

U-Net feature fusion for multi-class semantic segmentation of urban fabrics from Sentinel-2 imagery: an application on Grand Est Region, France

Romain Wenger¹, Anne Puissant¹, Jonathan Weber², Lhassane Idoumghar² and Germain Forestier²

¹LIVE UMR 7362 CNRS, University of Strasbourg, F-67000 Strasbourg, France;

²Université de Haute-Alsace, IRIMAS UR 7499, F-68100 Mulhouse, France

ABSTRACT

Urban areas are increasing since several years as a result of development of built-up areas, network infrastructure, industrial areas or other built-up areas. This urban sprawl has a considerable impact on natural areas by changing the functioning of ecosystems. Mapping and monitoring Urban Fabrics (UF) is therefore relevant for urban planning and management, risk analysis, human health or biodiversity. For this research, Sentinel-2 (level 2A) single-date images of the East of France, with a high spatial resolution (10m), are used to assess two semantic segmentation networks (U-Net) that we combined using feature fusion between a from scratch network and a pre-trained network on ImageNet. Moreover three spectral or textural indices have been added to the both networks in order to improve the classification results. The results showed a performance gain for the fusion methods in classifying several UF. However, there is a difference in performance depending on the urbanization gradient; highly urbanized areas provide a better distinction of some UF's classes than rural areas.

1. Introduction

By 2050, more than three out of four people will live in cities. For comparison, in 1950, it was only 33% of the population that lived in cities (United Nations Department of Economic and Social Affairs Population Division 2018), slightly more than one in four inhabitants. As a consequence, urban areas are increasing as a result of development of built-up areas, network infrastructure, industrial areas or other built-up areas. This urban sprawl triggers changes in landcover with the consumption of agricultural and natural areas, and has impacts on the ecosystems with important ecological, climate and social transformations (Irwin and Bockstael 2007; Zhu et al. 2019). Most studies quantify the dynamic of urban footprint (Puissant et al. 2011) which includes the road network, buildings, vegetation, and impervious surfaces (El Mendili et al. 2020). Few of them analyses the inner dynamics of urban areas through the changes of urban fabrics (UF) corresponding to a specific spatial organization of basic components of the city. Several works based on geographic object-based image analysis (GEOBIA) have been explored to obtain land cover land use (LULC) classifications (Souza-Filho et al. 2018; De Luca et al. 2019; Uddin, Abdul Matin, and Maharjan 2018) with a

higher accuracy than "per-pixels" methods, which are considered insufficient and do not take into account the neighbors of each pixel.

With the multiplication of Earth Observation satellites, the amount of acquired satellite images continue to grow exponentially. The evolution of computing power in computer science has motivated researchers to develop classification methods based on neural networks (Kamga et al. 2021) and running on GPUs instead of CPUs. Ma et al. (2019) demonstrate the renewed interest in these techniques by conducting a review using these neural network-based deep learning methods in remote sensing. Deep learning covers several fields in remote sensing, whether for image fusion (example of pan-sharpening) (Xing et al. 2018), scene classification and object detection (Zhong, Han, and Zhang 2018; Ding et al. 2018; Sumbul et al. 2019), LULC classification (Marcos et al. 2018; Zhu et al. 2018) or semantic segmentation (Chen et al. 2018; Kemker, Salvaggio, and Kanan 2018). These fields can be grouped into four main tasks : image preprocessing, change detection, accuracy assessment and classification (Ma et al. 2019).

Semantic segmentation methods are used in several domain applications from medical image segmentation to object detection on photographs (Han et al. 2019; Shin et al. 2016). In computer vision, CNN (Convolutional Neural Networks) are excellent networks to analyze images containing high level spatial features. CNNs, along with VGG networks, are used in remote sensing for slums detection (Wurm et al. 2019), object detection such as aircraft (Ding et al. 2018) or change detection (Amin Larabi et al. 2019). Moreover, encoder/decoder networks provide a higher accuracy than classical CNNs in detecting boundaries between objects (Chhor and Aramburu 2017). U-Net (Ronneberger, Fischer, and Brox 2015) and SegNet (Badrinarayanan, Kendall, and Cipolla 2017), two encoder/decoder networks have been developed for image classification. U-Net has shown excellent results in biomedical image segmentation and SegNet in scene classification. These networks have also been used in remote sensing for land cover classification of very high resolution images (Zhang et al. 2018) but also in the classification of UF using high resolution Sentinel-2 RGB imagery and pre-trained ImageNet (Russakovsky et al. 2015) weights (El Mendili et al. 2020). Sefrin, Riese, and Keller (2021) used an encoder/decoder like network combined with a LSTM (Long Short-Term Memory) to classify land cover into 8 classes from Sentinel-2 images and obtained high classification score in several land use classes.

Transfer learning (Lu et al. 2015) is also used in many semantic segmentation works. This technique consists in assigning the weights of a pre-trained network on a data source to the target network that we intend to train (Oquab et al. 2014). This result is a time saving in the training of the target network and also allows to counterbalance a dataset containing very few entries (Xie et al. 2016; Momeni, Aplin, and Boyd 2016). Kemker, Salvaggio, and Kanan (2018) showed the efficiency of this method by transferring weights from a source network trained on panchromatic images to a network treating multispectral data. Shendryk et al. (2019) trained a network for scene classification using planetescope imagery and then transferred the weights for Sentinel-2 image classification. Iglovikov and Shvets (2018) modified the decoder part of a U-Net in VGG11 to be able to use pre-trained weights on ImageNet (Russakovsky et al. 2015). Thanks to this fine-tuning technique, they obtained high-quality results that can be further improved by using deeper networks such as VGG16 or any other ResNet-like networks.

Many studies have shown the interest of fusion techniques between two networks. For instance, Audebert, Le Saux, and Lefèvre (2018) trained a modified SegNet using feature fusion method to detect UF classes from aerial images. They fuse the encoder

phase from two SegNet networks, one using exogenous indexes and the other IRRG (Infra-Red, Red and Green) bands using pretrained ImageNet weights. They obtained better results using this new method. Network fusion also applies when using CNN and ConvGRU (Convolutional Gated Recurrent Unit). Ienco et al. (2019) combined these two networks and two data sources for land cover mapping of Reunion Island and Koumbia and shows the efficiency of networks fusion from different sources. Hu et al. (2017) developed a two stream CNN for LULC classification using radar and hyperspectral images. In addition, many datasets have been developed to perform research on multi-modal fusion, whether between optical and radar imagery (Schmitt et al. 2019; Sumbul et al. 2021), aerial imagery and a Digital Surface Model (Vaihingen and Potsdam datasets developed for ISPRS 2D Semantic Labeling Challenge¹) or hyperspectral and LiDAR imagery (Houston2013 dataset ²).

The use of spectral and textural indexes when applying neural networks are increasingly used as an external addition of exogenous indexes calculated with the spectral bands of the sensor chosen for the study. These hand-crafted features allow for differentiation of different types of spatially structuring objects to accelerate network learning and improve classification results (Liu et al. 2017). Campos-Taberner et al. (2020) developed a method to determine the importance of Sentinel-2 bands and spectral and textural indices in a neural network. They noted that NDVI (Normalized Difference Vegetation Index), NIR (Near Infra-Red) and red bands, and entropy calculated on NDVI (*eNDVI*) are the features providing the most relevant information to the network. NDVI index is frequently used in the detection of UF because it allows to accentuate the distinction between these spaces and the vegetative areas due to the important difference in the spectrum of the materials constituting them. In fact, it is widely used in land cover/land use detection research (Ienco et al. 2019; Inglada et al. 2017). NDBI (Normalized Difference Building Index) (Zha, Gao, and Ni 2003), is an index developed to rapidly extract urban fabric (Yi and Jianhui 2016). It works like the NDVI by making a combination of different spectral bands, in this case for this index the MIR (Mid Infra-Red) and the NIR (Near Infra-Red).

In this context, our research focused on the contribution of feature fusion from Sentinel-2 high spatial resolution imagery to map UF based on a generic typology for France. In addition, this study aims to show that the use of exogenous indices (NDVI, NDBI and *eNDVI*) can improve the classification results of these UF. The paper is structured in five sections. Section 2 will describe the materials and methods. Section 3 will present the results for each method and study area. In section 4, we will discuss the results presented previously before concluding and developing our research perspectives in section 5.

2. Materials and Methods

In this section, the study sites is firstly described (section 2.1) followed by the presentation of both satellite and databases processed to obtain a reference dataset allowing to improve the UF map into five thematic classes (section 2.2). The proposed workflow is then explained to produce urban semantic segmentation based on Sentinel-2 mono-date image.

¹<http://www2.isprs.org/commissions/comm3/wg4/semantic-labeling.html>

²https://hyperspectral.ee.uh.edu/?page_id=459

2.1. Study Sites and Test areas

The 'Grand Est' region is an administrative French zone extended from Alsace in the East to the Ardennes and Marne in the West, covers 57,441 km² as well as many large urban areas such as Strasbourg, Metz, Nancy and Reims. Among the sixteen tiles covered by Sentinel-2, two of them are chosen for study sites for training and testing our classification models from imbalanced reference data (Figure 1). The land cover classes distribution of both tiles is representative of the whole region covered urban, peri-urban and rural areas with a diversity of UF. Three subsets of test with a gradient of urbanisation are selected in order to assess the robustness and precision of the methods tested for a diversity of case studies. The first test area is located on tile 32ULU North, one the most important city of the East French Region (Strasbourg with more than 500 000 inhabitants), the second one includes part of the city of Metz (one of the four cities with around 200,00 inhabitants) and the last one is located near the smaller city of Saint-Avold (tile 31UGQ) and is representative of cities with less than 50,000 inhabitants (Figure 1).

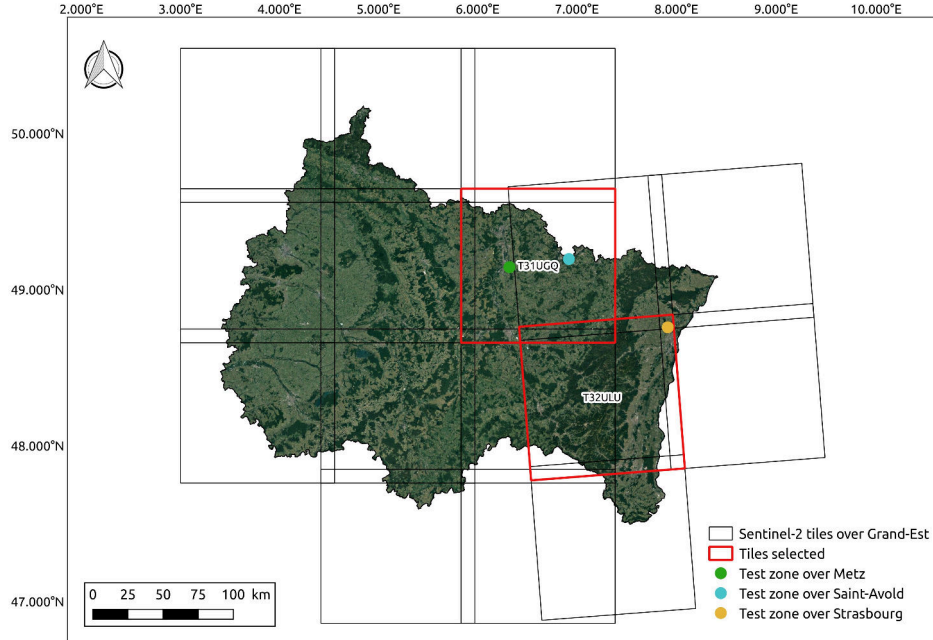


Figure 1. Grand Est region, France, including Sentinel-2 tiles and test areas where the experiments were made. (Coordinate system used is World Geodetic System 1984/EPSSG4326)

2.2. Datasets

2.2.1. Sentinel-2 Data

Sentinel-2 mission (Drusch et al. 2012) is composed of two satellites, Sentinel-2A and 2B respectively launched in June 2015 and in March 2017. They have a high revisit frequency of 5 days over the equator, and 2/3 days near mid-latitudes which is important to map land cover dynamics. Each sensor owns 13 spectral bands with different wavelengths, from the visible to the shortwave infrared at different spatial resolutions.

For this research paper, satellite data used come from the Theia/Muscate database (<https://www.theia-land.fr/>) and 10 spectral bands are available for each sensor (Table 1). This product is available on their dissemination platform or by automatic download by requesting their server. For this work, cloudless single-date Sentinel-2 images, from 24th July 2019 and 21th July 2020, are respectively chosen for tiles 32ULU and 31UGQ.

Table 1. Sentinel-2 bands available with 2A product from Theia/Muscate.

Band name	Central Wavelength (nm)	Spatial Resolution (m)
Band 2 - Blue	492.4	10
Band 3 - Green	559.8	10
Band 4 - Red	664.6	10
Band 5 - Vegetation Red Edge	704.1	20
Band 6 - Vegetation Red Edge	740.5	20
Band 7 - Vegetation Red Edge	782.8	20
Band 8 - Near Infra-Red	832.8	10
Band 8A - Vegetation Red Edge	864.7	20
Band 11 - SWIR	1613.7	20
Band 12 - SWIR	2202.4	20

2.2.2. UF Typology and Reference Dataset


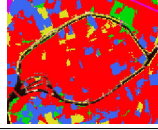

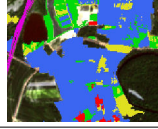



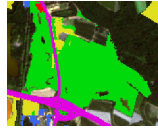
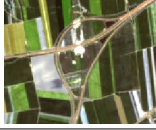
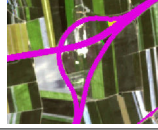


From High Spatial Resolution Imagery (10m), many research papers map UF in western cities in 4 classes (Table 2): dense (class 1) and sparse built-up (class 2) areas where the main difference consists in their relative density and the importance of vegetated areas and bare surfaces, specialized areas (class 3) characterizing industrial activities or waste land or open areas with a majority of artificial or bare surfaces and large scale road or rail network (class 5). This is the case for the well-known product, OSO (Occupation des Sols Opérationnelle - <http://osr-cesbio.ups-tlse.fr/oso/>) available at the national scale and produced from times series of Sentinel-2 or other research works (El Mendili et al. 2020). In order to improve this classification of UF, a fifth class has been added and describes specialized areas where green surfaces are dominant (more than 80%) (class 4) such as urban parks, cemeteries or vegetated sports or leisure complexes (outdoor sports fields). In this paper, we assume that a spatial resolution of 10m make it possible to map these five UF classes. A sixth class for non-urban areas (class 6) corresponds to any other non built-up areas as agricultural lands, forests or water surfaces.

To produce this reference dataset adapted to any type of city in France to map UF at 10m spatial resolution, two existing topographic databases are used : (1) the regional landuse/cover vector database (BDOCSGE2©GeoGrandEst³, 2019) and (2) the national topographic database with networks (BDTOPO©IGN⁴, 2014). The Minimum Mapping Unit (MMU) is less than 50m² for buildings and urban areas. The first database is produced by visual interpretation of aerial images (2018-2019) and maps each department of the region into 53 classes at level 4. The thematic classes are close to landuse that landcover classes. This database is a freely available at large scale (1:10,000) from the GeoGrandEst Data Infrastructure (www.geograndest.org) and the production is ongoing for the whole region. At the time of this research, only two departments were available (Bas-Rhin and Moselle) in relation with our tests sites. Since the satellite images and reference data are less than two years distant, we make the

³<https://www.datagrandest.fr/portail/fr/projets/occupation-du-sol>

⁴<https://geoservices.ign.fr/documentation/donnees/vecteur/bdtopo>

Table 2. List of the five UF classes (after pre-processing) used for this research (1/6000 scale).

Class	Subset on Sentinel-2	Reference data	Description
(1) Dense Built-Up			Surfaces mainly occupied by buildings and impervious surfaces. Vegetation and bare soil are scarce.
(2) Sparse Built-Up			Buildings and other artificial surfaces share the land with green surfaces and bare soil
(3) Specialized Built-Up Areas			Surfaces allocated to production, commercial, service and tertiary activities
(4) Specialized but Vegetative Areas			Surfaces containing at least 80% of vegetated areas and 20% of bare soil or impervious surfaces as urban park, sport leisure activities, cemetery or campgrounds
(5) Large Scale Networks			Primary road network and others associated areas, railways and train stations
(6) Others non-urban areas			Every non-urban areas such as agricultural land, forests, wetlands and water surfaces

hypothesis that changes are minor due to the low population dynamics in the region. The legend of BDOCSGE2 is organized into four levels of nomenclature where 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 16 classes at Level 3 and in 29 classes in level 4. In the BDOCSGE2 layer, all the roads have the same degree of importance which makes it impossible to remove the smallest polygons in the inner city that cannot be distinguished at 10 m spatial resolution (Table 2). In order to produce an adapted road network thematic class, the second database (BDTopo), produced by IGN describing lines in vector format with the degree of importance, is pre-processed. A buffer of 30 meters for the highways and 10 meters for the major roads is calculated to correspond to the real size of these networks which the only ones to be visible at 10 m. These data are then added to other classes of the vector reference datasets layer. We then summarize the fifth classes into a unique vector layer and rasterize all polygons at a 10 m spatial resolution. The UF classes are not evenly distributed on the reference data and represent 8.4% of the total area. Indeed, Dense Built-Up represents about 8.4% of the total UF area of the dataset, Sparse Built-Up 59.5%, Specialized Built-Up Areas 19%, Specialized but Vegetative Areas 8.3% and Large Scale Networks 4.8%. For the image processing test based on four classes, the class (4) describing specialized but vegetative areas has been removed and all these areas have been included in the last class (6).

2.3. Methods

The proposed workflow is proposed in three different steps (Figure 2.: (1) the pre-processing step for a training, validation and test patches data preparation, (2) the model training step where four different approaches built the same base network are compared and (3) the post-processing and evaluation step to predict and test every approach. All models have been trained and tested on a computer with an RTX Quadro 4000 with 8 Gb of VRAM, an Intel(R) Xeon(R) E-2246G processor clocked at 3.6 GHz for 6 physical cores and 32 Gb of rams. For implementation, we used Keras API (Chollet et al. 2015) built on top of TensorFlow 2.0.

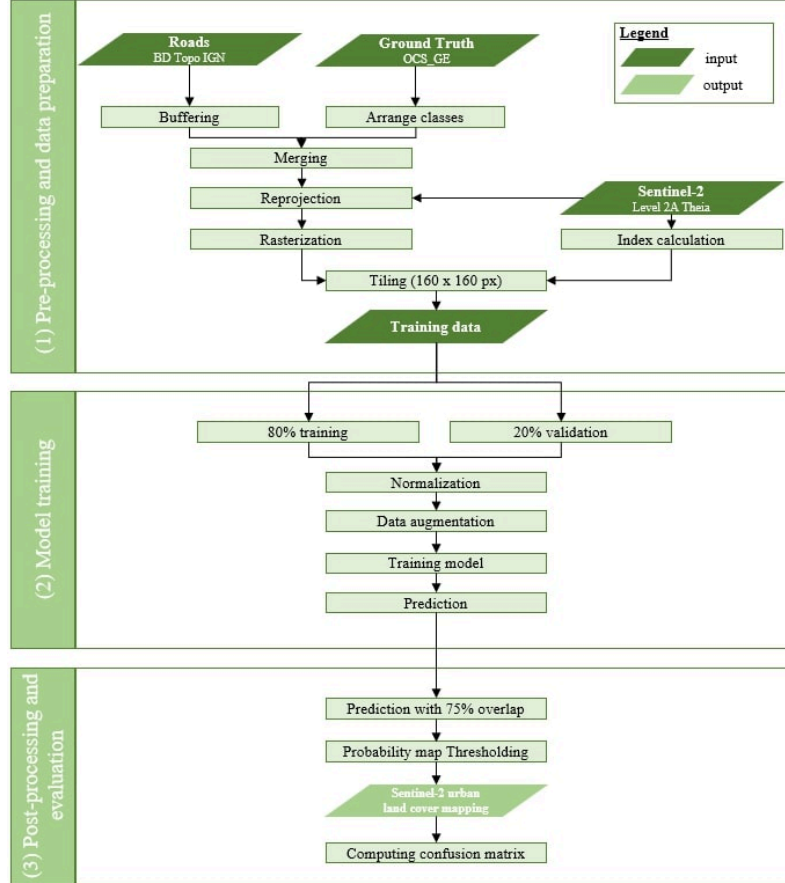


Figure 2. Workflow for automated UF mapping in Sentinel-2 L2A imagery over one tile. This workflow contains (1) preprocessing and data preparation where reference data, roads and satellite data are pre-processed, (2) model training where some networks are applied and (3) post-processing and evaluation where predictions are made.

2.3.1. Step 1: Pre-processing and data preparation

On the Sentinel-2 images, first combination of spectral bands, as input for all the Networks, are the 3 IRRG bands (Infrared, red and green). We kept these three spectral bands to merge the two networks with a similar depth and compare the 4 methods developed with equivalent input data. In addition, band 8 (Near Infra-Red) was preferred to band 2 (Blue) because it provides more information on vegetation and allows better distinction between urban and natural areas. The normalized difference vege-

tation index (NDVI) (Rouse 1973), the normalized difference built-up index (NDBI) (Zha, Gao, and Ni 2003) and the entropy (Haralick, Shanmugam, and Dinstein 1973) based on NDVI index are also calculated and will be inputs in Networks due to their relevance in urban studies (Huang, Yu, and Feng 2019; Mhangara and Odindi 2012; Su et al. 2008).

The reference data is pre-process in order to produce training and validation dataset for the different Networks used.

$$NDVI = \frac{Red - NIR}{Red + NIR} \quad (1)$$

$$NDBI = \frac{SWIR1 - NIR}{SWIR1 + NIR} \quad (2)$$

where *Red* is the red band of the Sentinel-2 image (band number 4), *NIR* is the near infrared band (band number 8) and *SWIR1* is the Short-Wave Infrared 1 band (band number 11). *SWIR1* is resampled at 10m spatial resolution to match spatial resolution with other spectral bands.

Haralick, Shanmugam, and Dinstein (1973) introduced the concept of Grey Level Co-occurrence Matrix (GLCM) in 1973. This technique of feature extraction is widely used in the field of image analysis. GLCM represents a histogram of co-occurring greyscale values at a given offset. Then in urban areas where the heterogeneity is high, the Entropy index representing the randomness of disorder present in the image is calculated. The entropy value is high when the elements of the co-occurrence matrix are the same and lower are the values of entropy and more unequal are the elements. Entropy (eNDVI) is calculated using NVDI. We defined a direction of 3 pixels and an offset of 1 to compute the GLCM.

$$eNDVI = - \sum_i \sum_j p_d(i, j) \ln p_d(i, j) \quad (3)$$

where $p_d(i, j)$ is the (i, j) th element of the normalized GLCM.

Data sources are split in training, validation and test zones following Saraiva et al. (2020) methodology which consists in selecting an area covered by the reference data, splitting it with a part for the training set and another for the validation set. The patches of each set are then cut in the selected areas by applying an overlap of 50% between each patch. Each area being independent, there can be no overlap between the partitions. Random patches are selected in training and validation areas which result of 5,517 training patches (80%) and 1,379 validation patches (20%) for 31UGQ tile and 8,942 training patches (80%) and 2,685 validation patches (20%) for 32ULU tile (Table 3). Test zone is splitted in patches with 75% overlap and reconstructed to predict and evaluate all the test zone. This method is applied for every network tested.

2.3.2. Models training

This section gives a detailed description of models training and selected networks. Fusion methods were successfully tested in our domain application for encoder/decoder

Table 3. Number of patch containing a class and number of pixels per class for each training set.

Class	32ULU		31UGQ	
	Patches	Pixels	Patches	Pixels
(1) Dense Built-Up	3,825	1,295,473	2,330	849,901
(2) Sparse Built-Up	7,154	10,724,491	4,374	4,698,560
(3) Specialized Built-Up Areas	5,050	3,141,083	4,168	1,985,748
(4) Specialized but Vegetative Areas	6,147	1,802,533	3,612	1,228,366
(5) Large Scale Networks	1,420	502,951	1,148	477,968
(6) Others non-urban areas	8,942	211,448,669	5,517	131,994,657

like networks (Hazırbaşı et al. 2016; Audebert, Le Saux, and Lefèvre 2018; Jie et al. 2020) to combine different input data sources for land cover classification.

In this paper, one main U-Net network with VGG-16 is used as the encoder to be able to apply pretrained ImageNet (Russakovsky et al. 2015) weights. We chose this backbone because it is one of the smallest networks and helps to limit overfitting. Pretrained weights and backbone VGG-16 network were obtained from keras built-in models. The encoding phase for this network allows to extract a set of features to detect the classes present within the patches. Conversely, the decoding phase allows to restore the spatial context of the patch. Skip connections between the two phases of the U-Net network allows to find more quickly characteristics discovered during the first blocks of the network without having to go through the deeper meshes.

Table 4. List of executions including networks used and spectral bands / index

Method	Network	Spectral bands/Index
U-Net-IRRG (Figure 4)	U-Net pretrained on ImageNet	3, 4, 8
U-Net-Index (Figure 4)	U-Net not pretrained	3, 4, 8, NDVI, NDBI, eNDVI
U-Net-Encoder (Figure 5)	Fusion of two U-Net	3, 4, 8 and NDVI, NDBI, eNDVI
U-Net-Decoder (Figure 6)	Fusion of two U-Net	3, 4, 8 and NDVI, NDBI, eNDVI

We have developed two fusion methods inspired by Hazırbaşı et al. (2016) works as it has proven to be effective in Audebert, Le Saux, and Lefèvre (2018) by combining exogenous indexes with reflectance data. First, a fusion of the encoders of two U-Net is performed (Figure 5) and then a fusion of the decoders (Figure 6). These fusion methods are compared to the classical U-Net network (Figure 4) with different input parameters described below and summarized in Table 4:

- **U-Net-IRRG (Figure 4):** U-Net with VGG-16 as a backbone, taking IRRG patches as inputs. This network is pretrained on ImageNet ;
- **U-Net-Index (Figure 4):** U-Net with VGG-16 as a backbone, taking six channels patches as inputs : green, red, infrared bands and NDVI, NDBI and eNDVI. This network could not be pretrained on ImageNet because it takes more than 3 channels as input ;
- **U-Net-Encoder (Figure 5):** Two U-Net with VGG-16 as as backbone. The main network takes 3 indexes patches (NDVI, NDBI and eNDVI) as inputs and it not pretrained on ImageNet. The second network takes IRRG as inputs and is pretrained on ImageNet. The contributions of each encoder are summed after each convolution block ;
- **U-Net-Decoder (Figure 6):** The same methodology as encoder fusion is applied but the fusion is executed during the decoder phase, including bottleneck in order to alter the final result as much as possible ;

These approaches consist in classifying satellite image patches of dimension $h \times w \times n_{channels}$ to obtain a classification of dimensions $h \times w \times n_{classes}$. All the operations described in Figures 4, 5 and 6 are explained below.

- **Convolution:** Each convolution block consists in applying a convolution with a kernel size of 3×3 and a stride of 1 pixel. *ReLu* (Rectified Linear Unit, $f(x) = \max(0, x)$) is the activation function used at the end of each convolution ;
- **Dropout:** This regularization technique is commonly used when developping a neural network (Srivastava et al. 2014). 50% dropout (Pelletier, Webb, and Petitjean 2019) is applied between the two convolution blocks during the decoding phase to limit overfitting. It consists of a random and temporary deactivation of some neurons to avoid complex co-adaptation. During the prediction phase, neurons are reactivated to test the new model. We decided to use dropout layers because our training data are small and imbalanced. Also, dropout has been used in different work as it provide restrictive regularization and enhance generalization (Rajaraman et al. 2020) ;
- **MaxPooling:** The MaxPooling bock allows you to reduce the size of the data during the encoding phase to detect different characteristics. For our network, we reduce the height and width of the features by 2 after all the convolution steps ;
- **Concatenation:** For this layer, the features of the same dimension $h \times w \times n$ of the decoding and encoding step are concatenated before the transpose step ;
- **Transpose:** For the decoding phase, a transpose layer was preferred to an Up-Sampling layer as it is commonly used in encoder/decoder architecture to perform semantic segmentation (Iglavikov and Shvets 2018; Ouyang and Li 2021). Indeed, it is a more complex operation that combines a convolution operation and an upsampling operation within the same layer ;

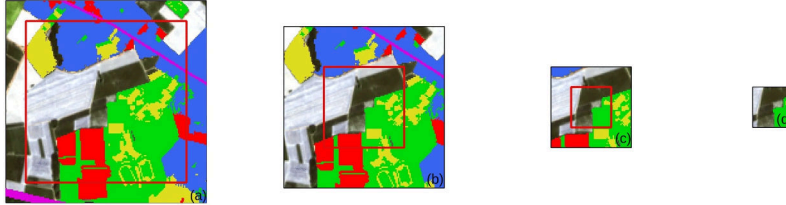


Figure 3. Size comparison of the patches: (a) 160x160 pixels, (b) 128x128 pixels, (c) 64x64 pixels and (d) 32x32 pixels.

The size of the input patches are $160 \times 160 \times 3$ for each of the two networks used, with 5 land use classes at the output. This size allows the network to get a wider spatial context of the land use classes represented in the image (Figure 3). This size allows for a large footprint and diversity of classes on each patch. At the end of the network, a \vec{x} vector of size 160×160 containing the semantic segmentation into 5 classes is produced. It is then normalized using a Softmax function defined below :

$$f(\vec{x})_i = \frac{e^{x_i}}{\sum_j e^{x_j}} \quad (4)$$

where i and j represent respectively the i^{th} and j^{th} class and x_i the probability of belonging to the i class.

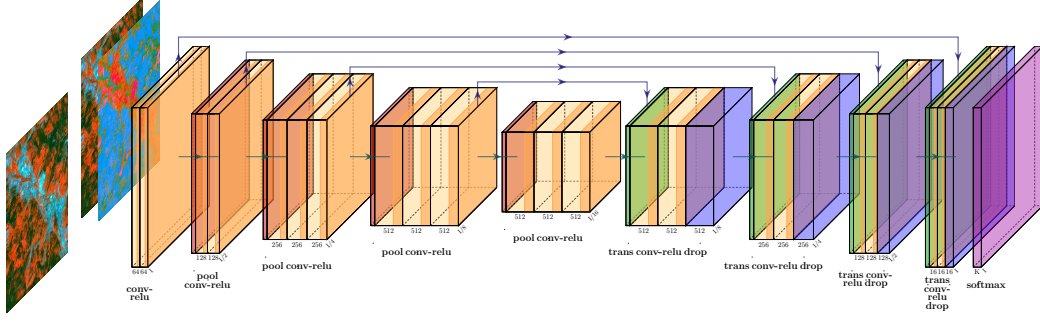


Figure 4. U-Net-IRRG and U-Net-Index architecture used for preliminary tests. This network uses, on the one hand, only IRRG images and, on the other hand, a combination of IRRG images and spectral and textural indexes (NDVI, NDBI and eNDVI).

Several data augmentation methods were applied to the training patches to enrich the dataset. Each patch is thus kept in its initial state and then augmented randomly either by rotations (90, 180, and 270 degrees) or flipping (from left to right or from top to bottom) using *numpy* library⁵. These data augmentation methods doubled the size of the initial training sets of each tile. We also add Dropout and L2 regularization during decoder process (Srivastava et al. 2014) to reduce the chance of overfitting due to the use of an imbalanced dataset and a low number of patches. L2 was preferred to L1 as it more stable by putting half of the weights on each input (Li et al. 2021).

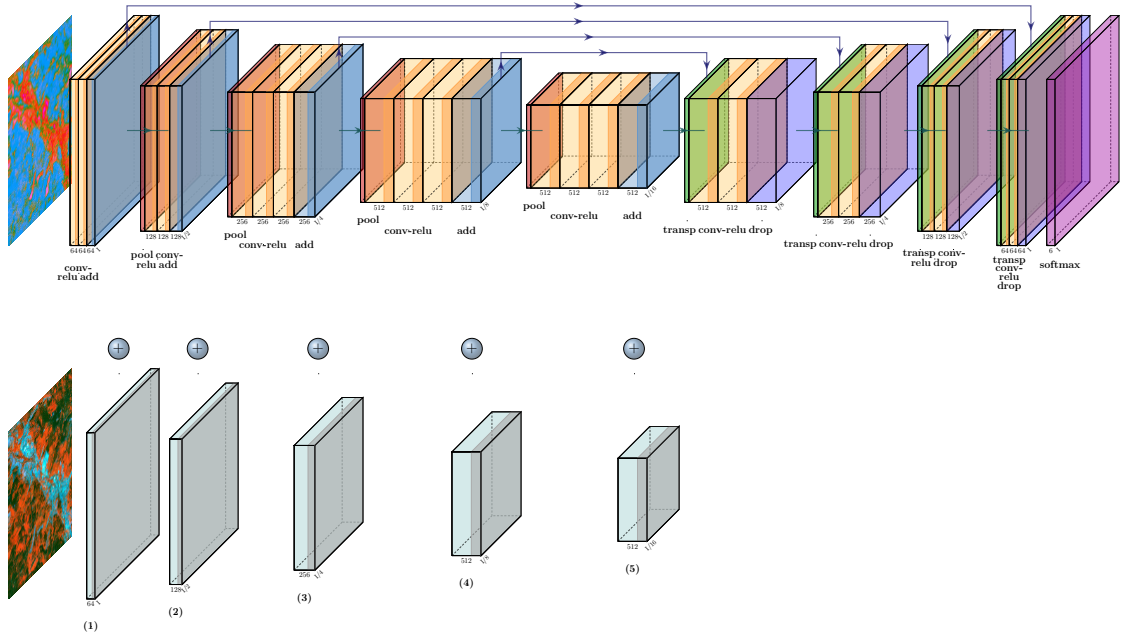


Figure 5. U-Net-Encoder architecture modified in order to apply encoder fusion. Features (1) to (5) are extracted from the second U-Net, pretrained on ImageNet with IRRG patches as inputs, and merged in the first U-Net during the encoding phase.

Every image is normalized by dividing the standard deviation of the reflectance of the spectral band by the difference of this one with the average of the reflectances.

⁵<https://numpy.org/>

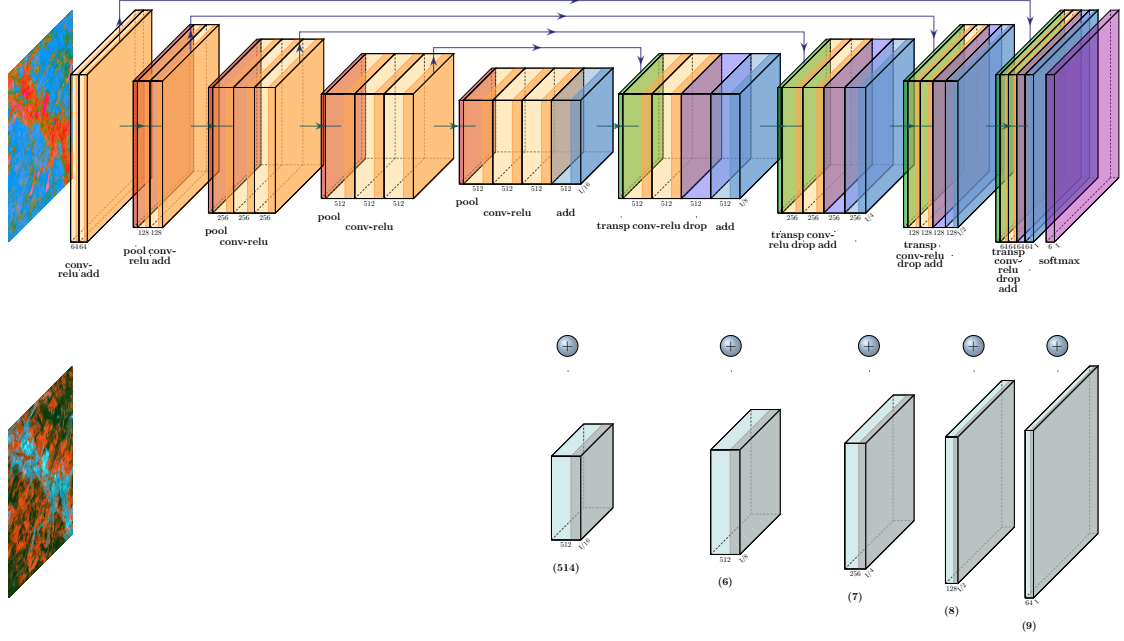


Figure 6. U-Net-Decoder architecture modified in order to apply decoder fusion. Features (5) to (9) are extracted from the second U-Net, pretrained on ImageNet with IRRG patches as inputs, and merged in the first U-Net during the decoding phase.

This normalization method allows the data to be centered on the same range of values so that our gradient remains stable. The normalization formula is developed below:

$$n = \frac{(b - \bar{b})}{\sigma_b} \quad (5)$$

where n represents the normalized spectral band, b the reflectance values of the spectral band, \bar{b} the mean of the reflectance values, and σ_b the standard deviation of the reflectance values.

The models were trained for 100 epochs with a Learning Rate (LR) of 1×10^{-4} and a batch size of 8. We reduced LR by 50% each time a plateau was reached for 5 epochs (Chhor and Aramburu 2017) using Keras callback ReduceLROnPlateau. Softmax was used as an activation function for the last layer of each model to predict multinomial probabilities (Figures 4, 5 and 6).

Training takes around 15 hours for each network.

2.3.3. Post-processing and evaluation

After training the model, predictions are made on selected test areas. The test image must first be reconstructed from 160 x 160-pixel patches. To do this, an overlap of 75% is applied to smooth the predictions and improve the classification results. The overlapped pixels within the overlap regions are averaged and then the index of the band with the highest probability is retained as a prediction. This method is applied to all the test images. Finally, a confusion matrix is computed over the entire image accompanied by a file with detailed statistics by land-use class. This prediction technique can also be applied to the whole image to provide an accurate mapping of land

use in urban structures.

2.4. Loss function

To take into account the low representativeness of certain classes in the dataset (imbalanced dataset), a weighted categorical cross-entropy loss is used by assigning higher weights to the classes with the least surface area. These are the inverse of the class frequency (Audebert, Le Saux, and Lefèvre 2018). This loss is commonly used in remote sensing multi-class supervised classification tasks (Ienco et al. 2017; Zhu et al. 2017). The loss has been taken from *segmentation_models* framework (Yakubovskiy 2019). On the other hand, the "Large Scale Networks" and "Specialized But Vegetative Areas" classes (rare objects composing the urban structures) occupy such small areas that we have decided to assign the lowest weights to another poorly represented "Dense Built-Up" class. Indeed, using too high weights can bias the loss and distort the learning process which would lead to prediction errors (Sefrin, Riese, and Keller 2021; Audebert, Le Saux, and Lefèvre 2018).

2.5. Evaluation metrics

We adopted different evaluation metrics (Maxwell, Warner, and Guillén 2021) to measure the quality assessments and the effectiveness of every network tested in our processing chain : Precision, Recall and $F1_{Score}$.

Precision (also know as User's Accuracy - UA) informs about the fraction of well-classified pixels in the classified image. It can be calculated by dividing True Positives (TP) values with the sum of True Positives (TP) and False Positives (FP).

$$Precision = \frac{TP}{TP + FP} \quad (6)$$

Recall (also know as Producer's Accuracy - PA) indicates the fraction of well-ranked pixels relative to the reference data. It can be calculated by dividing True Positives (TP) values with the sum of True Positives (TP) and False Negatives (FN).

$$Recall = \frac{TP}{TP + FN} \quad (7)$$

$F1_{Score}$ (also known as Dice) represents the harmonic mean between Precision and Recall. It is calculated by dividing twice the product of Precision and Recall by the sum of these same metrics. In order to have a global analysis metric of the results, the $F1_{Score}$ weighted is calculated for all the classes for each test.

$$F1_{Score} = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (8)$$

3. Results

In order to assess the results of the four networks the section 3.1 is dedicated to a global analysis comparing the test sites without distinction of UF. The section 3.2 presents the results for each study areas (for remind, three test sites with a gradient of urbanisation, respectively Strasbourg, Metz and St-Avold city - Figure 1). Then, section 3.3 focuses on the best fusion methods considering qualitative and quantitative results and the last section (3.4) compared this latest with a test mapping UF in four classes.

3.1. Global results analysis

The weighted $F1_{Score}$ and Overall Accuracy has been calculated for each test areas (Tables 5 and 6). We notice a slight advantage for the U-Net-Decoder method for the two least dense cities, Metz and Saint Avold (respectively 0.7488 and 0.7883 for weighted $F1_{Score}$ and 0.7343 and 0.7580 for Overall Accuracy). The U-Net-Encoder method obtains the high value for the city of Strasbourg with a weighted $F1_{Score}$ of 0.5990 and an Overall Accuracy of 0.5794. We can notice that the difference is however really close between U-Net-Encoder/Decoder. A detailed analysis for each UF is then necessary to conclude on their performance to map UF whatever the thematic classes.

Table 5. Results of weighted $F1_{Score}$ for each method in every study area.

U-Net-IRRG			U-Net-Index		
Strasbourg	Metz	St-Avold	Strasbourg	Metz	St-Avold
0.5133	0.7364	0.7716	0.5005	0.7214	0.7248
U-Net-Encoder			U-Net-Decoder		
Strasbourg	Metz	St-Avold	Strasbourg	Metz	St-Avold
0.5990	0.7479	0.7834	0.5894	0.7488	0.7883

Table 6. Results of Overall Accuracy for each method in every study area.

U-Net-IRRG			U-Net-Index		
Strasbourg	Metz	St-Avold	Strasbourg	Metz	St-Avold
0.5004	0.7067	0.7244	0.4737	0.6963	0.6701
U-Net-Encoder			U-Net-Decoder		
Strasbourg	Metz	St-Avold	Strasbourg	Metz	St-Avold
0.5794	0.7299	0.7513	0.5706	0.7343	0.7580

Training and validation loss has been plotted in order to monitor any possible overfitting (Figure 7).

3.2. Results analysis for each UF

A more detailed analysis of the evaluation metrics is presented from the most important city (Strasbourg) to the most rural test site (St-Avold). For each test areas, quantitative analysis based on Precision, Recall and $F1_{Score}$ measures is completed by a qualitative analysis of results with some zooms on the three test areas.

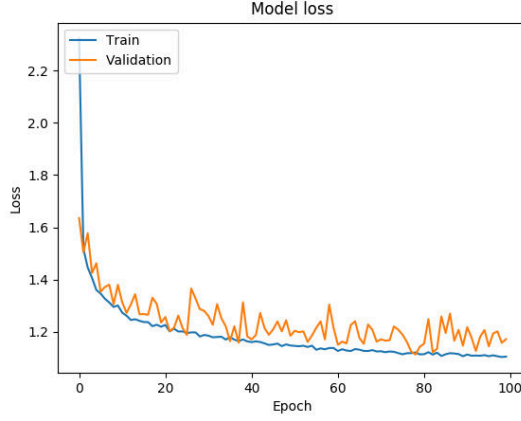


Figure 7. Training and validation learning curve for U-Net-IRRG method perform on 32ULU tile.

3.2.1. Semantic segmentation results for Strasbourg, Grand Est, 32ULU Tile

Table 7 summarizes the three evaluation metrics for the four different methods and for each UF (Table 2) based on Strasbourg test area. We notice an advantage for both fusion methods (U-Net-Encoder and U-Net-Decoder) where the best statistical results are found for both methods. More precisely, these fusion methods improve the results of the specialized but vegetated area (Class 4) and the large scale networks (Class 5), where the $F1_{Score}$ is respectively 0.4716 and 0.5038 for U-Net-Encoder and 0.4716 and 0.5781 for U-Net-Decoder.

Table 7. Results of all methods for the test zone located in Strasbourg, Grand-Est, France.

	U-Net-IRRG			U-Net-Index		
	Precision	Recall	F1	Precision	Recall	F1
Class 1	0.3928	0.6820	0.4985	0.5541	0.3929	0.4598
Class 2	0.6116	0.3752	0.4651	0.7412	0.2832	0.4098
Class 3	0.5045	0.5789	0.5391	0.3749	0.7388	0.4974
Class 4	0.3279	0.3313	0.3296	0.3988	0.3236	0.3573
Class 5	0.2887	0.7543	0.4176	0.2298	0.8325	0.3602
Class 6	0.9876	0.5199	0.6812	0.9798	0.5853	0.7328

	U-Net-Encoder			U-Net-Decoder		
	Precision	Recall	F1	Precision	Recall	F1
Class 1	0.4511	0.6836	0.5435	0.4797	0.6297	0.5446
Class 2	0.7415	0.4651	0.5717	0.7531	0.4179	0.5375
Class 3	0.5169	0.5878	0.5501	0.4773	0.7006	0.5678
Class 4	0.4057	0.5632	0.4716	0.3801	0.6210	0.4716
Class 5	0.3962	0.6917	0.5038	0.5099	0.6673	0.5781
Class 6	0.9882	0.6427	0.7789	0.9880	0.5987	0.7456

To complete these quantitative results, Figure 8 presents some subsets in the Strasbourg test area. Compared to the reference dataset, qualitative analysis shows that the methods U-Net-Encoder (subfigure e) and U-Net-Decoder (subfigure f) provide a better detection of large scale networks (class 5) than the reference data (Figure 8 b). Indeed, some roads and railway are classified while they do not appear on the image

(Figure 8 b). The specialized built-up areas (Class 3) is also well extracted by all methods by detecting areas with ongoing construction that are also not present on the reference data. On the other hand, we notice an overestimation of the Specialized but vegetative areas (class 4) which includes some cropping or forests areas in peri-urban area. U-Net-IRRG and U-Net-Index also produce some confusion between different classes, such as Specialized Built-up Areas (Class 3) and Specialized but vegetative Areas (Class 4), where the two fusion methods have much more unified results.

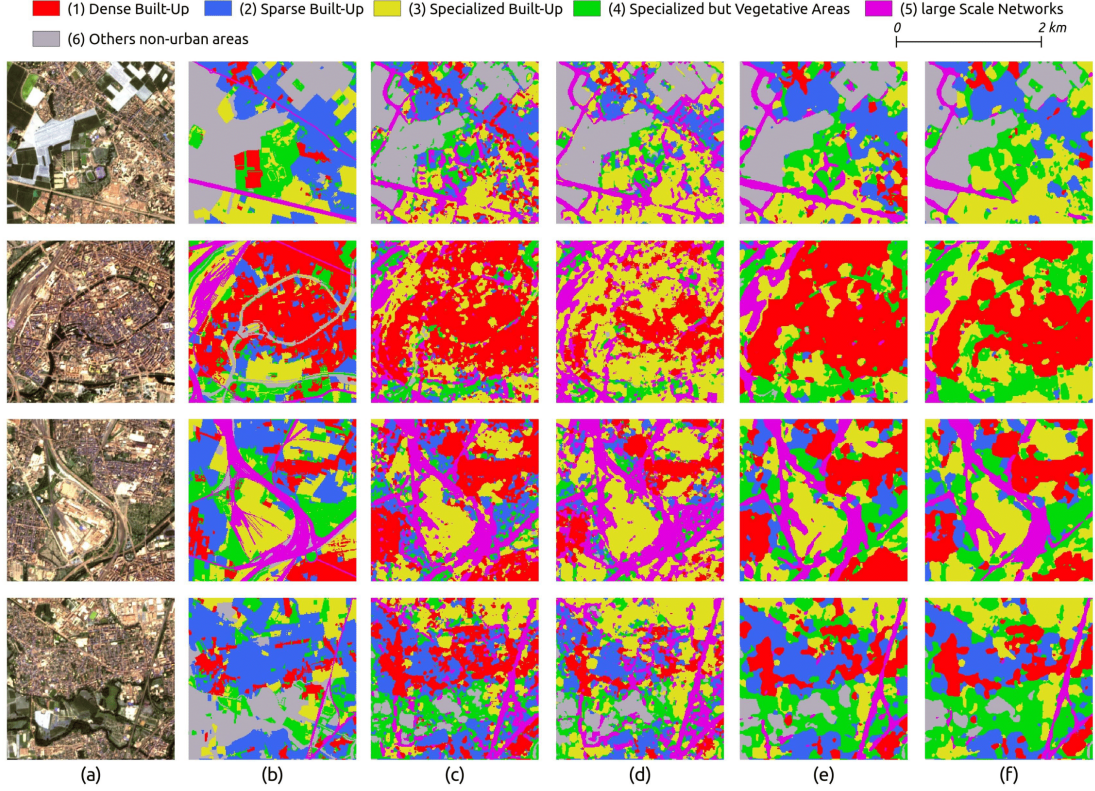


Figure 8. Semantic segmentation results for test zone over Strasbourg. (a) and (b) represent subset image and reference data respectively, (c) and (d) are respectively U-Net-IRRG and U-Net-Index and (e) and (f) are U-Net-Encoder and U-Net-Decoder.

3.2.2. Semantic segmentation Results for Metz, Grand Est, 31UGQ Tile

Table 8 summarized all the statistics calculated by class for each method for the second test area located in Metz and its surroundings.

The statistical results (Table 8) shows that the $F1_{score}$ scores values follow the same trends the results of test area 1 (Strasbourg - section 3.2.1), with class 6 (Others non-urban areas) always higher than the other classes (above 0.85) and classes 1, 2, 3 and 5 always equal or above 0.5. Large Scale network (Class 4) still has low F1 values, even slightly lower than test area 1. This can be explained by the urban morphology of the city characterized with a lot of sparse urban settlements. These trends are confirmed in the precision and recall values. They are not significantly different than in subset 1 but the maximum values are reached here by the U-Net-IRRG model in terms of precision and the U-Net-Encoder model for recall. More precisely, the U-Net-IRRG model underestimates classes 1 to 3 while the U-Net-Encoder model overestimates

Table 8. Results of all methods for the test zone located over Metz, Grand-Est, France.

	U-Net-IRRG			U-Net-Index		
	Precision	Recall	F1	Precision	Recall	F1
Class 1	0.5128	0.7021	0.5199	0.4577	0.5095	0.4822
Class 2	0.6228	0.6941	0.6565	0.5286	0.6586	0.5864
Class 3	0.6177	0.4329	0.5090	0.5385	0.5359	0.5372
Class 4	0.2092	0.6298	0.3141	0.1662	0.2666	0.2048
Class 5	0.4457	0.8559	0.5862	0.2812	0.9090	0.4295
Class 6	0.9643	0.7678	0.8549	0.9595	0.7758	0.8579

	U-Net-Encoder			U-Net-Decoder		
	Precision	Recall	F1	Precision	Recall	F1
Class 1	0.4510	0.5847	0.5092	0.3858	0.6592	0.4867
Class 2	0.6136	0.6948	0.6516	0.6195	0.5962	0.6076
Class 3	0.5228	0.5789	0.5494	0.5026	0.6186	0.5546
Class 4	0.2193	0.4168	0.2874	0.2674	0.3735	0.3117
Class 5	0.4638	0.8471	0.5994	0.4240	0.8791	0.5721
Class 6	0.9601	0.7933	0.8687	0.9587	0.8126	0.8796

classes 2 and 3 (Sparse and Specialized Built-Up areas).

Qualitative results (Figure 9) confirm that U-Net-Encoder and Decoder show better extraction of all the classes. Large Scale Networks (Class 5) is also better detected and new roads not visible in reference data are extracted. The U-Net-IRRG method clearly overestimated class (3). Specialized but Vegetative Areas (class (4) are much better detected visually for both fusion methods than for U-Net-IRRG.

3.2.3. Semantic segmentation results for Saint-Avoid, Grand Est, 31UGQ Tile

For this last test site (Table 9), Saint-Avoid, Grand Est, the trends of the statistical results are identical or even with lower $F1_{Score}$ values particularly for the specialized but Vegetated Areas which is the most complex class. This means that results are lower when the urbanisation gradient decrease, more sparse are urban settlement and more difficult is the extraction.

The qualitative analysis (Figure 10) confirm the quantitative results by showing that extraction of all classes are overestimated for the first two methods (U-Net-IRRG and U-Net-Index). As in the other both test areas, class 5 stays better than the reference data by considering the major part of the large scale networks. Moreover, the fusion methods (U-Net-Encoder and U-Net-Decoder) propose a much more unified and smoothed classification results than the two others methods which is also explained with the recall results.

In order to identify which of the both fusion methods is better to extract these 5 classes related to the test areas chosen for their urbanisation gradient, a more detailed analysis is proposed in the next section.

3.3. Encoder and Decoder fusion analysis

This section presents a detailed analysis based on 2D scatter plots for each metrics where the six classes are located in the 2D-space where the X-axis is the U-Net-Encoder and where the Y-axis is the U-Net-Decoder. Based on this Figure 11, we notice that,

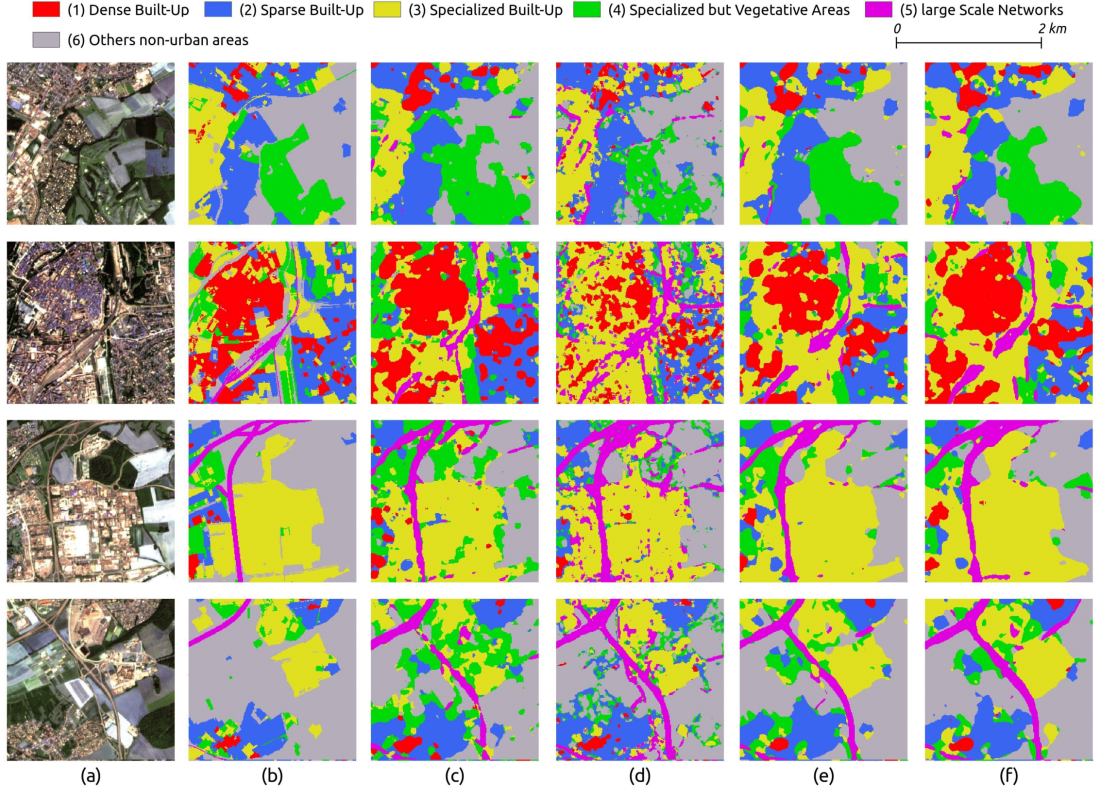


Figure 9. Semantic segmentation results for test zone over Metz. (a) and (b) represent subset image and reference data respectively, (c) and (d) are respectively U-Net-IRRG and U-Net-Index and (e) and (f) are U-Net-Encoder and U-Net-Decoder.

Table 9. Results of all methods for the test area located near Saint-Avoid, Grand-Est, France.

	U-Net-IRRG			U-Net-Index		
	Precision	Recall	F1	Precision	Recall	F1
Class 1	0.2597	0.8114	0.3934	0.3209	0.5785	0.4128
Class 2	0.6233	0.7010	0.6599	0.5202	0.6525	0.5788
Class 3	0.3902	0.4534	0.4194	0.3379	0.5544	0.4199
Class 4	0.0838	0.3414	0.1346	0.0348	0.1378	0.0556
Class 5	0.4571	0.9194	0.6106	0.3364	0.9558	0.4977
Class 6	0.9944	0.7493	0.8546	0.9896	0.6925	0.8148

	U-Net-Encoder			U-Net-Decoder		
	Precision	Recall	F1	Precision	Recall	F1
Class 1	0.2996	0.5668	0.3920	0.2593	0.6958	0.3778
Class 2	0.6367	0.7153	0.6737	0.6359	0.6114	0.6234
Class 3	0.2306	0.6118	0.3350	0.2350	0.6462	0.3446
Class 4	0.1186	0.2455	0.1600	0.1566	0.2320	0.1870
Class 5	0.4633	0.8944	0.6104	0.4363	0.9091	0.5896
Class 6	0.9933	0.7765	0.8716	0.9930	0.8034	0.8882

for all the metrics studied, UF classes are close to the diagonal line, which means that the both methods give close statistical results. Values of each metrics are concentrated around 0.5 with a Recall always slightly higher than $F1_{score}$ with the U-Net-Decoder.

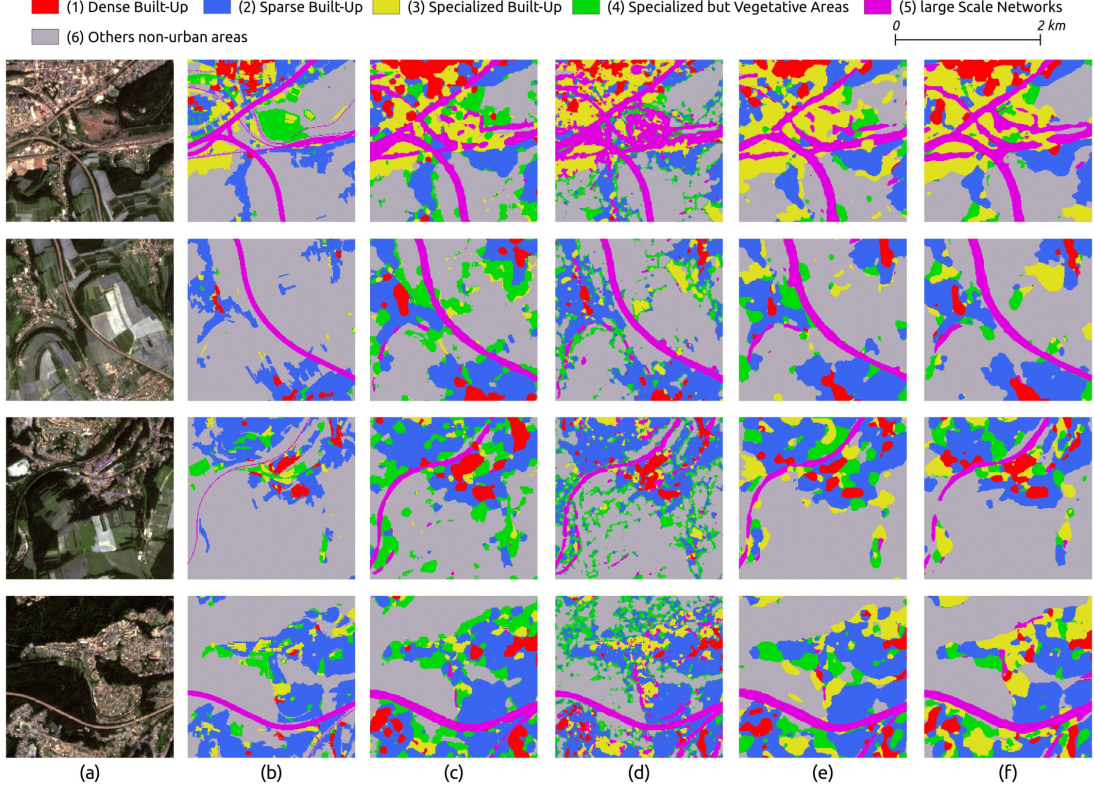


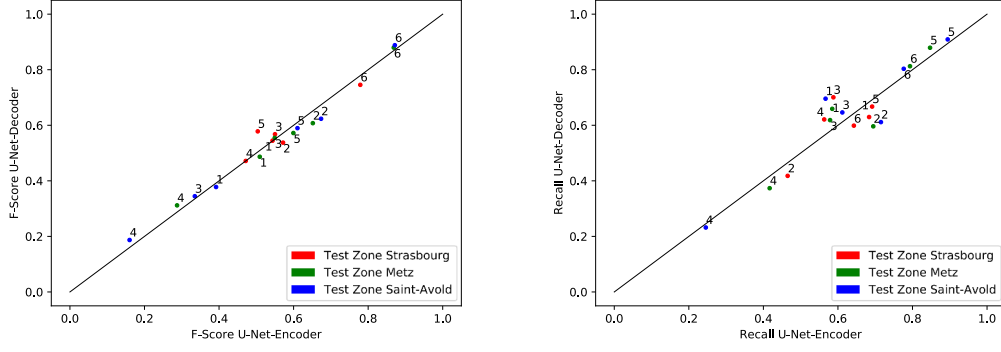
Figure 10. Semantic segmentation results for test zone over Saint-Avold. (a) and (b) represent subset image and reference data respectively, (c) and (d) are respectively U-Net-IRRG and U-Net-Index and (e) and (f) are U-Net-Encoder and U-Net-Decoder.

Only the precision values are similar for the classes 1,2,3,5 and confirm that the class 4 is very difficult to extract but the U-Net-Decoder is the model for which all metrics are higher (Figure 13).

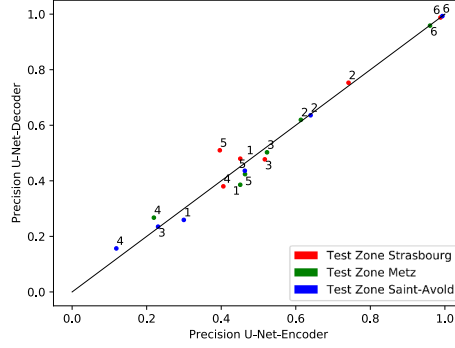
Figure 12 allows for a more in-depth analysis of the two fusion methods (U-Net-Encoder and U-Net-Decoder). The subset (i) focus on a road not mapped in the reference data because it is not considered as primary road. However, both models detect it due to his width. The subset (ii) highlights a better precision in the delimitation of the Specialized but Vegetative Areas (class 4) for the U-Net-Decoder method. This same observation also applies for the subsets (iii) and (iv) where the Dense Built-up (class 1) and Specialized Built-Up Areas (class 3) are better extracted for the U-Net-Decoder method.

3.4. Four-classes results for the Strasbourg study area

In order to analyse the impact of the number and choice of UF classes on the classification results, we have tested the both models (U-Net-Encoder and U-Net-Decoder) with only five classes by removing the most complex one at 10m spatial resolution (Specialized but Vegetative Areas - class 4) and grouping areas in the class 6 (Other). We notice that the results (Table 10) of $F1_{Score}$ are always higher than 0.5 for all classes only with U-Net-Encoder and U-Net-Decoder. Precision metrics shows that Large Scale Network (class 5) is always well detected and Specialized Built-Up Areas (class 3) is slightly better extracted. Recall metrics confirm that Dense Built-Up



(a) $F1_{Score}$ for each class/test zones of the two fusion methods (U-Net-Encoder and U-Net-Decoder) (b) Recall for each class/test zones of the two fusion methods (U-Net-Encoder and U-Net-Decoder)



(c) Precision for each class/test zones of the two fusion methods (U-Net-Encoder and U-Net-Decoder)

Figure 11. $F1_{Score}$, Recall and Precision for the analysis of the two fusion methods at the three study sites.

(class 1) and Specialized Built-Up Areas (class 3) are more overestimated with U-Net-Encoder than U-Net-Decoder. These statistical results are also visible in (Figure 14). Finally, quantitative and qualitative interpretation results show the trends in results with very similar analysis with four or five UF classes that confirmed the slight superiority of the U-Net-Decoder.

4. Discussion

The experiments performed in this study have tested two semantic segmentation networks (U-Net) that we combined using feature fusion between a from scratch network and a pre-trained network on ImageNet, to propose a generic UF mapping adapted to Grand-Est/France cities. First, the results showed the advantage of both fusion methods for the detection of these classes with quantitative and qualitative assessment showing better results for U-Net-Decoder method. Moreover, the contribution of exogenous indexes coupled with a pre-trained network allowed to refine the classification of the five UF classes.

Statistical and qualitative results confirm the good results of the U-Net-Decoder method over the U-Net-Encoder one to extract several UF classes with a mono-

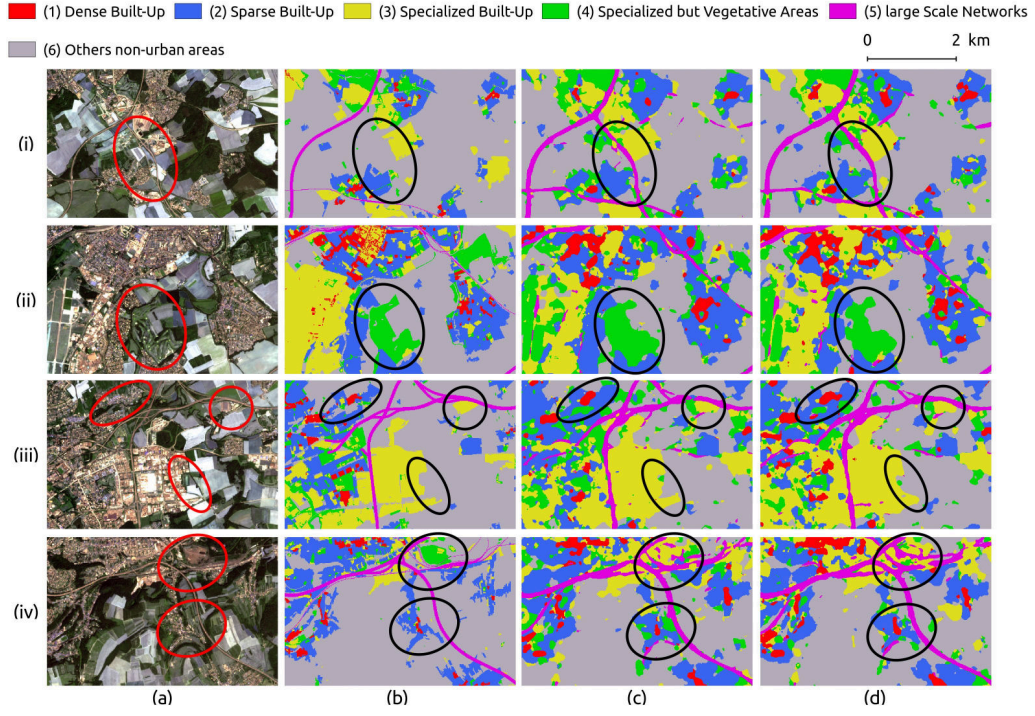


Figure 12. Four subsets are presented, from (1) to (4) with (a) Sentinel-2 zoom, (b) reference data, (c) U-Net-Encoder and (d) U-Net-Decoder.

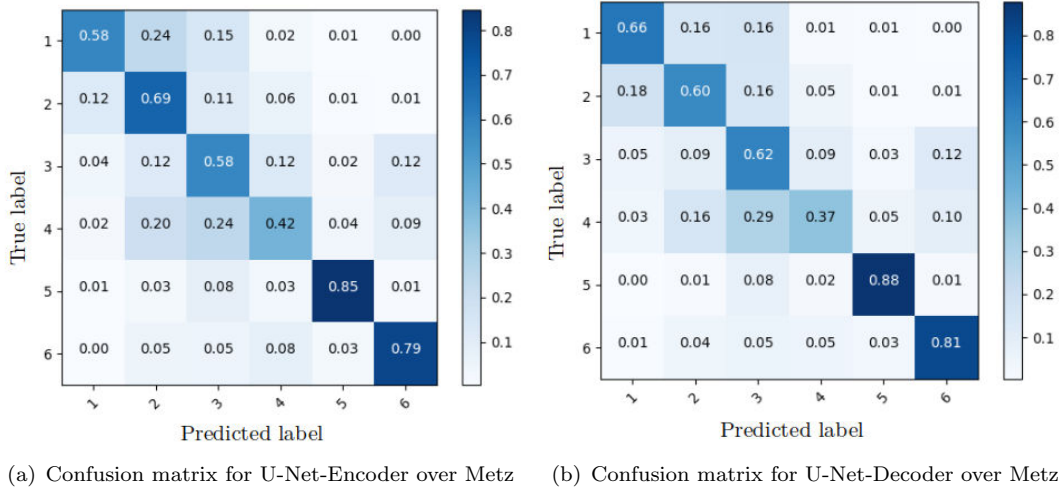


Figure 13. Confusion matrix (with Recall metric inside each cell) for U-Net-Encoder and U-Net-Decoder over the test area of Metz.

temporal image. It is possible to notice a better detection of some classes compared to the reference data. This is the case for Large Scale Networks (class 5) where roads and railroads, not yet or barely built at the time of the creation of the reference data, are well detected. Indeed, the selection of the Large Scale Networks was made according to a level of road in a French national database. Thus, there may be some roads of a

Table 10. Four UF classes results of all methods for the test area in Strasbourg, Grand-Est, France.

	U-Net-IRRG			U-Net-Index		
	Precision	Recall	F1	Precision	Recall	F1
Class 1	0.4674	0.6066	0.5281	0.4032	0.5232	0.4554
Class 2	0.6164	0.3534	0.4492	0.5035	0.3080	0.3822
Class 3	0.3696	0.6941	0.4824	0.3992	0.7110	0.5113
Class 5	0.2867	0.8279	0.4259	0.2667	0.8354	0.4043
Class 6	0.9844	0.5170	0.6779	0.9930	0.4908	0.6569

	U-Net-Encoder			U-Net-Decoder		
	Precision	Recall	F1	Precision	Recall	F1
Class 1	0.4426	0.6752	0.5347	0.4897	0.5544	0.5201
Class 2	0.5393	0.5841	0.5608	0.5629	0.5838	0.5732
Class 3	0.4268	0.6892	0.5271	0.3958	0.7362	0.5148
Class 5	0.4998	0.6276	0.5565	0.4585	0.6835	0.5488
Class 6	0.9907	0.4901	0.6558	0.9914	0.5060	0.6700

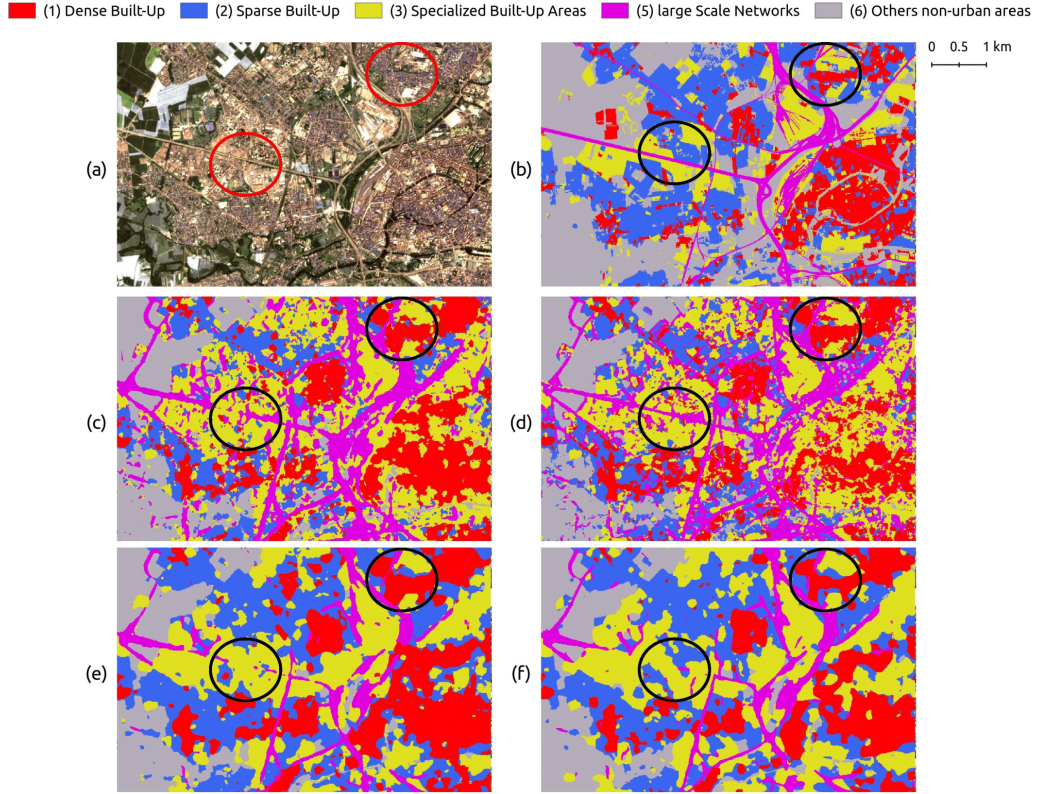


Figure 14. Semantic segmentation results for test area over Strasbourg in 4 UF classes with (a) S2, (b) reference data, (c) U-Net-IRRG, (d) U-Net-Index and respectively (e) and (f) with U-Net-Encoder and U-Net-Decoder.

lower level that are still visible at 10m spatial resolution. But for most of the roads available at this level, they are not visible at this resolution. On the other hand, Specialized But Vegetative Areas (class 4) offers rather low scores rarely exceeding 0.40 of $F1_{score}$. This is due to the complexity of this class at 10m where vegetated areas are dominant. The reflectance of these surfaces is quite similar to other vegetated

areas such as crops, grasslands or small forests summarized in the Other non-urban areas (class 6). However, Recall values reach a value of more than 0.6 with the U-Net-Decoder model close to the other Recall values. This class (4) is also best detected when the areas is highly urbanized such as in the center of Strasbourg (test area 1) or Metz (test area 2). Indeed, green areas inside cities are very often belong to class (4) as they are mostly urban parks, leisure activities areas, campgrounds or cemetery. These surfaces are rather composed of green areas than mineral surfaces. They are most often detected as Specialised but Vegetative areas (class 4).

We also notice that the first methods (U-Net-IRRG) have many geometric errors in the delimitation of classes. In fact, the delimitation of class (4) has a significant number of overestimation. Each fusion method proposes a better estimation of these surfaces and can be seen with the $F1_{score}$ metric. This is notably the case of U-Net-Decoder which takes them better into consideration by limiting the overestimation, as seen during the qualitative analysis. This is explained by the combination of intermediate features coming from the pre-trained network and bringing information in the delimitation of these surfaces. However, these surfaces remain complex to detect using mono-temporal imagery, even if the contribution of features from another network improves the qualitative and quantitative results. Moreover, the ImageNet weights have been initialized on RGB images. It could be interesting to test the impact of RGB bands compared to IRRG bands when using a pre-trained network.

Indeed at 10m, confusions appears between Dense Built-up (1) and Sparse Built-Up (2) classes due to the low distance between buildings not visible at this resolution. Others confusions exist between Specialized but Vegetative Areas (4) and Others non-urban areas (6) due to the presence of vegetation on the urban fringe. Despite the low number of Large Scale Networks patches (5) (Table 3), this class obtains encouraging scores and the networks are also able to classify roads that do not initially appear in the reference data (Figure 9). Overall, results remain encouraging despite the use of mono-temporal imagery. Nevertheless, the addition of multi-temporal and multi-source data should improve the current results even if urban environments have a low intra-temporal variability compared to natural areas.

The results from the both methods U-Net-IRRG and U-Net-Index give more heterogeneous classification results. The contribution of the feature fusion allows to obtain smoother UF classes with a better delimitation between the classes. This can be explained by the contribution of feature of a network taking in input only spectral and textural indices which allow to have a more precise information on the delimitation of urban classes. On the other hand, the network using spectral bands allows to extract more features allowing the distinction between the different UF classes. The U-Net has been designed to learn its own spatial filters. Thus, the contribution of eNDVI in conducted methods could be explored since the entropy is nothing else than a set of convolutions on a grayscale image with a sliding window.

Even with four classical classes rather than five UF classes, the results remain similar to the previous ones. This shows the interest of studying UFs as 5 classes rather than 4 because the Specialized but Vegetative Areas class is needed by end-users and is an integral part of urban areas.

5. Conclusions and perspectives

The objective of this research was to show the interest of semantic segmentation methods to help end-users to produce a relevant and up-to-date map of UF in five thematic

classes only with an optical mono-temporal and high spatial resolution image (Sentinel-2). Indeed, classically with a high spatial resolution (10m), UF are only mapped with four classes distinguishing network from dense, sparse built-up areas and activities. In this paper, two methods using features fusion techniques (U-Net-Encoder and U-Net-Decoder) between two networks have been developed using (i) three spectral bands (green, red, NIR) for the pre-trained network and (ii) three spectral and textural indexes (NDBI, NDVI and $eNDVI$) for the non-pre-trained network. These methods were compared with a U-Net taking into account either IRRG or IRRG bands and three spectral and textural indexes. The idea was to combine two types of data, each providing various information to improve the detection of urban surfaces proposed in five different UF classes adapted to Grand-Est/France cities: (1) Dense Built-Up, (2) Sparse Built-Up, (3) Specialized Built-Up Areas, (4) Specialized but Vegetative Areas, (5) Large Scale Networks. Methods have been tested on a gradient of urban areas in the East of France to ensure a generalisation of results.

Those highlight that both fusion methods, especially the U-Net-Decoder one for most of UF, offer the best results in refining the detection of the most UF classes. The U-Net-Decoder method showed an advantage in the delimitation of Specialized but Vegetative Areas and a better classification of these areas for highly urbanized areas (Strasbourg center and Metz). The qualitative analysis confirm these first analysis by showing an advantage for the U-Net-Decoder method with a better segmentation of the different UFs. On the other hand, the statistical results showed a better but close classification of the Dense Built-Up and Sparse Built-Up for the U-Net-Encoder method. Most of UF classes are always extracted with relevant evaluation metrics (greater than 0.5) and a qualitative interpretation of the results show the extraction homogeneous patches of classes. The weakness of the results for the Class 4 which is confused with other land cover classes due its relative complexity, could be improve by using multi-temporal imagery in order to take into account of the vegetation dynamics of this class. These first relevant results could be also applied to other study sites in France to confirm our conclusions.

This opens several perspectives for the use of optical times series imagery in order to take into account the spatio-temporal dynamic of UF. An other issue would be to integrate the multivariate properties of UF from Sentinel-1 imagery which has already demonstrated its interest for mapping urban footprint. Indeed, many works show the interest of using the radar amplitude for the detection of several UF classes. Our next research will focus on the addition of these multi-temporal optical and radar data for the mapping of these UF classes.

Acknowledgements

All the network figures were designed using PlotNeuralNet (Iqbal 2018). We thanks to the Spatial Data Infrastructure GeoGrandEst provided the reference data used in this study and the Theia Services and Data Infrastructure for the Sentinel-2A imagery.

Funding

This research is part of the PhD Thesis supported 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) and by the French TOSCA project AIMCEE (CNES, 2019-2022).

References

- Amin Larabi, Mohammed El, Souleyman Chaib, Khadidja Bakhti, and Moussa Sofiane Karoui. 2019. "Transfer Learning for Changes Detection in Optical Remote Sensing Imagery." In *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*, 1582–1585.
- Audebert, Nicolas, Bertrand Le Saux, and Sébastien Lefèvre. 2018. "Beyond RGB: Very High Resolution Urban Remote Sensing With Multimodal Deep Networks." *ISPRS Journal of Photogrammetry and Remote Sensing* 140: 20–32. <https://hal.archives-ouvertes.fr/hal-01636145>.
- Badrinarayanan, Vijay, Alex Kendall, and Roberto Cipolla. 2017. "SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39 (12): 2481–2495.
- Campos-Taberner, Manuel, Francisco Javier García-Haro, Beatriz Martínez, Emma Izquierdo-Verdiguier, Clement Atzberger, Gustau Camps-Valls, and María Amparo Gilabert. 2020. "Understanding deep learning in land use classification based on Sentinel-2 time series." *Scientific Reports* 10 (1): 17188. <https://doi.org/10.1038/s41598-020-74215-5>.
- Chen, Kaiqiang, Kun Fu, Menglong Yan, Xin Gao, Xian Sun, and Xin Wei. 2018. "Semantic Segmentation of Aerial Images With Shuffling Convolutional Neural Networks." *IEEE Geoscience and Remote Sensing Letters* 15 (2): 173–177.
- Chhor, Guillaume, and Cristian Bartolome Aramburu. 2017. "Satellite Image Segmentation for Building Detection using U-net." <http://cs229.stanford.edu/proj2017/final-posters/5148174.pdf>. [Online; accessed 20-March-2021].
- Chollet, François, et al. 2015. "Keras." <https://github.com/fchollet/keras>.
- De Luca, Giandomenico, João M. N. Silva, Sofia Cerasoli, João Araújo, José Campos, Salvatore Di Fazio, and Giuseppe Modica. 2019. "Object-Based Land Cover Classification of Cork Oak Woodlands using UAV Imagery and Orfeo ToolBox." *Remote Sensing* 11 (10). <https://www.mdpi.com/2072-4292/11/10/1238>.
- Ding, Peng, Ye Zhang, Wei-Jian Deng, Ping Jia, and Arjan Kuijper. 2018. "A light and faster regional convolutional neural network for object detection in optical remote sensing images." *ISPRS Journal of Photogrammetry and Remote Sensing* 141: 208–218. <https://www.sciencedirect.com/science/article/pii/S0924271618301382>.
- Drusch, M., U. Del Bello, S. Carlier, O. Colin, V. Fernandez, F. Gascon, B. Hoersch, et al. 2012. "Sentinel-2: ESA's Optical High-Resolution Mission for GMES Operational Services." *Remote Sensing of Environment* 120: 25–36. The Sentinel Missions - New Opportunities for Science, <https://www.sciencedirect.com/science/article/pii/S0034425712000636>.
- El Mendili, Lamiae, Anne Puissant, Mehdi Chougrad, and Imane Sebari. 2020. "Towards a Multi-Temporal Deep Learning Approach for Mapping Urban Fabric Using Sentinel 2 Images." *Remote Sensing* 12 (3). <https://www.mdpi.com/2072-4292/12/3/423>.
- Han, Chaoyi, Yiping Duan, Xiaoming Tao, and Jianhua Lu. 2019. "Dense Convolutional Networks for Semantic Segmentation." *IEEE Access* 7: 43369–43382.
- Haralick, Robert M., K. Shanmugam, and Its'Hak Dinstein. 1973. "Textural Features for Image Classification." *IEEE Transactions on Systems, Man, and Cybernetics* SMC-3 (6): 610–621.
- Hazirbas, Caner, Lingni Ma, Csaba Domokos, and Daniel Cremers. 2016. "FuseNet: Incorporating Depth into Semantic Segmentation via Fusion-Based CNN Architecture." 11.
- Hu, Jingliang, Lichao Mou, Andreas Schmitt, and Xiao Xiang Zhu. 2017. "FusioNet: A two-stream convolutional neural network for urban scene classification using PolSAR and hyperspectral data." In *2017 Joint Urban Remote Sensing Event (JURSE)*, 1–4.
- Huang, Fenghua, Ying Yu, and Tinghao Feng. 2019. "Automatic Extraction of Urban Imper-

- vious Surfaces Based on Deep Learning and Multi-Source Remote Sensing Data.” *Journal of Visual Communication and Image Representation* 60.
- Ienco, Dino, Raffaele Gaetano, Claire Dupaquier, and Pierre Maurel. 2017. “Land Cover Classification via Multitemporal Spatial Data by Deep Recurrent Neural Networks.” *IEEE Geoscience and Remote Sensing Letters* 14 (10): 1685–1689.
- Ienco, Dino, Roberto Interdonato, Raffaele Gaetano, and Dinh Ho Tong Minh. 2019. “Combining Sentinel-1 and Sentinel-2 Satellite Image Time Series for land cover mapping via a multi-source deep learning architecture.” *ISPRS Journal of Photogrammetry and Remote Sensing* 158: 11–22. <https://www.sciencedirect.com/science/article/pii/S0924271619302278>.
- Iglovikov, Vladimir, and Alexey Shvets. 2018. “TernausNet: U-Net with VGG11 Encoder Pre-Trained on ImageNet for Image Segmentation.” .
- Inglada, Jordi, Arthur Vincent, Marcela Arias, Benjamin Tardy, David Morin, and Isabel Rodes. 2017. “Operational High Resolution Land Cover Map Production at the Country Scale Using Satellite Image Time Series.” *Remote Sensing* 9 (1). <https://www.mdpi.com/2072-4292/9/1/95>.
- Iqbal, Haris. 2018. “HarisIqbal88/PlotNeuralNet v1.0.0.” <https://doi.org/10.5281/zenodo.2526396>.
- Irwin, Elena G., and Nancy E. Bockstael. 2007. “The evolution of urban sprawl: Evidence of spatial heterogeneity and increasing land fragmentation.” *Proceedings of the National Academy of Sciences* 104 (52): 20672–20677. <https://www.pnas.org/content/104/52/20672>.
- Jie, Yongshi, Xianhua Ji, Anzhi Yue, Jingbo Chen, Yupeng Deng, Jing Chen, and Yi Zhang. 2020. “Combined Multi-Layer Feature Fusion and Edge Detection Method for Distributed Photovoltaic Power Station Identification.” *Energies* 13 (24). <https://www.mdpi.com/1996-1073/13/24/6742>.
- Kamga, G. A. Fotso, L Bitjoka, T. Akram, A. Mengue Mbom, S. Rameez Naqvi, and Y. Bouroubi. 2021. “Advancements in satellite image classification : methodologies, techniques, approaches and applications.” *International Journal of Remote Sensing* 42 (20): 7662–7722. <https://doi.org/10.1080/01431161.2021.1954261>.
- Kemker, Ronald, Carl Salvaggio, and Christopher Kanan. 2018. “Algorithms for semantic segmentation of multispectral remote sensing imagery using deep learning.” *ISPRS Journal of Photogrammetry and Remote Sensing* 145: 60–77. <http://dx.doi.org/10.1016/j.isprsjprs.2018.04.014>.
- Li, Junjie, Yizhuo Meng, Donyu Dorjee, Xiaobing Wei, Zhiyuan Zhang, and Wen Zhang. 2021. “Automatic Road Extraction From Remote Sensing Imagery Using Ensemble Learning and Postprocessing.” *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 14: 10535–10547.
- Liu, Yansong, Sankaranarayanan Piramanayagam, Sildomar T. Monteiro, and Eli Saber. 2017. “Dense Semantic Labeling of Very-High-Resolution Aerial Imagery and LiDAR with Fully-Convolutional Neural Networks and Higher-Order CRFs.” In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 1561–1570.
- Lu, Jie, Vahid Behbood, Peng Hao, Hua Zuo, Shan Xue, and Guangquan Zhang. 2015. “Transfer learning using computational intelligence: A survey.” *Knowledge-Based Systems* 80: 14–23. 25th anniversary of Knowledge-Based Systems, <https://www.sciencedirect.com/science/article/pii/S0950705115000179>.
- Ma, Lei, Yu Liu, Xueliang Zhang, Yuanxin Ye, Gaofer Yin, and Brian Alan Johnson. 2019. “Deep learning in remote sensing applications: A meta-analysis and review.” *ISPRS Journal of Photogrammetry and Remote Sensing* 152: 166–177. <https://www.sciencedirect.com/science/article/pii/S0924271619301108>.
- Marcos, Diego, Michele Volpi, Benjamin Kellenberger, and Devis Tuia. 2018. “Land cover mapping at very high resolution with rotation equivariant CNNs: Towards small yet accurate models.” *ISPRS Journal of Photogrammetry and Remote Sensing* 145: 96–107. <http://dx.doi.org/10.1016/j.isprsjprs.2018.01.021>.

- Maxwell, Aaron E., Timothy A. Warner, and Luis Andrés Guillén. 2021. "Accuracy Assessment in Convolutional Neural Network-Based Deep Learning Remote Sensing Studies—Part 1: Literature Review." *Remote Sensing* 13 (13). <https://www.mdpi.com/2072-4292/13/13/2450>.
- Mhangara, Paidamwoyo, and John Odindi. 2012. "Potential of texture-based classification in urban landscapes using multispectral aerial photos." *South African Journal of Science* 109: 1–8.
- Momeni, Rahman, Paul Aplin, and Doreen S. Boyd. 2016. "Mapping Complex Urban Land Cover from Spaceborne Imagery: The Influence of Spatial Resolution, Spectral Band Set and Classification Approach." *Remote Sensing* 8 (2). <https://www.mdpi.com/2072-4292/8/2/88>.
- Oquab, Maxime, Leon Bottou, Ivan Laptev, and Josef Sivic. 2014. "Learning and Transferring Mid-level Image Representations Using Convolutional Neural Networks." In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 1717–1724.
- Ouyang, Song, and Yansheng Li. 2021. "Combining Deep Semantic Segmentation Network and Graph Convolutional Neural Network for Semantic Segmentation of Remote Sensing Imagery." *Remote Sensing* 13 (1). <https://www.mdpi.com/2072-4292/13/1/119>.
- Pelletier, Charlotte, Geoffrey I. Webb, and François Petitjean. 2019. "Temporal Convolutional Neural Network for the Classification of Satellite Image Time Series." *Remote Sensing* 11 (5). <https://www.mdpi.com/2072-4292/11/5/523>.
- Puissant, Anne, Nicolas Lachiche, Grzegorz Skupinski, Agnès Braud, Julien Perret, and Annabelle Mas. 2011. "Classification et évolution des tissus urbains à partir de données vectorielles." *Revue internationale de géomatique* 21: 513–532.
- Rajaraman, Sivaramakrishnan, Sudhir Sornapudi, Philip O Alderson, Les R Folio, and Sameer K Antani. 2020. "Interpreting Deep Ensemble Learning through Radiologist Annotations for COVID-19 Detection in Chest Radiographs." *medRxiv* <https://www.medrxiv.org/content/early/2020/07/16/2020.07.15.20154385>.
- Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. 2015. "U-Net: Convolutional Networks for Biomedical Image Segmentation." .
- Rouse, J.W. 1973. "Monitoring the Vernal Advancement and Retrogradation (Green Wave Effect) of Natural Vegetation." *NASA/GSFC Type III Final report* .
- Russakovsky, Olga, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, et al. 2015. "ImageNet Large Scale Visual Recognition Challenge." *International Journal of Computer Vision* 115 (3): 211–252. <http://dx.doi.org/10.1007/s11263-015-0816-y>.
- Saraiva, Marciano, Églen Protas, Moisés Salgado, and Carlos Souza. 2020. "Automatic Mapping of Center Pivot Irrigation Systems from Satellite Images Using Deep Learning." *Remote Sensing* 12 (3). <https://www.mdpi.com/2072-4292/12/3/558>.
- Schmitt, Michael, Lloyd Haydn Hughes, Chunping Qiu, and Xiao Xiang Zhu. 2019. "SEN12MS – A Curated Dataset of Georeferenced Multi-Spectral Sentinel-1/2 Imagery for Deep Learning and Data Fusion." .
- Sefrin, Oliver, Felix M. Riese, and Sina Keller. 2021. "Deep Learning for Land Cover Change Detection." *Remote Sensing* 13 (1). <https://www.mdpi.com/2072-4292/13/1/78>.
- Shendryk, Yuri, Yannik Rist, Catherine Ticehurst, and Peter Thorburn. 2019. "Deep learning for multi-modal classification of cloud, shadow and land cover scenes in PlanetScope and Sentinel-2 imagery." *ISPRS Journal of Photogrammetry and Remote Sensing* 157: 124–136. <https://www.sciencedirect.com/science/article/pii/S0924271619302023>.
- Shin, Hoo-Chang, Holger R. Roth, Mingchen Gao, Le Lu, Ziyue Xu, Isabella Nogues, Jianhua Yao, Daniel Mollura, and Ronald M. Summers. 2016. "Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning." .
- Souza-Filho, Pedro Walfir M., Wilson R. Nascimento, Diogo C. Santos, Eliseu J. Weber, Renato O. Silva, and José O. Siqueira. 2018. "A GEOBIA Approach for Multitemporal Land-Cover and Land-Use Change Analysis in a Tropical Watershed in the Southeastern

- Amazon." *Remote Sensing* 10 (11). <https://www.mdpi.com/2072-4292/10/11/1683>.
- Srivastava, Nitish, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. "Dropout: A Simple Way to Prevent Neural Networks from Overfitting." *Journal of Machine Learning Research* 15 (56): 1929–1958. <http://jmlr.org/papers/v15/srivastava14a.html>.
- Su, Wei, Jing Li, Yun Chen, Zhigang Liu, Jinshui Zhang, Tsuey Low, Inbaraj Suppiah, and Atikah Hashim. 2008. "Textural and local spatial statistics for the object-oriented classification of urban areas using high resolution imagery." *International Journal of Remote Sensing - INT J REMOTE SENS* 29: 3105–3117.
- Sumbul, Gencer, Marcela Charfuelan, Begum Demir, and Volker Markl. 2019. "Bigearthnet: A Large-Scale Benchmark Archive for Remote Sensing Image Understanding." *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium* <http://dx.doi.org/10.1109/IGARSS.2019.8900532>.
- Sumbul, Gencer, Arne de Wall, Tristan Kreuziger, Filipe Marcelino, Hugo Costa, Pedro Benedites, Mario Caetano, Begum Demir, and Volker Markl. 2021. "BigEarthNet-MM: A Large-Scale, Multimodal, Multilabel Benchmark Archive for Remote Sensing Image Classification and Retrieval [Software and Data Sets]." *IEEE Geoscience and Remote Sensing Magazine* 9 (3): 174–180. <http://dx.doi.org/10.1109/MGRS.2021.3089174>.
- Uddin, Kabir, Mir Abdul Matin, and Sajana Maharjan. 2018. "Assessment of Land Cover Change and Its Impact on Changes in Soil Erosion Risk in Nepal." *Sustainability* 10 (12). <https://www.mdpi.com/2071-1050/10/12/4715>.
- United Nations Department of Economic and Social Affairs Population Division. 2018. "The World's Cities in 2018." <https://population.un.org/wup/Publications/Files/WUP2018-Report.pdf>. [Online; accessed 09-February-2021].
- Wurm, Michael, Thomas Stark, Xiao Xiang Zhu, Matthias Weigand, and Hannes Taubenböck. 2019. "Semantic segmentation of slums in satellite images using transfer learning on fully convolutional neural networks." *ISPRS Journal of Photogrammetry and Remote Sensing* 150: 59–69. <https://www.sciencedirect.com/science/article/pii/S0924271619300383>.
- Xie, Michael, Neal Jean, Marshall Burke, David Lobell, and Stefano Ermon. 2016. "Transfer Learning from Deep Features for Remote Sensing and Poverty Mapping." .
- Xing, Yinghui, Min Wang, Shuyuan Yang, and Licheng Jiao. 2018. "Pan-sharpening via deep metric learning." *ISPRS Journal of Photogrammetry and Remote Sensing* 145: 165–183. Deep Learning RS Data, <https://www.sciencedirect.com/science/article/pii/S0924271618300212>.
- Yakubovskiy, Pavel. 2019. "Segmentation Models." https://github.com/qubvel/segmentation_models.
- Yi, Zhao, and Xu Jianhui. 2016. "Impervious surface extraction with Linear Spectral Mixture Analysis integrating Principal components analysis and Normalized Difference Building Index." In *2016 4th International Workshop on Earth Observation and Remote Sensing Applications (EORSA)*, 428–432.
- Zha, Yong, Jingqing Gao, and S. Ni. 2003. "Use of normalized difference built-up index in automatically mapping urban areas from TM imagery." *International Journal of Remote Sensing - INT J REMOTE SENS* 24: 583–594.
- Zhang, Pengbin, Yinghai Ke, Zhenxin Zhang, Mingli Wang, Peng Li, and Shuangyue Zhang. 2018. "Urban Land Use and Land Cover Classification Using Novel Deep Learning Models Based on High Spatial Resolution Satellite Imagery." *Sensors* 18 (11). <https://www.mdpi.com/1424-8220/18/11/3717>.
- Zhong, Yanfei, Xiaobing Han, and Liangpei Zhang. 2018. "Multi-class geospatial object detection based on a position-sensitive balancing framework for high spatial resolution remote sensing imagery." *ISPRS Journal of Photogrammetry and Remote Sensing* 138: 281–294. <https://www.sciencedirect.com/science/article/pii/S0924271618300492>.
- Zhu, Lin, Yushi Chen, Pedram Ghamisi, and Jón Atli Benediktsson. 2018. "Generative Adversarial Networks for Hyperspectral Image Classification." *IEEE Transactions on Geoscience*

- and Remote Sensing* 56 (9): 5046–5063.
- Zhu, Xiao Xiang, Devis Tuia, Lichao Mou, Gui-Song Xia, Liangpei Zhang, Feng Xu, and Friedrich Fraundorfer. 2017. “Deep Learning in Remote Sensing: A Comprehensive Review and List of Resources.” *IEEE Geoscience and Remote Sensing Magazine* 5 (4): 8–36.
- Zhu, Zhe, Yuyu Zhou, Karen C. Seto, Eleanor C. Stokes, Chengbin Deng, Steward T.A. Pickett, and Hannes Taubenböck. 2019. “Understanding an urbanizing planet: Strategic directions for remote sensing.” *Remote Sensing of Environment* 228: 164–182. <https://www.sciencedirect.com/science/article/pii/S0034425719301658>.