Building Sustainability: Using Artificial Intelligence for Estimating Construction Year from multi-modal street-view - EO dataset [AI4EO Challenge]

Aim and scope. The main aim of this document is to present the baseline model for the AI4EO Challenge. We examine the performance of the model for the training and testing with all three modalities, i.e. street-view images, aerial Very High Resolution (VHR) images, and Sentinel-2 data. The results include the confusion matrix, the average of the diagonal items of the confusion matrix, the off-diagonal items of the confusion matrix, the accuracy, precision, recall, and F1-score.

We also examine the results for the training and testing with only the street-view images. These evaluation results are compared to the results for the training and testing with all three modalities. We observe that there is benefit in using the aerial VHR images and the Sentinel-2 data in addition to the street-view images, as this improves performance. It is thus possible to benefit from all the modalities, specifically from aerial images and Sentinel-2 data, in addition to street-view images.

Furthermore, we examine the results for the testing without the street-view images using only two out of the three trained encoders. Here, we perform inference/ testing with only two modalities, i.e. aerial VHR images and Sentinel-2 data. These results are compared to the evaluation results for the training and testing with all three modalities. The model is the same for these two cases, i.e. same training.

Contents of this document

1. Introduction

Overview and importance. Sustainable buildings minimize energy and water consumption and are a key part of responsible and sustainable urban planning and development that seeks to effectively combat climate change. The building age plays an important role in building energy modelling and urban planning policies. However, despite the huge number of multi-modal datasets and deep neural network architectures that have been developed over the years, data of the construction period is not always publicly available. In this work, we aim to develop a method that will automatically estimate the construction period of buildings with improved performance using Artificial Intelligence (AI) techniques and cross-view datasets.

We use the new Map your City Dataset (MyCD) which includes building mappings of European cities and comprises cross-view inputs of: (a) Street-view images of buildings from seven classes for the building construction epoch, (b) The corresponding aerial top-view images, and (c) The corresponding satellite Sentinel-2 data. The images of this dataset are co-registered with respect to the specific house under study. The labels are the corresponding construction epoch class.

The MyCD dataset has data from 6 countries and 19 cities in Europe. The building construction epoch classes are 7, labelled from 0 to 6 where *Class 0* is the building construction age before

1930, *Class 1* the time period 1930-1945, *Class 2* the epoch 1946-1960, *Class 3* the period 1961-1976, *Class 4* the time epoch 1977-1992, *Class 5* the period 1993-2006, and *Class 6* after 2006.

Our main target is to develop a model to estimate the building construction epoch using deep neural networks. Moreover, our aim is to highlight the role and contribution of the aerial and satellite images to the model performance. Several different deep neural network architectures can be used to develop the building construction epoch model, including CNN, ResNet, and Transformer. The proposed model is trained and tested on the MyCD dataset. In particular, our model is tested on images from the MyCD dataset that include all the three modalities of street-view images, aerial images, and Sentinel-2 data, as well as on the two modalities of aerial images and Sentinel-2 data, i.e. without using the street-view images.

Brief description of the dataset. The dataset contains street-view images of buildings from seven classes/ categories for their construction epoch, e.g. “Before 1930” or “After 2006”. The dataset also contains *top-view images* of buildings (VHR data) where the center of such images is the house/ building in the corresponding street-view images.

In Figs. 1 and 2, we show the distribution of the image samples with respect to the country and the construction year, as well as with respect to the *city* and the construction year.

We note that the dataset has three categorizations: (1) Construction epoch which is the *class label* (i.e. 7 classes, 0 to 6), (2) Country, and (3) City.

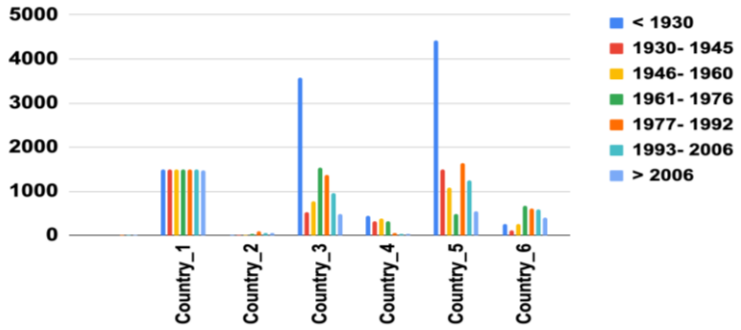


Fig. 1: Number of images per *country* and building construction epoch.

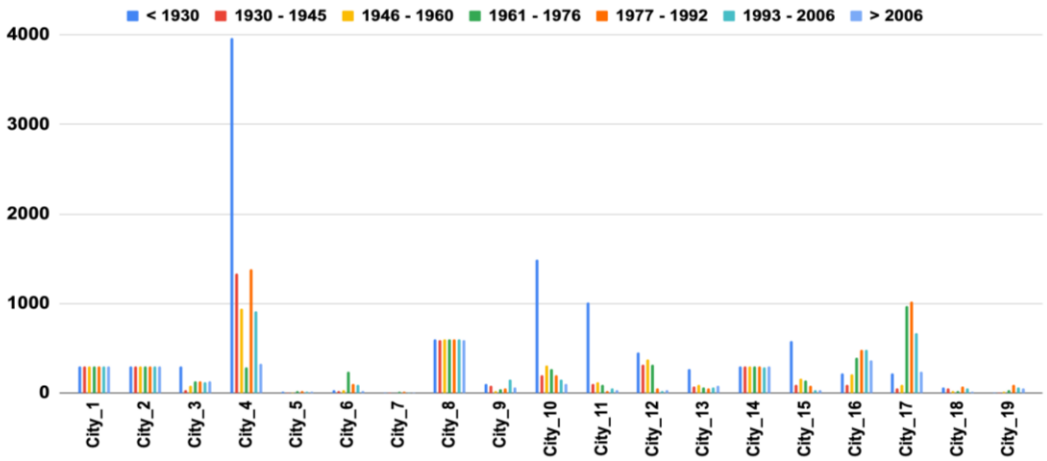


Fig. 2: Images per *city* and construction year class.

Main aim: Classification of the images in the 7 classes concerning their respective building construction epochs with good performance.

New dataset: Houses/ buildings in Europe: Several different cities: 6 countries and 19 cities

Near classes: Similar buildings because houses look alike

Task: Classification with 7 classes: Construction year/ age of buildings

AI4EO Challenge: (7) in http://www.bigdatafromspace2023.org/satellite-events, http://ai4eo.eu

Conference BiDS 2023: Big Data from Space (BiDS), in Vienna, Austria

Example images: *Street*-view image dataset: The images below show buildings/ houses from the first and last classes, i.e. Class 0 which is "< 1930" and Class 6 which is "> 2006".

The new dataset consists of several images of building facades in Europe, organised in folders and classified in 7 classes corresponding to their respective *construction epochs*.

Evaluation procedure:

Classification task setup. To examine and assess model generalisation, we use a leave-cities-out evaluation methodology. Specifically, we leave 4 cities out from training so that we can evaluate the models on these held-out cities and test the generalisation performance.

In this way, we examine the impact of region/ city change on the inference performance of the models.

Classes: Construction year: 7 classes. Class labels: 0 to 6

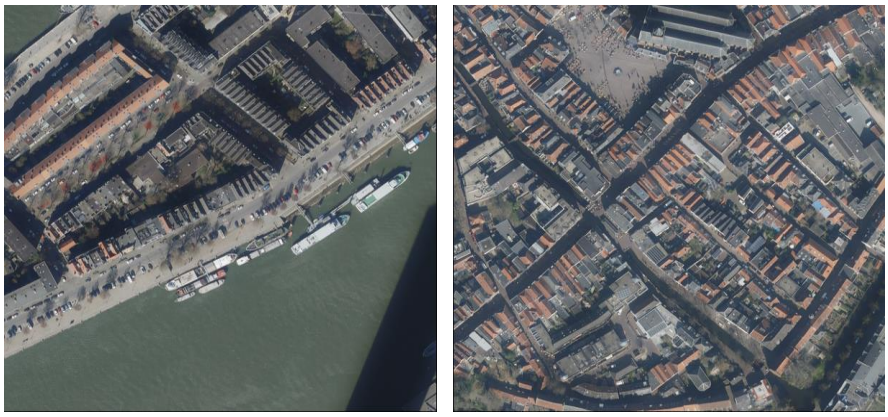
Tags: Countries and cities: 6 countries and 19 cities in Europe

Our problem setting: Class and tag imbalance

Evaluation of the models. The distribution of the *error*: Does the error have a bias towards old houses or new buildings?

Does the error have a bias towards specific cities and/or countries?

Example images: For the top-view image data



Center Crop (and/or downsampling) could be used in these top-view VHR images of the dataset.

2. Methodology

Development of the model. We design, train, and test the proposed Building Sustainability Model (BSM) that has as inputs the three modalities of the new MyCD dataset, namely (i) the street-view images, (ii) the aerial images, and (iii) the Sentinel-2 data.

We perform data fusion for the cross-view inputs at the feature level. In particular, we use an encoder for each of the three modalities and, then, we concatenate (i.e. concat) the latent features, thus performing late data fusion. The street-view images, as well as the aerial images, have three bands, i.e. RGB, while the Sentinel-2 data have 12 spectral bands.

The output of the model is the building construction epoch class, i.e. a classification task.

Flowchart diagram. The flowchart of the model is shown in Fig. 3.

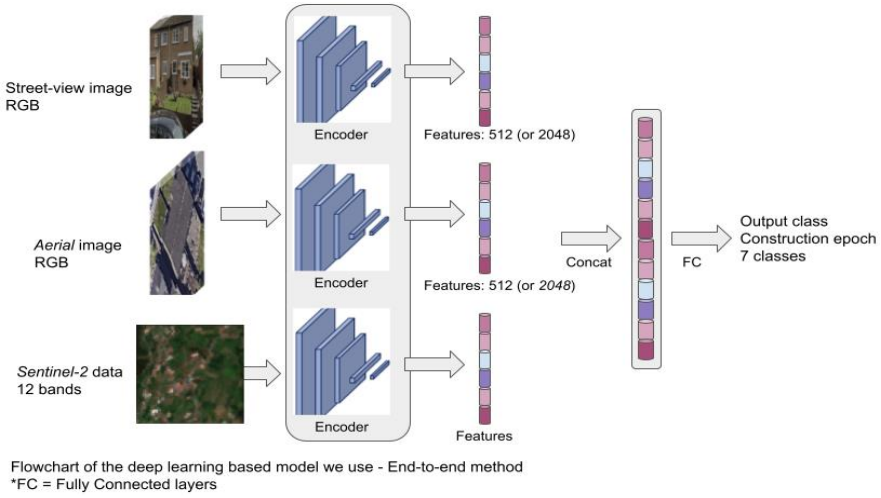


Fig. 3: Flowchart of the model, where we perform data fusion in the latent feature space.

Architecture. The encoder network for the ground/ street images is a ResNet-152 model that is pre-trained on the dataset ImageNet. The encoder for the aerial VHR images is also a ResNet-152 model that is pre-trained on ImageNet. The encoder for the Sentinel-2 data is a CNN Encoder taking as input all the 12 spectral bands at 10 m resolution. This encoder network for the multi-spectral images is randomly initialized, i.e. not pre-trained.

The model also uses data augmentation, horizontal flipping for street-view images and vertical and horizontal flipping for aerial VHR images.

We develop, train, and test the model in Fig. 3 and we note that for the accurate estimation of the construction epoch of buildings, the most important modality is the street-view images. In addition, the next most important cross-view input is the aerial VHR images as they provide useful global information (i.e. context) to the model.

3. Evaluation and experiments

We train the model using the optimizer Adaptive momentum (Adam) for several epochs, i.e. for 120 epochs. We also test the model and compare with other baseline models that are based on alternative decision choices.

3.1. BSM model evaluation. We train the proposed BSM model on the new MyCD dataset.

Results. The model is evaluated on the test set of the MyCD dataset. The test dataset contains images from cities not included in the training data. In this way, the generalization performance of the model is examined and assessed.

The accuracy on the test set is 58.47%, the precision is 58.22%, the recall is 58.47%, the F1-score is 58.15%, and the mean of the diagonal items of the confusion matrix is 58.44%. These results correspond to Fig. 4 where the confusion matrix is shown. In addition, these results are also summarized in Table 1.

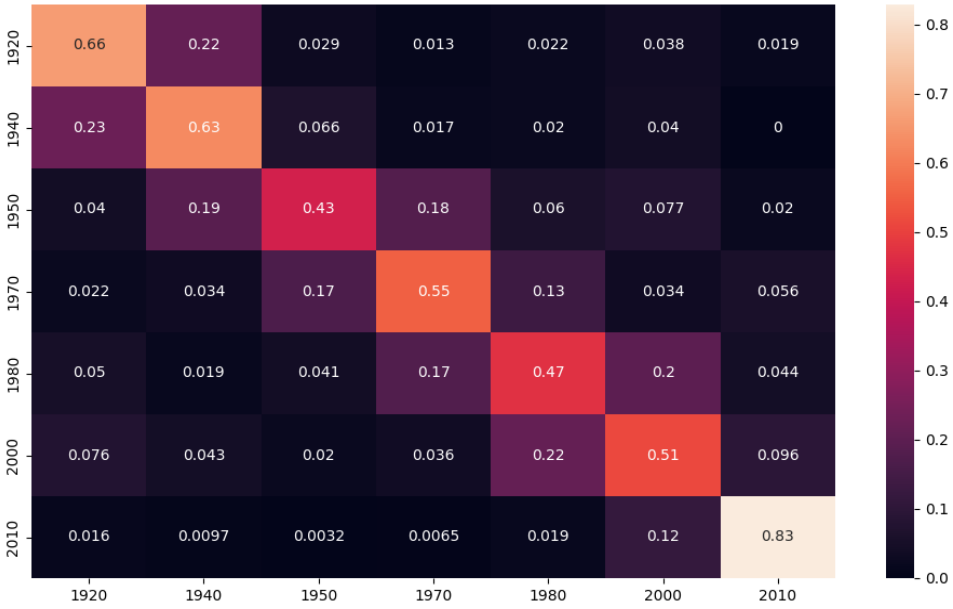


Fig. 4: Evaluation of the BSM model on the test set of the MyCD dataset using the confusion matrix, where we are mainly interested in the mean of the diagonal items.

BSM	Performance of model
Accuracy	58.47%
Precision	58.22%
Recall	58.47%
F1-score	58.15%
Mean of diagonals of confusion matrix	58.44%

Table 1: Results of the BSM model where the classification is for 7 classes for the building age.

The classification task we examine has 7 categories and the BSM model achieves an accuracy of 58.47%. We note that random chance/ guessing leads to an accuracy of approximately 14.29% (i.e. 1/7). In the following sections, we compare the proposed BSM model with other baseline models, also performing an ablation study. In particular, we examine the performance of models being trained (and tested) with and without aerial and Sentinel-2 images.

In addition, during testing/ inference only (i.e. and not during training), we also examine the performance of models using and not using street-view images. We perform such experiments because having street-view images of every single building of every city in Europe is not possible and, thus, does not scale. On the contrary, having aerial and Sentinel-2 images of buildings in Europe scales more favorably.

3.2. Comparison with testing without using the street-view images

During testing only (i.e. and not during training), we also examine the performance of models not using street-view images. We perform such experiments because having a street-view image of every building of every city in Europe (and in the world) during inference is not possible and, thus, does not scale. On the contrary, having aerial and Sentinel-2 images of buildings in Europe scales more favorably.

We evaluate the model and examine its performance on the test set of the MyCD dataset which has images of houses/ buildings from cities that are *not* included in the training data. When performing inference/ testing with only two modalities, i.e. using only aerial VHR images and Sentinel-2 data, it is possible to only use the two encoders for these two modalities. In particular, we use only two out of the three trained encoders, i.e. using the same training. The accuracy in this case on the test dataset that includes images of buildings from cities that are not included in the training data is 34.63%. The precision is 47.92%, the recall 34.63%, the F1-score 38.53%, and the mean of the diagonal items of the confusion matrix 26.40%. These results correspond to Fig. 5 where the confusion matrix is shown. In addition, these results are also summarized in Table 2.

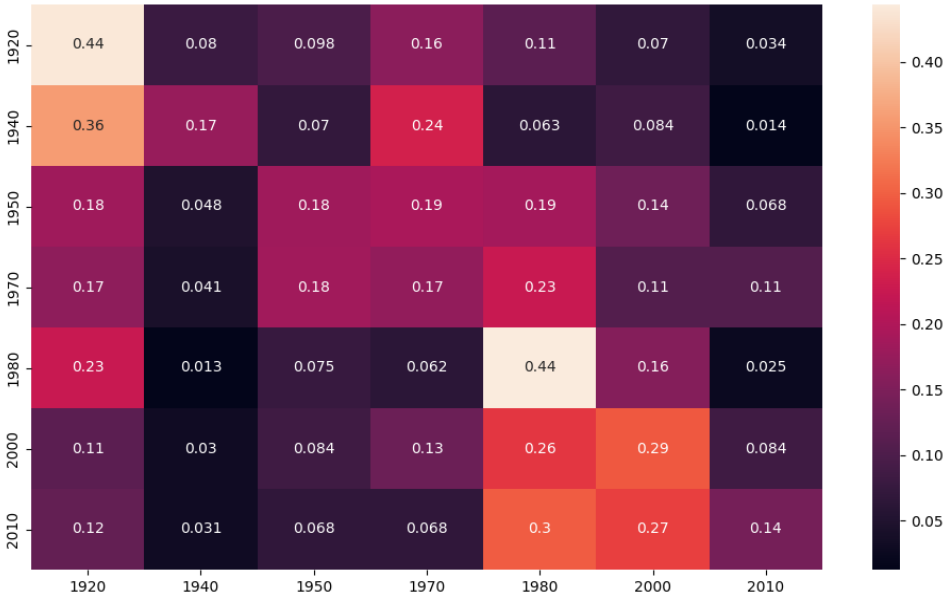


Fig. 5: Confusion matrix of the model that is tested using only two out of the three trained encoder networks.

BSM	Performance of model
Accuracy	34.63%
Precision	47.92%
Recall	34.63%
F1-score	38.53%
Mean of diagonals of confusion matrix	26.40%

Table 2: Evaluation results of the model that is tested here using only two out of the three trained encoders.

Further comparison with testing without using the street-view images. It is possible, for the testing without the street-view images, to use all the three trained encoders, by performing CutOut on the entire street-view image. The results for the testing without the street-view images are presented in Table 3. We note that during training, to emulate the testing conditions, we have trained on the three modalities of the MyCD dataset and we have performed the data augmentation method CutOut on the street-view images (set black colour) [1]-[3]. We have also done this on the entire image and not only on a part of the image, so as to be robust to inference without the street-view images. During training, we progressively increase the size of the CutOut data augmentation, for example like in curriculum learning, to ensure that testing without the modality of the street-view leads to good results. Then, after training the model, we perform the testing *without* the street-view images.

The accuracy of the model in this case is 43.40%. In addition, the precision is 48.87%, the recall 43.40%, the F1-score 45.16%, and the mean of the diagonal items of the confusion matrix 24.59%. Moreover, the confusion matrix is presented in Fig. 6.

Model	Performance of model
Accuracy	43.40%
Precision	48.87%
Recall	43.40%
F1-score	45.16%
Mean of diagonals of confusion matrix	24.59%

Table 3: Results of the model where the classification is for the 7 building age classes.

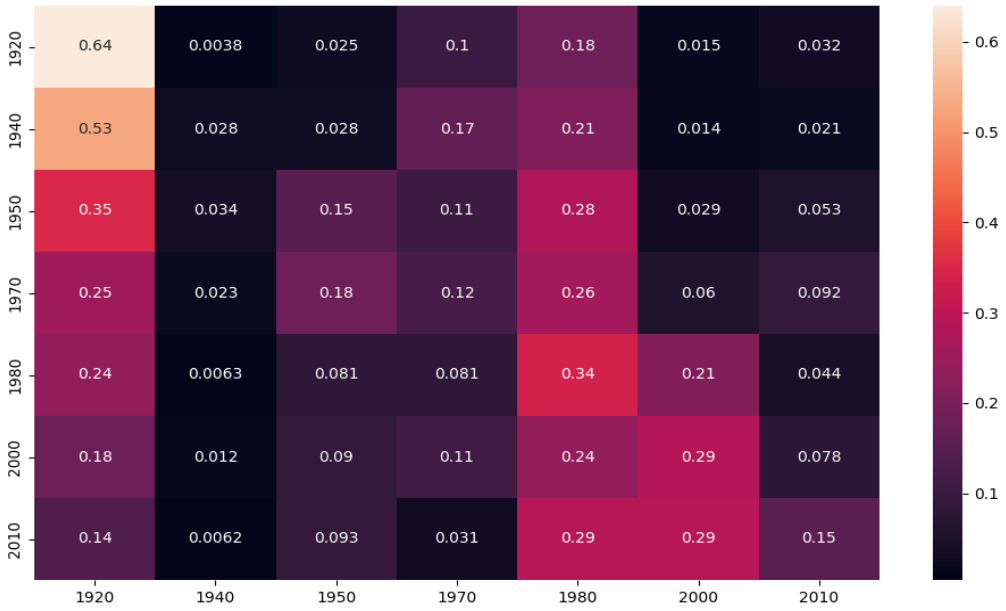


Fig. 6: Confusion matrix of the model when *not using* street-view images during testing.

3.3. Comparison with single-modality (i.e. street-view only) baseline model

Ablation study using only street-view images. Aiming at assessing the performance of our model, BSM, and the importance of the aerial and Sentinel-2 data on the estimation of the building construction epoch class, we have developed the BSM-s model that has a similar architecture as the BSM model and it is based *only* on street-view images. Hence, the BSL-s model is trained, validated, and tested only on the street-view images of the MyCD dataset.

BSM-s model evaluation. We design, train, and test the model BSM-s that has as inputs only the one modality, i.e. the street-view images, of the MyCD dataset.

Results. We evaluate the BSM-s model and we examine its performance on the test set of the MyCD dataset. We note that the test dataset has images of buildings from cities not included in the training data, so as to assess the generalization performance of the model. The accuracy of the BSM-s model on the test set is 54.78%, the precision is 54.31%, the recall is 54.78%, the F1-score is 54.40%, and the mean of the diagonal items of the confusion matrix is 54.72%, as also shown in Fig. 7 and also in Table 4.

In percentage terms, the performance improvement in the average of the diagonal items of the confusion matrix metric on the test dataset is approximately 6.71% when using all three modalities compared to using only the one modality of the street-view images.

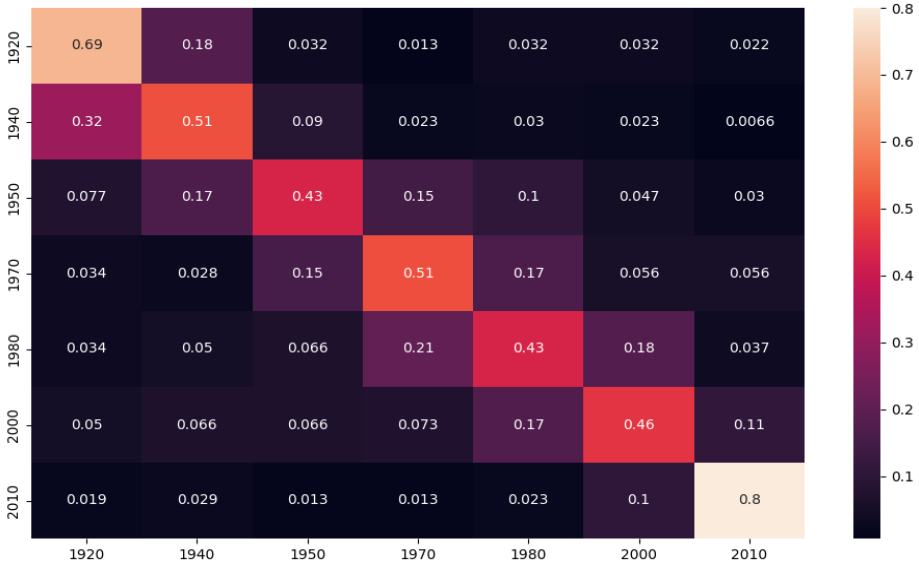


Fig. 7: Confusion matrix of the BSM-s model, where we are mainly interested in the mean of the diagonal items.

BSM-s	Performance of model
Accuracy	54.78%
Precision	54.31%
Recall	54.78%
F1-score	54.40%
Mean of diagonals of confusion matrix	54.72%

Table 4: Results of the BSM-s model where the classification is for the 7 building age classes.

4. Further suggestions to participants

Several different architectures and models can be trained and used to estimate the building construction epoch. In addition, for this problem, several different datasets (either labelled or unlabelled) and pre-training methods, either supervised or self-supervised (e.g. pre-text tasks), can also be used for model pre-training.

In addition:

(1) Because the models will be evaluated on left-out cities, the validation set can have a held-out city. In this way, it is possible to have a similar evaluation-like procedure in the validation set.

The validation set is used for hyper-parameter tuning and for early stopping.

(2) *Regression* rather than classification with the cross-entropy loss function can be performed, and this could potentially perhaps improve the model performance.

(3) Segmentation can be performed as a pre-processing step so that the training is done without not useful information, i.e. without *not helpful* information for the model. Segmentation to isolate the desired building can potentially improve the performance.

(4) Use of K-fold classification with, for example, K = 10 folds to improve performance.

Also, perform advanced regularisation techniques to prevent *overfitting* on the cities included in the training data, in order to be able to generalise to new/ previously unseen cities and/or countries in Europe.

(5) Address the issue of having a high error when the samples from the specific city or country are limited in number: Use of data augmentation mainly for the few-sample cities and/or countries.

Perform learned data augmentation using for example deep generative models, e.g. diffusion models, Generative Adversarial Networks (GAN), or flow-based models. Advanced data augmentation methods like generative models could improve the generalisation performance of the model.

Further data augmentation to be considered: Change/ alter the colours of the outside walls/ facade of the buildings.

Random/ stochastic data augmentation, e.g. Gaussian blur, colour jitter, and small-angle rotation, can also be performed. Here, *dense* random sampling can produce several samples for the limited-data cities and/or countries, while *less dense* random sampling can produce additional samples for the several-data cities.

(6) Use of model ensemble to improve performance and show that using aerial and satellite images has benefits in this setting.

Training an ensemble based on/ considering the correlations between countries, rather than just training a single model, might potentially lead to improved performance. See also (7) below.

(7) Use of city information during training to improve performance, i.e. the country and city information can be used during training.

Training different models and combining cities: Training a *model ensemble* by combining cities based on performing clustering in the feature space using a similarity metric to define the numerical criterion to combine cities, can improve performance. Here, using the t-SNE or the umap algorithms

to find correlations in the feature space between cities and/or countries in order to effectively combine cities can lead to benefits.

(8) Because of sometimes limited data for some cities and/or countries, consider the use of decision trees, including Random Forests. Such *machine learning* methods that are non-deep-learning-based models, i.e. non-neural-network-based, might show improved performance in the examined problem settings.

Logistic regression for classification can also be examined/ considered.

5. Appendix

Further evaluation of the BSM model. The accuracy of BSM on a different test dataset that includes images of buildings from cities that are included in the training data is 67.83%. Here, this test set has data from cities that are in the training data, while in Section 3.1, we examined the performance of the BSM model on the test set of the AI4EO Challenge that includes images of cities that are not in the training data. The precision in this case is 68.00%, the recall 67.84%, the F1-score 67.56%, and the mean of the diagonal items of the confusion matrix 64.03%. These results correspond to Fig. 8 where the confusion matrix is shown. In addition, these evaluation results are also summarized in Table 5.

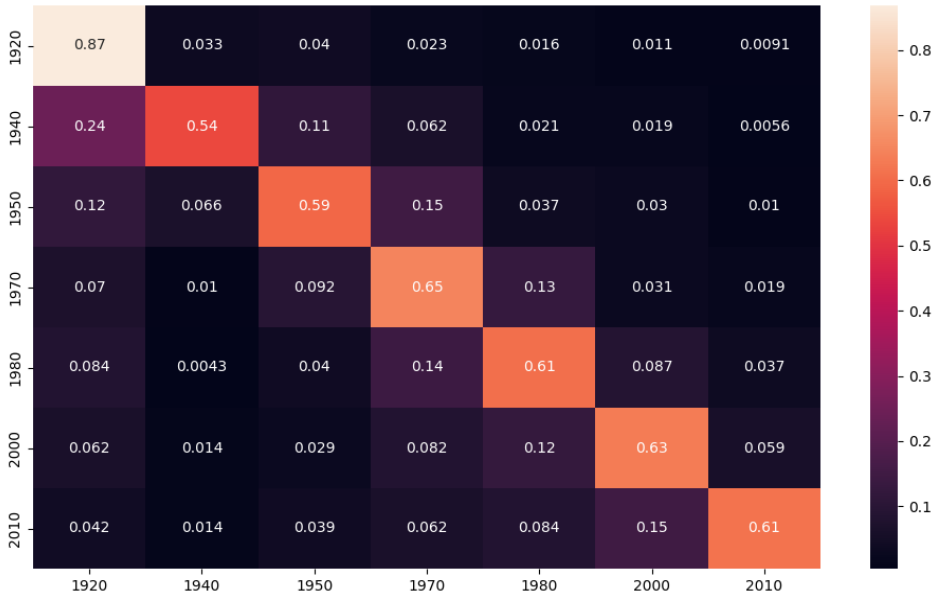


Fig. 8: Confusion matrix of the BSM model when tested on a different test dataset that includes images of houses/ buildings from cities that are included in the training data.

BSM	Performance of model
Accuracy	67.83%
Precision	68.00%

Recall	67.84%
F1-score	67.56%
Mean of diagonals of confusion matrix	64.03%

Table 5: Evaluation results of the BSM model on a different test set that includes images of buildings from cities that are included in the training data.

On this different test dataset that includes images of buildings from cities that are included in the training data, training and testing *only* on street-view images leads to an accuracy of 61.10%, precision 60.62%, recall 61.10%, F1-score 60.62%, and mean of diagonal items of confusion matrix 57.00%. Hence, in percentage terms, in the average of the diagonal items of the confusion matrix evaluation metric, the performance improvement is approximately 12.33% when using all three modalities of the MyCD dataset compared to using only a single modality.

Moreover, when comparing BSM with a model that is not based on any pre-trained network, the accuracy is 54.92%, the precision 54.34%, the recall 54.92%, the F1-score 54.33%, and the mean of the diagonal items of the confusion matrix 49.92%. Therefore, in percentage terms, in the average of the diagonals of the confusion matrix metric, the performance improvement is approximately 28.27% when using pre-trained networks compared to using only non-pretrained model architectures.

Frequently Asked Questions (FAQ):

(1) How can we contact you concerning the dataset?

The main contact emails are {Nicolas.Longepe, Bertrand.Le.Saux, and Nikolaos.Dionelis}@esa.int.

(2) Why do you use neural networks and deep learning models?

We would like to have a general framework for the problem and because we use neural networks for most of our problems, we would like to examine the effectiveness of deep learning models for this specific problem, i.e. the problem of the challenge. Participants are free to use any method, e.g. machine learning methods such as logistic regression or neural networks such as Residual Networks (ResNets).

(Regarding the *general* framework for the problem and using neural networks for most of our problems, we also develop Remote Sensing Foundation Models in our lab and this is why a deep neural network solution for the problem of the challenge could fit as a downstream task of such Foundation Models.)

ABOUT DOCUMENT: This document presents the model for the baseline method

This document is generated by Nikolaos Dionelis @ESA - LAST EDITED: 04/02/2024

References

[1] Terrance DeVries and Graham W. Taylor, "*Improved Regularization of Convolutional Neural Networks with Cutout*," arXiv:1708.04552, 2017. http://arxiv.org/pdf/1708.04552.pdf

[2] Kun He, Chang Liu, Stephen Lin, and John E. Hopcroft, "*Local Magnification for Data and Feature Augmentation*," arXiv:2211.07859, 2022. http://arxiv.org/pdf/2211.07859.pdf

[3] Shivang Agarwal, Jean Ogier du Terrail, and Frederic Jurie, "*Recent Advances in Object Detection in the Age of Deep Convolutional Neural Networks*," arXiv:1809.03193, 2019. http://arxiv.org/pdf/1809.03193.pdf