**Building Sustainability**

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 images and the Sentinel-2 data in addition to the street-view images, as this improves performance.

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**



## 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 building/ house under study. The labels are the corresponding construction epoch class.

The MyCD dataset comprises data for 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.

## 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. 1.

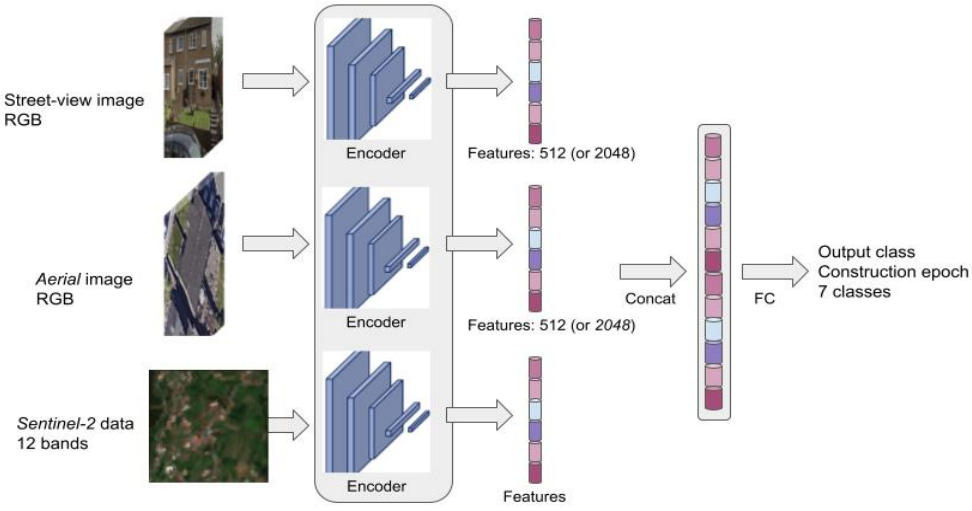


**Fig. 1:** Flowchart of the model, where we perform data fusion in the latent feature space.

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

## 3. Evaluation

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.

### Experiments

**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 58.42%, the precision is 58.17%, the recall is 58.42%, the F1-score is 58.10%, and the mean of the diagonal items of the confusion matrix is 58.39%. These results correspond to Fig. 2 where the confusion matrix is shown. In addition, these results are also summarized in Table 1.

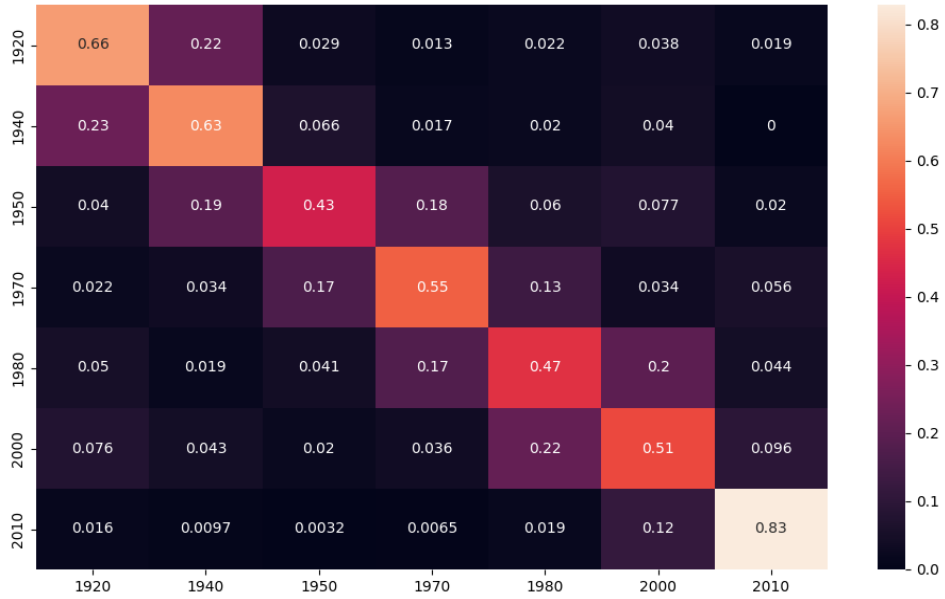


**Fig. 2:** Evaluation of the BSM model on the test set of the MyCD dataset using the confusion matrix, where we are mainly interested in the mean of the diagonal items.

| BSM | Performance of model |
|---|---|
| **Accuracy** | 58.42% |
| **Precision** | 58.17% |
| **Recall** | 58.42% |
| **F1-score** | 58.10% |
| **Mean of diagonals of confusion matrix** | 58.39% |

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

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. The results for the testing without the street-view images using however all the three trained encoders, by performing CutOut on the entire street-view image, are presented in Table 2. 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 56.21%. In addition, the precision is 55.77%, the recall 56.21%, the F1-score 55.80%, and the mean of the diagonal items of the confusion matrix 56.13%. Moreover, the confusion matrix is presented in Fig. 3.

| Model | Performance of model |
|---|---|
| **Accuracy** | 56.21% |

| | |
|---|---|
| **Precision** | 55.77% |
| **Recall** | 56.21% |
| **F1-score** | 55.80% |
| **Mean of diagonals of confusion matrix** | 56.13% |

**Table 2:** Results of the BSM model where the classification is for the 7 building age classes.

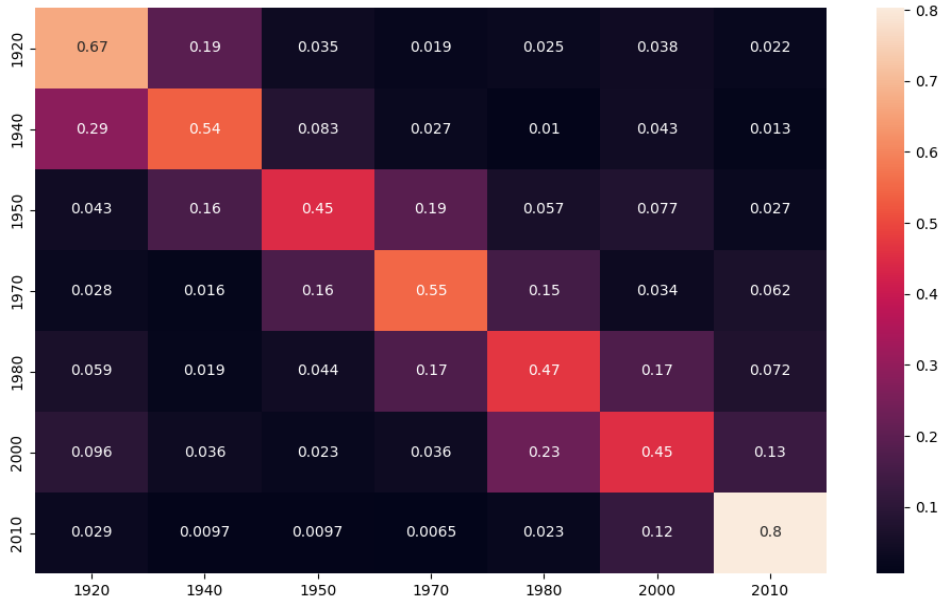


**Fig. 3:** Confusion matrix of the model when not using street-view images during testing.

### 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 shown in Fig. 4 and also in Table 3.

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.

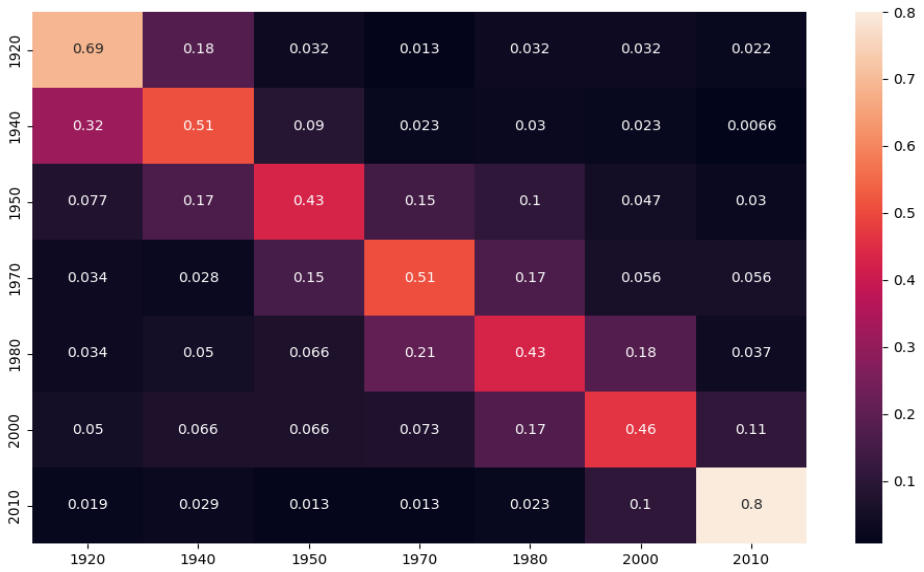


**Fig. 4:** Confusion matrix of the BSM-s model, where we are mainly interested in the mean of the diagonal items.

| 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 3:** 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.

**Further comparison with testing without using the street-view images.** It is possible, for the testing without the street-view images, to use only two out of the three trained encoders, i.e. using the same training. In particular, 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 the aerial VHR images and the Sentinel-2 data.

## 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. 5 where the confusion matrix is shown. In addition, these results are also summarized in Table 4.

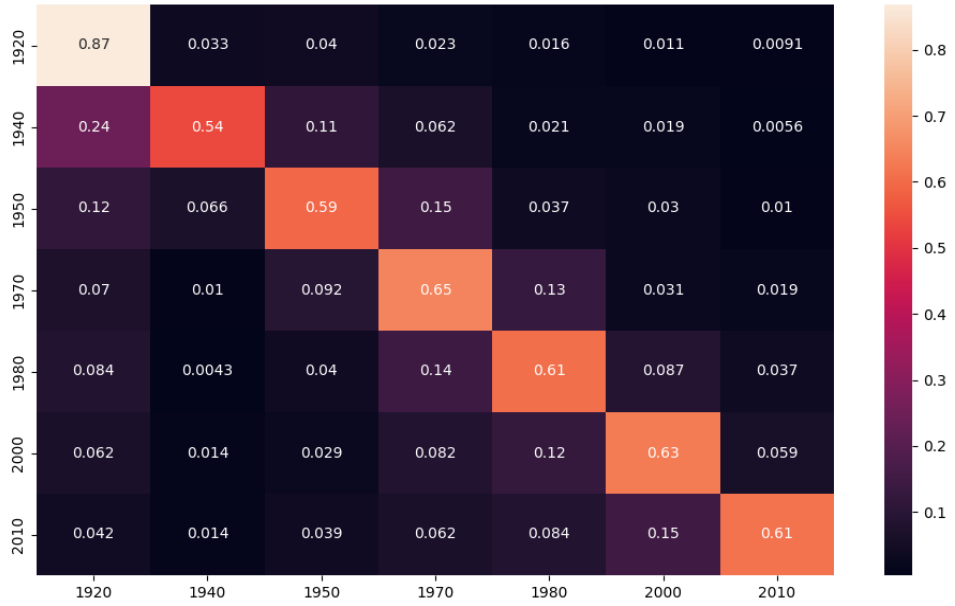


**Fig. 5:** Confusion matrix of the BSM model when tested on a different test dataset that includes images of houses/ buildings from cities that are included in the training data.

| 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 4:** 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 on the test dataset is approximately 12.33% when using all three modalities compared to using only a single modality.

## 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