

M-SpecGene: Generalized Foundation Model for RGBT Multispectral Vision

Kailai Zhou^{1,2}, Fuqiang Yang¹, Shixian Wang¹, Bihan Wen², Chongde Zi¹,
Linsen Chen^{1*}, Qiu Shen¹, Xun Cao^{1*}

¹Nanjing University, Nanjing, China ²Nanyang Technological University, Singapore

calayzhou@smail.nju.edu.cn {chenls, caoxun}@nju.edu.cn

Abstract

RGB-Thermal (RGBT) multispectral vision is essential for robust perception in complex environments. Most RGBT tasks follow a case-by-case research paradigm, relying on manually customized models to learn task-oriented representations. Nevertheless, this paradigm is inherently constrained by artificial inductive bias, modality bias, and data bottleneck. To address these limitations, we make the initial attempt to build a Generalized RGBT MultiSpectral foundation model (M-SpecGene), which aims to learn modality-invariant representations from large-scale broad data in a self-supervised manner. M-SpecGene provides new insights into multispectral fusion and integrates prior case-by-case studies into a unified paradigm. Considering the unique characteristic of information imbalance in RGBT data, we introduce the Cross-Modality Structural Sparsity (CMSS) metric to quantify the information density across two modalities. Then we develop the GMM-CMSS progressive masking strategy to facilitate a flexible, easy-to-hard, and object-centric pre-training process. Comprehensive experiments validate M-SpecGene’s generalizability across eleven datasets for four RGBT downstream tasks. The code will be available at <https://github.com/CalayZhou/M-SpecGene>.

1. Introduction

RGB sensors alone struggle to handle complex environmental conditions, including smog, low light, and high dynamic range scenarios. RGBT multispectral vision, with its all-weather, round-the-clock sensing capabilities, has emerged as a crucial technology in fields like autonomous driving, military defense, remote sensing, and industrial inspection.

Currently, most RGBT downstream tasks follow a case-by-case research paradigm. For a given task, task-oriented representations are learned via fully supervised learning on small, task-specific datasets, often using models pre-trained on ImageNet or trained from scratch. As illustrated in Fig. 1(a), existing methods commonly use two-stream

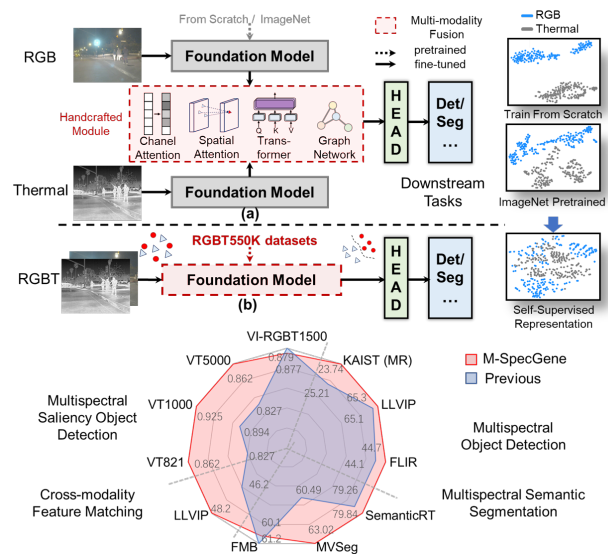


Figure 1. (a) Manually customized models: task-oriented representations are learned under a case-by-case research paradigm. (b) Generalized RGBT multispectral foundation model aims to learn modality-invariant representations by self-supervised learning. The t-SNE visualization of RGB and thermal features indicates M-SpecGene achieves superior cross-modality alignment.

branches to extract features from both RGB and thermal images, incorporating complex handcrafted modules in the intermediate feature space, such as channel attention [80], spatial attention [69], Transformer [41], and graph network [44]. However, this case-by-case paradigm has several limitations: **1) Artificial inductive bias:** Task-oriented, manually customized models, being optimized for a given task, are effective for that task but may lead to suboptimal results on others, thereby restricting both the scalability of the designed model and the generalizability of the learned representations. **2) Modality bias:** Due to inherent differences between RGB and thermal modalities, initializing the thermal branch with the ImageNet pretrained model inevitably introduces modality bias. This bias can potentially impair the encoded prior knowledge and result in suboptimal feature representations for the thermal modality. **3) Data bottleneck:** RGBT multispectral images are harder to obtain

than single RGB images, and high-quality manual annotation for large datasets is costly and time-intensive.

Recently, foundation models, with their capacity to encode extensive knowledge [2], offer a potential solution to above limitations. As shown in Fig. 1(b), we make an initial attempt to transform manually customized models into a generalized multispectral foundation model named M-SpecGene, which aims to explore a new RGBT fusion paradigm that learns modality-invariant representations in a self-supervised manner, therefore eliminating the need for handcrafted modules and facilitating multi-modality feature fusion in a simple yet effective way. However, the self-supervised pre-training of generalized multispectral foundation model is challenging, due to the lack of large-scale datasets and the inherent information imbalance in RGBT data. In contrast to RGB images, thermal images lack rich textures, colors, and fine details. Moreover, significant differences in imaging mechanisms introduce asymmetry in information density between the two modalities. Additionally, RGBT datasets are not object-centric like ImageNet [7]; instead, they tend to include smaller, less salient objects with dispersed and uneven information distribution.

To address above problems, M-SpecGene employs a Siamese architecture and a progressive masking strategy to promote consistent representations in latent space. Leveraging the unique correlations within multispectral images, we introduce cross-modality structural sparsity to quantify information density between two modalities. Then we develop a Gaussian Mixture Model (GMM) to fit the overall CMSS distribution of the whole pre-training datasets, enabling a flexible, modality-balanced masking strategy that progresses from easier to more difficult learning stages. Our GMM-CMSS progressive masking strategy alleviates the impact of information imbalance in self-supervised pre-training, enhancing the encoder’s ability to focus on consistent, modality-invariant, and object-centric representations.

M-SpecGene provides new insights into the RGBT fusion paradigm and offers the following advantages: **1) Simplified model design:** A single foundation model can effectively represent both RGB and thermal modalities, eliminating the need for complex handcrafted modules and facilitating the adaptation of single-modality RGB methods to RGBT two-modality tasks. **2) Generalized representation:** Self-supervised pre-training on large-scale data enables M-SpecGene to learn a versatile representation that overcomes limitations associated with artificial inductive and modality biases, making it adaptable to a diverse range of downstream tasks. **3) Enhanced data utilization:** M-SpecGene fully integrates self-supervised pre-training data from existing RGBT tasks without the need for human annotations. Our contributions are as follows:

- We make the first attempt to build a multispectral foundation model, M-SpecGene, exploring a new RGBT fusion

paradigm that eliminates the need for handcrafted modules.

- A high-quality, large-scale dataset, RGBT550K is carefully constructed for self-supervised pre-training.
- Considering the unique characteristic of RGBT datasets, we introduce a GMM-CMSS progressive masking strategy to mitigate the impact of information imbalance.
- M-SpecGene integrates prior case-by-case studies into a unified paradigm and demonstrates strong generalizability across eleven datasets for four RGBT downstream tasks.

2. Related Work

2.1. Task-Oriented RGBT Multispectral Vision

We first make an overview of the related RGBT multispectral vision tasks. **a) Multispectral Object Detection:** Previous methods can be divided into three categories: 1) Early fusion at the image level. 2) Halfway fusion at the feature level. 3) Late fusion in a post-process manner. Halfway fusion has emerged as a primary focus, involving an interaction module across modalities, such as channel attention [80], spatial attention [1, 69, 72], and Transformer [26, 41, 42]. **b) Multispectral Semantic Segmentation:** Early studies adopt straightforward strategies, such as concatenating RGB and thermal features [13] or integrating thermal features into the RGB encoder [6, 81]. Recent investigations explore weighted attention-based fusion strategies to achieve robust cross-modality fusion, utilizing techniques such as multi-scale spatial and channel context modules [77], explicit complement modeling framework [22], edge-aware guidance fusion [82], and spatio-temporal context integration [23]. **c) RGBT Cross-modality Feature Matching:** Modality-invariant representation plays a crucial role in cross-modality feature matching. Traditional handcrafted methods [30] design reliable filters that exhibit certain robustness to modality differences, while recent deep learning methods [8] leverage loss functions to supervise the extraction of features. Nevertheless, existing methods suffer from limited generalization and robustness. **d) Multispectral Salient Object Detection:** Compared to semantic segmentation, saliency object detection faces challenges such as background complexity and contextual understanding. Thus, technologies such as the manifold ranking algorithm [52], multi-interaction block [48], and multiple graph affinity interactive network [44] are proposed.

In conclusion, previous RGBT downstream tasks primarily follow a case-by-case research paradigm. In this paper, we explore the transformation of multispectral fusion paradigm from the perspective of foundation model.

2.2. Spectral Foundation Model

Foundation models are initially pretrained on large-scale broad data in a self-supervised manner, and can be adapted (e.g., fine-tuned) for a wide range of downstream tasks [2].

Foundation models driven by self-supervised learning for specialized data types have emerged in various areas, such as SARATR-X [32] for synthetic aperture radar, InfMAE [36] for infrared images, and EVA-X [63] for X-ray images. Research on spectral foundation models mainly focuses on hyperspectral images in remote sensing, including SpectralGPT [16] and HyperSIGMA [51]. Currently, there is a lack of research into the RGBT multispectral foundation model. A recent relevant work, UniRGB-IR [66], utilizes ViT-B as the pretrained foundation model and dynamically introduces richer RGB-IR features into the RGB-based pretrained model. Nevertheless, UniRGB-IR still requires the handcrafted fusion module and the adapter tuning design may not make adequate integration of two modalities. We make an initial attempt to develop multispectral foundation model, aiming to eliminate handcrafted modules by fully exploit large-scale RGBT data in a self-supervised manner.

2.3. Information-aware Masking Strategy

Compared to ImageNet [7], RGBT datasets exhibit a distinct characteristic of information imbalance. One solution involves an information-aware masking strategy, which aims to optimally choose what parts of the image to mask based on the informational value. For thermal images, InfMAE [36] implements information-aware masking based on gray values. For RGB images, previous methods rely on teacher-student framework [25, 53], semantic information learned by ViT [29], CLIP [17] or segmentation task pre-training [58] to measure information density distribution. However, these methods often necessitate extra components or incur higher computational costs. Furthermore, it should be noted that single-modality-based methods are difficult to adapt to multispectral images directly. We contend that the unique correlations between the two modalities can be leveraged to offer valuable clues for advanced information-aware masking.

3. RGBT550K Dataset

To pretrain a multispectral foundation model with robust generalization capabilities, we exert our utmost efforts to make a comprehensive collection of available RGBT datasets, resulting in three million RGBT samples (termed RGBT3M) drawn from 41 datasets and 10 multispectral tasks. Although RGBT3M offers substantial image quantity, we argue that diversity and quality are more critical. The RGBT3M dataset has several limitations: 1) Imbalance across datasets: RGBT detection and segmentation datasets [38, 71], typically contain fewer than 10,000 samples, while RGBT tracking datasets [28] often exceed 100,000 samples. 2) Temporal redundancy: Although tracking datasets contain hundreds of thousands of samples, they cover only a few hundred unique scenarios, leading to significant temporal redundancy; 3) Low image quality: Many datasets

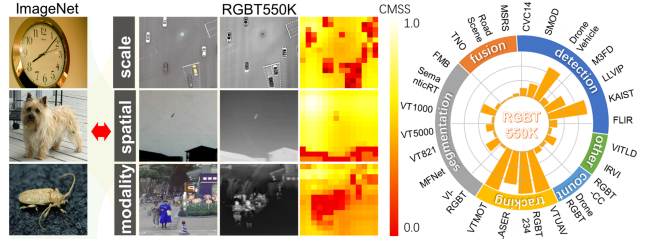


Figure 2. RGBT550K consists of diverse resources, it exhibits an imbalanced information distribution compared to ImageNet.

are captured in challenging conditions, such as nighttime or rainy scenes, resulting in lower imaging quality.

Thus, we refine the RGBT3M through the following steps: 1) Ensuring dataset balance: We prevent any single dataset to dominate an excessive proportion. 2) Removing redundancy: Temporal sampling is applied to RGBT video datasets to eliminate highly similar frames. 3) Evaluating image quality: Using objective metrics, we find that SSIM [56] is an effective measure of RGBT image quality. We remove samples with SSIM values below 0.80, as these images generally lack sufficient object information or are of poor quality. As shown in Fig. 2, our meticulous preprocessing yields RGBT550K, a comprehensive dataset comprising 548,238 high-quality samples. It encompasses diverse scenarios, tasks, lighting conditions, resolutions, and object categories, providing a solid foundation for the self-supervised pre-training of the multispectral foundation model. Further details can be found in the appendix.

4. Method

As shown in Fig. 3(a), our M-SpecGene adopts a Siamese-based architecture based on masked autoencoders [14] for cross-modality self-supervised learning. It begins with the GMM-CMSS progressive masking strategy, which dynamically selects masked patches based on information density. The complementary masked RGB and thermal patches are processed with a shared-weight ViT [10] encoder, a cross-attention layer is then employed to facilitate the propagation of complementary information in latent space. Finally, two modality-specific decoders with self-attention layers reconstruct the masked pixels for the RGB and thermal modalities independently. The Siamese-based architecture encourages both modalities to produce consistent representations. After self-supervised pre-training, we adopt the M-SpecGene ViT encoder for fine-tuning on downstream tasks, which will be explained in detail in Sec. 4.4.

The GMM-CMSS progressive masking strategy consists of three steps: 1) Given the uneven information distribution in RGBT datasets, we compute the CMSS metric for each RGBT image pair to quantify information density. 2) We employ Gaussian mixture modeling to estimate the overall CMSS distribution, which serves as a guide for subsequent information-aware masking. 3) A sampling function is de-

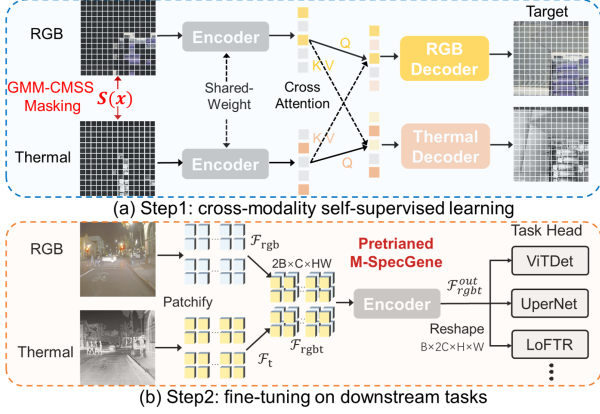


Figure 3. (a) The self-supervised pre-training of M-SpecGene. (b) The fine-tuning of M-SpecGene on downstream tasks.

signed based on GMM to implement the progressive masking strategy. With these steps, unmasked patches gradually move from foreground to background during pre-training.

4.1. Cross-modality Structural Sparsity

Fig. 2 shows a prominent characteristic of RGBT datasets is their pronounced information imbalance, reflected in the uneven distribution of object scales, spatial and modality information density. Unlike ImageNet [7], where objects are typically centered and occupy a larger portion of the image, RGBT datasets are not object-centered; they tend to contain smaller, less prominent objects with uneven spatial distribution. Additionally, differences in imaging mechanisms lead to modality imbalance [80], which means asymmetric information density between RGB and thermal modalities under varying conditions. Consequently, the random masking strategy used in MAE [14] may disproportionately focus on information-sparse regions, undermining effective self-supervised learning. Therefore, we aim to develop an adaptive masking strategy based on the measurement of information density across modalities. Specifically, we divide RGB and thermal images into $p \times p$ non-overlapping patch embeddings, denoted as $A_{rgb} = \{a_i\}_{i=1}^{p \times p}$, $B_t = \{b_i\}_{i=1}^{p \times p}$, respectively. Here, $a_i, b_i \in \mathbb{R}^{768}$ are feature vectors of the i -th patch embeddings. For each patch embedding pair (a, b) , we define cross-modality structural sparsity as follows:

$$m = CMSS(a, b) = \frac{1 + \langle \frac{a}{|a|}, \frac{b}{|b|} \rangle}{2\sigma_a^2 \sigma_b^2} \quad (1)$$

where the numerator represents the cosine similarity between RGB and thermal patch embeddings. The denominator consists of the structural variances of a and b . To facilitate post-processing, the value of m is normalized to the range $[0, 1]$. Fig. 2 shows that in low information density regions (e.g., sky), patch embedding pairs (a, b) exhibit high similarity and low structural variance, resulting in relatively high CMSS value. Conversely, in high-information density regions (e.g., pedestrians), (a, b) exhibit greater differ-

ences, yielding lower similarity but higher structural variance. Consequently, CMSS tends to have lower values in regions with rich semantic context. Thus, we employ the CMSS as a simple but effective metric to evaluate the information density across RGBT patch embedding pairs.

4.2. CMSS Gaussian Mixture Modeling

For the whole pre-training dataset comprising N image pairs, where each image pair contains $p \times p$ patch embeddings, the overall CMSS distribution can be denoted as $\mathbf{m} = \{m_i\}_{i=1}^{N \times p \times p}$. The primary problem is to develop an effective masking strategy based on this overall CMSS distribution \mathbf{m} . To address this, we first apply a Gaussian mixture model to estimate the whole CMSS distribution \mathbf{m} via maximum likelihood. After estimating the \mathbf{m} with Gaussian mixture model, we dynamically adjust masked patches based on the Gaussian model associated with specific CMSS distribution intervals. We model the observed CMSS m for the patch embedding (a, b) from the underlying distribution \mathbf{m} as:

$$p(m) = \sum_{k=1}^K \pi_k \mathcal{N}(m | \mu_k, \Sigma_k) \quad (2)$$

here, $p(m)$ represents the CMSS probability density function to be estimated by Gaussian mixture model; K denotes the number of Gaussian components, which is set to 3 by default; π_k is the weight of the k -th Gaussian component $\mathcal{N}(m | \mu_k, \Sigma_k)$ with mean μ_k and variance Σ_k . Calculating CMSS metrics for the entire pre-training dataset at once is computationally expensive. Moreover, the trainable linear projection parameters are continually updated during pre-training. Consequently, we aim to dynamically update the Gaussian mixture model estimation, synchronized with the pre-training process on an epoch-by-epoch basis. During each pre-training iteration, we calculate $B \times p \times p$ CMSS samples, denoted as $\mathbf{m}_{iter} = \{m_i\}_{i=1}^{B \times p \times p}$ for B image pairs in each batch. In the estimation step, the posterior probability of each CMSS sample m_i belonging to the k -th Gaussian model is estimated as follows:

$$\alpha_{ik} = \frac{\pi_k \mathcal{N}(m_i | \mu_k, \Sigma_k)}{\sum_{i=1}^K \pi_k \mathcal{N}(m_i | \mu_k, \Sigma_k)} \quad (3)$$

Using the posterior probability α_{ik} , we then update the parameters of the Gaussian mixture model $\{\mu_k, \Sigma_k, \pi_k\}$ in the maximization step:

$$\left\{ \begin{array}{l} \mu_k = \frac{\sum_{i=1}^{B \times p \times p} \alpha_{ik} m_i}{\sum_{i=1}^{B \times p \times p} \alpha_{ik}} \\ \Sigma_k = \frac{\sum_{i=1}^{B \times p \times p} \alpha_{ik} (m_i - \mu_k)(m_i - \mu_k)^T}{\sum_{i=1}^{B \times p \times p} \alpha_{ik}} \\ \pi_k = \frac{\sum_{i=1}^{B \times p \times p} \alpha_{ik}}{B \times p \times p} \end{array} \right. \quad (4)$$

| Methods | Near | Medium | Far | None | Partial | Heavy | Day | Night | All | FLIR | | | LLVIP | | |
|------------------------|-------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|-------------|-------------------|-------------------|-------------|-------------------|-------------------|
| | | | | | | | | | | mAP | mAP ₅₀ | mAP ₇₅ | mAP | mAP ₅₀ | mAP ₇₅ |
| ACF [20] | 28.74 | 53.67 | 88.20 | 62.94 | 81.40 | 88.08 | 64.31 | 75.06 | 67.74 | 35.8 | 71.5 | - | 55.1 | 91.4 | - |
| Halfway Fusion [37] | 8.13 | 30.34 | 75.70 | 43.13 | 65.21 | 74.36 | 47.58 | 52.35 | 49.18 | 37.4 | 74.7 | 31.3 | 55.8 | 94.0 | 60.2 |
| IATDNN+IASS [12] | 0.04 | 28.55 | 83.42 | 45.43 | 46.25 | 64.57 | 49.02 | 49.37 | 48.96 | 37.9 | 75.5 | 31.8 | 51.5 | 93.4 | 50.2 |
| CLAN [74] | 3.71 | 19.04 | 55.82 | 30.31 | 41.57 | 62.48 | 36.02 | 32.38 | 35.53 | 41.3 | 79.2 | 37.4 | 59.2 | 94.3 | 66.6 |
| MSDS-R-CNN [27] | 1.29 | 16.19 | 63.73 | 29.86 | 38.71 | 63.37 | 32.06 | 38.83 | 34.15 | - | - | - | 63.9 | - | - |
| AR-CNN [75] | 0.00 | 16.08 | 69.00 | 31.40 | 38.63 | 55.73 | 34.36 | 36.12 | 34.95 | 44.3 | 81.4 | 41.1 | 64.2 | 96.3 | 73.1 |
| MBNet [80] | 0.00 | 16.07 | 55.99 | 27.74 | 35.43 | 59.14 | 32.37 | 30.95 | 31.87 | 40.5 | 79.8 | 35.8 | 64.4 | 98.9 | 73.5 |
| TSFADet [65] | 0.00 | 15.99 | 50.71 | 25.63 | 37.29 | 65.67 | 31.76 | 27.44 | 30.74 | 44.0 | 78.1 | - | - | - | - |
| CMPD [31] | 0.00 | 12.99 | 51.22 | 24.04 | 33.88 | 59.37 | 28.30 | 30.56 | 28.98 | 41.4 | 79.2 | 36.9 | - | - | - |
| CAGTDet [67] | 0.00 | 14.00 | 49.40 | 24.48 | 33.20 | 59.35 | 28.79 | 27.73 | 28.96 | 42.1 | 79.3 | 38.5 | 65.1 | 97.4 | 75.4 |
| C2Former [64] | 0.00 | 13.71 | 48.14 | 23.91 | 32.84 | 57.81 | 28.48 | 26.67 | 28.39 | 41.4 | 81.1 | - | 59.2 | 94.3 | - |
| RSDet [79] | 0.00 | 12.13 | 39.80 | 20.49 | 33.25 | 57.60 | 25.83 | 26.48 | 26.02 | 44.1 | 81.4 | 40.2 | 63.2 | 96.1 | 72.2 |
| UniRGB-IR (ViT-B) [66] | 0.00 | 13.44 | 38.21 | 20.26 | 31.67 | 55.03 | 25.93 | 23.95 | 25.21 | 43.7 | 82.4 | 39.4 | 63.4 | 96.3 | 74.1 |
| M-SpecGene (ViT-S) | 0.03 | 16.00 | 40.54 | 22.70 | 33.92 | 55.91 | 28.28 | 25.15 | 27.28 | 44.7 | 84.8 | 40.1 | 65.3 | 97.4 | 75.4 |
| M-SpecGene (ViT-B) | 0.00 | 12.05 | 34.57 | 18.20 | 33.32 | 55.85 | 25.66 | 19.42 | 23.74 | | | | | | |

(a) Comparison results on nine test subsets of the KAIST dataset in terms of MR^2 .

(b) Evaluation on the FLIR and LLVIP datasets in terms of mAP.

Table 1. Evaluation of the proposed M-SpecGene on the KAIST, FLIR and LLVIP datasets for the multispectral object detection task.

| Methods | Bkg | Bike | Bicyclist | Car | Tricycle | Box | Pole | Curve | Person | mIoU (%) | Methods | | | Backbone | mIoU (%) |
|------------------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|-------------|--------------|----------------|----------|----------|
| | | | | | | | | | | | OCRNet [68] | LMANet [40] | DeepLabv3+ [4] | | |
| PSTNet [43] | 95.03 | 62.25 | 58.48 | 85.41 | 44.18 | 83.00 | 71.65 | 62.15 | 72.21 | 67.98 | ResNet-50 | 52.38 | | | |
| MFNet [13] | 96.31 | 65.87 | 64.07 | 89.70 | 62.10 | 83.93 | 77.14 | 66.18 | 80.29 | 74.08 | ResNet-50 | 52.73 | | | |
| RTFNet [46] | 96.40 | 67.96 | 67.41 | 90.39 | 65.96 | 85.91 | 78.02 | 67.22 | 78.90 | 75.48 | ResNet-50 | 51.59 | | | |
| EGFNet [82] | 96.57 | 71.26 | 70.86 | 90.52 | 71.51 | 85.41 | 76.49 | 66.92 | 83.74 | 77.44 | ResNet-50 | 54.52 | | | |
| ECM [22] | 96.55 | 75.04 | 75.50 | 90.26 | 74.01 | 85.61 | 77.23 | 68.28 | 85.02 | 79.26 | MiT-B5 | 57.90 | | | |
| UniRGB-IR (ViT-B) [66] | 96.33 | 68.72 | 64.79 | 90.33 | 69.43 | 85.57 | 76.44 | 65.56 | 79.79 | 75.21 | ViT-B | 56.46 | | | |
| M-SpecGene (ViT-S) | 96.74 | 73.82 | 71.17 | 91.01 | 73.08 | 85.87 | 77.95 | 68.51 | 84.64 | 78.42 | ViT-S | 60.49 | | | |
| M-SpecGene (ViT-B) | 96.81 | 75.99 | 75.51 | 91.11 | 76.79 | 86.05 | 78.41 | 68.64 | 85.66 | 79.84 | ViT-B | 63.02 | | | |

(a) Quantitative segmentation results on each class of the SemanticRT test set.

(b) Quantitative evaluation on the MVSeg dataset.

Table 2. Comparison of the M-SpecGene on the SemanticRT and MVSeg datasets for the multispectral semantic segmentation task.

flipping. By default, a 90% masking ratio is applied to both RGB and thermal images initially, and the AdamW optimizer is used with a base learning rate of 1.5×10^{-4} and a half-cycle cosine decay schedule on 8 GTX 4090 GPUs. Following previous studies [16, 32, 36], after self-supervised pre-training, M-SpecGene is full-parameter fine-tuned on downstream RGBT multispectral tasks.

5.2. RGBT Multispectral Object Detection

Experimental Settings: We validate M-SpecGene on the multispectral object detection across three datasets: KAIST [21], LLVIP [24], and FLIR [71]. We evaluate pedestrian detection on the KAIST dataset using the log-average Miss Rate over false positives per image (MR^2). For the LLVIP and FLIR datasets, we use mean Average Precision (mAP) for evaluation. To fully leverage the capabilities of the plain vision transformer, we use ViTDet [34] as the detector. Notably, RGB and thermal images undergo consistent data augmentation, and RGBT features are fused via simple concatenation of the ViT encoder outputs.

Results and Analyses: As shown in Tab. 1(a), our M-SpecGene achieves the best performance across the seven of the nine evaluation metrics on the KAIST dataset, outperforming the previous best method UniRGB-IR [66] by 1.47% on the “ALL” set. On the FLIR and LLVIP datasets, the ViT-S version of M-SpecGene achieves performance comparable to UniRGB-IR, while the ViT-B version demonstrates an enhanced ability to leverage founda-

tional model strengths in Tab. 1(b), achieving higher detection accuracy than previous methods. It should be noted that the ViT-B in UniRGB-IR is pretrained on COCO dataset first, while our M-SpecGene does not rely on the high-quality RGB detection dataset for extra improvement. With the learned self-supervised representation from large-scale data, our M-SpecGene can effectively fuse RGB and infrared modalities without complex handcrafted modules.

5.3. RGBT Multispectral Semantic Segmentation

Experimental Settings: Three recently released datasets which own high-quality samples are used for the validation on the multispectral semantic segmentation task. The SemanticRT [22], MVSeg [23] and FMB [38] datasets include 13, 26, and 15 categories, respectively. Mean Intersection over Union (mIoU) across all categories is used to evaluate semantic segmentation performance. Following MAE [14], we employ UperNet [59] as the base segmentation framework. The model architecture remains unchanged and only a simple concatenation operation is added.

Results and Analyses: We compare M-SpecGene with competitive methods on the SemanticRT dataset in Tab. 2(a) and the MVSeg dataset in Tab. 2(b). Quantitative results confirm the effectiveness of M-SpecGene on both datasets. MVNet [23] serves as simple baseline that uses multispectral video clips to leverage extra temporal information, while M-SpecGene achieves higher mIoU accuracy by only utilizing the frame-level information. Tab. 3 shows that on

| | Person | Truck | Vege. | Pole | mIoU (%) | | Methods | @3°↑ | @5°↑ | @10°↑ | | E_{ξ}^{\max} ↑ | F_{β}^{\max} ↑ | S_{α} ↑ | MAE↓ |
|------------------------|-------------|-------------|-------------|-------------|-------------|----------------|---------------|-------------|-------------|-------------|--------------|--------------------|----------------------|----------------|--------------|
| SegMiF [38] | 65.5 | 42.4 | 85.1 | 35.7 | 58.5 | Detector-based | RIFT [30] | 0.0 | 0.0 | 0.0 | MGFL [19] | 0.822 | 0.727 | 0.745 | 0.084 |
| MDRNet+ [78] | 67.0 | 27.0 | 82.7 | 45.3 | 55.5 | | POS-GIFT [18] | 0.0 | 0.0 | 0.4 | MIDD [49] | 0.928 | 0.859 | 0.867 | 0.049 |
| SGFNet [55] | 67.2 | 34.6 | 82.7 | 42.8 | 56.0 | | ReDFeat [8] | 0.0 | 0.0 | 0.0 | CGFNet [54] | 0.927 | 0.870 | 0.865 | 0.042 |
| MRFS [73] | 71.3 | 34.4 | 87.0 | 53.6 | 61.2 | | SP+LG [35] | 1.1 | 8.4 | 16.2 | ADF [50] | 0.892 | 0.815 | 0.830 | 0.074 |
| UniRGB-IR (ViT-B) [66] | 66.5 | 36.3 | 85.6 | 42.1 | 59.8 | Detector-free | SemLA [62] | 0.0 | 0.2 | 1.2 | MGAI [44] | 0.940 | 0.879 | 0.881 | 0.038 |
| M-SpecGene (ViT-S) | 68.8 | 22.6 | 86.2 | 50.0 | 56.5 | | LoFTR [45] | 18.8 | 29.7 | 46.2 | Ours (ViT-S) | 0.847 | 0.722 | 0.781 | 0.081 |
| M-SpecGene (ViT-B) | 65.6 | 44.4 | 86.9 | 52.8 | 60.1 | | Ours (ViT-B) | 20.5 | 31.7 | 48.2 | Ours (ViT-B) | 0.942 | 0.877 | 0.888 | 0.033 |

Table 3. Evaluation on the FMB segmentation dataset.

Table 4. RGBT feature matching evaluation.

Table 5. Test on VI-RGBT1500.

| Methods | VT821 | | | | VT1000 | | | | VT5000 | | | |
|------------------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
| | S ↑ | $adpE$ ↑ | $adpF$ ↑ | MAE ↓ | S ↑ | $adpE$ ↑ | $adpF$ ↑ | MAE ↓ | S ↑ | $adpE$ ↑ | $adpF$ ↑ | MAE ↓ |
| S2MA [39] | 0.811 | 0.813 | 0.709 | 0.098 | 0.918 | 0.912 | 0.848 | 0.029 | 0.853 | 0.864 | 0.743 | 0.053 |
| JLDCF [11] | 0.839 | 0.830 | 0.726 | 0.076 | 0.912 | 0.899 | 0.829 | 0.030 | 0.861 | 0.860 | 0.739 | 0.050 |
| MTMR [52] | 0.725 | 0.815 | 0.662 | 0.109 | 0.706 | 0.836 | 0.715 | 0.119 | 0.680 | 0.795 | 0.595 | 0.114 |
| FMSF [76] | 0.760 | 0.796 | 0.640 | 0.080 | 0.873 | 0.899 | 0.823 | 0.037 | 0.814 | 0.864 | 0.734 | 0.055 |
| MIDD [49] | 0.871 | 0.895 | 0.803 | 0.033 | 0.915 | 0.933 | 0.880 | 0.027 | 0.868 | 0.896 | 0.799 | 0.043 |
| ADF [50] | 0.810 | 0.842 | 0.717 | 0.077 | 0.910 | 0.921 | 0.847 | 0.034 | 0.864 | 0.891 | 0.778 | 0.048 |
| LSNet [83] | 0.877 | 0.911 | 0.827 | 0.033 | 0.924 | 0.936 | 0.887 | 0.022 | 0.876 | 0.916 | 0.827 | 0.036 |
| UniRGB-IR (ViT-B) [66] | 0.881 | 0.895 | 0.806 | 0.039 | 0.939 | 0.943 | 0.894 | 0.018 | 0.906 | 0.935 | 0.849 | 0.027 |
| M-SpecGene (ViT-S) | 0.783 | 0.826 | 0.703 | 0.079 | 0.867 | 0.889 | 0.827 | 0.043 | 0.853 | 0.892 | 0.803 | 0.044 |
| M-SpecGene (ViT-B) | 0.891 | 0.919 | 0.862 | 0.028 | 0.935 | 0.952 | 0.925 | 0.015 | 0.892 | 0.928 | 0.872 | 0.028 |

Table 6. Comparison of M-SpecGene on the VT821, VT1000 and VT5000 datasets for the multispectral salient object detection task.

the FMB dataset, M-SpecGene is superior to other competitive methods but falls short of MSRS [73] on certain metrics. Given that FMB is a small-scale dataset with only 280 validation samples, MSRS and UniRGB-IR, which incorporate complex fusion modules based on Segformer [61], tend to fit the FMB more easily than M-SpecGene, which only employs a simple concatenation operation for feature fusion. M-SpecGene tends to achieve superior performance, particularly in scenarios involving extensive category diversity, large-scale datasets, and high task complexity.

5.4. RGBT Cross-modality Feature Matching

Experimental Settings: Considering the high alignment quality, LLVIP [24] dataset is used to evaluate cross-modality feature matching. The Area Under the Curve (AUC) metric is used for evaluation. We adopt the widely recognized LoFTR [45] as the basic framework, with the backbone replaced by ViT-S. To enhance locality, we incorporate a convolutional stem [60].

Results and Analyses: Tab. 4 shows that traditional handcrafted feature descriptors struggle to handle complex scenes in the LLVIP dataset. Moreover, detector-based methods yield unsatisfactory results due to difficulties in extracting repeatable keypoints across two modalities. Our M-SpecGene significantly outperforms other methods at various thresholds, as the learned modality-invariant representation facilitates the RGBT feature matching with reduced modality characteristic differences in latent space.

5.5. RGBT Multispectral Salient Object Detection

Experimental Settings: The VT821 [52], VT1000 [47], VT5000 [50] and VI-RGBT1500 [44] are used for evaluation on the multispectral salient object detection. F-measure ($adpF$, F_{β}^{\max}), E-Measure ($adpE$, E_{ξ}^{\max}), S-Measure (S) and

Mean Absolute Error (MAE) are adopted as metrics. We employ the UperNet [59] as the basic framework and follow the common setting that 2,500 image pairs in the VT5000 dataset are treated as the training dataset, while the remaining and other datasets are used as the test sets.

Results and Analyses: Experiments in Tab. 5 and Tab. 6 show that M-SpecGene achieves better results than previous methods across eleven subset metrics, with particularly notable improvements on the VT821, VT1000, and VI-RGBT1500 datasets, rather than the VT5000 dataset. This highlights its superior generalization capability.

5.6. Ablation Study

Comparisons on Pretrained Models: In Tab. 7(a), we compare the performance of different pretrained models in multispectral object detection using KAIST dataset and cross-modality feature matching on LLVIP dataset. We observe that ViT trained from scratch performs poorly in terms of mAP on FLIR. While vanilla MAE-pretrained ViT improves mAP₅₀ from 40.6% to 43.0% compared to the Supervised (Sup.) pretrained ViT. M-SpecGene exhibits superior performance by further improving the mAP₅₀ to 44.8%. On the LLVIP dataset, M-SpecGene significantly boosts AUC@10° from 41.2 to 48.2, whereas both supervised and vanilla MAE pretrained ViT models lead to a decline in matching accuracy. We attribute this discrepancy to the inherent difference between detection and matching tasks. The detection task aims to leverage both modalities to generate complementary features, whereas the matching task focuses on identifying the common features shared by both modalities. Therefore, pre-training on the single-modality ImageNet dataset may disrupt symmetrical representations required for cross-modality feature matching. Overall, effective pre-training for modality-invariant representation is

| Methods | mAP | mAP ₅₀ | mAP ₇₅ | @3°↑ | @5°↑ | @10°↑ | Architecture | mAP ₅₀ | Masking | mAP ₅₀ | Blocks | mAP ₅₀ | Ratio | mAP ₅₀ |
|--------------|------|-------------------|-------------------|------|------|-------|--------------|-------------------|-----------|-------------------|--------|-------------------|-------|-------------------|
| From Scratch | 36.0 | 70.6 | 32.0 | 12.5 | 23.6 | 41.2 | Vanilla MAE | 83.1 | Random | 83.8 | 2 | 84.1 | 85% | 84.4 |
| Sup. (IN1K) | 40.6 | 79.3 | 34.0 | 12.5 | 23.3 | 40.3 | Concat | 80.1 | Low CMSS | 83.6 | 4 | 84.8 | 90% | 84.8 |
| MAE (IN1K) | 43.0 | 82.8 | 37.8 | 8.4 | 18.7 | 37.0 | Auxiliary | 83.5 | High CMSS | 83.4 | 8 | 84.5 | 95% | 84.1 |
| M-SpecGene | 44.7 | 84.8 | 40.1 | 20.5 | 31.7 | 48.2 | Siamese | 83.8 | GMM-CMSS | 84.8 | | | | |

(a) Comparisons on different pretrained models.

(b) Architecture.

(c) Masking way.

(d) Decoder Depth.

(e) Masking ratio.

Table 7. Ablation analysis of M-SpecGene in terms of pretrained model, architecture, masking strategy, decoder depth and masking ratio.

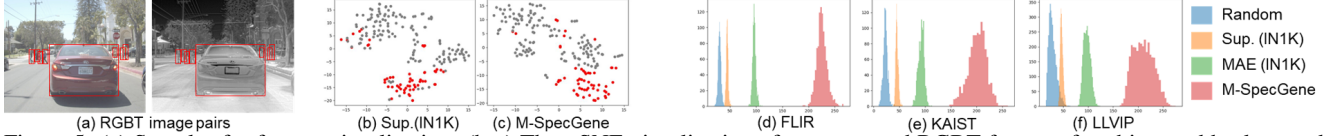


Figure 5. (a) Samples for feature visualization. (b-c) The t-SNE visualization of concatenated RGBT features for object and background regions. (d-f) The statistical distribution of the Wasserstein distance between object and background features on three detection datasets.

crucial for a generalized multispectral foundation model.

RGBT Representation Architecture: To investigate effective self-supervised representation architectures for both RGB and thermal modalities, we design four approaches: 1) Vanilla MAE [14]: RGB and thermal images are mixed in the input level, and a vanilla MAE is employed. 2) Channel concatenation: RGB and thermal images are concatenated along the channel dimension. 3) Auxiliary branch: Complementary masked RGB and thermal patches are processed with a shared-weight encoder, then thermal features serve as auxiliary information in the cross-attention layer to aid the RGB decoder in reconstructing the masked region. 4) Siamese-based: RGB and thermal modalities are encouraged to learn consistent representations with a shared-weight encoder, with independent decoders applied to each modality. Tab. 7(b) shows the Siamese-based architecture achieves the best results, which reserves the symmetry and fully utilizes cross-modality complementarity.

Masking Strategy: We compare four different masking strategies in Tab. 7(c): 1) Random masking. 2) Gaussian masking in the low-CMSS region. 3) Gaussian masking in the high-CMSS region. 4) GMM-CMSS progressive masking. Experimental results indicate that focusing on a single information density region leads to inferior performance. In contrast, GMM-CMSS progressive masking enables a flexible, easy-to-hard, and object-centered learning process, thereby producing more robust representations.

Decoder Depth: Tab. 7(d) shows a decoder depth of four achieves the best results, indicating that the default decoder depth of MAE [14] can be reduced under the Siamese-based architecture with two independent decoders.

Masking Ratio: Tab. 7(e) illustrates that a lower masking ratio, which reduces the reconstruction difficulty particularly for the thermal modality, leads to a decrease in mAP slightly. A higher masking ratio will negatively affect the effectiveness of the GMM-CMSS strategy. Therefore, we set the default masking ratio to 90%.

Feature Visualization and Statistical Analysis: We first concatenate the RGB and thermal features extracted by pretrained models and perform a visual analysis of the

concatenated object and background features. As shown in Fig. 5(b-c), the object features extracted by M-SpecGene exhibit greater discriminability compared to those from the Sup. (IN1K) pretrained model. Subsequently, we conduct a statistical analysis of the differences between object and background features across three detection datasets. Specifically, we compute the Wasserstein distance between object and background features for each sample and present the statistical Wasserstein distance distribution of different pretrained models. Fig. 5(d-f) show that the model trained from scratch exhibits smaller overall Wasserstein distances, whereas the distributions of MAE (IN1K) and Sup. (IN1K) show larger Wasserstein distances. Notably, our M-SpecGene achieves the largest Wasserstein distance distribution, indicating more significant feature differences between objects and backgrounds. This suggests that the GMM-CMSS progressive masking strategy facilitates the learning of more object-centric representations, thereby promoting the generation of more discriminative features.

6. Conclusion

We make the first attempt to build a multispectral foundation model, aiming to transform previous case-by-case studies into a unified paradigm. To mitigate the impact of information imbalance inherent in RGBT datasets, we introduce the CMSS metric to measure cross-modality information density and develop a GMM-CMSS progressive masking strategy to enable a flexible, easy-to-hard, and object-centric pre-training progress. The proposed M-SpecGene effectively represents both RGB and thermal modalities in the latent space, eliminating the need for handcrafted modules and offering new insights into multispectral fusion. Extensive experiments on eleven datasets across four tasks validate the generalizability of M-SpecGene, which can fully exploit the carefully constructed, high-quality RGBT550K dataset for self-supervised pre-training and seamlessly adapt RGB single-modality methods to RGBT two-modality tasks without extra modification. We hope this work will advance the application of multispectral vision from the perspective of generalized foundation model.

Acknowledgments

This research was supported by National Science Fund for Distinguished Young Scholars (62025108). In addition, this research was supported in part by the Agency for Science, Technology and Research (A*STAR) under its IAF-ICP Programme I2501E0041 and the Schaeffler-NTU Corporate Lab (SHARE@NTU), and in part by the National Research Foundation Singapore Competitive Research Program (award number CRP29-2022-0003). The work was done at Rapid-Rich Object Search (ROSE) Lab, School of Electrical & Electronic Engineering, Nanyang Technological University.

References

- [1] Zijia An, Chunlei Liu, and Yuqi Han. Effectiveness guided cross-modal information sharing for aligned rgb-t object detection. *IEEE Signal Processing Letters*, 29:2562–2566, 2022. 2
- [2] Muhammad Awais, Muzammal Naseer, Salman Khan, Rao Muhammad Anwer, Hisham Cholakkal, Mubarak Shah, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Foundational models defining a new era in vision: A survey and outlook. *arXiv preprint arXiv:2307.13721*, 2023. 2
- [3] Yue Cao, Junchi Bin, Jozsef Hamari, Erik Blasch, and Zheng Liu. Multimodal object detection by channel switching and spatial attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 403–411, 2023. 6
- [4] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018. 6
- [5] Yi-Ting Chen, Jinghao Shi, Zelin Ye, Christoph Mertz, Deva Ramanan, and Shu Kong. Multimodal object detection via probabilistic ensembling. In *European Conference on Computer Vision*, pages 139–158. Springer, 2022. 6
- [6] Fuqin Deng, Hua Feng, Mingjian Liang, Hongmin Wang, Yong Yang, Yuan Gao, Junfeng Chen, Junjie Hu, Xiyue Guo, and Tin Lun Lam. Feanet: Feature-enhanced attention network for rgb-thermal real-time semantic segmentation. In *2021 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pages 4467–4473. IEEE, 2021. 2
- [7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 2, 3, 4, 5
- [8] Yuxin Deng and Jiayi Ma. Redfeat: Recoupling detection and description for multimodal feature learning. *IEEE Transactions on Image Processing*, 32:591–602, 2022. 2, 7
- [9] Shaohua Dong, Yunhe Feng, Qing Yang, Yan Huang, Dongfang Liu, and Heng Fan. Efficient multimodal semantic segmentation via dual-prompt learning. *arXiv preprint arXiv:2312.00360*, 2023. 6
- [10] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3
- [11] Keren Fu, Deng-Ping Fan, Ge-Peng Ji, Qijun Zhao, Jianbing Shen, and Ce Zhu. Siamese network for rgb-d salient object detection and beyond. *IEEE transactions on pattern analysis and machine intelligence*, 44(9):5541–5559, 2021. 7
- [12] Dayan Guan, Yanpeng Cao, Jiangxin Yang, Yanlong Cao, and Michael Ying Yang. Fusion of multispectral data through illumination-aware deep neural networks for pedestrian detection. *Information Fusion*, 50:148–157, 2019. 6
- [13] Qishen Ha, Kohei Watanabe, Takumi Karasawa, Yoshitaka Ushiku, and Tatsuya Harada. Mfnet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5108–5115. IEEE, 2017. 2, 6
- [14] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 3, 4, 6, 8
- [15] Xiao He, Chang Tang, Xin Zou, and Wei Zhang. Multispectral object detection via cross-modal conflict-aware learning. In *Proceedings of the 31st ACM International Conference on Multimedia*, page 1465–1474, New York, NY, USA, 2023. Association for Computing Machinery. 6
- [16] Danfeng Hong, Bing Zhang, Xuyang Li, Yuxuan Li, Chenyu Li, Jing Yao, Naoto Yokoya, Hao Li, Pedram Ghamisi, Xiuping Jia, et al. Spectralgpt: Spectral remote sensing foundation model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 3, 6
- [17] Zejiang Hou, Fei Sun, Yen-Kuang Chen, Yuan Xie, and Sun-Yuan Kung. Milan: Masked image pretraining on language assisted representation. *arXiv preprint arXiv:2208.06049*, 2022. 3
- [18] Zhuolu Hou, Yuxuan Liu, and Li Zhang. Pos-gift: A geometric and intensity-invariant feature transformation for multimodal images. *Information Fusion*, 102:102027, 2024. 7
- [19] Liming Huang, Kechen Song, Jie Wang, Menghui Niu, and Yunhui Yan. Multi-graph fusion and learning for rgbt image saliency detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(3):1366–1377, 2021. 7
- [20] Soonmin Hwang, Jaesik Park, Namil Kim, Yookyung Choi, and In So Kweon. Multispectral pedestrian detection: Benchmark dataset and baseline. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1037–1045, 2015. 6
- [21] Soonmin Hwang, Jaesik Park, Namil Kim, Yookyung Choi, and In So Kweon. Multispectral pedestrian detection: Benchmark dataset and baseline. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1037–1045, 2015. 6
- [22] Wei Ji, Jingjing Li, Cheng Bian, Zhicheng Zhang, and Li Cheng. Semanticrt: A large-scale dataset and method for robust semantic segmentation in multispectral images. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 3307–3316, 2023. 2, 6

- [23] Wei Ji, Jingjing Li, Cheng Bian, Zongwei Zhou, Jiaying Zhao, Alan L Yuille, and Li Cheng. Multispectral video semantic segmentation: A benchmark dataset and baseline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1094–1104, 2023. 2, 6
- [24] Xinyu Jia, Chuang Zhu, Minzhen Li, Wenqi Tang, and Wenli Zhou. Llvip: A visible-infrared paired dataset for low-light vision. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3496–3504, 2021. 6, 7
- [25] Ioannis Kakogeorgiou, Spyros Gidaris, Bill Psomas, Yannis Avrithis, Andrei Bursuc, Konstantinos Karantzas, and Nikos Komodakis. What to hide from your students: Attention-guided masked image modeling. In *European Conference on Computer Vision*, pages 300–318. Springer, 2022. 3
- [26] Seungik Lee, Jaehyeong Park, and Jinsun Park. Crossformer: Cross-guided attention for multi-modal object detection. *Pattern Recognition Letters*, 179:144–150, 2024. 2, 6
- [27] Chengyang Li, Dan Song, Ruofeng Tong, and Min Tang. Multispectral Pedestrian Detection via Simultaneous Detection and Segmentation. *arXiv e-prints*, art. arXiv:1808.04818, 2018. 6
- [28] Chenglong Li, Wanlin Xue, Yaqing Jia, Zhichen Qu, Bin Luo, Jin Tang, and Dengdi Sun. Lasher: A large-scale high-diversity benchmark for rgbt tracking. *IEEE Transactions on Image Processing*, 31:392–404, 2021. 3
- [29] Gang Li, Heliang Zheng, Daqing Liu, Chaoyue Wang, Bing Su, and Changwen Zheng. Semmae: Semantic-guided masking for learning masked autoencoders. *Advances in Neural Information Processing Systems*, 35:14290–14302, 2022. 3
- [30] Jiayuan Li, Qingwu Hu, and Mingyao Ai. Rift: Multi-modal image matching based on radiation-variation insensitive feature transform. *IEEE Transactions on Image Processing*, 29:3296–3310, 2019. 2, 7
- [31] Qing Li, Changqing Zhang, Qinghua Hu, Huazhu Fu, and Pengfei Zhu. Confidence-aware fusion using dempster-shafer theory for multispectral pedestrian detection. *IEEE Transactions on Multimedia*, 25:3420–3431, 2022. 6
- [32] Weijie Li, Wei Yang, Yuenan Hou, Li Liu, Yongxiang Liu, and Xiang Li. Saratr-x: Toward building a foundation model for sar target recognition. *IEEE Transactions on Image Processing*, 34:869–884, 2025. 3, 6
- [33] Xiang Li, Wenhai Wang, Lijun Wu, Shuo Chen, Xiaolin Hu, Jun Li, Jinhui Tang, and Jian Yang. Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection. *Advances in Neural Information Processing Systems*, 33:21002–21012, 2020. 6
- [34] Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. In *European conference on computer vision*, pages 280–296. Springer, 2022. 5, 6
- [35] Philipp Lindenberger, Paul-Edouard Sarlin, and Marc Pollefeys. Lightglue: Local feature matching at light speed. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17627–17638, 2023. 7
- [36] Fangcen Liu, Chenqiang Gao, Yaming Zhang, Junjie Guo, Jinghao Wang, and Deyu Meng. Infmae: A foundation model in the infrared modality. In *European Conference on Computer Vision*, pages 420–437. Springer, 2025. 3, 6
- [37] Jingjing Liu, Shaoting Zhang, Shu Wang, and Dimitris N. Metaxas. Multispectral Deep Neural Networks for Pedestrian Detection. *arXiv e-prints*, art. arXiv:1611.02644, 2016. 6
- [38] Jinyuan Liu, Zhu Liu, Guanyao Wu, Long Ma, Risheng Liu, Wei Zhong, Zhongxuan Luo, and Xin Fan. Multi-interactive feature learning and a full-time multi-modality benchmark for image fusion and segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8115–8124, 2023. 3, 6, 7
- [39] Nian Liu, Ni Zhang, and Junwei Han. Learning selective self-mutual attention for rgb-d saliency detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13756–13765, 2020. 7
- [40] Matthieu Paul, Martin Danelljan, Luc Van Gool, and Radu Timofte. Local memory attention for fast video semantic segmentation. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1102–1109. IEEE, 2021. 6
- [41] Fang Qingyun, Han Dapeng, and Wang Zhaokui. Cross-modality fusion transformer for multispectral object detection. *arXiv preprint arXiv:2111.00273*, 2021. 1, 2
- [42] Jifeng Shen, Yifei Chen, Yue Liu, Xin Zuo, Heng Fan, and Wankou Yang. Icafusion: Iterative cross-attention guided feature fusion for multispectral object detection. *Pattern Recognition*, 145:109913, 2024. 2, 6
- [43] Shreyas S Shivakumar, Neil Rodrigues, Alex Zhou, Ian D Miller, Vijay Kumar, and Camillo J Taylor. Pst900: Rgb-thermal calibration, dataset and segmentation network. In *2020 IEEE international conference on robotics and automation (ICRA)*, pages 9441–9447. IEEE, 2020. 6
- [44] Kechen Song, Liming Huang, Aojun Gong, and Yunhui Yan. Multiple graph affinity interactive network and a variable illumination dataset for rgbt image salient object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(7):3104–3118, 2022. 1, 2, 7
- [45] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8922–8931, 2021. 5, 7
- [46] Yuxiang Sun, Weixun Zuo, and Ming Liu. Rtfnet: Rgb-thermal fusion network for semantic segmentation of urban scenes. *IEEE Robotics and Automation Letters*, 4(3):2576–2583, 2019. 6
- [47] Zhengzheng Tu, Tian Xia, Chenglong Li, Xiaoxiao Wang, Yan Ma, and Jin Tang. Rgb-t image saliency detection via collaborative graph learning. *IEEE Transactions on Multimedia*, 22(1):160–173, 2019. 7
- [48] Zhengzheng Tu, Zhun Li, Chenglong Li, Yang Lang, and Jin Tang. Multi-interactive encoder-decoder network for rgbt salient object detection. *arXiv e-prints*, pages arXiv:2005.2020. 2
- [49] Zhengzheng Tu, Zhun Li, Chenglong Li, Yang Lang, and Jin Tang. Multi-interactive dual-decoder for rgb-thermal salient

- object detection. *IEEE Transactions on Image Processing*, 30:5678–5691, 2021. 7
- [50] Zhengzheng Tu, Yan Ma, Zhun Li, Chenglong Li, Jieming Xu, and Yongtao Liu. Rgbt salient object detection: A large-scale dataset and benchmark. *IEEE Transactions on Multimedia*, 25:4163–4176, 2022. 7
- [51] Di Wang, Meiqi Hu, Yao Jin, Yuchun Miao, Jiaqi Yang, Yichu Xu, Xiaolei Qin, Jiaqi Ma, Lingyu Sun, Chenxing Li, et al. Hypersigma: Hyperspectral intelligence comprehension foundation model. *arXiv preprint arXiv:2406.11519*, 2024. 3
- [52] Guizhao Wang, Chenglong Li, Yunpeng Ma, Aihua Zheng, Jin Tang, and Bin Luo. Rgb-t saliency detection benchmark: Dataset, baselines, analysis and a novel approach. In *Image and Graphics Technologies and Applications, IGTA 2018, Beijing, China, April 8–10, 2018, Revised Selected Papers 13*, pages 359–369. Springer, 2018. 2, 7
- [53] Haochen Wang, Kaiyou Song, Junsong Fan, Yuxi Wang, Jin Xie, and Zhaoxiang Zhang. Hard patches mining for masked image modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10375–10385, 2023. 3
- [54] Jie Wang, Kechen Song, Yanqi Bao, Liming Huang, and Yunhui Yan. Cgfnct: Cross-guided fusion network for rgb-t salient object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(5):2949–2961, 2022. 7
- [55] Yike Wang, Gongyang Li, and Zhi Liu. Sgfnct: semantic-guided fusion network for rgb-thermal semantic segmentation. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(12):7737–7748, 2023. 7
- [56] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 3
- [57] Zeyu Wang, Fabien Colonnier, Jinghong Zheng, Jyotibha Acharya, Wenyu Jiang, and Kejie Huang. Tirdet: Mono-modality thermal infrared object detection based on prior thermal-to-visible translation. In *Proceedings of the 31st ACM International Conference on Multimedia*, page 2663–2672, New York, NY, USA, 2023. Association for Computing Machinery. 6
- [58] Jiantao Wu and Shentong Mo. Object-wise masked autoencoders for fast pre-training. *arXiv preprint arXiv:2205.14338*, 2022. 3
- [59] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Proceedings of the European conference on computer vision (ECCV)*, pages 418–434, 2018. 5, 6, 7
- [60] Tete Xiao, Mannat Singh, Eric Mintun, Trevor Darrell, Piotr Dollár, and Ross Girshick. Early convolutions help transformers see better. *Advances in neural information processing systems*, 34:30392–30400, 2021. 7
- [61] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34:12077–12090, 2021. 7
- [62] Housheng Xie, Yukuan Zhang, Junhui Qiu, Xiangshuai Zhai, Xuedong Liu, Yang Yang, Shan Zhao, Yongfang Luo, and Jianbo Zhong. Semantics lead all: Towards unified image registration and fusion from a semantic perspective. *Information Fusion*, 98:101835, 2023. 7
- [63] Jingfeng Yao, Xinggang Wang, Yuehao Song, Huangxuan Zhao, Jun Ma, Yajie Chen, Wenyu Liu, and Bo Wang. Eva-x: A foundation model for general chest x-ray analysis with self-supervised learning. *arXiv preprint arXiv:2405.05237*, 2024. 3
- [64] Maoxun Yuan and Xingxing Wei. C 2 former: Calibrated and complementary transformer for rgb-infrared object detection. *IEEE Transactions on Geoscience and Remote Sensing*, 2024. 6
- [65] Maoxun Yuan, Yinyan Wang, and Xingxing Wei. Translation, scale and rotation: cross-modal alignment meets rgb-infrared vehicle detection. In *European Conference on Computer Vision*, pages 509–525. Springer, 2022. 6
- [66] Maoxun Yuan, Bo Cui, Tianyi Zhao, and Xingxing Wei. Unirgb-ir: A unified framework for visible-infrared downstream tasks via adapter tuning. *arXiv preprint arXiv:2404.17360*, 2024. 3, 6, 7
- [67] Maoxun Yuan, Xiaorong Shi, Nan Wang, Yinyan Wang, and Xingxing Wei. Improving rgb-infrared object detection with cascade alignment-guided transformer. *Information Fusion*, 105:102246, 2024. 6
- [68] Yuhui Yuan, Xilin Chen, and Jingdong Wang. Object-contextual representations for semantic segmentation. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, pages 173–190. Springer, 2020. 6
- [69] Jun-Seok Yun, Seon-Hoo Park, and Seok Bong Yoo. Infusion-net: inter-and intra-weighted cross-fusion network for multispectral object detection. *Mathematics*, 10(21):3966, 2022. 1, 2
- [70] Yuqiao Zeng, Tengfei Liang, Yi Jin†, and Yidong Li. Mmidet: Exploring multi-modal integration for visible and infrared object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, pages 1–1, 2024. 6
- [71] Heng Zhang, Elisa Fromont, Sébastien Lefevre, and Bruno Avignon. Multispectral fusion for object detection with cyclic fuse-and-refine blocks. In *2020 IEEE International conference on image processing (ICIP)*, pages 276–280. IEEE, 2020. 3, 6
- [72] Heng Zhang, Elisa Fromont, Sébastien Lefèvre, and Bruno Avignon. Guided attentive feature fusion for multispectral pedestrian detection. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 72–80, 2021. 2, 6
- [73] Hao Zhang, Xuhui Zuo, Jie Jiang, Chunchao Guo, and Jiayi Ma. Mrfs: Mutually reinforcing image fusion and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26974–26983, 2024. 7
- [74] Lu Zhang, Zhiyong Liu, Shifeng Zhang, Xu Yang, Hong Qiao, Kaizhu Huang, and Amir Hussain. Cross-modality interactive attention network for multispectral pedestrian detection. *Information Fusion*, 50:20–29, 2019. 6

- [75] Lu Zhang, Xiangyu Zhu, Xiangyu Chen, Xu Yang, Zhen Lei, and Zhiyong Liu. Weakly Aligned Cross-Modal Learning for Multispectral Pedestrian Detection. *arXiv e-prints*, art. arXiv:1901.02645, 2019. 6
- [76] Qiang Zhang, Nianchang Huang, Lin Yao, Dingwen Zhang, Caifeng Shan, and Jungong Han. Rgb-t salient object detection via fusing multi-level cnn features. *IEEE Transactions on Image Processing*, 29:3321–3335, 2019. 7
- [77] Qiang Zhang, Shenlu Zhao, Yongjiang Luo, Dingwen Zhang, Nianchang Huang, and Jungong Han. Abmdrnet: Adaptive-weighted bi-directional modality difference reduction network for rgb-t semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2633–2642, 2021. 2
- [78] Shenlu Zhao, Yichen Liu, Qiang Jiao, Qiang Zhang, and Jungong Han. Mitigating modality discrepancies for rgb-t semantic segmentation. *IEEE Transactions on Neural Networks and Learning Systems*, 2023. 7
- [79] Tianyi Zhao, Maoxun Yuan, Feng Jiang, Nan Wang, and Xingxing Wei. Removal and selection: Improving rgb-infrared object detection via coarse-to-fine fusion. *arXiv preprint arXiv:2401.10731*, 2024. 6
- [80] Kailai Zhou, Linsen Chen, and Xun Cao. Improving multispectral pedestrian detection by addressing modality imbalance problems. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16*, pages 787–803. Springer, 2020. 1, 2, 4, 6
- [81] Wujie Zhou, Xinyang Lin, Jingsheng Lei, Lu Yu, and Jenq-Neng Hwang. Mffenet: Multiscale feature fusion and enhancement network for rgb–thermal urban road scene parsing. *IEEE Transactions on Multimedia*, 24:2526–2538, 2021. 2
- [82] Wujie Zhou, Shaohua Dong, Caie Xu, and Yaguan Qian. Edge-aware guidance fusion network for rgb–thermal scene parsing. In *Proceedings of the AAAI conference on artificial intelligence*, pages 3571–3579, 2022. 2, 6
- [83] Wujie Zhou, Yun Zhu, Jingsheng Lei, Rongwang Yang, and Lu Yu. Lsnet: Lightweight spatial boosting network for detecting salient objects in rgb-thermal images. *IEEE Transactions on Image Processing*, 32:1329–1340, 2023. 7