Are Deep Learning Models Pre-trained on RGB Data Good Enough for RGB-Thermal Image Retrieval?

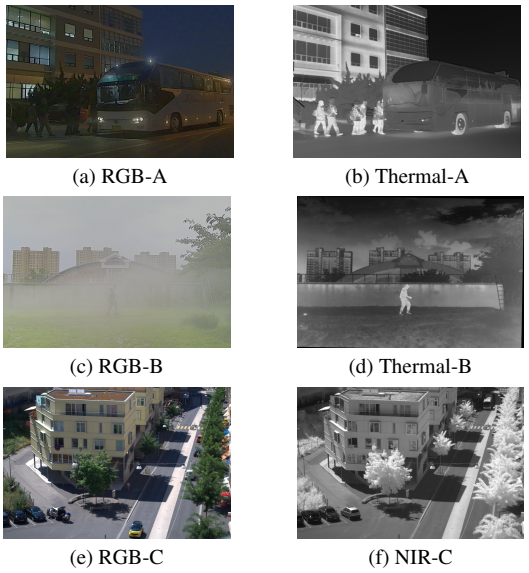
Amulya Pendota, Sumohana S. Channappayya
Indian Institute of Technology Hyderabad, India
{ee21mtech12003, sumohana@iith.ac.in}

Abstract

RGB-Thermal (RGB-T) image retrieval is crucial in scenarios where RGB data alone is insufficient for reliable decision-making. These include all-day, all-weather surveillance and security operations, search and rescue operations and autonomous navigation systems. However, RGB-T image retrieval remains underexplored due to the nature of the currently available datasets. Specifically, these datasets do not lend themselves to training models in the standard RGB visual place recognition (VPR) setting. Therefore, we explore and analyse the effectiveness of existing RGB pre-trained models in addressing the RGB-T image retrieval problem. In particular, we evaluate the performance of numerous pre-trained models on the RGB-T image retrieval task. The efficacy of the models is evaluated on eight RGB-T datasets. Quantitatively, recall rates, Central Kernel Alignment (CKA), and the proposed centroid condition are used for evaluation. Qualitative analysis uses distance plots, t-SNE plots and heatmaps like Saliency Based Similarity Maps (SBSM). Interestingly, and surprisingly, some of the pre-trained models deliver good cross-domain retrieval performance. To the best of our knowledge, this analysis is the first of its kind in RGB-T image retrieval with the available RGB-T datasets. We believe this will serve as a baseline for future work in this area of research.

1. Introduction

In video-based applications like surveillance and autonomous navigation systems that operate all-day and in all-weather conditions, the role of RGB and Thermal (RGB-T) imaging sensors cannot be overemphasized. RGB sensors capture the colour and texture information, while thermal sensors detect infrared radiation and capture the temperature variations in the scene. As seen in the first two rows of Fig. 1, their complementary capabilities address challenges due to low-light conditions (dawn, dusk, night) and adverse weather conditions (fog, mist, smoke etc.). Aligning features across these modalities can be complex due to visual inconsistencies like smoothened edges, lighting conditions, and the nature of information conveyed by each of the two modalities. Additionally, variabilities in the pixel intensities, distribution and domain shifts in RGB-T datasets add another layer of complexity to this task. The ultimate goal of using RGB-T data is to design systems that deliver high-quality and reliable performance irrespective of illumination and weather conditions.



(a) RGB-A (b) Thermal-A (c) RGB-B (d) Thermal-B (e) RGB-C (f) NIR-C

Figure 1. Rows 1 and 2: Corresponding RGB and thermal image pairs from the M3FD [31] dataset. Row 3: A corresponding pair from the VIS-NIR [7] dataset.

In this work, we are particularly interested in image retrieval solutions in scenarios where both modalities are crucial for comprehensive perception and decision-making. However, the state-of-the-art (SOTA) image retrieval literature [4], [18], [3], [50] focuses mainly on the Visual Place Recognition (VPR) task using RGB imagery. The problem of RGB-T image retrieval remains an open challenge due to

the lack of corresponding RGB-T datasets similar to VPR datasets. This shortcoming impedes the training of deep learning models on RGB-T data in VPR settings for the RGB-T image retrieval task. Furthermore, current SOTA VPR algorithms often struggle when evaluated on RGB-T datasets.

We make the following contributions in this work:

1. An extensive evaluation of ImageNet pre-trained models including AlexNet [29], VGG16 [42], SqueezeNet [22], ResNet-N [20] where $N = 18, 34, 50, 101, 152$ on the RGB-T image retrieval task on eight RGB-T datasets. These datasets serve as benchmarks for object detection, pedestrian detection, image translation, image fusion, and segmentation under challenging scenarios.
2. A demonstration that the above pre-trained models perform better than the SOTA VPR and multimodal techniques on the image retrieval task.
3. An explanation of the factors contributing to the superior performance of pre-trained models on RGB-T datasets. Specifically, we assess the models' ability to distinguish between the visually closest and farthest images using various quantitative measures such as CKA [27], the proposed centroid condition, and recall rates. Qualitative measures such as distance plots, feature visualizations using t-SNE [43] and heatmaps like SBSM [12] are used for model explainability.

2. Related Work

Existing studies on matching images from different modalities primarily focus on solving pair-wise image similarity rather than image matching through retrieval. This process involves assigning a similarity score for a given pair of images. Of late, matching multi-spectral images from wavelengths outside the visible spectrum has received significant attention. These multi-spectral images include VIS-NIR (visible - near-infrared), VIS-SWIR (visible - short wave infrared), VIS-LWIR (visible - long wave infrared), and VIS-thermal (or RGB-T). Among these, a majority of papers focus on matching VIS-NIR images (row 3 in Fig. 1). The literature can be broadly categorized into four types: 1) classical key-point-based techniques, 2) learning-based key-point extraction, 3) image patch matching/similarity, and 4) visual place recognition (VPR) methods.

2.1. Classical Key-Point-Based Techniques

The performance of traditional handcrafted methods like SIFT [33], SURF [6], ORB [40] and BRISK [30] in extracting corresponding keypoints in RGB-T images is subpar due to the colour and texture inconsistencies at the pixel level. However, some works try to address this issue on VIS-NIR datasets. For instance, MSIFT [7] extend the idea of opponent colour spaces to VIS-NIR, [34] uses SIFT [33] descriptors of VIS-LWIR along with time information. HCGEC [38] tries to enhance the rough structures and suppress the detailed texture information. HoDMs [14] uses DMs and DBMs to capture the common structure and texture properties between the multi-spectral images. DDCE [14] is based on consistent edge structures, and [46] proposes an OMF method based on self-similarity maps. HGEO [44] uses combined features of maximum gradient and edge-orientated histograms.

2.2. Learning-Based Key-Point Extraction

These methods extract key point correspondences between a given pair of images using a learning mechanism trained on specific data. A majority of the models in this category employ either a Siamese network (with shared weights) or a pseudo-Siamese network (with unshared weights) for the given spectral images. [41] trains a regression model to align the MN-SIFT descriptors of the VIS-NIR images before matching, [21] incorporates Unsuperpoint [9] with CLR [8] loss function to learn VIS-NIR feature point descriptors, [5] and CMM-Net [10] uses both Siamese and pseudo-Siamese networks to learn modality-shared and modality-specific features. S2-Net [36] employs a self-supervised strategy based on detect and describe methods to learn modality-invariant features. Since the above methods focus on finding the key point correspondences, the evaluation is based on the Number of Correctly Matched (NCM) corresponding key points, Correctly Matched Ratio (CMR), RMSE, and percentage of repeatability.

2.3. Image Patch Matching/Similarity

Multi-spectral image patch similarity methods utilize patches of varying sizes, including 32×32, 64×64, and 96×96. In our setting, however, we consider the complete image for analysis. Unlike feature point matching, these methods produce a binary output of whether a given patch pair is similar or not. [2] trains a two-channel network, Siamese and pseudo-Siamese networks independently with concatenated VIS-NIR inputs for a two-channel network. In contrast to regular methods that use higher-level features, AFD-Net [39] uses multi-level feature differences to enhance feature discrimination. MFD-Net [48] employs multiple feature difference networks to reduce the loss of feature difference information at multiple levels, SPIM-Net [26] performs domain translation and matching using two U-Net-based domain translation networks to translate each spectral image to the opposite domain (VIS-NIR) and a dual Siamese network for feature extraction. FIL-Net [49] introduces a feature interaction learning module to understand richer and deeper feature relations between multi-spectral images in a two-branch residual feature extraction network. [25] presents a review of multimodal image-matching methods and their applications.

2.4. Visual Place Recognition Methods

Both key point matching and image patch similarity methods might fail when applied to higher resolution images as these methods focus on extracting features in a local neighbourhood. Matching a higher resolution image (say 480×640, 1024×1280) as opposed to patch matching is quite challenging as it requires a deeper understanding of the semantic relationships between the objects and elements present in the scene. In this work, we focus on solving the RGB-T image retrieval problem with higher resolution images by retrieving the visually closest image from the reference database as a place recognition task. We evaluate the SOTA VPR models like NetVLAD [4] which has a trainable VLAD pooling layer to learn cluster centres for CNN feature maps, PatchNetVLAD [18] which uses global and local patch features at different scales from [4] to rerank the matches, MixVPR [3] which proposes an all-MLP aggregator to learn a compact global descriptor; and R^2Former [50] which uses correlation and attention scores of transformer tokens to rerank the retrievals. To verify the effectiveness of multimodal models that use ViT features, we evaluate Omnivore [16] trained for classifying images, videos and 3D data (all RGB) and Imagebind [17] which is capable of binding information from six modalities by learning a single embedding space. We use models trained on RGB and thermal images from Imagebind for our evaluation. We consider thermal geo-localization [45] as the current SOTA baseline in RGB-T image retrieval literature. The model is trained to match satellite (RGB) images with thermal imagery similar to the problem we are interested in. It uses a NetVLAD aggregator over a CNN backbone along with domain adaptation [15] loss function trained on the captured and synthetically generated Boson-nighttime [45] dataset.

3. RGB-T Image Retrieval

The majority of the multi-spectral methods mentioned above are trained on patches from the VIS-NIR dataset [7], which has images of nine categories. Some of these methods are trained on only one category and use the remaining categories and other VIS-NIR datasets for evaluation. Unlike traditional matching using descriptors, these methods treat it as a classification problem by evaluating whether a given patch pair belongs to a class or not using FPR95. Furthermore, all of these methods focus on specific local regions or features within a patch. Matching VIS-Thermal (RGB-T) images at higher resolutions is more complex than VIS-NIR. This is because NIR images are captured from a band that is very close to the visible spectrum and therefore have a high visual correlation in the information captured between VIS and NIR images. Thermal images, on the other hand, are far from the visible spectrum and have no direct visual correlation. As seen in row 3 of Fig. 1, the fine details in the NIR image are visible and are visually similar to those in an RGB image. These observations summarize the challenges in RGB-T image retrieval.

3.1. Problem Setting

In this work, we analyze the efficacy of the pre-trained models trained on RGB images for different tasks like classification, image retrieval and VPR on the RGB-T image retrieval task on various RGB-T datasets.

Let $\mathbb{Q}$ be a set of query images from the RGB domain and $\mathbb{R}$ be a set of corresponding thermal images treated as a reference database (or gallery). The objective is to retrieve the closest image $I_r \in \mathbb{R}$ for a given $I_q \in \mathbb{Q}$. Note that the query can also be a thermal image with its reference images as RGB images. A thermal image with the corresponding RGB query index is considered as a positive sample, while all other thermal images are considered as negative samples. It is expected that an ideal model should possess a shared latent space where the query images are much closer to the corresponding positive images and farther away from the negative samples in the feature domain. We use Euclidean

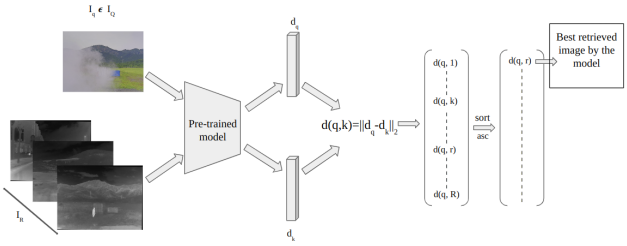


Figure 2. Overview of the RGB-T image retrieval pipeline used for analyzing the pre-trained models. The retrieved image is considered to be a successful match only when 'q' (query index) equals 'r' (retrieved image index).

distance as a metric to find the distance between the RGB query image $I_q \in \mathbb{Q}$ and thermal reference image $I_r \in \mathbb{R}$ by comparing their feature representations as

$$d(I_q, I_r) = \|f(I_q) - f(I_r)\|_2, \tag{1}$$

where $f(I_q)$ and $f(I_r)$ represent the normalized (in channel dimension) flattened feature maps obtained from the last convolutional layer, right before the classifier head, for all ImageNet pre-trained models. For all other models, they represent the output feature vectors for an RGB image and a thermal image, respectively. We compute the distance between the query image I_q and all reference images in $\mathbb{R}$ using Eq. (1). The distances are then sorted in ascending order and the top 10 indices with the lowest distance values are chosen to compute the Recall@N as discussed in Sec. 3.4, while the image with the least distance value is considered as the best match (top@1) retrieved by the model. Fig. 2 shows the evaluation pipeline.

3.2. Pre-trained Models

We evaluate the performance of the following ImageNet [11] pre-trained classification models: AlexNet [29], VGG16 [42], SqueezeNet [22], and ResNet-N [20]. We evaluate the following VPR models trained on RGB data for the RGB-T retrieval task: NetVLAD [4], PatchNetVLAD [18], MixVPR [3] and R^2Former [50]. Also, the efficacy of the following multimodal models is evaluated: Omnivore [16], Imagebind [17]. Finally, SGM [45] a state-of-the-art RGB-T retrieval method for geo-localization is used as a baseline for comparison.

3.3. Datasets

The lack of RGB-T image retrieval task-specific datasets in the literature led us to explore alternative RGB-T datasets designed for tasks like object detection, image fusion, object tracking, segmentation, I2I translation and pedestrian detection. These datasets have pixel-wise correspondences in both RGB and thermal images. The LLVIP [24] dataset has images captured in low-light conditions. The M3FD [31], Roadscene [47], and FLIR_ADAS [1] datasets are mainly used for image fusion and detection tasks. The INO [23] dataset is used for pedestrian detection and segmentation, while the VOT [28] and VTUAV [37] datasets are used for tracking applications. These datasets contain various categories, including people, vehicles like motorcycles, bikes, cars, trucks, vans, trees, and traffic lights, taken in different weather conditions. The Boson-nighttime [45] dataset is the only RGB-T dataset constructed for the thermal geo-localization task. To maintain consistency, we resize all higher resolution images to 320×320 as MixVPR [3] was trained on the same resolution images. For datasets that have high temporal correlation, we sample frames at regular intervals and include them for evaluation. Although thermal images are single-channel images, we use thermal images with repeated channels to ensure compatibility with the existing pre-trained models. Visual samples of all the datasets are given in the suppl. material Sec. 1.

3.4. Evaluation

We use Recall@N with N = 1, 5 for evaluating the accuracy of a model. We treat a prediction to be correct only when the model retrieves the exact corresponding image index from the opposite domain. We don't consider any bin or range to declare a prediction to be a correct match.

4. Experiments and Results

All the experiments have been carried out in inference mode only. No fine-tuning of any of the models has been done. As described in Sec. 3.2 we evaluate the popular ImageNet pre-trained models which are widely used as backbones for many downstream tasks in computer vision, the current SOTA VPR models and the multimodal models.

4.1. Quantitative Evaluation

4.1.1 Recall Rates

The recall rates are used to evaluate the models' ability to retrieve the visually closest images for a dataset. From Tab. 1, the R@1 recall rates of SqueezeNet [22] have outperformed all other models used for evaluation. This superior performance of SqueezeNet is consistent across all datasets, except the Boson-nighttime [45] dataset. The SGM_ResNet-18 [45] that was trained on the Boson-nighttime dataset has the highest recall rates on it. However, it is notable that SGM_ResNet-18 fails to generalize over other existing RGB-T datasets. On the other hand, SqueezeNet has demonstrated much better generalizability compared to other ImageNet pre-trained models, SOTA VPR models, multimodal models, and the thermal geo-localization SGM_Resnet-18 model. In datasets that have temporal correlation, we notice a significant increase in recall rates between R@1 and R@5. This is because we only consider a retrieved image to be a correct match if it matches the query image index exactly. When data is temporally correlated, the model may sometimes miss an exact location retrieval, leading to a decrease in the R@1 recall rate.

4.1.2 CKA

Central Kernel Alignment (CKA) [27] is an index that measures the similarity between the neural network representations that are learned during the training of the model. CKA helps us understand how correlated the pre-trained models' representations are in the feature space when evaluated on RGB and thermal images. Due to the space constraints, we chose to present results from the five models, including the top three ImageNet pre-trained models (SqueezeNet, VGG16, and AlexNet), the best VPR model (PatchNetVLAD), and the baseline SGM_ResNet-18. From Tab. 2, it is evident that the SqueezeNet feature representations of RGB are closer to thermal feature representations across all the datasets followed by VGG16 and AlexNet.

4.1.3 Centroid Condition

We propose a centroid condition to determine how well a model can distinguish visually similar and distinct location images. The analysis for the centroid condition is done in the higher-dimensional feature space. We consider RGB-T datasets LLVIP [24], INO [23] and VOT [28] that have temporally correlated images from different locations. We compute the mean of the feature embeddings of images from each location, representing it as a centroid feature representation of that location. We then calculate the Euclidean

Models/Datasets		LLVIP [24]	M3FD [31]	Roadscene [47]	Boson-nighttime [45]	FLIR ADAS [1]	INO [23]	VOT [28]	VTUAV [37]
AlexNet [29]	R@1 (%)	30.67	73.33	85.52	16.91	11.73	12.14	38.84	89.53
	R@5 (%)	64.84	91	93.67	30.83	39.47	32.24	71.35	96.51
VGG16 [42]	R@1 (%)	43.64	76.66	93.66	25.56	11.73	13.78	29.75	88.66
	R@5 (%)	68.33	**95**	95.64	45.49	42.93	32.71	58.5	97.38
SqueezeNet [22]	R@1 (%)	**51.8**	**80.67**	**96.38**	35.71	**14.13**	**15.65**	**41.87**	**94.77**
	R@5 (%)	**78.8**	94.67	**99.10**	56.39	**44.80**	**37.85**	**77.69**	**98.84**
ResNet-18 [20]	R@1 (%)	25.69	63.33	84.16	10.15	8.00	10.05	16.80	69.77
	R@5 (%)	54.86	87.33	95.02	18.05	28.80	27.80	43.25	91.28
ResNet-34 [20]	R@1 (%)	22.94	57.33	84.62	10.9	7.73	9.11	13.50	65.99
	R@5 (%)	48.88	84.33	95.02	21.80	25.60	23.60	32.78	87.21
ResNet-50 [20]	R@1 (%)	31.42	63.67	89.59	10.90	8.53	9.81	19.28	76.45
	R@5 (%)	55.11	89.33	96.83	23.31	30.13	28.04	40.50	90.99
ResNet-101 [20]	R@1 (%)	26.93	65.33	90.95	11.28	7.20	10.05	14.60	70.35
	R@5 (%)	55.11	92.00	96.38	23.68	28.53	26.40	38.29	89.83
ResNet-152 [20]	R@1 (%)	26.43	65.33	88.24	11.28	7.47	7.48	12.12	70.06
	R@5 (%)	57.61	91.00	97.29	25.19	29.87	23.60	36.36	89.24
NetVLAD [4]	R@1 (%)	11.47	45	67.42	6.02	5.6	4.67	5.79	36.05
	R@5 (%)	35.56	77.33	83.26	17.67	23.47	14.49	23.14	71.22
PatchNetVLAD [18]	R@1 (%)	38.65	70.33	93.67	10.15	12	8.8	19.56	85.47
	R@5 (%)	70.07	87.67	98.19	22.18	40.0	31.31	49.31	95.06
MixVPR [3]	R@1 (%)	15.71	66.00	87.78	21.43	10.4	4.90	11.84	67.44
	R@5 (%)	43.64	95.0	98.19	35.71	36.26	19.62	35.81	88.08
R^2Former [50]	R@1 (%)	12	55.7	68.8	7.9	9.9	4.4	9.1	43.6
	R@5 (%)	33.4	79.7	82.8	19.5	33.3	16.8	30.9	70.9
Omnivore [16]	R@1 (%)	3.24	15.66	14.93	3.75	1.06	2.1	4.95	2.90
	R@5 (%)	11.97	41.33	28.95	10.15	5.33	7.24	14.87	10.17
Imagebind [17]	R@1 (%)	2.99	0.66	0.45	0.37	0.26	0.46	0.27	0.29
	R@5 (%)	10.22	1.33	2.71	1.87	1.33	2.1	0.82	1.16
SGM_ResNet-18 [45]	R@1 (%)	22.69	61.66	84.61	**77.44**	12.26	7.24	18.45	61.91
	R@5 (%)	49.87	83.66	91.40	**98.87**	28.53	22.66	44.90	78.77

Table 1. Comparison of R@1 and R@5 recall rates on popular pre-trained models trained on different tasks. Clearly, SqueezeNet's R@1 scores outperform all other models on seven out of eight RGB-T datasets. The best score is indicated in **bold**, and the next best score is indicated in underline for each dataset.

Datasets/Models	LLVIP [24]	M3FD [31]	Roadscene [47]	Boson [45]	FLIR ADAS [1]	INO [23]	VOT [28]	VTUAV [37]
AlexNet [29]	0.9569	0.8530	0.7483	0.6135	0.9424	0.9894	0.9508	0.9139
VGG16 [42]	0.9732	0.8748	0.8481	0.6939	0.9558	0.9910	0.9614	0.9329
SqueezeNet [22]	**0.9785**	**0.8766**	**0.8738**	**0.7616**	**0.9563**	**0.9924**	**0.9688**	**0.9540**
PatchNetVLAD [18]	0.8483	0.7654	0.4718	0.4257	0.7915	0.9675	0.8482	0.8173
SGM_ResNet18 [45]	0.8022	0.6965	0.6306	0.6367	0.8406	0.9639	0.8369	0.8269

Table 2. The CKA similarity index is calculated by comparing the features of the query and reference database. A higher CKA score indicates that the RGB and thermal features of a particular model are similar. This shows the domain invariant characteristic of the model's feature representations

distance between the centroid feature representations of the RGB and thermal locations. The resulting matrix is expected to have lower distance values across diagonal elements, indicating that RGB features of a location are closer to its corresponding thermal images than to other locations in the opposite domain. A model that satisfies the centroid condition is awarded a score of $+1$ for each location. We repeat this process for all three datasets and sum up the scores to get a single cumulative score.

In Tab. 3, it is observed that SqueezeNet [22] can accurately identify 32 out of 42 locations. However, all the models fail to identify the corresponding locations in the LLVIP dataset. This is because the images in LLVIP have a static background location but have significant changes in the foreground when compared to the VOT and INO datasets.

4.2. Qualitative Evaluation

4.2.1 Distance Plots and top@1 Retrievals

The distance plots visualize how the distances vary for a query $I_q \in \mathbb{Q}$ with the reference images in $\mathbb{R}$ set using Eq. (1). A model with discriminative capability will ideally have a clear minimum at the corresponding query index in the distance plot.

Datasets / Models	No. of locations	AlexNet [29]	VGG16 [42]	SqueezeNet [22]	PatchNetVLAD [18]	SGM_ResNet18 [45]
LLVIP [24]	19	7	8	9	8	**13**
INO [23]	9	**9**	**9**	**9**	7	**9**
VOT [28]	14	13	12	**14**	7	13
Total_score	42	29	29	<u>32</u>	22	**35**

Table 3. Centroid condition scores comparison between the best-performing models. **Bold** indicates the highest scoring model which retrieved the maximum number of locations correctly out of the total number of locations. <u>Underline</u> indicates the next best model that satisfies the centroid condition.

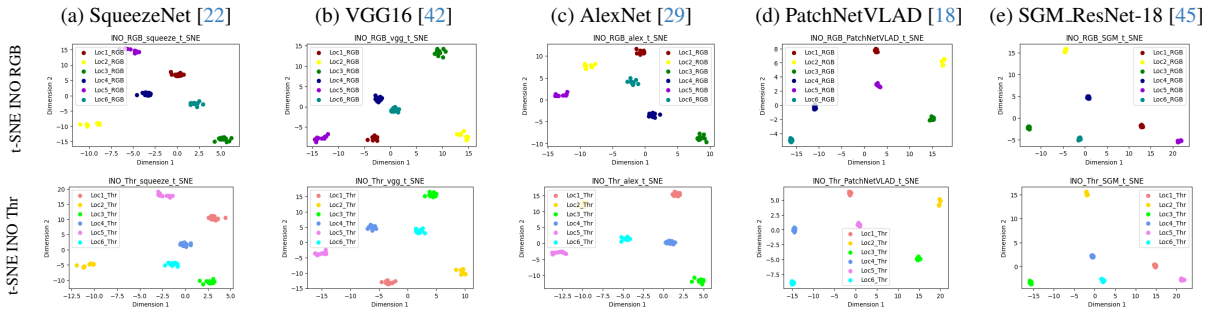


Figure 3. t-SNE plots for the RGB and thermal images of the INO dataset in the first and second rows. To help identify corresponding locations in the plot, similar colors have been used. For example, dark green is used to denote location 3 images in the INO RGB gallery, while light green is used for the corresponding location 3 images in the INO thermal gallery. PCA and UMAP visualizations can be seen in Sec. 3.3 in the suppl. material.

From Fig. 4, we observe that for a given query, the SqueezeNet [22] model has a clear minimum at the exact corresponding query index in the reference database. However, all other models fail to differentiate between similar and dissimilar images and end up with minima at incorrect indices resulting in a wrong retrieval.

Fig. 5 shows top@1 best retrieved visual results by the best-performing models under various scenarios. The qualitative results in rows 1 and 3 in Fig. 5 highlight that SqueezeNet is good at identifying even minute differences between neighbouring frames that are temporally correlated (which is very useful in security-related applications). Row 2 shows the visual results in abnormal weather conditions with the reference database in not-so-good quality. Even in these cases, SqueezeNet continues to retrieve the exact best thermal image for the given RGB query. Additionally, rows 1 and 4 serve as examples of low-illumination settings where SqueezeNet does well again. We also observe that SqueezeNet consistently retrieves the same location images as its second and third-best retrievals, while other models fail to retrieve the exact match even amongst their top three retrievals. Distance plots and the additional qualitative results can be found in the suppl. material in Sec. 3.1, 3.2.

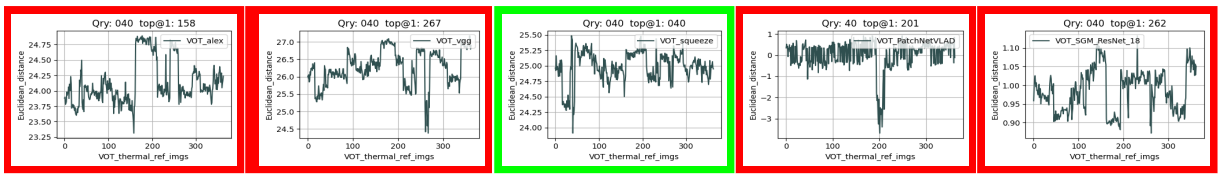
4.2.2 Feature Visualizations

We use t-SNE [43] to reduce the high-dimensional feature representations of the pre-trained models to a lower dimension and visualize them for different locations. We test the models on three datasets: LLVIP [24], INO [23], and VOT [28] as they have temporally correlated images from different locations. We expect the models to cluster the feature representations of each location together. This helps us evaluate the models' ability to compactly represent the features of a location and also understand the global and local relationship between the data samples.

Fig. 3 displays low-dimensional t-SNE feature visualizations for the INO dataset on the best-performing models. Different colour shades are used to indicate correspondence between RGB and thermal images at various locations. For most of the models, we see compact clusters for INO in t-SNE plots and VOT in UMAP [35] plots. The feature visualizations of PCA and UMAP on LLVIP and VOT datasets respectively can be found in Sec. 3.3 in the suppl. material.

4.2.3 Heatmaps

Heatmaps are useful to identify which parts of an image have influenced a model's decision when making a prediction. In this context, we use Saliency Based Similarity Maps (SBSM) [12] which are typically used to explain models in Content-Based Image Retrieval (CBIR) systems. SBSM works by highlighting the importance of a patch when masked, which affects the similarity between the query image and the reference image. To understand which regions of the images have contributed to the decision or prediction, we apply the SBSM technique only to the top@1 best-retrieved image by the model.



(a) AlexNet [29] (b) VGG-16 [42] (c) SqueezeNet [22] (d) PatchNetVLAD [18] (e) SGM_ResNet-18 [45]

Figure 4. Distance plots between a query [40] sample and the reference database of the VOT dataset using pre-trained models. SqueezeNet (highlighted in green) has a sharp clear minimum at the corresponding 40^{th} index, while other models have oscillating curves and have minima at incorrect indices. The title of each plot shows the query index and the top@1 best-retrieved image index by the respective model.

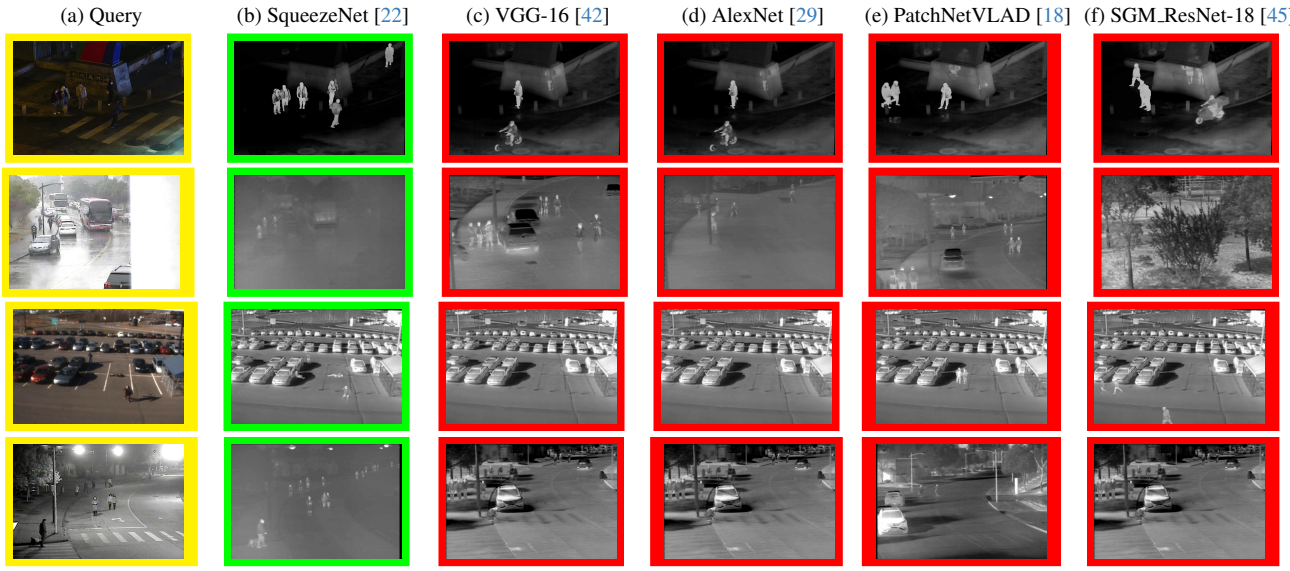


Figure 5. Qualitative examples with an RGB query sample vs top@1 retrieved thermal best match by the best performing models of Tab. 1. The samples are chosen to understand the models' behaviour under a variety of conditions. SqueezeNet continues to outperform other models. SqueezeNet is very good at capturing even the minute changes in the foreground despite the background being static.

The heatmaps in Fig. 6 show that the SqueezeNet top@1 retrieved SBSM maps (in (d)) are in correlation with the reference maps (in (b)) that are generated from the ground truth images. It is also clear that other models' features are not correlated with the ground truth.

4.3. Discussion

Based on the above analysis, we present plausible explanations for the success of ImageNet pre-trained classification models on the RGB-T image retrieval task.

The VPR models considered were primarily trained for place recognition task, which involves learning the unique characteristics of a location in a given scene. However, these models treat foreground objects like trees, traffic lights, vehicles, and people as occlusions and focus more on learning the background aspects of the image. On the other hand, the ImageNet pre-trained models address the classification task with 1000 classes, including many common classes such as motorbikes, cars, trucks, trees, poles, and people that are also present in the RGB-T datasets used for evaluation. Since the ImageNet [11] dataset is vast and contains a diverse range of classes, it may have helped the ImageNet pre-trained models create more powerful discriminative representations than other models, resulting in better recall rates than most SOTA VPR models.

In the case of the multimodal models, Omnivore uses a Swin Transformer [32] which is jointly trained on various classification tasks on images, videos and single-view 3D data (all RGB). On the other hand, Imagebind uses ViT [13] architectures fine-tuned for person classification on the LLVIP dataset. When evaluated with the pre-trained weights, these models with transformer features fail to compete with ImageNet pre-trained models. The SGM_ResNet-18 [45] model trained on the Boson-nighttime [45] dataset gives the highest recall rates for the same dataset. However, it fails to generalize well on all other RGB-T datasets because the Boson-nighttime dataset on which the model was trained, does not have variation or distinctive features as we

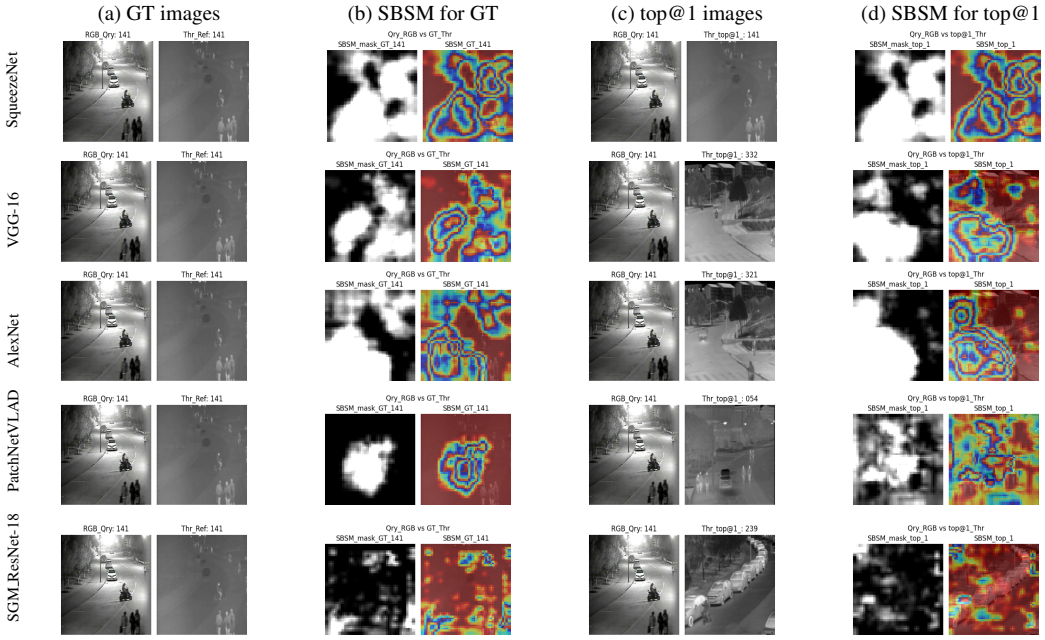


Figure 6. Column (a) shows an RGB query and its corresponding ground truth (GT) thermal image pair. Column (b) shows the mask and the resulting SBSM visualizations for the image pair in column (a). Similarly, column (c) displays the RGB query and the top@1 best-retrieved thermal image, while column (d) has a mask and a map for the corresponding image pair in column (c). This is to understand what is expected and what the model actually sees in a given RGB-T pair. SqueezeNet's SBSMs are correlated while others fail.

see in other datasets. The Boson-nighttime dataset was captured in the desert area and so is the reason for the SGM [45] model's poor generalizability.

The SqueezeNet's unique architecture may also contribute to its superior performance compared to other models. Unlike traditional CNN models such as VGG, AlexNet, and ResNet, SqueezeNet mainly uses fire modules instead of convolutional layers and incorporates delayed downsampling. The fire module significantly reduces the number of model parameters by replacing large-sized kernels with smaller 1×1 and 3×3 kernels and the larger activation maps due to delayed downsampling can lead to higher accuracies [19]. This makes SqueezeNet lightweight and much faster in inference mode, which is highly valued in real-time applications, particularly in resource-constrained environments. Tab. 4 shows the comparison of the parameters of the ImageNet pre-trained models. SqueezeNet has less than a million parameters while every other model used for evaluation has a higher number than SqueezeNet. The VPR, multimodal models and SGM_ResNet-18 use one of these ImageNet pre-trained models as their backbones and hence have a higher number of parameters than SqueezeNet.

Models	AlexNet [29]	VGG16 [42]	SqueezeNet [22]	ResNet-18 [20]
# params (M)	2.46	14.71	**0.72**	11.17
Models	ResNet-34 [20]	ResNet-50 [20]	ResNet-101 [20]	ResNet-152 [20]
# params (M)	21.28	23.50	42.50	58.14

Table 4. Comparison of model parameters in millions (omitting the classifier head). SqueezeNet, with less than a million parameters, is ideal for real-time applications.

5. Conclusion

We evaluated the suitability of pre-trained deep learning models trained on RGB data for the RGB-T image-retrieval task. Based on a thorough analysis of various factors, including model architectures, datasets, and quantitative and qualitative measures, we observe that ImageNet pre-trained models generalize well for the RGB-T image retrieval task. SqueezeNet [22] is a clear standout and is a good choice for real-time applications, consistently demonstrating superior performance and generalization capabilities when compared to other models. While our experiments with ImageNet pre-trained models focused on the features of the last convolutional layer, we would like to explore the capabilities of early CNN layers as they capture the low-level semantics like basic patterns, structures, edges and other attributes that may aid in improving the task performance in the future work and also investigate the recent foundational models for RGB-T image retrieval task. We also emphasize the need for VPR-related RGB-T datasets for the RGB-T image-retrieval task.

References

[1] Teledyne FLIR ADAS. Free teledyne flir thermal dataset for algorithm training, 2018. 4, 5

[2] Cristhian A Aguilera, Francisco J Aguilera, Angel D Sappa, Cristhian Aguilera, and Ricardo Toledo. Learning cross-spectral similarity measures with deep convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–9, 2016. 2

[3] Amar Ali-Bey, Brahim Chaib-Draa, and Philippe Giguere. Mixvpr: Feature mixing for visual place recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2998–3007, 2023. 1, 3, 4, 5

[4] Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Netvlad: Cnn architecture for weakly supervised place recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5297–5307, 2016. 1, 3, 4, 5

[5] Elad Ben Baruch and Yosi Keller. Joint detection and matching of feature points in multimodal images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10): 6585–6593, 2021. 2

[6] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. Speeded-up robust features (surf). *Computer vision and image understanding*, 110(3):346–359, 2008. 2

[7] Matthew Brown and Sabine Süsstrunk. Multi-spectral sift for scene category recognition. In *CVPR 2011*, pages 177–184. IEEE, 2011. 1, 2, 3

[8] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 2

[9] Peter Hviid Christiansen, Mikkel Fly Kragh, Yury Brodskiy, and Henrik Karstoft. Unsuperpoint: End-to-end unsupervised interest point detector and descriptor. *arXiv preprint arXiv:1907.04011*, 2019. 2

[10] Song Cui, Ailong Ma, Yuting Wan, Yanfei Zhong, Bin Luo, and Miaozhong Xu. Cross-modality image matching network with modality-invariant feature representation for airborne-ground thermal infrared and visible datasets. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–14, 2021. 2

[11] 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. 4, 7

[12] Bo Dong, Roddy Collins, and Anthony Hoogs. Explainability for content-based image retrieval. In *CVPR Workshops*, pages 95–98, 2019. 2, 6

[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 7

[14] Zhitao Fu, Qianqing Qin, Bin Luo, Chun Wu, and Hong Sun. A local feature descriptor based on combination of structure and texture information for multispectral image matching. *IEEE Geoscience and Remote Sensing Letters*, 16(1):100–104, 2018. 2

[15] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR, 2015. 3

[16] Rohit Girdhar, Mannat Singh, Nikhila Ravi, Laurens van der Maaten, Armand Joulin, and Ishan Misra. Omnivore: A single model for many visual modalities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16102–16112, 2022. 3, 4, 5

[17] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15180–15190, 2023. 3, 4, 5

[18] Stephen Hausler, Sourav Garg, Ming Xu, Michael Milford, and Tobias Fischer. Patch-netvlad: Multi-scale fusion of locally-global descriptors for place recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14141–14152, 2021. 1, 3, 4, 5, 6, 7

[19] Kaiming He and Jian Sun. Convolutional neural networks at constrained time cost. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5353–5360, 2015. 8

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2, 4, 5, 8

[21] Junchong Huang, Wei Tian, Yongkun Wen, Zhan Chen, and Yuyao Huang. Unsupervised multi-spectral image feature point matching under different lighting scenes. In *2021 5th CAA International Conference on Vehicular Control and Intelligence (CVCI)*, pages 1–6. IEEE, 2021. 2

[22] Forrest N. Iandola, Song Han, Matthew W. Moskewicz, Khalid Ashraf, William J. Dally, and Kurt Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and <0.5MB model size, 2017. 2, 4, 5, 6, 7, 8

[23] INO. Video analytics dataset, 2015. 4, 5, 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. 4, 5, 6

[25] Xingyu Jiang, Jiayi Ma, Guobao Xiao, Zhenfeng Shao, and Xiaojie Guo. A review of multimodal image matching: Methods and applications. *Information Fusion*, 73:22–71, 2021. 2

[26] Yeongmin Ko, Yong-Jun Jang, Vinh Quang Dinh, Hae-Gon Jeon, and Moongu Jeon. Spectral-invariant matching network. *Information Fusion*, 91:623–632, 2023. 2

[27] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network represen-

tations revisited. In *International conference on machine learning*, pages 3519–3529. PMLR, 2019. 2, 4

[28] Matej Kristan, Jiri Matas, Ales Leonardis, Michael Felsberg, Roman Pflugfelder, Joni-Kristian Kamarainen, Luka Cehovin Zajc, Ondrej Drbohlav, Alan Lukezic, Amanda Berg, Abdelrahman Eldesokey, Jani Kapyla, Gustavo Fernandez, Abel Gonzalez-Garcia, Alireza Memarmoghadam, Andong Lu, Anfeng He, Anton Varfolomieiev, Antoni Chan, Ardhendu Shekhar Tripathi, Arnold Smeulders, Bala Suraj Pedasingu, Bao Xin Chen, Baopeng Zhang, Baoyuan Wu, Bi Li, Bin He, Bin Yan, Bing Bai, Bing Li, Bo Li, Byeong Hak Kim, and Byeong Hak Ki. The seventh visual object tracking vot2019 challenge results. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, 2019. 4, 5, 6

[29] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012. 2, 4, 5, 6, 7, 8

[30] Stefan Leutenegger, Margarita Chli, and Roland Y Siegwart. Brisk: Binary robust invariant scalable keypoints. In *2011 International conference on computer vision*, pages 2548–2555. Ieee, 2011. 2

[31] 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. 1, 4, 5

[32] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 7

[33] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60:91–110, 2004. 2

[34] William Maddern and Stephen Vidas. Towards robust night and day place recognition using visible and thermal imaging. In *Proceedings of the RSS 2012 Workshop: Beyond laser and vision: Alternative sensing techniques for robotic perception*, pages 1–6. University of Sydney, 2012. 2

[35] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Grossberger. Umap: Uniform manifold approximation and projection. *The Journal of Open Source Software*, 3(29):861, 2018. 6

[36] Shasha Mei, Yong Ma, Xiaoguang Mei, Jun Huang, and Fan Fan. S2-net: Self-supervision guided feature representation learning for cross-modality images. *IEEE/CAA Journal of Automatica Sinica*, 9(10):1883–1885, 2022. 2

[37] Zhang Pengyu, 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 conference on computer vision and pattern recognition*, 2022. 4, 5

[38] Yueming Qin, Zhiguo Cao, Wen Zhuo, and Zhenghong Yu. Robust key point descriptor for multi-spectral image matching. *Journal of Systems Engineering and Electronics*, 25(4): 681–687, 2014. 2

[39] Dou Quan, Xuefeng Liang, Shuang Wang, Shaowei Wei, Yanfeng Li, Ning Huyan, and Licheng Jiao. Afd-net: Aggregated feature difference learning for cross-spectral image patch matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3017–3026, 2019. 2

[40] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. Orb: An efficient alternative to sift or surf. In *2011 International conference on computer vision*, pages 2564–2571. Ieee, 2011. 2

[41] Sajid Saleem and Abdul Bais. Visible spectrum and infrared image matching: a new method. *Applied Sciences*, 10 (3):1162, 2020. 2

[42] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015. 2, 4, 5, 6, 7, 8

[43] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9 (11), 2008. 2, 6

[44] Quan Wu and Shipeng Zhu. Multispectral image matching method based on histogram of maximum gradient and edge orientation. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2021. 2

[45] Jiuhong Xiao, Daniel Tortei, Eloy Roura, and Giuseppe Loianno. Long-range uav thermal geo-localization with satellite imagery. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5820–5827. IEEE, 2023. 3, 4, 5, 6, 7, 8

[46] Xin Xiong, Guowang Jin, Qing Xu, and Hongmin Zhang. Self-similarity features for multimodal remote sensing image matching. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:12440–12454, 2021. 2

[47] 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, 5

[48] Chuang Yu, Yunpeng Liu, Chenxi Li, Lin Qi, Xin Xia, Tianci Liu, and Zhuhua Hu. Multibranch feature difference learning network for cross-spectral image patch matching. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–15, 2022. 2

[49] Chuang Yu, Yunpeng Liu, Jinmiao Zhao, Shuhang Wu, and Zhuhua Hu. Feature interaction learning network for cross-spectral image patch matching. *IEEE Transactions on Image Processing*, 2023. 2

[50] Sijie Zhu, Linjie Yang, Chen Chen, Mubarak Shah, Xiaohui Shen, and Heng Wang. R2former: Unified retrieval and reranking transformer for place recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19370–19380, 2023. 1, 3, 4, 5