Thermal Infrared Image Georeferencing using an Ensemble of Land Cover Predictions

Mojgan Madadikhaljan, Michael Schmitt

University of the Bundeswehr Munich, Department of Aerospace Engineering, Munich, Germany -
(mojgan.madadikhaljan, michael.schmitt)@unibw.de

Key Words: TIR image georeferencing, geolocation, land cover classification, deep learning

Abstract

Precise georeferencing of remote sensing data is usually implemented in a post-processing fashion and is a crucial step for Earth observation applications such as change detection, natural hazard management, and ground target tracking. It is particularly important for small satellites intending to perform temperature monitoring and wildfire detection on a global scale wherein the precise location of fire is to be communicated. The cost-efficient navigating systems on board such satellites are often not capable of providing accurate geolocation information directly due to space and power limitations. Therefore, it is very important to have a globally applicable georeferencing refinement framework that is robust against illuminational and time-relevant scene changes. In this paper, we propose a georeferencing framework for thermal infrared images that consists of ensemble matching of deep learning-based land cover predictions to archival, well-georeferenced land cover maps. We verify the proposed framework on the georeferencing of single-band Landsat thermal imagery. Experimental results show the efficiency and practicality of the method with 72% of the test images geolocated within 1-pixel accuracy with no trajectory information available.

1. Introduction

Spaceborne remote sensing (RS) images often show an initial geo-positional misalignment ranging from meters (for Sentinel-2) to kilometers (for modern Cubesats). Different phenomena such as collision avoidance maneuvers, the inefficiency of low-cost navigating systems on board, or the misfunctioning of navigating systems due to temperature changes can lead to this degradation of the positional accuracy of the acquired images. The geo-positional misalignment is more detrimental for applications such as wildfire detection wherein information about the precise location of the fire is crucial for prompt action. Therefore, many research studies focus on improving the georeferencing of acquired RS images either by improving navigation systems or post-processing refinements (Mostafa and Schwarz, 2001; Aguilar et al., 2017; Leprince et al., 2007; Chen et al., 2022).

Since the improvement of onboard navigating systems of small satellites faces limitations due to space and power constraints, a vast majority of geopositioning refinement methods are realized by image-to-image matching strategies. These strategies intend to match the mislocated image - target image - to precisely georeferenced data - reference data - by either area-based, feature-based, or learning-based matching methods. By defining and comparing a similarity measure in the image (Li et al., 2015) or frequency domain (Reddy and Chatterji, 1996), area-based matching methods aim to find the most similar window in the reference image to the target image (Ma et al., 2021). The similarity measures differ from correlation-based approaches (Li et al., 2015) to domain transformation (Reddy and Chatterji, 1996) and mutual information (MI) based (Cao et al., 2020) methods. In feature-based methods, both reference and target images are searched for mutual feature points and feature descriptors that are used to find the transformation. To register multimodal RS data Ye et al. (2017) present a novel feature descriptor named the histogram of orientated phase congruency (HOPC), which is based on the structural properties of images. To tackle the problem of significant nonlinear intensity differences between multimodal RS data, Ye et al. (2019) suggest pixel-wise extraction of the histogram of oriented gradient (HOG) as descriptors. They define a fast similarity measure using the fast Fourier transform (FFT) and then apply a template-matching strategy to detect correspondences between images. Learning-based methods use deep learning to generate features, feature descriptors, or similarity metrics. Hughes et al. (2020) present a three-step DL-based framework for sparse image matching of SAR and optical imagery. They first predict matching-appropriate regions in each image and next, generate a correspondence heatmap, and then remove the outliers by classifying the correspondence surface as a positive or negative match. Aiming for low-cost RS image registration, Ye et al. (2022) propose a multi-scale framework with unsupervised learning, called MU-Net that directly learns the transformation parameters of image pairs. The stacks of several deep neural network (DNN) models on multiple scales in the MU-Net prevent the backpropagation trapping into a local extremum and resist significant image distortions.

Despite the aforementioned approaches to georeference RS data by image matching, the challenge persists for small satellites with TIR imaging cameras and a significant range of geopositioning error magnitudes. Area-based matching methods are sensitive to the similarity measure definition and the level of texture in the target image. The images acquired from CubeSats with large geopositioning errors require large search areas which intricates the definition of a robust similarity measure. In particular, when considering TIR images with very low textural details and daily illumination differences, the solely area-based methods fail to correctly match the target image. The feature-based matching methods on the other hand are highly dependent on appropriate search for feature points and definition of feature descriptors. In addition to being computationally expensive, feature-based methods are centered around low-level image features which are majorly disadvantageous in the TIR domain lacking sharp textural details such as salient points, lines, and regions. Daily thermal emission changes add to the complication of the feature extraction task. As fully data-driven methods,

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-3-2024-223-2024

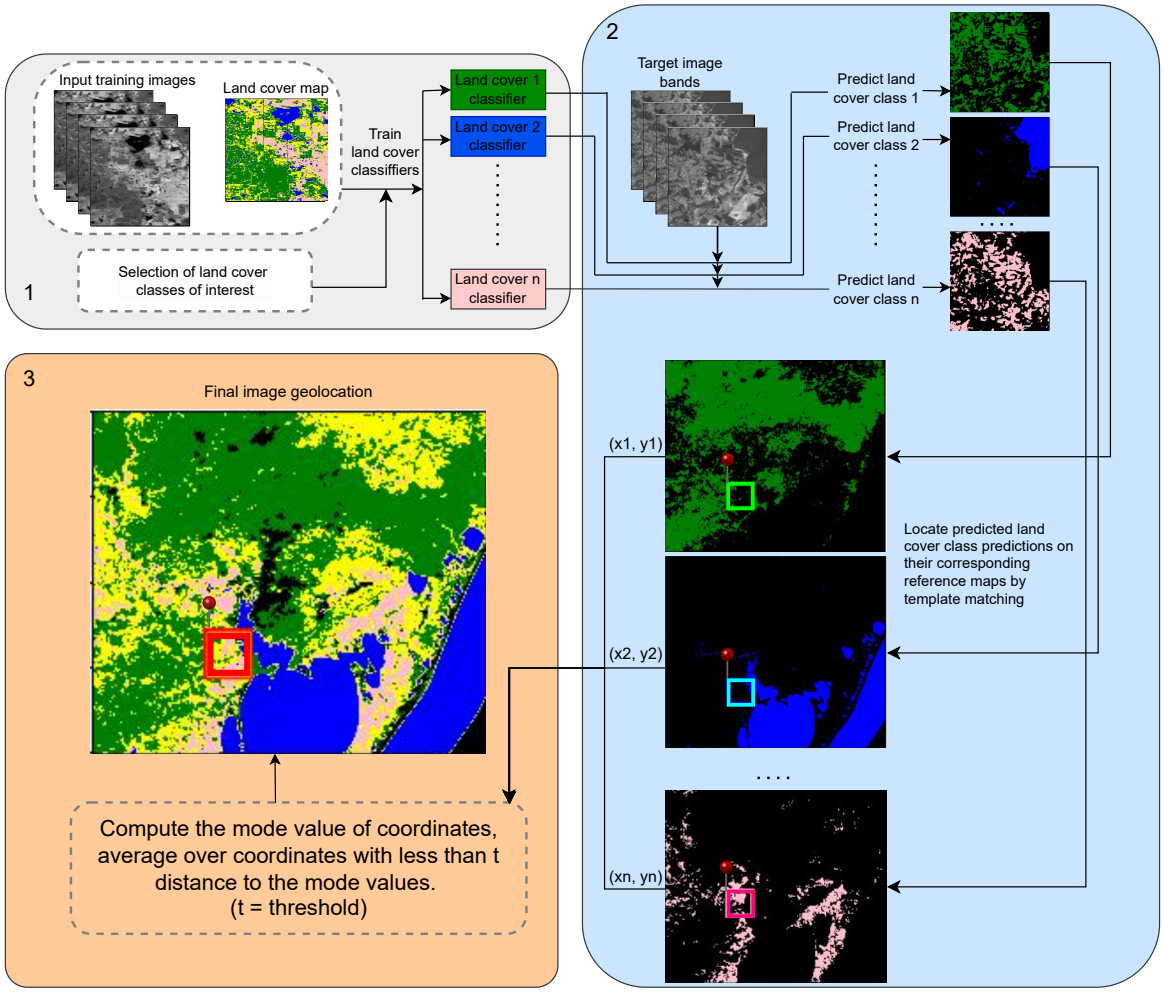

Figure 1. Overview of the 3 steps of the framework.

the learning-based approaches lack representative training data which covers the complexity and characteristics of multi-modal especially TIR RS data (Zhu et al., 2023). Additionally, many of the georeferencing techniques only focus on improving the geopositioning of a specific region (Van Ha et al., 2018; Khlopenkov et al., 2009; Aguilar et al., 2017; Hakim et al., 2018).

Building on top of our previous work (Madadikhaljan and Schmitt, 2023), with this paper, we propose a globally applicable georeferencing refinement framework (see Fig. 1) that focuses on high-level image features and is a composite of the three matching strategies reviewed above:

- Our method is **feature-based** since the target high-level features used for matching are land cover (LC) classes
- Our method is **learning-based** since we train multiple deep neural networks to predict LC maps from the target images
- Our method is **area-based** since the LC predictions are matched to an archival, well-georeferenced land cover map using a cross-correlation-based template matching strategy.

2. Proposed Georeferencing Framework

As shown in Fig. 1, the proposed framework consists of three main steps. The initial step includes training several LC classifiers. Next, the target image is put into the models, and corresponding LC maps are predicted. The predicted LC maps then are template-matched to their corresponding reference LC map and located based on the highest correlation. Finally, outlier-exclusive averaging is conducted to conclude the final geolocation of the image.

2.1 DL-based Land Cover Prediction

The first step (see Fig. 1, Part 1) of the framework consists of training different single-class LC classifiers. TIR images measure the thermal emission of the objects in the scene and are therefore highly correlated to the temperature and consequently the type of LC classes present in the scene. Compared to other feature-based methods, high-level features of LC class boundaries are independent of daily temperature changes and provide robust image textures. Also, with the availability of global world cover maps such as ESA world cover or dynamic world data (Brown et al., 2022), the presence of up-to-date high-resolution reference LC maps is guaranteed. A deep learning model of choice such as UNet or ResNet is trained to perform the pixel-wise binary segmentation of the input satellite image. The target classes are selected based on the geographical coverage of the images, the available data, the time of interest, and the information content of the input channel and their correlation to LC classes. The robustness of the proposed method concerning seasonal LC changes is obtained through careful choice of season-independent LC classes as features of interest. In the case of using ESA world cover, where a single LC label is as-

signed to each pixel for a period of the whole year, season-dependent LC classes such as seasonal rivers are to be avoided. While single-class LC predictions are used in the scope of this paper, Multiclass segmentation can be considered to be included in the decision fusion phase. To encourage true positives in the predictions, highly present and well-distributed LC classes are selected for training. (Schmitt and Zhu, 2016)

2.2 Template Matching

The target images are then inserted into the trained models from Step 1 and corresponding LC predictions are generated (see Fig. 1, Part 2). The predicted LC classes are then searched for in the corresponding reference LC maps. In contrast to traditional area-based methods that rely on image intensities, our approach utilizes LC predictions ensuring an enhanced and illumination-independent similarity comparison. The search area is often not the whole globe but a location defined by coarse direct georeferencing, considering a buffer of the expected georeferencing error of the devices. For instance, if the products of a satellite can have offsets up to 100 km, the reference map will have a buffer of around 100 km around the coordinates of direct georeferencing. The process of template matching is a simple computation of the two-dimensional correlation coefficient, which is very fast considering the reduced search area and binary-to-binary image comparison.

2.3 Decision Fusion of Matching Results

Each LC class will produce a location vote for the target image. The probability of mislocalization of the target image caused by imperfect LC predictions is reduced by averaging non-outlier geolocation votes. That means, within all the location votes $location \quad votes = \{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$ the mode values x_{mode} and y_{mode} is computed. All location votes are compared to the mode value x_{mode} and y_{mode} with a threshold of t. If a location vote is larger than t, they are considered outliers and are excluded from the final averaging. The definition and exclusion of outliers can be with one-pixel accuracy in case of high precision requirements. If no mode value exists (i.e., there is no repeating location vote), the mean will be the representative mode value and the estimated location vote is compared to the mean with a threshold t. The final geolocation of the image is derived from averaging over non-outlier location votes (see Fig. 1 part 3). The fusion of location votes robustifies the final positioning results.

3. Experiment and Results

We show the validity of the proposed framework in a complex case where the geolocation information of the onboard geopositioning systems error is relatively high and can be up to 500 pixels off and therefore, the image is to be geolocated in a large search area. In addition to significant geolocation error, we consider the images to be single-band thermal images to care for night-time imaging where RGB cameras do not provide any content for geolocation improvement.

The experiment consists of creating a global dataset with train and test images, training several classifiers, and geolocating the test data using correlation-based template matching methods.

3.1 Dataset

The dataset for training is created using globally distributed points for which Landsat band 10 (10.6 - 11.19 µm) together with the corresponding ESA World cover maps (Zanaga et al., 2022) are selected with a size of 512×512 pixels from 2021. ESA World Cover data are downsampling to the pixel spacing of 100 meters. The classes of interest are chosen as described below:

- ***Water bodies***. *water bodies* including rivers, seas, and coastlines are highly visible and detectable in thermal images and are often located close to wildfire-prone areas such as forests.
- ***Tree covers including forests***. This class is well-presented in the whole dataset and particularly plays an important role in wildfire-prone areas. Forests typically exhibit temperature differences with respect to their surroundings and therefore are also an interesting class for this experiment.
- ***Croplands***. Another well-presented class that can be interesting in case of farmland fires.
- ***Grasslands***. Also, a well-distributed and highly present class that is closely correlated to *cropland* and can be used for georeferencing

The classes such as shrublands, moss, and lichen, etc. are excluded in this experiment due to the very low amount of data available. Additionally, due to the fact that their typical areas are quite small, they don’t provide good matching features. The train set contains a varying number of images for each class (4720 for *tree cover*, 4231 for *grassland*, 2364 for *cropland*, and 1137 for *water bodies*) the test set consists of 30 carefully selected and globally distributed images (see Fig. 2). The images in the test set are selected in a way to includes all classes for a comprehensive evaluation.

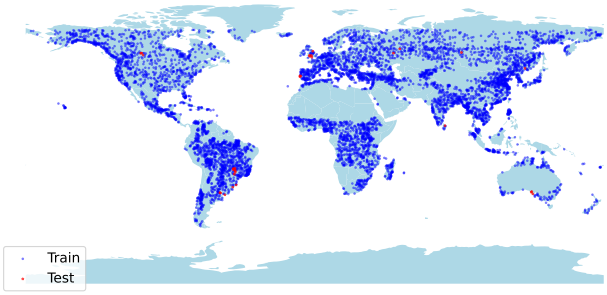

Figure 2. The global distribution of train and test data.

3.2 Training the classifier

To train a classifier, we choose the well-known U-Net (Ronneberger et al., 2015) due to its high performance in segmentation tasks. The Batch size of training is chosen as 64, with each image randomly cropped to 256×256 patches to prevent overfitting. The learning rate is 0.001, with 10% of validation data to tune the hyperparameters, trained for 500 epochs for each binary segmentation scenario. Table 1 shows the performance of the trained models on the validation set by the end of the training. The F1 score and Intersection over Union metrics are used as established measures for binary segmentation tasks to evaluate the performance of the models.

Figure 4 shows the output of the trained models for 4 sample test thermal images. The first row, from left to right illustrates the input image, and the ground truth LC maps, the *tree cover*, *water bodies*, *croplands*, and *grassland* predictions, respectively.

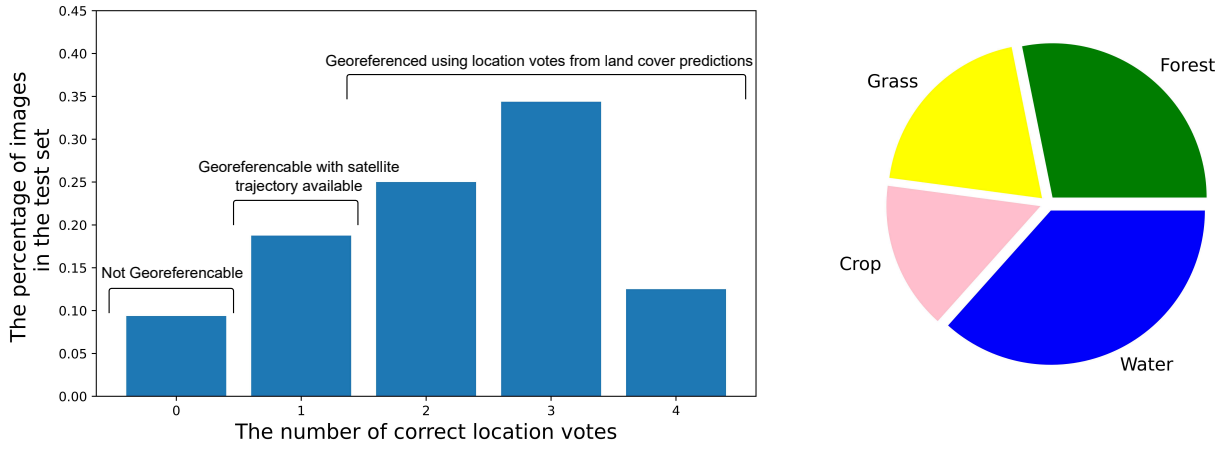

Figure 3. The performance evaluation charts. The left bar chart illustrates the percentage of test images with respect to their number of correct location votes from land cover predictions. The pie chart on the right shows how much each of the LC classes contributed to finding the correct geolocation.

Classes	Water	Forest	Crop	Grass
IOU	0.76	0.66	0.52	0.44
F1 score	0.86	0.79	0.67	0.61

Table 1. The classification performance of each land cover class on the validation set.

3.3 Template Matching of Land Cover Maps

The predicted LC maps are then searched for in a reference LC map. The search area is not the whole globe, but a buffered area of 200×200 km around the correct location of the image.

The search of predictions in the reference binary LC maps is conducted through a fast normalized cross correlation (Lewis, 1995). The window with the maximum correlation is considered as the geolocation outcome of the corresponding LC map. In Fig. 4 the geolocated predictions can be seen inside the corresponding reference maps for all samples and all LCs of interest.

3.4 Fusion for Final Geolocation Estimation

The average of non-outlier geolocation predictions is calculated to obtain the final geolocation of input thermal images. To exclude the outlier predictions, all x and y coordinate votes are compared to the mode value (or mean if no mode value is available) by a threshold of t. In our experiment, the threshold is set to one pixel. As depicted in Fig. 4, the final geolocation vote from all samples uses the votes from all class predictions except the last sample wherein the geolocation outcome from *grassland* is considered as an outlier and therefore excluded from the final vote. It means that a test image has to have at least 2 correct geolocation predictions from its LC template matches to be geo-locatable by the proposed framework. At the end of the experiments, we are able to geolocate 72% of the test images within a search area of 200×200 km to an accuracy of 1 pixel without trajectory information.

4. Discussion

From the results in the previous section, some key insights regarding the efficiency and applicability of the framework can be drawn:

Georeferencable	Without trajectory information	With trajectory information	Not
Nr. of test images	23/32	29/32	3/32

Table 2. The georeferencing results on test images with and without trajectory information.

- The bar plot in Fig. 3 shows the percentage of data with 0 to 4 correct location votes with an accuracy of one pixel. It can be seen that a high percentage of data receives at least 2 correct location votes while for 3% of test data, no correct location votes are found. For 20% of test images with only 1 correct location vote, the inaccurate votes can be excluded in the presence of trajectory information.

- The pie chart in Fig. 3 shows the contribution of each LC class to derive the accurate geolocation. It is to be seen that *water bodies* and *tree cover* or *forest* predictions have superior prediction performances with a single-band thermal image and are therefore more beneficial in locating the image. It is, therefore, crucial to select target LC class features that are well-presented and accurately detectable in the data. Rare LC classes are hard to train, difficult to find in the reference map - due to the large empty areas- and therefore less beneficial in the whole framework. The evaluation of the performance of each LC prediction can also be utilized to weight the LC classes with higher performances.

- According to Table 1, the highest IOU value and F1 Score belongs to the water classifier. It means, that predicting water classes from the thermal data is the most trivial task with respect to the other land cover classes of interest. Next, forest pixels can be detected with a relatively high performance. Grass pixels are however the most difficult ones to be detected from a single thermal band by this training setup, yet achieving rather satisfactory results. The overview of the model’s performances highly correlates to the pie chart in Figure 3. It means that the land cover classes with high-performance classifiers contribute more to finding the right geolocation.

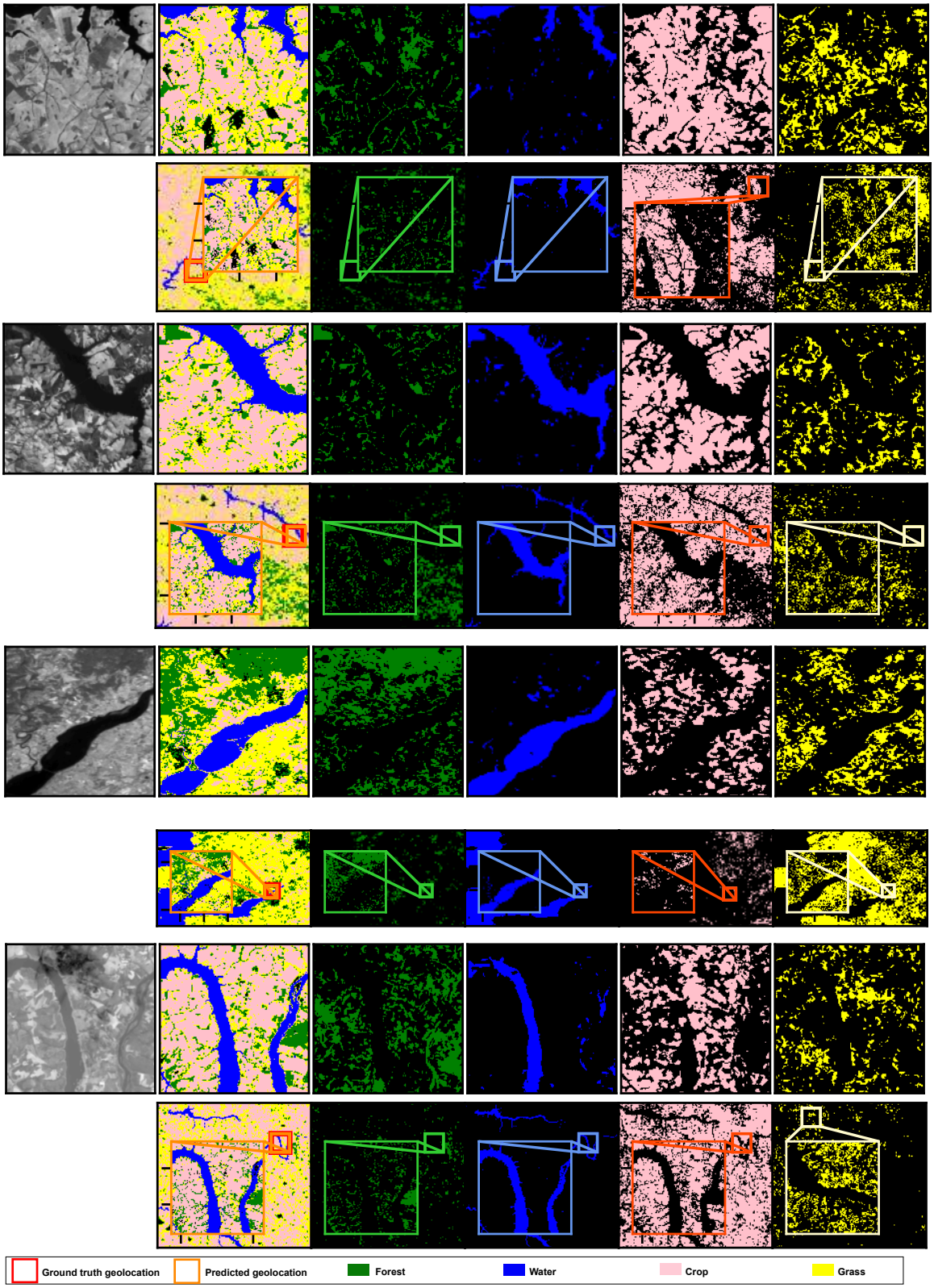

Figure 4. Illustration of geolocation results of 4 sample test images using the proposed framework. From left to right for each sample, the input thermal image, the ground truth land cover map, and the forest, water, cropland, and grassland predictions are shown in the first row. The second row includes the geolocated land cover predictions in the search area for each class (columns 2-5) and the visual comparison of the predicted final geolocation prediction and the ground truth geolocation in the first column.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-3-2024-223-2024 | © Author(s) 2024. CC BY 4.0 License.

- From Table 2, it can be seen that 23 out of 32 test images (72%) can be georeferenced solely based on their LC predictions without any trajectory information considered. The availability of satellite trajectory information enables the geolocating of images with at least one correct Lc prediction i.e., $23/32 = 90\%$.

- In Fig. 5 more challenging scenarios are demonstrated. The first sample from the top has low-performance predictions for *tree cover* and *cropland* classes and therefore receives its final geolocation results only from *water* and *grassland* classes. The second example test image in the middle, however, can only be correctly geolocated based on *water* class if the trajectory is available. Low coverage by the target land cover class leads to inaccurate location votes from *tree cover*, *cropland*, and *grassland* predictions. The last sample at the bottom of Fig. 5 can not be georeferenced due to the erroneous predictions from all LC classes. Low coverage of *water* class, together with low inaccurate LC predictions for other classes lead to failed georeferencing results.

- The focus of this study is on extracting robust patterns and the capability of the framework to perform the template match in large search areas. Within the experiments of this paper, no rotation, shearing, or scaling is involved. In case of drastic geometrical image distortions, the proposed framework will be insufficient in finding the pattern of the predicted LC class. However, in the case of rotation and scale changes, more complex template-matching methods can be utilized where the predicted land cover map will then be searched in the reference map at different scales and varying directions using the image pyramid technique (Chen et al., 2016). The main limitation of the proposed method is its sensitivity to high levels of shearing and image deformations.

- While the attention of this work centers around the thermal imagery domain, the proposed framework can also be well extended to the imagery from other modalities. The other superior modalities of RS (i.e., optical, SAR) are also physical measurements of the Earth's surface and hence, correlated to LC types present in the scene (Brown et al., 2022; Balzter et al., 2015; Eisavi et al., 2015).

- In case this methodology is used for novel Cubesat missions for which no data is available yet, the classifiers can be trained on existing data from established satellite missions and a domain adaptation approach will then adapt the performance of LC classifiers to the images acquired from the target CubeSat.

5. Summary & Conclusion

In this paper, we proposed a globally applicable framework to refine the georeferencing of thermal satellite images particularly suitable for CubeSat data with relatively large geopositioning errors. We examined the practicality of the framework on the created dataset and discussed the corresponding strengths and shortages. Due to the focus on high-level land cover features, the proposed method is robust against changes in season, daytime, illumination, and emission. By predicting multiple LC types, the proposed framework reduces the probability of erroneous geolocation results even in large search areas. The outlier-exclusive averaging allows compensation for the mislocation of images caused by imperfect predictions. The method is globally applicable and can further be employed in optical and SAR imagery domains.

References

Aguilar, M. A., Nemmaoui, A., Aguilar, F. J., Novelli, A., Garcia Lorca, A., 2017. Improving georeferencing accuracy of very high resolution satellite imagery using freely available ancillary data at global coverage. *International Journal of Digital Earth*, 10(10), 1055–1069.

Balzter, H., Cole, B., Thiel, C., Schmullius, C., 2015. Mapping CORINE land cover from Sentinel-1A SAR and SRTM digital elevation model data using random forests. *Remote Sensing*, 7(11), 14876–14898.

Brown, C. F., Brumby, S. P., Guzder-Williams, B., Birch, T., Hyde, S. B., Mazzariello, J., Czerwinski, W., Pasquarella, V. J., Haertel, R., Ilyushchenko, S. et al., 2022. Dynamic World, Near real-time global 10 m land use land cover mapping. *Scientific Data*, 9, 251.

Cao, S.-Y., Shen, H.-L., Chen, S.-J., Li, C., 2020. Boosting structure consistency for multispectral and multimodal image registration. *IEEE Transactions on Image Processing*, 29, 5147–5162.

Chen, C.-S., Huang, C.-L., Yeh, C.-W., Chang, W.-C., 2016. An accelerating CPU based correlation-based image alignment for real-time automatic optical inspection. *Computers & Electrical Engineering*, 49, 207-220.

Chen, J., Cheng, B., Zhang, X., Long, T., Chen, B., Wang, G., Zhang, D., 2022. A TIR-visible automatic registration and geometric correction method for SDGSAT-1 thermal infrared image based on modified RIFT. *Remote Sensing*, 14(6), 1393.

Eisavi, V., Homayouni, S., Yazdi, A. M., Alimohammadi, A., 2015. Land cover mapping based on random forest classification of multitemporal spectral and thermal images. *Environmental Monitoring and Assessment*, 187(291).

Hakim, P. R., Jayani, A. P. S., Sarah, A., Hasbi, W., 2018. Autonomous image georeferencing based on database image matching. *IEEE International Conference on Aerospace Electronics and Remote Sensing Technology (ICARES)*, IEEE.

Hughes, L. H., Marcos, D., Lobry, S., Tuia, D., Schmitt, M., 2020. A deep learning framework for matching of SAR and optical imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 169, 166–179.

Khlopenkov, K. V., Trishchenko, A. P., Luo, Y., 2009. Achieving subpixel georeferencing accuracy in the Canadian AVHRR processing system. *IEEE Transactions on Geoscience and Remote Sensing*, 48(4), 2150–2161.

Leprince, S., Barbot, S., Ayoub, F., Avouac, J.-P., 2007. Automatic and precise orthorectification, coregistration, and subpixel correlation of satellite images, application to ground deformation measurements. *IEEE Transactions on Geoscience and Remote Sensing*, 45(6), 1529–1558.

Lewis, J. P., 1995. Fast normalized cross-correlation. *Vision Interface*, 10, 120–123.

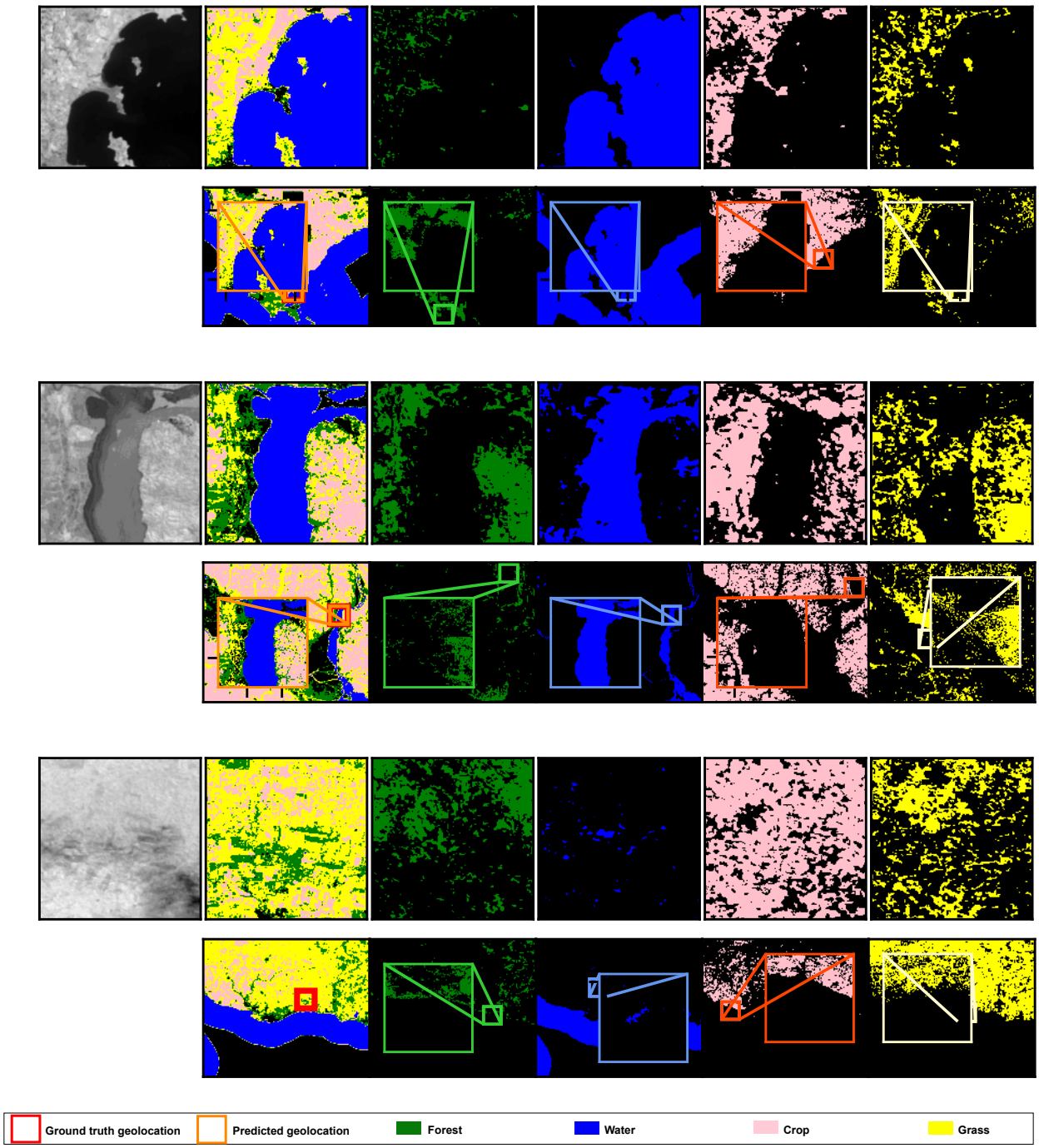

Figure 5. Illustration of geolocation results of 3 challenging test images using the proposed framework. From left to right for each sample, the input thermal image, the ground truth land cover map, and the forest, water, cropland, and grassland predictions are shown in the first row. The second row includes the geolocated land cover predictions in the search area for each class (columns 2-5) and the visual comparison of the predicted final geolocation prediction and the ground truth geolocation in the first column.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-3-2024-223-2024 | © Author(s) 2024. CC BY 4.0 License.

Li, Z., Mahapatra, D., Tielbeek, J. A., Stoker, J., van Vliet, L. J., Vos, F. M., 2015. Image registration based on autocorrelation of local structure. *IEEE Transactions on Medical Imaging*, 35(1), 63–75.

Ma, J., Jiang, X., Fan, A., Jiang, J., Yan, J., 2021. Image matching from handcrafted to deep features: A survey. *International Journal of Computer Vision*, 129, 23–79.

Madadikhaljan, M., Schmitt, M., 2023. Georeferencing thermal satellite images based on land cover information extraction. *IEEE International Geoscience and Remote Sensing Symposium*, IEEE, 6362–6365.

Mostafa, M. M., Schwarz, K.-P., 2001. Digital image georeferencing from a multiple camera system by GPS/INS. *ISPRS Journal of Photogrammetry and Remote Sensing*, 56(1), 1–12.

Reddy, B. S., Chatterji, B. N., 1996. An FFT-based technique for translation, rotation, and scale-invariant image registration. *IEEE Transactions on Image Processing*, 5(8), 1266–1271.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, Springer International Publishing, 234–241.

Schmitt, M., Zhu, X. X., 2016. Data fusion and remote sensing: An ever-growing relationship. *IEEE Geoscience and Remote Sensing Magazine*, 4(4), 6-23.

Van Ha, P., Thanh, N. T. N., Hung, B. Q., Klein, P., Jourdan, A., Laffly, D., 2018. Assessment of georeferencing methods on modis terra/aqua and viirs npp satellite images in vietnam. *International Conference on Knowledge and Systems Engineering (KSE)*, IEEE, 282–287.

Ye, Y., Bruzzone, L., Shan, J., Bovolo, F., Zhu, Q., 2019. Fast and robust matching for multimodal remote sensing image registration. *IEEE Transactions on Geoscience and Remote Sensing*, 57(11), 9059–9070.

Ye, Y., Shan, J., Bruzzone, L., Shen, L., 2017. Robust registration of multimodal remote sensing images based on structural similarity. *IEEE Transactions on Geoscience and Remote Sensing*, 55(5), 2941–2958.

Ye, Y., Tang, T., Zhu, B., Yang, C., Li, B., Hao, S., 2022. A multiscale framework with unsupervised learning for remote sensing image registration. *IEEE Transactions on Geoscience and Remote Sensing*, 60, 5622215.

Zanaga, D., Van De Kerchove, R., Daems, D., De Keersmaecker, W., Brockmann, C., Kirches, G., Wevers, J., Cartus, O., Santoro, M., Fritz, S., Lesiv, M., Herold, M., Tsendbazar, N.-E., Xu, P., Ramoino, F., Arino, O., 2022. ESA WorldCover 10 m 2021 v200.

Zhu, B., Zhou, L., Pu, S., Fan, J., Ye, Y., 2023. Advances and challenges in multimodal remote sensing image registration. *IEEE Journal on Miniaturization for Air and Space Systems*, 4(2), 165-174.