CISTA-NET: COUNTING MATTERS FOR CLOSELY-SPACED INFRARED SMALL TARGET UNMIXING

Anonymous authors
Paper under double-blind review

ABSTRACT

Closely-spaced infrared small target unmixing is an ill-conditioned inverse problem in which multiple sparse source configurations can produce highly similar observations under severe point-spread-function overlap. Existing deep unfolding methods adapt reconstruction parameters primarily from local or intermediate reconstruction features, but do not explicitly exploit scene-level source cardinality when deciding how aggressively sparse responses should be retained or suppressed. We investigate whether estimated target cardinality can serve as a global structural cue for iterative reconstruction, and instantiate this idea in CISTA-Net, a count-guided iterative shrinkage-thresholding network. CISTA-Net introduces a count-prior guided dynamic threshold generator (CP-DTG) that conditions stage-wise shrinkage on both local reconstruction features and a predicted count distribution. It further incorporates multi-stage image selective fusion (MSIF) to recover complementary intermediate responses and count-guided post-processing (CGPP) for count-consistent candidate selection. On CSIST-100K, CISTA-Net improves CSO-mAP over DISTA-Net from 46.74 to 47.28, 69.58 to 70.25, and 72.84 to 73.06 at sampling ratios $c = 3, 5, 7$, respectively, with the clearest gains under dense and moderately precise localization settings. These results suggest that explicit cardinality information provides complementary global guidance to locally adaptive deep unfolding.

1 INTRODUCTION

Infrared imaging supports detection in conditions where visible appearance is unreliable, but distant targets can occupy less than one detector pixel. When several sources are close together, their point-spread functions (PSFs) overlap and form a single blurred response. Identifying the individual sources then requires more than detecting the presence of a bright region. Closely-spaced infrared small target (CSIST) unmixing estimates the number, sub-pixel locations, and intensities of the underlying targets (Han et al., 2025). The difficulty is not merely low spatial resolution: under strong overlap, different source configurations may induce visually similar measurements, making the inverse mapping highly ambiguous (Kou et al., 2023).

A natural formulation models the observation as a noisy linear combination of point-source responses and discretizes candidate source locations, yielding an underdetermined sparse inverse problem. Iterative shrinkage-thresholding (ISTA) algorithms resolve this ambiguity by alternating data-consistency updates with sparsity-inducing shrinkage (Daubechies et al., 2004), while learned unfolding replaces fixed transformations and thresholds with trainable counterparts (Gregor & LeCun, 2010; Zhang & Ghanem, 2018). DISTA-Net further makes these transformations and thresholds input adaptive for CSIST unmixing (Han et al., 2025). More recently, DISTA-Net++ introduces target cardinality as an explicit semantic prior through a Count-Guided Prior (CGP), together with continuous coordinate rectification for off-grid localization (Xu et al., 2026).

These developments raise a more specific question that remains insufficiently explored: how should cardinality information interact with the iterative shrinkage process itself? Existing count guidance demonstrates that source cardinality is informative for resolving ambiguous mixtures, but treating count primarily as a global feature-level prior does not explicitly determine how strongly responses should be retained or suppressed at each reconstruction stage. Since the shrinkage threshold directly controls this sparsity trade-off, we hypothesize that cardinality should condition the

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

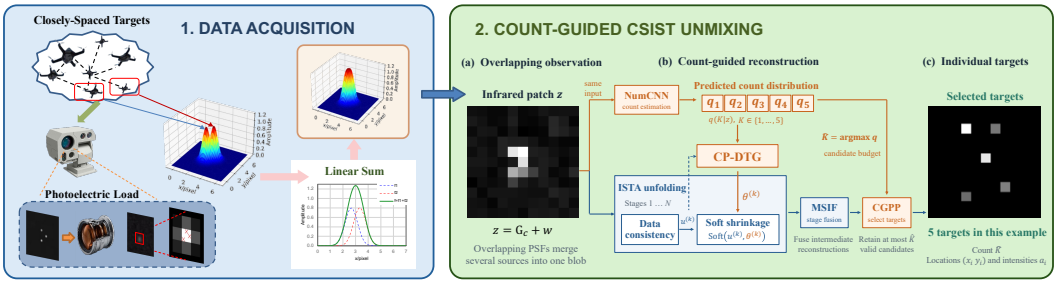


Figure 1: Overview of CSIST acquisition and count-guided unmixing. CISTA-Net uses the predicted target-count distribution as global guidance for stage-wise threshold generation, while MSIF preserves intermediate evidence and CGPP enforces count-consistent target selection.

threshold-generation mechanism itself. In other words, rather than merely informing the reconstructed representation, target count should directly modulate the optimization-inspired operator that governs stage-wise source preservation and suppression.

To test this hypothesis, as illustrated in Figure 1, we develop **CISTA-Net**. Its core mechanism, the **count-prior guided dynamic threshold generator (CP-DTG)**, injects the predicted count distribution directly into threshold generation at every unfolding stage, coupling scene-level cardinality with local reconstruction evidence. Multi-stage image selective fusion (MSIF) further recovers complementary responses attenuated during repeated shrinkage, while count-guided post-processing (CGPP) encourages count-consistent candidate selection. Thus, CISTA-Net focuses specifically on **cardinality-conditioned shrinkage**, complementing prior count-guided formulations by integrating count information more tightly into the iterative reconstruction dynamics.

The contributions of this work are threefold. (1) We study cardinality-conditioned shrinkage for CSIST unmixing, showing how a predicted source-count distribution can directly guide stage-wise threshold generation in learned iterative reconstruction. (2) We instantiate this idea in CISTA-Net, whose CP-DTG couples global cardinality with local reconstruction evidence at every unfolding stage, complemented by multi-stage image selective fusion and count-consistent candidate selection. (3) We evaluate CISTA-Net on CSIST-100K across multiple sampling-grid resolutions. The results show consistent improvements in overall CSO-mAP over DISTA-Net, while component ablations support the contributions of count conditioning and multi-stage fusion.

2 Related Work

Infrared small target detection. Infrared small target detection (IRSTD) aims to identify dim target regions against complex backgrounds. Classical approaches exploit low-rank background structure, local contrast, and hand-crafted priors, while recent learning-based methods improve target-background discrimination through multi-scale feature aggregation and attention mechanisms (Gao et al., 2013; Dai & Wu, 2017; Dai et al., 2021; Zhang et al., 2022). CSIST unmixing addresses a fundamentally different output space: when several point sources are merged by optical diffraction, the goal is not merely to detect the resulting blob, but to recover the number, sub-pixel locations, and intensities of the underlying sources. We therefore treat IRSTD as a preceding detection stage rather than a competing formulation.

Deep unfolding. Deep unfolding maps iterations of model-based optimization algorithms into trainable network stages, combining physically motivated update rules with learned representations. LISTA learns finite-depth approximations to sparse coding (Gregor & LeCun, 2010), while ADMM-Net and ISTA-Net extend this paradigm to imaging inverse problems (Yang et al., 2016; Zhang & Ghanem, 2018). Our work asks a more specific question: can global target count directly control the stage-wise shrinkage operator? CISTA-Net addresses this question by conditioning threshold generation on a predicted count distribution at every unfolding stage.

CSIST unmixing and count-guided reconstruction. DISTA-Net introduced dynamic deep unfolding for single-frame CSIST unmixing by adapting transformation and threshold parameters to each observation (Han et al., 2025), while SeqCSIST subsequently explored temporal information aggregation in a multi-frame setting (Zhai et al., 2025). More recently, DISTA-Net++ introduced a count-guided prior together with continuous coordinate rectification, using predicted target cardinality to provide global semantic guidance for dynamic unfolding and count-constrained inference (Xu et al., 2026). Our work shares the motivation of exploiting target cardinality, but focuses on a different aspect of count-guided reconstruction: *how global count information should interact with local reconstruction evidence inside the stage-wise threshold-generation process*. Thus, our emphasis is on the design of count-conditioned shrinkage and cross-stage reconstruction, rather than continuous coordinate rectification.

3 Problem Formulation and Count-Guided Shrinkage

An observed patch contains $U \times V$ detector pixels and K point targets. Let $\mathbf{z} \in \mathbb{R}^{UV}$ denote the vectorized measurement. Integrating a Gaussian PSF over each detector pixel yields a response vector for any hypothesized source location. With $c \times c$ sub-pixels per detector pixel, there are $L = UVc^2$ candidate locations on a grid Ω_c. The discretized forward model is

$$\mathbf{z} = \mathbf{G}_c\mathbf{s} + \mathbf{w}, \qquad \mathbf{G}_c \in \mathbb{R}^{UV\times L}, \quad \mathbf{s} \in \mathbb{R}^{L}, \tag{1}$$

where nonzero entries of $\mathbf{s}$ encode source amplitudes, and $\mathbf{w}$ includes measurement noise and discretization error. The physical amplitudes are nonnegative; the learned network described below uses signed feature-space soft-thresholding. The detailed imaging model and coordinate convention are given in Appendix A.

A standard sparse reconstruction objective is

$$\min_{\mathbf{s}} \ \frac{1}{2}\|\mathbf{z} - \mathbf{G}_c\mathbf{s}\|_2^2 + \lambda\|\mathbf{s}\|_1. \tag{2}$$

With step size ρ, ISTA updates the estimate through

$$\mathbf{r}^{(k)} = \mathbf{s}^{(k-1)} - \rho\mathbf{G}_c^\top(\mathbf{G}_c\mathbf{s}^{(k-1)} - \mathbf{z}), \tag{3}$$

$$\mathbf{s}^{(k)} = \mathrm{Soft}(\mathbf{r}^{(k)}, \rho\lambda), \tag{4}$$

where $\mathrm{Soft}(a,t) = \mathrm{sign}(a)\max(|a| - t, 0)$ for $t \geq 0$. The factor $1/2$ in Eq. (2) makes its gradient consistent with Eq. (3). Learned feature transformations and data-dependent thresholds generalize this construction, but are not necessarily exact proximal operators for a fixed objective.

The ℓ_1 regularization in Eq. (2) encourages a sparse solution, but it does not explicitly specify how many sources should be preserved. When multiple targets are closely spaced, severe PSF overlap can make different sparse source configurations produce similar measurements, leaving the reconstruction ambiguous. In ISTA-based unfolding, this ambiguity is closely related to the shrinkage operation: the threshold determines which candidate responses survive and which are suppressed. A threshold that is too large may eliminate weak but valid sources, whereas a small threshold may retain spurious responses.

If the source cardinality K were known, it would provide additional structural information about the support of the underlying source map,

$$|\mathrm{supp}(\mathbf{s})| \approx K, \tag{5}$$

where $\mathrm{supp}(\mathbf{s})$ denotes the set of active source locations. Rather than imposing this combinatorial constraint explicitly, we estimate a distribution over source cardinality from the observation and use it as global structural guidance for the learned shrinkage process. In particular, we ask whether the predicted cardinality distribution can be coupled with local reconstruction evidence to adapt the stage-wise threshold,

$$\boldsymbol{\theta}^{(k)} = f_\theta^{(k)}\left(\mathbf{u}^{(k)},\, q(K \mid \mathbf{z})\right), \tag{6}$$

where $\mathbf{u}^{(k)}$ denotes the reconstruction feature at stage k, and $q(K \mid \mathbf{z})$ is the predicted distribution over source counts. This formulation motivates the count-guided unfolding architecture introduced in the next section.

4 COUNT-GUIDED DEEP UNFOLDING

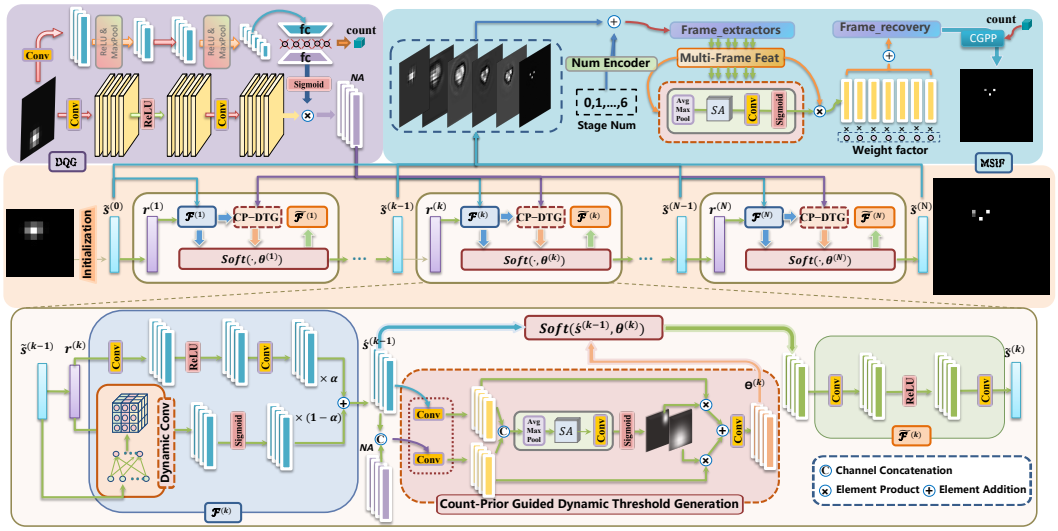


Figure 2: Overview of CISTA-Net. The predicted count distribution is embedded into count-aware features and injected into CP-DTG to guide stage-wise shrinkage. MSIF aggregates complementary intermediate reconstructions, followed by count-guided candidate selection.

The overall framework of CISTA-Net is illustrated in Figure 2. It consists of count-aware stage-wise shrinkage, multi-stage information fusion, and count-guided candidate selection. Specifically, CISTA-Net explicitly incorporates target-count information into stage-wise threshold generation. Given N reconstruction stages, we initialize $\mathbf{s}^{(0)} = \mathbf{Q}_{\text{init}}\mathbf{z}$. At stage k, the data-consistency update in Eq. (3) is transformed into a reconstruction feature,

$$\mathbf{u}^{(k)} = \mathcal{F}_d^{(k)}(\mathbf{s}^{(k-1)}, \mathbf{r}^{(k)}), \tag{7}$$

$$\boldsymbol{\theta}^{(k)} = \text{CP-DTG}^{(k)}(\mathbf{u}^{(k)}, \mathbf{A}), \tag{8}$$

$$\mathbf{s}^{(k)} = \widetilde{\mathcal{F}}^{(k)}\Big(\text{Soft}(\mathbf{u}^{(k)}, \boldsymbol{\theta}^{(k)})\Big). \tag{9}$$

Here $\mathbf{A}$ denotes the count-aware feature tensor. Unlike reconstruction-driven dynamic thresholding, Eq. (8) conditions the shrinkage threshold jointly on local reconstruction evidence and global target-count information. The learned synthesis transform $\widetilde{\mathcal{F}}^{(k)}$ approximately inverts the analysis transform under the consistency loss described in Sec. 4.4. The count estimate is computed from the same observation as the reconstruction and therefore adds no independent measurement. Its role is to bias a finite-depth solver toward a sparsity pattern learned from the training distribution, rather than to resolve non-identifiability by itself. The data-consistency update remains explicit at each stage, so count information changes the shrinkage pathway without replacing the forward model.

4.1 COUNT-PRIOR GUIDED DYNAMIC THRESHOLD GENERATION

CP-DTG is the core component of CISTA-Net. Its key idea is to use the predicted target-count distribution as global guidance for determining how aggressively responses should be preserved or suppressed at each unfolding stage.

Count estimation and embedding. Given an observed patch $\mathbf{Z}$, NumCNN predicts $\mathbf{q} = \text{NumCNN}(\mathbf{Z}) \in [0,1]^J$, where $J = 5$ corresponds to $K \in \{1, \ldots, 5\}$ and $\sum_j q_j = 1$. Rather than using only the most probable count, CP-DTG retains the full distribution $\mathbf{q}$ as a soft conditioning signal. A two-layer MLP converts $\mathbf{q}$ into channel-wise modulation weights:

$$\mathbf{g}_c = \text{Sigmoid}\left(\mathbf{W}_2\sigma(\mathbf{W}_1\mathbf{q})\right), \qquad \mathbf{A} = \mathbf{F}_0 \odot \mathbf{g}_c, \tag{10}$$

where σ denotes ReLU and $\mathbf{F}_0$ is a shallow image-feature representation. Thus, $\mathbf{A}$ combines image content with global count information and is reused across unfolding stages. For the concatenation below, $\mathbf{F}_0$ and $\mathbf{u}^{(k)}$ must be represented on the same $Uc \times Vc$ reconstruction grid; $\mathbf{g}_c$ is broadcast over the two spatial dimensions. Here the subscript c in $\mathbf{g}_c$ and in Avg_c denotes channels, whereas c alone denotes sub-pixels per detector pixel.

Dual-path threshold modulation. At stage k, CP-DTG combines the local reconstruction feature $\mathbf{u}^{(k)}$ with the count-aware feature $\mathbf{A}$. After concatenation and grouped convolution, two parallel projections produce $\mathbf{F}_1$ from the local reconstruction feature and $\mathbf{F}_2$ from the count-conditioned fused feature. Spatial weights are generated by channel-wise average and maximum pooling:

$$\mathbf{w} = \mathrm{Sigmoid}\left(\mathcal{C}_{3\times3}([\mathrm{Avg}_c(\mathbf{B}), \mathrm{Max}_c(\mathbf{B})])\right), \qquad \mathbf{B} = [\mathbf{F}_1, \mathbf{F}_2]. \tag{11}$$

The resulting two spatial maps $\mathbf{w}_1$ and $\mathbf{w}_2$ adaptively balance local reconstruction evidence and count-guided context. The stage-wise threshold is then generated as

$$\boldsymbol{\theta}^{(k)} = \mathbf{u}^{(k)} \odot \mathcal{C}_{1\times1}\left(\mathbf{F}_1 \odot \mathbf{w}_1 + \mathbf{F}_2 \odot \mathbf{w}_2\right). \tag{12}$$

This design allows the shrinkage strength to vary spatially while being conditioned on the global source-count estimate. Therefore, CP-DTG couples *where* candidate responses appear with *how many* responses are expected to survive, directly injecting count guidance into the iterative shrinkage process. The count distribution can distinguish two observations even when their most probable counts agree; the discrete count is used only after reconstruction. A high probability for a larger count does not impose an exact support size on any individual stage output.

4.2 Multi-Stage Image Selective Fusion

Repeated shrinkage may attenuate weak target responses that remain informative at intermediate stages. MSIF therefore aggregates intermediate reconstructions as a residual correction to the terminal output. For each stage $k \in \mathcal{I}$, a stage-specific Conv–BN–ReLU extractor and spatial attention module produce an attended feature $\mathbf{h}^{(k)}$. Learnable scalar logits ω_k determine normalized stage weights:

$$a_k = \frac{\exp(\omega_k)}{\sum_{j\in\mathcal{I}} \exp(\omega_j)}, \qquad \mathbf{f}_{\mathrm{fused}} = \sum_{k\in\mathcal{I}} a_k \mathbf{h}^{(k)}. \tag{13}$$

A recovery network $\mathcal{R}$ maps the fused representation to a residual correction,

$$\widehat{\mathbf{s}} = \mathbf{s}^{(N)} + \beta \mathbf{s}_{\mathrm{MSIF}}, \qquad \mathbf{s}_{\mathrm{MSIF}} = M_0 \tanh\left(\mathcal{R}(\mathbf{f}_{\mathrm{fused}})\right), \tag{14}$$

where $M_0 = 50$ and $\beta = 0.5$. MSIF complements CP-DTG by recovering useful evidence across unfolding stages rather than modifying the shrinkage rule itself. The scalar weights a_k are learned across the dataset, while the spatial attention inside $\mathbf{h}^{(k)}$ varies with the input. Because $|\tanh(x)| \le 1$, the residual term changes each grid response by at most $\beta M_0 = 25$ in the stated response units; this bound alone does not imply a localization improvement.

4.3 Count-Guided Candidate Selection

While CP-DTG uses the full count distribution during reconstruction, final source extraction requires a discrete target count. We therefore use

$$\widehat{K} = \arg\max_{j\in\{1,\ldots,J\}} q_j \tag{15}$$

to guide candidate selection from $\widehat{\mathbf{s}}$. With the response threshold $\tau_{\mathrm{cand}} = 50$ used in the reported evaluation, candidates are grid responses at or above τ_{cand}. They are sorted by response intensity and greedily retained subject to a minimum separation $d_{\min}$ until $\widehat{K}$ targets have been selected or no valid candidates remain. Thus, CP-DTG uses count information to modify the reconstructed source map, whereas candidate selection only enforces count consistency on the final responses. Distances are measured after mapping grid indices back to detector-pixel coordinates. The rule retains at most $\widehat{K}$ candidates: an underestimated count can exclude a true source, and a separation constraint can prevent selection of $\widehat{K}$ candidates even when that many responses are present. It does not reconstruct a source absent from the candidate set. The complete inference procedure is provided in Appendix B.

4.4 Learning Objective

For training observations $\mathbf{Z}_{\mathrm{tr}}$ and ground-truth source maps $\mathbf{S}_{\mathrm{tr}}$, the least-squares initializer is $\mathbf{Q}_{\mathrm{init}} = \mathbf{S}_{\mathrm{tr}}\mathbf{Z}_{\mathrm{tr}}^{\dagger}$. CISTA-Net is trained with reconstruction, transform-consistency, fusion, and count-estimation objectives:

$$\mathcal{L} = \mathcal{L}_{\mathrm{rec}} + \gamma\mathcal{L}_{\mathrm{sym}} + \mu\mathcal{L}_{\mathrm{fuse}} + \mathcal{L}_{\mathrm{count}}, \qquad \gamma = \mu = 0.01. \tag{16}$$

For batch size B and signal dimension L, the individual losses are

$$\mathcal{L}_{\mathrm{rec}} = \frac{1}{BL}\sum_{i=1}^{B} \|\mathbf{s}_i^{(N)} - \mathbf{s}_i^{\star}\|_2^2, \tag{17}$$

$$\mathcal{L}_{\mathrm{sym}} = \frac{1}{BL}\sum_{i=1}^{B}\sum_{k=1}^{N} \|\widetilde{\mathcal{F}}^{(k)}(\mathcal{F}_d^{(k)}(\mathbf{s}_i^{\star})) - \mathbf{s}_i^{\star}\|_2^2, \tag{18}$$

$$\mathcal{L}_{\mathrm{fuse}} = \frac{1}{BL}\sum_{i=1}^{B} \|\widehat{\mathbf{s}}_i - \mathbf{s}_i^{\star}\|_2^2, \tag{19}$$

$$\mathcal{L}_{\mathrm{count}} = -\frac{1}{B}\sum_{i=1}^{B} \log q_{i,K_i}. \tag{20}$$

The reconstruction loss supervises the terminal unfolding output, the fusion loss supervises the MSIF-corrected result, and cross-entropy trains the count estimator whose prediction conditions CP-DTG. Intermediate stage estimates supplied to MSIF are detached from the unfolding branch during fusion training.

5 Experiments

5.1 Setup and Evaluation

We evaluate the reported experiments on CSIST-100K (Han et al., 2025), using 11×11 grayscale patches and sampling ratios $c \in \{3, 5, 7\}$. The baseline setting is $c = 3$, yielding a 33×33 source map. The reported configuration uses six unfolding stages, batch size 64, Adam, and a fixed learning rate of 10^{-4}. The hyperparameter plot favors $\alpha = 0.5$ for the dynamic-transform branch weight. Appendix D records the available settings.

Metrics. CSO-mAP averages AP at localization tolerances $0.05, 0.10, 0.15, 0.20, 0.25$ in detector-pixel units. We retain the notation AP-05 through AP-25. Predicted grid locations are mapped back to detector coordinates before evaluation. The reported evaluation uses a response threshold of 50 for candidate extraction; the same candidate-threshold and matching protocol must be used in any fair post-processing comparison. We also report PSNR in dB and SSIM on the source maps, with SSIM values displayed as $100 \times$ SSIM.

Comparison scope. We compare the reported values for optimization, image super-resolution, and unfolding baselines. The main table emphasizes strong super-resolution baselines and closely related unfolding methods. All remaining rows and computational figures are preserved in Appendix C.

5.2 Quantitative Comparison

At $c = 3$, CISTA-Net achieves 47.28 CSO-mAP versus 46.74 for DISTA-Net, an increase of 0.54 percentage points (Table 1). AP-20 increases by 1.29 points and AP-25 by 0.65 points. AP-15 also improves relative to DISTA-Net, but is slightly below CFGN (43.22 versus 43.23). At AP-05, CISTA-Net is below both DISTA-Net and CFGN; at AP-10 it ties DISTA-Net and remains below CFGN. The improvement is therefore concentrated at selected tolerances, rather than uniformly improving localization precision.

Table 1 focuses on the primary detection metrics, while the complete comparison, including PSNR, SSIM, parameters, and FLOPs, is provided in Appendix C. CISTA-Net achieves the highest reported

Table 1: Comparison with state-of-the-art methods on the CSIST-100K dataset. The best results are highlighted in bold, and the second-best results are underlined. Please refer to Appendix C for the complete comparison with all methods and evaluation metrics.

Method	CSO-mAP	AP-05	AP-10	AP-15	AP-20	AP-25
Traditional Optimization						
ISTA Daubechies et al. (2004)	7.46	0.01	0.31	2.39	9.46	25.14
Image Super-Resolution						
SRFBN Li et al. (2019)	46.05	0.43	8.31	42.83	83.72	94.95
HAN Niu et al. (2020)	45.70	0.39	7.46	40.90	83.61	96.17
CFGN Dai et al. (2023)	46.67	**0.49**	**8.45**	**43.23**	85.03	96.13
Deep Unfolding						
ISTA-Net Zhang & Ghanem (2018)	45.16	0.41	7.71	40.57	82.58	94.53
ISTA-Net+ Zhang & Ghanem (2018)	46.06	0.42	7.66	41.58	84.46	96.17
FISTA-Net Xiang et al. (2021)	44.66	0.45	7.68	39.74	81.24	94.19
TiLISTA Daubechies et al. (2004)	14.95	0.06	1.23	7.72	22.50	46.23
DISTA-Net Han et al. (2025)	46.74	0.38	7.54	42.44	86.18	97.14
⋆ **CISTA-Net (Ours)**	**47.28**	0.36	7.54	43.22	**87.47**	**97.79**

CSO-mAP and PSNR in the full comparison. Specifically, its PSNR improves from 37.8747 dB for DISTA-Net to 37.8875 dB, while its SSIM increases slightly from 98.80 to 98.83, although the latter remains below the best reported SSIM (99.73 for CFGN). CISTA-Net requires 2.201M parameters and 35.247G FLOPs, compared with 2.179M parameters and 35.103G FLOPs for DISTA-Net, corresponding to only 1.01% and 0.41% increases, respectively. Thus, the improvements in detection and reconstruction performance are obtained with only marginal additional model complexity.

5.3 Ablation Study

To evaluate the contribution of each component, we conduct ablation experiments on the CSIST-100K dataset with c=3, as reported in Table 2. Using DISTA-Net as the baseline, incrementally incorporating CP-DTG and MSIF yields a monotonically increasing CSO-mAP, with the full CISTA-Net achieving 47.28—a gain of 0.54 over the baseline—confirming the effectiveness and synergistic benefit of both modules. Specifically, adding MSIF alone (CISTA-Net w/o CP-DTG) improves CSO-mAP from 46.74 to 46.84, with AP-20 and AP-25 rising to 86.78 and 97.38, demonstrating that multi-stage fusion improves cross-stage information utilization. Adding CP-DTG alone (CISTA-Net w/o MSIF) further pushes CSO-mAP to 47.01 with AP-15 reaching 43.01, indicating that count prior is particularly beneficial for medium-precision localization. When both modules are enabled, CISTA-Net achieves the best performance across mAP, AP-15, AP-20, and AP-25, with AP-20 reaching 87.47, highlighting the complementarity of the two modules in dense target localization. The additional computational cost—2.201M parameters and 35.247G FLOPs versus 2.179M and 35.103G for DISTA-Net—remains modest, preserving the lightweight nature of the overall architecture.

Table 2: The effect of different parts of our method. The best results are highlighted in bold, and the second-best results are underlined.

Method	CSO-mAP	AP-05	AP-10	AP-15	AP-20	AP-25
DISTA-Net	46.74	**0.38**	7.54	42.44	86.18	97.14
CISTA-Net w/o DQM	46.84	0.34	**7.61**	42.10	86.78	97.38
CISTA-Net w/o MSIF	47.01	0.36	7.52	43.01	86.74	97.43
⋆ **CISTA-Net (Ours)**	**47.28**	0.36	7.54	**43.22**	**87.47**	**97.79**

5.4 Parameter Analysis

(1) Performance vs. Sampling Grid. Table 3 compares methods across three sampling grid settings (c=3, 5, 7) on CSIST-100K. As c increases, all methods exhibit significant CSO-mAP gains, consistent with the reduced localization difficulty under sparser target distributions. CISTA-Net

Table 3: Comparison of methods across different Sampling Grids on the CSIST-100K dataset. The best results are highlighted in bold, and the second-best results are underlined.

Method	CSO-mAP	AP-05	AP-10	AP-15	AP-20	AP-25
c=5						
ISTA-Net Zhang & Ghanem (2018)	66.90	5.60	56.73	87.26	91.55	93.35
ISTA-Net+ Zhang & Ghanem (2018)	68.50	5.80	57.96	89.52	93.78	95.47
CFGN Dai et al. (2023)	67.95	**6.59**	58.08	88.35	92.60	94.15
DISTA-Net Han et al. (2025)	<u>69.58</u>	<u>6.28</u>	<u>60.95</u>	<u>90.85</u>	<u>94.26</u>	<u>95.57</u>
⋆ **CISTA-Net (Ours)**	**70.25**	6.10	**61.00**	**91.72**	**95.49**	**96.92**
c=7						
ISTA-Net Zhang & Ghanem (2018)	71.19	20.90	76.45	84.16	86.60	87.85
ISTA-Net+ Zhang & Ghanem (2018)	71.09	18.28	74.17	84.97	87.82	89.21
CFGN Dai et al. (2023)	70.38	18.48	74.10	84.05	86.92	88.34
DISTA-Net Han et al. (2025)	<u>72.84</u>	<u>21.32</u>	<u>78.47</u>	**86.09**	<u>88.55</u>	**89.76**
⋆ **CISTA-Net (Ours)**	**73.06**	**21.72**	**79.34**	<u>86.01</u>	**88.70**	<u>89.53</u>

consistently ranks first across all settings, achieving CSO-mAP of **47.28**, **70.25**, and **73.06** for c=3, 5, 7 respectively. At c=3 (dense), CISTA-Net improves CSO-mAP by 0.54 and AP-25 by 0.65 over DISTA-Net, demonstrating the effectiveness of count prior in highly cluttered scenes. At c=5 (moderate), CISTA-Net achieves a higher CSO-mAP (+0.67) while using *fewer* parameters (4.741M vs. 5.153M) and FLOPs (91.751G vs. 102.4G), reflecting a favorable efficiency-accuracy trade-off. At c=7 (sparse), CISTA-Net leads on CSO-mAP, AP-05, and AP-10, while marginally trailing DISTA-Net on AP-15 and AP-25; this is attributed to the larger inter-target spacing reducing the advantage of fine-grained count guidance at high-precision thresholds, with differences falling within normal variation. Overall, the proposed count-guided mechanism demonstrates consistent adaptability across target density regimes, with more pronounced gains in dense and moderate settings.

(2) Hyperparameter Sensitivity.

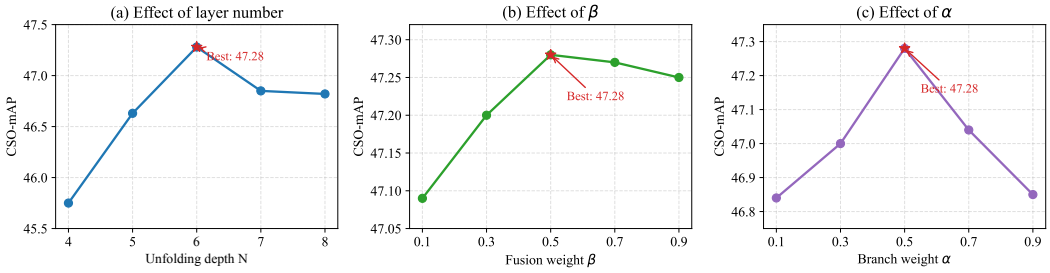


Figure 3: Sensitivity curves for unfolding depth N, fusion weight β, and transform branch weight α. The displayed peaks occur at $N = 6$, $\beta = 0.5$, and $\alpha = 0.5$.

We analyze the sensitivity to unfolding depth N, fusion weight β, and dynamic transform branch weight α in Figure 3. CSO-mAP increases from 45.75 at N=4 to a peak of 47.28 at N=6, then decreases for N=7, 8, so we set N=6. Performance also peaks at β=0.5 and α=0.5, indicating that moderate multi-stage fusion and a balanced dynamic/base transform contribution are most effective. Accordingly, we use N=6, β=0.5, and α=0.5 in all experiments.

5.5 Qualitative and Stage-wise Analysis

Figure 4 compares CISTA-Net with baseline on scenes containing five closely-spaced dim targets under different degrees of overlap. While both methods recover the target distribution under moderate mixing, CISTA-Net produces more concentrated responses and fewer residual activations under severe overlap, which facilitates subsequent target extraction. To further examine the unfolding process, Figure 5 visualizes intermediate reconstructions, showing that target responses become progressively sharper while background perturbations are suppressed across stages. These observations suggest

that count-guided shrinkage helps preserve weak target responses during iterative reconstruction, while MSIF complements this process by retaining useful intermediate information.

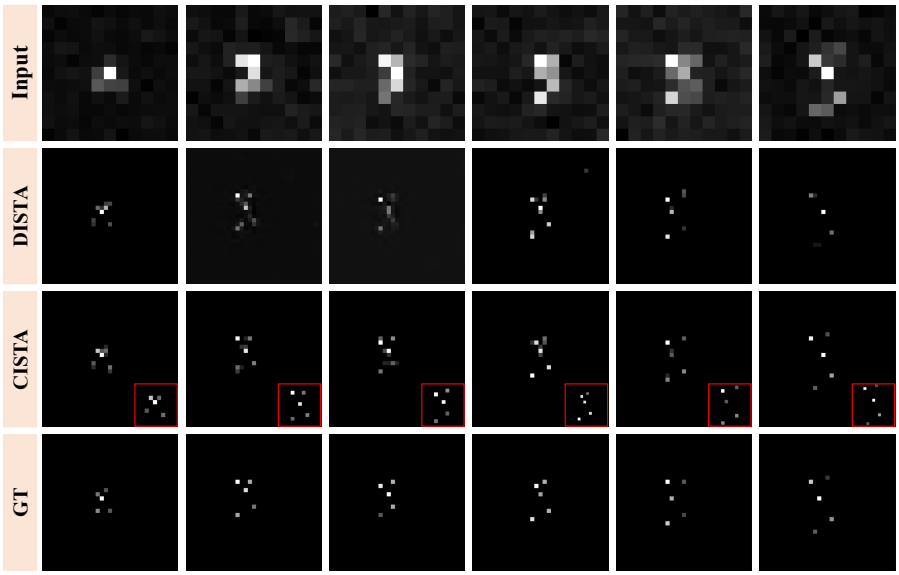


Figure 4: Qualitative comparison on real-world infrared observations with five closely-spaced small targets. The red-boxed regions show the local post-processed results, highlighting the separation performance for CSIST.

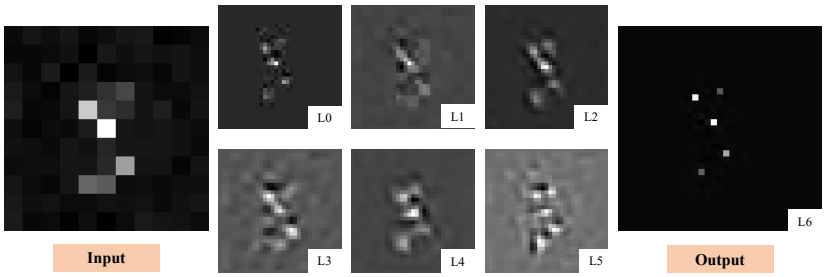


Figure 5: Stage-wise reconstruction of the supplied example. Successive estimates illustrate changes in the spatial concentration of source responses; they do not establish monotonic improvement or convergence for arbitrary inputs.

6 Conclusion

We present CISTA-Net, a count-guided deep unfolding framework for closely-spaced infrared small target unmixing. CISTA-Net introduces CP-DTG to couple the predicted count distribution with local reconstruction evidence, producing spatially adaptive shrinkage thresholds at each unfolding stage. MSIF further aggregates complementary intermediate reconstructions to recover weak responses that may be attenuated by repeated shrinkage, while count-guided candidate selection promotes consistency between the reconstructed responses and the predicted target count. Experiments on CSIST-100K show that CISTA-Net consistently improves overall CSO-mAP over DISTA-Net across different sampling-grid resolutions, with particularly clear gains at moderate and relaxed localization tolerances. The results suggest that global count information can provide useful guidance beyond locally adaptive reconstruction features, and that the manner in which count information is coupled with the shrinkage operator is an important design choice for deep unfolding. Future work will investigate more robust count estimation, uncertainty-aware count guidance, and generalization under PSF mismatch, noise variation, and real infrared observations.

AI USE STATEMENT

Generative AI tools were employed in this work solely to support language refinement and the identification of potentially relevant literature. All AI-assisted edits were carefully reviewed by the authors to ensure consistency with the intended meaning, and all suggested references were independently checked against the original sources. The authors assume full responsibility for the final manuscript, including its text, scientific claims, citations, and all associated research content.

ETHICS STATEMENT

This work uses publicly available datasets and benchmarks and involves no human-subject experiments or collection of personal data. Our study focuses on infrared small target unmixing and benchmark evaluation in a controlled research setting. While improved target reconstruction may benefit downstream sensing applications, potential misuse in surveillance or other sensitive scenarios should be considered when deploying such techniques in practice.

REPRODUCIBILITY STATEMENT

Sections 3–4.4 define the imaging model and reconstruction components. Algorithm 1 gives inference pseudocode, and Section 4.4 defines the losses; Appendix C retains the reported numerical results. Appendix D distinguishes available settings from details that the current draft does not specify. To support reproducibility, we provide our code and training datasets as supplementary materials on OpenReview.

REFERENCES

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)*, pp. 11065–11074, 2019.

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.

Yimian Dai and Yiquan Wu. Reweighted infrared patch-tensor model with both nonlocal and local priors for single-frame small target detection. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 10(8):3752–3767, 2017. doi: 10.1109/JSTARS.2017.2700023.

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. doi: 10.1109/TGRS.2020.3044958.

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.

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.

Chenqiang Gao, Deyu Meng, Yi Yang, Yongtao Wang, Xiaofang Zhou, and Alexander G. Hauptmann. Infrared patch-image model for small target detection in a single image. *IEEE Transactions on Image Processing*, 22(12):4996–5009, 2013. doi: 10.1109/TIP.2013.2281420.

Karol Gregor and Yann LeCun. Learning fast approximations of sparse coding. In Johannes Fürnkranz and Thorsten Joachims (eds.), *International Conference on Machine Learning (ICML)*, pp. 399–406, 2010. URL https://icml.cc/Conferences/2010/papers/449.pdf.

Shengdong Han, Shangdong Yang, Yuxuan Li, Xin Zhang, Xiang Li, Jian Yang, Ming-Ming Cheng, and Yimian Dai. Dista-net: Dynamic closely-spaced infrared small target unmixing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14655–14664, 2025.

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. ISSN 0031-3203. doi: https://doi.org/10.1016/j.patcog.2023.109788. URL https://www.sciencedirect.com/science/article/pii/S0031320323004867.

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)*, pp. 3867–3876, 2019.

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*, pp. 136–144, 2017.

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)*, pp. 191–207, 2020.

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.

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.

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.

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.

Mengze Xu, Zhu Liu, Weidong Sheng, Boyang Li, Yimian Dai, Ming-Ming Cheng, and Jian Yang. Dista-net++: Rethinking infrared small target unmixing beyond sub-pixel separation. *arXiv preprint arXiv:2609.18773*, 2026.

Yan Yang, Jian Sun, Huibin Li, and Zongben Xu. Deep admm-net for compressive sensing mri. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (eds.), *Neural Information Processing Systems (NeurIPS)*, volume 29, 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/file/1679091c5a880faf6fb5e6087eb1b2dc-Paper.pdf.

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. doi: 10.1109/TIM.2023.3304672.

Ximeng Zhai, Bohan Xu, Yaohong Chen, Hao Wang, Kehua Guo, and Yimian Dai. Seqcsist: Sequential closely-spaced infrared small target unmixing. *IEEE Transactions on Geoscience and Remote Sensing*, 2025.

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)*, pp. 1828–1837, 2018. doi: 10.1109/CVPR.2018.00196. URL http://openaccess.thecvf.com/content_cvpr_2018/html/Zhang_ISTA-Net_Interpretable_Optimization-Inspired_CVPR_2018_paper.html.

M. Zhang, R. Zhang, Y. Yang, H. Bai, J. Zhang, and J. Guo. ISNet: Shape matters for infrared small target detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 877–886, 2022.

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)*, pp. 286–301, 2018a.

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)*, pp. 2472–2481, 2018b.

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.

A Imaging Model and Coordinate Conventions

For a source at (x_t, y_t), the normalized Gaussian PSF is

$$p(x, y \mid x_t, y_t) = \frac{1}{2\pi\sigma_{\mathrm{PSF}}^2} \exp\left(-\frac{(x - x_t)^2 + (y - y_t)^2}{2\sigma_{\mathrm{PSF}}^2}\right). \tag{21}$$

Here σ_{PSF} is the standard deviation, rather than the variance. For a square pixel of width D centered at (x_{ij}, y_{ij}), the integrated response is

$$g_{ij}(x_t, y_t) = \int_{x_{ij}-D/2}^{x_{ij}+D/2} \int_{y_{ij}-D/2}^{y_{ij}+D/2} p(x, y \mid x_t, y_t)\, \mathrm{d}y\, \mathrm{d}x. \tag{22}$$

The integration variables match their bounds. Stacking pixel responses gives $\mathbf{g}(x_t, y_t) \in \mathbb{R}^{UV}$. For K sources,

$$\mathbf{z} = \sum_{k=1}^{K} s_k \mathbf{g}(x_k, y_k) + \mathbf{n}, \qquad \mathbf{n} \sim \mathcal{N}(\mathbf{0}, \sigma_n^2 \mathbf{I}). \tag{23}$$

Because the PSF is normalized, s_k denotes its amplitude coefficient (integrated source intensity before finite-window truncation). It is not the Gaussian peak height, which would be $s_k/(2\pi\sigma_{\mathrm{PSF}}^2)$. The radiometric convention of the dataset must be matched when constructing training targets.

Discretization. The grid Ω_c contains $L = UVc^2$ candidate sub-pixel centers. Its steering matrix has columns $\mathbf{g}(x_l, y_l)$, leading to Eq. (1). The maximum distance from a point inside a sub-pixel to its center is half the cell diagonal, $\sqrt{2}D/(2c)$. The original draft omitted the factor $1/2$. This bound assumes nearest-center assignment and that the source lies within the covered grid.

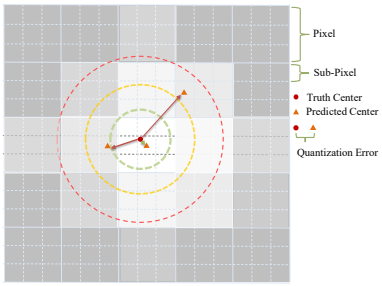


Figure 6: Sub-pixel discretization. The figure's $n \times n$ subdivision corresponds to $c \times c$ in the text. Refinement decreases representation error while increasing the number of candidate source locations.

Coordinates. For the odd sampling ratios used here, and detector-pixel centers indexed from zero, a source coordinate (x, y) corresponds to the continuous grid coordinate

$$(u, v) = \left(cx + \frac{c-1}{2},\ cy + \frac{c-1}{2}\right). \tag{24}$$

Conversely, a grid index (u, v) maps to detector coordinates

$$(x, y) = \left(\frac{u - (c-1)/2}{c},\ \frac{v - (c-1)/2}{c}\right). \tag{25}$$

Noninteger (u, v) requires an explicit rasterization rule. The draft does not specify rounding, interpolation, boundary handling, or how amplitudes are combined if multiple targets occupy one grid cell. Those choices affect supervision and evaluation and must be recovered from the implementation rather than inferred from Eq. (24).

ISTA and learned transforms. For the convex objective in Eq. (2), the gradient of the data term is $\mathbf{G}_c^\top(\mathbf{G}_c\mathbf{s} - \mathbf{z})$ and the proximal threshold is $\rho\lambda$. If an orthonormal transform $\mathbf{\Psi}$ replaces the identity in the regularizer, its proximal map is $\mathbf{\Psi}^\top \mathrm{Soft}(\mathbf{\Psi}\mathbf{r}, \rho\lambda)$. This expression does not generally hold for arbitrary nonlinear transforms. CISTA-Net adopts learned transforms and thresholds as a parameterized reconstruction architecture, without an accompanying proof that it minimizes a fixed convex objective.

B ALGORITHM, FUSION, AND TRAINING DETAILS

Algorithm 1 separates the count distribution from the count-conditioned feature tensor. The discrete class index is interpreted as a count in $\{1, \dots, 5\}$; zero-based implementations require the corresponding label conversion. CGPP is a deterministic post-processing rule, not an additional learned module in this description.

Algorithm 1 CISTA-Net inference with a predicted count

Input: Observation $\mathbf{z}$; grid steering matrix $\mathbf{G}_c$; initializer $\mathbf{Q}_{\text{init}}$; stages N; fusion index set $\mathcal{I}$; radius $d_{\min}$
Output: Reconstruction $\widehat{\mathbf{s}}$ and selected target set $\mathcal{S}$

$\mathbf{s}^{(0)} \leftarrow \mathbf{Q}_{\text{init}}\mathbf{z}$
$\mathbf{q} \leftarrow \mathrm{NumCNN}(\mathrm{reshape}(\mathbf{z}))$
Form count-aware features $\mathbf{A}$ using Eq. (10)
for $k = 1, \dots, N$ **do**
  $\mathbf{r}^{(k)} \leftarrow \mathbf{s}^{(k-1)} - \rho^{(k)}\mathbf{G}_c^{\top}(\mathbf{G}_c\mathbf{s}^{(k-1)} - \mathbf{z})$
  $\mathbf{u}^{(k)} \leftarrow \mathcal{F}_d^{(k)}(\mathbf{s}^{(k-1)}, \mathbf{r}^{(k)})$
  $\boldsymbol{\theta}^{(k)} \leftarrow \text{CP-DTG}^{(k)}(\mathbf{u}^{(k)}, \mathbf{A})$
  $\mathbf{s}^{(k)} \leftarrow \widetilde{\mathcal{F}}^{(k)}(\mathrm{Soft}(\mathbf{u}^{(k)}, \boldsymbol{\theta}^{(k)}))$
end for
Compute $\mathbf{s}_{\text{MSIF}}$ from stages $k \in \mathcal{I}$ using Eqs. (13)–(14)
$\widehat{\mathbf{s}} \leftarrow \mathbf{s}^{(N)} + \beta\mathbf{s}_{\text{MSIF}}$
$\widehat{K} \leftarrow \arg\max_{j \in \{1,\dots,5\}} q_j$
Extract candidates from $\widehat{\mathbf{s}}$; sort by decreasing response
$\mathcal{S} \leftarrow \varnothing$
for each candidate p in the sorted list **do**
  if $|\mathcal{S}| = \widehat{K}$ **then**
    Stop scanning candidates
  end if
  if $\|p - p'\|_2 \geq d_{\min}$ for every $p' \in \mathcal{S}$ **then**
    $\mathcal{S} \leftarrow \mathcal{S} \cup \{p\}$
  end if
end for
return $\widehat{\mathbf{s}}, \mathcal{S}$

Fusion implementation. Stage biases $\mathbf{p}^{(k)}$ are initialized from $\mathcal{N}(0, 0.01^2)$. Each stage uses an independent two-block Conv–BN–ReLU extractor with 32 output channels, followed by a projection to 16 channels. Channel-average and channel-maximum pooling are concatenated, transformed by a 3×3 convolution and sigmoid, and multiplied into the 16-channel features. Softmax-normalized scalar weights combine the attended stages. The reported recovery channel sequence is $16 \to 32 \to 32 \to 1$, followed by the scaled $\tanh$ in Eq. (14).

The draft is inconsistent about the fusion stage set: its algorithm uses $\{1, \dots, N\}$, one paragraph uses $\{0, \dots, N\}$, and its weighted sum uses $\{0, \dots, N-1\}$. We use $\mathcal{I}$ until the implementation resolves this discrepancy. Whether the final recovery layer includes ReLU also matters: applying ReLU before $\tanh$ would permit only nonnegative corrections. No change to that implementation is presumed here.

Least-squares initialization. Using only training pairs, define $\mathbf{Z}_{\text{tr}} = [\mathbf{z}_1, \dots, \mathbf{z}_T]$ and $\mathbf{S}_{\text{tr}} = [\mathbf{s}_1^{\star}, \dots, \mathbf{s}_T^{\star}]$. A least-squares solution is

$$\mathbf{Q}_{\text{init}} = \mathbf{S}_{\text{tr}}\mathbf{Z}_{\text{tr}}^{\dagger}. \tag{26}$$

If $\mathbf{Z}_{\text{tr}}\mathbf{Z}_{\text{tr}}^{\top}$ is invertible, this equals the expression in the draft, $\mathbf{S}_{\text{tr}}\mathbf{Z}_{\text{tr}}^{\top}(\mathbf{Z}_{\text{tr}}\mathbf{Z}_{\text{tr}}^{\top})^{-1}$. This construction minimizes training reconstruction error over linear maps; it does not establish faster convergence of the learned network.

The complete loss definitions are given in Section 4.4. The dynamic-transform conditioning and gradient paths discussed there must be matched when implementing the training objective.

C Complete Reported Numerical Results

Table 4: Comparison with state-of-the-art methods on the CSIST-100K dataset.

Method	#P ↓	FLOPs ↓	CSO-mAP						PSNR ↑	SSIM ↑
			mAP	AP-05	AP-10	AP-15	AP-20	AP-25		
Traditional Optimization										
ISTA Daubechies et al. (2004)	-	-	7.46	0.01	0.31	2.39	9.46	25.14	-	-
Image Super-Resolution										
ACTNet Zhang et al. (2023)	46.212M	62.798G	45.61	0.38	7.46	41.13	83.12	95.95	35.4526	98.00
CTNet Wang et al. (2021)	0.4M	2.756G	45.11	0.38	7.53	40.39	82.11	95.14	35.1499	98.19
DCTLSA Zeng et al. (2023)	0.865M	13.694G	44.51	0.39	7.35	39.35	81.15	94.34	34.6314	98.02
EDSR Lim et al. (2017)	1.552M	12.035G	45.32	0.33	7.07	40.58	83.24	95.41	35.3724	96.98
EGASR Qiu et al. (2023)	2.897M	17.732G	45.51	0.42	8.03	41.32	85.71	95.08	34.5681	96.56
FeNet Wang et al. (2022)	0.348M	2.578G	45.68	0.40	8.05	41.76	83.15	95.02	34.1543	97.70
RCAN Zhang et al. (2018a)	1.079M	8.243G	45.87	0.42	7.96	41.81	83.61	95.57	35.2119	98.47
RDN Zhang et al. (2018b)	22.306M	0.173T	45.81	0.35	7.11	41.07	84.07	96.43	36.4686	98.34
SAN Dai et al. (2019)	4.442M	34.05G	45.95	0.36	7.35	41.17	84.32	96.57	36.5037	98.12
SRCNN Dong et al. (2015)	19.393K	1.345G	29.06	0.23	4.10	21.65	49.95	69.39	28.7608	95.00
SRFBN Li et al. (2019)	0.373M	3.217G	46.05	0.43	8.31	42.83	83.72	94.95	34.0174	96.84
HAN Niu et al. (2020)	64.342M	0.495T	45.70	0.39	7.46	40.90	83.61	96.17	35.2703	99.71
CFGN Dai et al. (2023)	0.531M	4.068G	46.67	**0.49**	**8.45**	**43.23**	85.03	96.13	35.4731	99.73
Deep Unfolding										
ISTA-Net Zhang & Ghanem (2018)	0.171M	12.766G	45.16	0.41	7.71	40.57	82.58	94.53	33.9215	99.68
ISTA-Net+ Zhang & Ghanem (2018)	0.337M	24.328G	46.06	0.42	7.66	41.58	84.46	96.17	36.0892	99.72
LAMP Daubechies et al. (2004)	2.126M	0.278G	14.22	0.05	1.11	7.31	21.56	41.06	27.8299	96.89
LIHT Daubechies et al. (2004)	21.1M	1.358G	10.35	0.06	0.92	4.99	14.74	30.5	27.5107	96.42
LISTA Daubechies et al. (2004)	21.1M	1.358G	30.13	0.25	4.13	22.29	51.18	72.82	29.8936	99.12
FISTA-Net Xiang et al. (2021)	74.599K	18.958G	44.66	0.45	7.68	39.74	81.24	94.19	35.7519	99.67
TiLISTA Daubechies et al. (2004)	2.126M	0.278G	14.95	0.06	1.23	7.72	22.50	46.23	27.7038	90.12
DISTA-Net Han et al. (2025)	2.179M	35.103G	46.74	0.38	7.54	42.44	86.18	97.14	37.8747	**98.80**
⋆ **CISTA-Net (Ours)**	2.201M	35.247G	**47.28**	0.36	7.54	43.22	**87.47**	**97.79**	**37.8875**	**98.83**

Table 4 compares CISTA-Net with traditional optimization methods (ISTA), image super-resolution methods, and deep unfolding methods on the CSIST-100K dataset, evaluated across three dimensions: computational efficiency (#P / FLOPs), localization accuracy (CSO-mAP), and image reconstruction quality (PSNR / SSIM).

For localization accuracy, we adopt CSO-mAP with distance thresholds ranging from AP-05 to AP-25, where a smaller threshold indicates a stricter localization precision requirement. For instance, AP-05 evaluates detection accuracy within a 0.05-pixel distance, representing an extremely demanding precision criterion. Under practical scenarios where 0.25-pixel precision (AP-25) is sufficient, CISTA-Net achieves **97.79%** accuracy, outperforming all existing methods including DISTA-Net Han et al. (2025) (97.14%). Even under more stringent conditions (AP-20 and AP-15), our method maintains superior performance with **87.47%** and **43.22%** accuracy, respectively, demonstrating the effectiveness of count-guided supervision in reducing missed detections and mislocalization in dense multi-target scenarios. It is worth noting that at the most stringent thresholds (AP-05 and AP-10), CISTA-Net (0.36 / 7.54) is marginally below the best-performing super-resolution method SRFBN Li et al. (2019) (0.43 / 8.31). This is attributable to the fact that the count-guided mechanism in CISTA-Net primarily enforces global target quantity constraints and improves mid-to-high precision localization, with limited effect on sub-pixel accuracy under extremely tight tolerances. Addressing this limitation remains a direction for future work.

In terms of model efficiency, CISTA-Net achieves these results with moderate computational cost (2.201M parameters and 35.247G FLOPs), remaining comparable to DISTA-Net (2.179M, 35.103G)

despite the additional count prediction branch and multi-stage fusion module. This demonstrates that the proposed counting-guided framework introduces negligible computational overhead while delivering consistent performance gains. Furthermore, CISTA-Net achieves the highest PSNR (**37.8875 dB**) and SSIM (**98.83**) among all compared methods, confirming that the multi-stage count-guided fusion module effectively compensates for reconstruction detail and preserves structural fidelity.

These comprehensive results validate that CISTA-Net achieves an effective balance between computational efficiency, localization accuracy, and reconstruction quality, while demonstrating that explicit count supervision constitutes a meaningful and complementary contribution to the deep unfolding framework for CSIST unmixing.

D Implementation Settings

Table 5 summarizes the implementation settings explicitly reported in the current draft. Unless otherwise specified, experiments use 11×11 grayscale patches, six unfolding stages, the Adam optimizer with a fixed learning rate of 10^{-4}, and a batch size of 64. The sampling ratio is selected from $c \in \{3, 5, 7\}$, with the primary comparison conducted at $c = 3$. The target-count prediction covers one to five targets. For the fusion module, we set $C = 32$, $\beta = 0.5$, and $M_0 = 50$, while the loss weights are $\gamma = \mu = 0.01$. The candidate-selection threshold is fixed at 50 response units. The transform weight is set to $\alpha = 0.7$ in the default configuration and varied around this value in the sensitivity analysis.

We note that a small number of implementation records are not fully consistent across the current manuscript. In particular, the main comparison reports 2.201M parameters and 35.247G FLOPs for CISTA-Net at $c = 3$, whereas the ablation description reports 2.7M parameters and 37.986G FLOPs. Rather than selecting one set of values without verification, we retain both records here for transparency. Accordingly, the complexity figures used in the main text should be regarded as provisional until the final implementation statistics are verified.

Table 5: Settings explicitly reported in this paper.

Item	Reported setting
Input	11×11 grayscale patch
Sampling ratios	$c \in \{3, 5, 7\}$; primary comparison at $c = 3$
Count classes	One to five targets
Unfolding depth	$N = 6$
Optimizer	Adam; fixed learning rate 10^{-4}
Batch size	64
Fusion	$C = 32$, $\beta = 0.5$, $M_0 = 50$
Loss weights	$\gamma = \mu = 0.01$
Candidate threshold	50 response units
Transform weight α	0.7 in setup; 0.5 in sensitivity analysis
CISTA-Net cost at $c = 3$	2.201M / 35.247G in main table; 2.7M / 37.986G in ablation prose