Towards a Multi-Temporal Deep Learning Approach for Mapping Urban Fabric Using Sentinel 2 Images

Lamiae El Mendili 1, Anne Puissant 2,* Mehdi Chougrad 1 and Imane Sebari 1

1 School of Geomatics and Surveying Engineering, IAVH2, 10101 Rabat, Morocco; lamiae.elmendili@iav.ac.ma (L.E.M.); mehdi.chougrad11@iav.ac.ma (M.C.); i.sebari@iav.ac.ma (I.S.)
2 Laboratoire Image, Ville, Environnement—LIVE/CNRS UMR 7362, Department of Geography, University of Strasbourg, 3 rue de l’Argonne, 67000 Strasbourg, France
* Correspondence: anne.puissant@live-cnrs.unistra.fr

Received: 27 November 2019; Accepted: 24 January 2020; Published: 28 January 2020

Abstract: The major part of the population lives in urban areas, and this is expected to increase in the future. The main challenges faced by cities currently and towards the future are the rapid urbanization, the increase in urban temperature and the urban heat island. Mapping and monitoring urban fabric (UF) to analyze the environmental impact of these phenomena is more necessary than ever. This coupled with the increased availability of Earth observation data and their growing temporal capabilities leads us to consider using temporal features for improving land use classification, especially in urban environments where the spectral overlap between classes makes it challenging. Urban land use classification thus remains a central question in remote sensing. Although some research studies have successfully used multi-temporal images such as Landsat-8 or Sentinel-2 to improve land cover classification, urban land use mapping is rarely carried using the temporal dimension. This paper explores the use of Sentinel-2 data in a deep learning framework, by firstly assessing the temporal robustness of four popular fully convolutional neural networks (FCNs) trained over single-date images for the classification of the urban footprint, and secondly, by proposing a multi-temporal FCN. A performance comparison between the proposed framework and a regular FCN is also conducted. In this study, we consider four UF classes typical of many European Western cities. Results show that training the proposed multi-date model on Sentinel 2 multi-temporal data achieved the best results with a Kappa coefficient increase of 2.72% and 6.40%, respectively for continuous UF and industrial facilities. Although a more definitive conclusion requires further testing, first results are promising because they confirm that integrating the temporal dimension with a high spatial resolution into urban land use classification may be a valuable strategy to discriminate among several urban categories.

Keywords: urban fabric; Sentinel-2; deep-learning; multi-temporal approach

1. Introduction

In 2018, 55% of the world’s population lived in urban areas against only 30% in the 1950s. This value is expected to reach 68% in 2050 [1]. This growing urbanization makes urban areas very dynamic and changes the way people live, consume and exploit resources. For many years, it was monitored through mapping the urban footprint, which includes the road network, buildings, vegetation, and impervious surfaces. These elements structure the spatial layout of cities resulting in various urban fabrics (UF) [2]. UF can significantly affect the energy balance of the Earth’s surface, resulting in the urban heat island effect [3], and can also affect natural ecosystems through land use and land cover fragmentation [4]. Mapping UF is thus crucial for understanding and simulating urban dynamics [5].
The increasing availability of remote sensing data has greatly facilitated land use mapping. However, it is still challenging to ensure its accuracy [6], especially in urban environments where impervious materials have similar spectral properties. Spectral information alone is insufficient for this purpose. Consequently, object-based image analyses are preferred because of their ability to include contextual information. Numerous studies have used supervised object based methods to map land cover and land use [7]. However, these approaches usually require very high-resolution images since they use segments as processing units to compute spectral, geometrical, and spatial features to classify [6,8]. Even if these methods deliver satisfactory results, they are still limited as they are data-dependent and require an adequate set of features for each application.

Since recently, several papers have shown the efficiency of multi-temporal imagery to improve classification provided land use have time-variant properties [9–14]. For urban context, the urban elements spectral properties in each urban fabric classes vary along seasons due to different occurrences of each land cover element (vegetation, bare soil, building, road networks, and impervious surfaces). These patterns are already visible on high spatial resolution (10 m) images and the spectral variation can thus be characterized by temporal features [15]. However, unlike for crop mapping, the temporal properties we want to exploit are not seasonal patterns specific to each class but rather are related to a class’s temporal stability, which depends on its land cover elements. The presence/absence of vegetation along with the relative density of buildings make some classes more susceptible to change and others more stable. For instance, the class of discontinuous UF (Figure 1a) composed of buildings and other impervious surfaces with green areas and bare soil, is more susceptible to seasonal changes. However, other types of UF are less prone to change and will have more stable temporal properties such as dense and continuous UF (Figure 1b). This is the case when urban structures and transport networks are dominating the surface area. Based on this expert and qualitative analysis, we make the hypothesis that a classification algorithm capable of retrieving spatio-temporal features could be efficient in mapping UF compared to a classical supervised approach based on very high spatial resolution.

![Figure 1](image_url)

**Figure 1.** The temporal variation of two classes of urban fabric from winter to spring. (a) Discontinuous urban fabric shows important changes mainly due to the presence of vegetation; (b) continuous urban fabric remains almost the same from winter to spring due to the presence of buildings and the scarcity of vegetation.

Integrating the temporal information into a classification framework remains challenging. In the literature, several approaches exist. The first technique consists in applying a machine learning classifier to a stack of multi-temporal images. The second approach involves handcrafting adequate temporal features and then feeding them to a machine learning classifier [6]. These features can either be pixel
statistics (mean, minimum, etc.) or other metrics obtained from the time series [16]. This technique was used [12] to distinguish between different land cover classes by computing various statistical and phenological values. Even though these methods deliver satisfactory results, they are data-dependent as they require an adequate set of features for each classification. Other techniques commonly used are based on time series analytics [17]. They operate by using a classifier, for instance, the nearest neighbor classifier (NN), along with a similarity measure like the dynamic time warping distance (DTW) [18]. These techniques are, however, computationally expensive.

The shortcomings of these techniques make them less competitive in comparison with approaches capable of learning high-level features end-to-end like deep learning. In computer vision, convolutional neural networks (CNNs) are very good at analyzing images by learning high-level spatial features [19]. CNNs have been successfully applied to remote sensing data for tasks like scene classification [20], semantic segmentation [21], and object detection [22]. Ma et al. [23] offers a thorough review of deep learning applications in remote sensing. Encoder decoder architectures are the current state-of-the-art semantic segmentation models due to their numerous advantages like easier implementation, higher accuracy, and less computation complexity [24]. Numerous studies explored the use of encoder-decoder architectures for the mapping of urban environments. Audebert et al. [25] trained a variant of the SegNet [26] architecture over an urban area using aerial images and other heterogeneous sensors. Zhang et al. [27] distinguished between buildings, roads, water, and vegetation by implementing a multi-scale deep learning model based on the famous UNet architecture. Fu et al. [22] performed detailed urban land use mapping by improving FCN-8s, introducing Atrous convolutions and refined the model’s output using conditional random fields. This approach outperformed standard object-based image analysis techniques.

Although CNN’s are well-suited to processing 2D data, such as images, they are not able to manage sequential data like multi-temporal images or times series. Conversely, recurrent neural networks, and particularly LSTMs, can process such data. However, they are not adequate for our purposes, giving that land use mapping aims at producing one classification map per series [28]. Various studies investigated the use of CNNs for processing multi-temporal data. Benedetti et al. [29] proposed a deep learning framework for the fusion of multi-temporal Sentinel 2 imagery and very high-resolution satellite imagery to distinguish between various vegetation classes. Mauro et al. [30] used a multi-layer perceptron (MLP) and a CNN to classify land cover from multi-temporal Landsat images. Nogueira et al. [31] implemented a multi-branch CNN for vegetation mapping and confirmed its superiority to a traditional CNN operating on a temporal stack of images. Pelletier et al. [28] proposed a temporal CNN for crop classification where convolutions are applied in the temporal domain. The literature review reveals that few studies investigated the temporal dimension of high-resolution satellite imagery such as Sentinel-2 (10 m) for mapping UF. The use of very high-resolution images is often preferred given the complexity of urban environments and the great spectral overlap.

In this context, the goal of this study is two-fold: (i) to investigate whether UF mapping benefits from integrating temporal information; and (ii) to explore the use of CNNs on multi-temporal satellite imagery. More precisely, this study evaluates the potential contribution of temporal dimension of freely available high-resolution satellite imagery such as Sentinel-2 to map four urban classes: Continuous UF, discontinuous UF, industrial or tertiary facilities and the road and rail network. These are all classes with different vegetation and buildings densities and are thus well suited for this study. In this paper, we work on a multi-temporal sampling of a time series as a first step. The proposed framework can be later on extended to a full time series. This approach, exploiting the temporal dimension to compensate for the spatial resolution explores the use of CNNs which will provide us with an end-to-end learning framework without the need of engineering features (the process of using our knowledge of the data to create features that make machine learning algorithms work), often time consuming for experts. The paper is structured in five sections. Section 2 describes material and methods. First, a comparison baseline from mono-temporal images is presented by analyzing the performance of four encoder-decoder architectures for mapping the urban footprint. Secondly, based on
the best model, we perform UF mapping and we design a multi-temporal classifier by altering the previous model. Results (Section 3) are then detailed and allow assessing the benefits of using temporal data to discriminate UF. A discussion and conclusion are then provided respectively in Sections 4 and 5.

2. Material and Methods

2.1. Study Site and Reference Data

The study area is the Alsace region located in the East of France (Figure 2a). It covers an area of 8329 km$^2$ and includes four main land cover classes: urban footprint, agricultural areas, forest areas, and water surfaces. The urban footprint displays some level of class imbalance (Figure 2b). Discontinuous UF is the most common class, followed by industrial or tertiary facilities and then the road and rail network. The key reason for choosing the Alsace region as a study site is its morphological characteristics typical of many European Western cities.

![Figure 2. (a): Study area: (a) the Alsace region of France; (b) the percentage of each class in the dataset.](image)

The regional land use/cover vector database (BDOccsol©GeoGrandEst) produced by visual interpretation of aerial photography is used as reference data [33]. The most up-to-date version (2012) available through the regional data infrastructure GeoGrandEst, maps the study area into 55 thematic classes. The legend is organized into four levels of nomenclature. The first level categorizes land cover into four classes: (1) artificial surfaces, (2) agricultural areas, (3) forest areas, and (4) water surfaces. Artificial surfaces (Level 1) are further sub-divided into 12 classes at Level 3. Table 1 details the four generic urban thematic classes obtained after generalization.

Both continuous and discontinuous UF contain buildings and impervious surfaces. The main difference consists in their relative density and the importance of vegetated areas and bare surfaces. In continuous UF, more than 80% of the surface is covered by impervious elements (buildings, roads, and artificially surfaced areas). For discontinuous UF, roads and artificially areas range from 30% to 80%. This also presents for industrial facilities due to industrial wastelands and queries and the road and rail network due to the presence of open areas. These variations make these classes well...
suited for this study as we can already expect the densely built classes to have a more stable temporal response compared to the others (Figure 1). Reference data are rasterized at a 10 m spatial resolution. This process is applied at two levels: the first level with a binary map of the urban footprint, and the second level results in a multi-class map with the four UF classes (Figure 3—Step 1).

| Final Classes | Landcover Classes (Level 3) BDOccsol©GeoGrandEst | Description |
|---------------|-------------------------------------------------|-------------|
| (1) Continuous UF | Continuous housing blocks | Areas mainly occupied by buildings and other impervious surfaces. Vegetation and bare soil are exceptional |
| (2) Discontinuous UF | Discontinuous housing blocks | Areas where buildings, artificial surfaces, vegetated areas and bare surfaces are present and occupy significant surfaces in a discontinuous spatial pattern |
| (3) Industrial facilities | Commercial and industrial facilities, industrial wastelands, gravel quarries and sandpits, mining sites and construction sites. | Areas mainly occupied by industrial activities (trade, manufacturing, service and tertiary or extractive activities). |
| (4) Road and rail networks | Rail and road networks, airport and port sites, and other open associated areas. | Primary road network and associated installations (stations, platforms and parking areas), railways and train stations, areas associated with airport and port activities and buildings intended for agricultural use. |

### 2.2. Optical Satellite Data and Pre-Processing Steps

The Alsace region is covered by three Sentinel 2 tiles (T32ULU, T32ULV, and T32TLT) downloaded from the Theia Land Data Center (Figure 1a). Sentinel 2 is a satellite mission of the European earth surveillance program, Copernicus. It provides multispectral images with four bands at 10 m (blue, green, red, and near infrared), six bands at 20 m (four vegetation red edge bands and two short wave infrared bands) and three bands at 60 m (coastal aerosol, water vapor, and short wave infrared cirrus) [33]. In this paper, we use only the 10 m, spectral bands, as they are more adapted for UF mapping. Depending on the cloud cover, all images are selected to be as representative as possible of the year 2017, i.e., by varying the sensing seasons in order to add seasonal variations to our training data. Table 2 shows the distribution of the sensing periods.
Table 2. Image acquisition dates of the Sentinel 2 remote-sensing data.

| Sentinel 2 Tiles | Acquisition Dates |
|------------------|-------------------|
| T32ULU           | 02/2017           |
| T32ULV           | 05/2017           |
| T32TLT           | 11/2017           |
|                  | 02/2018           |
|                  | 04/2018           |
|                  | 08/2018           |

Image Normalization

All the images are normalized to make the learning process more stable by reducing the variability of training data and thus improving the generalization power of the model (Figure 3—Step1). This preprocessing step consists of computing the mean and the standard deviation of each channel separately and then normalizing across it:

$$z_i = \frac{x_i - \mu_i}{\sigma_i},$$ (1)

where $i$ represents the i-th channel, $z_i$ the pixel radiometric value after normalization, $x_i$ its value before normalization, $\mu_i$ the mean value of the channel $i$ and $\sigma_i$ its standard deviation.

2.3. Method

2.3.1. Problem Formulation

The purpose of this work is to study the benefit of using multi-temporal imagery for the classification of the urban fabric. For the classes of interest, the temporal dimension can supply additional information as certain classes experience seasonal variations, and others do not. In this context, temporal unstableness is as relevant as stableness to differentiate between the classes.

The first phase of our experiments aims at proposing a comparison baseline from mono-temporal images. The urban footprint is map (Figure 3—Step2) using four CNN models and the best architecture is selected to perform UF classification. Models are evaluated on test sets covering different geographical regions and different dates to assess their spatial robustness and temporal variations. The second step of our experiments (Figure 3—Step3) concerns mapping UF using multi-temporal imagery. Table 3 summarizes the different datasets used for training the three models: a mono-temporal model for (1) the classification of urban footprint, (2) the classification UF, and (3) a multi-temporal model for the classification of UF. The flowchart in Figure 3 details the proposed methodology to assess the best (mono or multi-temporal) model to map UF.

Table 3. The data used for training each model.

| Model                                           | Training Data                                      |
|-------------------------------------------------|---------------------------------------------------|
| (1) Mono-temporal model for the binary classification of urban footprint | One February tile (Figure 4—Step1 Urban footprint Train) |
| (2) Mono-temporal model for the classification of UF | Three February tiles (Figure 4—Step1 UF mono Train)   |
| (3) Multi-temporal model for the classification of UF | Three multi-temporal tiles (Figure 4—Step1 UF multi Train) |
2.3.2. Data Sampling

It consists of subdividing the image into 256 × 256 pixel patches (Figure 3—Step1). The size should be sufficiently large as not to lose spatial context but should also take into account the available computational resources. Images with no spectral information are then removed. We perform a random shuffling of the dataset before splitting it into training and validation sets using a 10% ratio for validation and 90% for training. Data augmentation on the training set is also applied by random rotations, translations, and zooms to make our models more robust to variations and prevent overfitting. Table 4 shows the number of image patches obtained for each model. The number of patches is proportional to the distribution of defined classes on the study site.

Table 4. Dataset size per UF class.

| Class               | Number of Patch (256 × 256) |
|---------------------|-----------------------------|
| Urban footprint     | 609                         |
| Continuous UF       | 1028                        |
| Discontinuous UF    | 1463                        |
| Industrial facilities | 888                      |
| Road and rail networks | 1575                 |

2.3.3. Choice of Model Architectures

It is essential to discuss how to approach multi-label classification. Two common strategies are often used: the first one consists in implementing a global model that outputs membership probabilities of all the classes of interest, whereas the second strategy, commonly referred to as the one-vs-all
Supervised classification of urban footprint using mono-temporal Sentinel 2 data.

Four distinct encoder-decoder architectures are implemented: FCN-32s [36], FCN-16s [36], SegNet [26], and UNet [37]. Unlike the original paper, we use VGG-19 [38] as the basis of the first two architectures. We initialize all encoders from the ImageNet weights and all decoders using Xavier Initialization [39]. According to [40], when changing sensors, transfer learning can fail. We, therefore, employ Sentinel 2 RGB bands analogously to the ImageNet dataset. We nonetheless train from scratch a variant of the SegNet model on multispectral data to explore the impact on classification performance. As a loss function, we use the dice loss (DSC) defined below. We then conduct a performance comparison between the implemented models to select the best model for UF mapping.

\[
DSC = \frac{2|Y \cap \hat{Y}|}{|Y| + |\hat{Y}|}
\]  

(2)

where DSC is the dice loss, Y is the ground truth, and \( \hat{Y} \) is the model prediction. We use Mini-batch gradient descent to optimize the objective function. To mitigate overfitting, we use numerous regularization techniques, including dropout and real-time data augmentation, by applying random rotations, translations, and zooms. We tune the dropout value and the batch size on the validation set, and we set an initial value for the learning rate using the method of [41]. We then reduce by 10% whenever the validation loss stops improving over more than five epochs of training. We train all models for 150 epochs. The outputs are class membership probabilities in the range 0 through 1. At this stage, we use 50% as a threshold to transform the results into a classified map.

Supervised classification of UF using mono and multi-temporal Sentinel 2 data.

The previous experiment showed that the UNet model offers the best performance. This will be further detailed in the results section. The model is then trained on mono-temporal data to map the four classes of UF. Afterward, we design a new architecture taking as a template the previous one to analyze multi-temporal data. The proposed architecture (Figure 4) takes as input a multi-temporal patch and outputs a classified map, and is capable of learning spatio-temporal features. The encoder consists of three VGG-16 branches carrying out the separate learning of spatial features for each timestamp. Then, a temporal-wise concatenation of the last convolutional layer is performed along the three branches for each VGG-16 block. The spatio-temporal features are then passed to the decoder via skip connections. The result of the last concatenation is up-sampled and further processed by the decoder. Finally, the output is fed to a sigmoid classifier to produce class membership probabilities. Thus, starting from a multi-temporal patch, our model outputs a single classification map.
Figure 4. The proposed architecture to perform multi-temporal learning.

We initialize and freeze the encoders using the corresponding mono-temporal weights learned previously. This aims at reducing the number of trainable parameters and reducing the risk of overfitting. We used a combination of the dice loss and the binary cross-entropy to train the model [43].

\[ L = \alpha H - (1 - \alpha)(1 - DSC) \]  

(3)

where DSC is the dice loss, \( Y \) is the ground truth, \( \hat{Y} \) is the model prediction, we set \( \alpha \) to a value of 0.7, and \( H \) is the binary cross-entropy function. We train the models for 300 epochs each. To define a threshold, we put into practice a systematic approach that consists in retrieving the break-even point defined as the point where the recall of the model equals its precision. This value represents an adequate choice given the Trade-off between recall and precision; a good model should, therefore, be able to compromise between these two quantities. Figure 5 below shows the precision recall curve for each classification model.

Figure 5. Precision recall curve per model.

Due to computational constraints, we train the multi-temporal UNet on two classes. To best assess the potential of temporal information for UF mapping, we chose a class with a large percentage of land covered by artificially surfaced areas, and another one with a small percentage. This is the case for continuous UF where the ground is almost entirely covered by buildings and the class of industrial facilities where vegetation is present.
To measure models accuracy, we use the intersection over union/Jaccard index (IoU) [43], overall accuracy (OA), and the confusion-matrix metrics that include precision, recall, and the F1 score. We also use the receiver operating characteristic (ROC) curve and the area under the ROC curve (AUROC) [44] as other performance indicators.

3. Experimental Results

Section 3.1 presents and discusses the results of the comparison between the four models for urban footprint classification, and Section 3.2 presents the results of UF classification using mono and multi-temporal images.

3.1. Results of Urban Footprint Classification

The loss function is plotted against time for each model and for both training and validation sets to monitor any possible overfitting (Figure 6). It appears that for both training and validation sets, UNet and RGB SegNet reached approximately a dice loss of -85% compared to a value approaching -75% for other models. Furthermore, the FCN-32s and the multi-spectral SegNet are prone to overfitting, which is shown by the increasing gap between the training loss and the validation loss. To make a more meaningful comparison, we plot the learning curves for each model on the same graph (Figure 7).

(a) FCN-32s  
(b) FCN-16s  
(c) SegNet  
(d) Multispectral SegNet

Figure 6. Cont.
We examined in more depth the variations of their sensitivity and specificity using the ROC curve. We will no longer consider the second one for the rest of the experiments. These findings support the importance of transfer learning for small datasets.

Another interesting remark is that the performance of the SegNet model trained on multi-spectral data is reduced compared with its RGB counterpart. Given the wide gap between the two variants of SegNet, we will no longer consider the second one for the rest of the experiments. These findings support the importance of transfer learning for small datasets. To further analyze the model’s performance, we examined in more depth the variations of their sensitivity and specificity using the ROC curve (Figure 8).

| Model   | IoU   | Precision | Recall | F1     | Kappa |
|---------|-------|-----------|--------|--------|-------|
| FCN-32s | 0.94  | 0.7279    | 0.8323 | 0.7766 | 0.7288|
| FCN-16s | 0.93  | 0.6348    | 0.8323 | 0.7766 | 0.7288|
| SegNet  | 0.99  | 0.9294    | 0.8283 | 0.8759 | 0.8538|

UNet achieved the best results with an F1 score of over 91%. As a last experiment, we conducted multispectral SegNet. Given the wide gap between the two variants of SegNet, we will no longer consider the second one for the rest of the experiments. These findings support the importance of transfer learning for small datasets. To further analyze the model’s performance, we examined in more depth the variations of their sensitivity and specificity using the ROC curve (Figure 8).

Figure 6. Training and validation learning curve per model: (a) FCN-32s; (b) FCN-16s; (c) SegNet; (d) multispectral SegNet; (e) UNet.

Figure 7. Learning curve per model.

UNet and RGB SegNet outperformed their counterparts in that their loss decreases more steeply and reaches lower values. Furthermore, it appears that the UNet model is holding a slight advantage. Another interesting remark is that the performance of the SegNet model trained on multi-spectral data is reduced compared with its RGB counterpart. Given the wide gap between the two variants of SegNet, we will no longer consider the second one for the rest of the experiments. These findings support the importance of transfer learning for small datasets. To further analyze the model’s performance, we examined in more depth the variations of their sensitivity and specificity using the ROC curve (Figure 8).

Figure 8. ROC curve per model.
UNet and SegNet outperformed the other models with a slight edge for the UNet. As an example, to reach 86% of true positives, UNet and SegNet only produce 1% of false positives compared with 7% and 10% for the FCN-16s and FCN-32s, respectively. The area under the ROC curve (AUROC) is also a meaningful indicator of a model’s performance. Given that each ROC curve must depart as much as possible from the diagonal, AUROC must be as close as possible to 1 and as far as possible from 0.5. Table 5 below shows the AUROC values for each model.

Table 5. AUROC values per model.

| Model   | AUROC |
|---------|-------|
| FCN-32s | 0.93  |
| FCN-16s | 0.94  |
| SegNet  | 0.99  |
| UNet    | 0.99  |

These results are very much in-line with those in Figures 8 and 9 with an on-going advantage for UNet and SegNet. Next, we examined the performance of each model on the validation set in terms of various evaluation metrics (Table 6).

Table 6. Evaluation metrics over the validation set per model.

| Model   | IoU    | Precision | Recall   | F1       | Kappa Coefficient | OA     |
|---------|--------|-----------|----------|----------|-------------------|--------|
| FCN-32s | 0.6303 | 0.7042    | 0.8572   | 0.7732   | 0.7078            | 0.8975 |
| FCN-16s | 0.6348 | 0.7279    | 0.8323   | 0.7766   | 0.7288            | 0.9210 |
| SegNet  | 0.7773 | 0.9294    | 0.8283   | 0.8759   | 0.8538            | 0.9617 |
| UNet    | 0.8437 | 0.9194    | 0.9111   | 0.9152   | 0.8941            | 0.9662 |

UNet achieved the best results with an F1 score of over 91%. As a last experiment, we conducted more rigorous tests enabling us to assess each model’s generalization ability. For this purpose, we constructed two test sets. In test set 1 (DiffT), we kept the same geographical zone while changing the sensing time. In contrast, we did the opposite for test set two (DiffZ), i.e., a different geographical zone while keeping sensing time unchanged. Results are reported in Table 7.

Table 7. Evaluation results on two test sets.

| Metrics   | UNet | SegNet | FCN-16s | FCN-32s |
|-----------|------|--------|---------|---------|
|           | DiffT| DiffZ  | DiffT   | DiffZ   | DiffT   | DiffZ   | DiffT   | DiffZ   |
| IoU       | 0.5106 | 0.5681  | 0.4721  | 0.5513  | 0.3439  | 0.4240  | 0.4369  | 0.5344  |
| Precision | 0.7039 | 0.7226  | 0.5915  | 0.8000  | 0.3823  | 0.4837  | 0.5231  | 0.6564  |
| Recall    | 0.6504 | 0.7265  | 0.7005  | 0.6395  | 0.7740  | 0.7747  | 0.7271  | 0.7418  |
| F1 score  | 0.6760 | 0.7245  | 0.6414  | 0.7108  | 0.5117  | 0.5956  | 0.6085  | 0.6965  |
| Kappa coefficient | 0.6465 | 0.6848  | 0.5999  | 0.6742  | 0.4401  | 0.5238  | 0.5595  | 0.6496  |
| OA        | 0.9485 | 0.9304  | 0.9251  | 0.9342  | 0.8588  | 0.8712  | 0.9106  | 0.9184  |

The performance is significantly better for the DiffZ test set for all the classification models with up to 9% difference for metrics like the F1 score and the IoU. This suggests that training a model over single date images might make it less robust to temporal variations and less efficient over images taken during another time of the year. The second observation is that, regardless of the test set, the UNet
model achieves the best results. For a more in-depth analysis, we plotted for each model, the differences in IoU, F1 score, and Kappa coefficient (Figure 9).

![Figure 9](image_url)

**Figure 9.** The difference between two test sets in terms of IoU, F1 score, and Kappa coefficient for each model.

The UNet model shows the smallest differences in the three-evaluation metrics. So, it is the least sensitive to seasonal variations. We finally conduct a visual examination of results (Figure 10).

![Figure 10](image_url)

**Figure 10.** The class of urban footprint (in blue) as predicted by each model at a scale of 1/80000: (a) portion of a Sentinel-2 image; (b) reference data; (c) UNet prediction; (d) SegNet prediction; (g) FCN-16s prediction; (f) FCN-32s prediction.

Overall, all the predictions closely resemble the ground truth. However, the UNet model does better when it comes to linear features like roads. If we take a closer look at each prediction (Figure 11), UNet’s and SegNet’s predictions are very refined compared to those of FCN-32s and FCN-16s that are more coarse.
Figure 10. The class of urban footprint (in blue) as predicted by each model at a scale of 1/80000: (a) UNet prediction; (b) SegNet prediction; (c) FCN-16s prediction; (d) FCN-32s prediction.

Overall, all the predictions closely resemble the ground truth. However, the UNet model does better when it comes to linear features like roads. If we take a closer look at each prediction (Figure 11), UNet’s and SegNet’s predictions are very refined compared to those of FCN-32s and FCN-16s that are more coarse.

Figure 11. The class of urban footprint (in blue) as predicted by each model at a scale of 1/30000: (a) UNet prediction; (b) SegNet prediction; (c) FCN-16s prediction; (d) FCN-32s prediction.

Each model’s architecture can explain these results. FCN-32s only carry out one upsampling operation. Because spatial information is lost as we go deeper into the network, the final prediction of FCN-32s is coarse. The improved performance of FCN-16s is due to the presence of a skip connection that makes spatial information flows from earlier layers to deeper ones. For SegNet, the use of a more substantial decoder allows a more progressive reconstruction of spatial information and, thus, a higher performance. As for the UNet model, the use of several layers in the decoder, along with skip connections, makes it the best model. Therefore, we chose the UNet to form the mono-temporal comparison baseline.

3.2. Results of UF Classification

Table 8 below shows the training duration for each UF class.

| Class                | Running Time per Epoch in Second | Total Running Time in Hours |
|----------------------|----------------------------------|----------------------------|
| Continuous UF        | 72                               | 6 h                        |
| Discontinuous UF     | 158                              | 13 h 9 min                 |
| Industrial facilities| 161                              | 13 h 24 min                |
| Rail and road networks| 61                              | 5 h 4 min                  |

ROC curves (Figure 12) reveal a comparable performance for all classes with AUCs between 0.97 and 0.99.
and, thus, a higher performance. As for the UNet model, the use of several layers in the decoder, along with skip connections, makes it the best model. Therefore, we chose the UNet to form the mono-temporal comparison baseline.

3.2. Results of UF classification

Table 7 below shows the training duration for each UF class.

Table 7. The duration of training per class.

| Class                        | Running time per epoch in seconds | Total running time in hours |
|------------------------------|----------------------------------|-----------------------------|
| Continuous UF                | 72                               | 6 hours                     |
| Discontinuous UF             | 158                              | 13 hours 9 minutes          |
| Industrial and tertiary facilities | 161                             | 13 hours 24 minutes         |
| Rail and road network        | 61                               | 5 hours 4 minutes           |

ROC curves (Figure 12) reveal a comparable performance for all classes with AUCs between 0.97 and 0.99.

Table 8 presents the evaluation metrics computed over the validation set, while Figure 13 provides visual results.

Table 8. Evaluation metrics computed over the validation set for each class.

| Model                                 | IoU    | Precision | Recall | F1     | Kappa Coefficient | OA     |
|---------------------------------------|--------|-----------|--------|--------|-------------------|--------|
| Continuous UF                         | 0.623  | 0.723     | 0.819  | 0.768  | 0.764             | 0.992  |
| Discontinuous UF                      | 0.668  | 0.789     | 0.813  | 0.801  | 0.786             | 0.972  |
| Industrial and tertiary facilities    | 0.654  | 0.843     | 0.745  | 0.791  | 0.785             | 0.988  |
| Rail and road network                 | 0.505  | 0.661     | 0.677  | 0.669  | 0.665             | 0.988  |

Figure 12. ROC curve per urban fabric (UF) class.

Table 9 presents the evaluation metrics computed over the validation set, while Figure 13 provides visual results.

Table 9. Evaluation metrics computed over the validation set for each class.

| Model                                 | IoU    | Precision | Recall | F1     | Kappa Coefficient | OA     |
|---------------------------------------|--------|-----------|--------|--------|-------------------|--------|
| Continuous UF                         | 0.623  | 0.723     | 0.819  | 0.768  | 0.764             | 0.992  |
| Discontinuous UF                      | 0.668  | 0.789     | 0.813  | 0.801  | 0.786             | 0.972  |
| Industrial and tertiary facilities    | 0.654  | 0.843     | 0.745  | 0.791  | 0.785             | 0.988  |
| Rail and road network                 | 0.505  | 0.661     | 0.677  | 0.669  | 0.665             | 0.988  |

Figure 13. Classification results on the validation set per class.

Furthermore, we tested each model over four multi-temporal Sentinel 2 images. To span the entire year, we chose the months of November, March, and August. F1 scores were then plotted along with those of the training dataset (Figure 14).
As we move away from the training dataset’s sensing period, the model’s performance gets worse. Moreover, the performance of discontinuous UF does not degrade as quickly as other classes. This can be explained by the number of available samples per class: a model trained over a large dataset has a better generalization capability. Nevertheless, it is still unclear why, even for discontinuous UF, performance degrades as we move further from training time. Afterward, we trained the multi-temporal UNet on two UF classes: Continuous UF and the class of industrial facilities. Results over the validation set are reported in Table 10.

Table 10. Evaluation metrics of the multi-temporal model computed over the validation set for two classes: continuous habitat and industrial facilities.

| Model                             | IoU  | Precision | Recall | F1   | Kappa Coefficient | OA   |
|-----------------------------------|------|-----------|--------|------|-------------------|------|
| Continuous UF                     | 0.612| 0.712     | 0.813  | 0.759| 0.755             | 0.993|
| Industrial and tertiary facilities| 0.591| 0.793     | 0.699  | 0.743| 0.736             | 0.985|

To answer the research question, we develop a testing methodology using multi-temporal Sentinel 2 imagery of the year 2018. We first train them separately, using the UNet, and then together using our multi-temporal framework. The results are listed in Table 11.

Table 11. Comparison between the mono and multi-temporal model in terms of evaluation metrics calculated over test images.

| Model                             | Test Set | IoU  | Precision | Recall | F1   | Kappa Coefficient | OA   |
|-----------------------------------|----------|------|-----------|--------|------|-------------------|------|
| Continuous UF                     | Feb2018  | 0.380| 0.457     | 0.693  | 0.551| 0.548             | 0.995|
|                                  | Mai2018  | 0.347| 0.491     | 0.5424 | 0.516| 0.513             | 0.996|
|                                  | Aug2018  | 0.313| 0.442     | 0.519  | 0.477| 0.475             | 0.995|
| Multi-temporal test set           |          | 0.406| 0.572     | 0.583  | 0.577| 0.576             | 0.996|
| Industrial and tertiary facilities| Feb2018  | 0.322| 0.710     | 0.370  | 0.486| 0.481             | 0.987|
|                                  | Apr2018  | 0.228| 0.694     | 0.254  | 0.371| 0.366             | 0.986|
|                                  | Aug2018  | 0.298| 0.524     | 0.409  | 0.459| 0.451             | 0.984|
| Multi-temporal test set           |          | 0.380| 0.672     | 0.467  | 0.551| 0.545             | 0.987|

It appears that for both categories of UF, all evaluation metrics show some degree of improvement. Continuous habitat has seen its Kappa coefficient go up by 2.72% by comparing the multi-temporal model with the most successful testing period (February 2018). The same goes for the class of industrial facilities for which an improvement of 6.4% was registered. The improvement further increases if we conduct the comparison with the least successful testing period (August 2018). To examine in more
depth how these quantitative improvements translate visually, we plot alongside the ground truth the predictions of the multi-temporal model and those of the mono-temporal model for the most successful testing period (Figures 15 and 16).

**Figure 15.** Comparison between predictions of the mono-temporal UNet and the multi-temporal UNet for the class of industrial facilities.

**Figure 16.** Comparison between predictions of the mono-temporal UNet and the multi-temporal UNet for the class of continuous UF.
Predictions made by the multi-temporal model are better than those produced by its mono-temporal counterpart for both the classes. Improvements were more critical for the class of industrial facilities, where the multi-temporal model was able to classify previously undetected portions of the image accurately. Similarly, for the class of continuous habitat, predictions made by the multi-temporal model are more refined and resemble more closely the ground truth.

4. Discussion

These experimentations evaluate if multi-temporal imagery might help discriminate between some UF classes with different occurrences of urban structures, vegetation, and bare surfaces. At first, we highlight that a mono-temporal model performed systematically better on data covering a different geographical zone than on data taken during a different time. This suggests that seasonal variations prevented models from recognizing certain UF classes. Moreover, when testing each class’s model on a multi-temporal test set, performance decreases as we move away from the training set’s sensing time. Both of these findings are complementary to each other: a mono-temporal model will not only do worse on multi-temporal data, but its performance will also continue to degrade as we move away from the training period. The proposed multi-temporal model did better than its mono-temporal counterpart. Namely, using three test images of the year 2018 simultaneously proved more effective than exploiting them separately. The quantitative, as well as the visual improvements, were more pronounced for the class of industrial facilities, compared to continuous UF. This is consistent with our general hypothesis: a multi-temporal approach can be fruitful to differentiate between classes with temporal characteristics. All of these results are consistent with each other as they all point to the temporal domain as possible useful input for urban land use classification. All things considered, a reasonable conclusion is that learning new features using multi-temporal data and deep learning can improve the classification accuracy for both continuous UF and industrial facilities. Another sound conclusion is that Sentinel 2 images, despite a spatial resolution of 10 m, can be used for urban land use mapping.

Nevertheless, the contribution of the temporal dimension to UF mapping requires further testing with larger datasets and longer time series. With this mind, we think the positive results achieved in this research are worth building upon in future studies to further investigate the temporal domain as an input to urban land use classification models. Conducting experiments on a wide range of UF classes, using longer time series, and bigger datasets all represent good avenues for future research.

5. Conclusions and Perspectives

In conclusion, this work explored the use of multi-temporal Sentinel-2 images for UF mapping. To this end, a novel multi-temporal fully convolutional network based on the famous UNet architecture was implemented. The proposed model learns a spatio-temporal representation of the data without the need for extensive feature engineering or heavy preprocessing and is trained in an end-to-end fashion. To investigate whether the temporal dimension can improve the classification performance of UF, we devised various testing methodologies using mono-temporal and multi-temporal images. We first assessed the robustness of four popular deep learning benchmarks to spatial and temporal variations and then went on to conduct a comparative analysis between the performance of a mono-temporal model and the proposed multi-temporal model. Experimental results showed that a mono-temporal model’s performance tends to decrease as the test set’s sensing period departs from that of the training set. Furthermore, the proposed multi-temporal framework yielded the best results with a Kappa coefficient increase of 2.72% and 6.4% for the classes of continuous UF and industrial facilities, respectively. These findings highlight the benefit of exploiting spatio-temporal properties of satellite images for urban land use mapping and pave the way to new approaches to leverage the power of multi-temporal data to discriminate among multiple urban categories.
Author Contributions: Conceptualization, L.E.M., M.C., A.P. and I.S.; methodology, A.P. and I.S.; software, L.E.M. and M.C.; formal analysis, L.E.M., M.C., A.P. and I.S.; writing—original draft preparation, L.E. and M.C.; writing—review and editing, A.P. and L.E.M.; supervision, A.P. and I.S.; funding acquisition, A.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the French funded project ANR TIMES ‘High-performance processing techniques for mapping and monitoring environmental changes from massive, heterogeneous and high frequency data times series’ (ANR-17-CE23-0015-01) and the CNES TOSCA program 2018.

Conflicts of Interest: The authors declare no conflict of interest.

References
1. United Nations Department of Economic and Social Affairs Population Division. World Urbanization Prospects 2018. Available online: https://population.un.org/wup/Publications/Files/WUP2018-Report.pdf (accessed on 24 December 2018).
2. Merlin, F.; Choay, F. Dictionnaire de l’urbanisme et de l’aménagement, 1st ed.; Presses Universitaires de France: Paris, France, 2005; p. 723.
3. Hulley, M. The urban heat island effect: causes and potential solutions. Metrop. Sustain. 2012, 79–98.
4. Irwin, E.G.; Bockstael, N.E. The evolution of urban sprawl: Evidence of spatial heterogeneity and increasing land fragmentation. Proc. Natl. Acad. Sci. USA 2007, 104, 20672–20677.
5. Puissant, A.; Lachiche, N.; Skupinski, G.; Braun, A.; Perret, J.; Mas, A. GeOpenSim: Classification et évolution des tissus urbains. Revue Internationale de Géomatique 2011, 21, 513–532. [CrossRef]
6. Zhai, Y.; Qu, Z.; Hao, L. Land Cover Classification Using Integrated Spectral, Temporal, and Spatial Features Derived from Remotely Sensed Images. Remote Sens. 2018, 10, 383. [CrossRef]
7. Ienco, D.; Gaetano, R.; Dupaquier, C.; Maurel, P. Land Cover Classification via Multitemporal Spatial Data by Deep Recurrent Neural Networks. IEEE Geosci. Remote Sens. Lett. 2017, 14, 1685–1689. [CrossRef]
8. Schäfer, P.; Pfugmacher, D.; Hostert, P.; Leser, U. Classifying land cover from satellite images using time series analytics. In Proceedings of the Workshops of the International Conference on Extending Database Technology and the International Conference on Database Theory, Austria, Vienna, 26–29 March 2018; pp. 10–15.
9. Inglanda, J.; Vincent, A.; Arias, M.; Tardy, B.; Morin, D.; Rodes, I. Operational High Resolution Land Cover Map Production at the Country Scale Using Satellite Image Time Series. Remote Sens. 2017, 9, 95. [CrossRef]
10. Ienco, D.; Gaetano, R.; Dupaquier, C.; Maurel, P. Land Cover Classification via Multitemporal Spatial Data by Deep Recurrent Neural Networks. IEEE Geosci. Remote Sens. Lett. 2017, 14, 1685–1689. [CrossRef]
11. Zhu, Z.; Woodcock, C.E.; Rogan, J.; Kellindorfer, J. Assessment of spectral, polarimetric, temporal, and spatial dimensions for urban and peri-urban land cover classification using Landsat and SAR data. Remote Sens. Environ. 2012, 117, 72–82. [CrossRef]
12. Carrão, H.; Gonçalves, P.; Caetano, M. Contribution of multispectral and multitemporal information from MODIS images to land cover classification. Remote Sens. Environ. 2008, 112, 986–997. [CrossRef]
13. Guerschman, J.P.; Paruelo, J.M.; Di Bella, C.; Giallorenzi, M.C.; Pacin, F. Land cover classification in the Argentine Pampas using multi-temporal Landsat TM data. Int. J. Remote Sens. 2003, 24, 3381–3402. [CrossRef]
14. Ball, J.E.; Anderson, D.T.; Chan, C.S. Comprehensive survey of deep learning in remote sensing: theories, tools, and challenges for the community. J. Appl. Remote Sens. 2017, 11, 1. [CrossRef]
15. Rußwurm, M.; Körner, M. Multi-Temporal Land Cover Classification with Sequential Recurrent Encoders. ISPRS Int. J. Geo-Inf. 2018, 7, 129.
16. Pelletier, C.; Valero, S.; Inglanda, J.; Champion, N.; Dedieu, G. Assessing the robustness of Random Forests to map land cover with high resolution satellite image time series over large areas. Remote Sens. Environ. 2016, 187, 156–168. [CrossRef]
17. Gómez, C.; White, J.C.; Wulder, M.A. Optical remotely sensed time series data for land cover classification: A review. ISPRS J. Photogramm. Remote Sens. 2016, 116, 55–72. [CrossRef]
18. Fawaz, H.I.; Forestier, G.; Weber, J.; Idoumghar, L.; Muller, P.-A. Deep learning for time series classification: a review. Data Min. Knowl. Discov. 2019, 33, 917–963. [CrossRef]
19. Khan, S.; Rahmani, H.; Shah, S.A.A.; Bennamoun, M. A Guide to Convolutional Neural Networks for Computer Vision. Synth. Lect. Comput. Vis. 2018, 8, 1–207. [CrossRef]
20. Hu, F.; Xia, G.-S.; Hu, J.; Zhang, L. Transferring Deep Convolutional Neural Networks for the Scene Classification of High-Resolution Remote Sensing Imagery. *Remote Sens.* **2015**, *7*, 14680–14707. [CrossRef]

21. Audebert, N.; Le Saux, B.; Lefèvre, S. Beyond RGB: Very high resolution urban remote sensing with multimodal deep networks. *ISPRS J. Photogramm. Remote Sens.* **2018**, *140*, 20–32. [CrossRef]

22. Yan, J.; Wang, H.; Yan, M.; Diao, W.; Sun, X.; Li, H. Iou-Adaptive Deformable R-CNN: Make Full Use Of Iou For Multi-Class Object Detection In Remote Sensing Imagery. *Remote Sens.* **2019**, *11*, 286. [CrossRef]

23. Ma, L.; Liu, Y.; Zhang, X.; Ye, Y.; Yin, G.; Johnson, B.A. Deep learning in remote sensing applications: A meta-analysis and review. *ISPRS J. Photogramm. Remote Sens.* **2019**, *152*, 166–177. [CrossRef]

24. Fu, G.; Liu, C.; Zhou, R.; Sun, T.; Zhang, Q. Classification for High Resolution Remote Sensing Imagery Using a Fully Convolutional Network. *Remote Sens.* **2017**, *9*, 498. [CrossRef]

25. Audebert, N.; Le Saux, B.; Lefèvre, S. Semantic Segmentation of Earth Observation Data Using Multimodal and Multi-scale Deep Networks. *ACCV 2016*.

26. Badrinarayanan, V.; Kendall, A.; Cipolla, R. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495. [CrossRef] [PubMed]

27. Zhang, P.; Ke, Y.; Zhang, Z.; Wang, M.; Li, P.; Zhang, S. Urban Land Use and Land Cover Classification Using Novel Deep Learning Models Based on High Spatial Resolution Satellite Imagery. *Sensors* **2018**, *18*, 3717. [CrossRef] [PubMed]

28. Pelletier, C.; Webb, G.I.; Pettiejan, F. Temporal Convolutional Neural Network for the Classification of Satellite Image Time Series. *Remote Sens.* **2019**, *11*, 523. [CrossRef]

29. Benedetti, P.; Ienco, D.; Gaetano, R.; Ose, K.; Pesna, R.; Dupuy, S. M3Fusion: A Deep Learning Architecture For Multiscale Multimodal Multitemporal Satellite Data Fusion. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **2018**, *11*(12), 4939–4949. [CrossRef]

30. Mauro, N.D.; Vergari, A.; Basile, T.M.; Ventola, F.G.; Esposito, F. End-to-end Learning of Deep Spatio-temporal Representations for Satellite Image Time Series Classification. In Proceedings of the PKDD-ECML Discovery Challenges, Skopje, Macedonia, 18 September 2017.

31. Nogueira, K.; Dos Santos, J.A.; Menini, N.; Silva, T.S.F.; Morellato, L.P.C.; Torres, R.D.S. Spatio-Temporal Vegetation Pixel Classification by Using Convolutional Networks. *IEEE Geosci. Remote Sens. Lett.* **2019**, *16*, 1665–1669. [CrossRef]

32. Theia - Land Data Center. Available online: https://theia.cnes.fr/atdistrib/rocket/#/search?collection=SENTINEL2 (accessed on 14 January 2019).

33. Cazaubiel, V.; Chorvalli, V.; Miesch, C. The multispectral instrument of the Sentinel2 program. In Proceedings of the International Conference on Space Optics — ICSEO - SPIE-Intl Soc Optical Eng, Toulouse, France, 14–17 October 2008; p. 95.

34. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. ImageNet Large Scale Visual Recognition Challenge. *Int. J. Comput. Vis.* **2015**, *115*, 211–252. [CrossRef]

35. Momeni, R.; Aplin, P.; Boyd, R.S. Mapping Complex Urban Land Cover from Spaceborne Imagery: The Influence of Spatial Resolution, Spectral Band Set and Classification Approach. *Remote Sens.* **2016**, *8*, 88. [CrossRef]

36. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.

37. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. *arXiv* arXiv:1505.04597.

38. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. Available online: https://arxiv.org/abs/1409.1556 (accessed on 14 January 2019).

39. Glorot, X.; Bengio, Y. Understanding the difficulty of training deep feedforward neural networks. In Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, Sardinia, Italy, 13–15 May 2010; p. 2010.

40. Pan, S.; Yang, Q. A Survey on Transfer Learning. *IEEE Trans. Knowl. Data Eng.* **2010**, *22*, 1345–1359. [CrossRef]
41. Smith, L.N. Cyclical Learning Rates for Training Neural Networks. In Proceedings of the 2017 IEEE Winter Conference on Applications of Computer Vision (WACV), Santa Rosa, CA, USA, 27–29 March 2017; pp. 464–472.

42. Iglovikov, V.; Seferbekov, S.; Buslaev, A.; Shvets, A. TernausNetV2: Fully Convolutional Network for Instance Segmentation. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Salt Lake City, UT, USA, 18–22 June 2018; pp. 228–2284.

43. García-García, A.; Orts-Escolano, S.; Oprea, S.; Villena-Martinez, V.; Martinez-González, P.; García-Rodríguez, J. A survey on deep learning techniques for image and video semantic segmentation. Appl. Soft Comput. 2018, 70, 41–65. [CrossRef]

44. Fawcett, T. An introduction to ROC analysis. Pattern Recognit. Lett. 2006, 27, 861–874. [CrossRef]