NEURAL INFORMATION PROCESSING SYSTEMS

Rethinking Evaluation of Infrared Small Target Detection

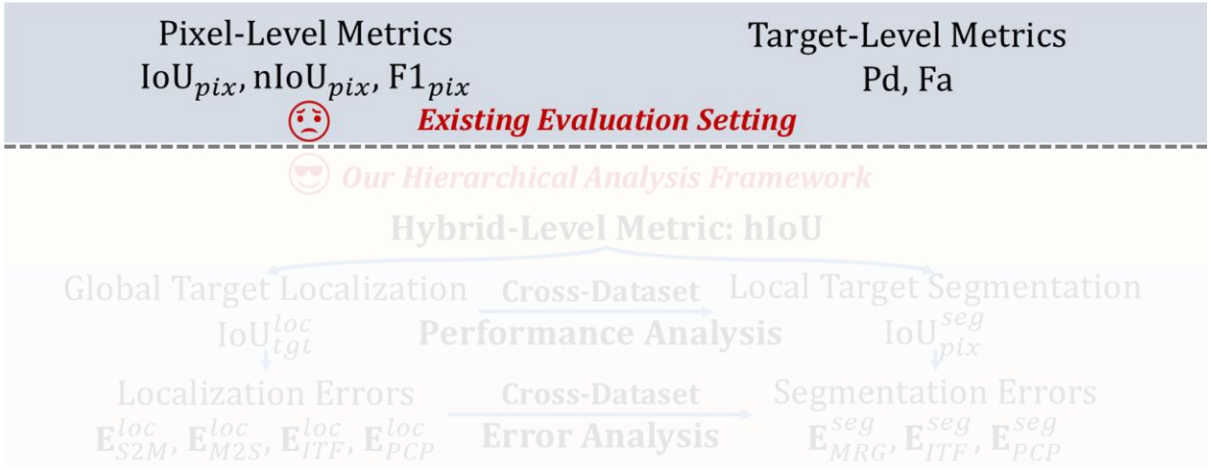
Youwei Pang[1,3], Xiaoqi Zhao[2], Lihe Zhang[1]✉, Huchuan Lu[1]
Georges El Fakhri[2], Xiaofeng Liu[2], Shijian Lu[3]

[1]*Dalian University of Technology*
[2]*Yale University*
[3]*Nanyang Technological University*

arXiv

GitHub

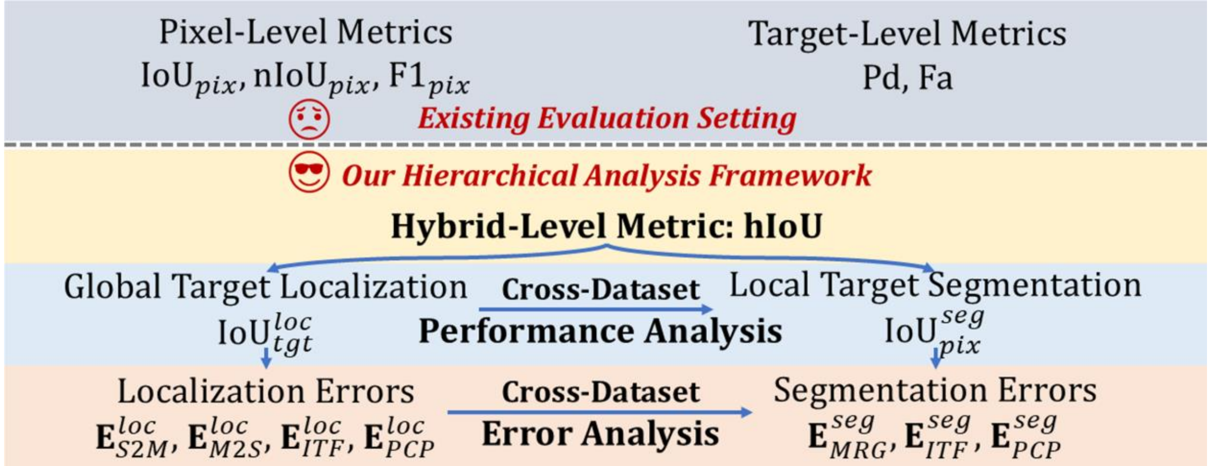
Limitations Of Current IRSTD Evaluation



- ***Metric Design***: Existing methods rely on fragmented pixel- and target-level specific metrics, which fails to provide a comprehensive view of model capabilities.
- ***Error Analysis***: An excessive emphasis on overall performance scores obscures crucial error analysis, which is vital for identifying failure modes and improving real-world system performance.
- ***Dataset-specific***: The field predominantly adopts dataset-specific training-testing paradigms, hindering the understanding of model robustness and generalization across diverse infrared scenarios.

Rethinking Evaluation of Infrared Small Target Detection, *Youwei Pang, Xiaoqi Zhao, Lihe Zhang, Huchuan Lu, Georges El Fakhri, Xiaofeng Liu, Shijian Lu, 2025*

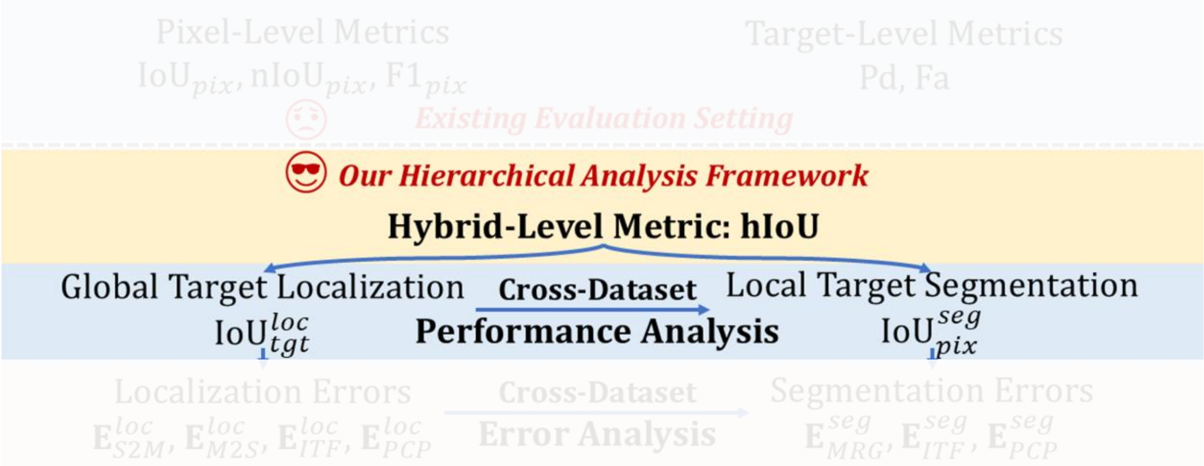
Our Contributions



- ☐ Expose limitations in current IRSTD evaluation protocols and *propose a hierarchical analysis framework*.
- ☐ Introduce a ***hybrid-level metric*** capturing IRSTD performance across target and pixel levels.
- ☐ Reveal limited ***cross-dataset generalization*** of IRSTD algorithms through detailed evaluation.
- ☐ First to ***systematically analyze errors*** in IRSTD by quantifying model limitations under our metric.
- ☐ Develop ***a universal and comprehensive evaluation toolkit*** to advance IRSTD research.

NEURAL INFORMATION PROCESSING SYSTEMS

Hierarchical Intersection over Union (hIoU)



- Evaluation practices in current IRSTD studies typically focus on either isolated pixel-level or target-level similarity measurements between predictions and GTs.
- We propose a new hybrid-level metric, *i.e.*, **hierarchical Intersection over Union (hIoU)**, which hierarchically combines both global target-level localization and local pixel-level segmentation performance:

$$\mathrm{hIoU} = \mathrm{IoU}_{tgt}^{loc} \times \mathrm{IoU}_{pix}^{seg} = \frac{\sum_{i=1}^{K} |\mathrm{TP}_{tgt}^{[i]}|}{\sum_{i=1}^{K} \left|\mathrm{TP}_{tgt}^{[i]}\right| + \left|\mathrm{FP}_{tgt}^{[i]}\right| + \left|\mathrm{FN}_{tgt}^{[i]}\right|} \times \frac{\sum_{(T_G^m, T_P^n \in \mathrm{TP}_{tgt})} (T_G^m \cap T_P^n) / (T_G^m \cup T_P^n)}{\sum_{i=1}^{K} \left|\mathrm{TP}_{tgt}^{[i]}\right|}$$

Overlap Priority with Distance Compensation

- Conventional IRSTD evaluation methods depend on strict distance-only matching, which often misjudges offset or fragmented predictions.
- The proposed OPDC (Overlap Priority with Distance Compensation) strategy first prioritizes overlap-based matching to ensure shape coherence, then applies distance compensation to handle small or low-overlap cases.
- This hierarchical design achieves more intuitive and robust target-level matching by combining morphological alignment with spatial proximity.

Algorithm 1: OPDC Matching Strategy

Input: $\{T_G^m\}_m^M$: the mask set of M targets extracted from GT map G; $\{T_P^n\}_n^N$: the mask set of N targets extracted from prediction map P; MAX: a extremely large value used to avoid the algorithm choosing irrational matches;

Output: S_{TP}: the matched index pair set; S_{FN}: the unmatched GT target index set; S_{FP}: the unmatched predicted target index set;

// 1. Overlap Priority Constraint

Valid indicator $\mathbb{I} \in \mathbb{R}^{M \times N}, \mathbb{I}_{m,n} \in \{0,1\}, \mathbb{I}_{m,n} = 0, \forall m, n;$

Distance matrix $D \in \mathbb{R}^{M \times N}, D_{m,n} = \text{MAX}, \forall m, n;$

for $m = 1$ **to** M **do**

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

    $D_{m,n} = \texttt{EuclideanDistance}(T_G^m, T_P^n);$

    if $|T_G^m \cap T_P^n| / |T_G^m \cup T_P^n| \geq 0.5$ **then** $\mathbb{I}_{m,n} = 1;$

Initial match $A = \texttt{Assignment}(D);$ *// scipy.optimize.linear_sum_assignment [5]*

$S_{TP}, S_{FN}, S_{FP} \leftarrow A \cap \mathbb{I};$ *// Consider only pairwise relations that satisfy constraints.*

// 2. Distance-based Compensation

Valid indicator $\hat{\mathbb{I}} \in \mathbb{R}^{|S_{FN}| \times |S_{FP}|}, \hat{\mathbb{I}}_{m,n} \in \{0,1\}, \hat{\mathbb{I}}_{m,n} = 0, \forall m, n;$

Distance matrix $\hat{D} \in \mathbb{R}^{|S_{FN}| \times |S_{FP}|}, \hat{D}_{m,n} = \text{MAX}, \forall m, n;$

for $m = 1$ **to** $|S_{FN}|$ **do**

  for $n = 1$ **to** $|S_{FP}|$ **do**

    if $D_{S_{FN}^m, S_{FN}^n} < 3$ **then** *// Following the setting in [17].*

      $\hat{D}_{m,n} = D_{S_{FN}^m, S_{FN}^n};$

      $\hat{\mathbb{I}}_{m,n} = 1;$

Compensation match $\hat{A} = \texttt{Assignment}(\hat{D});$ *// scipy.optimize.linear_sum_assignment [5]*

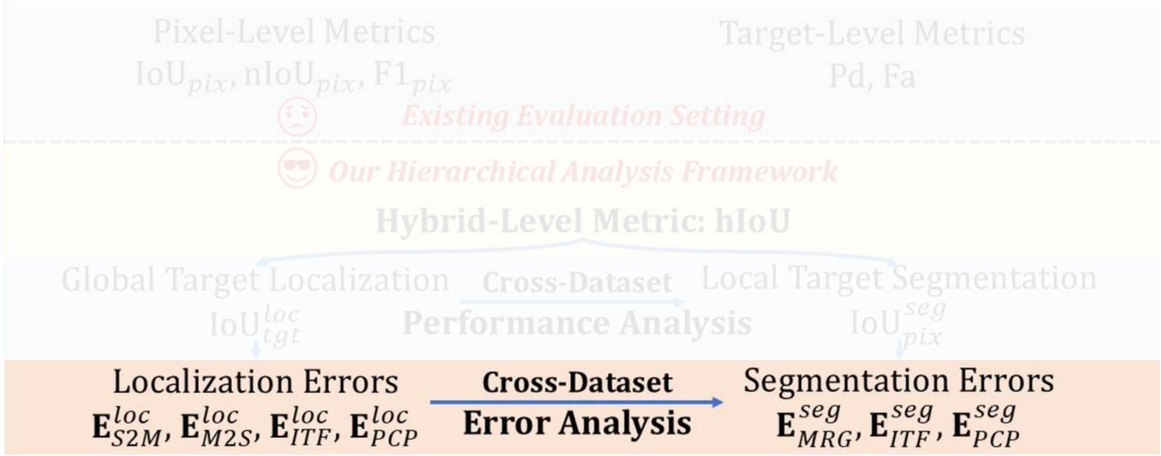
$\hat{S}_{TP}, \hat{S}_{FN}, \hat{S}_{FP} \leftarrow \hat{A} \cap \hat{\mathbb{I}};$

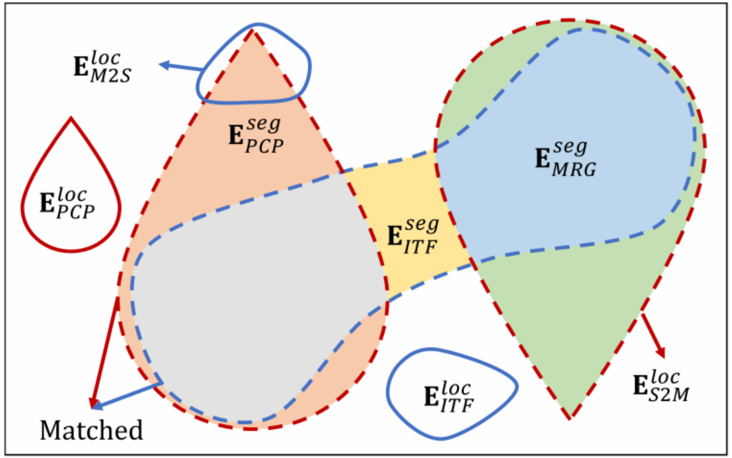
$S_{TP} = S_{TP} \cup \hat{S}_{TP};$ *// Construct final matched index pair set.*

$S_{FN} = S_{FN} \setminus \hat{S}_{TP};$ *// Construct final unmatched GT target index set.*

$S_{FP} = S_{FP} \setminus \hat{S}_{TP};$ *// Construct final unmatched predicted target index set.*

Systematic Error Analysis





Error types for three predicted and three GT targets. Blue contours denote predictions and red contours denote GT. Under our ***OPDC strategy***, only the middle prediction is matched to a GT and all others remain unmatched. We categorize prediction errors into two levels:

1. target-level localization error

$$\mathbf{E}^{loc} = 1 - \mathrm{IoU}^{loc}_{tgt} = \mathbf{E}^{loc}_{S2M} + \mathbf{E}^{loc}_{M2S} + \mathbf{E}^{loc}_{ITF} + \mathbf{E}^{loc}_{PCP}$$

2. pixel-level segmentation error

$$\mathbf{E}^{seg} = 1 - \mathrm{IoU}^{seg}_{pix} = \mathbf{E}^{seg}_{MRG} + \mathbf{E}^{seg}_{ITF} + \mathbf{E}^{seg}_{PCP}$$

NEURAL INFORMATION PROCESSING SYSTEMS

Cross-dataset Performance Comparison

Dataset	Metric	ACM$_{21}$ [9]	FC3Net$_{22}$ [37]	DNANet$_{22}$ [17]	ISNet$_{22}$ [38]	AGPCNet$_{23}$ [39]	UIUNet$_{23}$ [31]	RDIAN$_{23}$ [26]	MTU-Net$_{23}$ [30]	ABC$_{23}$ [22]	SeRankDet$_{24}$ [6]	MSHNet$_{24}$ [20]	MRF3Net$_{24}$ [41]	SCTransNet$_{24}$ [34]	RPCANet$_{24}$ [29]
	Params.	0.5M	6.9M	4.7M	1.1M	12.4M	50.5M	0.1M	4.1M	73.5M	108.9M	4.1M	0.5M	11.2M	0.7M
	FLOPs	2.0G	10.6G	56.1G	121.9G	327.5G	217.9G	14.8G	24.4G	332.6G	568.7G	24.4G	33.2G	67.4G	179.7G
	Trained on IRSTD1k$_{TR}$ [38].														
IRSTD1k$_{TE}$ [38]	IoU$_{pix}$ ↑	0.439	0.358	0.637	0.578	0.605	0.570	0.603	0.610	0.624	**0.642**	**0.650**	0.636	**0.644**	0.608
	nIoU$_{pix}$ ↑	0.476	0.531	**0.625**	0.518	0.580	0.600	0.605	0.607	0.595	0.621	0.620	**0.630**	**0.622**	0.579
	F1$_{pix}$ ↑	0.610	0.527	0.778	0.733	0.754	0.726	0.753	0.757	0.768	**0.782**	**0.788**	0.777	**0.783**	0.756
	Pd↑	0.798	0.865	0.912	0.919	0.916	0.906	0.902	**0.929**	0.916	**0.926**	**0.933**	0.899	0.912	0.886
	+OPDC	0.835	0.875	0.916	0.926	0.919	0.912	0.912	**0.939**	0.919	**0.933**	**0.936**	0.909	0.923	0.896
	Fa×10^6 ↓	95.178	237.365	**13.854**	**13.266**	15.354	51.147	21.503	28.012	16.815	44.638	**11.539**	17.441	16.834	28.145
	+OPDC	61.187	233.266	**10.476**	**11.444**	11.862	44.277	17.062	23.647	13.475	41.108	**7.686**	12.146	**10.476**	21.844
	hIoU↑	0.356	0.383	**0.557**	0.443	0.496	0.530	0.511	0.493	0.508	0.520	**0.549**	**0.553**	0.537	0.470
SIRST$_{TE}$ [9]	IoU$_{pix}$ ↑	0.472	0.234	0.676	0.712	**0.763**	0.696	0.658	0.701	0.708	**0.734**	0.649	**0.752**	0.629	0.543
	nIoU$_{pix}$ ↑	0.567	0.595	0.733	0.719	0.735	0.688	0.735	**0.749**	0.737	**0.742**	0.699	**0.763**	0.686	0.676
	F1$_{pix}$ ↑	0.642	0.380	0.807	0.832	**0.866**	0.821	0.794	0.824	0.829	**0.847**	0.787	**0.858**	0.772	0.704
	Pd↑	0.908	0.872	0.963	**0.982**	**0.991**	0.963	0.963	**0.982**	**0.972**	**0.982**	0.954	**0.982**	0.963	0.927
	+OPDC	0.927	0.881	0.963	**0.982**	**0.991**	**0.982**	0.963	**1.000**	**0.982**	**0.982**	0.954	**0.982**	0.963	0.927
	Fa×10^6 ↓	127.650	932.797	**3.754**	5.632	**2.560**	86.351	17.407	42.152	22.526	17.236	11.946	**4.608**	20.820	124.066
	+OPDC	121.165	932.456	**3.754**	5.632	**2.560**	30.376	17.407	38.397	21.844	17.236	11.946	**4.608**	20.820	124.066
	hIoU↑	0.418	0.390	**0.687**	0.674	0.682	0.644	0.645	**0.693**	0.672	0.684	0.623	**0.694**	0.596	0.598
NUDT$_{TE}$ [17]	IoU$_{pix}$ ↑	0.331	0.288	**0.504**	0.443	0.468	0.450	0.441	0.416	0.469	**0.494**	0.463	**0.512**	0.377	0.291
	nIoU$_{pix}$ ↑	0.452	0.443	**0.627**	0.538	0.556	0.541	0.554	0.528	**0.582**	0.581	0.561	**0.609**	0.502	0.430
	F1$_{pix}$ ↑	0.498	0.448	**0.670**	0.614	0.638	0.620	0.612	0.588	0.638	**0.661**	0.633	**0.678**	0.548	0.451
	Pd↑	0.757	0.752	**0.848**	0.759	0.785	**0.836**	0.804	0.769	0.808	**0.820**	0.771	0.815	0.748	0.701
	+OPDC	0.792	0.787	**0.886**	0.806	0.850	**0.867**	**0.871**	0.857	0.841	0.832	0.843	0.862	0.818	0.722
	Fa×10^6 ↓	253.092	273.488	121.206	71.920	**40.690**	114.695	71.869	96.690	133.464	**63.883**	65.562	**43.844**	110.728	190.989
	+OPDC	236.308	266.469	114.899	57.068	**31.789**	106.049	62.002	79.753	124.563	59.916	**47.913**	**35.502**	92.723	187.581
	hIoU↑	0.274	0.333	**0.526**	0.434	0.456	0.457	0.408	0.366	**0.484**	**0.466**	0.445	**0.484**	0.393	0.344

We benchmark 14 recent IRSTD methods using both conventional metrics and our hierarchical framework. Results reveal inconsistencies: models like MSHNet achieve top scores in standard metrics but underperform in holistic performance (hIoU), while DNANet ranks higher overall due to more balanced segmentation and localization. Incorporating the OPDC strategy further improves recall by addressing overly strict distance-only matching, demonstrating the value of shape- and distance-aware evaluation.

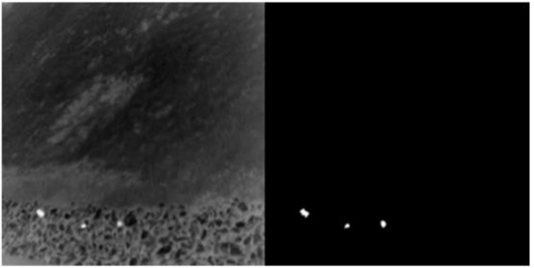
Cross-dataset Error Ratio Analysis

(a) Localization errors.

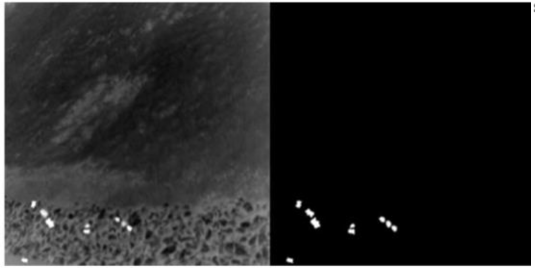
(b) Segmentation errors.

Our fine-grained error decomposition exposes specific weaknesses that conventional metrics obscure. The analysis highlights specific localization and segmentation errors (ITF & PCP) as major contributors, emphasizing the importance of detailed error breakdowns for understanding model limitations and guiding robust algorithm development.

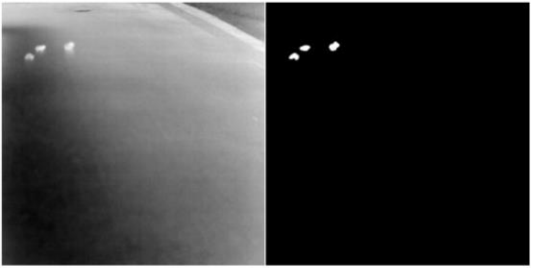
Synthetic Data Experiment



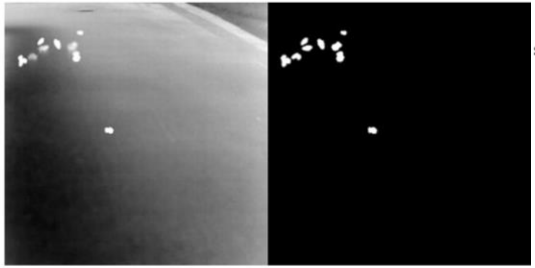
(a) Sample 1 (Original).



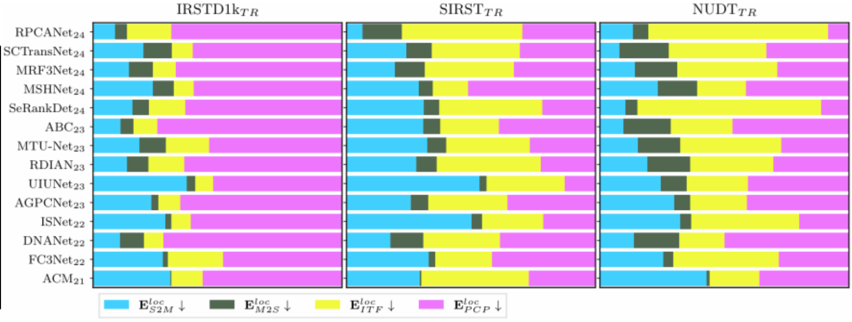
(b) Sample 1 (Augmented).



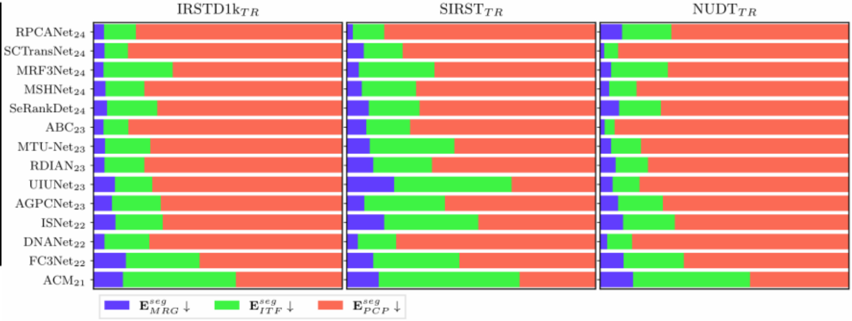
(c) Sample 2 (Original).



(d) Sample 2 (Augmented).



(a) Localization errors.



(b) Segmentation errors.

We introduce IRSTD1k$^{AUG}_{TE}$, a synthetic dataset built from IRSTD1k using copy-paste augmentation to ***increase target density and diversity while preserving scene coherence***. It provides challenging, high-density test cases. Results show that existing models degrade significantly on this dataset, exposing their weaknesses in complex scenarios.

NEURAL INFORMATION
PROCESSING SYSTEMS

Thanks!

arXiv

GitHub