Long-range UAV Thermal Geo-localization with Satellite Imagery

Jiuhong Xiao[1], Daniel Tortei[2], Eloy Roura[2], and Giuseppe Loianno[1]

***Abstract*— Onboard sensors, particularly cameras and thermal sensors, are gaining traction as alternatives to GPS in UAV navigation due to GPS's vulnerability to signal disruptions. This paper introduces a pioneering thermal geo-localization framework using satellite RGB imagery. By incorporating multiple domain adaptation strategies, we tackle the scarcity of paired thermal and satellite images. Experimental results on real UAV data affirm the robustness of our method, even in challenging thermal imaging conditions. We also unveil the *Boson-nighttime* dataset and associated code. This is, to our understanding, the first endeavor in thermal geo-localization using satellite RGB imagery for long-range UAV flights. Code and dataset: https://github.com/arplaboratory/satellite-thermal-geo-localization**

I. Introduction

Unmanned Aerial Vehicle (UAV) long-range flight navigation heavily relies on geo-localization methods such as the GPS, which is susceptible to signal loss and spoofing [1]. Alternative techniques, like Visual Inertial Odometry (VIO), offer onboard localization especially in GPS-compromised environments but encounter drift errors in long-range flights without loop closure detection [2]. Visual Geo-localization (VG) using satellite RGB imagery provides drift-free localization for long-range flights, matching UAV and satellite images. Our study focuses on VG in low-light, high-altitude UAV flights utilizing thermal imagery, addressing the challenges posed by limited satellite-thermal data pairs and the discrepancies between thermal and satellite RGB features.

VG with satellite imagery is considered an absolute self-localization method [1] in UAV localization research. For traditional image matching methods, direct alignment methods [3]–[5] densely match UAV images and satellite images with the highest pixel similarity. Image registration methods [6], [7] employ hand-crafted feature point descriptors to match feature points between satellite and UAV images. Recent works [8]–[12] also introduce deep neural networks. Specifically, previous works such as [9] use conditional generative adversarial nets [13] to synthesize a UAV-view image with a satellite image style. In [8], the authors combine visual odometry and a cross-view geo-localization module to predict the location of UAVs with a Kalman filter. Our proposed method dedicates to cross-domain geo-localization between satellite RGB and thermal images.

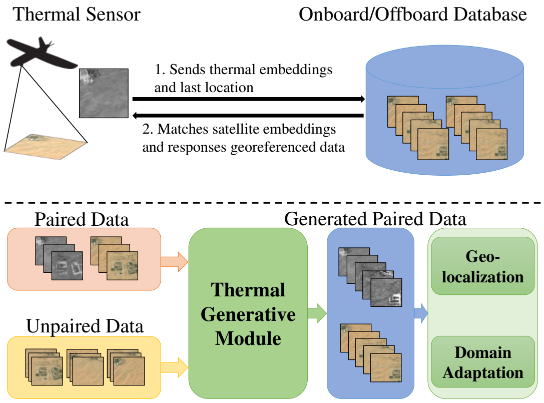
Thermal Sensor
Onboard/Offboard Database
1. Sends thermal embeddings and last location
2. Matches satellite embeddings and responses georeferenced data
Paired Data
Generated Paired Data
Unpaired Data
Thermal Generative Module
Geo-localization
Domain Adaptation

Fig. 1: Thermal geo-localization with satellite imagery. **Upper part:** Our method aims to perform geo-localization with a thermal sensor attached to a UAV in a low-illumination environment with a satellite image database. **Lower part:** The training of the geo-localization model leverages the thermal generative module to generate fake thermal images and the domain adaptation method.

Thermal Geo-localization (TG) is vital for UAV long-range nighttime flights, yet remains underexplored compared to VG with satellite imagery. Relevant works for UAV thermal localization [14]–[17] study the localization and navigation performance of Thermal-Inertial Odometry (TIO). UAV thermal stereo odometry is proposed in [14] to navigate short-range daytime and nighttime flights. The authors in [16] use top-down thermal camera settings to investigate localization performance at various times of the night. These works demonstrate that TIO performs on par with daytime VIO in short-range indoor or outdoor flights. Our proposed approach focuses on geo-localization by satellite-thermal matching for long-range high-altitude flights.

In this paper, we introduce a groundbreaking learning-based thermal geo-localization technique for long-range high-altitude UAV flights using satellite RGB imagery (Fig. 1). By integrating two domain adaptation methods [18], [19], we overcome challenges from limited thermal data. Our method's effectiveness is demonstrated on the **Boson-nighttime** dataset, particularly in areas with self-similar thermal features. Importantly, we publicly share our code and dataset, marking a novel contribution to long-range flights' geo-localization using combined satellite RGB and thermal imagery.

[1] The authors are with the New York University, Tandon School of Engineering, Brooklyn, NY 11201, USA. email: {jx1190, loiannog}@nyu.edu.

[2] The authors are with the Autonomous Robotics Research Center-Technology Innovation Institute, Abu Dhabi, UAE. email: {daniel.tortei, eloy.roura}.@tii.ae.

This work was supported by the Technology Innovation Institute, the NSF CAREER Award 2145277, the NSF CPS Grant CNS-2121391, Qualcomm Research, Nokia, and NYU Wireless. Giuseppe Loianno serves as consultant for the Technology Innovation Institute. This arrangement has been reviewed and approved by the New York University in accordance with its policy on objectivity in research.

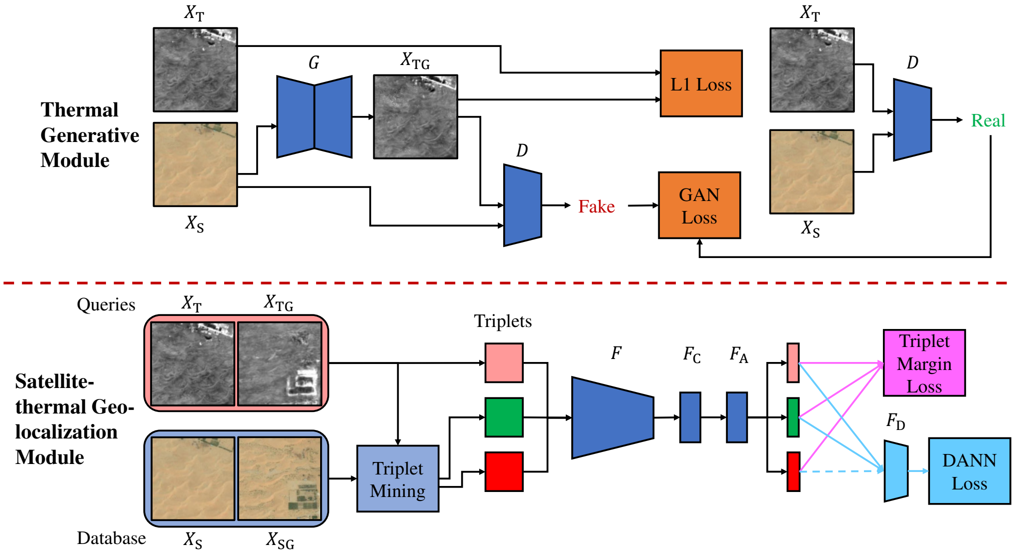

Fig. 2: The proposed framework of thermal geo-localization. Thermal Generative Module (TGM) is optimized by L1 Loss and least square GAN Loss. After training, the module generates fake thermal images X_{TG} from unpaired satellite images X_{SG}. Satellite-thermal Geo-localization Module (SGM) finds a positive sample (green) and a negative sample (red) relative to the query sample (pink). The triplets input the feature extractor F, the compression layer F_{C}, and the aggregation module F_{A}, and the outputs are 1-D embeddings. The Triplet Margin Loss (magenta) and DANN Loss (cyan) are optimized for geo-localization models F, F_{C}, and F_{A}. The dashed line for F_{D} means the negative sample is optional for DANN loss.

II. Methodology

The proposed TG framework, depicted in Fig. 2, has two main components: a Thermal Generative Module (TGM) and a Satellite-thermal Geo-localization Module (SGM). In this section, we describe them in detail.

A. Thermal Generative Module (TGM)

For the Thermal Generative Module (TGM), we leverage the pix2pix [19] model. Let $X_{\mathrm{S}}, X_{\mathrm{T}}$ denote the normalized satellite images and thermal images with height H, width W, and the number of channels C. G is the generator to generate a fake thermal image $X_{\mathrm{TG}} = G(X_{\mathrm{S}})$. D is the conditional discriminator where the first input is X_{T} or X_{TG} and the second input is X_{S}, and the output predicts if the input thermal image is a fake thermal image. The objectives of the module without the random noise are

$$\begin{aligned}\min_D L_{\mathrm{GAN}}(D) = &\frac{1}{2}\mathbb{E}_{X_{\mathrm{T}}}[(D(X_{\mathrm{T}})-b)^2] \\ &+\frac{1}{2}\mathbb{E}_{X_{\mathrm{S}}}[(D(G(X_{\mathrm{S}}))-a)^2],\end{aligned} \tag{1}$$

$$\min_G L_{\mathrm{GAN}}(G) = \mathbb{E}_{X_{\mathrm{S}}}[(D(G(X_{\mathrm{S}}))-c)^2], \tag{2}$$

$$\min_G L_1(G) = \mathbb{E}_{X_{\mathrm{S}},X_{\mathrm{T}}}[|G(X_{\mathrm{S}})-X_{\mathrm{T}}|], \tag{3}$$

where L_{GAN} is the least square GAN loss [20], L_1 is the L1 loss, a, b, c are the fake label, the real label and the label that G wants D to classify for fake data, respectively. The total loss L_{TGM} is the weighted sum of L_{GAN} and L_1 as

$$L_{\mathrm{TGM}}(G,D) = L_{\mathrm{GAN}}(G) + L_{\mathrm{GAN}}(D) + \lambda_1 L_1(G), \tag{4}$$

where λ_1 is the weight of L_1 loss and is set to 100.0. We train the G to minimize $L_{\mathrm{GAN}}(G)$ and $L_1(G)$, and train the D to minimize $L_{\mathrm{GAN}}(D)$.

B. Satellite-thermal Geo-localization Module (SGM)

The proposed Satellite-thermal Geo-localization Module (SGM) leverages the image retrieval workflow from deep-VG-benchmark [21]. Additionally, we improve it with a compression layer F_{C} to control the dimensionality of output descriptors, a module to train simultaneously with the paired dataset and the generated dataset, and a DANN loss [18] branch for domain adaption. $X_{\mathrm{S}}, X_{\mathrm{T}}$ are a pair of satellite and thermal images. X_{SG} denotes satellite images without paired thermal images and X_{TG} denotes the generated thermal images from X_{SG} using the trained generator G. During training, the framework samples thermal images from the thermal queries dataset. Triplet mining searches for the positive and the negative satellite database samples that have the lowest L_2 distance to the queries in the feature (embedding) space. The satellite database is built by tiling the satellite map with a certain stride. Each triplet consists of one query thermal image, one positive, and one negative satellite image. We input the triplets of $H \times W \times C$ (size and the number of channels) to the feature extractor F, and the output is feature maps $H/16 \times W/16 \times C'$ where C' is the number of channels for the feature map. The compression layer F_{C} consists of a 2-D convolution layer mapping from $H/16 \times W/16 \times C'$ to $H/16 \times W/16 \times C_{\mathrm{target}}$ with kernel size = 1 and a 2-D batch normalization layer. C_{target} is target channel number we want to control. The aggregation module

TABLE I: The results of different settings of CE, DANN loss, and Generated Dataset. The **bold** value shows the best result and the underline value shows the second-best result. The arrows with metrics show the direction of good values.

Backbone	CE	DANN	DANN Only Positive	Generated Dataset	$R@1\uparrow$	$R@5\uparrow$	$R_{512}@1\uparrow$	$R_{512}@5\uparrow$	$L_2^{512}(m)\downarrow$
ResNet-18					$\underline{69.3}$	$\underline{81.8}$	81.1	92.0	$\underline{58.9}$
		✓			61.1	74.5	75.5	87.7	77.8
		✓	✓		67.4	80.2	81.4	92.1	59.8
	✓				68.0	80.5	$\underline{82.1}$	$\underline{92.2}$	$\underline{58.9}$
	✓	✓			60.1	74.5	76.1	88.9	75.8
	✓	✓	✓		**72.1**	**83.9**	**84.5**	**93.7**	**52.2**
ResNet-18				✓	84.8	90.8	$\underline{95.5}$	$\underline{98.7}$	$\underline{19.0}$
		✓		✓	68.9	77.0	82.3	89.5	65.9
		✓	✓	✓	82.0	88.8	93.7	97.9	24.8
	✓			✓	$\underline{87.0}$	$\underline{91.8}$	94.2	97.6	24.2
	✓	✓		✓	81.2	88.7	90.3	95.7	34.6
	✓	✓	✓	✓	**92.1**	**96.9**	**96.5**	**99.1**	**14.7**

F_{A} uses NetVLAD [22] to map the compressed feature maps $H/16 \times W/16 \times C_{\text{target}}$ to 1-D embeddings $1 \times C_{\text{final}}$ where C_{final} is the final output dimension. Triplet margin loss L_{T} is used as

$$L_{\text{T}}(q,p,n) = (\|q-p\|_2 - \|q-n\|_2 + m)^+, \tag{5}$$

where q, p, n are the embeddings ($1 \times C_{\text{final}}$) of query, positive and negative samples. m is a positive scalar margin and is set to 0.1. L_{T} aims to decrease the L_2 embedding distance of q and p and increase that of q and n.

We also introduce DANN loss [18] into SGM for domain adaptation. F_{D} is a domain classifier and the goal is to classify q as the thermal label d and p, n as the satellite label e. A gradient reversal layer is attached to the beginning of F_{D}, and the DANN loss is a cross-entropy loss

$$\begin{aligned} L_{\text{DANN}}(q,p,n) = -\sum_{c\in\{d,e\}} [&y_{c,q}\log(o_{c,q}) \\ &+ y_{c,p}\log(o_{c,p}) + y_{c,n}\log(o_{c,n})], \end{aligned} \tag{6}$$

where $y_{d,q} = 1, y_{e,q} = 0$ for thermal embeddings, $y_{d,p} = 0, y_{e,p} = 1, y_{d,n} = 0, y_{e,n} = 1$ for satellite embeddings, $o_{c,q}, o_{c,p}, o_{c,n}$ are the output probability of q, p, n for the class c using F_{D}. The reversed gradients backpropagate to make the distribution of q, p, n similar. In the experiments, we find that using negative embeddings n for DANN loss affects the performance since DANN loss may conflict with triplet margin loss on q, n. DANN loss optimizes to increase the similarity between the distribution of q, n while triplet margin loss increases $\|q-n\|_2$. We take n as an optional input for F_{D}. The total loss L_{SGM} is the weighted sum of L_{T} and L_{DANN} as

$$L_{\text{SGM}} = L_{\text{T}} + \lambda_2 L_{\text{DANN}}, \tag{7}$$

where λ_2 is the weight of DANN loss and is set to 0.1.

III. Experimental Setup

In this section, we introduce the experimental setup of our proposed framework.

A. Datasets

In order to collect thermal aerial data, we used FLIR's Boson thermal imager (8.7 mm focal length, 640p resolution, and 50° horizontal field of view)[1]. The collected images are nadir at approx. 1m/px spatial resolution. We performed six flights from 9:00 PM to 4:00 AM and label this dataset as **Boson-nighttime**, accordingly. To create a single map, we first run a structure-from-motion (SfM) algorithm to reconstruct the thermal map from multiple views. Subsequently, orthorectification is performed by aligning the photometric satellite maps with thermal maps at the same spatial resolution. The ground area covered by Boson-nighttime measures 33 km^2 in total. The most prevalent map feature is the desert, with small portions of farms, roads, and buildings (Fig. 3).

The Bing satellite map[2] is cropped in the corresponding area as our satellite reference map. We tile the thermal map into 512×512 px thermal image crops with a stride of 35 px. Each thermal image crop pairs with the corresponding satellite image crop. Areas covered by three flights of Boson-nighttime are used for training and validation. The remaining areas, covered by the other three flights are used for testing. The train/validation/test splits for Boson-nighttime are 10256/13011/26568 pairs of satellite and thermal image crops, respectively. The generated dataset has 79950 pairs of satellite and generated thermal images.

B. Metrics

To evaluate the thermal geo-localization performance, we use the following metrics

- Recall@N ($R@N$): It measures the percentage of the query images from which the top-N retrieved database images are within 50 meters.
- Recall@N with prior location threshold d ($R_d@N$): It is $R@N$ with the search region limited to a radius of d meters from the queries. We show the results of $R_{512}@1$ and $R_{512}@5$.
- L_2 distance error with prior location threshold d (L_2^d): It measures the L_2 distance (meter) from the queries

[1] https://www.flir.es/products/boson/

[2] Bing satellite imagery is sourced from Maxar: https://www.bing.com/maps/aerial

to the estimated position from top-1 retrieved database images within a radius of d meters from the queries. We show the results of L_2^{512}.

IV. Results

The results of our proposed framework are shown in Tables I and Fig. 3 - 4.

A. CE and DANN Analysis

In the upper part of Table I, We compare the models without generated dataset and find that the models with CE have higher $R_{512}@1$ and $R_{512}@5$ and lower or equal L_2^{512} than those without CE. This reveals that using enhanced thermal images may boost geo-localization performances for low-contrast thermal features. *DANN only positive* means no negative samples are considered in the DANN loss. We look into the effectiveness of DANN loss and the necessity to remove negatives from DANN loss. We find that DANN loss with negatives always lowers the recall performance and increases L_2 distance error, which practically supports our assumption that DANN loss can conflict with Triplet margin loss in Section II-B. We also discover that DANN loss without negative samples typically works with enhanced thermal images. The model with CE and DANN (only positive) shows the best geo-localization performance among the models.

B. Generated Dataset Analysis

In the lower part of Table I, we observe that models with the generated dataset notably improve the recall performance and L_2 error. Our Best Model (ResNet-18 with CE, DANN only positive, and the generated dataset) exhibits accurate localization results, which significantly outperforms Baseline Model (ResNet-18). These findings highlight the ability of synthesized datasets to improve model performance when paired data is limited. Overall, these results provide compelling evidence for the effectiveness of our approach to thermal geo-localization.

In Fig. 3, we show examples of satellite, ground truth, and generated thermal images with and without CE. The visualized results show that the thermal features of the farm and building portion are mostly clear and consistent in the generated results, while those of roads and deserts are distorted. The results with CE show detectable textures on that portion. We recognize that the existence of clear textures impacts the performance of the geo-localization model.

C. Visualized Geo-localization Results

We compare the visualized SGM results of the Baseline Model and Our Best Model in Fig. 4. Our framework can retrieve accurate results on both farm regions and desert regions with the indistinct self-similar thermal feature, as shown in the examples. However, the baseline setting shows more errors than our best setting. We identify two types of failure cases: Offset error and localization failure. Offset error (The 1th and 2th columns of the 2nd row) results in the substantial offset ($\geq$ 50 m) of the retrieved satellite image from the true one. Localization failure (The 3th and 4th columns of the 2nd row) results in completely wrong retrieved results, which may cause a localization system failure. The Baseline Model suffers from limited paired data and self-similar low-contrast thermal features and makes more offset errors and localization failures, while Our Best Model mitigates the above problems.

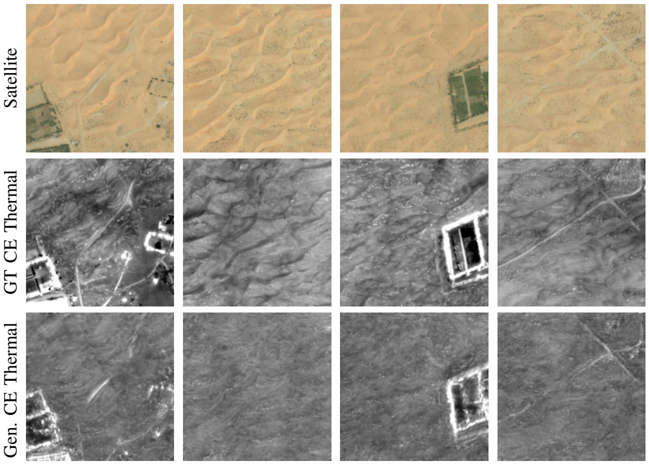

Fig. 3: Examples of satellite images, ground truth (GT) thermal images, and generated (Gen.) thermal images with CE in the test region.

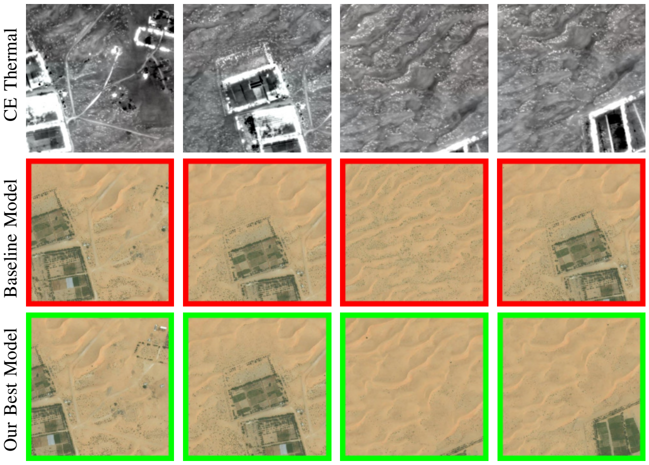

Fig. 4: Examples of ground truth thermal images with CE, correct and failed Top-1 retrieved satellite images in the test region. The correct and failed ones are bounded by green and red colors, respectively.

V. Conclusions

We introduced a thermal geo-localization framework for high-altitude UAV flights using satellite imagery. Addressing the challenge of limited paired satellite and thermal images, our approach blends adversarial domain adaptation with generative modeling, capitalizing on unpaired satellite images. The results demonstrate enhanced geo-localization accuracy, even in regions with low-contrast thermal features.

REFERENCES

[1] A. Couturier and M. A. Akhloufi, “A review on absolute visual localization for uav,” *Robotics and Autonomous Systems*, vol. 135, p. 103666, 2021.

[2] S. Weiss, M. W. Achtelik, S. Lynen, M. C. Achtelik, L. Kneip, M. Chli, and R. Siegwart, “Monocular vision for long-term micro aerial vehicle state estimation: A compendium,” *Journal of Field Robotics*, vol. 30, no. 5, pp. 803–831, 2013.

[3] G. J. V. Dalen, D. P. Magree, and E. N. Johnson, “Absolute localization using image alignment and particle filtering,” in *AIAA Guidance, Navigation, and Control Conference*, 2016.

[4] A. Yol, B. Delabarre, A. Dame, J.-E. Dartois, and E. Marchand, “Vision-based absolute localization for unmanned aerial vehicles,” in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2014, pp. 3429–3434.

[5] B. Patel, T. D. Barfoot, and A. P. Schoellig, “Visual localization with google earth images for robust global pose estimation of uavs,” in *IEEE International Conference on Robotics and Automation (ICRA)*, 2020, pp. 6491–6497.

[6] M. Shan, F. Wang, F. Lin, Z. Gao, Y. Z. Tang, and B. M. Chen, “Google map aided visual navigation for uavs in gps-denied environment,” in *IEEE International Conference on Robotics and Biomimetics (ROBIO)*, 2015, pp. 114–119.

[7] M. Mantelli, D. Pittol, R. Neuland, A. Ribacki, R. Maffei, V. Jorge, E. Prestes, and M. Kolberg, “A novel measurement model based on abbrief for global localization of a uav over satellite images,” *Robotics and Autonomous Systems*, vol. 112, pp. 304–319, 2019.

[8] A. Shetty and G. X. Gao, “Uav pose estimation using cross-view geolocalization with satellite imagery,” in *IEEE International Conference on Robotics and Automation (ICRA)*, 2019, pp. 1827–1833.

[9] X. Tian, J. Shao, D. Ouyang, and H. T. Shen, “Uav-satellite view synthesis for cross-view geo-localization,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 7, pp. 4804–4815, 2022.

[10] S. Chen, X. Wu, M. W. Mueller, and K. Sreenath, “Real-time geolocalization using satellite imagery and topography for unmanned aerial vehicles,” in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2021, pp. 2275–2281.

[11] M. Bianchi and T. D. Barfoot, “Uav localization using autoencoded satellite images,” *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 1761–1768, 2021.

[12] H. Goforth and S. Lucey, “Gps-denied uav localization using preexisting satellite imagery,” in *IEEE International Conference on Robotics and Automation (ICRA)*, 2019, pp. 2974–2980.

[13] M. Mirza and S. Osindero, “Conditional generative adversarial nets,” *arXiv preprint arXiv:1411.1784*, 2014.

[14] T. Mouats, N. Aouf, L. Chermak, and M. A. Richardson, “Thermal stereo odometry for uavs,” *IEEE Sensors Journal*, vol. 15, no. 11, pp. 6335–6347, 2015.

[15] S. Khattak, C. Papachristos, and K. Alexis, “Keyframe-based direct thermal–inertial odometry,” in *International Conference on Robotics and Automation (ICRA)*, 2019, pp. 3563–3569.

[16] J. Delaune, R. Hewitt, L. Lytle, C. Sorice, R. Thakker, and L. Matthies, “Thermal-inertial odometry for autonomous flight throughout the night,” in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2019, pp. 1122–1128.

[17] V. Polizzi, R. Hewitt, J. Hidalgo-Carrió, J. Delaune, and D. Scaramuzza, “Data-efficient collaborative decentralized thermal-inertial odometry,” *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 10 681–10 688, 2022.

[18] Y. Ganin and V. Lempitsky, “Unsupervised domain adaptation by backpropagation,” in *International Conference on Machine Learning (ICML)*, 2015, pp. 1180–1189.

[19] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, “Image-to-image translation with conditional adversarial networks,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1125–1134.

[20] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. Paul Smolley, “Least squares generative adversarial networks,” in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[21] G. Berton, R. Mereu, G. Trivigno, C. Masone, G. Csurka, T. Sattler, and B. Caputo, “Deep visual geo-localization benchmark,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[22] R. Arandjelović, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, “Netvlad: Cnn architecture for weakly supervised place recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 6, pp. 1437–1451, 2018.