Electronic Letters on Computer Vision and Image Analysis 24(2):355-372, 2025

Thermal Image Super-Resolution: Trends and Challenges

Rafael E. Rivadeneira[1], Angel D. Sappa[1,2] and Riad Hammoud[3]

[1] *Escuela Superior Politécnica Del Litoral, ESPOL, Campus Gustavo Galindo Km. 30.5 Vía Perimetral, Guayaquil, Ecuador*
[2] *Computer Vision Center, Campus UAB, 08193 Bellaterra, Barcelona, Spain*
[3] *Plus.ai, California, USA*

Received 10th of September, 2025; accepted 2nd of December, 2025

Abstract

Thermal Image Super-Resolution has become pivotal for security, autonomous driving, industrial inspection, and surveillance. This work consolidates six editions of the TISR challenge held within the Perception Beyond the Visible Spectrum workshop at CVPR from 2020 to 2025, detailing the evolution of tasks, datasets, evaluation protocols, and participation. The analysis traces a methodological shift from convolutional neural networks to transformer-based and hybrid architectures that better capture long-range dependencies, besides the emergence of cross-spectral guidance that leverages visible imagery to enhance thermal detail at large scale factors. Quantitative trends in PSNR and SSIM across editions show consistent improvement in results. Remaining challenges include robust cross-spectral alignment, computational efficiency for resource-constrained deployment, broader dataset diversity across conditions, and resilience to noise and environmental variation. The synthesis provides a unified reference for benchmarking progress and outlines actionable directions for future advances in thermal image reconstruction.

Key Words: Thermal imaging, Super-resolution, Cross-spectral guidance, Deep learning, Transformers, Benchmark datasets.

1 Introduction

Image Super-Resolution (SR) methods typically enhance image resolution on given low-resolution inputs. While SR in the visible spectrum has advanced rapidly, thermal imaging presents unique challenges because of its inherently low resolution, sensitivity to environmental conditions, and fundamental differences from visible spectrum images.

Thermal vision supports several real-world tasks where visible imaging is limited. In security and surveillance, thermal cameras provide reliable detection under poor illumination, smoke or fog; super-resolution helps recover sharper contours for tracking, re-identification priors, and incident analysis [1]. In autonomous driving, thermal cues complement RGB and LiDAR at night or in adverse weather; higher thermal detail improves road-side object delineation and may reduce false alarms in perception stacks [2]. In industrial inspection and

Correspondence to: rrivaden@espol.edu.ec

Recommended for acceptance by Josep Lladós
https://doi.org/10.5565/rev/elcvia.2374
ELCVIA ISSN:15108-5097
Published by Computer Vision Center / Universitat Autònoma de Barcelona, Barcelona, Spain

predictive maintenance, small thermal anomalies indicate overheating in machinery, power lines or pipelines; super-resolution facilitates early fault localization on low-cost sensors [3]. In clinical screening and physiological monitoring, thermal patterns correlate with perfusion or inflammation; enhancing thermal detail may aid delineation of vascular or inflammatory regions (with appropriate clinical validation) [4]. These scenarios illustrate where accurate thermal SR is useful, especially at large scaling factors where context from guided modalities and robust reconstruction are critical.

Recognizing these challenges, the Thermal Image Super-Resolution (TISR) challenge began in 2020 as part of the Perception Beyond the Visible Spectrum (PBVS) workshop at the Conference on Computer Vision and Pattern Recognition (CVPR). This event quickly became an essential reference point for researchers globally, attracting increased participation every year. Across six years, the TISR challenges made significant progress in methods, datasets, and overall thermal imaging performance.

Throughout these editions, various datasets were introduced. Initially, the datasets featured thermal images captured at different resolutions (low, mid, and high resolution) with added noise conditions [5]. In later editions, especially from 2024, the challenges employed the Cross-spectral Image Dataset for Image Super-resolution (CIDIS) [6], consisting of pairs of thermal images and high-resolution RGB images, to facilitate guided super-resolution approaches. This dataset evolution supported considerable improvements in super-resolution accuracy and robustness. Methodological trends also shifted notably during these years, transitioning from classical CNN approaches to more advanced transformer-based and hybrid convolutional-transformer architectures. This transition aligns with broader advancements in the field of deep learning.

This paper reviews these six years of PBVS TISR challenges, discussing progression, technical trends, dataset development, and ongoing issues. By synthesizing insights from the challenge results annually, the review aims to offer a consolidated overview and guide future directions for thermal image super-resolution research.

2 Background and Challenge Overview

Image Super-Resolution is a technique aimed at reconstructing high-resolution images from low-resolution inputs. Early SR methods relied primarily on classical image processing algorithms, such as interpolation and reconstruction from multiple images. However, with the emergence of deep learning, SR has significantly improved, particularly within the visible spectrum. Deep learning-based methods, including convolutional neural networks (CNNs), quickly became the dominant approach due to their superior performance in capturing complex patterns from data.

In the context of thermal images, SR poses unique challenges due to distinct physical properties of thermal radiation, sensor limitations, and lower inherent resolutions compared to visible images. Initial SR methods for thermal imagery adapted techniques from visible image SR but quickly evolved to consider specific characteristics of infrared data. Early editions of the TISR challenge (2020-2022) primarily utilized CNN-based methods, achieving incremental improvements each year [7, 8, 9].

From 2023, a significant methodological shift occurred, moving towards transformer-based architectures. Transformers have shown remarkable success in capturing global context and intricate dependencies within images, which significantly benefits thermal SR tasks [10]. Additionally, hybrid models combining CNN and transformer architectures emerged, effectively leveraging the strengths of both methods [11, 12].

Regarding datasets, initial challenges utilized datasets containing images at multiple resolutions (low, medium, high) with varying noise conditions, designed to simulate real-world scenarios [7, 8, 9]. These datasets were instrumental in establishing early benchmarks and provided a robust evaluation framework. However, recognizing the importance of cross-spectral guidance, the CIDIS dataset was introduced in 2024 [6], featuring thermal and RGB image pairs captured simultaneously. This dataset allowed for guided super-resolution techniques that significantly advanced performance benchmarks [11].

Evaluation metrics for TISR challenges consistently employed Peak Signal-to-Noise Ratio (PSNR) and

Structural Similarity Index (SSIM) [13], providing standardized and objective measures of performance across different methodologies and datasets.

This background sets the foundation for detailed analysis and comparison of methods and results in subsequent sections, providing a comprehensive understanding of the advancements and current state in thermal image super-resolution research.

2.1 Overview

The Thermal Image Super-Resolution challenges have consistently evolved since their inception at the PBVS workshop in CVPR 2020, becoming a central event for benchmarking advancements in thermal imaging technology. Each year, the challenges have featured improvements in datasets, evaluation methodologies, and participant engagement.

TISR - PBVS 2020. The inaugural challenge introduced a novel dataset with 1021 thermal images captured at three distinct resolutions (low, mid, high) using different thermal cameras. Two evaluation tasks were proposed: (i) SR from downsampling and noisy LR/MR/HR at varying factors ($\times 2$, $\times 3$, $\times 4$), and (ii) MR-to-HR SR at $\times 2$ assessed against semi-registered HR images. A total of 51 participants initially registered, with 6 teams reached the final validation phase [7].

TISR - PBVS 2021. The second edition of the challenge continued using the original dataset, specifically emphasizing mid-resolution and high-resolution image sets. The evaluation methodology remained consistent with the previous edition, employing two separate evaluations: (i) Evaluation 1 measured just $\times 4$ SR from a downsampled noisy high-resolution images; (ii) Evaluation 2 measured $\times 2$ SR task from mid-resolution images compared to semi-registered HR computed on a 90% central crop. In this edition, 9 team reached the final phase and improved performance metrics (PSNR, SSIM) compared to the previous edition, indicating methodological advancements [8].

TISR - PBVS 2022. The third TISR challenge maintained both evaluations and the same thermal-only dataset from earlier editions. Nearly 100 teams registered, and more than 50 submitted results, showcasing increasing global interest. Evaluation protocols were aligned with 2021, except the central crop for MR$\rightarrow$HR was set to 80%. The competition demonstrated further incremental improvements in image reconstruction accuracy, highlighting advancements in CNN-based super-resolution methods popular at that time [9].

TISR - PBVS 2023. In its fourth edition, the challenge introduced a significant change by adding a new dataset comprising 200 cross-spectral pairs of thermal and visible images. Track-1 retained the traditional thermal-only approach using the original dataset, while Track-2 involved guided super-resolution at $\times 8$ using registered visible and thermal image pairs without noise from new dataset. More than 150 teams participated, indicating growing community interest and methodological diversity. This edition significantly emphasized cross-spectral registration challenges, broadening the scope and complexity of evaluations [10].

TISR - PBVS 2024. The fifth edition marked another substantial evolution by introducing the Cross-spectral Image Dataset for Image Super-resolution (CIDIS). This dataset contain 1000 pairs registered of thermal and high-resolution RGB images, standardized to 640$\times$480 pixels resolution and split into 700/200/100 (train/val/test). This new dataset facilitated guided super-resolution at scale factors of $\times 8$ and $\times 16$, vastly expanding the technical possibilities and evaluation robustness. Track-1 evaluated single-image thermal SR at $\times 8$; Track-2 evaluated guided thermal SR at $\times 8$ and $\times 16$. Over 175 teams participated, and the event set new performance benchmarks. Transformer-based architectures started gaining notable prominence from this edition onward [11].

TISR - PBVS 2025. In the sixth and latest edition, the CIDIS dataset remained central to the challenge, featuring the same structure of 1000 thermal-RGB image pairs for consistency. The competition maintained two tracks: single thermal image super-resolution ($\times 8$ factor) and guided thermal image super-resolution ($\times 8$ and $\times 16$ factors). This edition attracted the highest participation yet, with 214 teams across both tracks. Innovations in transformer-based and hybrid convolutional-transformer models dominated, significantly improving thermal image reconstruction accuracy. A detailed comparative analysis with previous editions demonstrated substantial methodological progress, solidifying transformer-based methods as state-of-the-art solutions in thermal SR [12].

2.2 Summary of Evolution

Throughout these six editions, the PBVS TISR challenge has evolved substantially in terms of technical complexity, methodological innovation, dataset quality, and participant engagement. Early editions relied heavily on classical CNN-based approaches, gradually shifting towards transformer-based methods and guided cross-spectral super-resolution techniques. The consistent use of standard evaluation metrics (PSNR and SSIM) allowed clear tracking of progress and facilitated comparative analyses across different methods and datasets.

The original Thermal-only benchmarks (2020–2022) comprised 3063 thermal images captured with three cameras and used two evaluation protocols. Evaluation 1 trained on DS+Gaussian noise (10%) with scale factors depending on the source camera; Evaluation 2 evaluated MR$\rightarrow$HR at $\times 2$ with feature-based registration and central-crop metrics [7, 8, 9]. The Cross-spectral benchmarks (2023–2025) introduced registered RGB–thermal pairs for guided SR at $\times 8$ (160/30/10 splits) [10]. CIDIS (2024-–2025) scaled to 1000 registered pairs (700/200/100) at 640$\times$480, enabling SISR $\times 8$ and guided SR at $\times 8$ and $\times 16$ with consistent protocols and public CodaLab leaderboards [11, 12].

Table 1 condenses the salient elements of each edition. The introduction and continuous refinement of cross-spectral datasets such as CIDIS have significantly influenced methodological advancements, enabling the community to address more complex scenarios involving multi-modal inputs and cross-spectral alignment challenges. The annual increase in participation and quality of submissions reflects the community's robust engagement and underscores the sustained relevance of the TISR challenges.

Challenge websites (CodaLab). For transparency and reproducibility, the public leaderboards and evaluation bundles are available on CodaLab for the most recent editions:

- 2022: `https://codalab.lisn.upsaclay.fr/competitions/1990`
- 2023:
  - Track 1: `https://codalab.lisn.upsaclay.fr/competitions/9649`
  - Track 2: `https://codalab.lisn.upsaclay.fr/competitions/9666`
- 2024:
  - Track 1: `https://codalab.lisn.upsaclay.fr/competitions/17013`
  - Track 2: `https://codalab.lisn.upsaclay.fr/competitions/17014`
- 2025:
  - Track 1: `https://codalab.lisn.upsaclay.fr/competitions/21247`
  - Track 2: `https://codalab.lisn.upsaclay.fr/competitions/21248`

Year	Dataset	Tracks/Evals	Scales	Participation
2020 [7]	3 cams, 1021/img/cam	E1: DS+noise & E2: MR→HR	×2,×3,×4 & ×2	∼50 reg.; 6 final
2021 [8]	Same (MR, HR)	E1: HR ×4 & E2: MR→HR	×4 & ×2 (90% crop)	∼70 reg.; 9 final
2022 [9]	Same (MR, HR)	E1/E2 as 2021	×4 & ×2 (80% crop)	∼100 reg.; 5 final
2023 [10]	Thermal + cross-spectral	Track-1 (as 2022) — Track-2 guided	×4,×2 — ×8	∼150 reg.; 28/23 finals
2024 [11]	CIDIS (1000 pairs)	Track-1 (as 2023) — Track-2 guided	×8 — ×8,×16	∼175 reg.; 19/16 finals
2025 [12]	CIDIS (same)	Track-1/Track-2 (as 2024)	×8 — ×8, ×16	∼214 reg.; 34/14 finals

Table 1: Condensed overview of PBVS TISR editions. DS: downsampled; reg.: registered.



Figure 1: Illustrations of thermal images dataset most used for Challenges 2020–2023 [5].

3 Technical Analysis and Trends

This section presents a detailed technical analysis of the methodologies used throughout the six editions of the PBVS TISR challenges, highlighting significant trends and technological shifts in thermal image super-resolution.

3.1 Dataset Evolution

Datasets play a crucial role in benchmarking the effectiveness of super-resolution algorithms. Over six PBVS-CVPR editions, three dataset families were used: a thermal-only benchmark (2020–2022), a first cross-spectral set (2023), and the standardized Cross-spectral Image Dataset for Image Super-resolution (CIDIS, 2024–2025). This section details sensing hardware, image content, splits, and the degradation and registration protocols that define each benchmark.

Thermal-only benchmark (2020–2022). The original benchmark comprises 3063 thermal images captured with three distinct cameras at low-, mid-, and high-resolution settings, totaling 1021 images per camera. From each resolution, 951 images were used for training and 50 for testing; 20 images were held out for official evaluations [7]. Images include outdoor scenes across varied conditions and object categories, with the cameras mounted on a rig to minimize parallax such that viewpoints are almost registered [7, 8, 9]. Table 2 summarizes camera specifications reported by the organizers.

Degradation model and splits. Evaluation 1 uses bicubic downsampling with added Gaussian noise (σ=10%), at ×2 ×3 ×4 for LR/MR/HR respectively (2020), then consolidated to HR ×4 in 2021–2022. Evaluation 2 compares MR→HR at ×2 against semi-registered HR images from a different camera. Metrics are PSNR and SSIM. The cropping policy evolved from an unspecified central region (2020) to 90% (2021) and 80% (2022–2023) of the central crop for MR→HR [7, 8, 9, 10].

Set	Camera	Native res.	FOV	Focal length	# images
LR	Axis Domo P1290	160×120	35.4°	4 mm	1021
MR	Axis Q2901-E	320×240	35°	9 mm	1021
HR	FLIR FC-6320/FC-632O	640×512	32°	19 mm	1021

Table 2: Thermal-only benchmark cameras and specifications (HR images cropped to 640×480 for evaluation).



Figure 2: Illustrations of the cross-spectral dataset, thermal and visible registered images used in Track-2 for challenge 2023 [10].

Registration protocol. For MR→HR, 2020 used keypoint-based registration with SIFT [14], SURF [15] and ORB; subsequent editions retained feature-based alignment before computing metrics on the cropped region [7, 8, 9].

Scene content. The dataset covers most daily outdoor scenes with varied objects (vegetation, people, vehicles, buildings), enabling robust training and evaluation under realistic thermal conditions [8, 9]. Illustrative examples for 2020–2023 are shown in Fig. 1.

First cross-spectral dataset (2023). To enable guided thermal SR, 2023 introduced a registered visible–thermal dataset of 200 pairs captured under daylight. Images were acquired with a Basler visible camera and a FLIR TAU2 thermal camera and standardized to 640×480 after registration with Elastix [16]. The official split is 160/30/10 for train/val/test. Track-2 evaluates guided ×8 SR with no added noise; PSNR and SSIM are computed on the test pairs [10]. Figure 2 shows examples.

CIDIS cross-spectral dataset (2024–2025). CIDIS generalizes the 2023 setup to 1000 registered pairs with the same sensing configuration (Basler RGB and FLIR TAU2 thermal), all standardized at 640×480. Registration employs a pipeline that combines Elastix [16], MATLAB `imregister` [17], LightGlue [18], and Nemar-based [19] models to ensure sub-pixel alignment across modalities. The official split is 700/200/100 for train/val/test. The 100-image test set is partitioned as 20 images for Track-1 (thermal-only SISR at ×8) and 40+40 images for Track-2 guided SR at ×8 and ×16, respectively; downsampling is bicubic and no additional noise is added [11]. The 2025 edition maintains CIDIS and the same track definitions, and also reports auxiliary reference metrics (e.g., RMSE and perceptual proxies) alongside the primary PSNR/SSIM rankings [12]. Examples are depicted in Fig. 3.

3.2 Methodological Evolution

The methodologies in the PBVS TISR challenges evolved markedly. From 2020–2022, CNN-centric pipelines delivered steady gains through deeper residual designs, robust pixelwise losses, and improved augmentation [7,



Figure 3: Illustrations of the cross-spectral dataset, thermal and visible registered images used in challenges 2024 and 2025 [6].

8, 9]. In 2023, transformer architectures that capture long-range dependencies surpassed CNNs [10]. In 2024–2025, hybrid CNN–transformer models, often coupled with cross-spectral guidance, achieved the strongest PSNR/SSIM [11, 12].

2020–2022: CNN-centric pipelines for thermal-only SR. Early editions were dominated by convolutional neural networks tailored to the two official evaluations: (i) SR from synthetically downsampled and noisy inputs and (ii) MR→HR super-resolution with semi-registered pairs [7, 8, 9]. Typical backbones followed residual and multi-scale recipes widely adopted in SR:

- *Backbones and upsampling.* Residual-in-residual blocks, channel attention, and densely connected modules were common; upsampling was usually implemented with sub-pixel convolution (pixel-shuffle) or transposed convolution, sometimes in cascaded stages for large factors.
- *Loss functions.* L1 (or Charbonnier) remained the main reconstruction loss due to stable convergence and superior PSNR. Several teams added SSIM loss or gradient/edge consistency terms to preserve structural content under thermal noise. Adversarial and perceptual losses were seldom prioritized in rankings focused on PSNR/SSIM.
- *Data preparation.* Cropped patches (e.g., 128×128–256×256) and classical augmentations (flip/rotate) were standard. To address the synthetic degradation gap, some pipelines injected mild noise or blur during training beyond the official protocol.
- *Registration-awareness.* For MR→HR, feature-based pre-alignment was assumed by the protocol; models increasingly incorporated robust objectives (Charbonnier, huberized SSIM) to tolerate residual misalignment [8, 9].

2023: Introduction of cross-spectral guidance and first transformer backbones. The 2023 edition added a cross-spectral track with registered RGB–thermal pairs and evaluated guided $\times8$ SR [10]. This change accelerated the adoption of attention mechanisms and transformer-style modules:

- *Transformers for thermal-only and guided SR.* Self-attention improved modeling of long-range dependencies, particularly at high scale factors where context aggregation becomes critical. Lightweight windowed attention and hierarchical tokenization were used to balance capacity and efficiency.
- *Fusion strategies.* Three families emerged for RGB guidance: *early fusion* (concatenation of RGB and thermal inputs), *mid/late fusion* (separate encoders with feature fusion through cross-attention or gating),

and *pyramidal fusion* (multi-scale aggregation with shared decoders). Cross-attention offered stronger structure transfer from RGB while keeping thermal radiometry as the primary target modality.

- *Robustness to misregistration.* Although the dataset was registered, several solutions introduced deformable alignment layers, spatial transformers, or confidence masks to mitigate residual offsets, reducing edge ghosting and improving SSIM.

2024–2025: Standardization with CIDIS and consolidation of hybrid CNN–Transformer designs. With the introduction of CIDIS (1000 registered RGB–thermal pairs at 640×480 and the evaluation at $\times 8$ and $\times 16$, the landscape converged toward hybrid designs that combine locality-sensitive convolutions and global-context transformers [11, 12]:

- *Hybrid encoders and decoders.* Common patterns include convolutional stem layers to stabilize low-level statistics, transformer blocks for non-local reasoning, and convolutional refinement heads to sharpen details without introducing checkerboard artifacts. Windowed or shifted-window attention remained popular for memory efficiency at 640×480.

- *Guided fusion modules.* Cross-spectral fusion advanced with cross-attention and deformable cross-attention, modulated by edge or gradient priors extracted from RGB. Several solutions employed multi-branch encoders (thermal, RGB, and joint) with feature selection gates that downweight RGB in textureless or low-light regions to avoid transferring artifacts.

- *Loss design at large scale factors.* Composite objectives combined L1/Charbonnier with SSIM and frequency- or gradient-domain penalties to stabilize training at $\times 16$. Perceptual losses were used selectively as auxiliary terms to enhance structure without hurting PSNR.

- *Alignment-aware training.* Even with high-quality registration, training incorporated warping losses between RGB edges and the SR thermal prediction, sparse keypoint consistency, or patch-level spatial jitter to harden models against tiny residual misalignments.

- *Efficiency techniques.* Memory/computation reduction was addressed through low-rank attention, token pruning, mixed precision, sliding-window inference for large images, exponential moving average (EMA) of weights, and checkpoint averaging. Knowledge distillation from large teacher models to compact student backbones appeared in entries targeting real-time feasibility.

Common training and inference practices across editions. Regardless of architecture or track, several practices became widespread:

- *Optimization.* Adam/AdamW with cosine or step-wise decay schedules, warm-up epochs, and gradient clipping. Training commonly used tens to hundreds of epochs on cropped patches with moderate batch sizes due to memory constraints at high scale factors.

- *Normalization and color spaces.* Thermal inputs were normalized to fixed dynamic ranges before feeding the network. In guided settings, RGB was standardized with per-channel normalization; fusion took place in feature space rather than directly in the image domain.

- *Self-ensemble and TTA.* Test-time augmentation with flips and rotations and self-ensemble averaging provided consistent, albeit modest, gains on PSNR/SSIM, especially for $\times 8$ and $\times 16$.

- *Regularization and robustness.* Mixup/CutMix variants are uncommon in SR; instead, robustness relied on mild blur/noise injections during thermal-only training and on confidence-gated fusion under cross-spectral guidance.

Observed impact on accuracy. From 2020 to 2022, improvements were primarily attributed to deeper residual backbones, refined training recipes, and better handling of synthetic degradations [7, 8, 9]. The 2023 shift to attention-driven and fusion-based approaches delivered notable gains for the guided track at $\times 8$ [10]. In 2024 and 2025, hybrid CNN–Transformer architectures, alignment-aware objectives, and stronger fusion modules consolidated state-of-the-art results on CIDIS for both $\times 8$ and $\times 16$ guided SR while also improving single-image thermal SR at $\times 8$ [11, 12].

Takeaway. The methodological trajectory reflects a transition from local, single-modality restoration to global-context, cross-spectral reconstruction. Progress has depended on three levers: stronger backbones (transformers and hybrids), better exploitation of RGB guidance via attention and deformable alignment, and training strategies explicitly robust to residual misregistration and large upscaling factors.

3.3 Critical Analysis: method behavior and trade-offs

Gains across editions are explained by three factors. **(i) Context modeling.** At large scaling factors, long-range reasoning is vital; transformer blocks improve SSIM near structured regions where purely convolutional models tend to overfit local textures. **(ii) Guidance quality.** In guided tracks, RGB helps recover edges and geometry, but residual misregistration can transfer wrong structures. Deformable alignment, cross-attention, and confidence gating reduce ghosting and preserve edge continuity. **(iii) Objectives and stability.** At $\times 8$ and $\times 16$, combining L1/Charbonnier with SSIM and gradient/frequency penalties limits ringing; alignment-aware losses mitigate boundary artifacts. Systems ignoring these aspects underperform near high-contrast edges and motion.

3.4 Evaluation Metrics

The evaluation metrics remained consistent throughout all editions, employing PSNR and SSIM to objectively assess the performance of super-resolution methods, computed on sequestered test sets with edition-specific regions of interest and alignment procedures. PSNR is defined from the mean squared error (MSE) between the super-resolved output and the high-resolution (HR) reference as

$$\mathrm{PSNR} = 10 \log_{10} \left(\frac{\mathrm{MAX}^2}{\mathrm{MSE}} \right),$$

where MAX denotes the maximum possible pixel value; larger PSNR indicates higher fidelity. SSIM complements PSNR by assessing structural similarity (luminance, contrast, and structure terms), which correlates better with perceived quality in textured regions.

SSIM. Given two image patches x and y over a local window W, the Structural Similarity Index is

$$\mathrm{SSIM}(x, y) = \frac{(2\,\mu_x\,\mu_y + C_1)\,(2\,\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)\,(\sigma_x^2 + \sigma_y^2 + C_2)},$$

where μ_x and μ_y are local means, σ_x^2 and σ_y^2 are local variances, and σ_{xy} is the local covariance, all computed over a window W (typically a Gaussian window of size 11×11 with $\sigma{=}1.5$).

2020–2022 (thermal-only benchmark). Two evaluations were reported each year: (i) SR from bicubic downsampled HR images augmented with Gaussian noise (first evaluation) and (ii) MR→HR at $\times 2$ using semi-registered pairs from different cameras (second evaluation). PSNR/SSIM were averaged over the official test images to determine rankings [7, 8, 9]. Because the MR→HR task involves cross-camera comparisons, metrics were computed on a central cropped region to mitigate border and alignment artifacts. The cropping policy was standardized in 2021 to a 90% central crop and updated to 80% in 2022, a setting kept for the Track-1 MR→HR evaluation in 2023 [8, 9, 10].

2023 (first cross-spectral track). Track-1 preserved the thermal-only evaluations and PSNR/SSIM protocol (including the central-crop policy for MR→HR). The new Track-2 evaluated guided $\times 8$ SR on registered RGB–thermal pairs without added noise, with rankings again based on mean PSNR/SSIM over the test subset [10].

2024–2025 (CIDIS). With CIDIS, Track-1 considered single-image thermal SR at $\times 8$; Track-2 evaluated guided SR at $\times 8$ and $\times 16$, both without noise. The metrics that continued to be decided strictly by mean PSNR and SSIM across the corresponding test splits, enabling direct comparison across teams and years [11, 12]. The move to higher scale factors and precise cross-spectral registration led to noticeable gains in both metrics relative to pre-2024 editions.

Additional reference metrics (2025). The 2025 report also listed auxiliary indicators for analysis on the competition site—Root Mean Square Error (RMSE) and learned perceptual proxies labeled ALEX, VGG, and SQUEEZE—while preserving PSNR/SSIM as the only official criteria for ranking [12]. These reference metrics help interpret edge fidelity and perceptual consistency but did not affect leaderboards.

Reporting conventions and fairness controls. All editions computed PSNR/SSIM on thermal imagery (single-channel) using the official downsampling protocol and, when applicable, feature-based pre-alignment to limit bias from residual misregistration. The central-crop policy in early editions ensured that comparisons were made on reliable overlapping content. From 2024 onward, standardized registration in CIDIS reduced alignment-induced variance, improving metric stability at $\times 8$ and $\times 16$. Where relevant, organizers provided CodaLab [20] pages with per-team averages to ensure transparency and reproducibility.

3.5 Key Trends Summary

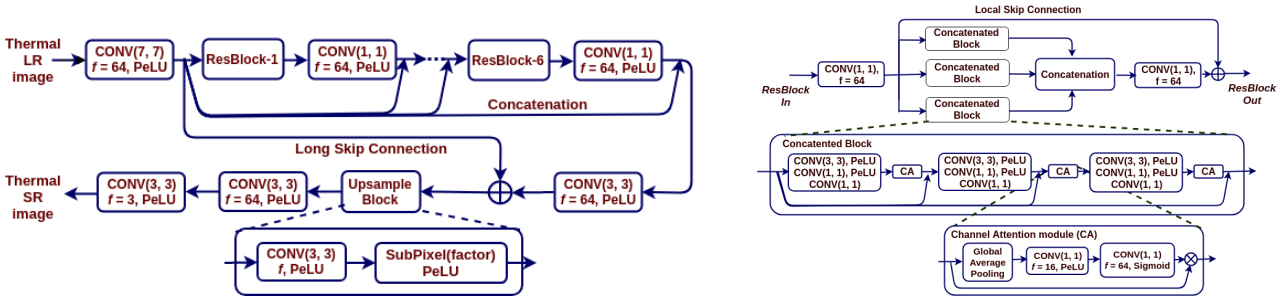
Across the six PBVS TISR editions, several trends became clear: methods shifted from pure CNNs to transformer-based and hybrid CNN–transformer architectures; dataset scale and quality increased, culminating in the introduction of CIDIS for standardized guided thermal super-resolution; evaluation scores in PSNR and SSIM improved steadily due to stronger models, better data, and refined protocols; and cross-spectral guidance gained prominence, leveraging visible-spectrum information to enhance thermal detail. Together, these developments indicate substantial progress and outline promising directions for future research and deployment in thermal image super-resolution.

4 Representative Methods

This section summarizes representative, top-ranked architectures reported throughout the PBVS TISR challenges, focusing on their core design choices (backbones, attention and fusion mechanisms, upsampling, and training strategies). The descriptions are based on the official challenge reports and publicly available technical summaries submitted by teams. For complete author attributions and implementation details, readers are referred to the corresponding annual references.

4.1 In evaluation1 $\times 4$ at PBVS2020

The contribution targets the thermal-only Evaluation 1 setting at $\times 4$ by designing a residual backbone with Inception-style parallel filtering [21] and channel attention [22]. Instead of relying on a single receptive field per block, the Inception-inspired residual unit aggregates features at multiple effective scales and then re-weights them channel-wise, encouraging the network to emphasize thermally informative responses while suppressing redundancies. Stability is addressed through two levels of skip connections: intra-block shortcuts that preserve low-level cues and ease gradient flow, and long-range shortcuts across groups of six blocks that directly bridge



(a) Proposed architecture. (b) ResBlock design.

Figure 4: Proposed architecture and ResBlock in in PBVS 2020 [7].

early and late layers to counteract vanishing or exploding gradients. PeLU activations [23] are used to retain nonlinearity while maintaining smooth gradients, and the spatial upscaling stage employs sub-pixel convolution to project features into the high-resolution grid (see Fig. 4a and 4b).

The implementation follows a straightforward training recipe aligned with the official protocol. The network is built in TensorFlow and optimized with Adam using an L_1 objective (learning rate 10^{-4}, β=0.5), trained for 50,000 iterations with batch size 4. Data augmentation consists of flips and rotations on training patches to promote invariance to canonical viewpoint changes. At inference, the upsampling module uses an internal channel factor f=64 for the $\times 4$ branch, mapping compact latent features into the reconstructed thermal image. Experiments were conducted on a workstation with an Intel Core i7-7700K, 32 GB RAM, and an NVIDIA GeForce GTX 1070 (8 GB), providing a reproducible reference environment for the reported pipeline.

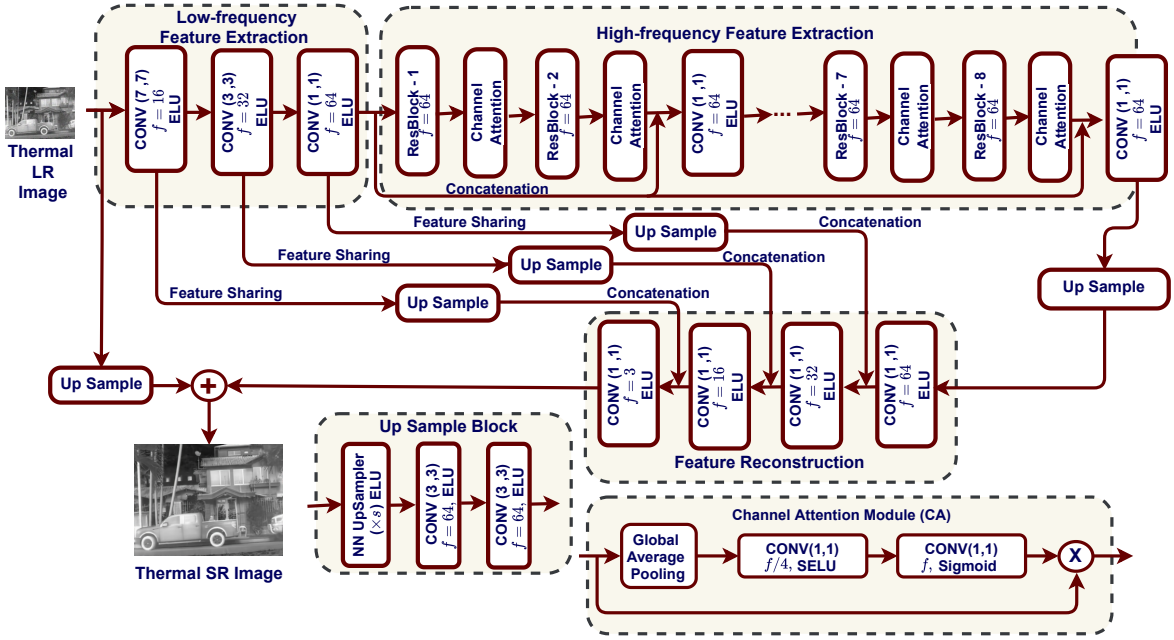
4.2 In evaluation1 x4 at PBVS2021

Designed for the thermal-only benchmark at $\times 2$ and $\times 4$, the architecture in Fig. 5(a) processes the LR thermal input through dedicated low- and high-frequency branches to extract salient content, using ELU activations for stable optimization [24]. Its core component is a compact residual block, shown in Fig. 5(b), that preserves high-frequency detail with a limited parameter budget. The block employs multi-branch convolutions with 1×1, 3×3, and 5×5 kernels to capture both local and broader structures, introduces depthwise convolution to reduce computation [25], and applies channel attention to adaptively re-scale channel-wise responses.

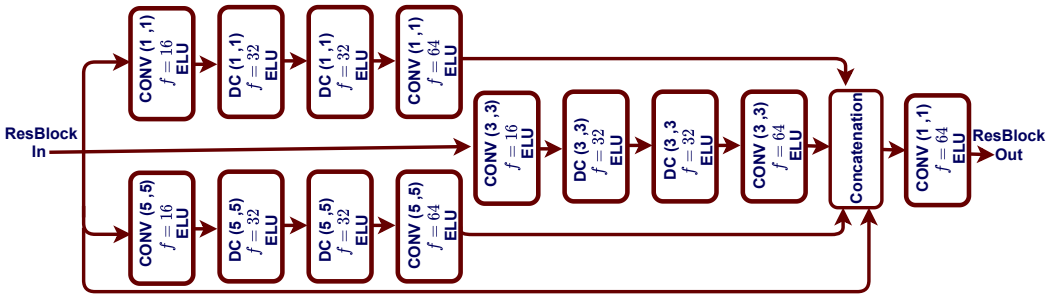
Upsampling adopts nearest-neighbor interpolation to the target scale ($\times 2$ or $\times 4$) followed by convolutional refinement, a choice that avoids checkerboard artifacts typically associated with transposed convolutions. A feature-sharing pathway connects intermediate representations to the reconstruction module, guiding the decoder toward semantically relevant structures and stabilizing gradients. Together, these components form a lightweight yet effective architecture for Evaluation 1 at $\times 4$, aligning efficiency with fidelity under the 2021 protocol.

4.3 In evaluation1 $\times 4$ at PBVS2022

As illustrated in Fig. 6, the Convolution Attached Transformer Super-resolution (CATS) network couples convolutional processing with transformer-based reasoning by interleaving *convolutional blocks (CB)* and *transformer blocks (TB)*. Each CB replaces the vanilla RCAB with the Detail-Fidelity Attention Module (DeFiAM) [26], which explicitly targets both low-frequency structural cues and high-frequency detail preservation. The pipeline first routes the thermal input through five successive DeFiAM-based CBs to stabilize statistics and enhance edges in a CNN-friendly regime. A transitional stage composed of a Swin-based TB [27] followed by a CB then bridges toward non-local modeling while maintaining robust low-level representations. This staging reduces optimization difficulty by letting early convolutions handle denoising and structure reinforcement before attention modules aggregate longer-range dependencies.



(a) The block schematic of the proposed architecture



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

Figure 5: Proposed architecture in PBVS 2021 [8].

Subsequently, a stack of five Swin TBs refines the feature tensor using windowed self-attention with hierarchical receptive fields, enabling efficient context capture at $\times 4$ without excessive memory growth. Features available before each block are concatenated along a long skip pathway, encouraging multi-depth aggregation and improving gradient flow. The reconstruction head upsamples via a dedicated *Re-Scale* module and maps latent features back to image space with a final convolution. Supervision relies on a pixel-wise mean squared error objective between the predicted super-resolved output and the high-quality reference, aligning the optimization with the PSNR-focused evaluation protocol.

4.4 In track2-guided at PBVS2023

For cross-spectral guided super-resolution in Track 2, the approach adopts a two-stream design that processes a low-resolution thermal image together with a high-resolution RGB counterpart. Each input first passes through a shallow feature extractor; the resulting features are concatenated and fed to a fusion trunk composed of cascaded NAF blocks [28], after which a reconstruction head produces the high-resolution thermal output (see Fig. 7). Training follows a two-stage schedule that stabilizes convergence and improves final fidelity: an initial phase optimized with an L_1 loss, followed by fine-tuning with an MSE objective. The reported setup uses PyTorch on two NVIDIA RTX 3090 GPUs with batch size 8 and 32×32 patches, reflecting a compact training

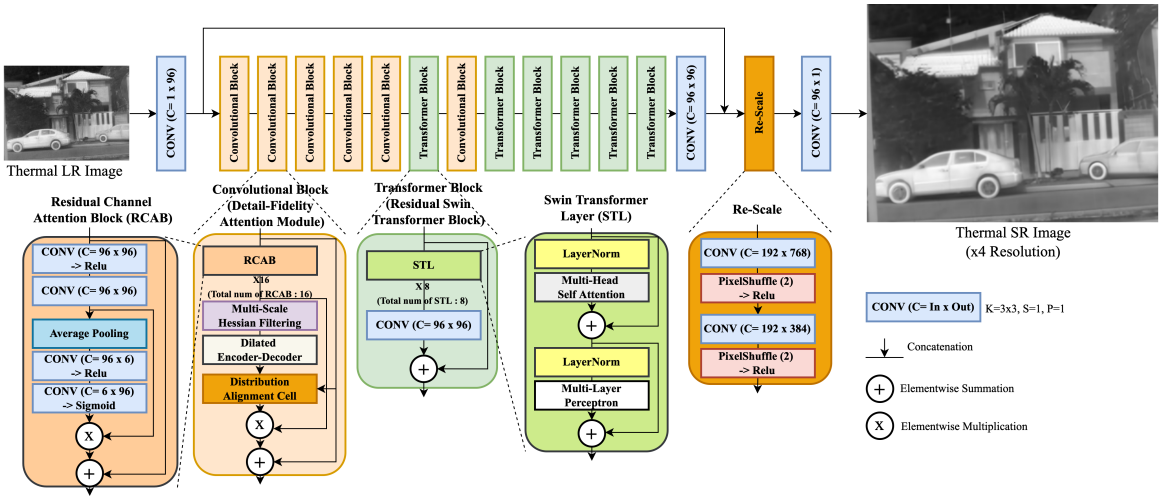


Figure 6: Architecture proposed in PBVS 2022 [9].

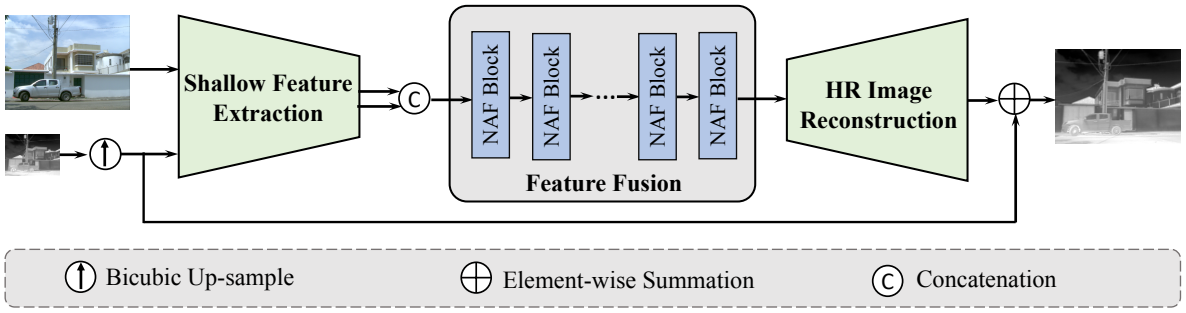
regime aligned with the Track 2 protocol.

A mixture-of-experts (MoE) mechanism is introduced to increase model diversity and robustness when transferring structure from RGB to thermal [29, 30, 31]. Here, the term *experts* refers to multiple, identical sub-models (built from NAFNet-style modules [28]) that process deterministic variants of the same paired input. Concretely, four experts receive the RGB–thermal pair under fixed rotations (e.g., $0°$, $90°$, $180°$, $270°$), each producing a super-resolved prediction in its own orientation; these predictions are then realigned to the canonical orientation and passed to a fifth expert that serves as a fusion module to integrate them (see Fig. 7). This design is conceptually similar to a built-in, rotation-based ensemble: by exposing sub-models to different canonical views, the system captures complementary geometric and textural evidence that a single pass might miss, while the fusion stage consolidates consistent details. The same strategy is applicable to both $\times 8$ and $\times 16$ guided evaluations under the standardized Track 2 settings.

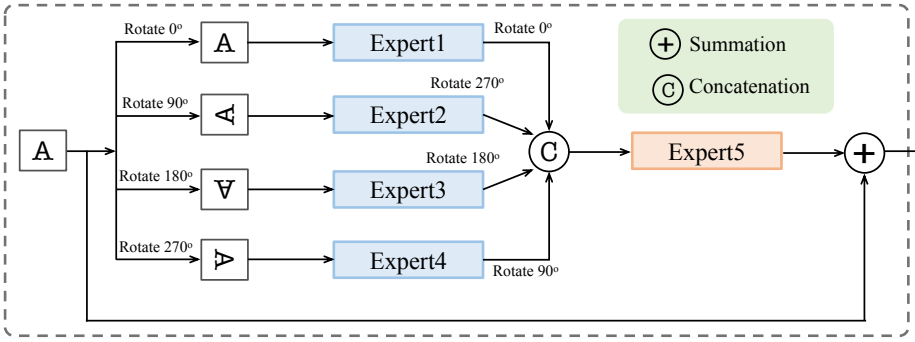
4.5 In track2-guided at PBVS2024

Building on a multi-scale design inspired by [32], the pipeline (Fig. 8) first bicubically upsamples the LR thermal frame to the HR RGB resolution for GTISR at $\times 8$ and $\times 16$. The aligned thermal–RGB pair is concatenated and dispatched into two pixel-unshuffled streams at $2\times$ and $4\times$ scales, which form a compact feature pyramid without expensive strided convolutions. Each stream begins with shallow extractors composed of two deformable convolutions with a PReLU in between, allowing spatially adaptive sampling that can accommodate residual cross-spectral misalignment or parallax while keeping early computation lightweight.

Feature fusion is realized by a recurrent block repeated N times. At each iteration, the $4\times$-scale features are upsampled to the $2\times$ domain and concatenated to promote cross-scale interaction. The fused tensor is refined by a residual unit with enhanced channel attention [33], which adaptively re-weights channels to emphasize informative responses from both modalities. A channel transformer [34] then aggregates long-range dependencies and balances multi-scale context across channels, complementing the locality of the deformable convolutions. After the recurrent fusion, the $2\times$ branch feeds a reconstruction head to synthesize the super-resolved thermal output. This modular layout—pre-alignment via bicubic upsampling, pyramid formation by pixel unshuffling, deformable and attention-based refinement, and transformer-style channel mixing—aims to couple robustness to geometric variation with strong detail transfer from RGB guidance, while remaining scalable to both $\times 8$ and $\times 16$ GTISR settings.



(a) Proposed architecture



(b) Details of proposed approach.

Figure 7: Proposed architecture in PBVS 2023 [10].

4.6 In track2-guided at PBVS2025

The guided thermal SR pipeline in Fig. 9 leverages both a low-resolution thermal image and a corresponding high-resolution RGB image. Inputs are concatenated for early fusion and passed through a patch-embedding stage that maps local neighborhoods into a latent feature space. The backbone then applies cascaded Gated CNN Blocks inspired by recent sequence-modeling advances [35] together with deep residual learning principles [36]. The gating mechanism modulates information flow across channels and spatial locations, enabling the network to highlight structures conveyed by RGB guidance while preserving modality-specific thermal cues. Reconstruction proceeds by patch unembedding to recover the image grid, and a global residual pathway from the input thermal image stabilizes optimization and helps retain low-frequency content throughout the network depth.

Training adopts an L_1 reconstruction loss between predictions and ground-truth HR thermal images. The reported configuration includes approximately 132M parameters with mini-batches of 16 patches of size 128×128, trained for two days on an NVIDIA A100 GPU using PyTorch. In practice, the combination of early fusion, gated convolutional blocks, and residual connections offers a balance between long-range context aggregation and computational efficiency, and is aligned with the standardized Track 2 evaluation at $\times8$ and $\times16$.

5 Results: Quantitative and Qualitative Trends

Across six PBVS TISR editions, performance improved steadily and protocols became more consistent. Figures 10 and 11 summarize the official metrics (PSNR/SSIM) by year and track; report trends rather than full leaderboards to keep the survey concise. Detailed per-team scores remain available on the public CodaLab pages and annual reports cited earlier in Section 2.2.

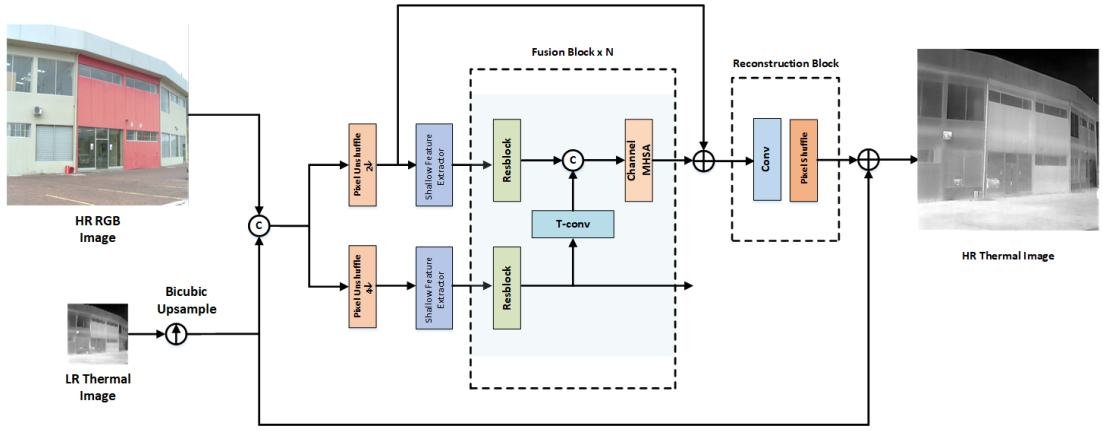


Figure 8: Architecture proposed in PBVS 2024 [11].

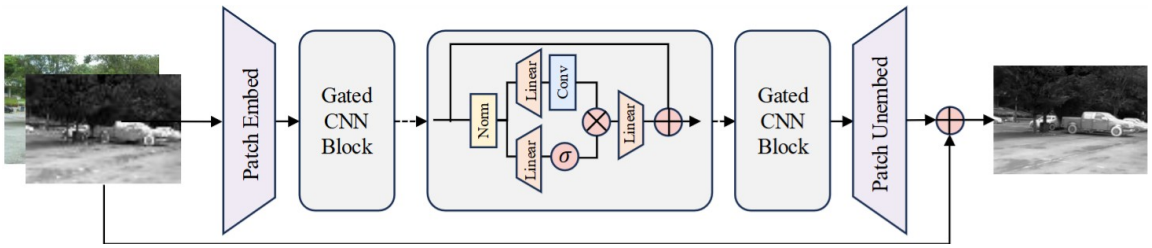


Figure 9: Architecture proposed in PBVS 2025 [12].

Quantitative. From 2020–2022, residual CNNs on thermal-only data yielded gradual gains. The 2023 introduction of attention and cross-spectral guidance produced a clear jump for the guided $\times 8$ track. With CIDIS in 2024–2025, hybrid CNN–Transformer designs and standardized registration further increased mean PSNR/SSIM for both SISR $\times 8$ and guided $\times 8/\times 16$ (see Figs. 10–11).

Qualitative. Top entries preserve sharper edges and thin structures (e.g., poles, window frames) and reduce ringing near high-contrast boundaries. Guided methods improve contour continuity versus single-image SR, although slight RGB–TIR misregistration may cause faint ghosting at occlusions; leading systems mitigate this with deformable alignment and confidence-gated fusion. Efficiency-wise, windowed attention and lightweight hybrid blocks control memory/latency with minor accuracy costs; test-time self-ensembles provide small but consistent gains.

Deployment considerations. For surveillance, driving, and industrial monitoring, accuracy alone is insufficient. Reporting parameters, FLOPs, and inference latency on a reference device—together with robustness probes (mild misregistration, sensor-like noise)—would improve comparability and encourage deployable designs.

6 Conclusions and Future Directions

Across six PBVS TISR editions (2020–2025), the community advanced from residual CNNs on thermal-only data to hybrid CNN–Transformer architectures with cross-spectral guidance. These developments were accompanied by steady improvements in PSNR and SSIM, as well as more consistent evaluation protocols. The introduction of the CIDIS dataset, comprising registered RGB–thermal pairs, enabled guided super-resolution at $\times 8$ and $\times 16$ and played a key role in standardizing benchmarks. At the same time, the challenge reports con-

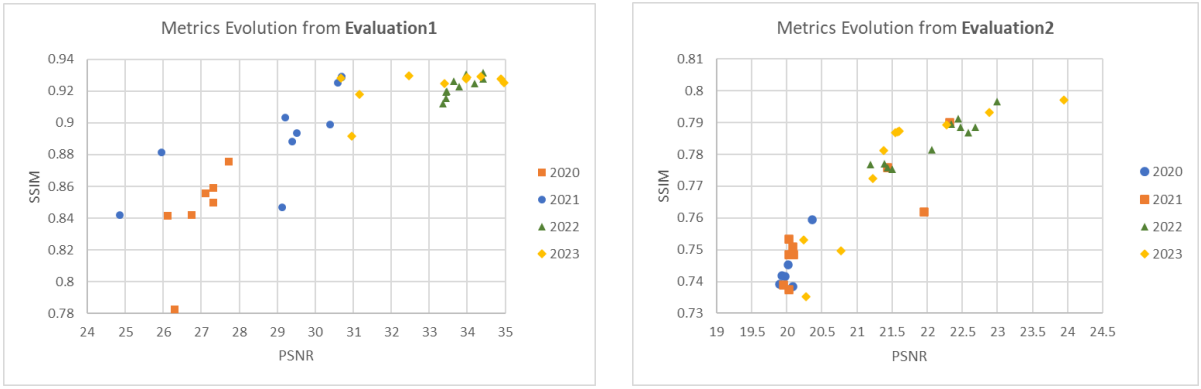


Figure 10: Metrics evolution through in 2020, 2021, 2022 & 2023 challenges [10].

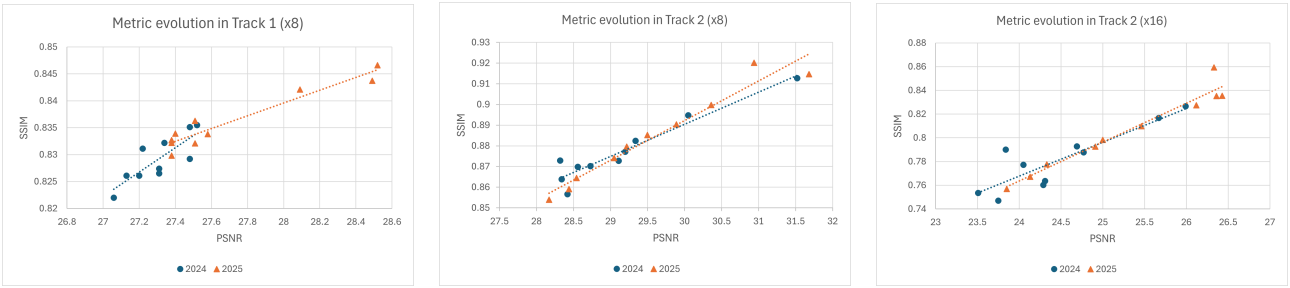


Figure 11: Metrics evolution through all tracks in 2024 & 2025 challenges [12].

sistently highlight several open issues: sensitivity to residual misregistration, the computational cost of large models, and the limited diversity of existing datasets when facing extreme or adverse conditions.

Future editions of the challenges are expected to evolve in several complementary directions. Expanding data coverage remains crucial, particularly toward multimodal acquisitions in challenging scenarios such as fog, rain, heat haze, night-time scenes, fast motion, and varied backgrounds. Publishing registration error budgets for each pair would further support fair comparisons. Including diffusion-based super-resolution baselines could enrich methodological diversity, as these models have shown promise for recovering fine details at large scale factors while maintaining data consistency through guidance and back-projection mechanisms. Another relevant aspect is the systematic reporting of computational efficiency and robustness, including parameters, FLOPs, inference latency on reference devices, and evaluations under sensor-like noise or slight misregistration. Such standardized reporting would improve comparability and encourage methods that balance accuracy with deployability. Finally, ensuring clear attribution and reproducibility remains important: pointing readers to the annual references when describing team innovations helps preserve credit and supports long-term research continuity.

The datasets released for these challenges avoid personally identifiable information in thermal imagery, and visible frames used for guidance follow restricted capture protocols with masking or blurring where appropriate. Public releases exclude metadata that could enable identification or geolocation and comply with institutional and local data-protection norms.

In summary, the PBVS TISR series has documented major advances in thermal image super-resolution while providing a structured basis for future comparisons. Continued attention to data diversity, generative baselines, efficiency and robustness evaluation, and reproducible reporting will further strengthen the impact of upcoming editions.

References

[1] J. Lu, C. Zhao, Y. Zhang, J. Zhu, and J. Zheng, "Gtf: Generator for pairwise thermal face image synthesis," in *International Conference on Intelligent Computing*, pp. 72–83, Springer, 2025.

[2] S. Hazra and U. Bhattacharya, "Cross-modal attention with adaptive and hierarchical fusion for robust rgb-t image segmentation for safe driving," in *2025 IEEE International Conference on Image Processing (ICIP)*, pp. 2139–2144, IEEE, 2025.

[3] E. Mihailov, D. Choshnova, M. Ivanova, and M. Asenova, "Investigation of the possibilities for infrared diagnosis of pierce-smith converters in non-ferrous metallurgy," 2025.

[4] A. M. Sharafaddini, K. K. Esfahani, and N. Mansouri, "Deep learning approaches to detect breast cancer: a comprehensive review," *Multimedia Tools and Applications*, vol. 84, no. 21, pp. 24079–24190, 2025.

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

[6] R. E. Rivadeneira, H. O. Velesaca, and A. D. Sappa, "Cross-spectral image registration: a comparative study and a new benchmark dataset," in *Proceedings of International Conference on Innovations in Computational Intelligence and Computer Vision (ICICV)*, 2024.

[7] R. E. Rivadeneira, A. D. Sappa, B. X. Vintimilla, L. Guo, J. Hou, A. Mehri, P. Behjati Ardakani, H. Patel, V. Chudasama, K. Prajapati, K. P. Upla, R. Ramachandra, K. Raja, C. Busch, F. Almasri, O. Debeir, S. Nathan, P. Kansal, N. Gutierrez, B. Mojra, and W. J. Beksi, "Thermal image super-resolution challenge-PBVS 2020," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 96–97, 2020.

[8] 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-pbvs 2021," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4359–4367, 2021.

[9] 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-pbvs 2022," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 418–426, 2022.

[10] R. E. Rivadeneira, A. D. Sappa, B. X. Vintimilla, D. Bin, L. Ruodi, L. Shengye, Z. Zhong, X. Liu, J. Jiang, and C. Wang, "Thermal image super-resolution challenge results-pbvs 2023," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 470–478, 2023.

[11] R. E. Rivadeneira, A. D. Sappa, C. Wang, J. Jiang, Z. Zhong, P. Chen, and S. Wang, "Thermal image super-resolution challenge results-pbvs 2024," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3113–3122, 2024.

[12] R. E. Rivadeneira, A. D. Sappa, J. R. iad Hammoud, H. Zhong, Y. Wang, S. Zhao, Z. Zhong, Y.-H. Li, S. Wang, Q. Shen, H. Wang, and X. Zhang, "Thermal image super-resolution challenge results-pbvs 2025," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 0–0, 2025.

[13] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[14] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[15] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," in *European conference on computer vision*, pp. 404–417, Springer, 2006.

[16] K. Marstal, F. Berendsen, M. Staring, and S. Klein, "Simpleelastix: A user-friendly, multi-lingual library for medical image registration," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 134–142, 2016.

[17] D. Muthukumaran and M. Sivakumar, "Medical image registration: a matlab based approach," *Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol*, vol. 2, no. 1, pp. 29–34, 2017.

[18] P. Lindenberger, P.-E. Sarlin, and M. Pollefeys, "Lightglue: Local feature matching at light speed," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 17627–17638, 2023.

[19] M. Arar, Y. Ginger, D. Danon, A. H. Bermano, and D. Cohen-Or, "Unsupervised multi-modal image registration via geometry preserving image-to-image translation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13410–13419, 2020.

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

[21] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1251–1258, 2017.

[22] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 286–301, 2018.

[23] L. Trottier, P. Gigu, B. Chaib-draa, *et al.*, "Parametric exponential linear unit for deep convolutional neural networks," in *16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pp. 207–214, IEEE, 2017.

[24] A. Shah, E. Kadam, H. Shah, S. Shinde, and S. Shingade, "Deep residual networks with exponential linear unit," in *Proceedings of the International Symposium on Computer Vision and the Internet*, pp. 59–65, 2016.

[25] R. Ye, F. Liu, and L. Zhang, "3d depthwise convolution: Reducing model parameters in 3d vision tasks," in *Proceedings of the Canadian Conference on Artificial Intelligence*, pp. 186–199, Springer, 2019.

[26] Y. Huang, J. Li, *et al.*, "Interpretable Detail-Fidelity Attention Network for Single Image Super-Resolution," *IEEE Transactions on Image Processing*, vol. 30, pp. 2325–2339, 2021.

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

[28] L. Chen, X. Chu, X. Zhang, and J. Sun, "Simple baselines for image restoration," in *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VII*, pp. 17–33, Springer, 2022.

[29] Z. Zhong, X. Liu, J. Jiang, D. Zhao, and X. Ji, "Deep attentional guided image filtering," *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[30] Z. Zhong, X. Liu, J. Jiang, D. Zhao, and X. Ji, "Guided depth map super-resolution: A survey," *ACM Comput. Surv.*, vol. 55, jul 2023.

[31] R. E. Rivadeneira, A. D. Sappa, B. X. Vintimilla, J. Kim, D. Kim, Z. Li, Y. Jian, B. Yan, L. Cao, F. Qi, H. Wang, R. Wu, L. Sun, Y. Zhao, L. Li, K. Wang, Y. Wang, X. Zhang, H. Wei, C. Lv, Q. Sun, X. Tian, Z. Jia, J. Hu, C. Wang, Z. Zhong, X. Liu, and J. Jiang, "Thermal image super-resolution challenge results - pbvs 2022," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 417–425, 2022.

[32] B. Kathariya, Z. Li, and G. Van der Auwera, "Joint pixel and frequency feature learning and fusion via channel-wise transformer for high-efficiency learned in-loop filter in vvc," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2023.

[33] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "Eca-net: Efficient channel attention for deep convolutional neural networks," 2020.

[34] Y. Cai, J. Lin, Z. Lin, H. Wang, Y. Zhang, H. Pfister, R. Timofte, and L. V. Gool, "Mst++: Multi-stage spectral-wise transformer for efficient spectral reconstruction," 2022.

[35] W. Yu and X. Wang, "Mambaout: Do we really need mamba for vision?," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

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