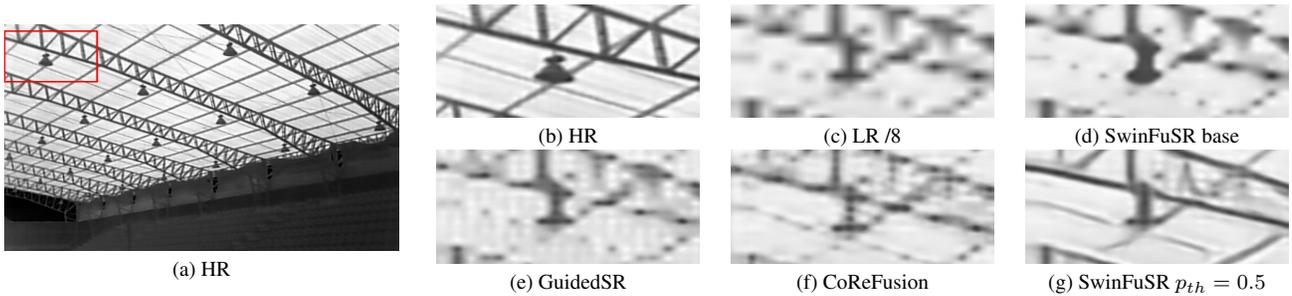


Figure 6. Unguided super resolution on image 044.02_D1 from PBVS 2024 Track 2 dataset.

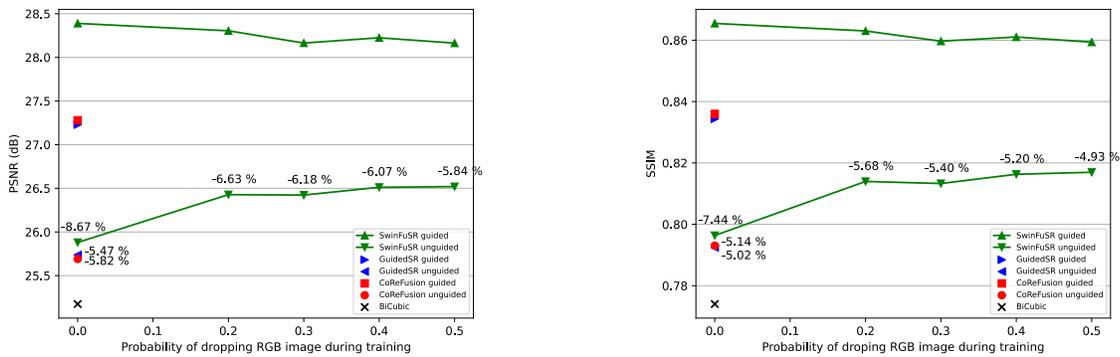


Figure 7. Effect of training parameter p_{th} on performance with (SwinFuSR guided) and without (SwinFuSR unguided) RGB input images at inference on the PBVS24 validation set.

5.2. Robustness to missing RGB modality

To evaluate our proposed training regime to improve robustness to missing RGB input, we trained our network five times, each with a different probability threshold p_{th} . Figure 7 illustrates the performance with and without RGB guide images at inference on the PBVS24 validation set. $p_{th} = 0$ means that during the training, all RGB guide images were used during the training.

First of all, we note that GuidedSR and CoReFusion have a smaller drop in performance than SwinFuSR when removing the RGB guide images. We can explain this by the fact that their baseline performance in guided SR is much lower than SwinFuSR.

Second, we see that increasing p_{th} substantially improves the performance of SwinFuSR when no guide image is used for inference (from -8.67% to -6.63% for PSNR and from -7.44% to -5.68% for SSIM) when p_{th} goes from 0 to 0.2, while only slightly reducing performance for guided SR in terms of both metrics. This result suggests that dropping RGB images during training with a certain probability enables a trade-off between maintaining good performance in guided SR and improving results in the absence of guide images.

Figure 6 confirms visually that this training strategy can increase performance in unguided SR. Indeed, Figure 6g is much clearer than Figure 6d.

5.3. Discussion

Our model is much smaller in terms of parameters than the two competing methods (CoReFusion and GuidedSR), but is slower at inference (1.3s to go from 80x56 to 640x448 running on a PC equipped with an RTX 3080 GPU and 12 GB of VRAM). This limitation restricts the use of SwinFuSR for real-time inference. Moreover, model selection (varying the number of blocks) is costly in terms of VRAM usage, and required us to run those experiments on a GPU cluster. These two drawbacks are probably due to the high proportion of transformers in the network, which are known to be particularly resource-hungry.

Another aspect to consider in order to efficiently use the proposed architecture on other datasets is the fact that the IR and RGB images must be registered. For this purpose, several algorithms are available, such as the one proposed in [13] or Elastix [22], the method used in the PBVS competition. In future work, we will study the robustness of the proposed model to IR-RGB registration errors.

6. Conclusion

This article proposes a new method for RGB guided thermal image super resolution. Our solution, named SwinFuSR, was submitted to Track 2 of the PBVS 2024 Thermal Image Super-Resolution Challenge and achieved better quali-

tative and quantitative results than other state-of-the-art architectures. We also present a novel training strategy that improves robustness to missing guide images at inference time. By randomly dropping a portion of the RGB images during training, the model's performance in unguided SR improves significantly compared to the guided SR baseline. In future work, we will explore how to make better use of the RGB image data, for instance by generating pseudo-IR images. In addition, we will examine how super resolution can improve the performance of related tasks such as estimating in-bed human pose.

7. Acknowledgments

We thank Philippe Debanné for his valuable help in editing this paper. The project was supported by L. Seoud's NSERC Discovery grant. This research was enabled in part by support provided by the Digital Research Alliance of Canada (alliancecan.ca).

References

- [1] Kiyoharu Aizawa, Takashi Komatsu, and Takahiro Saito. Acquisition of very high resolution images using stereo cameras. In *Visual Communications and Image Processing'91: Visual Communication*, pages 318–328. SPIE, 1991. 2
- [2] Khaled Bayouh, Raja Knani, Fayçal Hamdaoui, and Abdelatif Mtibaa. A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets. *The Visual Computer*, 38(8):2939–2970, 2022. 2
- [3] Vincent Boivin, Mana Shahriari, Gaspar Faure, Simon Mel-lul, Edem Donatien Tiassou, Philippe Jovet, and Rita Noumeir. Multimodality video acquisition system for the assessment of vital distress in children. *Sensors*, 23(11):5293, 2023. 1
- [4] Armelle Bridier, Monisha Shcherbakova, Atsushi Kawaguchi, Nancy Poirier, Carla Said, Rita Noumeir, and Philippe Jovet. Hemodynamic assessment in children after cardiac surgery: A pilot study on the value of infrared thermography. *Frontiers in Pediatrics*, 11, 2023. 1
- [5] Yang Cao, Guo Long Li, Yuan Kai Luo, Qi Pan, and Shao Ying Zhang. Monitoring of sugar beet growth indicators using wide-dynamic-range vegetation index (wdrvi) derived from uav multispectral images. *Computers and Electronics in Agriculture*, 171:105331, 2020. 1
- [6] Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. Simple Baselines for Image Restoration, 2022. arXiv:2204.04676 [cs]. 2
- [7] Xiangyu Chen, Xintao Wang, Jiantao Zhou, Yu Qiao, and Chao Dong. Activating more pixels in image super-resolution transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22367–22377, 2023. 2, 3
- [8] Deng Cheng-Zhi, Tian Wei, Chen Pan, Wang Sheng-Qian, Zhu Hua-Sheng, and Hu Sai-Feng. Infrared image super-resolution via locality-constrained group sparse model. *Acta Physica Sinica*, 63(4):044202–044202, 2014. 2

- [9] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(2):295–307, 2015. [2](#)
- [10] Chao Dong, Chen Change Loy, and Xiaoou Tang. Accelerating the super-resolution convolutional neural network. In *Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part II 14*, pages 391–407. Springer, 2016. [2](#)
- [11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. [2](#)
- [12] Kefeng Fan, Kai Hong, and Fei Li. Infrared image super-resolution via progressive compact distillation network. *Electronics*, 10(24):3107, 2021. [2](#)
- [13] Nils Genser, Jürgen Seiler, and André Kaup. Camera Array for Multi-Spectral Imaging. *IEEE Transactions on Image Processing*, 29:9234–9249, 2020. Conference Name: IEEE Transactions on Image Processing. [7](#)
- [14] Yuchong Gu, Zitao Zeng, Haibin Chen, Jun Wei, Yaqin Zhang, Binghui Chen, Yingqin Li, Yujuan Qin, Qing Xie, Zhuoren Jiang, and Yao Lu. MedSRGAN: medical images super-resolution using generative adversarial networks. *Multimedia Tools and Applications*, 79(29-30):21815–21840, 2020. [1](#)
- [15] Paul M. Harvey, Joseph D. Adams, Terry L. Herter, George Gull, Justin Schoenwald, Luke D. Keller, James M. De Buizer, William Vacca, William Reach, and E. E. Becklin. First Science Results from SOFIA/FORCAST: Super-resolution Imaging of the S140 Cluster at 37 μm . , 749(2): L20, 2012. [1](#)
- [16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. [2](#)
- [17] Yongsong Huang, Zetao Jiang, Rushi Lan, Shaoqin Zhang, and Kui Pi. Infrared Image Super-Resolution via Transfer Learning and PSRGAN. *IEEE Signal Processing Letters*, 28: 982–986, 2021. Conference Name: IEEE Signal Processing Letters. [2](#)
- [18] Yongsong Huang, Tomo Miyazaki, Xiaofeng Liu, and Shinichiro Omachi. Infrared Image Super-Resolution: Systematic Review, and Future Trends, 2022. arXiv:2212.12322 [cs, eess]. [1](#), [2](#)
- [19] Andrey Ignatov, Radu Timofte, Maurizio Denna, Abdel Younes, Ganzorig Gankhuyag, Jingang Huh, Myeong Kyun Kim, Kihwan Yoon, Hyeon-Cheol Moon, Seung-ho Lee, Yoonsik Choe, Jinwoo Jeong, Sungjei Kim, Maciej Smyl, Tomasz Latkowski, Pawel Kubik, Michal Sokolski, Yujie Ma, Jiahao Chao, Zhou Zhou, Hongfan Gao, Zhengfeng Yang, Zhenbing Zeng, Zhengyang Zhuge, Chenghua Li, Dan Zhu, Mengdi Sun, Ran Duan, Yan Gao, Lingshun Kong, Long Sun, Xiang Li, Xingdong Zhang, Jiawei Zhang, Yaqi Wu, Jinshan Pan, Gaocheng Yu, Jin Zhang, Feng Zhang, Zhe Ma, Hongbin Wang, Hojin Cho, Steve Kim, Huaen Li, Yanbo Ma, Ziwei Luo, Youwei Li, Lei Yu, Zhihong Wen, Qi Wu, Haoqiang Fan, Shuaicheng Liu, Lize Zhang, Zhikai Zong, Jeremy Kwon, Junxi Zhang, Mengyuan Li, Nianxiang Fu, Guanchen Ding, Han Zhu, Zhenzhong Chen, Gen Li, Yuanfan Zhang, Lei Sun, Dafeng Zhang, Neo Yang, Fitz Liu, Jerry Zhao, Mustafa Ayazoglu, Bahri Batuhan Bilecen, Shota Hirose, Kasidis Arunruangsirilert, Luo Ao, Ho Chun Leung, Andrew Wei, Jie Liu, Qiang Liu, Dahai Yu, Ao Li, Lei Luo, Ce Zhu, Seongmin Hong, Dongwon Park, Joonhee Lee, Byeong Hyun Lee, Seunggyu Lee, Se Young Chun, Ruiyuan He, Xuhao Jiang, Haihang Ruan, Xinjian Zhang, Jing Liu, Garas Gendy, Nabil Sabor, Jingchao Hou, and Guanghui He. Efficient and accurate quantized image super-resolution on mobile npus, mobile ai & aim 2022 challenge: Report. In *Computer Vision – ECCV 2022 Workshops*, pages 92–129, Cham, 2023. Springer Nature Switzerland. [1](#)
- [20] Aditya Kasliwal, Pratinav Seth, Sriya Rallabandi, and Sanchit Singhal. CoReFusion: Contrastive Regularized Fusion for Guided Thermal Super-Resolution. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 507–514, Vancouver, BC, Canada, 2023. IEEE. [2](#), [4](#), [5](#)
- [21] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1646–1654, 2016. [2](#)
- [22] S. Klein, M. Staring, K. Murphy, M.A. Viergever, and J. Pluim. Elastix: A Toolbox for Intensity-Based Medical Image Registration. *IEEE Transactions on Medical Imaging*, 29(1):196–205, 2010. [7](#)
- [23] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017. [2](#)
- [24] Jingyun Liang, Jiezhong Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1833–1844, 2021. [2](#), [3](#), [4](#)
- [25] ShuBo Liang, Kechen Song, Wenli Zhao, Song Li, and Yunhui Yan. DASR: Dual-Attention Transformer for infrared image super-resolution. *Infrared Physics & Technology*, 133: 104837, 2023. [2](#), [4](#)
- [26] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 136–144, 2017. [2](#)
- [27] Jinyuan Liu, Xin Fan, Zhanbo Huang, Guanyao Wu, Risheng Liu, Wei Zhong, and Zhongxuan Luo. Target-aware Dual Adversarial Learning and a Multi-scenario Multi-Modality Benchmark to Fuse Infrared and Visible for Object Detection. In *2022 IEEE/CVF Conference on Computer Vision*

- and *Pattern Recognition (CVPR)*, pages 5792–5801, New Orleans, LA, USA, 2022. IEEE. 3
- [28] Shuangjun Liu, Yu Yin, and Sarah Ostadabbas. In-bed pose estimation: Deep learning with shallow dataset. *IEEE Journal of Translational Engineering in Health and Medicine*, 7: 1–12, 2019. 5, 6
- [29] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 2
- [30] Jiayi Ma, Linfeng Tang, Fan Fan, Jun Huang, Xiaoguang Mei, and Yong Ma. Swinfusion: Cross-domain long-range learning for general image fusion via swin transformer. *IEEE/CAA Journal of Automatica Sinica*, 9(7):1200–1217, 2022. 3, 4
- [31] Mengmeng Ma, Jian Ren, Long Zhao, S. Tulyakov, Cathy Wu, and Xi Peng. Smil: Multimodal learning with severely missing modality. *ArXiv*, abs/2103.05677, 2021. 3
- [32] Xiao-Jiao Mao, Chunhua Shen, and Yu-Bin Yang. Image restoration using convolutional auto-encoders with symmetric skip connections. *arXiv preprint arXiv:1606.08921*, 2016. 2
- [33] Sen Pan, Si-Bao Chen, and Bin Luo. A super-resolution-based license plate recognition method for remote surveillance. *Journal of Visual Communication and Image Representation*, 94:103844, 2023. 1
- [34] Cheng Peng, S. Kevin Zhou, and Rama Chellappa. DA-VSR: Domain Adaptable Volumetric Super-Resolution For Medical Images, 2022. *arXiv:2210.05117 [cs, eess]*. 1
- [35] Haixia Qi, Bingyu Zhu, Zeyu Wu, Yu Liang, Jianwen Li, Leidi Wang, Tingting Chen, Yubin Lan, and Lei Zhang. Estimation of peanut leaf area index from unmanned aerial vehicle multispectral images. *Sensors*, 20(23), 2020. 1
- [36] Defu Qiu, Yuhu Cheng, and Xuesong Wang. Medical image super-resolution reconstruction algorithms based on deep learning: A survey. *Computer Methods and Programs in Biomedicine*, 238:107590, 2023. 1
- [37] Seunghyeon Rhee and Moon Gi Kang. Discrete cosine transform based regularized high-resolution image reconstruction algorithm. *Optical Engineering*, 38(8):1348–1356, 1999. 2
- [38] Rafael E. Rivadeneira, Angel D. Sappa, Boris X. Vintimilla, Jin Kim, Dogun Kim, Zhihao Li, Yingchun Jian, Bo Yan, Leilei Cao, Fengliang Qi, Hongbin Wang, Rongyuan Wu, Lingchen Sun, Yongqiang Zhao, Lin Li, Kai Wang, Yicheng Wang, Xuanming Zhang, Huiyuan Wei, Chonghua Lv, Qigong Sun, Xiaolin Tian, Zhuang Jia, Jiakui Hu, Chenyang Wang, Zhiwei Zhong, Xianming Liu, and Junjun Jiang. Thermal Image Super-Resolution Challenge Results - PBVS 2023. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 417–425, New Orleans, LA, USA, 2022. IEEE. 2, 3, 5
- [39] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015. 2
- [40] Fatih Mehmet Senalp and Murat Ceylan. A new approach for super-resolution and classification applications on neonatal thermal images. *Quantitative InfraRed Thermography Journal*, pages 1–18, 2023. 1
- [41] Monisha Shcherbakova, Rita Noumeir, Michaël Levy, Armelle Bridier, Victor Lestrade, and Philippe Jovet. Optical thermography infrastructure to assess thermal distribution in critically ill children. *IEEE Open Journal of Engineering in Medicine and Biology*, PP:1–1, 2021. 1
- [42] Huanfeng Shen, Liangpei Zhang, Bo Huang, and Pingxiang Li. A map approach for joint motion estimation, segmentation, and super resolution. *IEEE Transactions on Image processing*, 16(2):479–490, 2007. 2
- [43] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1874–1883, 2016. 2
- [44] Patricia L. Suárez, Dario Carpio, and Angel D. Sappa. Enhancement of guided thermal image super-resolution approaches. *Neurocomputing*, 573:127197, 2024. 2
- [45] Radu Timofte, Eirikur Agustsson, Luc Van Gool, Ming-Hsuan Yang, Lei Zhang, Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, Kyoung Mu Lee, Xintao Wang, Yapeng Tian, Ke Yu, Yulun Zhang, Shixiang Wu, Chao Dong, Liang Lin, Yu Qiao, Chen Change Loy, Woong Bae, Jaejun Yoo, Yoseob Han, Jong Chul Ye, Jae-Seok Choi, Munchurl Kim, Yuchen Fan, Jiahui Yu, Wei Han, Ding Liu, Haichao Yu, Zhangyang Wang, Honghui Shi, Xinchao Wang, Thomas S. Huang, Yunjin Chen, Kai Zhang, Wangmeng Zuo, Zhimin Tang, Linkai Luo, Shaohui Li, Min Fu, Lei Cao, Wen Heng, Giang Bui, Truc Le, Ye Duan, Dacheng Tao, Ruxin Wang, Xu Lin, Jianxin Pang, Jinchang Xu, Yu Zhao, Xiangyu Xu, Jinshan Pan, Deqing Sun, Yujin Zhang, Xibin Song, Yuchao Dai, Yueying Qin, Xuan-Phung Huynh, Tiantong Guo, Hwijat Seyed Mousavi, Tiej Huu Vu, Vishal Monga, Cristovao Cruz, Karen Egiazarian, Vladimir Katkovnik, Rakesh Mehta, Arnav Kumar Jain, Abhinav Agarwalla, Ch V. Sai Praveen, Ruofan Zhou, Hongdiao Wen, Che Zhu, Zhiqiang Xia, Zhengtao Wang, and Qi Guo. Ntire 2017 challenge on single image super-resolution: Methods and results. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1110–1121, 2017. 2
- [46] Joaquim Torra, Felipe Viela, Diego Megías, Begonia Sot, and Cristina Flors. Versatile near-infrared super-resolution imaging of amyloid fibrils with the fluorogenic probe cranad-2. *Chemistry - A European Journal*, 28, 2022. 1
- [47] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2
- [48] Jing Wang, Jason F Ralph, and John Y Goulermas. An analysis of a robust super resolution algorithm for infrared imaging. In *2009 Proceedings of 6th International Symposium on Image and Signal Processing and Analysis*, pages 158–163. IEEE, 2009. 2
- [49] Shenlong Wang, Lei Zhang, Yan Liang, and Quan Pan. Semi-coupled dictionary learning with applications to image super-resolution and photo-sketch synthesis. In *2012*

- IEEE Conference on computer vision and pattern recognition*, pages 2216–2223. IEEE, 2012. 2
- [50] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European conference on computer vision (ECCV) workshops*, pages 0–0, 2018. 2, 3
- [51] Sangmin Woo, Sumin Lee, Yeonju Park, Muhammad Adi Nugroho, and Changick Kim. Towards good practices for missing modality robust action recognition, 2023. 3
- [52] Najia Yakob, Sandrine Laliberte, Philippe Doyon-Poulin, Philippe Jouvret, and Rita Noumeir. Data representation structure to support clinical decision-making in the pediatric intensive care unit: Interview study and preliminary decision support interface design. *JMIR Formative Research*, 8, 2024. 1
- [53] Masayuki Yamaguchi, Kazunori Akiyama, Takashi Tsukagoshi, Takayuki Muto, Akimasa Kataoka, Fumie Tazaki, Shiro Ikeda, Misato Fukagawa, Mareki Honma, and Ryohei Kawabe. Super-resolution imaging of the protoplanetary disk hd 142527 using sparse modeling. *The Astrophysical Journal*, 895(2):84, 2020. 1
- [54] Jianchao Yang, Zhaowen Wang, Zhe Lin, Scott Cohen, and Thomas Huang. Coupled dictionary training for image super-resolution. *IEEE transactions on image processing*, 21(8):3467–3478, 2012. 2
- [55] Yifan Yang, Qi Li, Chenwei Yang, Yannian Fu, Huajun Feng, Zhihai Xu, and Yueting Chen. Deep networks with detail enhancement for infrared image super-resolution. *IEEE Access*, 8:158690–158701, 2020. 2
- [56] Dafeng Zhang, Feiyu Huang, Shizhuo Liu, Xiaobing Wang, and Zhezhu Jin. SwinFIR: Revisiting the SwinIR with Fast Fourier Convolution and Improved Training for Image Super-Resolution, 2023. arXiv:2208.11247 [cs]. 2, 4
- [57] Haichao Zhang, Yanning Zhang, and Thomas S Huang. Efficient sparse representation based image super resolution via dual dictionary learning. In *2011 IEEE International Conference on Multimedia and Expo*, pages 1–6. IEEE, 2011. 2
- [58] Haikun Zhang, Yueli Hu, and Ming Yan. Thermal Image Super-Resolution Based on Lightweight Dynamic Attention Network for Infrared Sensors. *Sensors*, 23(21):8717, 2023. 2
- [59] Lanfeng Zhou and Shuaijie Feng. A review of deep learning for single image super-resolution. In *2019 International Conference on Intelligent Informatics and Biomedical Sciences (ICIIBMS)*, pages 139–142, 2019. 3
- [60] Yan Zou, Linfei Zhang, Qian Chen, Bowen Wang, Yan Hu, and Yuzhen Zhang. An infrared image super-resolution imaging algorithm based on auxiliary convolution neural network. In *Other Conferences*, 2020. 2
- [61] Yan Zou, Linfei Zhang, Chengqian Liu, Bowen Wang, Yan Hu, and Qian Chen. Super-resolution reconstruction of infrared images based on a convolutional neural network with skip connections. *Optics and Lasers in Engineering*, 146:106717, 2021. 2, 3