Distance transform regression for spatially-aware deep semantic segmentation
Nicolas Audebert, Alexandre Boulch, Bertrand Le Saux, Sébastien Lefèvre

To cite this version:
Nicolas Audebert, Alexandre Boulch, Bertrand Le Saux, Sébastien Lefèvre. Distance transform regression for spatially-aware deep semantic segmentation. Computer Vision and Image Understanding, 2019, 189, pp.102809. 10.1016/j.cviu.2019.102809. hal-02277621

HAL Id: hal-02277621
https://hal.science/hal-02277621
Submitted on 3 Sep 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Distance transform regression for spatially-aware deep semantic segmentation

Nicolas Audebert\textsuperscript{a,b,\textdagger}, Alexandre Boulch\textsuperscript{a}, Bertrand Le Saux\textsuperscript{a}, Sébastien Lefèvre\textsuperscript{b}

\textsuperscript{a}DTIS, ONERA, Université Paris Saclay, F-91123 Palaiseau - France
\textsuperscript{b}Univ. Bretagne-Sud, UMR 6074, IRISA, F-56000 Vannes, France

\textbf{ABSTRACT}

Understanding visual scenes relies more and more on dense pixel-wise classification obtained via deep fully convolutional neural networks. However, due to the nature of the networks, predictions often suffer from blurry boundaries and ill-segmented shapes, fueling the need for post-processing. This work introduces a new semantic segmentation regularization based on the regression of a distance transform. After computing the distance transform on the label masks, we train a FCN in a multi-task setting in both discrete and continuous spaces by learning jointly classification and distance regression. This requires almost no modification of the network structure and adds a very low overhead to the training process. Learning to approximate the distance transform back-propagates spatial cues that implicitly regularizes the segmentation. We validate this technique with several architectures on various datasets, and we show significant improvements compared to competitive baselines.

\textcopyright 2019 Elsevier Ltd. All rights reserved.

1. Introduction

Semantic segmentation is a task that is of paramount importance for visual scene understanding. It is often used as the first layer to obtain representations of a scene with a high level of abstraction, such as listing objects and their shapes. Fully Convolutional Networks (FCNs) have proved themselves to be very effective for semantic segmentation of all kinds of images, from multimedia images Everingham et al. (2014) to remote sensing data Rottensteiner et al. (2012), medical imaging Ulman et al. (2017) and autonomous driving Cordts et al. (2016). However, a recurrent issue often raised by the practitioners is the fact that FCN tend to produce blurry or noisy segmentations, in which spatial transitions between classes are not as sharp as expected and objects sometimes lack connectivity or convexity, and therefore the results need to be regularized using some post-processing Zheng et al. (2015); Chen et al. (2018). This has led the computer vision community to investigate many post-processing and regularization techniques to sharpen the visual boundaries and enforce spatial smoothness in semantic maps inferred by FCN. Yet these methods are often either graphical models

\textdagger Corresponding author: e-mail: nicolas.audebert@onera.fr (Nicolas Audebert)
added on top of deep neural networks Liu et al. (2018); Zheng et al. (2015) or based on sophisticated prior knowledge Le et al. (2018); Bertasius et al. (2016). In this work, we propose a much straightforward approach by introducing a simple implicit regularization embedded in the network loss function. We consider the distance transform of the segmentation masks in a regression problem as a proxy for the semantic segmentation task. The distance transform is a continuous representation of the label masks, where one pixel becomes represented not only by its belonging to a class, but by its spatial proximity to all classes. This means that the gradient back-propagated contains more information about the underlying spatial structure of the data compared to traditional classification. As such, the network learns a smoother segmentation with a very low complexity overhead. Moreover, this is straightforward to implement and does not rely on any additional priors, but only on an alternative representation of the ground truth. Therefore any deep segmentation architecture can be adapted in this fashion without any structural alteration. We validate our method with several architectures on diverse application domains on which we obtain significant improvements w.r.t strong baselines: urban scene understanding, RGB-D images and Earth Observation.

2. Related work

Semantic segmentation is a longstanding task in the computer vision community. Several benchmarks have been introduced on many application domains such as COCO Lin et al. (2014) and Pascal VOC Everingham et al. (2014) for multimedia images, CamVid Brostow et al. (2009) and Cityscapes Cordts et al. (2016) for autonomous driving, the ISPRS Semantic Labeling Rottensteiner et al. (2012) and INRIA Aerial Image Labeling Maggiori et al. (2017) datasets for aerial image, and medical datasets Ulman et al. (2017), which are now dominated by the deep fully convolutional networks. Many applications rely on a pixel-wise semantic labeling to perform scene understanding, such as object-instance detection and segmentation He et al. (2017); Arnab and Torr (2017) in multimedia images, segment-before-detect pipelines for remote sensing data processing Audebert et al. (2017); Sommer et al. (2017) and segmentation of medical images for neural structure detection and gland segmentation Ronneberger et al. (2015); Chen et al. (2016a).

State-of-the-art architectures are all derived from the Fully Convolutional Network paradigm Long et al. (2015), which introduced the possibility to perform pixel-wise classification using convolutional networks that were previously restricted to image-wide classification. Many models building upon this idea were then proposed, e.g. DeepLab Chen et al. (2018), dilated convolutional networks Yu and Koltun (2015) or auto-encoder inspired architectures such as Seg-Net Badrinarayanan et al. (2017) and U-Net Ronneberger et al. (2015). The introduction of the residual learning framework He et al. (2016) also introduced many new models for semantic segmentation, most notably the PSPNet Zhao et al. (2017) that incorporates multi-scale context in the final classification using a pyramidal module.
However, one common deficiency of the FCNs is the lack of spatial-awareness that adversely affects the classification maps and makes spatial regularization a still active field of research Garcia-Garcia et al. (2017). Indeed, predictions often tend to be blurry along the object edges. As FCN perform pixel-wise classification where all pixels are independently classified, spatial structure is fundamentally implicit and relies only on the use of convolutional filters. Although this has given excellent results on many datasets, this often leads to noisy segmentations, where artifacts might arise in the form of a lack of connectivity of objects and even salt-and-pepper noise in the classifications. Those problems are especially critical in remote sensing applications, in which most objects are fundamentally groups of convex structures and where connectivity and inter-class transitions are a requirement for better mapping.

To address this issue, several approaches for smoothing have been suggested. Graphical models methods, such as dense Conditional Random Fields (CRF), have been used to spatially regularize the segmentation and sharpen the boundaries as a post-processing step Lin et al. (2016). However, this broke the end-to-end learning paradigm, and led to several reformulations in order to couple more tightly the graphical models with deep networks. To this end, Zheng et al. (2015); Liu et al. (2018) rewrote respectively the Conditional Random Field (CRF) and the Markov Random Field (MRF) graphical models as trainable neural networks. In a similar concept, Le et al. (2018) reformulates the Variational Level Set method to solve it using an FCN, while Chen et al. (2016b) uses CNN to perform domain transform filtering. Those methods all are revisiting traditional vision techniques adapted to fit into the deep learning framework. However, they require heavy network modification and are computationally expensive.

A more straightforward strategy consists in performing a data-driven regularization by enforcing new constraints on the model in the form a special loss penalty. Notably, this area has been investigated in the literature for edge detection Yang et al. (2016). For instance, Kokkinos (2015); Bertasius et al. (2016) introduce a carefully crafted loss especially tailored for object boundary detection. CASENet Yu et al. (2017) tries to leverage semantics-related priors into the edge prediction task by learning the classes that are adjacent to each boundary, while the COB strategy Maninis et al. (2018) incorporates geometric cues by predicting oriented boundaries. Multi-scale approaches such as Liu et al. (2017) tune the network architecture to fuse activations from multiple layers and improve edge prediction by mixing low and high-level features. In the case of semantic segmentation, object shapes benefit from a better spatial regularity as most shapes are often clean, closed sets. Therefore, better boundaries often help by closing the contours and removing classification noise. To this end, models such as DeepContours Shen et al. (2015) explicitly learn both the segmentation and the region boundaries using a multi-task hand-crafted loss. A similar approach with an ensemble of models has been suggested in Marmanis et al. (2017), especially tailored for aerial images. Chen et al. (2016a); Cheng et al. (2017)
still use a multi-task loss with explicit edge detection, but also fuse feature maps from several layers for more precise boundaries, with applications in gland segmentation and aerial image labeling, respectively. The SharpMask Pinheiro et al. (2016) approach uses a multi-stage network to successively learn refinements of the segmented shapes.

These methods all try to alleviate the classification noise by incorporating spatial-awareness in the semantic segmentation pipeline. However, they share a common drawback as they introduce an explicit hand-crafted loss term to sharpen boundaries and spatially regularize the segmentation, either in the form of a regularization loss penalty, a heavy network modification or a graphical model post-processing. This stems from the fact that segmentation labels are often an aggregation of binary masks that have a low spatial-expressiveness. In Hayder et al. (2017), a distance transform was introduced to allow an instance segmentation to infer shapes outside the original bounding box of the object. Indeed, the distance transform conveys proximity meaning along the edges and even further. This allows the network to learn more precise information than only “in” or “out” as would do one-hot encoding and therefore feeds cues about the spatial structure to the network.

Inspired by this recent idea, we introduce a distance transform regression loss in a multi-task learning framework, which acts as a natural regularizer for semantic segmentation. This idea was tested independently from us in Bischke et al. (2017), although only for building footprint extraction using a quantized distance transform that was roughly equivalent to standard multi-class classification task. Our method is simpler as it directly works on the distance transform using a true regression. While previous methods brought additional complexity, either in the form of a hand-crafted loss function or an alternative network design, our approach remains straightforward and fully data-driven. It requires nearly almost no network modification as it only adds a regression target, in the form of the distance transformed labels, to the original classification task.
3. Distance transform regression

In this work, we suggest to use the signed distance transform (SDT) to improve the learning process of semantic segmentation models. The SDT transforms binary sparse masks into equivalent continuous representations. We argue that this representation is more informative for training deep networks as one pixel now owns a more precise representation of its spatial proximity with various semantic classes. We show that using a multi-task learning framework, we can train a FCN to perform both semantic segmentation by traditional classification and SDT regression, and this helps the network infer better structured semantic maps.

3.1. Signed-distance transform

We use the signed distance transform (SDT) Ye (1988), which assigns to each pixel of the foreground its distance to the closest background point, and to each pixel of the background the opposite of its distance to the closest foreground point. If $x_{i,j}$ are the input image pixel values and $M$ the foreground mask, then the pixels $d_{i,j}$ of the distance map are obtained with the following equation:

$$
\forall i,j, \quad d_{i,j} = \begin{cases}
+\min_{z \in M}(\|x_{i,j} - z\|), & \text{if } x_{i,j} \in M, \\
-\min_{z \in M}(\|x_{i,j} - z\|), & \text{if } x_{i,j} \notin M.
\end{cases}
$$ (1)

Considering that semantic segmentation annotations can be interpreted as binary masks, with one mask per class, it is possible to convert the labels into their signed-distance transform counterparts. In this work, we apply class-wise the signed Euclidean distance transform to the labels using a linear time exact algorithm Maurer et al. (2003).

In order to avoid issues where the nearest point is outside the receptive field of the network, we clip the distance to avoid long-range spatial dependencies that would go out of the network field-of-view. The clipping value is set globally for all classes. We then normalize the SDTs of each class to constrain them in the $[-1;1]$ range. This can be seen as feeding the SDT into a non-linear saturating activation function hardtanh. The visual representations are illustrated in Fig. 1. The same processing is applied to the distances estimated by the network.

3.2. Multi-task learning

Signed distance transform maps are continuous representations of the labels (classes). We can train a deep network to approximate these maps using a regression loss.

However, preliminary experiments show that training only for regression does not bring any improvement compared to traditional classification and even degrades the results. Therefore, we suggest to use a multi-task strategy, in which the network learns both the classification on the usual one-hot labels and the regression on all SDTs. More precisely, we alter the network to first predict the SDTs and we then use an additional convolutional layer to fuse the last layer features.
and the inferred SDTs to perform the final classification. In this way, the network is trained in a cascaded multi-task fashion, where the distance transform regression is used as a proxy, i.e. an intermediate task, before classification.

Therefore, the network modification can be summarized as follows. Instead of using the last layer and feeding it into a softmax, we now use the last layer as a distance prediction. As distances are normalized between -1 and 1, these distances pass through a hardtanh non-linearity. Then, we concatenate the previous layer features maps and the distance predictions to feed both into a convolutional and a softmax layer. The complete architecture is illustrated in Fig. 2.

In this work, we keep the traditional cross-entropy loss for classification, in the form of the negative log-likelihood (NLL). As our regression results are constrained in $[-1; 1]$, we use the L1 loss to preserve relative errors.

Assuming that $Z_{seg}, Z_{dist}, Y_{seg}, Y_{dist}$ respectively denote the output of the segmentation softmax, the regressed distance, the ground truth segmentation labels and the ground truth distances, the final loss to be minimized is:

$$L = NLLoss(Z_{seg}, Y_{seg}) + \lambda L1(Z_{dist}, Y_{dist})$$

where $\lambda$ is an hyper-parameter that controls the strength of the regularization.

4. Experiments

4.1. Baselines

We first obtain baseline results on various datasets using SegNet or PSPNet for semantic segmentation, either using the cross entropy for label classification or the L1 loss for distance regression. However, note that our method is not architecture-dependent. It consists in a straightforward modification of the end of the network that would fit any architecture designed for semantic segmentation.

SegNet Badrinarayanan et al. (2017) is a popular architecture for semantic segmentation, originally designed for autonomous driving. It is designed around a symmetrical encoder-decoder architecture based on VGG-16 Simonyan and Zisserman (2015). The encoder results in downsampled feature maps at 1:32 resolution. These maps are then upsampled
and projected in the label space by the decoder using unpooling layer. The unpooling operation replaces the decoded activations into the positions of the local maxima computed in the encoder maxpooling layers.

PSPNet Zhao et al. (2017) is a recent model for semantic segmentation that achieved new state-of-the-art results on several datasets Cordts et al. (2016); Everingham et al. (2014). It is based on the popular ResNet He et al. (2016) model and uses a pyramidal module at the end to incorporate multi-scale contextual cues in the learning process. In our case, we use PSPNet, that encodes the input into feature maps at 1:32 resolution, which are then upsampled using transposed convolutions.

4.2. Datasets

We validate our method on several datasets in order to show its generalization capacity on multi and mono-class segmentation of both ground and aerial images.

**ISPRS 2D Semantic Labeling.** The ISPRS 2D Semantic Labeling Rottensteiner et al. (2012) datasets consist in two sets of aerial images. The Vaihingen scene is comprised of 33 infrared-red-green (IRRG) tiles with a spatial resolution of 9cm/px, with an average size of 2000 × 1500px. Dense annotations are available on 16 tiles for six classes: impervious surfaces, buildings, low vegetation, trees, cars and clutter, although the latter is not included in the evaluation process. The Potsdam scene is comprised of 38 infrared-red-green-blue (IRRGB) tiles with a spatial resolution of 5cm/px and size of 6000 × 6000px. Dense annotations for the same classes are available on 24 tiles. Evaluation is done by splitting the datasets with a 3-fold cross-validation.

**INRIA Aerial Image Labeling Benchmark.** The INRIA Aerial Image Labeling dataset Maggiori et al. (2017) is comprised of 360 RGB tiles of 5000 × 5000px with a spatial resolution of 30cm/px on 10 cities across the globe. Half of the cities are used for training and are associated to a public ground truth of building footprints. The rest of the dataset is used only for evaluation with a hidden ground truth.

**SUN RGB-D.** The SUN RGB-D dataset Song et al. (2015) is comprised of 10,335 RGB-D images of indoor scenes acquired from various sensors, each capturing a color image and a depth map. These images have been annotated for 37 semantic classes such as “chairs”, “floor”, “wall” or “table”, with a few pixels unlabeled.

**Data Fusion Contest 2015.** The Data Fusion Contest 2015 Campos-Taberner et al. (2016) is comprised of 7 aerial RGB images of 10,000 × 10,000px with a spatial resolution of 5cm/px on the city of Zeebruges, Belgium. A dense set of annotations on 8 classes (6 from ISPRS dataset plus “water” and “boat”) is given. Two images are reserved for testing, we use one image for validation and the rest for training.
The CamVid dataset Brostow et al. (2009) is comprised of 701 fully annotated still frames from urban driving videos, with a resolution of 360 \times 480px. We use the same split as in Badrinarayanan et al. (2017), i.e. 367 training images, 101 validation images and 233 test images. The ground truth covers 11 classes relevant to urban scene labeling, such as “building”, “road”, “car”, “pedestrian” and “sidewalk”. A few pixels are assigned to a void class that is not evaluated.

4.3. Experimental setup

We experiment with the SegNet and PSPNet models.

SegNet is trained for 50 epochs with a batch size of 10. Optimization is done using Stochastic Gradient Descent (SGD) with a base learning rate of 0.01, divided by 10 after 25 and 45 epochs, and a weight decay set at 0.0005. Encoder weights are initialized from VGG-16 Simonyan and Zisserman (2015) trained on ImageNet Deng et al. (2009), while decoder weights are randomly initialized using the policy from He et al. (2015). For SUN RGB-D, in order to validate our method in a multi-modal setting, we use the FuseNet Hazirbas et al. (2016) architecture. This model consists in a dual-stream SegNet that learns a joint representation of both the color image and the depth map. We train it using SGD with a learning rate of 0.01 on resized 224 \times 224 images. On aerial datasets, we randomly extract 256 \times 256 crops (384 \times 384 on the INRIA Labeling dataset), augmented with flipping and mirroring. Inference is done using a sliding window of the same shape with a 75% overlap.

We train a PSP-Net on CamVid for 750 epochs using SGD with a learning rate of 0.01, divided by 10 at epoch 500, a batch size of 10 and a weight decay set at 0.0005. We extract random 224 \times 224 crops from the original images and we perform random mirroring to augment the data. We fine-tune on full scale images for 200 epochs, following the practice from Jégou et al. (2017). Our implementation of PSPNet is based on ResNet-50 pre-trained on ImageNet and does not use the auxiliary classification loss for deep supervision Zhao et al. (2017).

Finally, we use median-frequency balancing to alleviate the class unbalance from SUN RGB-D and CamVid.

For a fair comparison, the same additional convolutional layer required by our regression is added to the previous classification baselines, so that both models have the same number of parameters.

All experiments are implemented using the PyTorch library noa (2016). SDT is computed on CPU using the Scipy library Jones et al. (2001) and cached on-memory or on-disk, which slows down training during the first epoch and uses system resources. Online SDT computation using a fast GPU implementation Zampirolli and Filipe (2017) would strongly alleviate those drawbacks.
<table>
<thead>
<tr>
<th>Method</th>
<th>Dataset</th>
<th>OA</th>
<th>Roads</th>
<th>Buildings</th>
<th>Low veg.</th>
<th>Trees</th>
<th>Cars</th>
</tr>
</thead>
<tbody>
<tr>
<td>SegNet (SDT regression)</td>
<td>Vaihingen</td>
<td>89.49</td>
<td>91.03</td>
<td>95.60</td>
<td>81.23</td>
<td>88.31</td>
<td>0.00</td>
</tr>
<tr>
<td>SegNet (classification)</td>
<td>Potsdam</td>
<td>91.85</td>
<td>94.12</td>
<td>96.09</td>
<td>88.48</td>
<td>85.44</td>
<td>96.62</td>
</tr>
<tr>
<td>SegNet (+ SDT)</td>
<td>Potsdam</td>
<td>92.22</td>
<td>94.33</td>
<td>96.52</td>
<td>88.55</td>
<td>86.55</td>
<td>96.79</td>
</tr>
<tr>
<td>SegNet (+ SDT)</td>
<td>Vaihingen</td>
<td>90.31 ± 0.12</td>
<td>91.55 ± 0.24</td>
<td>95.75 ± 0.21</td>
<td>78.80 ± 0.35</td>
<td>90.10 ± 0.12</td>
<td>81.59 ± 0.71</td>
</tr>
</tbody>
</table>

Table 1: Results on the ISPRS datasets. F1 scores per class and overall accuracy (OA) are reported.

4.4. Results

**ISPRS dataset.** The cross-validated results on the ISPRS Vaihingen and Potsdam datasets are reported in Table 1. For Vaihingen dataset, the validation set comprises 4 images out of 16 and 5 images out of 24 for Potsdam. All classes seem to benefit from the distance transform regression. On Potsdam, the class "trees" is significantly improved as the distance transform regression forces the network to better learn its closed shape, despite the absence of leaves that make the underlying ground visible from the air. Two example tiles are shown in Fig. 3 and Fig. 4, where most buildings strongly benefit from the distance transform regression, with smoother shapes and less classification noise. Moreover, we also tested to perform regression only on the Vaihingen dataset, which slightly improved the results on several classes, although it missed all the cars and had a negative impact overall. It is also worth noting that our strategy succeeds while CRF did not improve classification results on this dataset as reported in Marmanis et al. (2017).

**INRIA Aerial Image Labeling Benchmark.** The results on the test set of the INRIA Aerial Image Labeling benchmark are reported in Table 2. Our results are competitive with those from other participants to the contest. Using the distance transform regression improves the intersection over union (IoU) by 0.47 and makes many errors disappear. As shown
Fig. 4: Excerpt of the results on the ISPRS Potsdam dataset. Legend: white: impervious surfaces, blue: buildings, cyan: low vegetation, green: trees, yellow: vehicles, red: clutter, black: undefined.

<table>
<thead>
<tr>
<th>Method</th>
<th>Bellingham</th>
<th>Bloomington</th>
<th>Innsbruck</th>
<th>San Francisco</th>
<th>East Tyrol</th>
<th>IoU</th>
<th>OA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inria1</td>
<td>52.91</td>
<td>46.08</td>
<td>58.12</td>
<td>57.84</td>
<td>59.03</td>
<td>55.82</td>
<td>93.54</td>
</tr>
<tr>
<td>Inria2</td>
<td>56.11</td>
<td>50.40</td>
<td>61.03</td>
<td>61.38</td>
<td>62.51</td>
<td>59.31</td>
<td>93.93</td>
</tr>
<tr>
<td>TeraDeep</td>
<td>58.08</td>
<td>53.38</td>
<td>59.47</td>
<td>64.34</td>
<td>62.00</td>
<td>60.95</td>
<td>94.41</td>
</tr>
<tr>
<td>RMIT</td>
<td>57.30</td>
<td>51.78</td>
<td>60.70</td>
<td>66.71</td>
<td>59.73</td>
<td>61.73</td>
<td>94.62</td>
</tr>
<tr>
<td>Raisa Energy</td>
<td>64.46</td>
<td>56.63</td>
<td>66.99</td>
<td>67.74</td>
<td>69.21</td>
<td>65.94</td>
<td>94.36</td>
</tr>
<tr>
<td>DukeAMLL</td>
<td>66.90</td>
<td>58.48</td>
<td>69.92</td>
<td>75.54</td>
<td>72.34</td>
<td>70.91</td>
<td>95.70</td>
</tr>
<tr>
<td>NUS</td>
<td>65.36</td>
<td>58.50</td>
<td>68.45</td>
<td>71.17</td>
<td>71.58</td>
<td>68.36</td>
<td>95.18</td>
</tr>
<tr>
<td>SegNet* (classification)</td>
<td>63.42</td>
<td>62.74</td>
<td>63.77</td>
<td>66.53</td>
<td>65.90</td>
<td>65.04</td>
<td>94.74</td>
</tr>
<tr>
<td>SegNet* (+SDT)</td>
<td>68.92</td>
<td>68.12</td>
<td>71.87</td>
<td>71.17</td>
<td>74.75</td>
<td>71.02</td>
<td>95.63</td>
</tr>
</tbody>
</table>

Table 2: Results on the test set of the INRIA Aerial Image Labeling Benchmark when our results were submitted (11/14/17). The multi-task framework consistently improves the standard SegNet results. We report the overall accuracy (OA) and the intersection over union (IoU) for each city. Best results are in **bold**, second best are in *italics*.

<table>
<thead>
<tr>
<th>Method</th>
<th>IoU (val)</th>
<th>OA (val)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SegNet</td>
<td>72.57</td>
<td>95.66</td>
</tr>
<tr>
<td>SegNet (multi-task)</td>
<td>73.00</td>
<td>95.73</td>
</tr>
<tr>
<td>SegNet* (classification)</td>
<td>73.70</td>
<td>95.91</td>
</tr>
<tr>
<td>SegNet* (+SDT)</td>
<td><strong>74.17</strong></td>
<td><strong>96.03</strong></td>
</tr>
</tbody>
</table>

Table 3: Results on the validation set of the INRIA Aerial Image Labeling Benchmark for comparison to Bischke et al. (2017). We report the overall accuracy (OA) and the intersection over union (IoU).

Fig. 5: Excerpt of the results on the INRIA Aerial Image Labeling dataset. Correctly classified pixels are in green, false positive are in pink and false negative are in blue. The multi-task framework allows the network to better capture the spatial structure of the buildings.
in Fig. 5, multi-task prediction yields more regular building shapes and no mis-classified "holes" within the building inner part. Although no additional buildings are detected, those that were already segmented become cleaner. Note
Table 4: Results on the SUN RGB-D dataset on $224 \times 224$ images. We report the overall accuracy, average intersection over union (AIoU) and average precision (AP). We retrained our own reference FuseNet. Best results are in **bold**, second best are in *italics*.

<table>
<thead>
<tr>
<th>Model</th>
<th>OA</th>
<th>AIoU</th>
<th>AP</th>
</tr>
</thead>
<tbody>
<tr>
<td>DFCN-DCRF Jiang et al. (2017)</td>
<td>76.6</td>
<td>39.3</td>
<td>50.6</td>
</tr>
<tr>
<td>3D Graph CNN Qi et al. (2017)</td>
<td>-</td>
<td>42.0</td>
<td>55.2</td>
</tr>
<tr>
<td>3D Graph CNN Qi et al. (2017) (MS)</td>
<td>-</td>
<td><strong>43.1</strong></td>
<td>55.7</td>
</tr>
<tr>
<td>FuseNet* Hazirbas et al. (2016)</td>
<td>76.8</td>
<td>39.0</td>
<td>55.3</td>
</tr>
<tr>
<td>FuseNet* (+SDT)</td>
<td><strong>77.0</strong></td>
<td>38.9</td>
<td><strong>56.5</strong></td>
</tr>
</tbody>
</table>

Table 5: Results on the Data Fusion Contest 2015 dataset. We report F1 scores per class and the overall accuracy (OA).

<table>
<thead>
<tr>
<th>Method</th>
<th>OA</th>
<th>Roads</th>
<th>Buildings</th>
<th>Low veg.</th>
<th>Trees</th>
<th>Cars</th>
<th>Clutter</th>
<th>Boat</th>
<th>Water</th>
</tr>
</thead>
<tbody>
<tr>
<td>AlexNet</td>
<td>83.32</td>
<td>79.10</td>
<td>75.60</td>
<td>78.00</td>
<td>79.50</td>
<td>50.80</td>
<td>63.40</td>
<td>44.80</td>
<td>98.20</td>
</tr>
<tr>
<td>SegNet (classification)</td>
<td>86.67</td>
<td><strong>84.05</strong></td>
<td><strong>82.21</strong></td>
<td>82.24</td>
<td>69.10</td>
<td>79.27</td>
<td>65.78</td>
<td><strong>56.80</strong></td>
<td>98.93</td>
</tr>
<tr>
<td>SegNet (+SDT)</td>
<td><strong>87.31</strong></td>
<td>84.04</td>
<td>81.71</td>
<td><strong>83.88</strong></td>
<td><strong>80.04</strong></td>
<td><strong>80.27</strong></td>
<td><strong>69.25</strong></td>
<td>50.83</td>
<td><strong>98.94</strong></td>
</tr>
</tbody>
</table>

that several missing buildings are actually false positive in the ground truth. We also present a comparison to another multi-task approach which uses a distance transform Bischke et al. (2017) in table 3, this time on their custom validation set. It shows that regression on SDT is better than SDT discretization followed by classification.

**SUN RGB-D.** We report in Table 4 test results on the SUN RGB-D dataset. Switching to the multi-task setting improves the overall accuracy and the average precision by respectively 0.33 and 1.06 points, while very slightly decreasing the average IoU. This shows that the distance transform regression also generalizes to a multi-modal setting on a dual-stream network. Note that this result is competitive with the state-of-the-art 3D Graph CNN from Qi et al. (2017) that leverages 3D cues.

**Data Fusion Contest 2015.** Table 5 details the results on the Data Fusion Contest 2015 dataset compared to the best result from the original benchmark Campos-Taberner et al. (2016). Most classes benefit from the distance transform regression, with the exception of the “boat” class. The overall accuracy is improved by 0.64% in the multi-task setting. Similarly to the Potsdam dataset, trees and low vegetation strongly benefit from the distance transform regression. Indeed, vegetation is often annotated as closed shapes even if it is possible to see what lies underneath. Therefore, filter responses to the pixel spectrometry can be deceptive. Learning distances forces the classifier to integrate spatial features into the decision process.

**CamVid.** The test results on the CamVid dataset are reported in Table 6 that also includes a comparison with other methods from the state-of-the-art, notably Jégou et al. (2017). We report here the results obtained by training two architectures: a deeper PSPNet Zhao et al. (2017) based on ResNet-101 He et al. (2016) and a fully convolutional DenseNet...
Using skip connections from DenseNet Huang et al. (2017) with a UNet-inspired encoder-decoder structure Ronneberger et al. (2015). We use median frequency balancing for both the classification and the regression losses.

On the DenseUNet architecture, embedding the distance transform regression in the network improves the mean IoU by 2.1% and the overall accuracy by 1.0%, with consistent moderate improvements on all classes and significant improvements on cyclists, pedestrians and traffic signs thanks to the class balancing. On the PSPNet-101, our method improves the IoU by 1.9% and the overall accuracy by 0.7%, with per class metrics consistent with the other models,
i.e. significant improvements on all classes except roads and sky. Some examples are shown in Fig. 7 where the distance transform regression once again produces smoother segmentations.

The PSPNet baseline is competitive with those other methods and its mean IoU is improved by 0.5 by switching to the multi-task setting including the distance transform regression. Most classes benefit from the distance transform regression, with the exception of the “road” and “sky” classes. This is due to the void pixels, that are concentrated on those classes and that result in noisy distance labels.

Overall, our results are competitive with other state-of-the-art methods, with only Jégou et al. (2017) obtaining a better segmentation. However we were not able to reproduce their results and therefore unable to test the effect of the distance regression on their model, although it is expected to behave similarly to our DenseUNet reference.

4.5. Discussion

Hyperparameter tuning. In order to better understand how the weight of each loss impacts the learning process, we train several models on the ISPRS Vaihingen dataset using different values for $\lambda$. This adjusts the relative influence of the distance transform regression compared to the cross-entropy classification loss. As reported in Table 1, we compared the regression + classification framework to both individual regression and classification. It is worth noting that SDT regression alone performs worse than the classification. This justifies the need to concatenate the inferred SDT with the last layers features in order to actually improve the performances. Classification alone can be interpreted as $\lambda = 0$. As can be seen in the results, incorporating the SDT regression by increasing $\lambda$ helps the network significantly. Improvements obtained with several values of $\lambda$ are detailed in Fig. 8. There are two visible peaks: one around 0.5 and one around 2. However, these two are not equivalent. The 0.5 peak is unstable and presents a high standard deviation in overall accuracy, while the $\lambda = 2$ peak is even more robust than the traditional classification. Interestingly, this value is equivalent to rescaling the gradient from the distance regression so that its norm is approximately equal to the gradient.
coming from the classification. Indeed, experiments show that there is a ratio of 2 between both gradients and that they decrease roughly at the same speed during training. Therefore, it seems that better results are obtained when both tasks are given similar weights in the optimization process. Nonetheless, all values of $\lambda$ in the test range improved the accuracy and reduced its standard deviation, making it a fairly easy hyperparameter to choose.

Finally, we also investigate the impact of using the distance transform regression compared to performing the regression on the label binary masks, which can be seen as a clipped-SDT with a threshold of 1. We experimented this on CamVid, as reported in Table 6, on lines PSPNet-50* (classification, +mask, +SDT). Using the L1 regression on the masks does not improve the segmentation and even worsens it on many classes. This is not surprising, as the regularization brought by the SDT regression relies on spatial cues that are absent from the binary masks.

Effect of the multi-task learning. The multi-task learning incorporating the distance transform regression in the semantic segmentation model helps the network to learn spatial structures. More precisely, it constrains the network not only to learn if a pixel is in or out a class mask, but also the Euclidean distance of this pixel w.r.t the mask. This information can be critical when the filter responses are ambiguous. For example, trees from birdview might reveal the ground underneath during the winter, as there are no leaves, although annotations still consider the tree to have a shape similar to a disk. Spatial proximity helps in taking these cases into account and removing some of the salt-and-pepper classification noise that it induces, as shown on the ISPRS Vaihingen and Potsdam and DFC2015 datasets. Moreover, as the network has to assign spatial distances to each pixel w.r.t the different classes, it also learns helpful cues regarding the spatial structure underlying the semantic maps. As illustrated in Figs. 5 and 6, the predictions become more coherent with the original structure, with sharper boundaries and less holes when shapes are supposed to be closed.

It can be noted that multi-task is processed by concatenation of SDTs and feature maps. Indeed, concatenation with convolution generalizes the weighted sum operator. It ensures some balance in the influence of SDT and feature maps. However, other mechanisms could have been considered. For example, multiplication would intricate SDTs and feature maps. This is relevant when SDTs are well-estimated but can degrade dramatically results otherwise. For instance, in Table 1, regression fails to estimate the car SDT (which yields a null F1-score for this class) so multi-task with multiplication would also fail.

5. Conclusion

In this work, we looked into semantic segmentation using Fully Convolutional Networks. Semantic segmentation is the first block of many computer vision pipelines for scene understanding and therefore is a critical vision task. Despite their
excellent results, FCNs often lack spatial-awareness without specific regularization techniques. This can be done using various methods ranging from graphical models to ad hoc loss penalties. We investigated an alternative ground truth representation for semantic segmentation tasks, in the form of the signed distance transform. We show how using both label classification and distance transform regression in a multi-task setting helps state-of-the-art networks to improve the segmentation. Especially, constraining the network to learn distance cues helps the segmentation by including spatial information. This implicit method for segmentation smoothing is fully data-driven and relies on no prior, while adding only a very low overhead to the training process. Using the distance transform regression as a regularizer, we obtained consistent quantitative and qualitative improvements in several applications of semantic segmentation for urban scene understanding. RGB-D semantic segmentation and aerial image labeling. We argue that this method can be used straightforwardly for many use cases and could help practitioners to qualitatively improve segmentation results without using CRF or other ad hoc graphical models.

References


