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

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

We present additional results in support of the claims made in the main paper.

1. Dataset Comparison

The comparison between the RGB-T datasets is shown in Fig. 1, which includes an RGB-T pair from each evaluated dataset. The title of each image indicates the name of the dataset from its respective RGB and thermal (Thr) galleries. Additionally, Fig. 2 displays some samples from the VIS-NIR [4] dataset. It is worth noting that the near-infrared images have clear detailing of the elements, similar to the RGB images, except for the color information. The luminance variations can be easily observed in the near-infrared images, which is not the case for the thermal images. In thermal images, we can only see the temperature aspects of an object without additional detailing. Thus, these figures visually demonstrate the complexity of the RGB-thermal datasets compared to VIS-NIR datasets present in the literature.

2. Recall Rates on VIS-NIR Dataset

To further support that VIS (RGB)-NIR image retrieval is much easier than RGB-T image retrieval problem, we evaluate the ImageNet pre-trained models on VIS-NIR [4] dataset on each of the nine categories present in it. From Tab. 1 we see that most of the models perform well on all the categories during inference. The numbers in the table indicate that VIS-NIR is a less complex dataset compared to RGB-T datasets for the image retrieval task.

3. Additional Results

3.1. Distance plots

Fig. 3 show the distance plots for all the pre-trained models for query 141 in the VOT dataset. All the models shown in red retrieve a wrong match, while SqueezeNet continues to retrieve a correct match.

3.2. Qualitative Plots

Fig. 4 shows top@1 retrieval for all the models used for evaluation. A green bounding box indicates a correct retrieval while red indicates incorrect retrieval by a model.

Fig. 5 shows a few other visual cases of query vs top@1 best retrieved image for the best performing models as discussed in the main paper. Fig. 4 shows the top@1 best thermal retrievals by all the models for a given RGB query image. Fig. 6 shows the comparison of the top three best-retrieved images among the best-performing models for a given same query image. The captions of each image indicate the corresponding retrieved index by the respective model. We see that SqueezeNet is consistent in retrieving the images from similar locations along with exact match in its top@1 while other models fail to retrieve the correct match even in its three best retrievals.

3.3. Feature Visualisation Using PCA and UMAP

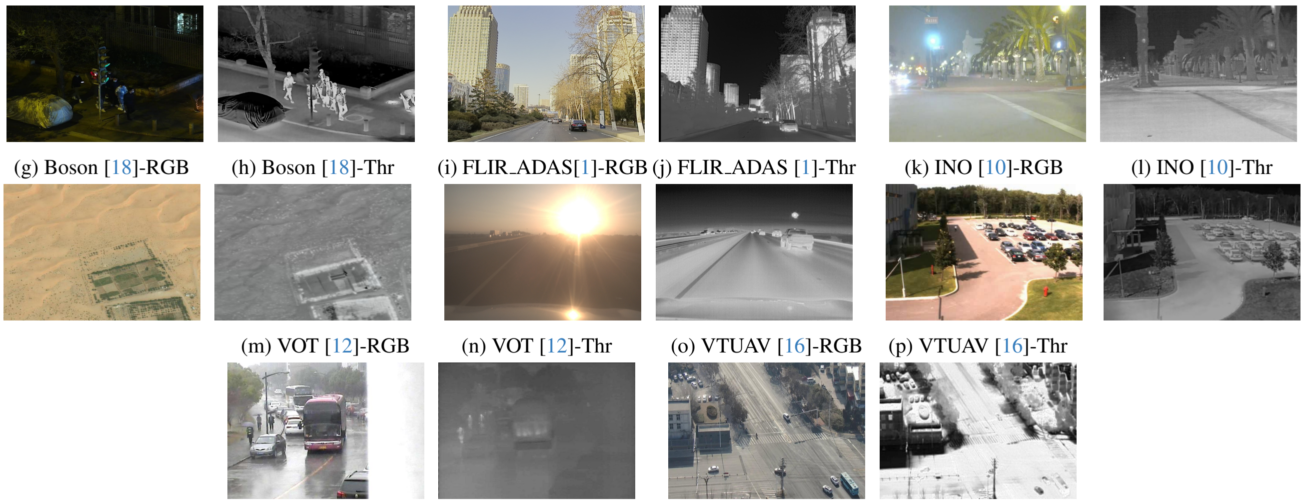
Fig. 7 show the feature visualization using PCA on the LLVIP dataset and UMAP [15] on the VOT dataset for the best-performing models. The colour coding is chosen to help identify the corresponding location clusters in RGB and thermal PCA and UMAP plots. The LLVIP dataset does not show well-defined PCA clusters for all the models because of major foreground changes in the images.

References

[1] Teledyne FLIR ADAS. Free teledyne flir thermal dataset for algorithm training, Year. 2

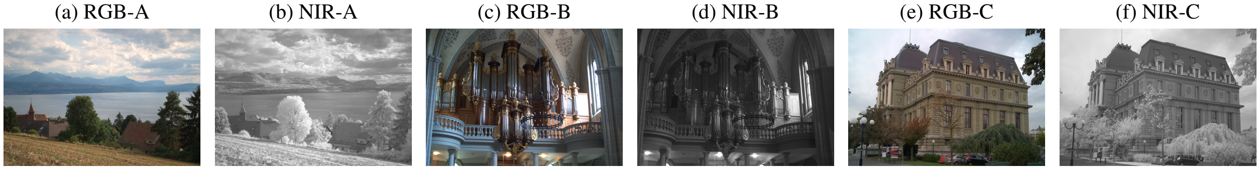
[2] Amar Ali-Bey, Brahim Chaib-Draa, and Philippe Giguere. Mixvpr: Feature mixing for visual place recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2998–3007, 2023. 3, 4

[3] 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. 3, 4



(a) LLVIP [11]-RGB (b) LLVIP [11]-Thr (c) M3FD [14]-RGB (d) M3FD[14]-Thr (e) Roadscene [19]-RGB (f) Roadscene [19]-Thr
(g) Boson [18]-RGB (h) Boson [18]-Thr (i) FLIR_ADAS[1]-RGB (j) FLIR_ADAS [1]-Thr (k) INO [10]-RGB (l) INO [10]-Thr
(m) VOT [12]-RGB (n) VOT [12]-Thr (o) VTUAV [16]-RGB (p) VTUAV [16]-Thr

Figure 1. Sample RGB-Thermal pairs from all the RGB-T datasets considered for evaluation.



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

Figure 2. Sample visible-near-infrared (VIS-NIR) pair from VIS-NIR dataset.

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

[5] 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

[6] 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

[7] 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. 3, 4, 6

[8] 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. 3, 4

[9] 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¡ 0.5 mb model size. *arXiv preprint arXiv:1602.07360*, 2016. 3, 4, 6

[10] INO. Video analytics dataset, 2015. 2

[11] 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. 2

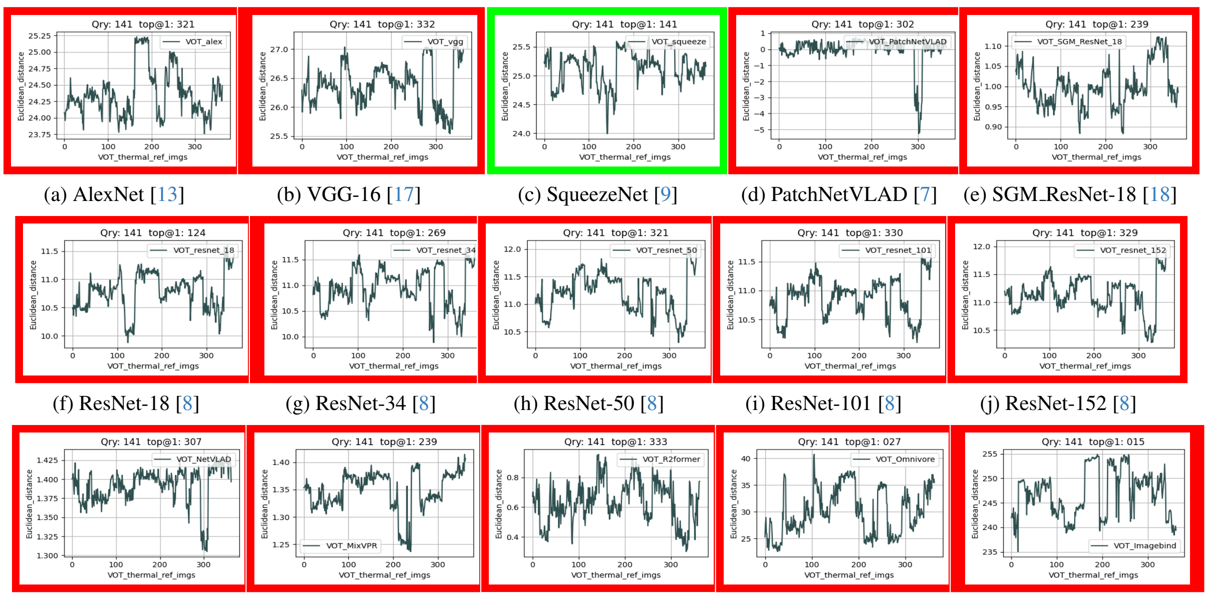
[12] Matej Kristan, Jiri Matas, Ales Leonardis, Michael Felsberg, Roman Pflugfelder, Joni-Kristian Kamarainen, Luka Cehovin Zajc, Ondrej Drbohlav, Alan Lukezic, Amanda Berg, Abdelrahman Eldesokey, Jani Kapyla, Gustavo Fernandez, Abel Gonzalez-Garcia, Alireza Memarmoghadam, Andong Lu, Anfeng He, Anton 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. 2

[13] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012. 3, 4, 6

[14] 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

Datasets/Models	Country	Field	Forest	Indoor	Mountain	Oldbuilding	Street	Urban	Water
AlexNet [13]	96.15%	92.16%	100%	100%	100%	100%	100%	100%	98.04%
VGG16 [17]	96.15%	90.2%	100%	100%	98.18%	100%	100%	100%	98.04%
SqueezeNet [9]	98.08%	98.04%	100%	100%	98.18%	100%	100%	100%	100%
ResNet-18 [8]	96.15%	84.31%	90.57%	100%	94.55%	100%	100%	100%	98.04%
ResNet-34 [8]	94.23%	84.31%	86.79%	100%	98.18%	100%	100%	100%	100%
ResNet-50 [8]	98.08%	92.16%	98.11%	100%	98.18%	100%	100%	100%	100%
ResNet-101 [8]	100%	92.16%	92.45%	100%	98.18%	100%	100%	100%	100%
ResNet-152 [8]	100%	92.16%	90.57%	100%	100%	100%	100%	100%	98.04%

Table 1. Fairly high recall rates were observed on both visible and near-infrared images for all categories in the VIS-NIR dataset when evaluated on ImageNet pre-trained models.



(a) AlexNet [13] (b) VGG-16 [17] (c) SqueezeNet [9] (d) PatchNetVLAD [7] (e) SGM_ResNet-18 [18]

(f) ResNet-18 [8] (g) ResNet-34 [8] (h) ResNet-50 [8] (i) ResNet-101 [8] (j) ResNet-152 [8]

(k) NetVLAD [3] (l) MixVPR [2] (m) R^2former [20] (n) Omnivore [5] (o) Imagebind [6]

Figure 3. Distance plots of all the pre-trained models for a query from the VOT dataset. The green colour represents correct retrieval, while the red colour represents a wrong retrieval.

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. 2

[15] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018. 1

[16] 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. 2

[17] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 3, 4, 6

[18] 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. 2, 3, 4, 6

[19] 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. 2

[20] 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. 3, 4

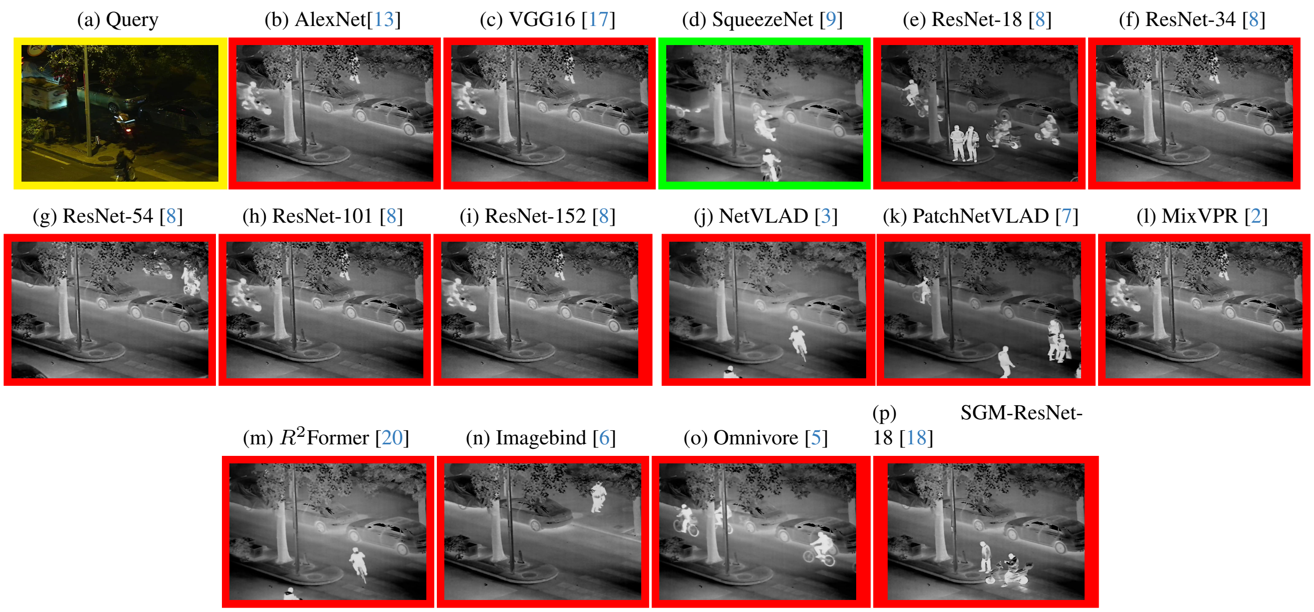


Figure 4. Few other visual examples of best-retrieved images by all the models for a given RGB query image.

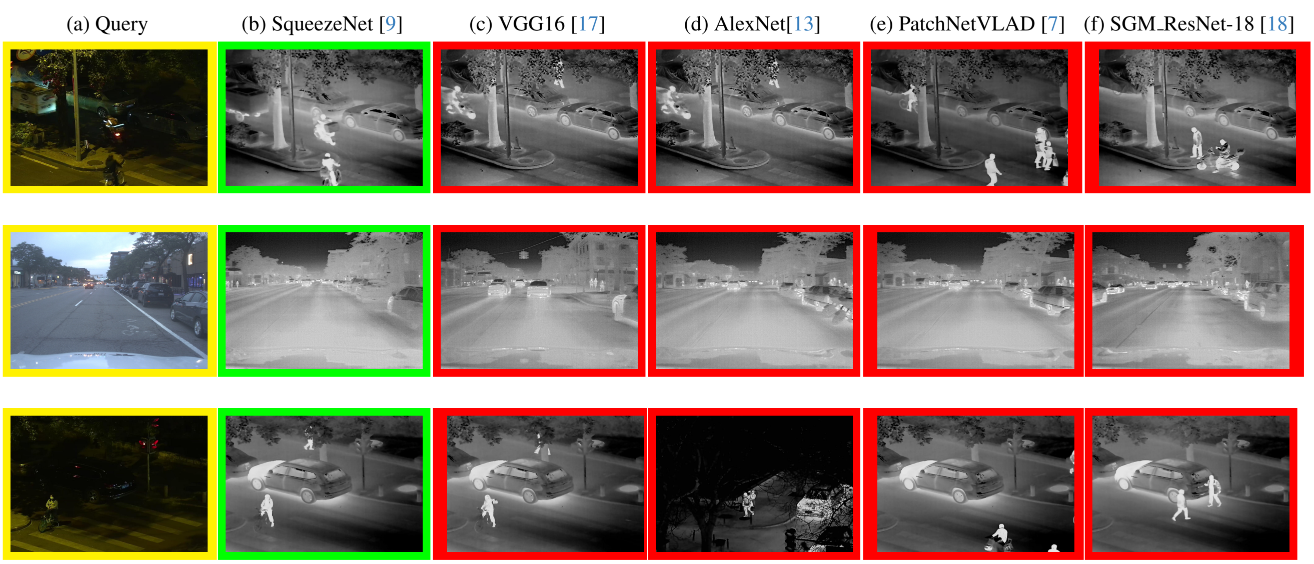


Figure 5. A few more visual examples of best-retrieved images in continuation to the results in the main paper.

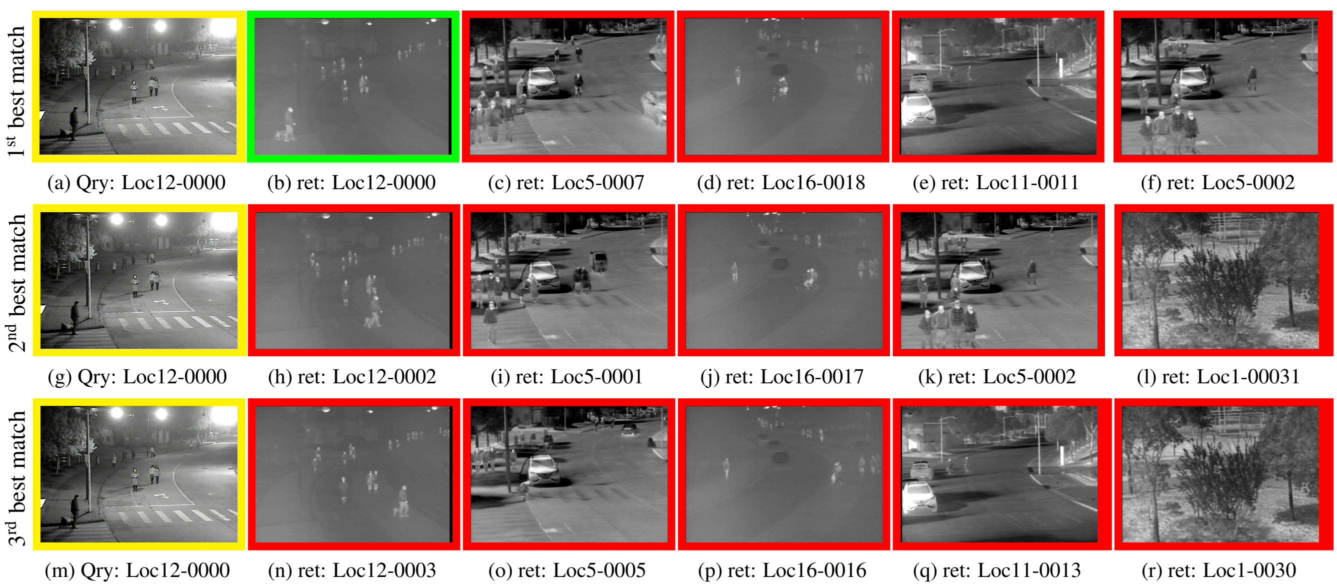


Figure 6. Top three thermal images retrieved by the best-performing pre-trained models listed in the main paper. The retrieved images correspond to a single RGB query. The first, second, and third best-retrievals are in rows 1, 2, and 3 respectively. The results indicate that SqueezeNet performs the best in retrieving the exact match with the query image. Additionally, SqueezeNet consistently retrieves the same location images as its second and third-best-retrievals. On the other hand, other models fail to retrieve the exact match even amongst their top three retrievals.

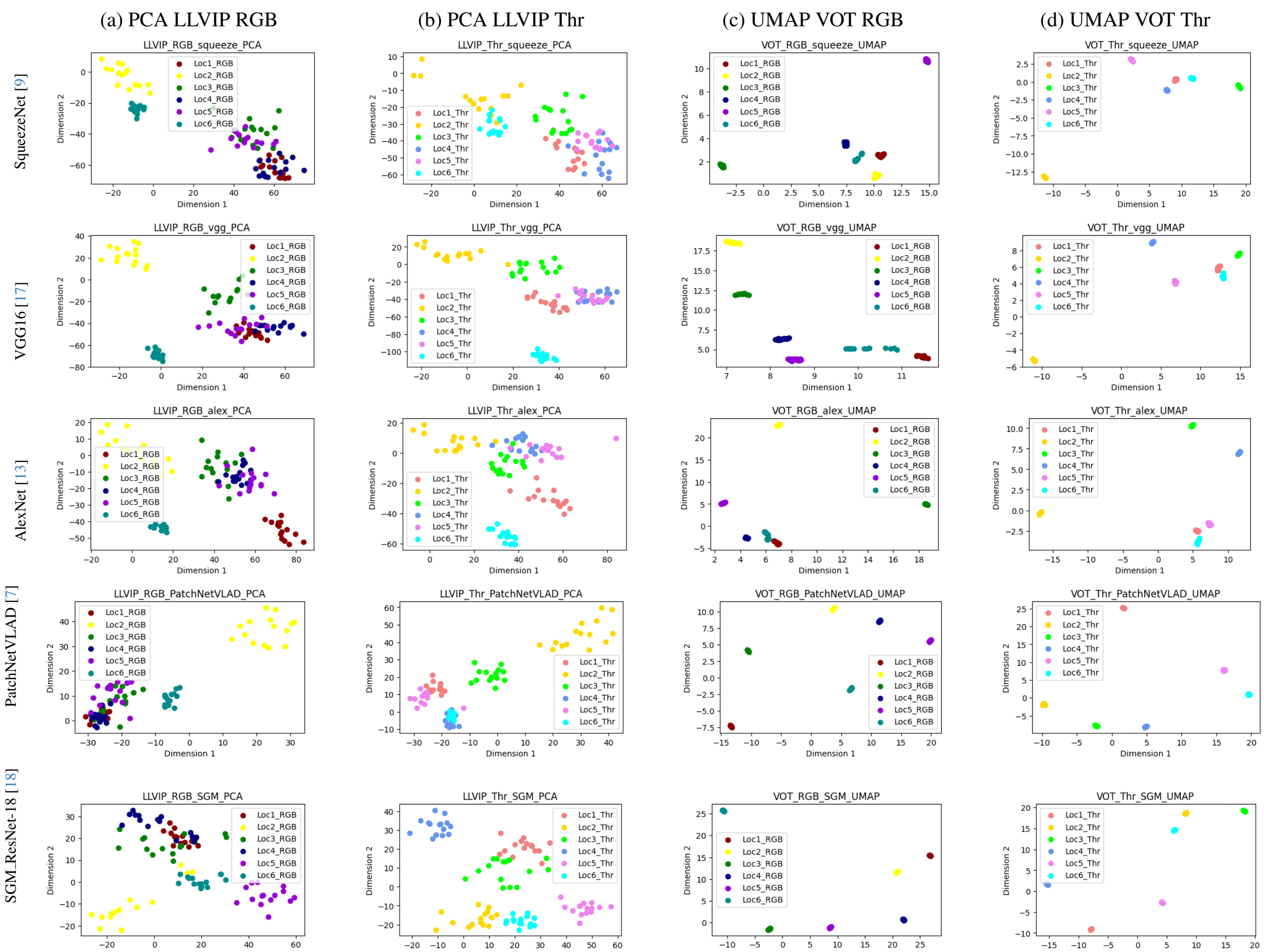


Figure 7. PCA and UMAP visualisations for the RGB and thermal images from the LLVIP and VOT datasets respectively.