

A Satellite-Aerial Dataset Collection Process

Our data collection process involves the following steps: (1) deploying UAVs over selected areas to capture thermal image patches embedded with GPS metadata; (2) generating a unified thermal orthomosaic¹; (3) cropping and aligning corresponding regions from satellite imagery with a spatial resolution of 1m/pixel; (4) excluding regions with invalid thermal data; and (5) applying grid sampling to extract square image patches for model training and evaluation (following the setup in [54], we adopt a sampling stride of 35m and a crop size of 512×512). This process ensures the precise alignment between satellite RGB and thermal images.

B Dataset Preprocessing Details

For large-scale training, all datasets are converted into the WebDataset² format, except for the satellite-aerial datasets. These datasets consist of paired, spatially aligned thermal and satellite RGB maps, from which we randomly sample regions across the entire map. We also detail the specific preprocessing steps applied to each dataset to ensure reproducibility:

- **NII-CU**: Resized RGB images to match thermal image dimensions.
- **TARDAL**: Created an 80%/20% train/test split by random sampling since no official split was provided in the dataset.
- **Freiburg**: Removed the black padding from thermal images on both right and left sides.

C Additional Training Details

We conduct all training and evaluation using a single NVIDIA A100 or H100 GPU. For training the thermal encoder and decoder, we employ a batch size of 16 and use the AdamW [32] optimizer with a learning rate of 6×10^{-5} and a weight decay of 1×10^{-3} , over a total of 200k training steps. All other configurations follow the default settings of the Latent Diffusion Model [37]. In training the flow-based generative models, we use a batch size of 64 with AdamW optimizer at a learning rate of 1×10^{-4} and no weight decay, for 200k training steps.

D Dataset Parameters Comparison

A comparison between the Boson-night dataset [54] and our datasets is provided in Table 5, highlighting our contributions in terms of broader geographic coverage, increased diversity of thermal sensors, and the inclusion of both daytime and nighttime imagery.

Table 5: **Comparison of dataset parameters between Boson-night [54] and our datasets.** The differences are highlighted regarding area coverage, satellite map sources, thermal sensors used, collection years, and the number of surveyed regions.

Dataset Name	Area	Satellite Map	Thermal Sensor	Collection Year	Number of Regions
Boson-night [54]	33km ²	Bing	Boson	2021	1
DJI-day (Ours)	29km ²	ESRI ³	DJI Zenmuse H20T	2023	1
BosonPlus-day (Ours)	85km ²	ESRI ³	BosonPlus	2024	3
BosonPlus-night (Ours)	94km ²	ESRI ³	BosonPlus	2024	3

E Thermal Auto-encoder Training Details and Reconstruction Performance

This section details the training procedure and reconstruction performance of the thermal encoder E_T and decoder D_T . Both components are trained using the KL-VAE framework [37] on our curated RGB-T dataset collection, optimized with a combination of reconstruction losses (L1 and LPIPS [57])

¹We used Agisoft Metashape to generate the orthomosaic: <https://www.agisoft.com/>

²<https://github.com/webdataset/webdataset>

³ESRI satellite map: <https://www.esri.com/en-us/home>

Table 6: Reconstruction FID and PSNR performance across multiple RGB-T datasets									
Method	FLIR	LLVIP	AVIID	MSRS	NII-CU	M ³ FD	BosonPlus-day	BosonPlus-night	Boson-night
FID ↓									
klvae w/o GAN loss	18.51	2.94	18.16	14.37	9.07	5.50	16.63	6.48	4.24
klvae w/ GAN loss	14.66	3.14	12.21	12.20	8.67	5.33	6.73	3.52	2.00
PSNR ↑									
klvae w/o GAN loss	30.10	37.63	31.29	38.73	45.73	40.12	30.46	41.32	43.53
klvae w/ GAN loss	28.74	36.59	30.07	37.60	45.24	39.03	29.15	40.84	43.16

795 and KL regularization on the latent space. After 300 epochs (200k steps) of training, we optionally
 796 fine-tune the final decoder layers for 75 epochs using a GAN loss plus the loss mentioned above,
 797 as in the original LDM [37], to enhance FID scores without altering the latent representation, but it
 798 will slightly negatively affect PSNR metrics. We note that the results reported in the paper use the
 799 decoder **without** GAN loss.