DISTA-Net: Dynamic Closely-Spaced Infrared Small Target Unmixing

Shengdong Han[1*] Shangdong Yang[1*] Yuxuan Li[2] Xin Zhang[2]
Xiang Li[2,3] Jian Yang[2] Ming-Ming Cheng[2,3] Yimian Dai[2,3†]
[1] School of Computer Science, Nanjing University of Posts and Telecommunications
[2] VCIP, CS, Nankai University [3]NKIARI, Futian, Shenzhen

Abstract

Resolving closely-spaced small targets in dense clusters presents a significant challenge in infrared imaging, as the overlapping signals hinder precise determination of their quantity, sub-pixel positions, and radiation intensities. While deep learning has advanced the field of infrared small target detection, its application to closely-spaced infrared small targets has not yet been explored. This gap exists primarily due to the complexity of separating superimposed characteristics and the lack of an open-source infrastructure. In this work, we propose the Dynamic Iterative Shrinkage Thresholding Network (DISTA-Net), which reconceptualizes traditional sparse reconstruction within a dynamic framework. DISTA-Net adaptively generates convolution weights and thresholding parameters to tailor the reconstruction process in real time. To the best of our knowledge, DISTA-Net is the first deep learning model designed specifically for the unmixing of closely-spaced infrared small targets, achieving superior sub-pixel detection accuracy. Moreover, we have established the first open-source ecosystem to foster further research in this field. This ecosystem comprises three key components: (1) CSIST-100K, a publicly available benchmark dataset; (2) CSO-mAP, a custom evaluation metric for sub-pixel detection; and (3) GrokCSO, an open-source toolkit featuring DISTA-Net and other state-of-the-art models, available at https://github.com/GrokCV/GrokCSO.

1. Introduction

Infrared imaging plays a pivotal role in various long-distance detection and surveillance tasks [26], due to its exceptional sensitivity to thermal radiation and independence from illumination conditions. However, the radiation intensity captured from remote targets is inherently weak due to their long-range distance from the imaging system [11]. **This challenge is exacerbated when targets appear in spatially close dense clusters, as the closely-spaced objects (CSO)** [24], **resulting in overlapping blob-like spots.**

As shown in Fig. 1, such overlap makes it impossible to resolve the targets independently via human vision, thus obscuring the perception of target count, precise locations, and radiation intensities [46], presenting a substantial impediment for Infrared Search and Tracking (IRST) systems in their subsequent phases of detection, tracking, and identification. Therefore, exploring effective techniques for unmixing and reconstruction of such **closely-spaced infrared small targets (CSIST)**, to accurately discern their exact locations and radiant intensities, holds great significance.

Despite the critical importance of CSIST unmixing in various applications, **research addressing this specific task remains exceedingly scarce**. These approaches typically formulate the problem as a parameter estimation task and employ optimization algorithms to solve it [21]. Target sparsity on the imaging plane was leveraged to devise a discretized sampling-based sparse reconstruction method [39], utilizing an over-complete dictionary and solving a second-order cone programming problem under the ℓ_1 norm regularization. However, the performance of these optimization-based models is highly dependent on meticulous hyperparameter tuning [1], which poses significant challenges in real-world scenarios. Variations in target quantity or location further complicate the selection of optimal hyperparameters, limiting the generalizability and practicality of these methods [29]. Consequently, there is a pressing need for the development of unmixing algorithms that exhibit greater robustness to hyperparameter variations and can be effectively applied in diverse real-world settings.

While deep learning has revolutionized image super-resolution [13], its application to CSIST unmixing remains unexplored, primarily due to fundamentally different task objectives and ecosystem limitations. Unlike generic super-resolution that enhances clarity through high-frequency detail restoration [33], infrared small target unmixing requires precise estimation of overlapping targets' counts, locations, and radiation intensities - essentially mapping images to specific target attributes rather than conventional pixel-space

*Equal contribution.
†Corresponding author: yimian.dai@gmail.com.

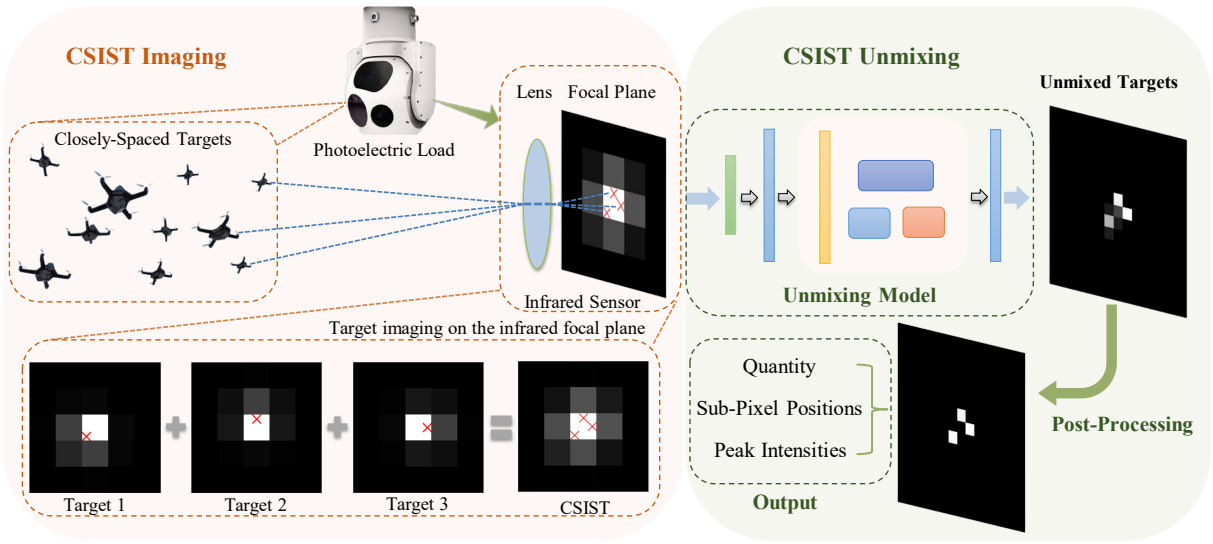

Figure 1. Conceptual illustration of imaging and unmixing processes for closely-spaced infrared small targets (CSIST). CSIST unmixing aims to disentangle and accurately estimate the count, positions, and intensities of overlapping targets.

super-resolution. This challenge is further compounded by the absence of standardized benchmark datasets, task-specific evaluation metrics, and open-source implementations, creating significant barriers to developing and comparing deep learning approaches in this specialized domain.

To address the aforementioned challenges, in this paper, we propose a novel deep unfolding network for CSIST unmixing, termed as the Dynamic Iterative Shrinkage Thresholding Network (DISTA-Net). We reformulate the traditional sparse reconstruction approach into a dynamic deep learning framework, which adaptively generates convolution weights and thresholding parameters to tailor the reconstruction process in real time. Distinct from prior methods, the parameters associated with the proximal mapping (nonlinear transforms and shrinkage thresholds) are dynamically adapted to the input data, rather than being hand-crafted or fixed after training. *To the best of our knowledge, this represents the first deep learning-based effort towards CSIST unmixing.*

Furthermore, we establish a comprehensive open-source ecosystem to facilitate research in this domain, including CSIST-100K, an open benchmark dataset comprising 100,000 pairs of CSIST images and exact annotations of location and radiation intensities; CSO-mAP, a custom evaluation metric inspired by the mean average precision (mAP) from object detection, calibrated to evaluate the quantity, spatial positioning, and radiation intensity of the unmixed infrared targets; and GrokCSO, an open-source PyTorch-based toolkit encapsulating our DISTA-Net alongside other state-of-the-art models, empowering researchers to effortlessly leverage the CSIST-100K dataset.

Our contributions can be categorized into **Four** aspects: 1. We reformulate CSIST unmixing as an interpretable deep unfolding problem. To our knowledge, this is the first deep learning based attempt for this task. 2. Our proposed DISTA-Net is a dynamic deep unfolding network, which adaptively generates convolution weights and thresholding parameters to tailor the reconstruction process conditioned on the input data. 3. We establish the first open-source ecosystem for this task, including the CSIST-100K dataset, the CSO-mAP metric, and the GrokCSO toolkit. 4. We provide a comprehensive analysis of our approach, validating the importance of dynamic deep unfolding and the effectiveness of the DISTA-Net in addressing the CSIST unmixing task.

2. Related Work

2.1. Infrared Small Target Detection

Driven by a range of open-source datasets [4, 6, 15], infrared small target detection has garnered significant research attention in recent years. Current research mainly focuses on developing multi-scale feature fusion models to counteract the scarcity of intrinsic target features [35]. Dai *et al.* introduced an asymmetric contextual modulation module that bridges high-level semantics with low-level details using a bottom-up pathway via point-wise channel attention [5]. Wang *et al.* merged reinforcement learning with pyramid feature fusion and proposed a global context boundary attention module to mitigate localized bright noise [32]. Cheng *et al.* proposed a difference-aware attention module with a dual-temporal aggregation module for global feature learning and channel activation, and a difference-attention module for multi-scale detection via local correlations [20]. Tong *et al.* adopted an encoder–decoder structure, enhancing feature extraction with an atrous spatial pyramid pooling and

a dual-attention module, and using multiscale labels to focus on target edges and internal features [28]. Zhang *et al.* proposed an infrared small target detection framework integrating visual-textual information via CLIP-prompted SAM adaptation with a denoising module, achieving enhanced generalization capability for the infrared domain [41].

Our work focuses on infrared small targets but differs in two key aspects. First, while infrared small target detection precedes our study, we prioritize CSIST unmixing, where detecting overlapped targets is crucial for sub-pixel localization and radiation intensity prediction. Second, our task goes beyond detection by enabling sub-pixel-level localization and radiative intensity estimation, providing finer target characterization than binary detection.

2.2. Deep Unfolding

Deep unfolding, as delineated in [22], originated in 2010 with the Learned Iterative Shrinkage-Thresholding Algorithm (LISTA) [9], which reinterprets ISTA [7] as a fully connected feed-forward neural network. This approach generalizes effectively to new samples, achieving ISTA-like accuracy with fewer iterations. Subsequent works, such as ADMM-Net [36], have unfolded the steps of the Alternating Direction Method of Multipliers (ADMM) into a deep learning framework, thereby improving the accuracy and efficiency of MRI reconstruction through a compressive sensing model optimized via end-to-end discriminative training. Likewise, ISTA-Net [40] has adopted an end-to-end learning approach for proximal mapping, enhancing the performance of compressive sensing for natural image reconstruction.

Motivated by such advances, deep unfolding has found applicability in a variety of computer vision tasks. Notably, Li *et al.* transformed a generalized gradient-domain total variation algorithm into a deep interpretable network for blind image deblurring, delivering superior performance with learned parameters [14]. For image super-resolution, Guo *et al.* incorporated trainable convolutional layers into the Discrete Cosine Transform framework, effectively mitigating artifacts and enabling learning from limited data [10]. Solomon *et al.* unfolded robust principal component analysis into a deep network, improving the distinction between microbubble and tissue signals in ultrasound imaging [27].

Unlike previous methods with static parameters [37, 40], our DISTA-Net dynamically adapts proximal mapping weights based on the input, enabling an adaptive reconstruction process that caters to varying scenarios.

3. CSIST Benchmark, Metric, and Toolkit

CSIST-100K Dataset. In our study, we set σ_{PSF} at 0.5 pixels. Simulations include 1–5 overlapping targets per image, each defined by 2D coordinates and radiation intensity (220–250 units), randomly placed within an 11×11 grid while maintaining ≥ 0.52 Rayleigh units separation. We

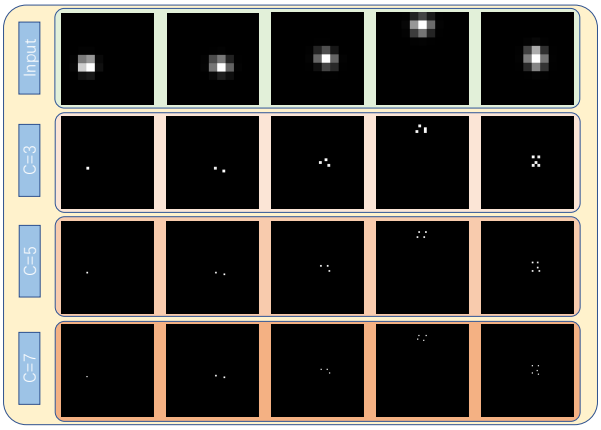

Figure 2. CSIST Visualization: The top row shows 1 to 5 overlapping targets, and the following rows display unmixing results for sub-pixel division factors of $3\times$, $5\times$, and $7\times$.

generated the **CSIST-100K** dataset with 100,000 samples: 80,000 for training, and 20,000 split equally for validation and testing. As shown in Fig. 2, closely spaced targets diffuse, with energy concentrated in a 3×3 pixel area, causing significant overlap that complicates target counting and coordinate determination. (See **Supplementary** for optical modeling details.)

CSO-mAP Metric. Traditional bounding box metrics fail when target separation falls below the Rayleigh criterion (due to Airy spot interference). Our CSO-mAP with strict sub-pixel position/intensity matching for CSIST evaluation, redefining TP/FP as follows:

$$\mathbb{1}_k\left(\hat{t}_j, t_i\right) = \begin{cases} 1, & \text{if } d\left(\hat{t}_j, t_i\right) < \delta_k, \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

Here, $\delta_k \in \{0.05, 0.1, 0.15, 0.2, 0.25\}$ is a series of distance thresholds, with $k = 1, 2, 3, 4, 5$, used to control the desired localization accuracy. Precision-recall (PR) curves are generated through intensity-prioritized matching, with Average Precision (AP) computed at each δ_k. The final CSO-mAP metric is derived by averaging AP values across all thresholds, explicitly quantifying performance under varying spatial resolution demands (details in **Supplementary**).

GrokCSO Toolkit. To address the lack of specialized tools in this domain, we introduce GrokCSO, an open-source toolkit for CSIST unmixing. Built on PyTorch, GrokCSO provides pre-trained models, reproducibility scripts, and specialized evaluation metrics tailored for CSO challenges. The detailed architecture and features of this toolkit are provided in **Supplementary**.

4. Method

In this section, we introduce the imaging model for closely-spaced infrared small targets, traditional sparse reconstruction approaches, and the proposed DISTA-Net architecture.

4.1. Imaging and Unmixing Framework

CSIST Imaging Model. Given the significant distance between targets and the infrared detector, targets can be approximated as point sources. The optical system's diffraction spreads the energy across adjacent pixels, described by a two-dimensional Gaussian point spread function (PSF) [18]:

$$p(x,y) = \frac{1}{2\pi\sigma_{\mathrm{PSF}}^2} \exp\left[-\frac{(x-x_t)^2 + (y-y_t)^2}{2\sigma_{\mathrm{PSF}}^2}\right], \quad (2)$$

where σ_{PSF}^2 is the diffusion variance and (x_t, y_t) the target coordinates. On an infrared focal plane of $U \times V$ pixels, each pixel integrates the PSF within its boundaries:

$$g_{i,j}(x_t, y_t) = \int_{x_{i,j}-1/2D}^{x_{i,j}+1/2D} \int_{y_{i,j}-1/2D}^{y_{i,j}+1/2D} p(x,y)\mathrm{d}x\,\mathrm{d}y, \quad (3)$$

where $(x_{i,j}, y_{i,j})$ is the pixel's center and D the pixel width. The focal plane measurement model is vectorized as:

$$\mathbf{z} = \mathbf{G}(\mathbf{x}, \mathbf{y})\mathbf{s} + \mathbf{n}, \quad (4)$$

where $\mathbf{G}(\mathbf{x}, \mathbf{y})$ is the steering matrix, $\mathbf{s}$ represents target intensities, and $\mathbf{n}$ denotes Gaussian white noise.

CSIST Unmixing via Sparse Reconstruction. Given permissible quantization error, target positions in closely-spaced infrared small targets can be discretized into a finite set of sub-pixel locations $\Omega = \{(x_l, y_l)\}_{l=1,\cdots,L}$, where actual target positions form a sparse subset.

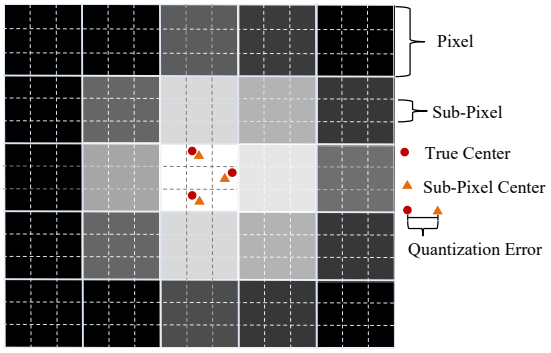

Figure 3. Division of each pixel into an $n \times n$ grid of sub-pixels, representing potential target positions.

As shown in Fig. 3, each pixel is divided into an $n \times n$ grid, resulting in $L = UVn^2$ sub-pixels. With sufficient grid resolution, each sub-pixel contains at most one target, with maximum position deviation of $\sqrt{2}D/n$. The measurement model can be extended to Ω:

$$\mathbf{z} = \mathbf{G}(\Omega)\tilde{\mathbf{s}} + \mathbf{w}, \quad (5)$$

where $\mathbf{G}(\Omega)$ contains steering vectors from Ω, and $\tilde{\mathbf{s}} \in \mathrm{R}^L$ is sparse with $L \gg UV$. The CSIST unmixing problem can then be formulated as a sparse reconstruction with ℓ_1 regularization:

$$\min_{\tilde{\mathbf{s}}} \|\mathbf{z} - \mathbf{G}(\Omega)\tilde{\mathbf{s}}\|_2^2 + \lambda\|\tilde{\mathbf{s}}\|_1, \quad (6)$$

where λ is the regularization parameter. The solution $\tilde{\mathbf{s}}$ directly yields target attributes: its non-zero entries indicate target count and intensities, while their corresponding positions in Ω provide sub-pixel coordinates.

Optimization Solution. The ISTA [7] solves the sparse reconstruction problem through two alternating steps:

$$\mathbf{r}^{(k)} = \tilde{\mathbf{s}}^{(k-1)} - \rho\mathbf{G}^\top\left(\mathbf{G}\tilde{\mathbf{s}}^{(k-1)} - \mathbf{z}\right), \quad (7)$$

$$\tilde{\mathbf{s}}^{(k)} = \arg\min_{\tilde{\mathbf{s}}} \frac{1}{2}\left\|\tilde{\mathbf{s}} - \mathbf{r}^{(k)}\right\|_2^2 + \lambda\|\mathbf{\Psi}\tilde{\mathbf{s}}\|_1, \quad (8)$$

where $\mathbf{\Psi}$ denotes the transform matrix, k the iteration index, and ρ the step size. The second step represents a proximal mapping:

$$\mathrm{prox}_{\lambda\phi}(\mathbf{r}) = \arg\min_{\tilde{\mathbf{s}}} \frac{1}{2}\|\tilde{\mathbf{s}} - \mathbf{r}\|_2^2 + \lambda\phi(\tilde{\mathbf{s}}). \quad (9)$$

While ISTA with orthogonal transforms (e.g., wavelets) has efficient solutions, it faces challenges with complex transforms and requires numerous iterations. To address these limitations, ISTA-Net replaces $\mathbf{\Psi}$ with a trainable non-linear transform $\mathcal{F}(\cdot)$:

$$\tilde{\mathbf{s}}^{(k)} = \arg\min_{\tilde{\mathbf{s}}} \frac{1}{2}\left\|\mathcal{F}(\tilde{\mathbf{s}}) - \mathcal{F}\left(\mathbf{r}^{(k)}\right)\right\|_2^2 + \theta\|\mathcal{F}(\tilde{\mathbf{s}})\|_1, \quad (10)$$

where θ is a learnable parameter. However, ISTA-Net's static network weights post-training limit its adaptability to input data, particularly in CSIST unmixing scenarios where input sensitivity is crucial.

4.2. DISTA-Net: A Dynamic Framework

The details of our DISTA-Net are illustrated in Fig. 4. Building upon ISTA-Net's framework, we introduce two key improvements. First, we design a data-adaptive nonlinear transformation function $\mathcal{F}_d(\cdot)$ that maps images to richer dimensional representations while emphasizing significant image regions. The transformation follows:

$$\mathcal{F}_d(\tilde{\mathbf{s}}^{(k)}) = \mathrm{Soft}(\mathcal{F}_d(\mathbf{r}^{(k)}), \theta_d), \quad (11)$$

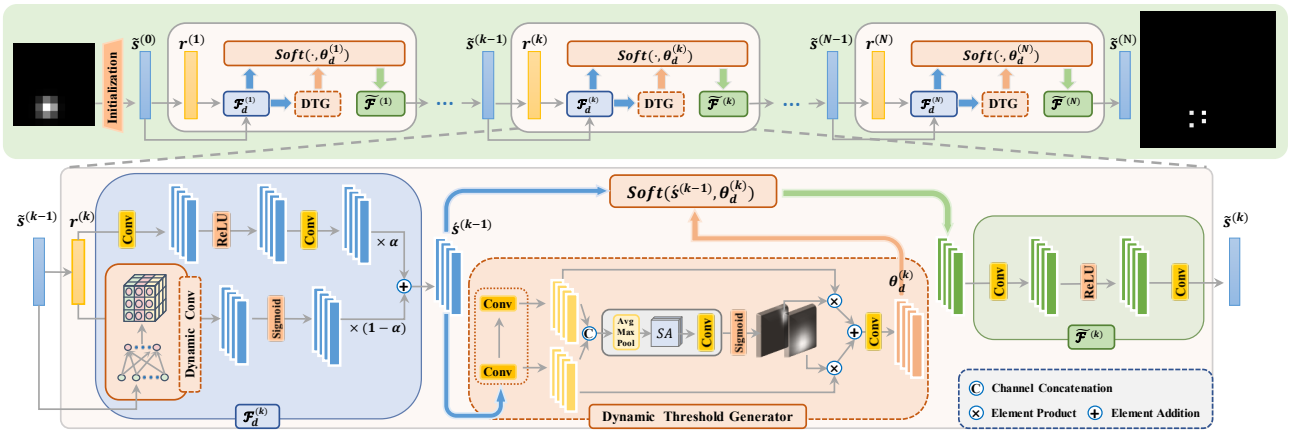

Figure 4. Architecture of the proposed DISTA-Net. The overall framework consists of multiple cascaded stages. Each stage contains three main components: a dual-branch dynamic transform module ($\mathcal{F}_d^{(k)}$) for feature extraction, a dynamic threshold module ($\Theta_d^{(k)}$) for feature refinement, and an inverse transform module ($\tilde{\mathcal{F}}^{(k)}$) for reconstruction.

Algorithm 1 DISTA-Net

Input: CSIST image $\mathbf{z}$, steering matrix $\mathbf{G}(\mathbf{x},\mathbf{y})$, initial matrix Q_{init}, number of stages N, step size $\{\rho^{(k)}\}_{k=1}^{N}$

Output: Reconstructed result $\tilde{\mathbf{s}}^{(N)}$

Learnable parameters:

$\{\rho^{(k)}\}_{k=1}^{N}$, $\{\text{DTG}^{(k)}\}_{k=1}^{N}$, $\{\mathcal{F}_d^{(k)}\}_{k=1}^{N}$, $\{\tilde{\mathcal{F}}^{(k)}\}_{k=1}^{N}$

$(\tilde{\mathcal{F}}^{(k)} \circ \mathcal{F}_d^{(k)} = \mathbf{I})$

Initialization:

$\tilde{\mathbf{s}}^{(0)} \leftarrow Q_{\text{init}}\mathbf{z}$

Iterative reconstruction:

for $k = 1$ **to** N **do**

  $\mathbf{r}^{(k)} \leftarrow \tilde{\mathbf{s}}^{(k-1)} - \rho^{(k)}\mathbf{G}^{\top}(\mathbf{G}\tilde{\mathbf{s}}^{(k-1)} - \mathbf{z})$

  $\acute{\mathbf{s}}^{(k)} \leftarrow \mathcal{F}_d(\tilde{\mathbf{s}}^{(k-1)}, \mathbf{r}^{(k)})$

  $\theta_d^{(k)} \leftarrow \text{DTG}^{(k)}(\acute{\mathbf{s}}^{(k)})$

  $\tilde{\mathbf{s}}^{(k)} \leftarrow \tilde{\mathcal{F}}(\text{Soft}(\mathcal{F}_d(\mathbf{r}^{(k)}), \theta_d^{(k)}))$

end for

where $\text{Soft}(\cdot,\theta_d)$ denotes soft-thresholding with learnable parameter θ_d, and k is the stage index. The left inverse $\tilde{\mathcal{F}}$ satisfies $\tilde{\mathcal{F}}(\cdot)\circ\mathcal{F}_d(\cdot) = \mathbf{I}$, without requiring structural symmetry to $\mathcal{F}_d(\cdot)$, yielding:

$$\tilde{\mathbf{s}}^{(k)} = \tilde{\mathcal{F}}(\text{Soft}(\mathcal{F}_d(\mathbf{r}^{(k)}), \theta_d)). \tag{12}$$

Second, we introduce a dynamic threshold module that adapts to input image variations, addressing the sensitivity of sparse vector perturbations in image generation. This flexible thresholding mechanism improves upon fixed parameters that can be either too strict or too lenient. As shown in Fig. 4, DISTA-Net comprises N stages. Each stage k ($k > 1$) contains three components: $\mathcal{F}_d^{(k)}$, $\text{Soft}(\cdot,\theta_d^{(k)})$, and $\tilde{\mathcal{F}}^{(k)}$. The input $\mathbf{r}^{(k)}$ is derived from $\tilde{\mathbf{s}}^{(k-1)}$ via Eq. (7), processed through these components sequentially to generate $\tilde{\mathbf{s}}^{(k)}$ for the next iteration. Overall, the proposed DISTA-Net can be referenced in Algorithm 1.

Dynamic Transform. While trainable non-linear transformations overcome limitations of handcrafted methods, their fixed post-training parameters result in static transformation patterns. We address this by introducing a dual-branch *Dynamic Transform* module $\mathcal{F}_d^{(k)}$ at the k-th stage.

To handle the sensitivity of sparse image $\tilde{\mathbf{s}}^{(k-1)}$ perturbations, we design an auxiliary branch that guides $\mathbf{r}^{(k)}$ through a dynamic convolutional kernel. This approach enhances feature representation adaptively.

The module first processes $\tilde{\mathbf{s}}^{(k-1)}$ through a fully connected network to generate a weight vector:

$$W = f(\tilde{\mathbf{s}}^{(k-1)}). \tag{13}$$

The *Dynamic Conv* module applies this weight vector as an adaptive convolutional kernel to $\mathbf{r}^{(k)}$:

$$w_r = C(W, \mathbf{r}^{(k)}). \tag{14}$$

The final output combines a Conv-ReLU-Conv branch with the sigmoid-activated auxiliary branch:

$$\mathcal{F}_d^{(k)} = \alpha \cdot A(\text{ReLU}(B(\mathbf{r}^{(k)}))) + (1-\alpha)\cdot \text{sigmoid}(w_r), \tag{15}$$

where $A(\cdot)$ and $B(\cdot)$ are convolution operations and $\alpha \in [0,1]$ governs the contribution between the two branches.

Dynamic Soft-Thresholding. Unlike ISTA-Net's fixed threshold θ, we propose a *Dynamic Thresholding Generator(DTG)* module that adapts θ_d based on input image information. This approach better handles densely overlapped targets and spatial context variations.

As shown in Fig. 4, the module employs dual convolutional layers to capture multi-scale features as in [12, 16]. The $\mathcal{F}_d^{(k)}$ output passes through two 3×3 convolutions, generating feature maps $\tilde{U}_1$ and $\tilde{U}_2$. These are concatenated to form $\tilde{U} = [\tilde{U}_1, \tilde{U}_2]$. Spatial relationships are captured through parallel pooling operations:

$$SA_{\text{avg}} = P_{\text{avg}}(\tilde{U}), \quad SA_{\text{max}} = P_{\text{max}}(\tilde{U}). \tag{16}$$

The pooled features are processed through the convolution:

$$(\hat{SA}) = \text{Conv}^{2\to N}([\text{SA}_{\text{avg}}; \text{SA}_{\text{max}}]), \tag{17}$$

followed by a sigmoid activation to generate spatial selective masks:

$$(\tilde{SA})_i = \text{sigmoid}((\hat{SA})_i). \tag{18}$$

The dynamic threshold θ_d is then computed by combining these masks with multi-scale feature maps through a final convolution $C(\cdot)$:

$$\theta_d = C(\sum_{i=1}^{N} (\tilde{SA})_i \cdot \tilde{U}_i). \tag{19}$$

Initialization and Learning Objectives. DISTA-Net employs linear initialization similar to iterative sparse coding algorithms, where the initial estimate $\tilde{s}^{(0)}$ is obtained through an optimal linear projection Q_{init} learned from training pairs $\{(z_i, s_i)\}$ (details in **Supplementary**).

The training objective combines reconstruction fidelity with structural preservation:

$$\mathcal{L} = \mathcal{L}_{\text{discrepancy}} + \gamma \mathcal{L}_{\text{constraint}}, \tag{20}$$

where:

$$\mathcal{L}_{\text{discrepancy}} = \frac{1}{MN_s} \sum_{i=1}^{M} \|\tilde{s}_i^{(N)} - s_i\|_2^2, \tag{21}$$

$$\mathcal{L}_{\text{constraint}} = \frac{1}{MN_s} \sum_{i=1}^{M} \sum_{k=1}^{N} \|\tilde{\mathcal{F}}^{(k)}(\mathcal{F}_d^{(k)}(s_i)) - s_i\|_2^2. \tag{22}$$

Here, $\mathcal{L}_{\text{discrepancy}}$ computes the MSE between reconstructed $\tilde{s}_i^{(N)}$ and ground truth s_i, while $\mathcal{L}_{\text{constraint}}$ enforces multi-stage identity constraints through $\tilde{\mathcal{F}}^{(k)} \circ \mathcal{F}_d^{(k)} \approx \mathbf{I}$, with $\gamma = 0.01$ balancing these objectives.

5. Experiments

5.1. Experimental Settings

Training. Using CSIST-100K images as input, we apply Sec. 4.1's method to perform sub-pixel division with a sampling grid ratio of c. This generates an unmixed high-resolution grid as ground truth, where for each target (x_i, y_i, g_i), the intensity g_i is assigned to the pixel at position $\left(c \cdot x_i + \frac{c-1}{2}, c \cdot y_i + \frac{c-1}{2}\right)$ while other pixels remain zero. The selected c value ensures each target appears as a distinct point, enabling accurate spatial separation.

Testing. We (1) apply post-processing with an intensity threshold of 50 to identify predicted targets; (2) project the unmixed grid back to the original 11×11 space via $\left(\frac{x_i - \lfloor \frac{c-1}{2} \rfloor}{c}, \frac{y_i - \lfloor \frac{c-1}{2} \rfloor}{c}\right)$. These mapped coordinates are then compared with ground truth target positions to compute CSO-mAP. Additionally, we calculate PSNR and SSIM by directly comparing the full predicted and ground truth high-resolution images.

Hyperparameters. Our configuration uses: $c = 3$ (baseline grid ratio), batch size of 64, DISTA-Net with 6 stages ($N = 6$), and Dynamic Transform branch coefficient $(1-\alpha) = 0.3$.

5.2. Comparison with State-of-the-Art Methods

Experimental Results. Table 1 compares DISTA-Net with traditional optimization (ISTA), image super-resolution, and deep unfolding methods on the CSIST-100K dataset, evaluated by computational efficiency (#P/FLOPs), localization accuracy (CSO-mAP), and image quality (PSNR/SSIM).

For localization accuracy, we adopt CSO-mAP with different distance thresholds (from AP-05 to AP-25), where a smaller threshold indicates a stricter localization precision requirement. For instance, AP-05 evaluates the detection accuracy within 0.05-pixel distance, representing an extremely high precision demand. Considering the inherent error of 0.236 pixel width at subpixel division factor $c = 3$, AP-20 and AP-25 (with 0.20 and 0.25 precision requirements respectively) serve as primary performance benchmarks. DISTA-Net achieves remarkable accuracy rates of 86.18% and 97.14% accuracy, outperforming most existing methods. On the mAP metric reflecting average CSIST unmixing performance, our method maintains a leading advantage with 46.74% accuracy, demonstrating its robust localization capability across different precision requirements.

In terms of model efficiency, DISTA-Net achieves these results with moderate computational costs (2.179M parameters and 35.103G FLOPs), showing better efficiency compared to methods like ACTNet (46.212M, 62.798G) and HAN (64.342M, 0.495T). Additionally, DISTA-Net demonstrates superior image quality with the highest PSNR (37.8747) and SSIM (99.79) scores among all methods, indicating its excellent capability in preserving image details.

These comprehensive results validate that DISTA-Net achieves an effective balance between computational effi-

Method	#P ↓	FLOPs ↓	CSO-mAP						PSNR ↑	SSIM ↑
			mAP	AP-05	AP-10	AP-15	AP-20	AP-25		
Traditional Optimization										
ISTA [7]	-	-	7.46	0.01	0.31	2.39	9.46	25.14	-	-
Image Super-Resolution										
ACTNet [45]	46.212M	62.80G	45.61	0.38	7.46	41.13	83.12	95.95	35.4526	99.70
CTNet [30]	0.400M	2.756G	45.11	0.38	7.53	40.39	82.11	95.14	35.1499	99.70
DCTLSA [38]	0.865M	13.69G	44.51	0.39	7.35	39.35	81.15	94.34	34.6314	99.65
EDSR [19]	1.552M	12.04G	45.32	0.33	7.07	40.58	83.24	95.41	35.3724	99.71
EGASR [25]	2.897M	17.73G	45.51	0.42	8.03	41.32	85.71	95.08	34.5681	99.66
FeNet [31]	0.348M	2.578G	45.77	0.42	8.19	42.13	83.30	94.80	34.1531	99.66
RCAN [43]	1.079M	8.243G	45.87	0.42	7.96	41.81	83.61	95.57	35.2119	99.69
RDN [44]	22.306M	173.0G	45.81	0.35	7.11	41.07	84.07	96.43	36.4686	99.74
SAN [2]	4.442M	34.05G	45.95	0.36	7.35	41.17	84.32	96.57	36.5037	99.74
SRCNN [8]	0.019M	1.345G	29.06	0.23	4.10	21.65	49.95	69.39	28.7608	98.44
SRFBN [17]	0.373M	3.217G	46.05	0.43	**8.31**	**42.83**	83.72	94.95	34.0174	99.68
HAN [23]	64.342M	495.0G	45.70	0.39	7.46	40.90	83.61	96.17	35.2703	99.71
HiT-SNG [42]	0.952M	13.324G	45.01	0.39	7.34	40.19	81.98	95.17	35.1390	99.71
Deep Unfolding										
ISTA-Net [40]	0.171M	12.77G	45.16	0.41	7.71	40.57	82.58	94.53	33.9215	99.68
ISTA-Net+ [40]	0.337M	24.33G	46.06	0.42	7.66	41.58	84.46	96.17	36.0892	99.72
LAMP [7]	2.126M	0.278G	14.22	0.05	1.11	7.31	21.56	41.06	27.8299	96.89
LIHT [7]	21.10M	1.358G	10.35	0.06	0.92	4.99	14.74	30.5	27.5107	96.42
LISTA [7]	21.10M	1.358G	30.13	0.25	4.13	22.29	51.18	72.82	29.8936	99.12
FISTA-Net [34]	0.074M	18.96G	44.66	**0.45**	7.68	39.74	81.24	94.19	35.7519	99.67
TiLISTA [7]	2.126M	0.278G	14.95	0.06	1.23	7.72	22.50	46.23	27.7038	97.40
⋆ **DISTA-Net (Ours)**	2.179M	35.10G	**46.74**	0.38	7.54	42.44	**86.18**	**97.14**	**37.8747**	**99.79**

Table 1. Comparison with SOTA methods on the **CSIST-100K** dataset.

ciency, localization accuracy, and reconstruction quality.

Visual comparison. Fig. 5 compares reconstruction results with $3\times$ sub-pixel division across methods for scenes containing $3 \sim 5$ targets, assessing performance in dense multi-target scenarios. Existing methods struggle to reconstruct closely spaced targets, exhibiting blurred boundaries or merged detections, with degradation worsening for higher target counts (e.g., five-target cases). In contrast, DISTA-Net preserves both target quantity and sub-pixel positions while maintaining sharp boundaries and accurate spatial distributions, even under extreme density.

5.3. Ablation Study

Method	CSO-mAP	AP-05	AP-10	AP-15	AP-20	AP-25
ISTA-Net [40]	45.16	0.41	7.71	40.57	82.58	94.53
DISTA-Net w/o DT	46.32	0.34	6.83	40.76	86.18	**97.50**
DISTA-Net w/o Thres.	46.17	**0.44**	**7.77**	42.18	84.67	95.79
⋆ **DISTA-Net (Ours)**	**46.74**	0.38	7.54	**42.44**	**86.18**	97.14

Table 2. The effect of different components.

Effect of different components. We conduct ablation studies to evaluate the contribution of each component in DISTA-Net, with results shown in Table 2. The second row (DISTA-Net w/o DT) corresponds to the model variant without Dynamic Transform and the third row (DISTA-Net w/o Thres.) represents the model without Dynamic Soft-Thresholding. Compared to the baseline ISTA-Net, our complete model shows notable improvements in both CSO-mAP (45.16% to 46.74%) and AP-20 (82.58% to 86.18%).

Removing the Dynamic Transform leads to a slight performance decrease (CSO-mAP drops to 46.32%), highlighting its role in enhancing the model's performance. The removal of Dynamic Soft-Thresholding results in the most significant performance degradation (CSO-mAP decreases to 46.17%), emphasizing its key role in ensuring accuracy.

Method	#P ↓	FLOPs ↓	CSO-mAP		
			mAP	AP-10	AP-15
c=5					
ISTA-Net [40]	0.171M	39.544G	66.90	56.73	87.26
ISTA-Net+ [40]	0.225M	48.158G	68.50	57.96	89.52
CFGN [3]	0.538M	4.122G	67.95	58.08	88.35
⋆ **DISTA-Net (Ours)**	5.153M	102.4G	**69.58**	**60.95**	**90.85**
c=7					
ISTA-Net [40]	0.171M	89.51G	71.19	76.45	84.16
ISTA-Net+ [40]	0.225M	103.0G	71.09	74.90	84.90
CFGN [3]	0.548M	4.202G	70.38	73.88	83.97
⋆ **DISTA-Net (Ours)**	6.409M	142.3G	**72.84**	**78.47**	**86.09**

Table 3. Comparison of methods across different Sampling Grids on the **CSIST-100K** dataset.

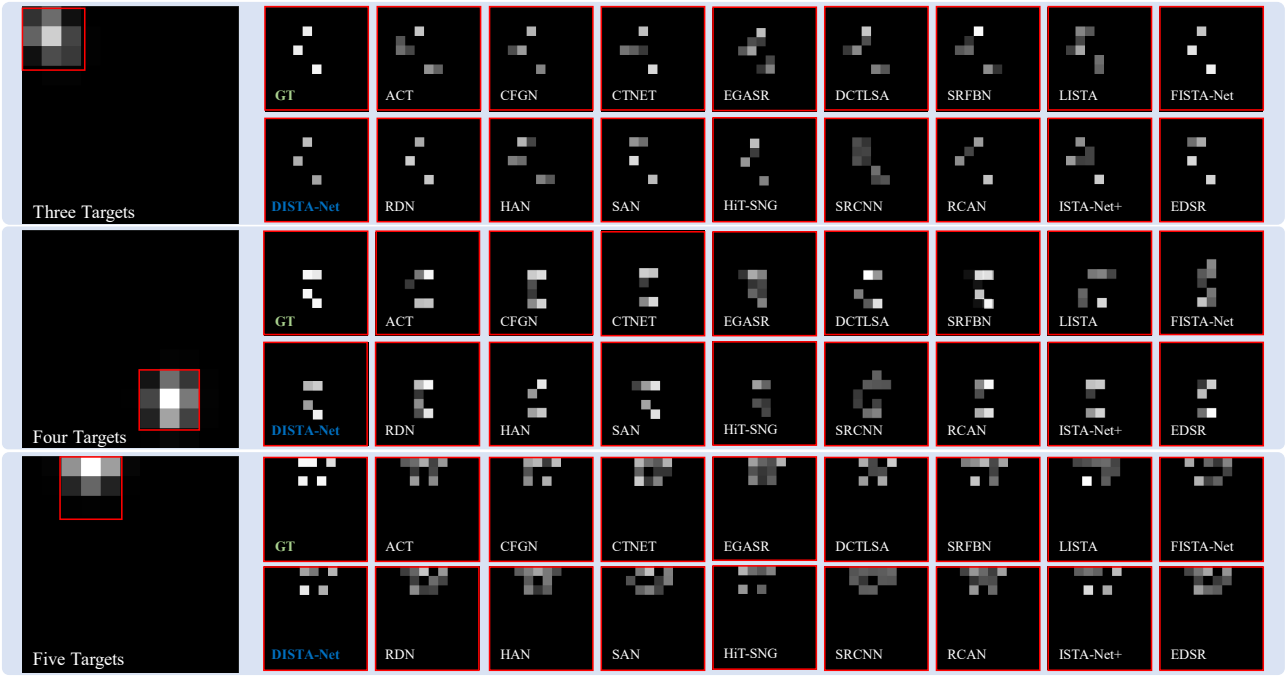

Figure 5. Visual comparison of $3\times$ sub-pixel division reconstruction for scenes containing different numbers of closely-spaced infrared small targets. The red boxes highlight regions where targets exhibit significant sub-pixel characteristics.

Model Performance and Sampling Grid. We evaluate sampling ratios $c = 5$ and 7 across methods, analyzing their impact on performance and efficiency. As the sampling grid ratio increases, all methods show improved detection performance due to better target positioning precision. DISTA-Net consistently outperforms other methods across all configurations. While ISTA-Net and ISTA-Net+ exhibit similar trends with lower overall performance, our proposed method achieves greater accuracy improvements for equivalent increases in sampling ratio. This advantage is particularly pronounced in the AP-10 ($c = 5$) and AP-15 ($c = 7$) metrics (see Table 3 for complete results). However, these performance gains are accompanied by substantially increased computational complexity. We recommend selecting an appropriate sampling grid ratio that aligns with specific application requirements to achieve an optimal balance between detection accuracy and computational efficiency.

Ours vs Super-Resolution + Detector Pipeline. The unmixing stage typically serves as a refinement step following the detection phase. For experimental rigor, we conducted experiments implementing the "SR + Detector" pipeline. We used YOLOv11 as the detector following leading SR methods (retrained on our IR data with unmixing GT for point-source super-resolution). The results demonstrate DISTA-Net's continued advantage: DISTA-Net + YOLOv11 achieved a CSO-mAP of 47.82, outperforming SRFBN + YOLOv11 (45.74) and CFGN + YOLOv11 (46.71). Visual analysis reveals that both conventional SR methods and our unmixing approach generate high-resolution images with well-resolved peaks, enabling effective target separation through simple thresholding. This demonstrates the dominant role of the unmixing stage in this task.

Hyperparameters Analysis. We analyzed model performance versus stage numbers and dynamic branch coefficients in **Supplementary**, demonstrating robust design.

6. Conclusion

In this paper, we present the Dynamic Iterative Shrinkage Thresholding Network (DISTA-Net) to address CSIST unmixing task, which features adaptive generation of both convolution weights and thresholding parameters. Extensive experiments demonstrate that DISTA-Net achieves superior performance in both sub-pixel target detection accuracy and image reconstruction quality. To advance research in this domain, we introduce the CSIST dataset, CSO-mAP metric, and GrokCSO toolkit.

Acknowledgement This research is supported by the NSFC (NO.62206133, 62301261, 62206134, U24A20330, 62361166670, 62225604) and the Shenzhen Science and Technology Program (JCYJ20240813114237048). Computation is supported by the Supercomputing Center of Nankai University.

References

[1] Xiaohan Chen, Jialin Liu, Zhangyang Wang, and Wotao Yin. Hyperparameter tuning is all you need for lista. In *Neural Information Processing Systems (NeurIPS)*, pages 11678–11689, 2021. 1

[2] Tao Dai, Jianrui Cai, Yongbing Zhang, Shu-Tao Xia, and Lei Zhang. Second-order attention network for single image super-resolution. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11065–11074, 2019. 7

[3] Tao Dai, Mengxi Ya, Jinmin Li, Xinyi Zhang, Shu-Tao Xia, and Zexuan Zhu. Cfgn: A lightweight context feature guided network for image super-resolution. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 8(1):855–865, 2023. 7

[4] Yimian Dai, Yiquan Wu, Fei Zhou, and Kobus Barnard. Attentional local contrast networks for infrared small target detection. *IEEE Transactions on Geoscience and Remote Sensing*, 59(11):9813–9824, 2021. 2

[5] Yimian Dai, Yiquan Wu, Fei Zhou, and Kobus Barnard. Asymmetric contextual modulation for infrared small target detection. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 949–958, 2021. 2

[6] Yimian Dai, Xiang Li, Fei Zhou, Yulei Qian, Yaohong Chen, and Jian Yang. One-stage cascade refinement networks for infrared small target detection. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–17, 2023. 2

[7] Ingrid Daubechies, Michel Defrise, and Christine De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications on Pure and Applied Mathematics*, 57(11):1413–1457, 2004. 3, 4, 7

[8] 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. 7

[9] Karol Gregor and Yann LeCun. Learning fast approximations of sparse coding. In *International Conference on Machine Learning (ICML)*, pages 399–406, 2010. 3

[10] Tiantong Guo, Hojjat Seyed Mousavi, and Vishal Monga. Adaptive transform domain image super-resolution via orthogonally regularized deep networks. *IEEE Transactions on Image Processing*, 28(9):4685–4700, 2019. 3

[11] Renke Kou, Chunping Wang, Zhenming Peng, Zhihe Zhao, Yaohong Chen, Jinhui Han, Fuyu Huang, Ying Yu, and Qiang Fu. Infrared small target segmentation networks: A survey. *Pattern Recognition*, 143:109788, 2023. 1

[12] Xiang Li, Wenhai Wang, Xiaolin Hu, and Jian Yang. Selective kernel networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 510–519, 2019. 6

[13] Xin Li, Weisheng Dong, Jinjian Wu, Leida Li, and Guangming Shi. Superresolution image reconstruction: Selective milestones and open problems. *IEEE Signal Processing Magazine*, 40(5):54–66, 2023. 1

[14] Yuelong Li, Mohammad Tofighi, Junyi Geng, Vishal Monga, and Yonina C. Eldar. Efficient and interpretable deep blind image deblurring via algorithm unrolling. *IEEE Transactions on Computational Imaging*, 6:666–681, 2020. 3

[15] Yuxuan Li, Xiang Li, Yunheng Li, Yicheng Zhang, Yimian Dai, Qibin Hou, Ming-Ming Cheng, and Jian Yang. Sm3det: A unified model for multi-modal remote sensing object detection. *arXiv preprint arXiv:2412.20665*, 2024. 2

[16] Yuxuan Li, Xiang Li, Yimain Dai, Qibin Hou, Li Liu, Yongxiang Liu, Ming-Ming Cheng, and Jian Yang. Lsknet: A foundation lightweight backbone for remote sensing. *International Journal of Computer Vision*, 133(3):1410–1431, 2025. 6

[17] Zhen Li, Jinglei Yang, Zheng Liu, Xiaomin Yang, Gwanggil Jeon, and Wei Wu. Feedback network for image super-resolution. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3867–3876, 2019. 7

[18] T Liaudat, J Bonnin, J-L Starck, MA Schmitz, A Guinot, M Kilbinger, and SDJ Gwyn. Multi-ccd modelling of the point spread function. *Astronomy & Astrophysics*, 646:A27, 2021. 4

[19] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) workshops*, pages 136–144, 2017. 7

[20] Jie Mei, Yi-Bo Zheng, and Ming-Ming Cheng. D2anet: Difference-aware attention network for multi-level change detection from satellite imagery. *Computational Visual Media*, 9(3):563–579, 2023. 2

[21] Florian Meyer and Jason L. Williams. Scalable detection and tracking of geometric extended objects. *IEEE Transactions on Signal Processing*, 69:6283–6298, 2021. 1

[22] Vishal Monga, Yuelong Li, and Yonina C. Eldar. Algorithm unrolling: Interpretable, efficient deep learning for signal and image processing. *IEEE Signal Processing Magazine*, 38(2): 18–44, 2021. 3

[23] Ben Niu, Weilei Wen, Wenqi Ren, Xiangde Zhang, Lianping Yang, Shuzhen Wang, Kaihao Zhang, Xiaochun Cao, and Haifeng Shen. Single image super-resolution via a holistic attention network. In *European Conference on Computer Vision (ECCV)*, pages 191–207, 2020. 7

[24] K. Pruett, N. McNaughton, and M. Schneider. Closely Spaced Object Classification Using MuyGPyS. In *Advanced Maui Optical and Space Surveillance (AMOS) Technologies Conference*, page 158, 2023. 1

[25] Zhonghang Qiu, Huanfeng Shen, Linwei Yue, and Guizhou Zheng. Cross-sensor remote sensing imagery super-resolution via an edge-guided attention-based network. *ISPRS Journal of Photogrammetry and Remote Sensing*, 199:226–241, 2023. 7

[26] José A. Sobrino, Fabio Del Frate, Matthias Drusch, Juan C. Jiménez-Muñoz, Paolo Manunta, and Amanda Regan. Review of thermal infrared applications and requirements for future high-resolution sensors. *IEEE Transactions on Geoscience and Remote Sensing*, 54(5):2963–2972, 2016. 1

[27] Oren Solomon, Regev Cohen, Yi Zhang, Yi Yang, Qiong He, Jianwen Luo, Ruud J. G. van Sloun, and Yonina C. Eldar. Deep unfolded robust PCA with application to clutter

suppression in ultrasound. *IEEE Transactions on Medical Imaging*, 39(4):1051–1063, 2020. 3

[28] Xiaozhong Tong, Shaojing Su, Peng Wu, Runze Guo, Junyu Wei, Zhen Zuo, and Bei Sun. Msaffnet: A multiscale label-supervised attention feature fusion network for infrared small target detection. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–16, 2023. 3

[29] Zhiqiang Wan, Haibo He, and Bo Tang. A generative model for sparse hyperparameter determination. *IEEE Transactions on Big Data*, 4(1):2–10, 2018. 1

[30] Shunzhou Wang, Tianfei Zhou, Yao Lu, and Huijun Di. Contextual transformation network for lightweight remote-sensing image super-resolution. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–13, 2021. 7

[31] Zheyuan Wang, Liangliang Li, Yuan Xue, Chenchen Jiang, Jiawen Wang, Kaipeng Sun, and Hongbing Ma. Fenet: Feature enhancement network for lightweight remote-sensing image super-resolution. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–12, 2022. 7

[32] Zhe Wang, Tao Zang, Zhiling Fu, Hai Yang, and Wenli Du. Rlpgb-net: Reinforcement learning of feature fusion and global context boundary attention for infrared dim small target detection. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–15, 2023. 2

[33] Gang Wu, Junjun Jiang, Kui Jiang, and Xianming Liu. Fully $1\times$ 1 convolutional network for lightweight image super-resolution. *Machine Intelligence Research*, 21(6):1062–1076, 2024. 1

[34] Jinxi Xiang, Yonggui Dong, and Yunjie Yang. Fista-net: Learning a fast iterative shrinkage thresholding network for inverse problems in imaging. *IEEE Transactions on Medical Imaging*, 40(5):1329–1339, 2021. 7

[35] Hai Xu, Sheng Zhong, Tianxu Zhang, and Xu Zou. Multiscale multilevel residual feature fusion for real-time infrared small target detection. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–16, 2023. 2

[36] yan yang, Jian Sun, Huibin Li, and Zongben Xu. Deep admm-net for compressive sensing mri. In *Neural Information Processing Systems (NeurIPS)*, 2016. 3

[37] Di You, Jingfen Xie, and Jian Zhang. Ista-net++: Flexible deep unfolding network for compressive sensing. In *IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, 2021. 3

[38] Kun Zeng, Hanjiang Lin, Zhiqiang Yan, and Jinsheng Fang. Densely connected transformer with linear self-attention for lightweight image super-resolution. *IEEE Transactions on Instrumentation and Measurement*, 72:1–12, 2023. 7

[39] Hui Zhang, Hui Xu, and Liangkui Lin. Super-resolution method of closely spaced objects based on sparse reconstruction using single frame infrared data. *Acta Optica Sinica*, 33 (4):0411001–8, 2013. 1

[40] Jian Zhang and Bernard Ghanem. Ista-net: Interpretable optimization-inspired deep network for image compressive sensing. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1828–1837, 2018. 3, 7

[41] Mingjin Zhang, Qian Xu, Yuchun Wang, Xi Li, and Haojuan Yuan. Mirsam: multimodal vision-language segment anything model for infrared small target detection. *Visual Intelligence*, 3(1):1–13, 2025. 3

[42] Xiang Zhang, Yulun Zhang, and Fisher Yu. Hit-sr: Hierarchical transformer for efficient image super-resolution. In *European Conference on Computer Vision (ECCV)*, pages 483–500, 2024. 7

[43] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *European Conference on Computer Vision (ECCV)*, pages 286–301, 2018. 7

[44] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2472–2481, 2018. 7

[45] Zheng Zhang, Fanchen Liu, Changan Liu, Qing Tian, and Hongquan Qu. Actnet: A dual-attention adapter with a cnn-transformer network for the semantic segmentation of remote sensing imagery. *Remote Sensing*, 15(9):2363, 2023. 7

[46] Mingjing Zhao, Wei Li, Lu Li, Jin Hu, Pengge Ma, and Ran Tao. Single-frame infrared small-target detection: A survey. *IEEE Geoscience and Remote Sensing Magazine*, 10(2):87–119, 2022. 1