Aerial Road Segmentation in the Presence of Topological Label Noise

Corentin Henry
Remote Sensing Technology Institute
German Aerospace Center (DLR)
Oberpfaffenhofen, Germany
Email: corentin.henry@dlr.de

Friedrich Fraundorfer
Institute of Computer Graphics and Vision
Graz University of Technology (TUG)
Graz, Austria
Email: fraundorfer@icg.tugraz.at

Eleonora Vig
Amazon
Berlin, Germany
Email: eleonov@amazon.com

Abstract*—The availability of large-scale annotated datasets has enabled Fully-Convolutional Neural Networks to reach outstanding performance on road extraction in aerial images. However, high-quality pixel-level annotation is expensive to produce and even manually labeled data often contains topological errors. Trading off quality for quantity, many datasets rely on already available yet noisy labels, for example from OpenStreetMap. In this paper, we explore the training of custom U-Nets built with ResNet and DenseNet backbones using noise-aware losses that are robust towards label omission and registration noise. We perform an extensive evaluation of standard and noise-aware losses, including a novel Bootstrapped DICE-Coefficient loss, on two challenging road segmentation benchmarks. Our losses yield a consistent improvement in overall extraction quality and exhibit a strong capacity to cope with severe label noise. Our method generalizes well to two other fine-grained topology delineation tasks: surface crack detection for quality inspection and cell membrane extraction in electron microscopy imagery.

I. Introduction

In the era of digitization, we set the goal of mapping the entire surface of the world. Our motivation to monitor human infrastructures is driven by the rapid expansion of urban areas: transportation networks and high-definition maps in particular are being intensively investigated thanks to their relevance in autonomous driving. Although road extraction in aerial and satellite images has already been studied for decades, it remains a complex topic. Unlike other object types such as buildings and vehicles, roads are continuous objects often arbitrarily shaped and organized in a complex topology. In 2013, a Convolutional Neural Network (CNN) first surpassed traditional road extraction algorithms, leveraging a huge dataset of annotated satellite images, the Massachusetts Roads Dataset [1]. Since then, more datasets of increasing difficulty were released [2], [3]. On the methods side, a clear trend shows a preference for variations of the renowned U-Net [4], a Fully-Convolutional Neural Network (FCNN) introduced back in 2015, over state-of-the-art architectures in semantic segmentation, such as DeepLabv3+ [5] and DenseASPP [6].

At present, a major obstacle to improving the performance of U-Net-like FCNNs is the annotation quality of benchmark datasets. As partly described in [7], road label inaccuracies come from several sources. First, *omission noise* when objects of interest are missed by the annotators. Second, *registration noise* when labels are offset compared to the object beneath them. Third, *over-simplification of the labels* when, for instance, variable-thickness objects are annotated with fixed-thickness lines. In the context of road segmentation, all three issues sometimes come from the use of OpenStreetMap (OSM) rasterized vector data, a less resource-intensive labeling process than pixel-wise annotation. It may also occur when a ground truth is drawn on top of one image source, say optical RGB images, and transferred to another image source of the same geographic area, say radar images. In this case, some roads might appear in slightly different positions, while others might not be visible anymore. Label noise results in a greater difficulty to train FCNNs, but more importantly it eventually

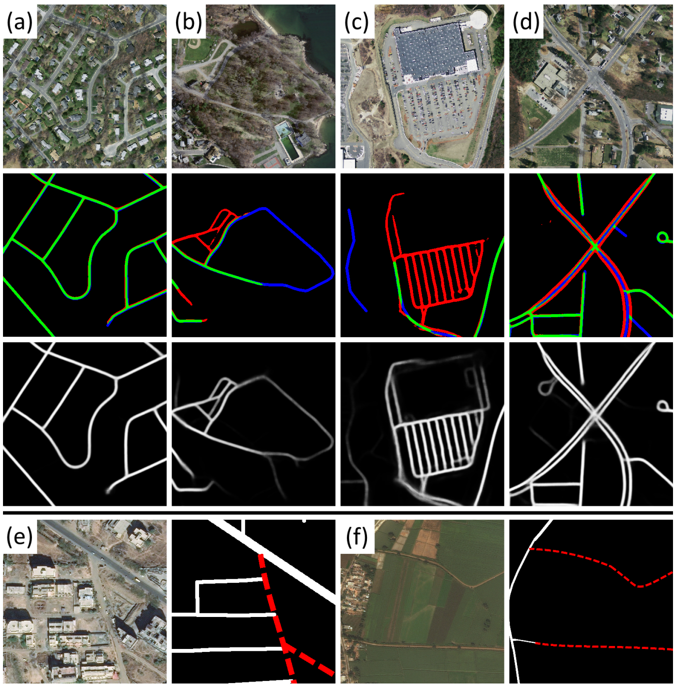

Fig. 1. Visualization of different types of label noise and their effect on the predictions. (a-d) Triplets of RGB, confusion, and probability samples of our Dense-U-Net-121 on the test set of Massachusetts. In confusion maps, green are true positives, blue false negatives, and red false positives. (a) typical almost perfect predictions; (b) inconsistent labels; (c) non-annotated parking lot; (d) incorrect label topology (single instead of dual-lane). (e-f) RGB and label samples from the training set of DeepGlobe. The non-annotated roads are overlaid with red dash-lines.

makes the benchmark less reliable if the test set is also noisy. Concrete examples of noise-effects can be found in Figure 1. For all these reasons, FCNNs must be trained robustly to label noise. In the machine learning literature, one solution is to use *noise-aware losses* [8] which re-balance the confidence granted to the ground truth in favor of the predictions.

In this paper, we introduce such noise-aware losses in the context of pixel-wise labeling, more precisely for road segmentation in aerial images. To this end, **(1)** to benchmark existing segmentation losses for road extraction, we first train custom U-Nets with ResNet and DenseNet backbones using several (metric-based) losses, alone or in combinations. **(2)** We then adapt several *noise-aware losses* from image classification, such as the Sigmoid, Unhinged, Savage, and Bootstrapped BCE, and demonstrate their benefits, especially in the presence of more severe topological noise. **(3)** We also introduce a novel noise-aware loss, the *Soft-Bootstrapped DICE*, which is most efficient in recovering from high levels of label noise. **(4)** We experimentally validate our findings and reach competitive performance on the challenging Massachusetts [1] and DeepGlobe [2] road extraction datasets, and **(5)** show good generalization on two other topology delineation tasks from materials science and medical imaging.

II. Related work

Road segmentation is a long-studied topic in computer vision, which is often formulated as a two-phase extraction: (1) a *binary pixel-wise segmentation* of road candidates with a deep segmentation network as a starting point [1], [9], [10], [11], [12], supporting (2) the *inference of a topological graph* in the form of nodes and edges. This is the approach taken in DeepRoadMapper [13], RoadTracer [14] and [15]. While only the SpaceNet dataset [3] covers both stages, several challenging benchmarks are available for binary road segmentation, namely the Massachusetts Roads [1], the DeepGlobe Road Extraction Challenge [2], and TorontoCity [16] (not released yet). Our focus is on stage one: binary road segmentation.

Aerial road extraction presents unique difficulties as opposed to ground imagery, especially on lower-resolution satellite imagery with Ground Sampling Distances (GSD) of 1 m to 30 cm/pixel. Class imbalance due to roads' thinness and sparsity in the images is usually solved in three ways: using a weighted-loss to emphasize certain classes [17], [18], using relaxed or specialized metrics during evaluation to increase the spatial tolerance towards small mistakes [19], [1], [20], [9], or using special CNNs, losses or post-processing [13], [9], [21], [22]. Somewhat surprisingly, the current benchmark state of the art is not based on complex segmentation architectures found in ground imagery [5], [6], but derived from U-Net [4], a much simpler FCNN for medical image analysis. The top-4 performers on the DeepGlobe challenge are D-LinkNet [10], U-Net-like Resnet [11], Residual Inception SkipNet [12], and EOSResUNet [23], all derivatives of U-Net. It makes all the more sense since tasks in the medical field are subject to the same imbalance, topology and annotation issues as in road extraction. Similarly to FC-DenseNet [24], we explore the fusion of ResNets [25] and DenseNets [26] with U-Net, but contrary to this previous work, we use the original architectures as backbone so as to be able to use the widely available ImageNet pre-trained weights.

Learning from noisy labels is tackled by several lines of research [27]. Many works focus on designing *noise-aware* or *noise-corrected losses* [28], [8], [29], [30], [31], [32], including the Unhinged [28], the Sigmoid [29], the Ramp [32], and the Savage [30] losses. Some of these losses, however, require a noise rate estimation, in the form of the noise transition matrix [8], which makes more sense in a multi-class (especially in the fine-grained) classification setting, and less in binary classification. Reed *et al.* [31] proposed a bootstrapping scheme to combine training labels and the current model's prediction to generate new training targets, thus avoiding the explicit modeling of the noise distribution.

Noise-robustness has almost exclusively been studied in the context of image classification and rarely in pixel-level labeling. An early exception is [7], whose deep-learning method for binary road labeling explicitly handles omission and registration noise. Here, we adapt noise-correcting losses and perform label noise reduction, all initially proposed in the context of image-level labeling, to the task of semantic segmentation and more specifically to aerial road labeling. We perform an exhaustive evaluation of standard losses, existing noise-correcting losses, and our new loss, the Soft-Bootstrapped DICE loss, and show that our new loss is optimal in the face of severe topology noise.

III. Methods

A. Res- and Dense-U-Nets for road segmentation

As demonstrated by the state of the art in three public benchmarks [1], [2], [3], U-Nets constitute a competitive baseline for road segmentation in aerial images. However, the original U-Net suffers from a performance bottleneck. It features a unique backbone which is not used by any other FCNN, hence is rarely pre-trained on ImageNet. Yet research has shown the importance of such pre-training when training data is limited. We overcome this limitation by designing a FCNN in all points similar to U-Net, only differing in what backbone network is used. We replace both the encoder and the decoder by either a ResNet [25] or a DenseNet [26]. As opposed to very deep and more complex FCNNs, these two networks still perform best on aerial image understanding tasks because (with fewer pooling layers) they preserve fine-grained details better than the typical state-of-the-art classification CNNs.

A U-Net should maintain a symmetry between its encoder and its decoder. In our FCNN, a given residual or dense block in the encoder is connected via a skip-connection to its corresponding block in the decoder. Both blocks are identically configured (same architecture and number of layers and output features). In each decoder block, the feature map is first up-sampled using a transposed convolution and then concatenated with the feature map from the corresponding encoder dense block. Before being processed by a decoder dense block, its

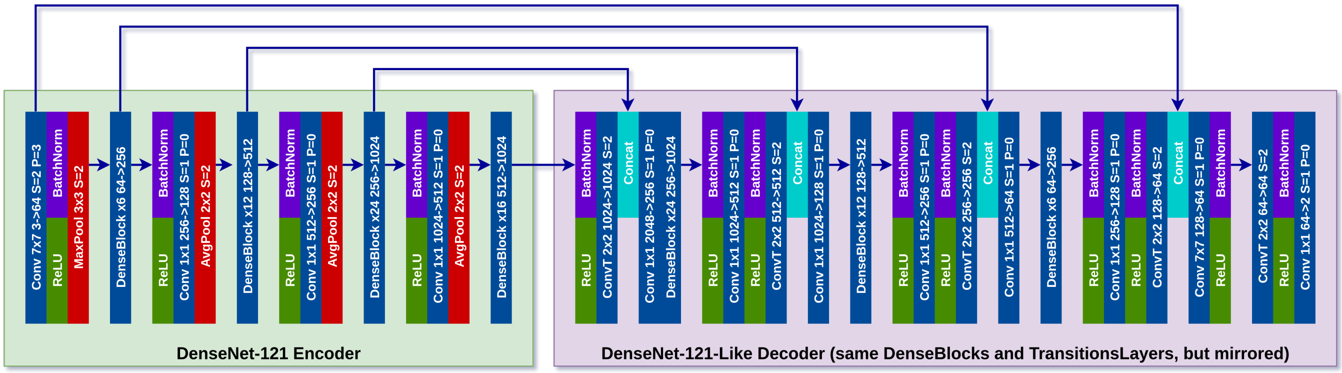

Fig. 2. The Dense-U-Net-121 architecture. Blocks sequence order: left to right, top to bottom. Nomenclature: ConvT stands for Transposed Convolution, NxN for kernel size, S=N for stride size, P=N for padding size, IN→OUT for number of input and output features. We refer the readers to the original DenseNet paper [26] for details on the DenseBlocks.

depth is reduced by a 1×1 convolution layer, down to the input feature size of the dense block. Figure 2 shows the resulting architecture with DenseNet-121 used as encoder, and the individual U-Net blocks replaced with DenseNet blocks in the decoder. Depending on their backbone, we refer to such U-Nets as *Dense-U-Nets* and *Res-U-Nets*, respectively.

B. Segmentation losses for aerial road labeling

The standard training loss used in single-class segmentation is Binary Cross-Entropy (BCE). It has one defect however, particularly visible when facing a high class imbalance: when detecting linear structures as in road segmentation tasks, the boundaries of small and thin objects need to reach pixel-accuracy, for which BCE requires extensive fine-tuning. To emphasize the errors on these sensitive borders, the community has been using differentiable versions of quality metrics as losses: the DICE coefficient (or F1-score) loss [33] and the Soft-IoU loss [13]. Like their metric counterparts, they clearly reveal the margin for improvement around imprecisely extracted objects.

Due to the sparse nature of our road labels, it is necessary to take precautions regarding the numerical stability of the losses. They must account for the possibility that some output or ground truth patches might be empty. The BCE is resilient by default, but the IoU and DICE losses must be slightly modified. The expected behavior is the following: shall both the labels and the predicted patches be all zeros, the denominators should be different from zero and the losses should be zero. In equations 2 and 3, we ensure this by the addition of a 1 outside the numerators' and denominators' sums. In all other cases, these additional 1s have a negligible impact on the loss compared to the thousands of pixels they are summed up with. In the loss formulas below, $y_{ik} \in \{0, 1\}$ is the value of the binary ground truth for pixel i and class $k \in C = \{0, 1\}$, 1 for road and 0 for background, $p_{ik} \in [0, 1]$ is the predicted probability for pixel i and class k, p_{i1} for roads and p_{i0} for background, and N is the total number of pixels in the image:

$$\mathrm{BCE}(y, p) = -\sum_{k}^{C} \sum_{i}^{N} y_{ik} \log(p_{ik}) \tag{1}$$

$$\mathrm{DICE}(y, p) = 1 - \frac{1 + \sum_i^N 2 y_i p_i}{1 + \sum_i^N (y_i^2 + p_i^2)} \tag{2}$$

$$\mathrm{IoU}(y, p) = 1 - \frac{1 + \sum_i^N y_i p_i}{1 + \sum_i^N (y_i + p_i - y_i p_i)} \tag{3}$$

We also experiment with combinations of losses, in which case we compute the unweighted sum of individual losses, after having confirmed that these individual losses operate roughly in the same range.

C. Noise-aware losses

As stated in the introduction, road labels are often subject to omission and registration noise. In such cases the ground truth cannot be entirely trusted, and the predictions might actually hold more accurate information. Similarly to what is done in multi-class classification tasks, where object categories may be swapped, it is possible to balance the confidence between the predictions and the labels in segmentation, so that good road candidates or absence thereof are not penalized because of a faulty ground truth. We therefore adapt several noise-aware losses from the classification literature to assess their potential on our segmentation task. Some of the losses work by blurring the line between labels and predictions by giving them a symmetrical influence on the result, meaning that swapping them in the formula would result in the same loss: the Unhinged [28], Sigmoid [29], Ramp [32], and Savage [30] losses [8]:

$$\mathrm{Unhinged}(y, p) = \frac{1}{N} \sum_{i}^{N} (1 - y_i p_i) \tag{4}$$

$$\mathrm{Ramp}(y, p) = \frac{1}{N} \sum_{i}^{N} \max\left(0, \min(1, 1 - \beta y_i p_i)\right) \tag{5}$$

$$\mathrm{Sigmoid}(y, p) = \frac{1}{N} \sum_{i}^{N} \mathrm{Sigmoid}(-\beta y_i p_i) \tag{6}$$

$$\mathrm{Savage}(y,p) = \frac{1}{N}\sum_{i}^{N} \frac{1}{\left[1 + e^{2y_i \max\left(\epsilon, \min(1-\epsilon, p_i)\right)}\right]^2} \tag{7}$$

We arbitrarily set $\epsilon = 10^{-6}$. A parameter β is used to control the degree of confidence given to either the ground truth or the predictions. Other losses downgrade the importance of the labels and proportionately increase the confidence attributed to the predictions.

In addition to the above losses, we use the Hard- (HB) and Soft-Bootstrapped (SB) BCE losses from [31], and introduce a novel Soft-Bootstrapped DICE Coefficient loss:

$$\mathrm{BCE}_{\mathrm{HB}}(y,p,\hat{p}) = -\sum_{k}^{C}\sum_{i}^{N}\left(\beta y_{ik} + (1-\beta)\hat{p}_{ik}\right)\log(p_{ik}) \tag{8}$$

$$\mathrm{BCE}_{\mathrm{SB}}(y,p) = -\sum_{k}^{C}\sum_{i}^{N}[\beta y_{ik} + (1-\beta)p_{ik}]\log(p_{ik}) \tag{9}$$

$$\mathrm{DICE}_{\mathrm{SB}}(y,p) = 1 - \frac{1 + \sum_i^N 2[\beta y_i + (1-\beta)p_i]p_i}{1 + \sum_i^N [\beta y_i + (1-\beta)p_i]^2 + p_i^2} \tag{10}$$

In Eq. 8, $\hat{p}_{ik} \in \{0,1\}$ is the value of p_{ik} thresholded at 0.5. To prevent numerical instability in the Soft-Bootstrapped DICE loss, we resort to the same strategy as in Sec. III-B.

IV. Experiments

We first describe the two road extraction datasets used in our experiments and provide implementation details. Then, we evaluate several different FCNN architectures and perform an ablative analysis of the above losses. Next, we artificially corrupt labels to study how well the different noise-correcting losses can recover from noisy training data. Finally, we perform a generalization study on two other topology delineation tasks.

A. Road datasets

Massachusetts Roads Dataset (MA): [1] contains 1171 satellite images of size 1500×1500 acquired at a GSD of 1 m/px. The labels are binary images where roads are represented as constant 7-pixel thick lines generated by rasterizing and dilating OpenStreetMap vector centerlines. The dataset is split into 1108, 14, and 49 training/validation/test images, corresponding to 94.6%, 1.2%, and 4.2% of the images. Validation and test splits are thus statistically under-representative of the whole dataset. To increase our results' consistency, we perform our ablation study on a re-split with an 80%, 10%, and 10% split and report our final results on the official split in Sec. IV-E.

DeepGlobe'18 Road Extraction Challenge (DG): [2] contains 8570 satellite images of size 1024×1024 acquired at a GSD of 50 cm/px. Labels are binary images with manually-annotated variable-width roads. DG is split into 6226 training, 1243 validation, and 1101 test images, and labels for validation and test splits are not public. Our ablative analysis is therefore conducted on a re-split (4983 training, 1243 validation images) of the official training set. We perform the final study on the official split in Sec. IV-E.

B. Implementation details

We implement several versions of our method: Res-U-Net18, Res-U-Net34, Res-U-Net50, Res-U-Net101, Dense-U-Net-121, Dense-U-Net-169, and Dense-U-Net-201, by varying the backbones used, e.g. ResNet-18, ResNet-34, DenseNet-121, etc.

Weights initialization: we use ImageNet pre-trained weights for all encoders except in two baselines: U-Net and DeepLabv3+. The decoders, as well as the encoders in U-Net and DeepLabv3+, are initialized as proposed in [34].

Data augmentation: we perform random horizontal flips and rotations in $\{90°, 180°, 270°\}$.

Losses: for the Soft-Bootstrapped losses, we cross-validate the β parameter in the range $\{0.4, 0.5, ..., 0.9\}$.

Training: we train our FCNNs over 40 epochs and with a fixed learning rate of 10^{-4} with the ADAM optimizer and no L2 weight decay.

Post-processing: to improve the segmentation performance for our best models, we perform Test-Time Augmentation (TTA). We apply eight combinations of flips and 90-degree rotations to the input images, run the model and revert the transformations on the softmax outputs (values in $[0.0, 1.0]$). We perform a sum-merge (values in $[0.0, 8.0]$) and threshold at 4.0 to obtain the binary masks.

Metrics: we report IoU, DICE/F1-score, Precision, and Recall (in percentages). Additionally, we compute the road metrics from [19], namely Correctness, Completeness, and Quality, which measure the topological proximity between the ground truth and predicted centerlines (also in percentages). They allow for spatial tolerance controlled by a buffer width (3 pixels). Due to space constraints, we sometimes only report IoU/F1/Quality, please refer to the supplementary materials for the full tables. There, we also provide details on the road metrics.

C. Analysis of various FCNN architectures

For our analysis of the optimal FCNN architecture, we compare our networks with the following state-of-the-art semantic segmentation baselines: DeepLabv3+ (with Xception65 backbone) [5], DenseASPP (DenseNet-121 backbone) [6], U-Net (with BatchNorm) [4], D-LinkNet34, D-LinkNet50, and D-LinkNet101 [10]. These networks are trained using a BCE loss and the optimizers recommended by their respective authors, with an initial learning rate of 10^{-4} and learning rate schedules scaled to 40 iterations.

In Table I, we report the performance obtained by the best baseline FCNNs. As anticipated, the state-of-the-art DeepLabv3+ and DenseASPP are losing to U-Net-like architectures. Overall, Dense-U-Nets reach higher scores than Res-U-Nets and the other U-Net-like networks: we consider Dense-U-Net-121 for the rest of the study as it reaches the highest IoU and F1 on both MA and DG. Finally, we confirm the benefits of pre-training on ImageNet, with +1.63% IoU on MA and +3.01% IoU on DG. Additional models are included in the tables in the supplementary materials.

TABLE I
BASELINE RESULTS ON MA/DG RE-SPLITS (* NOT PRE-TRAINED).

Model	MA test re-split			DG valid. re-split		
	IoU	F1	Qual.	IoU	F1	Qual.
DeepLabv3+*	52.95	69.35	66.74	59.65	75.19	64.80
DenseASPP	46.63	64.44	62.32	61.46	76.78	69.14
D-LinkNet50	54.90	71.01	68.25	58.12	74.04	64.75
U-Net*	55.92	71.91	69.25	61.97	76.82	68.65
Res-U-Net50	56.93	72.74	69.89	64.55	78.62	71.59
Dense-U-Net-121	**57.12**	**73.03**	**70.06**	**65.13**	**79.19**	**72.43**
Dense-U-Net-121*	55.49	71.44	68.69	62.12	76.99	69.75

TABLE II
LOSSES COMPARISON FOR DENSE-U-NET-121 ON MA/DG RE-SPLITS (WITH PER-MODEL OPTIMAL β PARAMETERS CROSS-VALIDATED ON VALIDATION RE-SPLIT).

Loss	MA test re-split			DG valid. re-split		
	IoU	F1	Qual.	IoU	F1	Qual.
BCE	57.12	73.03	70.06	65.13	79.19	72.43
DICE	57.50	73.53	69.76	65.18	79.02	72.08
IoU	57.19	73.38	69.40	63.19	77.58	68.01
BCE + DICE	57.65	73.43	70.08	65.43	79.37	72.08
BCE + IoU	58.12	73.84	70.48	63.83	78.35	70.73
BCE + DICE + IoU	57.99	73.71	70.20	65.57	79.42	72.29
BCE + Sigmoid	57.88	73.51	70.60	65.76	79.63	73.02
BCE + Unhinged	57.72	73.47	69.82	**65.99**	79.63	73.15
BCE + Savage	57.56	73.15	70.06	65.89	**79.65**	73.14
BCE HB	57.54	73.08	69.90	65.80	79.58	73.12
BCE SB	57.93	73.54	70.29	65.87	79.58	**73.28**
DICE SB [OURS]	**58.26**	**73.91**	**70.74**	65.29	79.08	71.74

D. Loss analysis

Table II reports our results for the different metric-based and noise-aware losses when tested in isolation and in various combinations.

Individual metric-based losses: DICE outperforms BCE and IoU on both datasets, but the advantage is minimal.

Combining metric-based losses: as is common practice in road segmentation challenges, we search for the best combination of BCE, Soft-IoU and DICE Coeff. losses: these are $BCE + IoU$ on MA (+1.00% IoU over BCE) and $BCE + DICE + IoU$ on DG (+0.44% IoU over BCE).

Adding noise-aware losses: we find that adding our noise-robust losses to BCE consistently improves the results (see Table II). For the Sigmoid, Ramp, and Savage losses, we set $\beta = 1$, so the Ramp and Unhinged losses become equivalent. We use them in combination with BCE, as experiments showed that alone they were not sufficient to reach high segmentation accuracy. The Hard-Bootstrapped BCE (**BCE HB**), the Soft-Bootstrapped BCE (**BCE SB**) and the Soft-Bootstrapped DICE (**DICE SB**) use $\beta = 0.7$. On MA, we gain 1.14% IoU by using our novel DICE SB loss. On DG, we gain 0.86% IoU by using BCE + Unhinged. This is as mild an improvement as reported in previous works [8], yet a significant increase in a competitive context. We also cannot expect a substantial improvement as long as the test set is noisy.

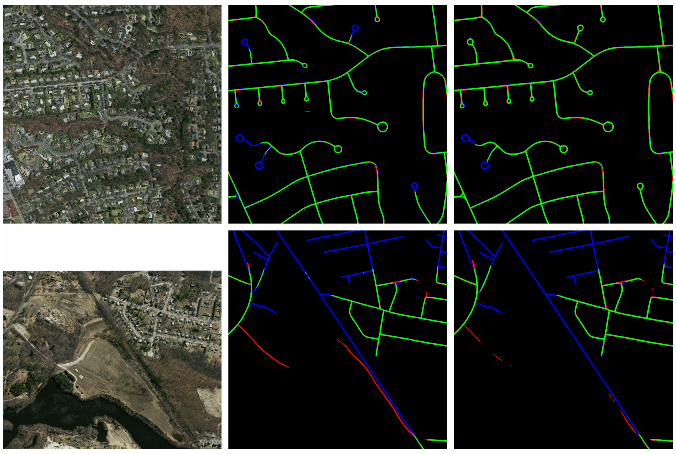

Fig. 3. Results on the test re-split of Massachusetts. Left to right: RGB, Dense-U-Net-121 + BCE confusion, and Dense-U-Net-121 + DICE SB confusion. Some false positives were erased and a few road segments were completed thanks to the noise-aware Soft-Bootstrapped DICE loss. A blank region is visible in the second image, but is still annotated with roads in the ground truth. In total, 13% of the pixels in the dataset belong to blank areas, fortunately almost none in the official test set.

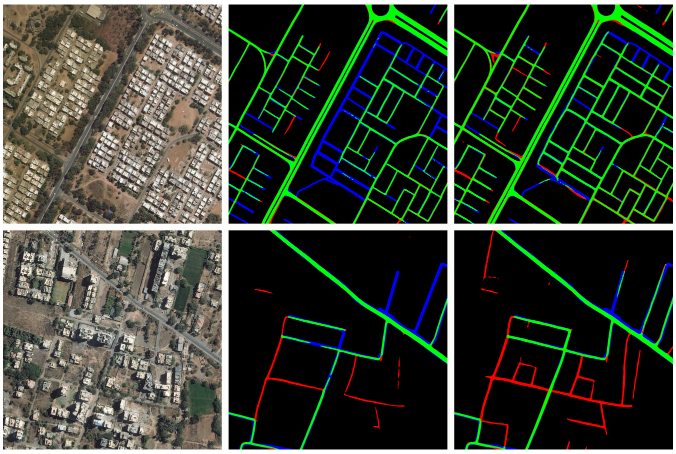

Fig. 4. Results on the validation re-split of DeepGlobe. Left to right: RGB, Dense-U-Net-121 + BCE confusion, and Dense-U-Net-121 + BCE + Unhinged confusion. Top: many additional good road candidates were added by the noise-aware model. Bottom: although many predictions seem to be good road candidates, the ground truth does not consider them as valid.

Qualitative results for BCE and the best noise-correcting losses are shown in Fig. 3 for MA and Fig. 4 for DG.

E. Results on the official splits

We now test our best models on the official benchmarks. For MA (cf. Table IV), we train on the training+validation set and report the performance on the test set. For DG (cf. Table V), we train on the public training set and test on the private validation set through the online evaluation server. With test-time augmentation (TTA), our rather simple U-Net-based model ranks second on MA, on par with the more complicated FCNN architecture RDRCNN. On DG, our model ranks fourth; the performance gap is due to the more

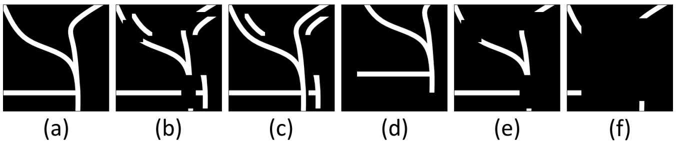

Fig. 5. Examples of synthetic label noise: (a) no noise, (b) ROI shift, (c) ROI duplication, (d) image-wise shift, (e) ROI ablation, (f) half ablation. (Shifts exaggerated for easier visualization)

sophisticated post-processing and model ensembling adopted by the competing methods, which are engineering practices indispensable to perform well on such a challenge, but are beyond the scope of the current paper. We also show that our noise-correcting losses are *independent of the network architecture*: when we replace the loss with the noise-aware BCE SB in the 2nd best D-LinkNet (for which code is available), we obtain a significant boost (cf. Table V in gray).

F. Synthetic label noise

We have seen in Sec. IV-D that, although the two datasets are noisy (cf. Fig. 1), our noise-aware losses do not have a significant edge over standard losses because the test sets are also noisy. Nevertheless, to be able to study the extent to which the ground truth becomes too corrupted to be usable and how much our models can recover from such noisy data, we resort to artificial label corruption.

To this end, we systematically introduce different levels of synthetic label noise that either mimic slight human errors or outright wrong labels. On the one hand, we perform **(1) label shifts** either *image-wise* or on several small *Regions of Interest (RoI)* drawn randomly. The RoI may be *translated* or *duplicated*, the direction is random and the shift distance is either *random* or *fixed*. On the other hand, we perform **(2) label ablation** by zeroing *several small RoI* or a *single large RoI* covering half of the image. See Figure 5 for an illustration. We experiment with the following noise types of increasing severity:

1) *Label shift*: **Shift1** – duplicated small ROIs with random shifts; **Shift2** – translated small ROIs with random shifts; **Shift3** – whole-image translation with small random shift (shift in 0-7 pixels); **Shift4** - whole-image translation with fixed 7 pixel shift.
2) *Label ablation*: **Ablation1** – removing several small RoIs; **Ablation2** – removing a large RoI (half image).
3) *Label shift and ablation*: **Shift2 + Ablation1** and **Shift2 + Ablation2**.

TABLE III
PERFORMANCE OF DENSE-U-NET-121 TRAINED WITH SYNTHETIC LABEL NOISE ON THE RE-SPLITS OF MA AND DG, WITHOUT NOISE CORRECTION (BCE LOSS) AND WITH CORRECTION (BCE SB OR DICE SB LOSS). SEE IV-F FOR A DESCRIPTION OF NOISE TYPES AND LEVELS.

Noise	Loss	Noise	MA test re-split			DG valid. re-split		
Type/Level		Corr.	IoU	F1	Qual.	IoU	F1	Qual.
No Noise	BCE	×	57.12	73.03	70.06	65.13	79.19	72.43
	BCE	✓	57.87	**73.53**	70.02	**65.87**	**79.58**	**73.28**
	DICE	✓	**57.91**	73.30	**70.22**	64.88	79.00	71.69
Shift1	BCE	×	57.19	73.10	70.27	66.05	79.49	73.36
	BCE	✓	**57.54**	**73.30**	**69.90**	66.36	79.85	72.94
	DICE	✓	56.85	72.85	69.24	**68.03**	**81.13**	**74.77**
Shift2	BCE	×	57.01	72.92	**70.45**	66.00	79.65	73.61
	BCE	✓	**57.16**	72.99	69.51	66.03	79.61	72.66
	DICE	✓	56.79	**73.02**	68.97	**67.72**	**80.91**	**74.89**
Shift3	BCE	×	47.26	64.47	63.42	61.14	76.27	65.87
	BCE	✓	**53.80**	**70.84**	**65.23**	**63.01**	**77.65**	68.55
	DICE	✓	41.63	60.56	64.72	59.34	69.71	**75.21**
Shift4	BCE	×	4.18	8.25	5.95	34.28	51.10	27.47
	BCE	✓	12.16	22.58	12.43	41.80	58.45	34.48
	DICE	✓	**23.24**	**39.39**	**20.83**	**42.58**	**59.71**	**42.76**
Ablation1	BCE	×	56.09	71.91	69.82	65.47	79.32	73.37
	BCE	✓	57.15	72.89	**69.91**	65.88	79.69	73.08
	DICE	✓	**57.67**	**73.53**	69.85	**67.04**	**80.54**	**74.46**
Ablation2	BCE	×	0.00	0.00	0.01	0.03	0.06	0.41
	BCE	✓	38.34	55.35	63.47	45.90	63.04	60.87
	DICE	✓	**57.11**	**70.11**	**72.92**	**64.41**	**78.73**	**71.31**
Sh.2+Ab.1	BCE	✓	56.93	72.80	69.08	65.00	78.92	72.42
	DICE	✓	**57.42**	**73.22**	**69.31**	**67.30**	**80.78**	**74.61**
Sh.2+Ab.2	BCE	✓	40.20	57.31	65.92	47.87	64.86	64.29
	DICE	✓	**57.06**	**69.64**	**72.60**	**64.83**	**79.05**	**71.69**

Results for different noise levels on the test re-splits of MA and DG are reported in Table III. There, for each noise type, we compare the standard BCE (no noise-correction) with the noise-corrected BCE and our new loss, the Soft-Bootstrapped DICE (DICE SB). We list here the key findings: (a) Noise correction is always beneficial when the training set is corrupted. (b) When the noise level is not too severe (as in Shift1, Shift2, and Shift3), on MA the BCE SB has a slight advantage over DICE SB. On DeepGlobe, however, DICE SB outperforms BCE SB for all noise types and levels. (c) When faced with more severe noise (such as full-image label shifts, or missing annotations of any size – Shift4 and Ablations), DICE SB is the optimal countermeasure outperforming BCE SB. Surprisingly, despite such drastic corruptions as removing annotations for half of the image (Ablation2), our DICE SB can fully recover to match the performance of models trained on non-corrupted training sets (cf. 'No Noise', 57.11% vs 57.91% IoU for MA, 64.41% vs 64.88% IoU for DG), whereas BCE SB lags behind (MA: 38.34%, DG: 45.90%). Without noise correction, the training fails (0% IoU). (d) Interestingly, when noise-aware losses are used, slight shifts (or jitters) in the training labels (as in Shift1 and Shift2) lead to improvements w.r.t the non-corrupted data and standard losses, which is more pronounced for DG: observe the big jump from 65.13% IoU to 68.03% (Shift1) and 67.72% (Shift2). These synthetic shifts can thus be interpreted as a novel way of data augmentation.

G. Generalization studies

To validate our findings on similar topological delineation tasks with widely different image sources, we ran two generalization studies on (1) surface crack detection and on (2) cell membrane detection in EM imagery.

CrackTree Dataset (CT): [43] contains 206 pavement images of size 800 × 600 featuring various kinds of surface cracks, and are challenging due to the presence of shadows, occlusion, low contrast, and noise. As no training/test split is provided, we perform a 187/19 split (not performing hyperparameter

TABLE IV
RESULTS OF DENSE-U-NET-121 + BCE SB + IOU AND STATE-OF-THE-ART ON OFFICIAL MA TEST SET; * NO POST-PROCESSING; BEST AND SECOND BEST; TTA - TEST-TIME AUGMENTATION.

Models	IoU	F1	Prec.	Rec.
RSRCNN [35]	49.46	66.20	60.60	72.90
Modified U-Net [36]	59.76	74.54	74.15	75.48
JointNet [37]	64.00	78.05	71.90	**85.36**
WRAU-Net [38]	64.58	78.48	74.50	**82.90**
MFPN [21]	65.70	79.30	**85.10**	74.20
RDRCNN* [39]	66.28	79.72	84.64	75.33
RDRCNN [39]	**67.10**	**80.31**	**85.35**	75.75
OURS w/o TTA	65.16	78.89	79.55	78.25
OURS w/ TTA	**66.61**	**79.98**	81.67	78.35

TABLE V
RESULTS OF DENSE-U-NET-121 + BCE SB + RAMP AND LEADERBOARD ON OFFICIAL DG VALIDATION SET; OUR REPLICATION OF D-LINKNET RESULTS, ALSO MODIFIED TO USE BCE SB LOSS.

Models	IoU
Stacked U-Nets [40]	60.00
Ensemble U-Nets [41]	60.58
ResNet50-D2S [42]	60.60
U-Net-like ResNet34 [11]	64.00
D-LinkNet [10]	64.12
D-LinkNet [10]	63.29
D-LinkNet [10] + BCE SB	64.36
EOSResUNet [23]	**65.60**
OURS w/ TTA	63.52

validation). The labels are binary images with single-pixel centerlines, which we dilated to 5-pixel thick lines as in [9]. For the Correctness, Completeness, and Quality metrics, we use a 2 pixel buffer width.

In Table VI, we report results for our Dense-U-Net-121 trained with different losses. We achieve excellent performance, though not directly comparable to other works (such as [9]) because there is no official dataset split. Highest IoU is reached with a joint BCE+IoU loss, however DICE DB comes close giving highest Completeness value. Qualitative results are shown in Fig. 6.

Electron Microscopy (EM) dataset of the ISBI'12 challenge [44] contains 60 images of size 512 × 512 from neural tissue and the corresponding binary cell boundary labels, among which 30 were kept private. Similar to [9] we split the 30 training images into 15 for training / 15 for test.

We report quantitative results in Table VII and show qualitative ones in Fig. 7. We obtain good results, close to [9], though again not directly comparable because of the different split. The BCE+Unhinged loss outperforms the BCE loss by more than 2% IoU. As visible in the qualitative results, the cells are diffuse and the annotation a matter of human perception, which makes the dataset a challenge for current state-of-the-art methods.

TABLE VI
RESULTS OF DENSE-U-NET-121 TRAINED WITH DIFFERENT LOSSES ON OUR OWN TEST SPLIT OF CT, COMPARED TO A NOT PRE-TRAINED U-NET (* NOT DIRECTLY COMPARABLE AS USES A DIFFERENT SPLIT).

Loss	IoU	F1	Corr.	Comp.	Qual.
Topology loss [9]*	-	-	88.44	95.13	84.61
BCE [U-Net]	81.79	89.83	**93.71**	91.37	86.48
BCE	81.96	89.96	92.22	92.94	86.54
BCE + IoU	**82.90**	**90.53**	92.64	93.41	87.31
BCE + Sigmoid	82.41	90.20	91.29	93.34	86.15
BCE SB $\beta = 0.5$	82.32	90.18	93.43	93.24	**87.82**
DICE SB $\beta = 0.4$	82.53	90.31	91.00	**93.79**	86.17

TABLE VII
RESULTS OF DENSE-U-NET-121 TRAINED WITH DIFFERENT LOSSES ON OUR OWN TEST SPLIT OF EM, COMPARED TO D-LINKNET50 (* NOT DIRECTLY COMPARABLE AS USES A DIFFERENT SPLIT).

Loss	IoU	F1	Corr.	Comp.	Qual.
Topology loss [9]*	-	-	72.27	73.58	57.22
BCE [D-LinkNet50]	64.64	78.53	67.62	72.32	53.69
BCE	65.41	79.08	66.18	**73.33**	53.34
IoU	66.55	79.93	**71.53**	71.75	**55.79**
BCE + Unhinged	**67.55**	**80.64**	68.27	72.94	54.47
BCE SB $\beta = 0.5$	66.41	79.83	70.22	72.29	55.31
DICE SB $\beta = 0.5$	66.80	80.09	68.18	72.31	54.08

V. CONCLUSION

Our work shows the advantages of using noise-aware losses when training segmentation models with noisy labels. We report consistent performance increases over two challenging road segmentation datasets, and show that our method generalizes well to datasets from other fields (materials quality and medical imagery). More importantly, we show that our new Soft-Bootstrapped DICE loss is especially robust toward high levels of label noise. Furthermore, light synthetic noise proves to be a good data augmentation technique, particularly efficient when used in combination with noise-aware losses, enabling us to reach competitive performance. We further confirm the trend as to which U-Net-like networks are best suited for thin object delineation, and show that they natively cope well with noisy labels during training.

REFERENCES

[1] V. Mnih, "Machine Learning for Aerial Image Labeling," Ph.D. dissertation, University of Toronto, 2013.

[2] I. Demir *et al.*, "DeepGlobe 2018: A Challenge to Parse the Earth Through Satellite Images," in *CVPR Workshops*, 2018.

[3] (2018) SpaceNet on Amazon Web Services (AWS). Datasets. The SpaceNet Catalog.

[4] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Medical Image Computing and Computer-Assisted Intervention*, 2015, pp. 234–241.

[5] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation," *ECCV*, 2018.

[6] M. Yang, K. Yu, C. Zhang, Z. Li, and K. Y. Deepmotion, "DenseASPP for Semantic Segmentation in Street Scenes," in *CVPR*, 2018.

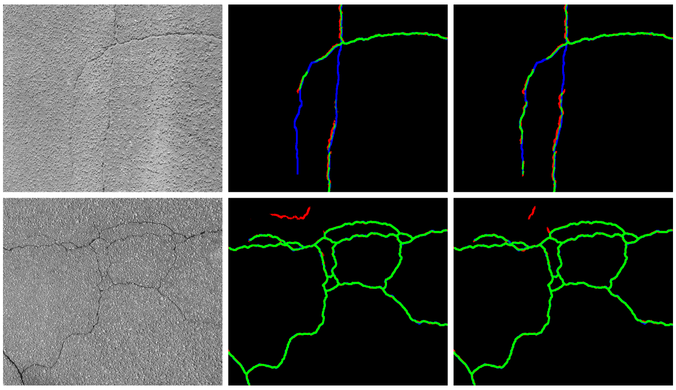

Fig. 6. Results on samples from the CrackTree test set. Left to right: RGB image of cracked pavement, Dense-U-Net-121 + BCE confusion and Dense-U-Net-121 + BCE + IoU confusion, where false positives were removed and more crack pixels were detected.

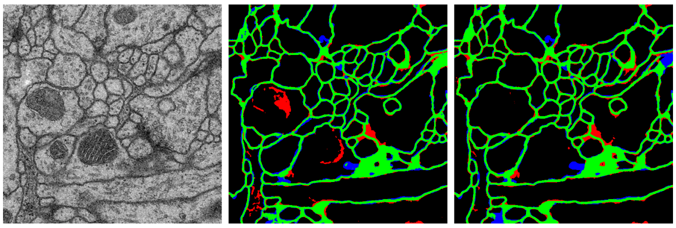

Fig. 7. Results on samples from the Electron Microscopy test set. Left to right: grayscale image of a neuronal tissue slice, Dense-U-Net-121 + BCE confusion and Dense-U-Net-121 + BCE + Unhinged confusion, where false positives were removed.

[7] V. Mnih and G. Hinton, "Learning to label aerial images from noisy data," in *ICML*, 2012.

[8] G. Patrini, A. Rozza, A. Krishna Menon, R. Nock, and L. Qu, "Making deep neural networks robust to label noise: A loss correction approach," in *CVPR*, 2017.

[9] A. Mosinska, P. Marquez-Neila, M. Kozinski, and P. Fua, "Beyond the Pixel-Wise Loss for Topology-Aware Delineation," in *CVPR*, 2018.

[10] L. Zhou, C. Zhang, and M. Wu, "D-LinkNet: LinkNet with Pretrained Encoder and Dilated Convolution for High Resolution Satellite Imagery Road Extraction," in *CVPR Workshops*, 2018.

[11] A. Buslaev, S. Seferbekov, V. Iglovikov, and A. Shvets, "Fully convolutional network for automatic road extraction from satellite imagery," in *CVPR Workshops*, 2018.

[12] J. Doshi, "Residual inception skip network for binary segmentation," in *CVPR Workshops*, 2018.

[13] G. Mattyus, W. Luo, and R. Urtasun, "DeepRoadMapper: Extracting Road Topology from Aerial Images," in *ICCV*, 2017.

[14] F. Bastani *et al.*, "RoadTracer: Automatic Extraction of Road Networks from Aerial Images," in *CVPR*, 2018.

[15] D. Costea, A. Marcu, M. Leordeanu, and E. Slusanschi, "Creating Roadmaps in Aerial Images with Generative Adversarial Networks and Smoothing-Based Optimization," in *ICCV Workshops*, 2017.

[16] S. Wang *et al.*, "Torontocity: Seeing the world with a million eyes," in *CVPR*, 2016.

[17] D. Eigen and R. Fergus, "Predicting Depth, Surface Normals and Semantic Labels with a Common Multi-scale Convolutional Architecture," in *ICCV*, 2015.

[18] C. Henry, S. M. Azimi, and N. Merkle, "Road segmentation in sar satellite images with deep fully convolutional neural networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 12, pp. 1867–1871, 2018.

[19] C. Wiedemann, C. Heipke, H. Mayer, and O. Jamet, "Empirical Evaluation Of Automatically Extracted Road Axes," in *Empirical Evaluation Techniques in Computer Vision*, 1998, pp. 172–187.

[20] T. Panboonyuen, K. Jitkajornwanich, S. Lawawirojwong, P. Srestasathiern, and P. Vateekul, "Road Segmentation of Remotely-Sensed Images Using Deep Convolutional Neural Networks with Landscape Metrics and Conditional Random Fields," *Remote Sens.*, vol. 9, no. 7, 2017.

[21] X. Gao *et al.*, "An end-to-end neural network for road extraction from remote sensing imagery by multiple feature pyramid network," *IEEE Access*, vol. 6, pp. 39 401–39 414, 2018.

[22] C. Ventura, J. Pont-Tuset, S. Caelles, K.-K. Maninis, and L. Van Gool, "Iterative deep learning for road topology extraction." 2018.

[23] O. Filin, A. Zapara, and S. Panchenko, "Road detection with eosresunet and post vectorizing algorithm," in *CVPR Workshops*, 2018.

[24] S. Jégou, M. Drozdzal, D. Vázquez, A. Romero, and Y. Bengio, "The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation," *arXiv:1611.09326*, 2016.

[25] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *CVPR*, 2016.

[26] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," in *CVPR*, 2017.

[27] B. Frenay and M. Verleysen, "Classification in the presence of label noise: A survey," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 5, pp. 845–869, 2014.

[28] B. v. Rooyen, A. K. Menon, and R. C. Williamson, "Learning with symmetric label noise: The importance of being unhinged," in *NeurIPS*, 2015.

[29] A. Ghosh, N. Manwani, and P. Sastry, "Making risk minimization tolerant to label noise," *Neurocomput.*, vol. 160, no. C, pp. 93–107, 2015.

[30] H. Masnadi-Shirazi and N. Vasconcelos, "On the design of loss functions for classification: Theory, robustness to outliers, and savageboost," in *NeurIPS*, 2008.

[31] S. Reed *et al.*, "Training Deep Neural Networks on Noisy Labels with Bootstrapping," *arXiv:1412.6596*, 2014.

[32] J. P. Brooks, "Support vector machines with the ramp loss and the hard margin loss," *Oper. Res.*, vol. 59, no. 2, pp. 467–479, 2011.

[33] F. Milletari, N. Navab, and S. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," *arXiv:1606.04797*, 2016.

[34] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *ICCV*, 2015.

[35] Y. Wei, Z. Wang, and M. Xu, "Road structure refined cnn for road extraction in aerial image," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 5, pp. 709–713, 2017.

[36] A. Constantin, J. Ding, and Y. Lee, "Accurate road detection from satellite images using modified u-net," in *IEEE Asia Pacific Conference on Circuits and Systems*, 2018, pp. 423–426.

[37] Z. Zhang and Y. Wang, "Jointnet: A common neural network for road and building extraction," *Remote Sens.*, vol. 11, no. 6, 2019.

[38] M. Yuan, Z. Liu, and F. Wang, "Using the wide-range attention u-net for road segmentation," *Remote Sens. Lett.*, vol. 10, no. 5, pp. 506–515, 2019.

[39] L. Gao, W. Song, J. Dai, and Y. Chen, "Road extraction from high-resolution remote sensing imagery using refined deep residual convolutional neural network," *Remote Sens.*, vol. 11, no. 5, 2019.

[40] T. Sun, Z. Chen, W. Yang, and Y. Wang, "Stacked u-nets with multi-output for road extraction," in *CVPR Workshops*, 2018.

[41] D. Costea, A. Marcu, E. Slusanschi, and M. Leordeanu, "Roadmap generation using a multi-stage ensemble of deep neural networks with smoothing-based optimization," in *CVPR Workshops*, 2018.

[42] S. Aich, W. van der Kamp, and I. Stavness, "Semantic binary segmentation using convolutional networks without decoders," in *CVPR Workshops*, 2018.

[43] Q. Zou, Y. Cao, Q. Li, Q. Mao, and S. Wang, "Cracktree: Automatic crack detection from pavement images," *Pattern Recognition Letters*, vol. 33, no. 3, pp. 227 – 238, 2012.

[44] I. Arganda-Carreras *et al.*, "Crowdsourcing the creation of image segmentation algorithms for connectomics," *Front Neuroanat.*, vol. 9, p. 142, 2015.