CAN SPOT-6/7 CNN SEMANTIC SEGMENTATION IMPROVE SENTINEL-2 BASED LAND COVER PRODUCTS? SENSOR ASSESSMENT AND FUSION

Olivier Stocker, Arnaud Le Bris*

LASTIG, Univ. Gustave Eiffel, ENSG, IGN, F-94160 Saint-Mande, France
stocker.olivier@gmail.com,arnaud.le-bris@ign.fr

KEY WORDS: Land cover, semantic segmentation, satellite, deep learning, CNN, SPOT-6/7, Sentinel-2

ABSTRACT:

Needs for fine-grained, accurate and up-to-date land cover (LC) data are important to answer both societal and scientific purposes. Several automatic products have already been proposed, but are mostly generated out of satellite sensors like Sentinel-2 (S2) or Landsat. Metric sensors, e.g. SPOT-6/7, have been less considered, while they enable (at least annual) acquisitions at country scale and can now be efficiently processed thanks to deep learning (DL) approaches. This study thus aimed at assessing whether such sensor can improve such land cover products. A custom simple yet effective U-net - Deconv-Net inspired DL architecture is developed and applied to SPOT-6/7 and S2 for different LC nomenclatures, aiming at comparing the relevance of their spatial/spectral configurations and investigating their complementarity. The proposed DL architecture is then extended to data fusion and applied to previous sensors. At the end, the proposed fusion framework is used to enrich an existing S2 based LC product, as it is generic enough to cope with fusion at distinct levels.

1. INTRODUCTION

Needs for fine-grained, accurate and up-to-date land cover (LC) data are important to answer both public policies monitoring issues and scientific purposes. Indeed LC is a mandatory knowledge for various uses, such as monitoring urban or forest sprawl or providing inputs to (e.g. climate) simulations. LC products exist at various scales (ranging from global to local level). Some of them are generated automatically (Inglada et al., 2017a, Pesaresi et al., 2016), but often out of Landsat 8 or Sentinel 2 (S2) satellite data. Others are manually plotted to ensure quality and exhibit a higher level of details (when created out of Very High spatial Resolution (VHR) imagery), but generating and updating such products is tedious, long and expensive. Thus using VHR data for LC semantic segmentation would be a way to at least provide change detection inputs to the updating process of a LC map (Matikainen et al., 2019) or to enrich (improve delineation, add smaller objects but also better discriminate textured classes) the content of a product existing at a lower resolution.

A strong trend in remote sensing during the last decade is the outstanding availability of an unprecedented amount of satellite sensors exhibiting multi-modal and complementary characteristics. VHR sensors enable the delineation of small features and the use of texture information, but are generally limited to 4 spectral bands (red - green - blue - near infrared) and rare acquisitions, which reduces their ability to distinguish fine LC types. On the other hand, sensors such as S2 have more bands and an important revisit frequency (time series) but a less geometric resolution. Due to their availability at large scale, S2 or Landsat8 time series have often been considered for automatic land cover classification (Inglada et al., 2017a, Lefebvre et al., 2016, Pelletier et al., 2019, Pesaresi et al., 2016).

In the mean time, the advent of deep learning (DL) for semantic segmentation has enabled to efficiently process VHR data, exploiting texture and context information in a better way than with classic machine learning methods as long as a sufficient training data sets are available. Thanks to the availability of several benchmarks (such as ISPRS Vaihingen and Postdam ones), many DL studies have been devoted to submetric VHR data (Volpi, Tuia, 2017, Marmanis et al., 2016), in particular when 3D information is also available (Paisitkriangkrai et al., 2016, Audebert et al., 2018). However, metric VHR sensors, e.g. SPOT-6/7, have been less considered, while they enable (at least annual) acquisitions at country scale (Postadjian et al., 2017, Gaetano et al., 2018).

Thus, this study is mainly dedicated to this kind of metric remote sensing data available regularly at country scale, aiming at identifying to which extent an enhanced spatial resolution can improve LC maps. Several issues are considered. First, a DL architecture is developed to process them, and assessed for different LC nomenclatures on SPOT-6/7 (SPOT-6/7) data. A comparison between SPOT-6/7 and Sentinel 2 (S2) spatial/spectral configurations is then carried out: both images are processed by the proposed DL architecture to assess their relevance to discover different classes and investigate their complementarity. Second, the previous DL network is modified to perform data fusion and applied to previous sensors. At the end, the proposed fusion framework is used to enrich an existing S2 based LC product, as it is generic enough to cope with fusion at distinct levels, as long as training data are available.

2. SEMANTIC SEGMENTATION METHOD

2.1 Short overview of existing methods

Deep learning approaches and especially Convolutional Neural Networks (CNN) are at present the most efficient semantic segmentation methods as long as a sufficient training dataset is available (LeCun et al., 2015). Indeed, compared to more traditional machine learning methods involving handcrafted features, such methods better cope with texture and context information and show increased generalisation capacities. The most efficient CNN semantic segmentation architectures are fully

* Corresponding author

convolutional encoder-decoder ones (Long et al., 2015). U-net (Ronneberger et al., 2015) is one such architecture. Compared to basic encoder-decoder, it involves skip connections between corresponding encoder and decoder layers to progressively re-introduce high frequency information in the decoder to better delineate objects. (Zhou et al., 2018) proposed a variant named U-net+ involving 1d convolution in the skip connections to keep only the most relevant high frequency information from the encoder layers. More complex approaches as (Chen et al., 2017) exist but involve heavier architectures.

DL approaches have already been applied to mono-date SPOT-6/7 like imagery, from earlier experiments with a patch-based method (Postadjian et al., 2017) to a two-entries network using separately panchromatic and multispectral bands (Gaetano et al., 2018). (Maggiori et al., 2017) proposed a multi-resolution approach to better cope with context information, that can also be tackled using atrous convolution (Chen et al., 2017).

DL architectures were also proposed for remote sensing data fusion. They generally consist in as many encoders as data source. Most of them were dedicated to image-DSM fusion, e.g. (Audebert et al., 2018, Koppanyi et al., 2019), but S2 time series and VHR image fusion has also been tackled (Benedetti et al., 2018).

2.2 One sensor CNN architecture

For this study, a light yet efficient network is required. Indeed, a light architecture reduces the number of parameters to optimize, and is less training set greedy. To lower computing times, a fully convolutional network was preferred to the ones involving fully connected layers as (Gaetano et al., 2018).
At the end, the U-net like CNN presented in figure 1 was adopted. It is slightly different from the original U-net (Ronneberger et al., 2015). First, its depth is reduced to take into account the resolution of the images in respect to the small size of targeted objects. Second, to improve spatial information, as in (Noh et al., 2015), transposed convolution is used in the decoder and Max pooling from the encoder transfer their max-indices to homologous Unpooling layers in the decoder.
Convolutions are done with mirror padding and each floor is composed of two block containing a 2D convolution, a Batch Normalization (Ioffe, Szegedy, 2015) and a ReLU activation. Both Max pooling and unpooling operations have a ratio of 2.

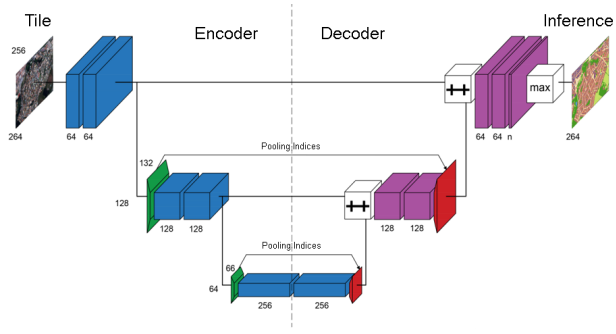

Figure 1. Proposed architecture based on Deconv-Net and U-Net. Blue : 3×3 convolution, green : 2×2 max pooling, red : 2×2 unpooling, Violet : 3×3 deconvolution, ++ : concatenation, max : softmax.

2.3 CNN architecture extension towards fusion

The previous architecture is extended to sensor fusion, especially to merge SPOT-6/7 and S2. It is simply turned into a double entry CNN. The previous encoder branch was simply duplicated. The merging point of these two encoding branches is set before the deepest layer. Skip connections and max pooling indices transfer, supposed to enhance spatial information flow are only kept for SPOT entry, as it's the main provider of spatial information. Both sensors are expected to have been resampled at a same GSD for reasons given in 3.2. Keeping the same encoding and decoding operations as in the previous mono sensor CNN also enables to fairly compare them.
It can here be noted that this architecture can cope with fusion at different levels (early or late) as inputs can be raw images, classification probabilities or label maps.

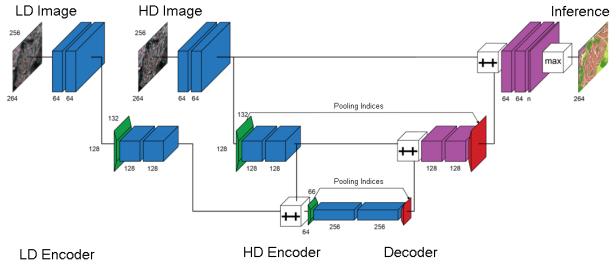

Figure 2. Proposed architecture for fusion. Same color legend as figure 1

2.4 Implementation details

The forward operation is fed with image tiles of 264 × 256 pixels, which are normalized given the 2% and 98% quantiles of the radiometric distribution of the whole image. Training and evaluation is done out of a set of couples of images and ground truth tiles at the same size. Training, validation and test sets represent respectively 60%, 20% and 20% of the whole dataset.

"Xavier" initialization (Glorot, Bengio, 2010) and "Adam" optimizer (Kingma, Ba, 2014) are used for the learning step. The latter follows a step decreasing learning rate from 10^{-2} with $\gamma = 0.7$ each 50 epochs. The loss function is a weighted cross entropy. It aims at taking into account the strongly heterogeneous class representation in the dataset. Each weight $w[class]$ is calculated out of its class proportion $p[class]$ as $w[class] = 0$ when $p[class] = 0$ or $w[class] = \frac{1}{p[class] \sum_{i=1}^{n} p[i]^2}$ (with n the number of classes) else.

Both networks are implemented in https://pytorch.org framework.

3. ASSESSMENT DATASET

As mentioned earlier, this study is mainly dedicated to metric VHR multispectral SPOT-6/7 imagery available regularly at country scale and aims at assessing to which extent an enhanced spatial resolution is relevant for LC classification. In particular, its complementarity with commonly used S2 sensor is studied for different LC classification problems (see 3.4). Having a fair comparison between both sensors also implies defining some requirements concerning images are processed (see 3.2).

3.1 Study area

Experiments are performed over a study area of 576 km^2. It covers the dense urban area of Toulouse city (south-western France) and its rural surroundings. Thus, it contains various land cover classes.

3.2 Image data : sensors, requirements and restrictions

SPOT-6/7 and S2 sensors are considered:

- SPOT-6/7 image has a final resolution of 1.5 m. It results from pansharpening out of native panchromatic and multispectral (red-green-blue-near infrared) data.
- S2 image is a cloud-free gap filled synthesis generated out of a month long time series. All bands are considered except B1, B9 and B10.

However, specific precautions have to be taken in order to guarantee an equivalent processing and a fair comparison of both sensor:

- Only one monodate image per sensor is considered. Both were captured at the same period, respectively end of June 2018 and July 2018 for SPOT-6/7 and S2, so that different results can be related only to their spatial or spectral configurations. It must here be reminded that the present study is for the moment only dedicated to the assessment of the spectral/spatial configurations of both sensors. Thus, though they are already known to enhance crops and vegetation discrimination, the multi-temporal characteristics of S2 and its ability to deliver time series is not investigated further.
- S2 image has been upsampled from its native 10m or 20m GSD to SPOT-6/7 1.5m GSD to cope with the spatial resolution gap between both sensors. Though slightly artificial, this resampling step ensures the comparison is fair and the receptive field of the architecture corresponds to the same ground area in both imagery.

3.3 Training and reference LC data

Training and reference LC maps are generated from a selection of layers from existing national national topographic, forest and agriculture vector geodatabases. The topographic and forest database describes buildings, roads, water areas and forests. Its current version was last updated in 2016. The agriculture DB corresponds to the crops in 2018 and is thus completely coherent with images. All DBs were rasterized and tiled to match their corresponding satellite tiles. It must also be kept in mind that these generated LC maps do not partition the whole study areas as some parts are not described in these original DB.

3.4 Land cover nomenclatures

Thanks to its fine spatial resolution, SPOT is expected to improve the detection of classes with thin specific texture features or consisting in small objects (e.g. urban objects). On the other hand the enhanced spectral configuration of S2 should enable a better discrimination for classes with specific optical features (e.g. vegetation species). Thus considered LC classification problems have to involve these different kinds of classes.

Two land cover problems are considered. Their nomenclatures as well as the representation of the different classes over the data set are presented in table 1.
6 class nomenclature: this legend involving 5 basic topographic classes (buildings, roads, high vegetation, low vegetation/crops, water areas) is suitable to almost all landscapes. An "around building buffer" class is added to help to enhance building delineation. Indeed, building surroundings are generally not described in the previous databases. This additional class introduces constraints in such area preventing building overdetection over their direct unlabelled neighbourhood. A 5m radius was used to create this buffer from the buildings.
18 class nomenclature: It is more complex aims at understanding specific class type detection properties and fusion potential. Except for "cemeteries", its items mostly correspond to subclasses of the 6 class legend. For instance, roads is split into a "main roads" (containing highway, 1 and 2 traffic lanes roads, and their respective junctions) and a "path" (all other smaller roads, paths, etc.) classes. As the study focuses on the enrichment possibilities brought by SPOT-6/7 , classes adapted to VHR information were considered (e.g. buildings or road types), while the the crops class was not extended.

6 class		18 class	
Buffer (Buf)	11.1%	Buffer (Buf)	11.1%
Water (Wat)	2.9%	Water (Wat)	2.9%
Crops (Cro)	47.8%	Crops (Cro)	47.8%
Roads (Roa)	6.0%	Paths (Pat)	0.3%
		Main roads (MR)	4.3%
		Airstrip (Air)	0.6%
		Parking (Par)	0.9%
/		Cemeteries (Cem)	0.2%
Vegetation (Veg)	23.9%	Conifer (Con)	0.6%
		Deciduous (Dec)	16.3%
		Open Forest (OF)	0.5%
		Hedges (Hed)	4.6%
		Heaths (Hea)	1.5%
		Poplars (Pop)	0.2%
		Orchards (Orc)	0.1%
		Vines (Vin)	0.1%
Buildings (Bui)	8.1%	Residential (Res)	6.2%
		Industrial (Ind)	1.9%

Table 1. 6 class and 18 class data distribution, acronyms and color legend

4. SENSOR COMPARISON: RESULTS AND DISCUSSION

This first set of experiments aims at assessing the relevance and complementarity of SPOT-6/7 and S2 sensors for land cover classification problems. Inference maps are displayed on polarsensing.com/isprs2020.

4.1 Quality assessment

Results are compared to ground truth maps at pixel level. Accuracy scores are derived out of confusion matrices: per class Intersection over Union (IoU), their mean Intersection over Union (mIoU) and the Overall Accuracy (OA). However, these figures must sometimes be handle with care. Indeed, reference DB are not completely up-to-date (see section 3.3) and are not a full partition of the image. Besides, some classes are also land use ones and can contain different land covers, as the cemetery example of figure 4.1 that also contains vegetation parts. Thus, visual assessment is also considered to get a better comprehension of the results.

4.2 6 class comparison

The results for the 6 class nomenclature are shown in table 2. Overall metrics already show better discrimination capabilities of SPOT-6/7 sensors. The score gap with S2 is mainly due to the better classification of small topographic elements like roads, buildings and buffers. However, the delineation of such small objects out of S2 is better than expected (fig 3.3). High

vegetation and crops usually relying more on spectral information show less SPOT-6/7 / S2 differences. Very High spatial Resolution (VHR) SPOT-6/7 sensor generally improves object delineation.

6 class				18 class			
	SPOT	**S2**	**Fu.**		**SPOT**	**S2**	**Fu.**
mIoU	78.4	74.5	**79.7**	**mIoU**	48.9	42.2	**52.1**
OA	91.6	89.5	**92.2**	**OA**	**86.3**	81.5	**86.5**
	IoU				**IoU**		
Buf	64.2	58.0	**66.7**	**Buf**	**64.2**	55.9	**64.2**
Wat	87.3	**87.1**	**88.3**	**Wat**	**86.7**	**86.5**	**86.2**
Cro	**94.6**	93.4	**94.9**	**Cro**	**93.7**	88.2	**93.5**
				Pat	**12.7**	5.3	11.0
Roa	72.1	67.7	**74.6**	**MR**	**65.0**	59.9	**65.4**
				Par	42.1	36.5	**50.1**
				Air	69.9	64.9	**71.8**
	/	/	/	**Cem**	30.2	16.9	**37.1**
				Con	37.0	34.7	**46.6**
				Dec	**80.6**	76.8	**80.9**
				OF	**9.0**	4.7	8.3
Veg	**85.6**	84.0	**86.2**	**Hed**	**44.3**	42.8	**44.9**
				Hea	26.1	21.1	**27.1**
				Pop	27.9	43.2	**45.2**
				Orc	2.3	1.0	**6.7**
				Vin	85.0	25.6	**87.1**
Bui	66.3	57.0	**67.6**	**Res**	56.9	49.2	**59.1**
				Ind	46.0	46.1	**52.5**

Table 2. Classification results (in %) for simple and complex legends from both sensors and for the proposed fusion CNN. "Fu." stands for "Fusion".

4.3 18 class comparison

The first obvious results of nomenclature enrichment, displayed in table 2, is the global metric drop. Both SPOT-6/7 and S2 inferences suffer a 30% loss of mIoU score. Per class IoU reveal that it is mainly caused by vegetation class dilatation.
Several situations can be distinguished.

Classes with strong textural aspect but with few examples (vines, cemeteries,...) These classes are by nature more prone to be better detected out of SPOT-6/7 than S2. Vine class is a good example (fig. 3.1). S2 confusion matrix underlines that spectral information is inadequate to separate vines from crops. Inversely SPOT-6/7 VHR is able to use the characteristic texture of vineyards. However, it must be reminded that vine is the less represented class and thus its great detection score could be the consequence of a optimal setup in the dataset.Orchards present the lowest IoU for both sensors. Many orchards are classified as deciduous trees. Orchards is a land use class, and its exploitation layout is the only difference with deciduous trees. However, it is here unsufficient to distinguish it from other forest classes, even at SPOT-6/7 resolution. IoU score of S2 inference confirm that no other discriminating attributes can be derived from spectral information. Thus, sensor fusion will probably not help its detection. Poplar class follows the same exploitation layout statement as orchards, but demonstrates better SPOT results than the latter, meaning more reliable texture features are identified by the CNN. Low metrics are explained by visual assessment: several stands labeled as poplars were cut, resulting in open fields similar to heath.S2 sensors shows better discrimination capabilities, even without textural analysis. As poplar layout frequency is not visible in S2 imagery (6 to 9 m between trees) it reveals useful superspectral features. Hence, sensor fusion will probably improve poplar . Similar good fusion *a priories* can be stated for cemeteries class. S2 show quite good IoU scores even if tombstones texture frequency is less lower than its resolution , and SPOT also provide good tombstone detection (fig. 4.1), even though, cemeteries tend to be over detected in urban areas. It is mostly a visual noise with low effect on metrics as the ground truth is not a full partition. More generally, using cross entropy weights according to class proportion forces the CNN to strongly associate certain texture to low represented classes, that can bring classification noises where closely related textures appear.

Classes with high semantic proximity Industrial buildings class performance is equivalent for both sensors (fig. 3.2). This fact highlights the relevance of superspectral features derived from S2. Indeed, Short-Wave Infrared (SWIR) bands of S2 help to discriminate the hydrocarbon surfaces on industrial building (as it was assessed by an additional experiment involving only the 4 native 10 m GSD multispectral bands of S2). Thus, unified industrial complex is better segmented (fig. 3.2.d). Yet SPOT sensor is able to get cleaner border object delineation (fig. 3.2.c). These two facts underline good fusion *a priories* for this class. Nevertheless such facilities only appear in industrial and commercial areas. Thus this S2 property does not provide any help to discriminate residential from industrial buildings in dense urban contexts, where buildings are generally similar. Confusion matrix analysis validates that the latter point constitute their main confusions source. Roads enrichment suffers mainly from the "path" class detection. Its thin geometry makes it harder to be detected out of S2. Results from both sensors display a non semantic way of bordering paths and mains roads (fig. 3.3). Proposed CNN is probably not deep enough to understand the need of placing roads border on intersection.If SPOT classification metrics on roads enriched classes is higher than S2, the latter still presents close scores. This fact, linked to the latter discussed S2 capabilities of detecting hydrocarbon surfaces, tends towards good fusion performances *a priories*. This exclude paths which do not mainly exhibit such materials.

Heterogeneous compound land cover classes This situation was already mentioned for the very specific cemeteries example of fig. 4.1: this class corresponds to a land use and do not label only tombstones but also vegetated areas, while the CNN has identified a cemeteries specific texture and is not able to associate the latter surroundings to cemeteries classes. Data distribution pressures the learning towards labelling already seen elements to the most present class. Open forest is a mix of deciduous, coniferous and heath classes, but has no specific texture for the CNN to hang on, like tombstones for cemeteries. Two ground truth specification points complicate its classification: (1) it is defined as a large percentage windows of tree coverage by square meters into a certain polygon and (2) there is a huge disparities of how this percentage is balanced into the latter (close forest and a glade, or evenly distributed). This results to a bad overall classification that would probably not be improved by sensor fusion.

Other classes A high misclassification rate is observed for vegetation classes as deciduous trees. As data distribution is mainly in favor of the latter, its IoU does not underline this phenomenon. More generally, it is generally difficult to conclude about the sensor relevance to retrieve forest classes because of reference DB's specification (important mimimum collect unit) and temporal shift that can generate learning and evaluation biases. Results are even difficult to be visually assessed. Conifer are mainly discriminated by a high near infrared absorption. But this spectral feature can disappear, especially for young conifer stands, leading to heaths or deciduous trees misclassi-

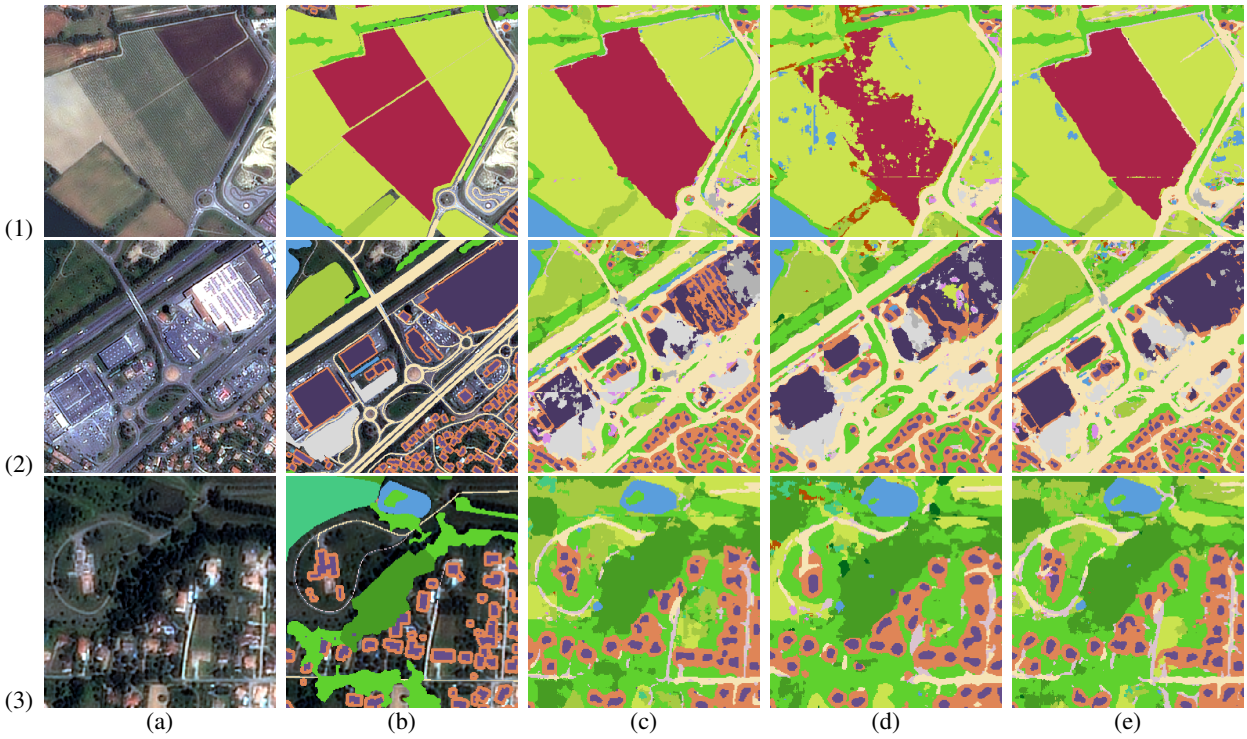

Figure 3. Rows : (1) Vineyards, (2) Industrial and commercial area, (3) Paths in diffuse urban area. Columns : (a) SPOT image, (b) Ground truth, (c) SPOT inference, (d) S2 inference, (e) Fusion inference

fication (fig. 4.2) compared to the database for which such parcel remains a conifer stand even though young conifers can indeed really be lost among these classes. Moreover forest reality does not display straight border like DBs polygons. Therefore, these mismatch impacts accuracy metrics, even when CNN inference corresponds to ground reality. Hedges and heaths suffer from the same labelling problems. In general both sensors achieved same classification performances on forest elements, even though results are different. Visual analysis shows a better element delineation from SPOT.
Last, the three unextended classes (water, crops and building buffer) only undergo small loss. Crops class lost 2 IoU percent with S2 sensors, where confusions between industrial buildings and green-housed crops fields appears.

4.3.1 Stability assessment A strong correlation when aligning the 18 class nomenclature onto the 6 class one. This aggregation was done by assigning each of the 18 class its hierarchically correspondent class from the simple nomenclature (except for cemeteries that were just dismissed and considered as unlabelled pixels). The aggregated classification can then be evaluated according to the simple nomenclature ground truth. The results shown in table 3 point out the disappearance of mIoU and OA drop between complex and simple nomenclature classifications observed in table 2. This indicates that prime misclassification faults occur between related classes. A more detailed nomenclature only contributes to a 1.99% mIoU decline. This reduction is mainly induced by a drop in the roads and buildings classification. Misclassification between industrial buildings and airstrip/parkings can also be noticed from confusion matrices. These errors are inter-classes and still exist after a nomenclature aggregation. At the end, rich nomenclature shows a reliable overall stability compared to simple one, but is still sensitive to spectral or textural similar but non-related classes.

	6 class	**18 class**		**6 class**	**18 class**
mIoU	**78.4**	76.4	OA	**91.6**	90.6
	IoU			IoU	
Buff.	**64.2**	63.0	Roads	**72.1**	68.7
Water	**87.3**	**86.6**	Veget.	**85.6**	83.9
Crops	**94.6**	93.4	Build.	**66.3**	62.7

Table 3. Results (in %) for simple and transposed to 6 class complex nomenclature with SPOT images.

5. SPOT - S2 FUSION: RESULTS AND DISCUSSION

Sensor fusion was performed according to the proposed CNN architecture fed with SPOT-6/7 and S2 images. Fusion classification improved the previous best mIoU (from SPOT) by 3 %. Per class accuracy scores (see table 2) reveal sensor fusion has better discrimination for low represented classes at the cost of more errors on highly represented ones.

5.1 Pros

Fusion mostly profits to classes with specific spectral properties in NIR or SWIR domain. This explains the IoU increase of hydrocarbon surfaces, in particular on parkings, airstrips and industrial buildings. For the latter class, very good building detection is achieved (fig. 3.2). Fusion improves the detection of deciduous, conifer, heath and poplar classes, confirming previous intuitions from sensor assessment. CNN semantic segmentation errors on the latter class (fig. 4.3) helps to understand its discrimination mechanisms. It is based onto textural pattern with a 7 to 10 m frequency. On figure 4.3.e, poplar over-detection is of this latter order of magnitude. This reveals the fusion CNN can extract new discriminant features from both sensors.

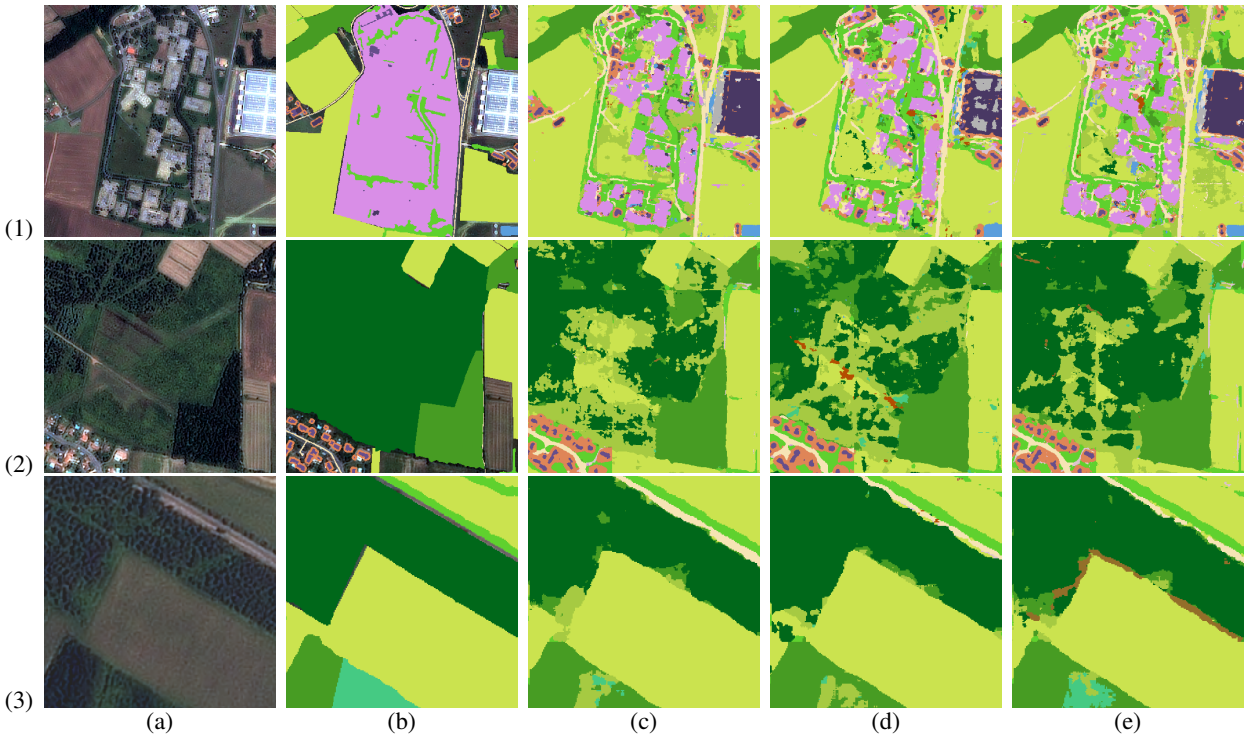

Figure 4. Rows : (1) Cemetery, (2) Forest area, (3) Countryside area. Columns : (a) SPOT image, (b) Ground truth, (c) SPOT inference, (d) S2 inference, (e) Fusion inference

5.2 Cons

As already pointed, the main mIoU improvement comes with several classification faults. The main problem consists in a decrease of spatial information stability, a coarser object delineation. This is easily confirmed by a visual comparison between SPOT and Fusion inferences. This is why roads class did not undergo a similar detection increase as other hydrocarbon surfaces (fig. 3.3). Hedges and mainly paths classes are also affected by this spatial issue. Misclassification of agricultural furrows as paths also increases. Classes that presented bad semantic segmentation score on mono-sensor inferences do not show better results. This confirms intuitions about sensor complementarity from section 4 but indicates that Fusion network's learned feature stay closely related to sensors'performances.

6. TOWARDS ON DEMAND LC MAPS: FUSION TO ENRICH AN EXISTING LC PRODUCT

Previous fusion tests only concern two single-date images, but S2 time series were not evaluated, while they are known to contain discriminant information for instance to classify (crops or forest) vegetation species. This powerful characteristic of S2 is here indirectly taken into account, through the use of an already existing classification results (a label map) inferred from S2 time series. This experiments aims at improving the latter product using VHR SPOT imagery through a reclassification process. Indeed, the proposed fusion CNN framework is generic enough to cope with fusion at distinct levels (early or late) as its inputs can be raw images, classification probabilities or label maps, as long as training data are available. Thus, it is used here to merge a label map and a VHR SPOT-6/7 image.

6.1 Data

6.1.1 Inputs The already existing semantic segmentation is the OSO map (Inglada et al., 2017b). OSO is automatically generated each year over the whole French territory by a Random Forest classification of a year long gap-filled S2 time series. OSO products are freely available at osr-cesbio.ups-tlse.fr/∼oso/ OSO map is a raster product (i.e. a byte single-channel image containing a value corresponding to the attributed label per pixel).

SPOT image is the same than previous tests. Thus, the 2018 version of OSO map is considered to be consistent with it.

6.1.2 Ground truth and nomenclature issues The **ground truth** is generated out of from the same geodatabases presented in 3, which are also approximately the same than are used to train the OSO classifier.

Nomenclature adaptations However, the original OSO nomenclature has to be adapted, as described in 4.

- On one hand, some OSO classes, like glaciers or dunes, are not present in the study area and thus become irrelevant. Hence they are discarded.
- On the other hand, dense, diffuse and industrial & commercial urban area classes display a coarser level of description than SPOT can achieve. Thus they were replaced with new finer classes : residential and industrial buildings. A urban vegetation class was also added. It serves as an equivalent role of the previous building buffer class, namely helping buildings instantiation.

Indeed, a strength of the proposed fusion strategy states in its ability to reclassify the original OSO map for a different nomenclature, taking into account SPOT image information.

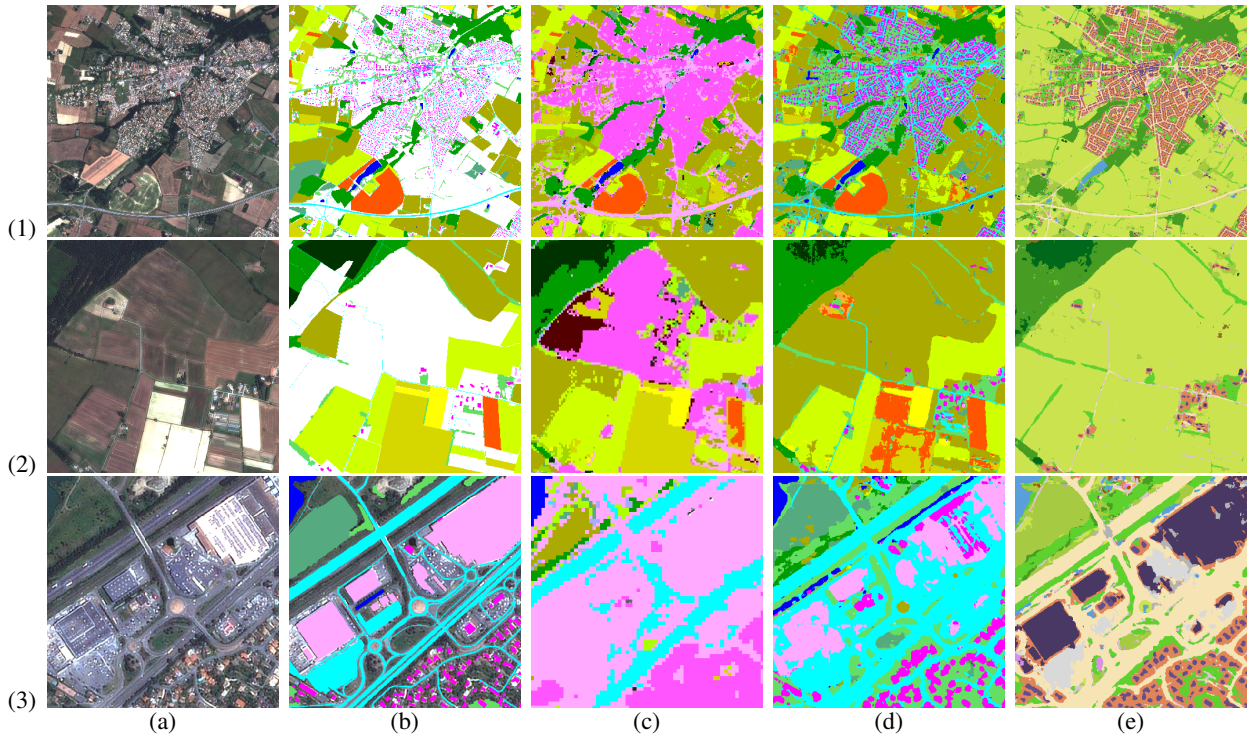

Figure 5. Rows : (1) Road network overview, (2) Countryside area, (3) Industrial and commercial area. Columns : (a) SPOT image, (b) Ground truth, (c) OSO map, (d) Reclassification inference, (e) Fusion inference

Kept		**New**	
Water (Wat)	3.6%	Residential (Res)	7.9%
Rapeseeds (Rap)	6.1%	Industrial (Ind)	2.4%
Cereals (Cer)	18.6%	Urban vegetation (UV)	8.2%
Protein plants (PP)	0.4%		
Soya (Soy)	1.9%	**Discarded**	
Sunflower (Sun)	4.4%	Dense urban areas	
Corn (Cor)	6.7%	Diffuse urban areas	
Grasslands (Gra)	8.3%	Ind. & com. areas	
Orchards (Orc)	0.1%	Rice	
Vines (Vin)	0.1%	Root & tubers	
Deciduous (Dec)	20.4%	Lawns	
Conifer (Con)	0.7%	Mineral surfaces	
Heath (Hea)	1.9%	Beaches & dunes	
Roads (Roa)	7.9%	Snows & glaciers	

Table 4. OSO nomenclature adaptations with discraded, kept and new classes, their data distribution, acronyms and color legend

	OSO	**Fusion**			**Fusion**
Kept mIoU	39.8	**67.6**	**mIoU**		65.6
Kept OA	62.3	**88.2**	**OA**		85.6
IoU	**OSO**	**Fusion**		**OSO**	**Fusion**
Wat	69.5	**87.4**	Con	19.3	**46.0**
Rap	79.4	**85.9**	Hea	4.2	**30.6**
Cer	73.4	**87.3**	*UV*	/	*55.6*
PP	9.1	**42.1**	Roa	8.4	**74.2**
Soy	55.8	**70.1**	Orc	0.7	**6.2**
Sun	8.7	**78.6**	Vin	0.8	**88.5**
Cor	75.0	**77.3**	*Res*	/	*79.4*
Gra	21.6	**70.6**	*Ind*	/	*54.8*
Dec	64.2	**80.6**			

Table 5. Results (in %) for OSO classification and its reclassified version. Kept mIoU and Kept OA omitted incomparable classes in *italic*.

6.2 Results

6.2.1 Quality assessment Original OSO LC map is first evaluated according to the ground truth. Discarded classes were treated as no data. mIoU and OA were calculated only for kept classes. Obviously, no score could be calculated for new classes, and thus they will only be considered for a visual comparison.
The reclassified LC map obtained from the fusion of OSO map and SPOT imagery is evaluated over all classes of the new reclassification nomenclature.
Obtained quality metrics are presented in table 5.

6.2.2 Main improvements First visual assessment confirms a huge detection improvement of urban areas basic elements from OSO map (fig. 5.1). Contrary to original OSO product, road network is no longer limited to highways. Finer road network previously submerged with dense and diffuse urban areas class is now well depicted. This is confirmed by a more than 60% increase in IoU scores.
The second main improvement states in cleaner border delineation between well detected semantic elements. Crops fields, forestry and other objects all benefit from the finer textural analysis brought by the SPOT VHR.
Moreover the latter also helps crops discrimination. Many crops classes show high IoU progress, which can not only be linked to more accurate borders. One such improvement concern the suppression of buildings and crops misclassifications present in OSO map. Indeed, the fusion CNN shows some great capacities of finding the good crop label when OSO maps previously labelled them as buildings (fig 5.2). It seems to have learned the previous algorithms weaknesses, and having linked specific crops texture to RF often misclassified labels.

6.2.3 Mono temporal fusion comparison As the ground truth and methods are similar, the OSO reclassification result

can be compared to the previous fusion one using mono temporal images (cf section 5). One can first notice that both results have the same IoU scores on deciduous trees. This is mainly explained by the fact that deciduous trees are already well detected by SPOT. On the other hand conifer IoU score of previous fusion overtake the current one. As the image fed into the network is only a semantic segmentation map exhibiting a poor conifer detection performance, it has lost all specific spectral information which could improve the result. This is why the reclassification result is closer to SPOT alone classification score on this specific class. The already stated difficulties (see section 4.3) to draw strong conclusions about forest classes discrimination were encountered here. Concerning building detection, accuracy metrics display a great residential and industrial growth. Especially for residential buildings, the absence of the building buffer class suppresses a huge constraint and detected buildings tend to be overdetected over their (unlabeled in the ground truth) surroundings. Thus this gain is only metric-wise and has no visual impact on inference maps. For industrial facilities, the *a priories* brought by industrial & commercial areas class from OSO map help avoiding previous crops misclassification. Nevertheless the absence of further spectral information does not make possible the reconstruction of industrial buildings as previous fusion achieved (fig. 5.3).

6.2.4 Other errors By comparing the 6 class classification result from SPOT to the reclassification one, one can state that the fusion CNN does not exploit all SPOT capabilities in terms of spatial information. Small roads are less detected, and border generally appear more blurry than for the simple nomenclature segmentation. Some reclassification errors can also be added in some crops that were well classified in OSO map. In particular a confusion between corn and sunflowers is observed as in the lower part of fig. 5.2 images.

7. CONCLUSION AND PERSPECTIVES

This study was wondering about the use of metric sensors, as SPOT-6/7, to improve Sentinel-2 based land cover products. A custom simple yet effective U-net - Deconv-Net inspired DL architecture was set to classify such data. SPOT-6/7 and S2 spatial/spectral configurations were evaluated for two different LC nomenclatures, a simple and a fine-grained ones. As expected, SPOT VHR enables to better retrieve small topographic object and textured classes, while S2 additional bands are helpful for some classes. The proposed approach was tested over 4 SPOT images, but the model was trained for each of them (fine-tuning). Thus, it would be relevant to limit its dependence on training data. The proposed DL architecture was then extended to data fusion. It was first applied to previous sensors, leading to globally improved results even though object delineation can be slightly coarser than using only SPOT. It was also used to enrich an existing S2 based LC product (label map), tackling the on-demand land cover issue. Obtained result is globally improved, but some errors also occurred that could be avoided using also S2 imagery. Thus, the CNN should be modified to have a third entry corresponding to one S2 image, or to directly integrate S2 time series. Last, object delineation could also probably be improved considering U-net + skip connections.

ACKNOWLEDGEMENTS

This work was done within the PARCELLE project funded by the French CNES TOSCA program and within the MAESTRIA project supported by the French National Research Agency under the grant ANR-18-CE23-0023.

REFERENCES

Audebert, N., Le Saux, B., Lefèvre, S., 2018. Beyond RGB: Very high resolution urban remote sensing with multimodal deep networks. *ISPRS Journal of Photogrammetry and Remote Sensing*, 140, 20–32.

Benedetti, P., Ienco, D., Gaetano, R., Ose, K., Pensa, R. G., Dupuy, S., 2018. M3Fusion: A Deep Learning Architecture for Multi-{Scale/Modal/Temporal} satellite data fusion. *CoRR*, abs/1803.01945. http://arxiv.org/abs/1803.01945.

Chen, L.-C., Papandreou, G., Schroff, F., Adam, H., 2017. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*.

Gaetano, R., Ienco, D., Ose, K., Cresson, R., 2018. A Two-Branch CNN Architecture for Land Cover Classification of PAN and MS Imagery. *Remote Sensing*, 10(11), 1746. https://hal.archives-ouvertes.fr/hal-01931435.

Glorot, X., Bengio, Y., 2010. Understanding the difficulty of training deep feedforward neural networks. *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, 249–256.

Inglada, J., Vincent, A., Arias, M., Tardy, B., Morin, D., Rodes, I., 2017a. Operational High Resolution Land Cover Map Production at the Country Scale Using Satellite Image Time Series. *Remote Sensing*, 9(1), 95.

Inglada, J., Vincent, A., Arias, M., Tardy, B., Morin, D., Rodes, I., 2017b. 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.

Ioffe, S., Szegedy, C., 2015. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *CoRR*, abs/1502.03167. http://arxiv.org/abs/1502.03167.

Kingma, D. P., Ba, J., 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Koppanyi, Z., Iwaszczuk, D., Zha, B., Saul, C. J., Toth, C. K., Yilmaz, A., 2019. *Multimodal Scene Understanding - Algorithms, Applications and Deep Learning*. Elsevier, chapter Multi-Modal Semantic Segmentation: Fusion of RGB and Depth Data in Convolutional Neural Networks.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. *nature*, 521(7553), 436–444.

Lefebvre, A., Sannier, C., Corpetti, T., 2016. Monitoring Urban Areas with Sentinel-2A Data: Application to the Update of the Copernicus High Resolution Layer Imperviousness Degree. *Remote Sensing*, 8(7).

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. *CVPR*.

Maggiori, E., Tarabalka, Y., Charpiat, G., Alliez, P., 2017. Convolutional neural networks for large-scale remote-sensing image classification. *IEEE TGRS*, 55(2), 645-–657.

Marmanis, D., Wegner, J., Galliani, S., Schindler, K., Datcu, M., Stilla, U., 2016. Semantic segmentation of aerial images with an ensemble of fully convolutional neural networks. *ISPRS An. of Phot., Rem. Sens. and Spat. Inf. Sc.*, III-3.

Matikainen, L., Pandzic, M., Li, F., Karila, K., Hyyppä, J., Litkey, P., Kukko, A., Lehtomäki, M., Karjalainen, M., Puttonen, E., 2019. Toward utilizing multitemporal multispectral airborne laser scanning, Sentinel-2, and mobile laser scanning in map updating. *Journal of Applied Remote Sensing*, 13(4), 1 – 35.

Noh, H., Hong, S., Han, B., 2015. Learning Deconvolution Network for Semantic Segmentation. *CoRR*, abs/1505.04366. http://arxiv.org/abs/1505.04366.

Paisitkriangkrai, S., Sherrah, J., Janney, P., Van Den Hengel, A., 2016. Semantic labeling of aerial and satellite imagery. *IEEE JSTARS*, 9(7).

Pelletier, C., Webb, G. I., Petitjean, F., 2019. Temporal Convolutional Neural Network for the Classification of Satellite Image Time Series. *Remote Sensing*, 11(5).

Pesaresi, M., Corbane, C., Julea, A., Florczyk, A., Syrris, V., Soille, P., 2016. Assessment of the Added-Value of Sentinel-2 for Detecting Built-up Areas. *Remote Sensing*, 8(4).

Postadjian, T., Le Bris, A., Sahbi, H., Mallet, C., 2017. Investigating the Potential of Deep Neural Networks for Large-Scale Classification of Very High Resolution Satellite Images. *ISPRS An. of the Phot., Rem. Sens. and Spat. Inf. Sc.*, IV-1-W1, 183–190.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. *CoRR*, abs/1505.04597. http://arxiv.org/abs/1505.04597.

Volpi, M., Tuia, D., 2017. Dense semantic labeling of subdecimeter resolution images with convolutional neural networks. *IEEE TGRS*, 55(2).

Zhou, Z., Siddiquee, M., Tajbakhsh, N., Liang, J., 2018. Unet++: A nested u-net architecture for medical image segmentation. *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, Springer, 3–11.