SAIST: Segment Any Infrared Small Target Model Guided by Contrastive Language-Image Pretraining

Supplementary Material

1. Dataset Creation and Composition

In the field of infrared small target detection (IRSTD), although several large-scale datasets have been developed to advance the research, these datasets primarily consist of image data with target masks, which limits the ability to detect small targets in complex backgrounds. Recently, multimodal learning methods, particularly text-guided detection approaches, have shown significant performance improvements. Inspired by these developments, we propose and create the MIRSTD dataset, which combines infrared images with textual descriptions to enhance the performance of small target detection in complex scenarios.

To ensure broad accessibility, the dataset will be hosted on both GitHub and a cloud storage platform. It is made available under the Creative Commons Attribution-ShareAlike 4.0 International (CC BY-SA 4.0) license, which permits use, modification, and distribution for both academic research and industrial applications, provided that appropriate credit is given and any derived works are distributed under the same license. The dataset, along with the associated code, will be publicly accessible and can be retrieved from the dataset and code link. As the authors, we assume full responsibility for the dataset's release, including addressing any potential intellectual property concerns or legal issues that may arise from its distribution. All necessary precautions have been taken to ensure compliance with relevant legal and ethical standards.

1.1. Dataset Creation Motivation and Purpose

The MIRSTD dataset was created with the goal of enabling research in multimodal learning for small target detection. To address the challenge of weak target signals in the presence of strong background noise, we augment infrared image data with textual descriptions that provide semantic context. This multimodal approach is expected to improve detection accuracy. Specifically, MIRSTD builds upon three existing datasets—IRSTD-1k, NUAA-SIRST, and NUDT-SIRST—by adding concise textual descriptions for each infrared image, summarizing the scene and aiding in training vision-language models.

1.2. Dataset Composition

Each record in the dataset consists of three components:

- **Infrared Image:** The image data, after the target mask is removed, is used as the input for the model.
- **Text Description:** A natural language description of the scene in the image. The descriptions are concise, typically no more than 25 words, and follow a format such as "a photo of a [scene] with [object], [object], and [object]".
- **Target Mask:** The mask information of the target within the image.

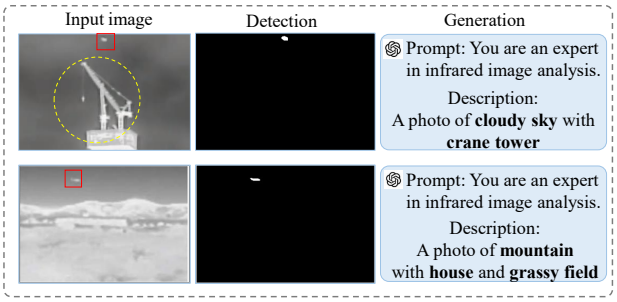

Figure 1. Generating descriptions of infrared images on MIRSTD.

The dataset contains a total of 427, 1327 and 1,001 records, stored in separate subdirectories (refer to Section 2.3 of the main paper for further details). Each image is encoded in Base64 and stored alongside its textual description for multimodal processing. The textual descriptions are generated by ChatGPT and manually verified to ensure their accuracy and brevity. As shown in Fig 1, We provided some examples, offering a better task description, which enabled GPT-4 Vision to generate more accurate text descriptions.

1.3. Dataset Development and Academic Collaboration

The MIRSTD dataset was developed by an academic team, including both students and faculty members. Through this multimodal dataset, we aim to provide a novel tool for small target detection research, especially in tasks where targets must be detected against complex backgrounds. By combining both image and text information, we provide richer context to the model, which can enhance its ability to identify and locate small targets. In conclusion, the MIRSTD dataset combines infrared images with textual descriptions, providing new opportunities for multimodal learning in small target detection. Its carefully designed structure aims to support future research in multimodal small target detection, particularly in complex scenarios, and contribute to the advancement of this field.

References