Caltech Aerial RGB-Thermal Dataset in the Wild Supplementary Material

Connor Lee*, Matthew Anderson*, Nikhil Raganathan, Xingxing Zuo,
Kevin Do, Georgia Gkioxari, and Soon-Jo Chung

California Institute of Technology
{clee, matta, nrangana, zuox, kdo, georgia, sjchung}@caltech.edu

S1 Comparison of Related Benchmarks and Datasets

A detailed comparison of related datasets and benchmarks is shown in Tab. S1. Included datasets must be captured from aerial vehicles, surface vehicles (boats), or cars. We included pure RGB datasets only if they are related to semantic segmentation or object detection and captured from aerial vehicles. Other datasets included for comparison must contain thermal imagery and be related to semantic segmentation or object detection.

Table S1: Comparison of datasets captured from aerial platforms or depicting thermal scenes.

Dataset	Platform	Task	RGB	Thermal	GPS	IMU	Setting	Location	# Samples	Camera pose
Ours	UAV	Sem. Seg.	✓	✓	✓	✓	River, Coast, Lake, Mountain, Desert	USA	4,195	20°, 45°, < 120 m
AeroScapes [26], UDD [4], UAVid [23], VDD [3], IDD [3]	UAV	Sem. Seg.	✓	✗	✗	✗	Urban	China	205 - 3269	30 - 90°, < 100 m
Semantic Drone Dataset [25]	UAV	Sem. Seg.	✓	✓	✗	✓	Urban	Germany	400	90°, < 30m
Swiss and Okutama Drone Datasets [37]	UAV	Sem. Seg.	✓	✗	✗	✗	Urban	Switzerland, Japan	191	90°
NII-CU Multispectral Aerial Person Detection [37]	UAV	Person Det.	✓	✓	✗	✗	Urban	Japan	5,880	45°
WIT-UAS [16]	UAV	Obj. Det.	✗	✓	✓	✓	Forest	USA	6,951	30°, 90°, < 120 m
VisDrone [47]	UAV	Obj. Det.	✓	✗	✗	✗	Urban	China	10,209	Variable
MassMIND [27]	USV	Sem. Seg.	✗	✓	✓	✓	Harbor	USA	2,900	0°, 0 m
Flir ADAS [7]	Car	Obj. Det.	✓	✓	✗	✗	Urban	USA	10,228	0°, 0 m
BIRDSAI [1]	UAV	Obj. Track.	✗	✓	✗	✗	Savannas	S. Africa	62,000	Off-nadir, 60-120 m
HIT-UAV [38]	UAV	Obj. Det.	✓	✓	✗	✗	Urban	China	2,898	30°-90°, 60-130 m
KAIST Multispectral [5]	Car	Obj. Det.	✓	✓	✓	✓	Urban	S. Korea	4750	0°, 0 m
MFNet [10]	Car	Sem. Seg.	✓	✓	✗	✗	Urban	Japan	1,569	0°, 0 m
M^3FD [20]	Car	Obj. Det.	✓	✓	✗	✗	Urban	China	4,200	0°, 0 m
Freiburg Thermal [40]	Car	Sem. Seg.	✓	✓	✗	✗	Urban	Germany	20,656‡	0°, 0 m
SODA [18], SCUT-Seg [44]	Car	Sem. Seg.	✗	✓	✗	✗	Indoor/Urban	—	∼2,000	0°, 0 m
STheReO [45], MS^2 [35]	Car	SLAM / Depth Est.	✓	✓	✓	✓	Urban/Suburban	S. Korea	—	0°, 0 m
PST900 [36]	UGV	Sem. Seg.	✓	✓	✗	✗	Subterranean	USA	894	0°, 0 m
LLVIP [15]	Building (fixed)	Pedestrian Detection	✓	✓	✗	✗	Urban	China	14,588	Variable

† Camera angle of 90°is nadir-pointing ‡ Annotations provided for only 64 test set images.

* These authors contributed equally to this work.

S2 Dataset Information

Our dataset consists of 37 aerial and ground trajectories captured from diverse natural landscapes across the USA. Further details are shown in Tab. S2. Visualizations of several flight trajectories are shown in Fig. S1.

Table S2: Dataset capture locations and settings.

Location of Capture	Terrain Type	Capture Method	Motion	Time of Day	# Seq.	Total Time
Kentucky River, KY	River	UAV Flight	Large	Afternoon	3	17m10s
Colorado River, CA	River	UAV Flight	Large	Sunrise	4	32m50s
Castaic Lake, CA	Lake	UAV Flight	Large	Midday	4	58m56s
Duck, NC	Coast	UAV Flight	Large	Day/Night	7	91m26s
Big Bear Lake, CA	Lake	Ground	Minor	Mid-morning	8	15m52s
Arroyo Seco, CA	Stream	Ground	Still	Afternoon	8	10m22s
Idyllwild, CA	Mountain	Ground	Minor	Day	2	5m58s
Joshua Tree, CA	Desert	Ground	Minor	Day	2	5m58s
North Field (Caltech), CA	Urban	UAV Flight	Large	Day	5	26m13s

S2.1 Data Capture

All data is stored as ROS1 rosbags as this is a natural format for robotics work. As the rosbag format may not be preferred for all users, extraction tools for generating csv and images files are provided at *link_provided_after_review*. Within the rosbags, the synchronized data contains timestamps for both the time of trigger (*sync/rate_**) and time the data was received (header stamp in the topic), allowing the data to be aligned in post-processing regardless of transport delays in the system. Where available, position data should be taken from the UAV (*uav1/mavros/local_position/** topics) and orientation data from the VN100 (*imu/imu*) to provide the best available estimate. Finally, while mostly complete, not all datasets contain all sensors due to issues when collecting data.

S2.2 Sensor Calibration

We calibrate the three cameras and IMU via Kalibr [8] by conducting three independent calibrations. We do this to isolate any difficulties caused by the thermal camera. We first calibrate the thermal camera and IMU, and then perform stereo calibration for each possible camera pairing. Thermal calibrations are done using a 10×10 circle grid (1″ diameter). To ensure calibration image sharpness in the thermal domain, we place the calibration board in direct sunlight for 2-3 minutes prior to data collection and keep it illuminated (reflecting sunlight) during the entire collection process. The RGB and Mono cameras were calibrated using a standard 6x6 April tag grid board.

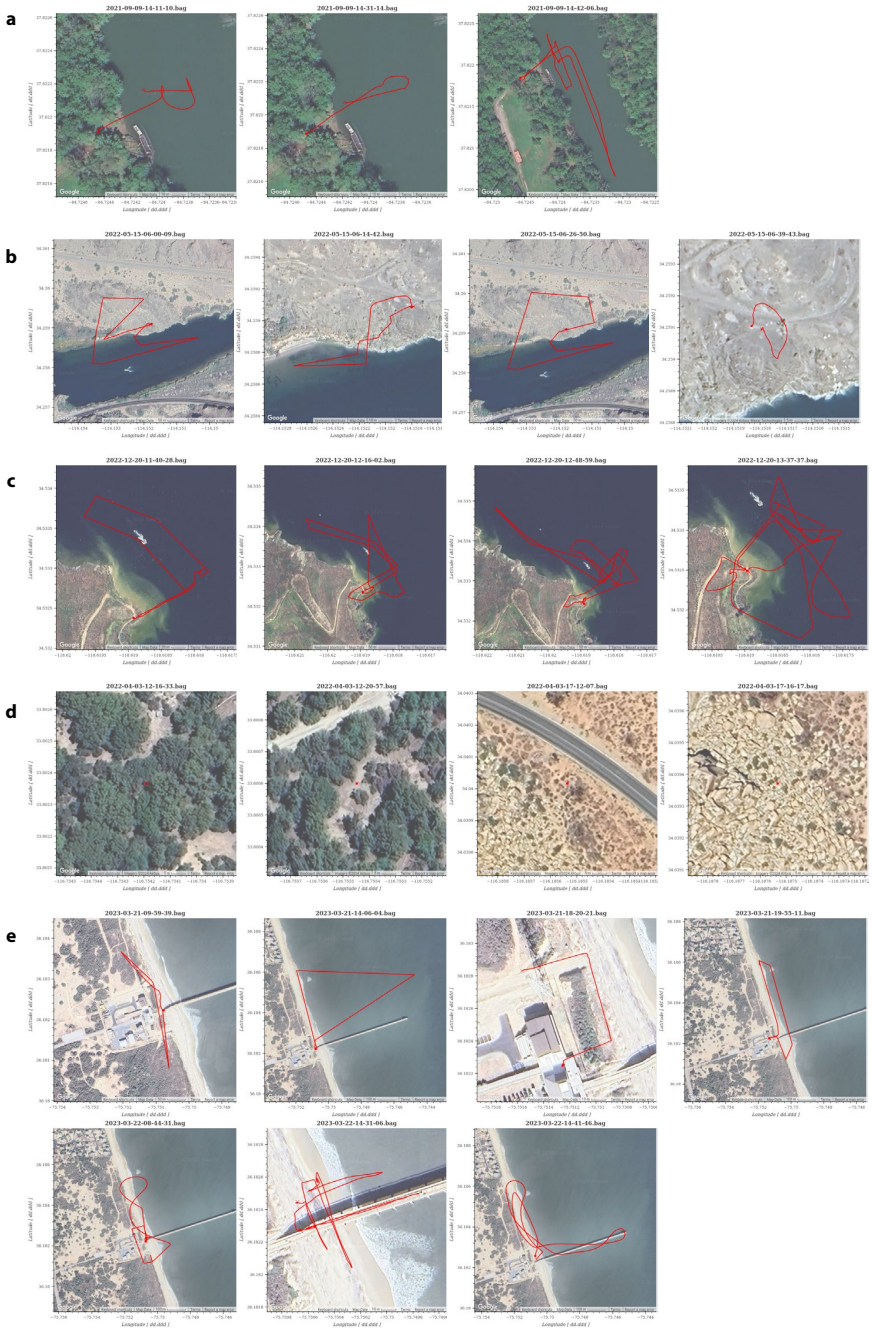

Fig. S1: Trajectories from our dataset: **(a)** Kentucky River **(b)** Colorado River **(c)** Castaic Lake **(d)** Idyllwild and Joshua Tree **(e)** Duck. North Field trajectories not shown.

S2.3 Labeling Process for Semantic Segmentation Annotations

The thermal images were labeled via manual annotation. Annotators were asked to read a guide with descriptions of each trajectory. Descriptions contained lists of possible classes, a expert-labeled examples provided by the authors, and GPS coordinates (for look-up in Google Earth). During labeling, annotators were presented with a side-by-side copy of a thermal image and its corresponding RGB image, the trajectory it came from, and were asked to annotate the thermal image. The annotations were reviewed by authors, with rejected annotations sent back for re-annotation. A total of 3 rounds of annotation were conducted.

S3 Implementation and Training Details

Our baselines were implemented using code from public Github repositories shown in Tab. S3. Official code was used whenever possible. We trained and tested all networks using a single Nvidia A6000 ADA GPU (48 GB).

Table S3: Public repositories used in this work to create our baselines

Baselines	Public Repositories
Thermal Segmentation	
FastSCNN [31]	`Tramac/Fast-SCNN-pytorch`
EfficientViT [2]	`mit-han-lab/efficientvit`
Segformer [42]	`NVlabs/SegFormer`
DINOv2 [28]	`facebookresearch/dinov2`
FTNet [29]	`shreyaskamathkm/FTNet`
Everything else [11, 12, 22, 24, 39, 43]	`huggingface/pytorch-image-models` + `qubvel/segmentation_models.pytorch`
RGB-T Segmentation	
EAEFNet [19]	`FreeformRobotics/EAEFNet`
CRM [34]	`UkcheolShin/CRM_RGBTSeg`
CMNeXt [46]	`jamycheung/DELIVER`
RGB-T Image Translation	
UNIT [21]	`NVlabs/imaginaire`
MUNIT [13]	`NVlabs/imaginaire`
Edge-guided RGB-T [17]	`RPM-Robotics-Lab/sRGB-TIR`
Pix2Pix [14]	`junyanz/pytorch-CycleGAN-and-pix2pix`
Pix2PixHD [41]	`NVIDIA/pix2pixHD`
VQ-GAN [6]	`CompVis/taming-transformers`
Palette [33]	`Janspiry/Palette-Image-to-Image-Diffusion-Models`
VIO/SLAM	
VINS-Fusion [32]	`HKUST-Aerial-Robotics/VINS-Fusion`
OpenVINS [9]	`rpng/open_vins`

S3.1 Thermal Baselines

Most networks used in our thermal baseline experiments (Tab. 2) employed a DeepLabV3+ segmentation head with the exception of FastSCNN [31], EfficientViT [2], Segformer [42], DINOv2, and FTNet. For DINOv2, we implemented the linear and nonlinear multi-scale segmentation heads ourselves, with their architectures shown in Tab. S4. All networks, besides DINOv2 and ConvNeXt (CLIP), were trained starting from pretrained ImageNet weights.

Table S4: Multi-scale segmentation head architecture used with DINOv2. The architecture is described using PyTorch syntax [30].

DINOv2 Linear Head	DINOv2 Nonlinear Head
`nn.Conv2d(384*4, 10, 1)`	`nn.Conv2d(384*4, 512, 3, padding=1)`
	`nn.GELU`
	`nn.Conv2d(512, 10, 1)`

Networks were trained for 300 epochs or until validation loss plateued. We used the Adam optimizer with a `1e-3` learning rate that decayed at an exponential rate of `0.99` and a batch size of 64. We augmented our 16-bit thermal training data using random contrast stretches (within the lower and upper 5^{th} percentiles) and CLAHE with a random clip limit (see Sec. 3.2). Photometric augmentations were followed by horizontal flips, rotations (within 10°), scaling between 1 and 1.5, and random crops to 512×512.

S3.2 RGB-T Segmentation, Image Translation, and VIO/SLAM

We used the repositories listed in Tab. S3 and followed their training procedures. However, we increased the batch size to maximize GPU memory usage whenever possible.

S4 Additional Results

S4.1 Relative Thermal Pixel Intensities Throughout the Day

In order to better understand how different objects and entities appear in thermal imagery at different times of day, we plot the thermal pixel intensities of each class (using the ground truth annotations) as a function of their recorded capture time (Fig. S2). Since our thermal camera was not radiometric, we plotted the relative thermal pixel values post-normalization using the normalization scheme described in Sec. 3.2. The plots show cases of thermal inversion between certain classes, notably *bare ground* and *water*.

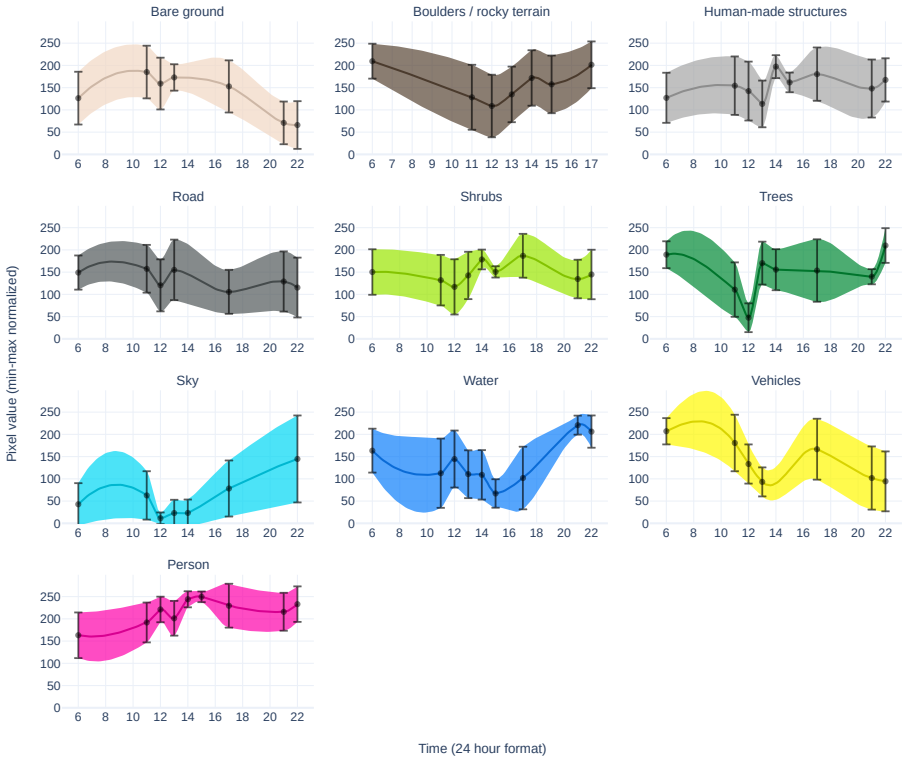

Fig. S2: Change in normalized pixel values per class throughout the day.

S5 Examples

S5.1 Thermal Inversion

Thermal inversion occurs when object pixel values change depending on its temperature. One notable example of this is thermal inversion of water and land as seen in Fig. S3. In this example, the two images were taken over the same location (Duck, NC) but at different times of day.

References

1. Bondi, E., Jain, R., Aggrawal, P., Anand, S., Hannaford, R., Kapoor, A., Piavis, J., Shah, S., Joppa, L., Dilkina, B., 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. pp. 1747–1756 (2020)
2. Cai, H., Gan, C., Han, S.: Efficientvit: Enhanced linear attention for high-resolution low-computation visual recognition. arXiv preprint arXiv:2205.14756 (2022)
3. Cai, W., Jin, K., Hou, J., Guo, C., Wu, L., Yang, W.: Vdd: Varied drone dataset for semantic segmentation. arXiv preprint arXiv:2305.13608 (2023)

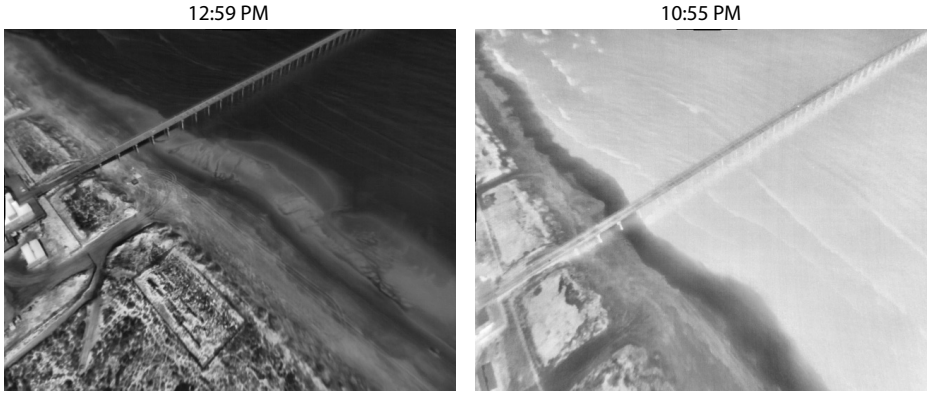

Fig. S3: Example of thermal inversion of water and land classes at Duck, NC.

4. Chen, Y., Wang, Y., Lu, P., Chen, Y., Wang, G.: Large-scale structure from motion with semantic constraints of aerial images. In: Pattern Recognition and Computer Vision: First Chinese Conference, PRCV 2018, Guangzhou, China, November 23-26, 2018, Proceedings, Part I 1. pp. 347–359. Springer (2018)
5. Choi, Y., Kim, N., Hwang, S., Park, K., Yoon, J.S., An, K., Kweon, I.S.: Kaist multi-spectral day/night data set for autonomous and assisted driving. IEEE Transactions on Intelligent Transportation Systems **19**(3), 934–948 (2018)
6. Esser, P., Rombach, R., Ommer, B.: Taming transformers for high-resolution image synthesis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12873–12883 (2021)
7. Teledyne flir adas dataset, `https://www.flir.com/oem/adas/adas-dataset-form/`, Last accessed on 2023-10-27
8. Furgale, P., Rehder, J., Siegwart, R.: Unified temporal and spatial calibration for multi-sensor systems. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (2013)
9. Geneva, P., Eckenhoff, K., Lee, W., Yang, Y., Huang, G.: Openvins: A research platform for visual-inertial estimation. In: 2020 IEEE International Conference on Robotics and Automation (ICRA). pp. 4666–4672. IEEE (2020)
10. Ha, Q., Watanabe, K., Karasawa, T., Ushiku, Y., Harada, T.: 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). pp. 5108–5115. IEEE (2017)
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
12. Howard, A., Sandler, M., Chu, G., Chen, L.C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., et al.: Searching for mobilenetv3. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1314–1324 (2019)
13. Huang, X., Liu, M.Y., Belongie, S., Kautz, J.: Multimodal unsupervised image-to-image translation. In: European Conference on Computer Vision (ECCV) (2018)

14. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1125–1134 (2017)
15. Jia, X., Zhu, C., Li, M., Tang, W., Zhou, W.: Llvip: A visible-infrared paired dataset for low-light vision. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3496–3504 (2021)
16. Jong, A., Yu, M., Dhrafani, D., Kailas, S., Moon, B., Sycara, K., Scherer, S.: Wit-uas: A wildland-fire infrared thermal dataset to detect crew assets from aerial views. In: 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 11464–11471. IEEE (2023)
17. Lee, D.G., Jeon, M.H., Cho, Y., Kim, A.: Edge-guided multi-domain rgb-to-tir image translation for training vision tasks with challenging labels. In: 2023 IEEE International Conference on Robotics and Automation (ICRA). pp. 8291–8298. IEEE (2023)
18. Li, C., Xia, W., Yan, Y., Luo, B., Tang, J.: Segmenting objects in day and night: Edge-conditioned cnn for thermal image semantic segmentation. IEEE Transactions on Neural Networks and Learning Systems **32**(7), 3069–3082 (2020)
19. Liang, M., Hu, J., Bao, C., Feng, H., Deng, F., Lam, T.L.: Explicit attention-enhanced fusion for rgb-thermal perception tasks. IEEE Robotics and Automation Letters (2023)
20. Liu, J., Fan, X., Huang, Z., Wu, G., Liu, R., Zhong, W., Luo, Z.: 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. pp. 5802–5811 (2022)
21. Liu, M.Y., Breuel, T., Kautz, J.: Unsupervised image-to-image translation networks. In: Conference on Neural Information Processing Systems (NeurIPS) (2017)
22. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11976–11986 (2022)
23. Lyu, Y., Vosselman, G., Xia, G.S., Yilmaz, A., Yang, M.Y.: Uavid: A semantic segmentation dataset for uav imagery. ISPRS Journal of Photogrammetry and Remote Sensing **165**, 108 – 119 (2020). https://doi.org/https://doi.org/10.1016/j.isprsjprs.2020.05.009, http://www.sciencedirect.com/science/article/pii/S0924271620301295
24. Mehta, S., Rastegari, M.: Separable self-attention for mobile vision transformers. arXiv preprint arXiv:2206.02680 (2022)
25. Mostegel, C., Maurer, M., Heran, N., Pestana Puerta, J., Fraundorfer, F.: Semantic drone dataset (Jan 2019), http://dronedataset.icg.tugraz.at/, Last accessed on 2023-10-27
26. Nigam, I., Huang, C., Ramanan, D.: Ensemble knowledge transfer for semantic segmentation. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 1499–1508. IEEE (2018)
27. Nirgudkar, S., DeFilippo, M., Sacarny, M., Benjamin, M., Robinette, P.: Massmind: Massachusetts maritime infrared dataset. The International Journal of Robotics Research **42**(1-2), 21–32 (2023)
28. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Howes, R., Huang, P.Y., Xu, H., Sharma, V., Li, S.W., Galuba, W., Rabbat, M., Assran, M., Ballas, N., Synnaeve, G., Misra, I., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: Dinov2: Learning robust visual features without supervision (2023)

29. Panetta, K., Shreyas Kamath, K.M., Rajeev, S., Agaian, S.S.: Ftnet: Feature transverse network for thermal image semantic segmentation. IEEE Access **9**, 145212–145227 (2021). https://doi.org/10.1109/ACCESS.2021.3123066
30. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems **32** (2019)
31. Poudel, R., Liwicki, S., Cipolla, R.: Fast-scnn: Fast semantic segmentation network. In: Sidorov, K., Hicks, Y. (eds.) Proceedings of the British Machine Vision Conference (BMVC). pp. 187.1–187.12. BMVA Press (September 2019). https://doi.org/10.5244/C.33.187, https://dx.doi.org/10.5244/C.33.187
32. Qin, T., Li, P., Shen, S.: Vins-mono: A robust and versatile monocular visual-inertial state estimator. IEEE Transactions on Robotics **34**(4), 1004–1020 (2018)
33. Saharia, C., Chan, W., Chang, H., Lee, C., Ho, J., Salimans, T., Fleet, D., Norouzi, M.: Palette: Image-to-image diffusion models. In: ACM SIGGRAPH 2022 Conference Proceedings. pp. 1–10 (2022)
34. Shin, U., Lee, K., Kweon, I.S.: Complementary random masking for rgb-thermal semantic segmentation. In: IEEE International Conference on Robotics and Automation (2024)
35. Shin, U., Park, J., Kweon, I.S.: Deep depth estimation from thermal image. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1043–1053 (2023)
36. Shivakumar, S.S., Rodrigues, N., Zhou, A., Miller, I.D., Kumar, V., Taylor, C.J.: Pst900: Rgb-thermal calibration, dataset and segmentation network. In: 2020 IEEE international conference on robotics and automation (ICRA). pp. 9441–9447. IEEE (2020)
37. Speth, S., Goncalves, A., Rigault, B., Suzuki, S., Bouazizi, M., Matsuo, Y., Prendinger, H.: Deep learning with rgb and thermal images onboard a drone for monitoring operations. Journal of Field Robotics **39**(6), 840–868 (2022)
38. Suo, J., Wang, T., Zhang, X., Chen, H., Zhou, W., Shi, W.: Hit-uav: A high-altitude infrared thermal dataset for unmanned aerial vehicle-based object detection. Scientific Data **10**, 227 (2023)
39. Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: International conference on machine learning. pp. 6105–6114. PMLR (2019)
40. Vertens, J., Zürn, J., Burgard, W.: 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). pp. 8461–8468. IEEE (2020)
41. Wang, T.C., Liu, M.Y., Zhu, J.Y., Tao, A., Kautz, J., Catanzaro, B.: High-resolution image synthesis and semantic manipulation with conditional gans. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2018)
42. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: Segformer: Simple and efficient design for semantic segmentation with transformers. In: Neural Information Processing Systems (NeurIPS) (2021)
43. Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1492–1500 (2017)
44. Xiong, H., Cai, W., Liu, Q.: Mcnet: Multi-level correction network for thermal image semantic segmentation of nighttime driving scene. Infrared Physics & Technol-

ogy p. 103628 (2021). https://doi.org/https://doi.org/10.1016/j.infrared.2020.103628
45. Yun, S., Jung, M., Kim, J., Jung, S., Cho, Y., Jeon, M.H., Kim, G., Kim, A.: Sthereo: Stereo thermal dataset for research in odometry and mapping. In: 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 3857–3864. IEEE (2022)
46. Zhang, J., Liu, R., Shi, H., Yang, K., Reiß, S., Peng, K., Fu, H., Wang, K., Stiefelhagen, R.: Delivering arbitrary-modal semantic segmentation. In: CVPR (2023)
47. Zhu, P., Wen, L., Du, D., Bian, X., Fan, H., Hu, Q., Ling, H.: Detection and tracking meet drones challenge. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(11), 7380–7399 (2021)