Distance transform regression for spatially-aware deep semantic segmentation

Nicolas Audebert, Alexandre Boulch, Bertrand Le Saux, Sébastien Lefèvre

Distance transform regression for spatially-aware deep semantic segmentation

Nicolas Audebert[a,b,*], Alexandre Boulch[a], Bertrand Le Saux[a], Sébastien Lefèvre[b]

[a] *DTIS, ONERA, Université Paris Saclay, F-91123 Palaiseau - France*
[b] *Univ. Bretagne-Sud, UMR 6074, IRISA, F-56000 Vannes, France*

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.

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

*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

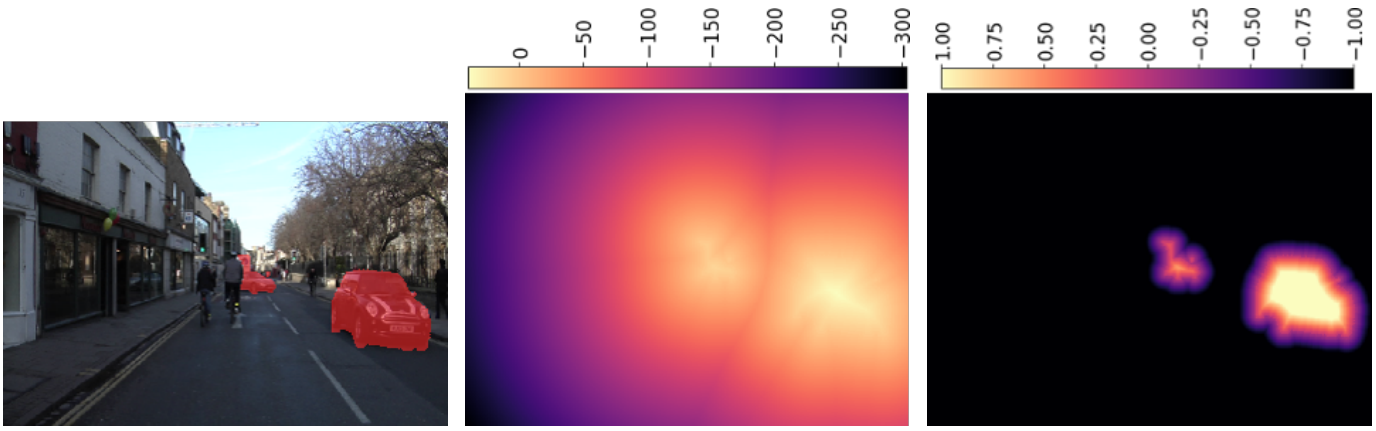
Semantic segmentation is a longstanding task in the computer vision community. Several benchmarks have been introduced on many application domains such as COCO Lin et al. (2014) and Pascal VOC Everingham et al. (2014) for multimedia images, CamVid Brostow et al. (2009) and Cityscapes Cordts et al. (2016) 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 SegNet 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)



(a) Car segmentation binary mask. (b) Signed distance transform (SDT). (c) Truncated and normalized SDT.

Fig. 1: Different representations of a segmentation label.

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:

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

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

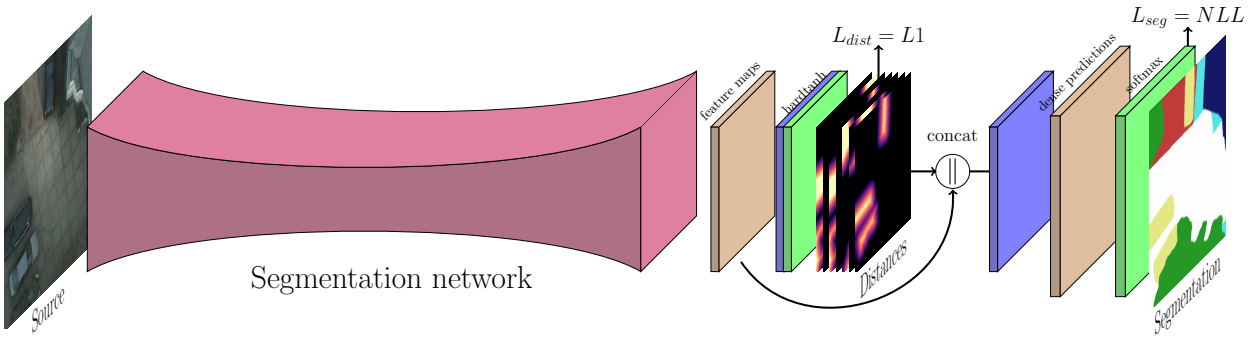


Fig. 2: Multi-task learning framework by performing both distance regression and pixel-wise classification. Convolutional layers are in blue and non-linear activations are in green, while feature maps are in brown.

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:

$$L = NLLLoss(Z_{seg}, Y_{seg}) + \lambda L1(Z_{dist}, Y_{dist}) \tag{2}$$

where λ 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 $10{,}000 \times 10{,}000$px 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.

CamVid. The CamVid dataset Brostow et al. (2009) is comprised of 701 fully annotated still frames from urban driving videos, with a resolution of 360×480px. 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 224×224 images. On aerial datasets, we randomly extract 256×256 crops (384×384 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 224×224 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).

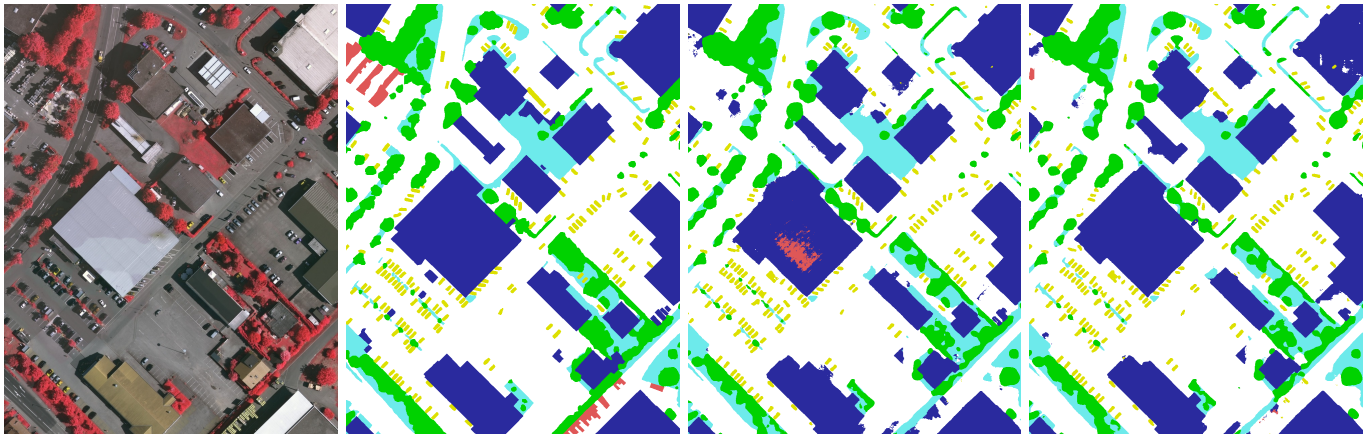
Finally, we use median-frequency balancing to alleviate the class unbalance from SUN RGB-D and CamVid.

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.

Method	Dataset	OA	Roads	Buildings	Low veg.	Trees	Cars
SegNet (SDT regression)	Vaihingen	89.49	91.03	95.60	81.23	88.31	0.00
SegNet (classification)		90.11 ± 0.11	91.31 ± 0.14	95.59 ± 0.14	78.43 ± 0.22	89.99 ± 0.14	**82.37** ± 1.05
SegNet (+ SDT)		**90.31** ± 0.12	**91.55** ± 0.24	**95.75** ± 0.21	**78.80** ± 0.35	**90.10** ± 0.11	81.59 ± 0.71
SegNet (classification)	Potsdam	91.85	94.12	96.09	88.48	85.44	96.62
SegNet (+SDT)		**92.22**	**94.33**	**96.52**	**88.55**	**86.55**	**96.79**

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



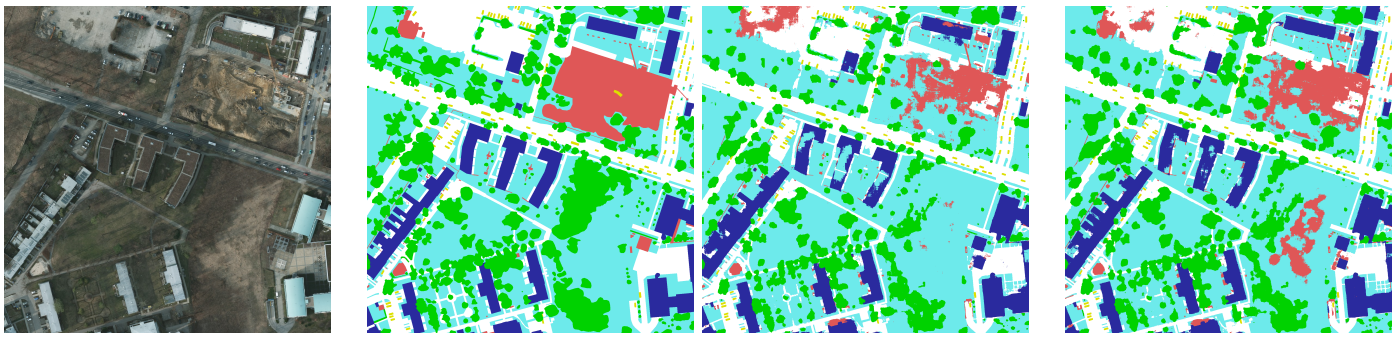
(a) IRRG image (b) Ground truth (c) SegNet (classification) (d) SegNet (multi-task)

Fig. 3: Excerpt of the results on the ISPRS Vaihingen dataset. Legend: white: impervious surfaces, blue: buildings, cyan: low vegetation, green: trees, yellow: vehicles, red: clutter, black: undefined.

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



(a) RGB image (b) Ground truth (c) SegNet (classification) (d) SegNet (multi-task)

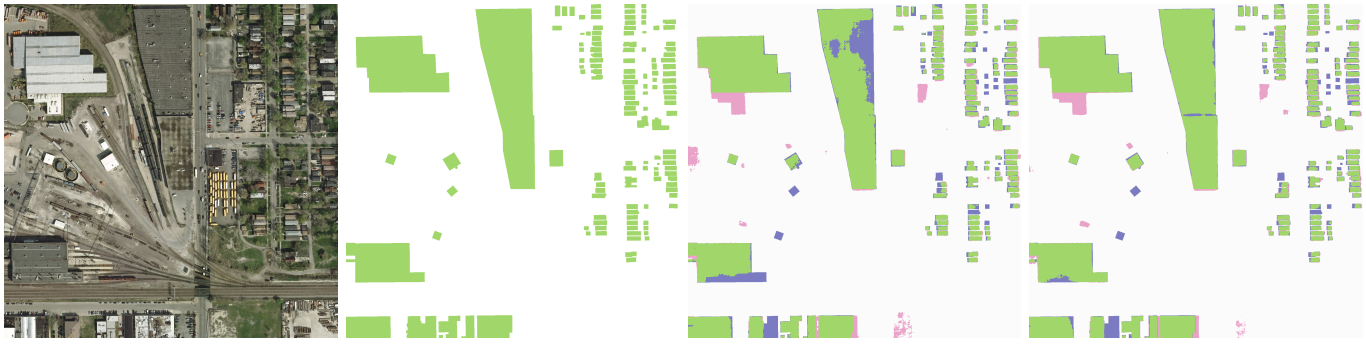
Fig. 4: Excerpt of the results on the ISPRS Potsdam dataset. Legend: white: impervious surfaces, blue: buildings, cyan: low vegetation, green: trees, yellow: vehicles, red: clutter, black: undefined.

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

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*.

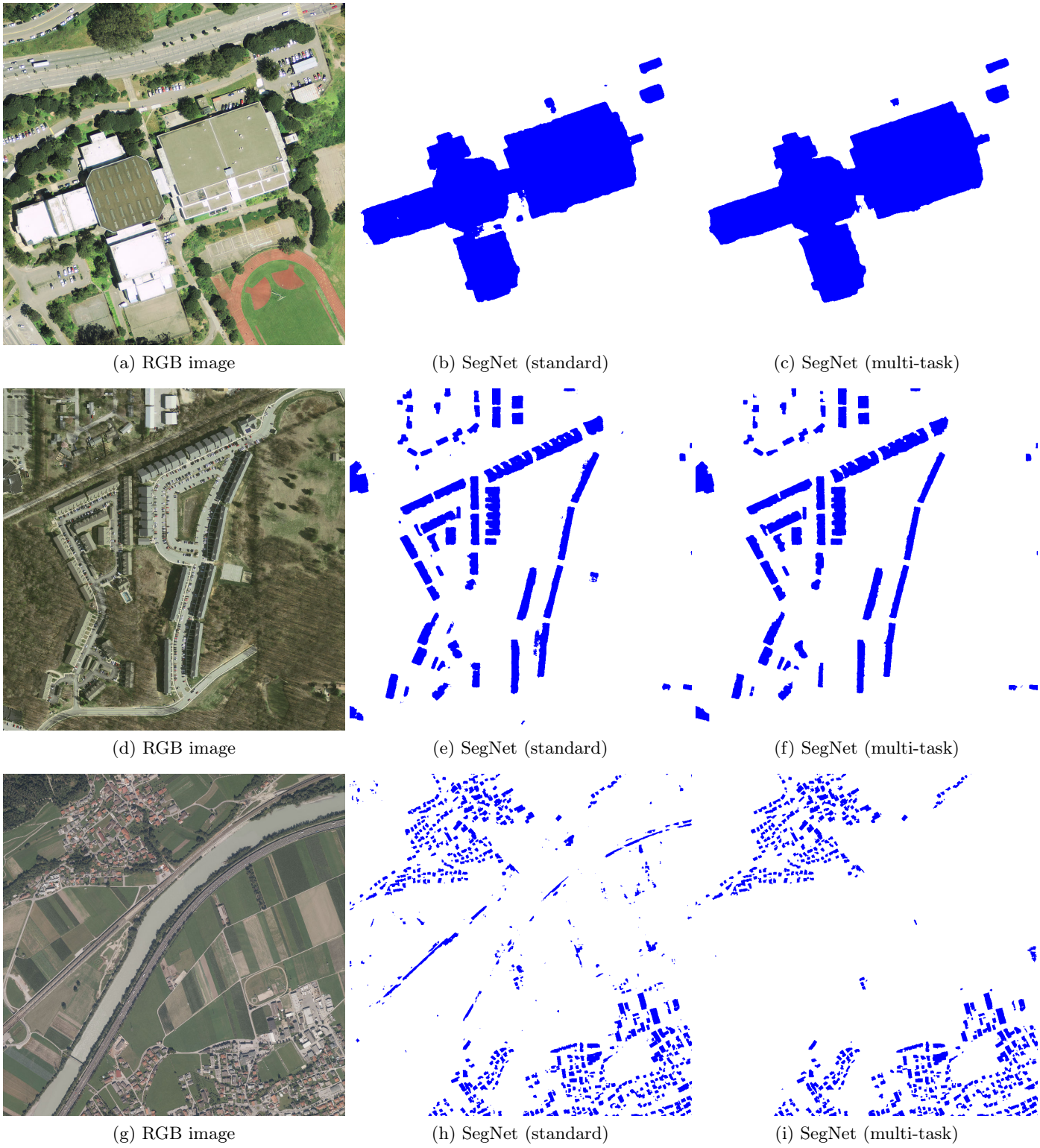
Method	IoU (val)	OA (val)
SegNet Bischke et al. (2017)	72.57	95.66
SegNet (multi-task) Bischke et al. (2017)	73.00	95.73
SegNet* (classification)	73.70	95.91
SegNet* (+SDT)	**74.17**	**96.03**

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).



(a) RGB image (b) Ground truth (c) SegNet (standard) (d) SegNet (multi-task)

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.



(a) RGB image (b) SegNet (standard) (c) SegNet (multi-task)

(d) RGB image (e) SegNet (standard) (f) SegNet (multi-task)

(g) RGB image (h) SegNet (standard) (i) SegNet (multi-task)

Fig. 6: Excerpt of the results on the INRIA Aerial Image Labeling test set. The multi-task framework filters out noisy predictions and cleans the predictions. Its effect is visible at multiple scales, both on a single building (more accurate shape) and on large areas (reduces the number of false positive 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

Model	OA	AIoU	AP
DFCN-DCRF Jiang et al. (2017)	76.6	39.3	50.6
3D Graph CNN Qi et al. (2017)	-	*42.0*	55.2
3D Graph CNN Qi et al. (2017) (MS)	-	**43.1**	*55.7*
FuseNet* Hazirbas et al. (2016)	*76.8*	39.0	55.3
FuseNet* (+SDT)	**77.0**	38.9	**56.5**

Table 4: Results on the SUN RGB-D dataset on 224 × 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*.

Method	OA	Roads	Buildings	Low veg.	Trees	Cars	Clutter	Boat	Water
AlexNet (patch) Campos-Taberner et al. (2016)	83.32	79.10	75.60	78.00	79.50	50.80	63.40	44.80	98.20
SegNet (classification)	86.67	**84.05**	**82.21**	82.24	69.10	79.27	65.78	**56.80**	98.93
SegNet (+SDT)	**87.31**	84.04	81.71	**83.88**	**80.04**	**80.27**	**69.25**	50.83	**98.94**

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

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

Model	mIoU	OA	Building	Tree	Sky	Car	Sign	Road	Pedest.	Fence	Pole	Sidewalk	Cyclist
SegNet Badrinarayanan et al. (2017)	46.4	62.5	68.7	52.0	87.0	58.5	13.4	86.2	25.3	17.9	16.0	60.5	24.8
DeepLab Chen et al. (2018)	61.6	–	*81.5*	*74.6*	89.0	**82.2**	42.3	92.2	*48.4*	27.2	14.3	75.4	50.1
Tiramisu Jégou et al. (2017)	58.9	88.9	77.6	72.0	*92.4*	73.2	31.8	*92.8*	37.9	26.2	*32.6*	*79.9*	31.1
Tiramisu Jégou et al. (2017)	**66.9**	**91.5**	**83.0**	**77.3**	**93.0**	77.3	*43.9*	**94.5**	**59.6**	37.1	**37.8**	**82.2**	**50.5**
PSPNet-50* (classif.)	60.2	89.9	76.3	67.7	89.2	71.0	37.8	91.5	44.0	33.7	26.9	76.6	47.4
PSPNet-50* (+ SDT)	60.7	90.1	76.9	69.7	88.7	72.7	38.1	90.6	44.0	36.6	27.1	75.6	47.7
PSPNet-50* (+ mask)	60.0	89.8	75.6	67.1	89.6	71.4	37.3	92.8	44.4	36.1	27.6	75.7	42.6
PSPNet101* (classif.)	60.3	89.3	74.7	64.1	89.0	71.8	36.6	90.8	44.5	*38.5*	25.4	77.4	*50.3*
PSPNet101* (+ SDT)	*62.2*	90.0	76.2	66.4	88.8	78.0	37.6	90.7	47.2	**40.1**	28.6	78.9	51.2
DenseUNet* (classif.)	59.5	89.6	75.8	68.6	90.9	75.3	37.3	90.0	42.1	26.5	30.1	74.1	43.7
DenseUNet* (+ SDT)	61.6	*90.6*	77.5	69.7	91.1	*78.9*	**44.0**	90.7	46.9	23.7	31.6	77.4	46.2

Table 6: Results on CamVid reporting Intersection over Union (IoU) per class, the mean IoU (mIoU) and the overall accuracy (OA). Models with an "*" are ours. The top part of the table shows several state-of-the-art methods, while the bottom part shows how the distance transform regression consistently improved the metrics on several models. Best results are in **bold**, second best are in *italics*.

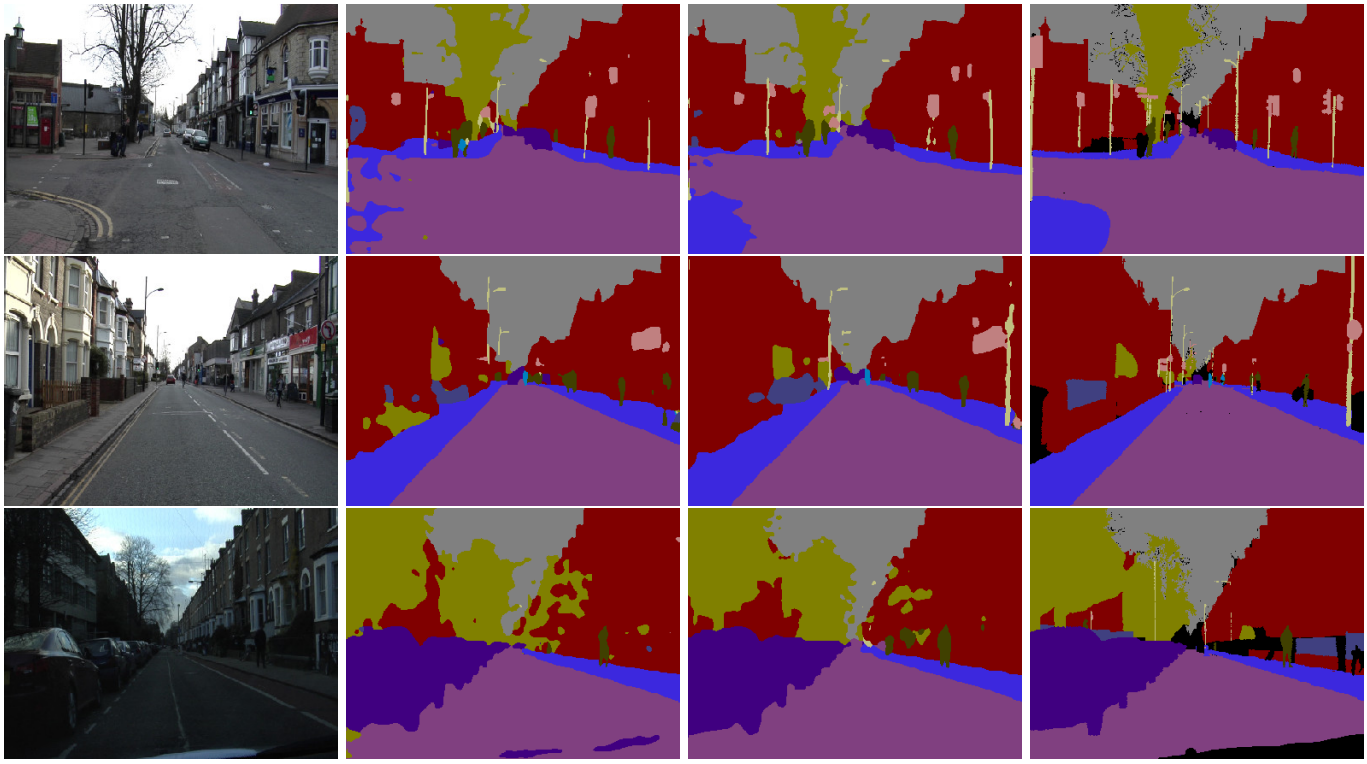


Fig. 7: Example of segmentation results on CamVid using PSPNet. From left to right: RGB image, PSPNet (classification), PSPNet (multi-task), ground truth. The distance transform regression helps improve the consistency and smoothness of the sidewalks, trees, poles and traffic signs.

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,

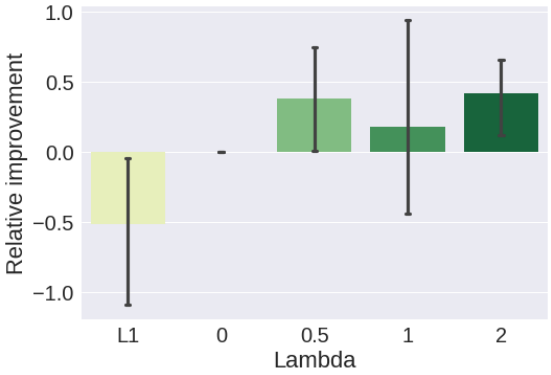


Fig. 8: Exploration of several values for the trade-off factor of the distance transform regularization on the ISPRS Vaihingen dataset and influence on the relative improvement. Results are obtained using a 3-fold cross-validation.

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 oly 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 λ. 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 λ helps the network significantly. Improvements obtained with several values of λ 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 λ 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

PyTorch: Tensors and Dynamic neural networks in Python with strong GPU acceleration, 2016. http://pytorch.org/. 8

Anurag Arnab and Philip H. S. Torr. Pixelwise Instance Segmentation With a Dynamically Instantiated Network. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 441–450, 2017. 2

Nicolas Audebert, Bertrand Le Saux, and Sébastien Lefèvre. Segment-before-Detect: Vehicle Detection and Classification through Semantic Segmentation of Aerial Images. *Remote Sensing*, 9(4):368, April 2017. doi: 10.3390/rs9040368. 2

Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Scene Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(12):2481–2495, December 2017. ISSN 0162-8828. doi: 10.1109/TPAMI.2016.2644615. 2, 6, 8, 13

Gedas Bertasius, Jianbo Shi, and Lorenzo Torresani. Semantic Segmentation With Boundary Neural Fields. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3602–3610, 2016. 2, 3

Benjamin Bischke, Patrick Helber, Joachim Folz, Damian Borth, and Andreas Dengel. Multi-Task Learning for Segmentation of Building Footprints with Deep Neural Networks. *arXiv:1709.05932 [cs]*, September 2017. 4, 10, 12

Gabriel J. Brostow, Julien Fauqueur, and Roberto Cipolla. Semantic object classes in video: A high-definition ground truth database. *Pattern Recognition Letters*, 30(2):88–97, January 2009. ISSN 0167-8655. doi: 10.1016/j.patrec.2008.04.005. 2, 8

Manuel Campos-Taberner, Adriana Romero-Soriano, Carlo Gatta, Gustau Camps-Valls, Adrien Lagrange, Bertrand Le Saux, Anne Beaupère, Alexandre Boulch, Adrien Chan-Hon-Tong, Stéphane Herbin, Hicham Randrianarivo, Marin Ferecatu, Michal Shimoni, Gabriele Moser, and Devis Tuia. Processing of Extremely High-Resolution LiDAR and RGB Data: Outcome of the 2015 IEEE GRSS Data Fusion Contest Part A: 2-D Contest. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 9(12):5547–5559, December 2016. ISSN 1939-1404. doi: 10.1109/JSTARS.2016.2569162. 7, 12

Hao Chen, Xiaojuan Qi, Lequan Yu, and Pheng-Ann Heng. DCAN: Deep Contour-Aware Networks for Accurate Gland Segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2487–2496, 2016a. 2, 3

L. C. Chen, J. T. Barron, G. Papandreou, K. Murphy, and A. L. Yuille. Semantic Image Segmentation with Task-Specific Edge Detection Using CNNs and a Discriminatively Trained Domain Transform. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4545–4554, June 2016b. doi: 10.1109/CVPR.2016.492. 3

Liang-Chieh Chen, Georges Papandreou, Murphy, Kevin, and Yuille, Alan. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):834–848, April 2018. ISSN 0162-8828. doi: 10.1109/TPAMI.2017.2699184. 1, 2, 13

D. Cheng, G. Meng, S. Xiang, and C Pan. FusionNet: Edge Aware Deep Convolutional Networks for Semantic Segmentation of Remote Sensing Harbor Images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 10(12):5769–5783, December 2017. ISSN 1939-1404. doi: 10.1109/JSTARS.2017.2747599. 3

Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The Cityscapes Dataset for Semantic Urban Scene Understanding. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3213–3223, Las Vegas, United States, June 2016. doi: 10.1109/CVPR.2016.350. 1, 2, 7

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, June 2009. doi: 10.1109/CVPR.2009.5206848. 8

Mark Everingham, S. M. Ali Eslami, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The Pascal Visual Object Classes Challenge: A Retrospective. *International Journal of Computer Vision*, 111(1):98–136, June 2014. ISSN 0920-5691, 1573-1405. doi: 10.1007/s11263-014-0733-5. 1, 2, 7

Alberto Garcia-Garcia, Sergio Orts-Escolano, Sergiu Oprea, Victor Villena-Martinez, and Jose Garcia-Rodriguez. A Review on Deep Learning Techniques Applied to Semantic Segmentation. *arXiv:1704.06857 [cs]*, April 2017. 3

Zeeshan Hayder, Xuming He, and Mathieu Salzmann. Boundary-aware Instance Segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 4

Caner Hazirbas, Lingni Ma, Csaba Domokos, and Daniel Cremers. FuseNet: Incorporating Depth into Semantic Segmentation via Fusion-Based CNN Architecture. In *Computer Vision – ACCV 2016*, pages 213–228. Springer, Cham, November 2016. doi: 10.1007/978-3-319-54181-5_14. 8, 12

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1026–1034, December 2015. doi: 10.1109/ICCV.2015.123. 8

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, Las Vegas, United States, June 2016. doi: 10.1109/CVPR.2016.90. 2, 7, 12

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *Proceedings of the International Conference on Computer Vision*, March 2017. 2

Gao Huang, Zhuang Liu, Kilian Q. Weinberger, and Laurens van der Maaten. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, page 3, 2017. 13

Simon Jégou, Michael Drozdzal, David Vazquez, Adriana Romero, and Yoshua Bengio. The One Hundred Layers Tiramisu: Fully Convolutional DenseNets for Semantic Segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1175–1183, Honolulu, United States, July 2017. doi: 10.1109/CVPRW.2017.156. 8, 12, 13, 14

Jindong Jiang, Zhijun Zhang, Yongqian Huang, and Lunan Zheng. Incorporating Depth into both CNN and CRF for Indoor Semantic Segmentation. *arXiv:1705.07383 [cs]*, May 2017. 12

Eric Jones, Travis Oliphant, Pearu Peterson, and others. SciPy: Open source scientific tools for Python. http://www.scipy.org/, 2001. 8

Iasonas Kokkinos. Pushing the Boundaries of Boundary Detection using Deep Learning. *arXiv:1511.07386 [cs]*, November 2015. 3

TT. Hoang Ngan Le, Kha Gia Quach, Khoa Luu, Chi Nhan Duong, and Marios Savvides. Reformulating Level Sets as Deep Recurrent Neural Network Approach to Semantic Segmentation. *IEEE Transactions on Image Processing*, 27(5):2393–2407, May 2018. ISSN 1057-7149. doi: 10.1109/TIP.2018.2794205. 2, 3

Guosheng Lin, Chunhua Shen, Anton Van Den Hengel, and Ian Reid. Efficient piecewise training of deep structured models for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3194–3203, Las Vegas, United States, 2016. doi: 10.1109/CVPR.2016.348. 3

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common Objects in Context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, number 8693 in Lecture Notes in Computer Science, pages 740–755. Springer International Publishing, September 2014. ISBN 978-3-319-10601-4 978-3-319-10602-1. doi: 10.1007/978-3-319-10602-1_48. 2

Yun Liu, Ming-Ming Cheng, Xiaowei Hu, Kai Wang, and Xiang Bai. Richer Convolutional Features for Edge Detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3000–3009, 2017. 3

Ziwei Liu, Xiaoxiao Li, Ping Luo, Chen Change Loy, and Xiaoou Tang. Deep Learning Markov Random Field for Semantic Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(8):1814–1828, August 2018. ISSN 0162-8828. doi: 10.1109/TPAMI.2017.2737535. 2, 3

Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3431–3440, June 2015. doi: 10.1109/CVPR.2015.7298965. 2

Emmanuel Maggiori, Yuliya Tarabalka, Guillaume Charpiat, and Pierre Alliez. Can Semantic Labeling Methods Generalize to Any City? The Inria Aerial Image Labeling Benchmark. In *Proceedings of the IEEE International Symposium on Geoscience and Remote Sensing (IGARSS)*, July 2017. doi: 10.1109/IGARSS.2017.8127684. 2, 7

K. K. Maninis, J. Pont-Tuset, P. Arbelaez, and L. Van Gool. Convolutional Oriented Boundaries: From Image Segmentation to High-Level Tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):819–833, April 2018. ISSN 0162-8828. doi: 10.1109/TPAMI.2017.2700300. 3

Dimitrios Marmanis, Konrad Schindler, Jan Dirk Wegner, Silvano Galliani, Mihai Datcu, and Uwe Stilla. Classification With an Edge: Improving Semantic Image Segmentation with Boundary Detection. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2017. doi: 10.1016/j.isprsjprs.2017.11.009. 3, 9

Calvin R. Maurer, Rensheng Qi, and Vijay Raghavan. A linear time algorithm for computing exact Euclidean distance transforms of binary images in arbitrary dimensions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(2):265–270, February 2003. ISSN 0162-8828. doi: 10.1109/TPAMI.2003.1177156. 5

Pedro O. Pinheiro, Tsung-Yi Lin, Ronan Collobert, and Piotr Dollár. Learning to Refine Object Segments. In *Computer Vision – ECCV 2016*, Lecture Notes in Computer Science, pages 75–91. Springer, Cham, October 2016. ISBN 978-3-319-46447-3 978-3-319-46448-0. doi: 10.1007/978-3-319-46448-0_5. 4

Xiaojuan Qi, Renjie Liao, Jiaya Jia, Sanja Fidler, and Raquel Urtasun. 3D Graph Neural Networks for RGBD Semantic Segmentation. In *Proceedings of the International Conference on Computer Vision*, 2017. 12

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III*, pages 234–241. Springer International Publishing, Cham, 2015. ISBN 978-3-319-24574-4. doi: 10.1007/978-3-319-24574-4_28. 2, 13

Franz Rottensteiner, Gunho Sohn, Jaewook Jung, Markus Gerke, Caroline Baillard, Sebastien Benitez, and Uwe Breitkopf. The ISPRS benchmark on urban object classification and 3D building reconstruction. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 1:3, 2012. 1, 2, 7

Wei Shen, Xinggang Wang, Yan Wang, Xiang Bai, and Z. Zhang. DeepContour: A deep convolutional feature learned by positive-sharing loss for contour detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3982–3991, June 2015. doi: 10.1109/CVPR.2015.7299024. 3

Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *Proceedings of the International Conference on Learning Representations (ICLR)*, May 2015. 6, 8

L. Sommer, K. Nie, A. Schumann, T. Schuchert, and J. Beyerer. Semantic labeling for improved vehicle detection in aerial imagery. In *IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6, August 2017. doi: 10.1109/AVSS.2017.8078510. 2

S. Song, S. P. Lichtenberg, and J. Xiao. SUN RGB-D: A RGB-D scene understanding benchmark suite. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 567–576, June 2015. doi: 10.1109/CVPR.2015.7298655. 7

Vladimír Ulman, Martin Maska, Klas E. G. Magnusson, Olaf Ronneberger, Carsten Haubold, Nathalie Harder, Pavel Matula, Petr Matula,

David Svoboda, Miroslav Radojevic, Ihor Smal, Karl Rohr, Joakim Jaldén, Helen M. Blau, Oleh Dzyubachyk, Boudewijn Lelieveldt, Pengdong Xiao, Yuexiang Li, Siu-Yeung Cho, Alexandre C. Dufour, Jean-Christophe Olivo-Marin, Constantino C. Reyes-Aldasoro, Jose A. Solis-Lemus, Robert Bensch, Thomas Brox, Johannes Stegmaier, Ralf Mikut, Steffen Wolf, Fred A. Hamprecht, Tiago Esteves, Pedro Quelhas, Ömer Demirel, Lars Malmström, Florian Jug, Pavel Tomancak, Erik Meijering, Arrate Muñoz-Barrutia, Michal Kozubek, and Carlos Ortiz-de-Solorzano. An objective comparison of cell-tracking algorithms. *Nature Methods*, advance online publication, October 2017. ISSN 1548-7091. doi: 10.1038/nmeth.4473. 1, 2

J. Yang, B. Price, S. Cohen, H. Lee, and M. Yang. Object Contour Detection with a Fully Convolutional Encoder-Decoder Network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 193–202, June 2016. doi: 10.1109/CVPR.2016.28. 3

Q. Z. Ye. The signed Euclidean distance transform and its applications. In *[1988 Proceedings] 9th International Conference on Pattern Recognition*, pages 495–499 vol.1, November 1988. doi: 10.1109/ICPR.1988.28276. 5

Fisher Yu and Vladlen Koltun. Multi-Scale Context Aggregation by Dilated Convolutions. In *Proceedings of the International Conference on Learning Representations (ICLR)*, November 2015. 2

Zhiding Yu, Chen Feng, Ming-Yu Liu, and Srikumar Ramalingam. CASENet: Deep Category-Aware Semantic Edge Detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5964–5973, 2017. 3

F. d A. Zampirolli and L. Filipe. A Fast CUDA-Based Implementation for the Euclidean Distance Transform. In *International Conference on High Performance Computing Simulation*, pages 815–818, July 2017. doi: 10.1109/HPCS.2017.123. 8

Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid Scene Parsing Network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2881–2890, Honolulu, United States, July 2017. doi: 10.1109/CVPR.2017.660. 2, 7, 8, 12

Shuai Zheng, Sadeep Jayasumana, Bernardino Romera-Paredes, Vibhav Vineet, Zhizhong Su, Dalong Du, Chang Huang, and Philip H. S. Torr. Conditional Random Fields as Recurrent Neural Networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1529–1537, December 2015. doi: 10.1109/ICCV.2015.179. 1, 2, 3