M-SpecGene: Generalized Foundation Model for RGBT Multispectral Vision Supplementary Material

Kailai Zhou[1,2], Fuqiang Yang[1], Shixian Wang[1], Bihan Wen[2], Chongde Zi[1],
Linsen Chen[1*], Qiu Shen[1], Xun Cao[1*]
[1]Nanjing University, Nanjing, China [2]Nanyang Technological University, Singapore
{calayzhou, fuqiangyang, 211180050}@smail.nju.edu.cn bihan.wen@ntu.edu.sg
{zichongde, chenls, shenqiu, caoxun}@nju.edu.cn

In the supplementary materials, we first present the additional experimental analysis and introduce the construction of the Thermal480K and RGBT550K datasets, detailing their sources, constitutions, and a selection of visualized samples. Next, we explain the process of filtering samples from the RGBT3M dataset to form the RGBT550K dataset using the Structural Similarity Index Measure (SSIM). We then provide additional visualizations of Cross-Modality Structural Sparsity (CMSS), demonstrating its reliability as a measure of information density in RGBT multispectral images. Finally, we present the probability density functions fitted by the Gaussian Mixture Model (GMM) and conduct a feature visualization analysis for M-SpecGene.

1. Experimental Analysis

Modality Bias: 1) Qualitative analysis: Fig. 1 shows two multi-object scenes with rich modality-invariant features (e.g., mid-level: edge, relative location, local descriptors; high-level: semantic information). M-SpecGene's t-SNE reveals significantly closer cross-modality feature clustering. 2) Quantitative analysis: Single-modality pretrained models in Tab. 1 underperform in multispectral tasks, as their pretrained parameters are suboptimal for dual-modality training. By fully leveraging the complementary characteristics, M-SpecGene encoder is driven to focus on learning modality-invariant representations, thereby effectively mitigating modality bias.

Methods	Models	Pretrain	mAP	mAP$_{50}$	mAP$_{75}$
InfMAE	ConViT-B	Infrared	39.7	76.6	35.6
MAE	ViT-B	RGB	43.0	82.8	37.8
M-SpecGene	ViT-B	RGB+IR	**44.7**	**84.8**	**40.1**

Table 1. Quantitative analysis of the modality bias.

Pre-trained Foundation Models Comparison: In Tab. 2, we supply InfMAE evaluated under identical fine-tuning protocols as M-SpecGene. InfMAE's inferior results stem from: 1) infrared-only pretraining and 2) limited data. Notably, M-SpecGene represents a pioneering exploration of multispectral foundation models, establishing a simple yet scalable baseline for this area.

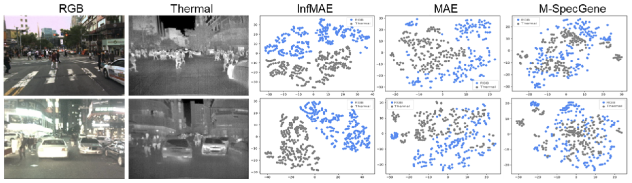

Figure 1. Qualitative analysis of the modality bias, which shows the t-SNE visualization of RGB and thermal features extracted by InfMAE, MAE and M-SpecGene.

	FLIR	LLVIP	SemanticRT	MVSeg
InfMAE	39.7	59.5	75.95	52.16
UniRGB-IR	44.1	63.2	75.21	56.46
M-SpecGene	**44.7**	**65.3**	**79.84**	**63.02**

Table 2. Comparison of pre-trained foundation models.

GMM-CMSS Masking Strategy: Survey [34] identifies four key directions for improving Masked Image Modeling (MIM): masking strategy, encoder, target, and head. We clarify that innovations in masking strategy are a key part of MIM framework development. Moreover, for the initial exploration of self-supervised learning on multispectral images, we believe that leveraging data characteristics effectively is more meaningful than introducing complex architectural changes. Therefore, we retain the core framework design, such as the standard ViT, for better generalizability and easier adaptation to other methods and tasks. While the proposed sampling strategy is specifically tailored to address the unique challenge of information imbalance.

The design of GMM-CMSS masking strategy is motivated by three key considerations: 1) Information density estimation: Given the uneven information distribution in RGBT datasets, we aim to progressively mask from high- to low-density regions. Thus, the CMSS is designed to serve as an objective metric for density measurement. 2) Sampling strategy based on CMSS: To balance high- and low-density areas and not just focus on specific regions, we adopt a Gaussian-based sampling function. 3) How to determine GMM parameters: The EM algorithm is employed to iteratively estimate $\{\mu_k, \Sigma_k, \pi_k\}$ for fitting the overall CMSS distribution. Notably, this adds minimal computational overhead. The prominent features of GMM-CMSS masking strategy are interpretability and lightweight design. Tab. 5 validates the effectiveness of GMM-CMSS by ablation studies across four datasets, especially in object-centric tasks such as detection.

	Random	GMM-CMSS
FLIR$_{\mathrm{mAP}\uparrow}$	83.80	**84.80**$\uparrow$ 1.0
KAIST$_{MR^{-2}\downarrow}$	26.33	**23.74**$\downarrow$ 2.59
MVSeg$_{\mathrm{mIoU}\uparrow}$	62.66	**63.06**$\uparrow$ 0.4
VI-RGBT1500$_{\mathrm{F}_\beta^{\max}\uparrow}$	0.864	**0.877**$\uparrow$ 0.013

Table 3. Ablation studies of the GMM-CMSS masking strategy.

Limitations: We highlight that M-SpecGene delivers impressive performance compared to complex customized models, particularly without any handcrafted modules. Yet its advantages are limited by: 1) Dataset constraints: Many benchmarks (e.g., FMB) suffer from small scale, low quality, limited diversity, and issues like coarse annotations and train-test overlap. 2) Modality imbalance: As seen in LLVIP, pedestrian detection relies heavily on the thermal modality. 3) Low task complexity: In the simpler salient object detection task, the performance gain is relatively less. Overall, M-SpecGene excels in large-scale, diverse, and modality-balanced multispectral datasets that demand strong generalization. However, the lack of high-quality RGBT datasets (e.g., ImageNet-scale benchmark) currently limits full evaluation of its potential.

2. Thermal480K Dataset

Most available RGBT downstream datasets contain only several thousand to tens of thousands of samples, which is insufficient to train a foundational model with robust generalization capabilities. To address this data bottleneck, we aim to maximize the utility of existing data: 1) Leveraging unimodal datasets: Since collecting aligned RGBT images pairs is more difficult, while RGB or thermal unimodal datasets are abundant. To this end, we collect a single-modality thermal dataset named Thermal480K, consisting of 486,024 thermal image samples. 2) RGBT multimodality datasets: We have consolidated approximately three million RGBT image samples (termed RGBT3M) from various datasets across diverse tasks and scenarios.

We believe that combining datasets from diverse scenarios enhances the heterogeneity and richness of the thermal data. Thus, we extensively collect unimodal thermal data through a comprehensive review of publicly available datasets in the fields of artificial intelligence and computer vision. After performing further post-processing to eliminate redundant and highly similar samples, we construct the Thermal480K dataset. As depicted in Fig. 3, the Thermal480K dataset comprises samples acquired by different imaging devices, covering a variety of resolutions, object types, distances, tasks, and fields of view. This ensures a robust and diverse collection of high-quality data, providing strong support for self-supervised learning. As shown in Tab. 4, these datasets cater to a variety of applications, including military operations, surveillance, industrial monitoring, denoising, and scientific research. They span a wide range of resolutions, from 32 × 64 to 1080 × 1920, and include diverse objects such as pedestrians, vehicles, trees, mountains, and animals. The datasets are captured using various infrared detectors, such as the FLIR Tau 320 and Infrec R500, among others.

3. RGBT550K Dataset

We further curate publicly available datasets containing paired RGBT image samples, which are summarized in Tab. 5. Based on the tasks targeted by these datasets, they can be broadly categorized into RGBT multispectral fusion and matching, object detection, semantic segmentation, saliency object detection, crowd counting, object tracking, and other applications. Among these, RGBT multispectral object detection is one of the most actively researched areas, with the largest number of open-source datasets. Object tracking datasets typically offer the highest volume of frame-level samples, often exceeding 100,000 frames. Thus we avoid allowing any single dataset to dominate an excessive proportion and temporal sampling is applied to RGBT video datasets to eliminate highly similar frames.

As shown in Fig. 4, the scenes captured in these datasets vary significantly depending on the intended task. Datasets for RGBT multispectral object detection and semantic segmentation primarily focus on outdoor environments, such as those encountered in drone monitoring, surveillance, and autonomous driving. In contrast, RGBT multispectral saliency object detection datasets include a mix of indoor and outdoor scenes, featuring more diverse target objects. Notably, most RGBT datasets encompass both daytime and nighttime conditions, leveraging the complementary characteristics of RGB and thermal modalities to enable more robust perception in complex environments.

To consolidate the samples from all the aforementioned

Num	Name	Year	Frames	Classes	Resolution	Sensor	Bit	Application
1	AAlart Data [68]	2018	771	pedestrian, vehicle	640×513	Catherine MP LWIR	8 HE	Mil., Surv.
2	AAU RainSnow [6]	2018	4.5K	vehicle	640×480	not specified	8*HE	Mil., Surv.
3	All-Ther [5]	2022	20K	vehicle, pedestrian	640×512 1280×1024	notspecified	8*HE	Mil., Surv.
4	ASL-TIR [45]	2014	4381	human, cat,horse	324×256	FLIRTau 320	8/16 HE/RAW	Mil., Surv.
5	Bird [3]	2022	302	bird	416×416	notspecified	8*HE	Sci.
6	BIRDSAI [8]	2020	160K	human, animal	640×480	FLIR Tamarisk640	8HE	Sci.
7	BU-TIV[63]	2014	35K 19K	motorcycle, runner, car, etc.	up to 1024×1024	FLIRSC8000	16RAW	Mil., Surv. Sci.
8	CAMEL [18]	2018	44.5K	biker,vehi cle, pedes trian	336×256	FLIRVuePro	8*HE	Mil., Surv.
9	CSIRCSIO [2]	2014	3650	vehicle, human, dog,bird	640×480	Uncool.μ-bol	8*HE	Mil., Surv.
10	CVC-09 [51]	2013	10K	pedestrian	640×480	notspecified	8HE	Mil., Surv.
11	Indoor-OutdoorIR[41]	2007	400	person, vehicle, etc.	384×288	Thermotek Miricle	8*HE	Mil., Surv.
12	InfAR [16]	2016	3.6M	person	293×25	GUIDIR IR300	8*HE	Mil., Surv. Sci.
13	LR-MR-HRFIR [49]	2020	3063	person, animal, vehicle, objects	up to 640×512	Axis, FLIR	8HE	Mil., Surv.
14	LSI [43]	2013	20K	pedestrian	32×64 164×128	IndigoOmega	14RAW	Mil., Surv.
15	LSOTBTIR [39]	2020	600K	animal, vehicle, aircraft, etc.	not specified	notspecified	8HE	Mil., Surv.
16	LTIR [7]	2015	11K	rhino, human, horse, etc.	up to 1920×480	FLIR	8/16 HE/RAW	Mil., Surv.
17	MSFocus [73]	2013	420	building, car, corridor, etc.	640×480	Canon FLIR	8HE	Industrial
18	Mov-Tar [55]	2021	150K	vegetation, building	640×512	notspecified	8HE	Mil., Surv.
19	PTB-TIR [38]	2019	30128	vehicle, pedestrian	up to 1280×720	8dif.cams.	8HE	Mil., Surv.
20	RGB-NIR [9]	2011	477	building, mountain, tree, etc.	1024×768	CanonT1i	8*HE	Sci.
21	RIFIR [40]	2014	20K	pedestrian	640×480	not specified	8*HE	Mil., Surv.
22	Roboflow-P [47]	2022	13K	person	640×512	notspecified	8*HE	Mil., Surv.
23	SCUT-FIR [65]	2019	211K	pedestrian	720×576	NV628	8HE	Mil., Surv.
24	SG-Ship [46]	2017	24K	ship	1080×1920	Canon70D	8*HE	Mil., Surv.
25	Soccer [15]	2018	3000	people	1920×480	AXIS Q1922	8HE	Mil., Surv.
26	TIDOC [4]	2021	6892	car, cat, pedestrian	300×400 1080×1440	FLIR SeekThermal	8*HE	Mil., Surv.
27	Transformer [42]	2021	255	transformer, induction motors	320×240	Dali-tech T4/T8	8*HE	Industrial
28	Valle-Aerial[17]	2020	110	road,car	336×256	Zenmuse XT	8*HE	Mil., Surv.
29	Anti-UAV410 [22]	2023	438K	small object	-	-	-	Mil.
30	IOD-Video [70]	2022	141K	gas	320×240	-	8HE	Industrial
31	TIVID [10]	2024	159K	road, car, pedestrian	320×240	-	8HE	Denoising

Table 4. To construct the Thermal480K dataset, we extensively collect unimodal thermal data through a comprehensive review of publicly available datasets.

datasets, we create the RGBT3M dataset. However, the RGBT3M dataset exhibits several limitations, including 1) imbalance across datasets, 2) temporal redundancy, and 3) low image quality. To address these issues, we rebalance the sample proportions across datasets, remove temporally redundant and low-quality samples, and conduct meticulous preprocessing. This process yields the RGBT550K dataset, a high-quality, large-scale, and diverse dataset encompassing a wide range of scenarios, tasks, lighting conditions, resolutions, and object categories.

4. Sample Filtering

Due to the presence of a certain proportion of low-quality samples across datasets, manually filtering the RGBT3M dataset, which comprises hundreds of thousands of samples, would be highly time-consuming. Therefore, we seek to automate the removal of low-quality samples using objective evaluation metrics. Among the various image fusion metrics we evaluate, Structural Similarity Index Measure (SSIM) demonstrate superior performance.

As illustrated in Fig. 5, samples with SSIM values below a certain threshold predominantly exhibit the following issues: 1) Low information density: For example, scenes consisting entirely of the sky without additional objects. 2) Extremely low-light conditions: RGB images appear completely black, rendering target objects indistinct. 3) Thermal scenes with limited dynamic range: In the absence of significant heat sources, thermal images lack sufficient contrast and clarity.

We determine that such low-quality samples could degrade the self-supervised learning process. Consequently,

	Number	Name	Year	Video/Image	Number of samples	Train	Test	Scene Type	Day/Night	Publication
Image Fusion and Matching	1	multispectraldata [13]	2021	video	53	-	-	Indoor/Outdoor	Both	ICRA
	2	CVC-15 [11]	2016	image	100	-	-	Urban Scenarios	Both	-
	3	CVC-lghd [1]	2015	image	44	-	-	Driving	Both	ICIP
	4	RoadScene [64]	2020	image	221	-	-	Driving	Both	AAAI
	5	MSRS [56]	2022	image	1444	-	-	Driving	Both	Inf. Fusion
	6	TNO [57]	2014	image	60	25	21\-	Military Scene	Both	Data Brief
	7	CATS [58]	2017	image	1400	-	-	Indoor/Outdoor	Both	CVPR
	8	M^{3}FD [35]	2022	image	4200	-	-	Overcast	Both	CVPR
Object Detection	9	VEDAI [48]	2016	image	1210	-	-	Aerial	Both	HAL
	10	KAIST [24]	2015	video	95k	7601	2251	Driving	Both	CVPR
	11	LLVIP [27]	2021	image	15448	12025	3463	Surv.	Night	ICCVW
	12	MFNet [21]	2021	image	1569	785	784	Driving	Both	IROS
	13	FLIR [14]	2018	image	10228	8862	1366	Driving	Both	NeurIPS
	14	CVC14 [19]	2016	image	8518	7085	1433	Pedestrian Scene	Both	MDPI
	15	MAVD [61]	2021	image	113282	-	-	Driving	Both	CVPR
	16	DroneVehicle [71]	2022	image	28439	17990	10339	Drone	Both	TCSVT
	17	SMOD [12]	2024	image	8676	-	-	Driving	Both	arXiv
	18	DVTOD [54]	2023	image	4358	-	-	Indoor/Outdoor	Both	TIV
	19	RGBT-Tiny [67]	2024	video	93K (115videos)	-	-	Indoor/Outdoor	Both	arXiv
Semantic Segmentation	20	PST900 [50]	2020	image	894	606	288	Subterranean	Both	arXiv
	21	Freiburg Thermal[62]	2020	image	20647	20583	64	Outdoor	Both	arXiv
	22	RoadScene-seg [66]	2022	image	1326	1105	221	Road Scene	Both	NeuroComp.
	23	SemanticRT [25]	2023	image	11371	6830	4541	Urban Scene	Both	ACM MM
	24	MVSeg [26]	2023	video	53k (738 videos)	452	286	Driving	Both	CVPR
	25	FMB [36]	2023	image	1500	-	-	Driving	Both	ICCV
Salient Object Detection	26	VT821 [29]	2017	image	821	-	-	Indoor/Outdoor	Both	IGTA
	27	VT1000 [59]	2020	image	1000	-	-	Indoor/Outdoor	Both	TMM
	28	VT5000 [60]	2022	image	5000	2500	2500	Indoor/Outdoor	Both	TMM
	29	VI-RGBT1500 [52]	2022	image	1500	-	-	Indoor/Outdoor	Both	TCSVT
	30	UAV RGB-T 2400 [53]	2023	image	2400	-	-	Drone	Both	TGRS
Crowd Counting	31	RGBT-CC [37]	2021	image	2030	1030	1000	Indoor/Outdoor	Both	CVPR
	32	DroneRGBT [44]	2020	image	3600	1807	1806	Drone	Both	ACCV
Object Tacking	33	VT-MOT [72]	2024	video	401068 (582 videos)	-	-	Surv., drone	Both	arXiv
	34	VTUAV [69]	2022	video	1700k (500 videos)	-	-	Tracking	Both	CVPR
	35	RGBT234 [31]	2019	video	116.7k (234 videos)	-	-	Tracking	Both	Pattern Rec.
	36	LasHeR [32]	2021	video	738.8k (1224 videos)	-	-	Tracking	Both	TIP
	37	GTOT [28]	2016	video	(50 videos)	-	-	Tracking	Both	TIP
	38	RGBT210 [30]	2017	video	(210 videos)	-	-	Tracking	Both	ACM MM
Others	39	VITLD [20]	2022	image	880	280	400	Indoor/Outdoor	Both	TII
	40	RGB-T-Glass [23]	2022	image	5551	4427	1124	Indoor/Outdoor	Both	IEEE
	41	IRVI [33]	2020	video	24352 (12 videos)	6	6	Monitoring	Both	ACM MM

Table 5. To construct the RGBT550K dataset, we curate publicly available datasets containing paired RGBT image samples, which cover a wide range of tasks.

as shown in Fig. 5, we exclude image pairs with SSIM values below 0.80, retaining only those with SSIM values above 0.80. These retained samples exhibit superior image clarity, object saliency, and texture features, providing higher-quality data for self-supervised pre-training.

5. More CMSS Visualization Samples

Given that thermal images lack the color and texture information inherent in RGB images, the existence of modality imbalance leads to an asymmetry in information density between RGB and thermal modalities. Additionally, un-

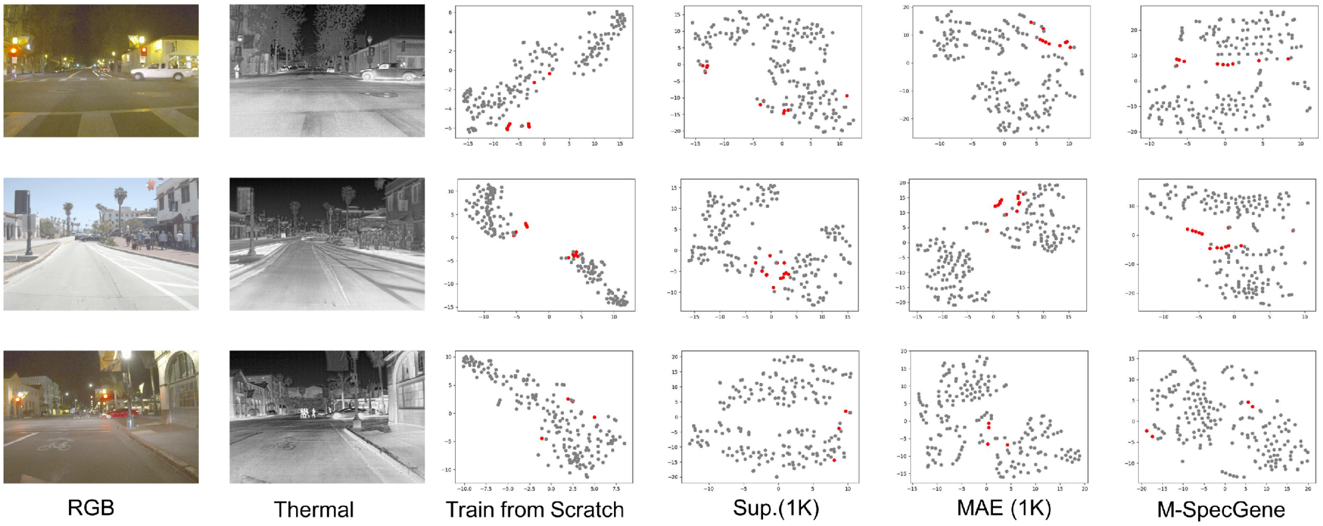

Figure 2. The t-SNE visualization of concatenated RGBT features for object and background regions from different pretrained models.

like ImageNet, the RGBT550K dataset is not object-centric. Consequently, the random sampling strategy employed by the MAE tends to allocate excessive attention to regions with low information density.

To address this, we adopt an information-aware sampling strategy, leveraging a simple yet effective Cross-Modality Structural Sparsity (CMSS) metric to evaluate the information density across RGBT modalities. In regions of high information density, RGBT patch embedding pairs exhibit lower similarity and greater structural variance, resulting in smaller CMSS values. Conversely, in regions of low information density, RGBT patch embedding pairs demonstrate higher similarity and lower structural variance, leading to larger CMSS values.

Due to space limitations in the main text, we present additional CMSS evaluation samples in Fig. 6. Our proposed CMSS metric generalizes well across diverse scenarios (e.g., drone imagery, surveillance, autonomous driving), varying lighting conditions (daytime and nighttime), and settings with modality imbalance. This robust generalization provides effective guidance for a progressive, easy-to-hard masking strategy.

6. GMM Estimation Visualization

We visualize the probability density functions fitted by the Gaussian Mixture Model (GMM) during the initial stages of self-supervised pre-training. As illustrated in Fig. 7, the parameters $\{\mu_k, \Sigma_k, \pi_k\}$ estimated by the GMM exhibit minimal variation from the 1st epoch to the 20th epoch, fluctuating only slightly within a narrow range. This observation indicates that the overall CMSS distribution reaches a relatively stable state early in the pre-training process, enabling the Gaussian Mixture Model to provide a consistent and optimal fit for $p(m)$.

7. Feature Visualization Analysis

We first concatenate the thermal and RGB features extracted by different pre-training models, including Train from Scratch, Sup.(1K), MAE (1K), and M-SpecGene. In Fig. 2, we present additional visual analyses of object and background features. It is evident that the object and background features derived from models initialized with random weights are highly entangled. In contrast, our M-SpecGene model demonstrates significantly greater discriminative capability between object and background features compared to the Train from Scratch, Sup. (1K), and MAE (1K) pre-training models. Statistical analyses conducted on the FLIR, KAIST, and LLVIP datasets further corroborate this observation. We attribute this improvement to the GMM-CMSS progressive sampling strategy, which promotes the learning of object-centric representations and facilitates the generation of more distinctive features.

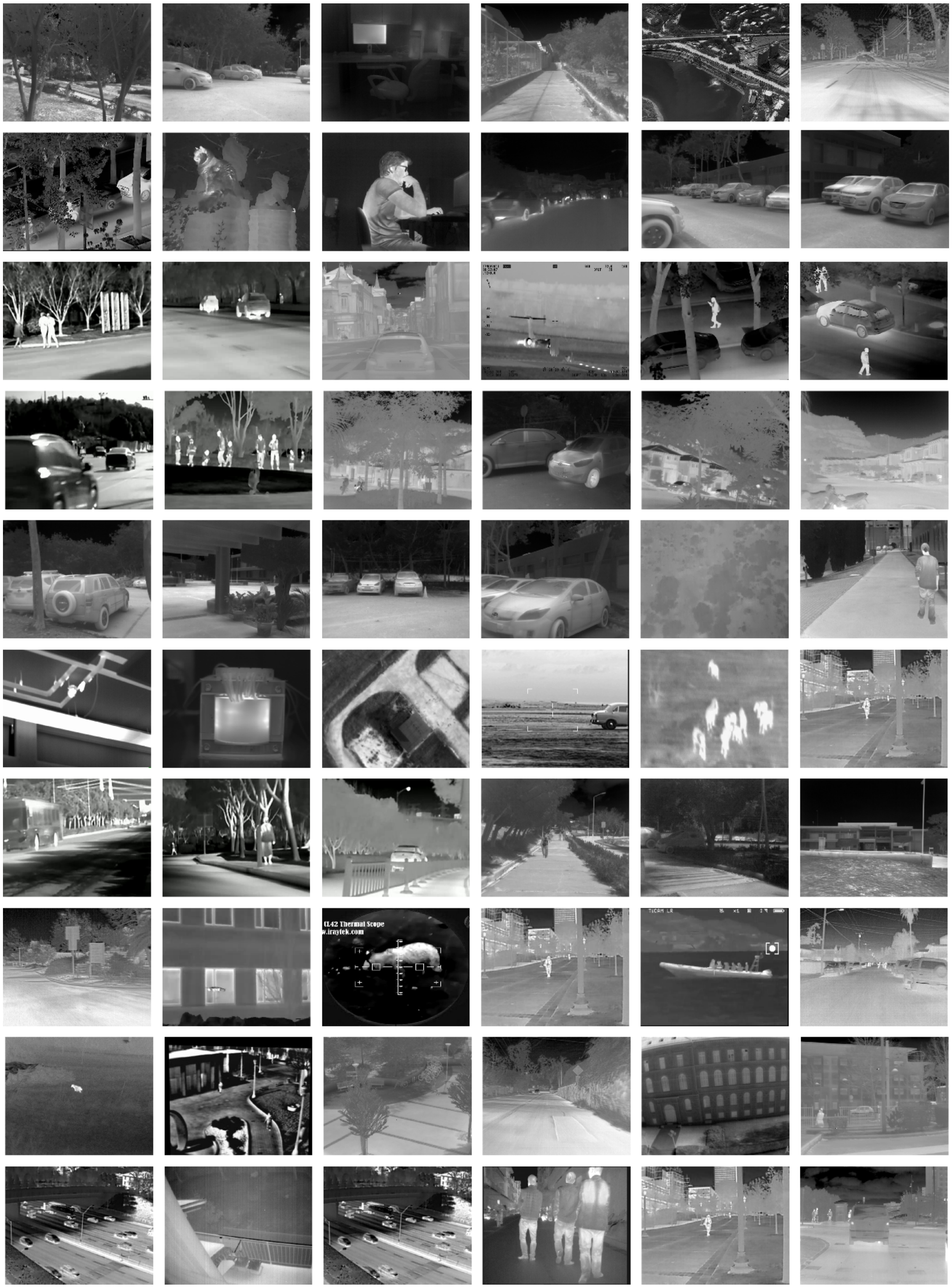

Figure 3. The visualization of the Thermal480K dataset, which covers a variety of resolutions, object types, distances, tasks, and fields of view.

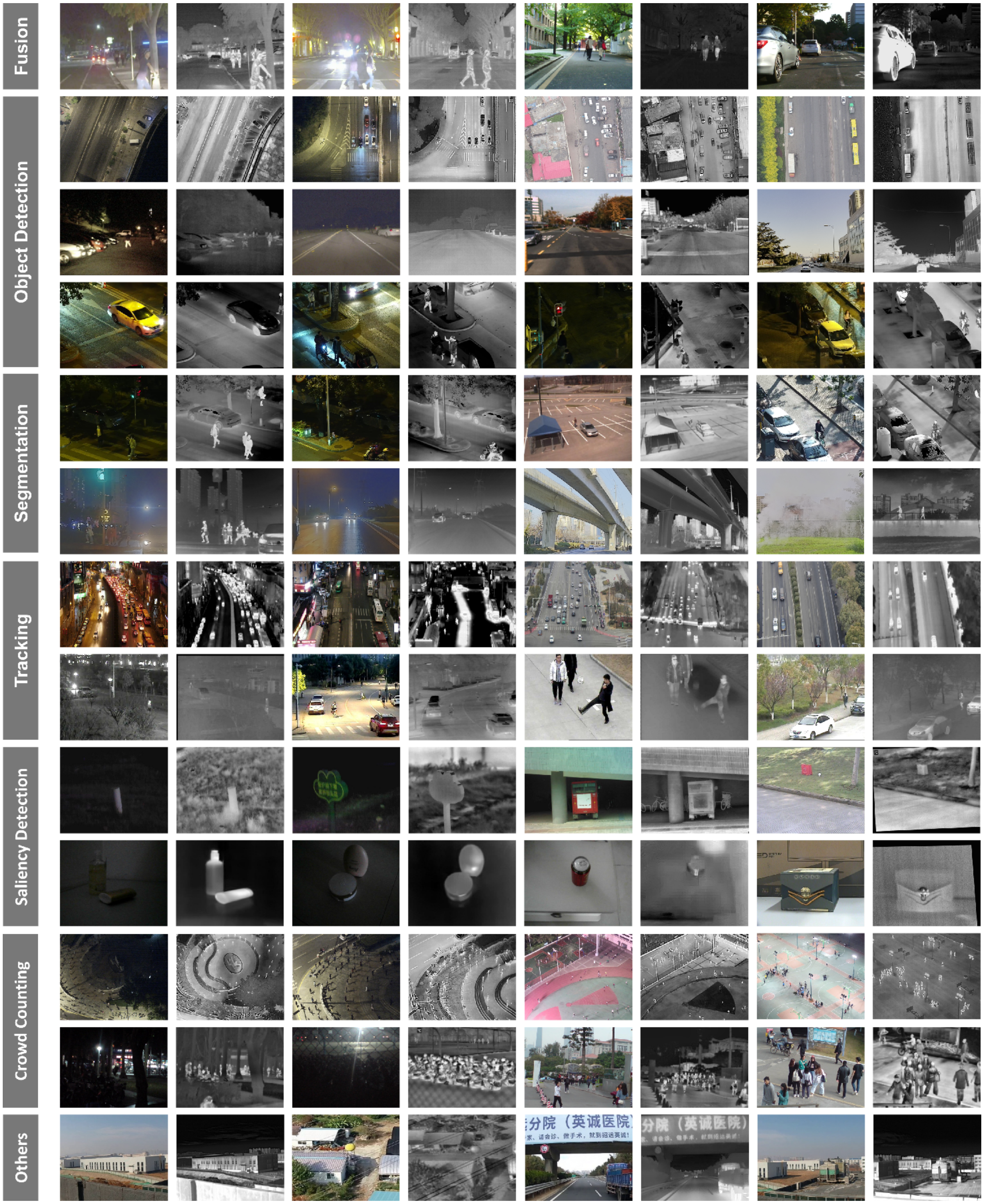

Figure 4. The visualization of the RGBT550K dataset, which contains samples from various RGBT multispectral downstream tasks.

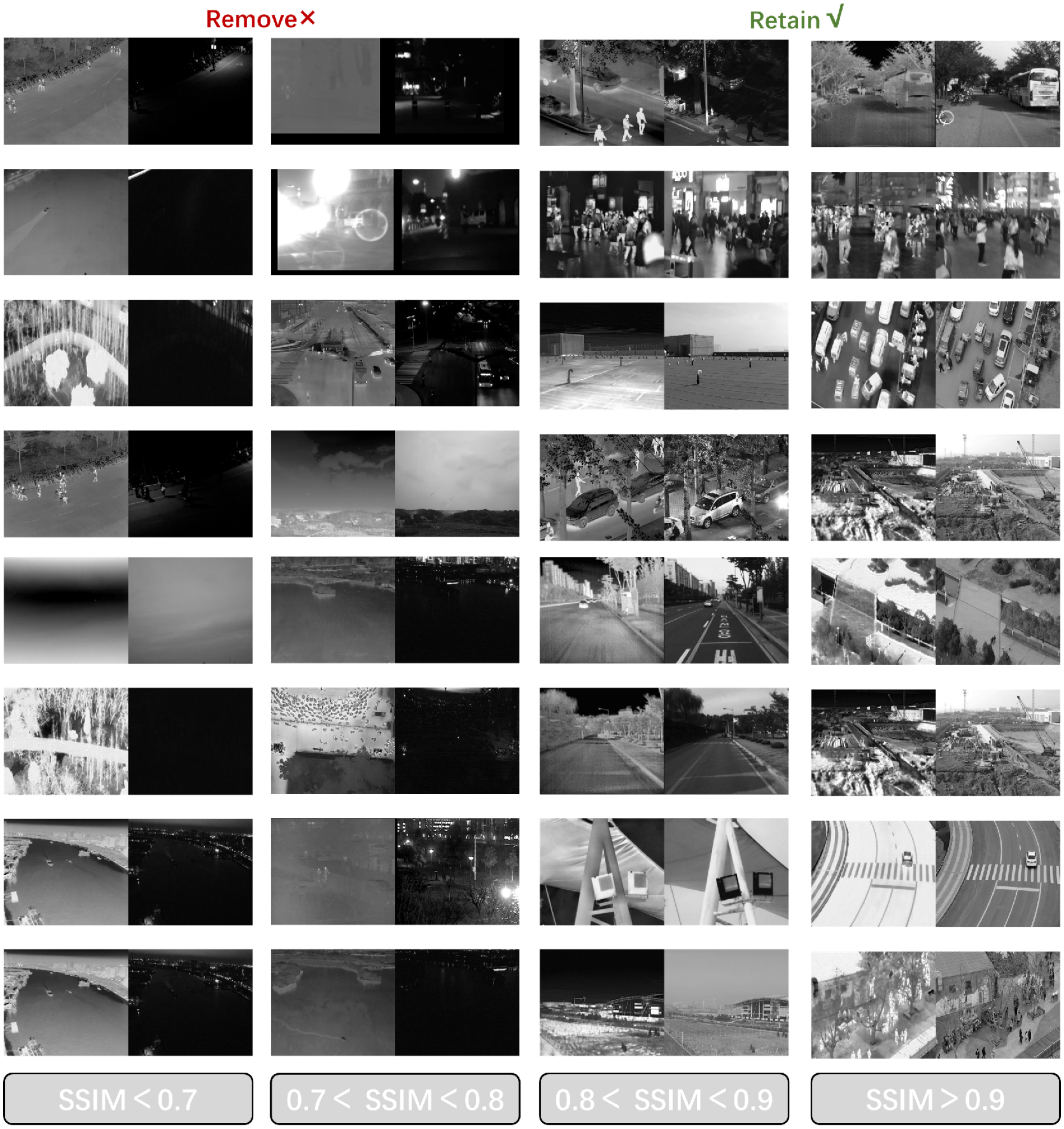

Figure 5. The RGB images are converted to grayscale to compute the SSIM with thermal images. Samples with SSIM values below a certain threshold predominantly exhibit the following issues: 1) Low information density. 2) Extremely low-light conditions. 3) Thermal scenes with limited dynamic range.

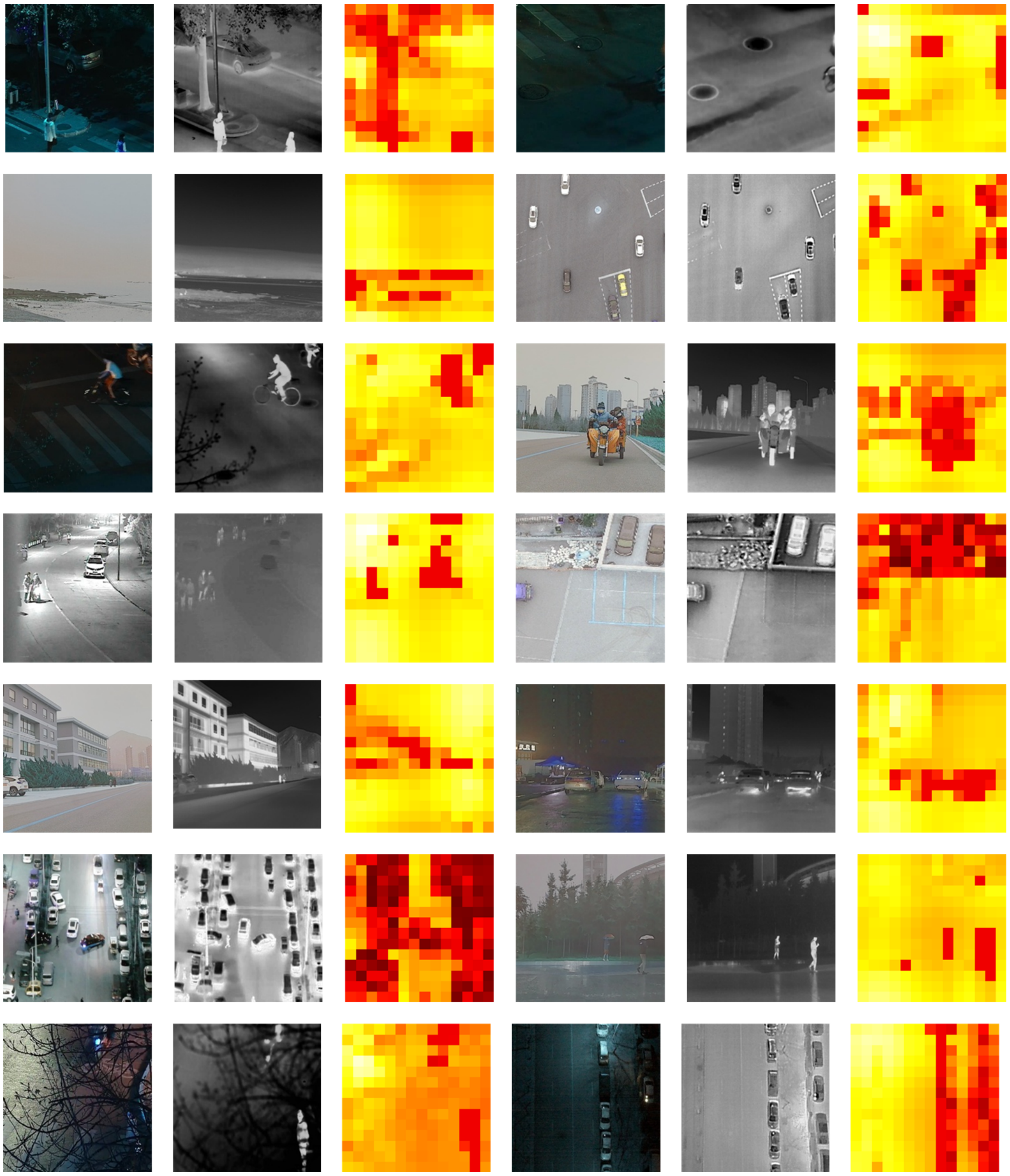

Figure 6. More visualization samples of the Cross-Modality Structural Sparsity (CMSS), which serves as a simple yet effective metric to measure the information density across RGB and thermal modalities.

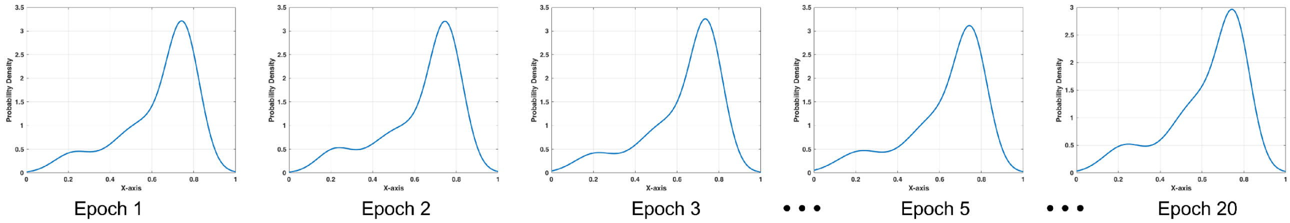

Figure 7. The visualization of the probability density functions fitted by the Gaussian Mixture Model (GMM) during the initial stages of self-supervised pre-training.

References

[1] Cristhian Aguilera, Angel D. Sappa, and Ricardo Toledo. Lghd: A feature descriptor for matching across non-linear intensity variations. In *Image Processing (ICIP), 2015 IEEE International Conference on*, page 5. IEEE, 2015. 4

[2] Aparna Akula, Nidhi Khanna, Ripul Ghosh, Satish Kumar, Amitava Das, and HK Sardana. Adaptive contour-based statistical background subtraction method for moving target detection in infrared video sequences. *Infrared Physics & Technology*, 63:103–109, 2014. 3

[3] AntiUAV 9 Aniket. bird dataset, 2022. visited on 2024-11-19. 3

[4] Qirat Ashfaq, Usman Akram, and Roshaan Zafar. Thermal image dataset for object classification, 2021. 3

[5] AVpublic. All_ther dataset. Roboflow, 2022. Accessed: 2024-11-20. 3

[6] Chris H Bahnsen and Thomas B Moeslund. Rain removal in traffic surveillance: Does it matter? *IEEE Transactions on Intelligent Transportation Systems*, 20(8):2802–2819, 2018. 3

[7] Amanda Berg, Jörgen Ahlberg, and Michael Felsberg. A thermal object tracking benchmark. In *2015 12th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6. IEEE, 2015. 3

[8] Elizabeth Bondi, Raghav Jain, Palash Aggrawal, Saket Anand, Robert Hannaford, Ashish Kapoor, Jim Piavis, Shital Shah, Lucas Joppa, Bistra Dilkina, et al. Birdsai: A dataset for detection and tracking in aerial thermal infrared videos. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1747–1756, 2020. 3

[9] Matthew Brown and Sabine Süsstrunk. Multi-spectral sift for scene category recognition. In *CVPR 2011*, pages 177–184. IEEE, 2011. 3

[10] Lijing Cai, Xiangyu Dong, Kailai Zhou, and Xun Cao. Exploring video denoising in thermal infrared imaging: Physics-inspired noise generator, dataset and model. *IEEE Transactions on Image Processing*, 2024. 3

[11] Computer Vision Center. Cvc-15 multimodal stereo dataset. http://adas.cvc.uab.es/elektra/enigma-portfolio/cvc-15-multimodal-stereo-dataset-2/, 2016. 4

[12] Zizhao Chen, Yeqiang Qian, Xiaoxiao Yang, Chunxiang Wang, and Ming Yang. Amfd: Distillation via adaptive multimodal fusion for multispectral pedestrian detection, 2024. 4

[13] Weichen Dai, Yu Zhang, Shenzhou Chen, Donglei Sun, and Da Kong. A multi-spectral dataset for evaluating motion estimation systems. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5560–5566. IEEE, 2021. 4

[14] Christian Dallago, Jody Mou, Kadina E Johnston, Bruce J Wittmann, Nicholas Bhattacharya, Samuel Goldman, Ali Madani, and Kevin K Yang. Flip: Benchmark tasks in fitness landscape inference for proteins. *bioRxiv*, pages 2021–11, 2021. 4

[15] Rikke Gade and Thomas B Moeslund. Constrained multi-target tracking for team sports activities. *IPSJ Transactions on Computer Vision and Applications*, 10:1–11, 2018. 3

[16] Chenqiang Gao, Yinhe Du, Jiang Liu, Jing Lv, Luyu Yang, Deyu Meng, and Alexander G Hauptmann. Infar dataset: Infrared action recognition at different times. *Neurocomputing*, 212:36–47, 2016. 3

[17] Lina García, Jean Díaz, Humberto Loaiza, and Andrés Restrepo. Thermal and visible aerial imagery, 2020. 3

[18] Evan Gebhardt and Marilyn Wolf. Camel dataset for visual and thermal infrared multiple object detection and tracking. In *2018 15th IEEE international conference on advanced video and signal based surveillance (AVSS)*, pages 1–6. IEEE, 2018. 3

[19] A. Gonzalez Alzate, Z. Fang, Y. Socarras, et al. Pedestrian detection at day/night time with visible and fir cameras: A comparison. *Sensors*, 16(6):820, 2016. 4

[20] Xiaodong Guo, Wujie Zhou, and Tong Liu. Multilevel attention imitation knowledge distillation for rgb-thermal transmission line detection. *Expert Systems with Applications*, 260:125406, 2025. 4

[21] 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. 4

[22] Bo Huang, Jianan Li, Junjie Chen, Gang Wang, Jian Zhao, and Tingfa Xu. Anti-uav410: A thermal infrared benchmark and customized scheme for tracking drones in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 3

[23] Dong Huo, Jian Wang, Yiming Qian, and Yee-Hong Yang. Glass segmentation with rgb-thermal image pairs. *IEEE Transactions on Image Processing*, 32:1911–1926, 2023. 4

[24] Soonmin Hwang, Jaesik Park, Namil Kim, Yukyung 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. 4

[25] 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. 4

[26] 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. 4

[27] 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. 4

[28] Chenglong Li, Hui Cheng, Shiyi Hu, Xiaobai Liu, Jin Tang, and Liang Lin. Learning collaborative sparse representation for grayscale-thermal tracking. *IEEE Transactions on Image Processing*, 25(12):5743–5756, 2016. 4

[29] Chenglong Li, Guizhao Wang, Yunpeng Ma, Aihua Zheng, Bin Luo, and Jin Tang. A unified rgb-t saliency detection benchmark: dataset, baselines, analysis and a novel approach. *arXiv preprint arXiv:1701.02829*, 2017. 4

[30] Chenglong Li, Nan Zhao, Yijuan Lu, Chengli Zhu, and Jin Tang. Weighted sparse representation regularized graph learning for rgb-t object tracking. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1856–1864, 2017. 4

[31] Chenglong Li, Xinyan Liang, Yijuan Lu, Nan Zhao, and Jin Tang. Rgb-t object tracking: Benchmark and baseline. *Pattern Recognition*, 96:106977, 2019. 4

[32] 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. 4

[33] Shuang Li, Bingfeng Han, Zhenjie Yu, Chi Harold Liu, Kai Chen, and Shuigen Wang. I2v-gan: Unpaired infrared-to-visible video translation. In *Proceedings of the 29th ACM international conference on multimedia*, pages 3061–3069, 2021. 4

[34] Siyuan Li, Luyuan Zhang, Zedong Wang, Di Wu, Lirong Wu, Zicheng Liu, Jun Xia, Cheng Tan, Yang Liu, Baigui Sun, et al. Masked modeling for self-supervised representation learning on vision and beyond. *arXiv preprint arXiv:2401.00897*, 2023. 1

[35] Jinyuan Liu, Xin Fan, Zhanbo Huang, Guanyao Wu, Risheng Liu, Wei Zhong, and Zhongxuan Luo. Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5802–5811, 2022. 4

[36] 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. 4

[37] Lingbo Liu, Jiaqi Chen, Hefeng Wu, Guanbin Li, Chenglong Li, and Liang Lin. Cross-modal collaborative representation learning and a large-scale rgbt benchmark for crowd counting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4823–4833, 2021. 4

[38] Qiao Liu, Zhenyu He, Xin Li, and Yuan Zheng. Ptb-tir: A thermal infrared pedestrian tracking benchmark. *IEEE Transactions on Multimedia*, 22(3):666–675, 2019. 3

[39] Qiao Liu, Xin Li, Zhenyu He, Chenglong Li, Jun Li, Zikun Zhou, Di Yuan, Jing Li, Kai Yang, Nana Fan, et al. Lsotb-tir: A large-scale high-diversity thermal infrared object tracking benchmark. In *Proceedings of the 28th ACM international conference on multimedia*, pages 3847–3856, 2020. 3

[40] Alina Dana Miron. *Multi-modal, multi-domain pedestrian detection and classification: proposals and explorations in visible over stereovision, fir and swir*. PhD thesis, INSA de Rouen; Universitatea Babeș-Bolyai (Cluj-Napoca, Roumanie), 2014. 3

[41] Nigel JW Morris, Shai Avidan, Wojciech Matusik, and Hanspeter Pfister. Statistics of infrared images. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–7. IEEE, 2007. 3

[42] Mohamad Najafi, Yasser Baleghi, and Seyyed Mehdi Mirimani. Thermal images dataset, transformer, 1 phase dry type, 2021. 3

[43] Daniel Olmeda, Cristiano Premebida, Urbano Nunes, Jose Maria Armingol, and Arturo de la Escalera. Pedestrian detection in far infrared images. *Integrated Computer-Aided Engineering*, 20(4):347–360, 2013. 3

[44] Tao Peng, Qing Li, and Pengfei Zhu. Rgb-t crowd counting from drone: A benchmark and mmccn network. In *Proceedings of the Asian conference on computer vision*, 2020. 4

[45] Jan Portmann, Simon Lynen, Margarita Chli, and Roland Siegwart. People detection and tracking from aerial thermal views. In *2014 IEEE international conference on robotics and automation (ICRA)*, pages 1794–1800. IEEE, 2014. 3

[46] Dilip K Prasad, Deepu Rajan, Lily Rachmawati, Eshan Rajabally, and Chai Quek. Video processing from electro-optical sensors for object detection and tracking in a maritime environment: A survey. *IEEE Transactions on Intelligent Transportation Systems*, 18(8):1993–2016, 2017. 3

[47] Roboflow Universe Projects. People detection - thermal dataset. `https://universe.roboflow.com/roboflow-universe-projects/people-detection-thermal`, 2022. visited on 2024-11-19. 3

[48] Sebastien Razakarivony and Frederic Jurie. Vehicle detection in aerial imagery (vedai): a benchmark. Technical report, Tech. Rep., 2015. 2. 4

[49] Rafael E Rivadeneira, Angel D Sappa, and Boris Xavier Vintimilla. Thermal image super-resolution: A novel architecture and dataset. In *VISIGRAPP (4: VISAPP)*, pages 111–119, 2020. 3

[50] 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. 4

[51] Yainuvis Socarrás, Sebastian Ramos, David Vázquez, Antonio M López, and Theo Gevers. Adapting pedestrian detection from synthetic to far infrared images. In *ICCV Workshops*, 2013. 3

[52] 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. 4

[53] Kechen Song, Hongwei Wen, Xiaotong Xue, Liming Huang, Yingying Ji, and Yunhui Yan. Modality registration and object search framework for uav-based unregistered rgb-t image salient object detection. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–15, 2023. 4

[54] Kechen Song, Xiaotong Xue, Hongwei Wen, Yingying Ji, Yunhui Yan, and Qinggang Meng. Misaligned visible-thermal object detection: A drone-based benchmark and baseline. *IEEE Transactions on Intelligent Vehicles*, 2024. 4

[55] X Sun, L Guo, W Zhang, Z Wang, Y Hou, Z Li, and X Teng. A dataset for small infrared moving target detection under clutter background. *Chin. Sci. Data*, 5(6):8, 2021. 3

[56] Linfeng Tang, Jiteng Yuan, Hao Zhang, Xingyu Jiang, and Jiayi Ma. Piafusion: A progressive infrared and visible image fusion network based on illumination aware. *Information Fusion*, 2022. 4

[57] Alexander Toet. The tno multiband image data collection. *Data in brief*, 15:249–251, 2017. 4

[58] Wayne Treible, Philip Saponaro, Scott Sorensen, Abhishek Kolagunda, Michael O'Neal, Brian Phelan, Kelly Sherbondy, and Chandra Kambhamettu. Cats: A color and thermal stereo benchmark. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2961–2969, 2017. 4

[59] 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. 4

[60] 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. 4

[61] Francisco Rivera Valverde, Juana Valeria Hurtado, and Abhinav Valada. There is more than meets the eye: Self-supervised multi-object detection and tracking with sound by distilling multimodal knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11612–11621, 2021. 4

[62] Johan Vertens, Jannik Zürn, and Wolfram Burgard. Heatnet: Bridging the day-night domain gap in semantic segmentation with thermal images. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 8461–8468. IEEE, 2020. 4

[63] Zheng Wu, Nathan Fuller, Diane Theriault, and Margrit Betke. A thermal infrared video benchmark for visual analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 201–208, 2014. 3

[64] Han Xu, Jiayi Ma, Zhuliang Le, Junjun Jiang, and Xiaojie Guo. Fusiondn: A unified densely connected network for image fusion. In *proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence*, 2020. 4

[65] Zhewei Xu, Jiajun Zhuang, Qiong Liu, Jingkai Zhou, and Shaowu Peng. Benchmarking a large-scale fir dataset for on-road pedestrian detection. *Infrared Physics & Technology*, 96:199–208, 2019. 3

[66] Shi Yi, Junjie Li, Xi Liu, and Xuesong Yuan. Ccaffmnet: Dual-spectral semantic segmentation network with channel-coordinate attention feature fusion module. *Neurocomputing*, 482:236–251, 2022. 4

[67] Xinyi Ying, Chao Xiao, Ruojing Li, Xu He, Boyang Li, Zhaoxu Li, Yingqian Wang, Mingyuan Hu, Qingyu Xu, Zaiping Lin, Miao Li, Shilin Zhou, Wei An, Weidong Sheng, and Li Liu. Visible-thermal tiny object detection: A benchmark dataset and baselines. *arXiv preprint arXiv:2406.14482*, 2024. 4

[68] Huaizhong Zhang, Chunbo Luo, Qi Wang, Matthew Kitchin, Andrew Parmley, Jesus Monge-Alvarez, and Pablo Casaseca-De-La-Higuera. A novel infrared video surveillance system using deep learning based techniques. *Multimedia tools and applications*, 77:26657–26676, 2018. 3

[69] Pengyu Zhang, Jie Zhao, Dong Wang, Huchuan Lu, and Xiang Ruan. Visible-thermal uav tracking: A large-scale benchmark and new baseline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8886–8895, 2022. 4

[70] Kailai Zhou, Yibo Wang, Tao Lv, Yunqian Li, Linsen Chen, Qiu Shen, and Xun Cao. Explore spatio-temporal aggregation for insubstantial object detection: benchmark dataset and baseline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3104–3115, 2022. 3

[71] Pengfei Zhu, Yiming Sun, Longyin Wen, Yu Feng, and Qinghua Hu. Drone based rgbt vehicle detection and counting: A challenge. *arXiv e-prints*, pages arXiv–2003, 2020. 4

[72] Yabin Zhu, Qianwu Wang, Chenglong Li, Jin Tang, and Zhixiang Huang. Visible-thermal multiple object tracking: Large-scale video dataset and progressive fusion approach, 2024. 4

[73] Martin Zukal, Jiri Mekyska, Petr Cika, and Zdenek Smekal. Interest points as a focus measure in multi-spectral imaging. *Radioengineering*, 22(1):68–81, 2013. 3