IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 13 May 2024, accepted 22 May 2024, date of publication 27 May 2024, date of current version 3 June 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3405559

RESEARCH ARTICLE

An Enhanced Loss Function for Semantic Road Segmentation in Remote Sensing Images

LORIS NANNI [1], **SHERYL BRAHNAM** [2], **AND ANDREA LOREGGIA** [3]

[1] Department of Information Engineering (DEI), University of Padova, 35131 Padua, Italy
[2] Information Technology and Cybersecurity, Missouri State University, Springfield, MO 65804, USA
[3] Department of Information Engineering (DII), University of Brescia, 25123 Brescia, Italy

Corresponding author: Loris Nanni (loris.nanni@unipd.it)

ABSTRACT The analysis of road continuity in satellite images is a complex challenge. This is due to the difficulty in identifying the directional vector of road sections, especially when the satellite view of roads is obstructed by trees or other structures. Today, most research focuses on optimizing the deep learning network topology, however, the accuracy of segmentation is affected by the loss function used in training; currently, little research has been published on ad-hoc loss functions for road segmentation. To solve this problem, we proposed loss functions based on topological pixel analysis, in which more weight is given to problematic pixels representing non-real road breaks. We report the results of different tests, obtaining state-of-the-art performance among convolution neural network-based approaches. For instance, on the Massachusetts Roads dataset, our method achieved a Dice score of 75.34% and an IoU of 60.44%, compared to the best baseline scores of 74.64% and 59.51% achieved by GapLoss. Similarly, on the DeepGlobe Roads dataset, our method obtained a Dice score of 79.78% and an IoU of 66.36%, outperforming the best baseline scores of 78.62% and 64.47% by GapLoss. Both the code and information for replicating our experiments are available at https://github.com/LorisNanni/An-Enhanced-Loss-Function-for-Semantic-Road-Segmentation-in-Remote-Sensing-Images, so as to enable future reliable comparisons.

INDEX TERMS Convolutional neural networks, road segmentation, optimization, ensemble.

I. INTRODUCTION

Roads and highways are fundamental to various human activities and industries, encapsulating aspects like traffic control, electronic cartography, urban design, and more. Within the discipline of geospatial analysis, roads, and highways are principal components of geographical information. The advent of advanced remote sensing technologies has led to significant enhancements in the geospatial analysis of these roads [57].

One of the reasons for road segmentation from aerial images stems from the unprecedented surge in road construction, with projections indicating a staggering 25 million kilometers of new paved roads by mid-century compared to 2010 levels [16], [35]. The great majority of this construction activity is happening in developing countries, particularly in tropical and subtropical regions known for their rich biodiversity. This surge in road building, often poorly regulated, is opening up previously remote natural areas, leading to a sharp increase in environmental disruption caused by activities like logging, mining, and land clearing.

Planning and zoning for road development have historically been insufficient in remote rural areas, wilderness frontiers, and semi-forested regions, where road construction tends to be chaotic and environmentally damaging. Many (legal and illegal) roads in these areas remain unmapped. Consequently, studies in regions like the Brazilian Amazon and the Asia-Pacific consistently reveal far more road length than officially reported, highlighting the challenges environmental governance and conservation efforts face due to incomplete and outdated information on road development [50]. Enhancing the capacity for automatic road segmentation from remote sensing imagery could play a crucial role in addressing these challenges.

Traditionally, the extraction of road and highway details from remote sensing imagery has been performed manually, but manual approaches are labor-intensive and have been shown to be prone to human biases [33], [55]. As a result, automated methods for effectively extracting road information from remote sensing imagery is a highly sought

The associate editor coordinating the review of this manuscript and approving it for publication was Yanli Xu.

after technology [19], [62], [63]. A plethora of research using remote sensing images has led to the development of numerous methodologies with varying degrees of accuracy. During the early development of automated road extraction techniques, the primary focus was on utilizing the spectral features of remote sensing imagery. This was often accomplished via a morphological algorithm, followed by the selection of an appropriate threshold for road segmentation [26], [31]. The advent of deep learning (DL) technology has significantly disrupted traditional research methodologies. Numerous deep learning based techniques have been suggested to enhance the efficiency of road extraction from remote sensing imagery, including methods like Deep Road Mapper [32], Topology Loss [39], Improve Road Connectivity [5]. Among them, several prominent convolutional neural networks follow the encoder-decoder architecture, exemplified by LinkNet [9] proposed by Chaurasia and Culurciello. Leveraging LinkNet's structure and dilation convolution, Zhou et al. introduced D-LinkNet [71], which emerged as the winning approach in the DeepGlobe 2018 Road Extraction Challenge [14]. These studies serve as significant reference points for road extraction from remote sensing imagery, actively driving advancements in the convergence of computer vision and remote sensing. Given the recent advancements in road extraction methods, different approaches are emerging. For instance, Bastani and Madden [4] proposed shifting focus towards the practical task of map updating, which involves modifying existing maps by adding, removing, and relocating roads while preserving the accuracy of unchanged areas. They introduced MUNO21, a dataset specifically designed for the map update task. While with Topo-boundary [59], Xu et al. proposed another dataset for offline topological road-boundary detection and they also designed a new entropy-based metric for connectivity evaluation.

Very recently, a new model called MSMDFF-Net [56] has been introduced to enhance model generalization across various resolutions and reduce road extraction fragmentation.

It is worth noting that the task analyzed in our work is consistently different from visual scene understanding commonly used in self-driving cars and augmented reality. Self-driving cars, for instance, perform an online visual scene understanding using on-vehicle cameras/Lidars. In our work, we perform offline detection using aerial images. Moreover, in visual scene understanding, such as road scenes, images are from the street level and comprehend roads, vehicles, possibly people, and objects in the background, such as sky and buildings. These tasks are addressed by designing and developing specific deep learning approaches trained on road scene datasets (see for instance [49], [66]).

The remainder of the paper is structured as follows: Section II reports the literature focused on road segmentation; Section III describes the loss functions tested/proposed in this paper; Section IV describes the experimental environments, including the datasets used for experiments, the testing protocols, and the performance indicators; moreover, we suggest and discuss a set of experiments to evaluate our suggested approach comparing it with the literature. We also report the results of an exploratory experiment on a different domain to test whether our approach can generalize to different yet similar tasks. Section V concludes this work and provides some further research on this topic.

The contributions of this article are emphasized as follows:

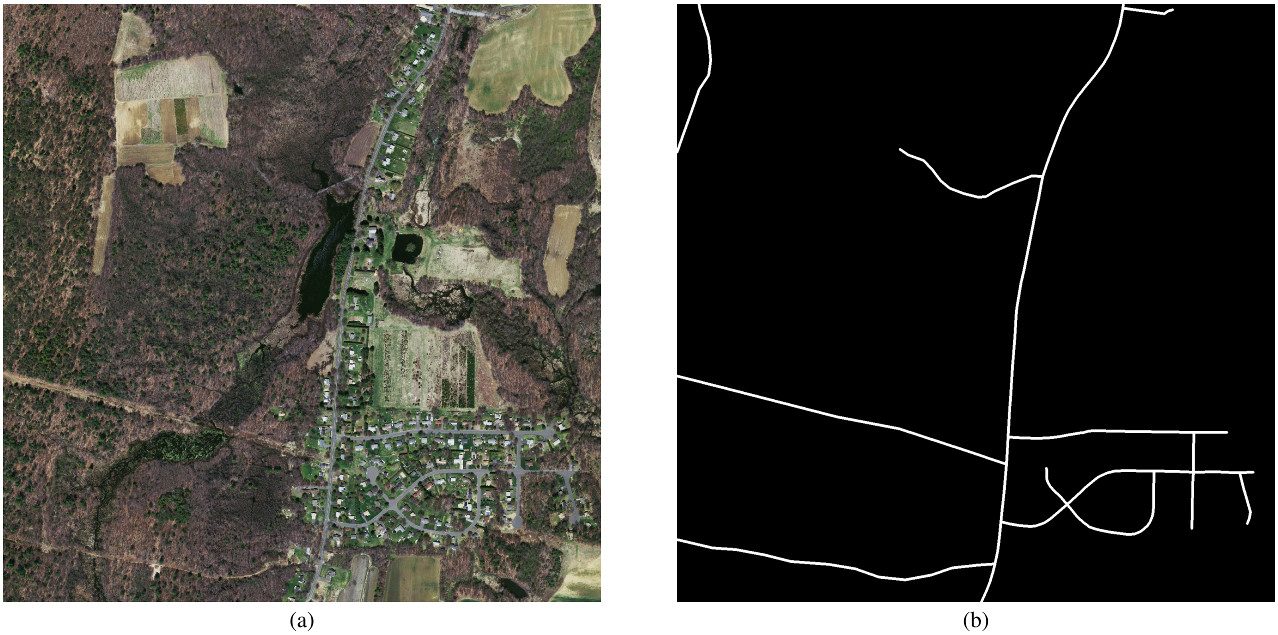
i) We show that in different datasets the best performance is achieved by a different loss function, and the fusion of them achieves SOTA (among CNN-based methods) performance, for reproducing our tests the code of the proposed approach is available on https://github.com/LorisNanni/An-Enhanced-Loss-Function-for-Semantic-Road-Segmentation-in-Remote-Sensing-Images.

ii) in the literature different protocols are used to compare the various methods, we propose replicable comparisons, specifying which images are used for training and which for test sets; also reporting cross-domain tests, which are essential to validate network performance.

II. RELATED WORK

In this section, we briefly introduce the literature focused on road segmentation.

In [70], authors were among the first to apply DL to road segmentation. These researchers utilized Fully Convolutional Networks (FCN) to extract roads and buildings in aerial imagery, attaining reliable outcomes on the Massachusetts Roads Dataset [38]. An example of an aerial image and the mask that detects the roads that are present in the image is reported in Figure 1. [45] enhanced road segmentation outcomes by applying the SegNet network with an Exponential Linear Unit (ELU) activation function. Reference [57] improved road recognition accuracy by increasing the pixel weight near labels with a cross-entropy function. In an attempt to capture the topology of roads in aerial images, [32] employed ResNet as an encoder and a full deconvolutional network as a decoder. Similarly, [17] proposed a multiple-feature pyramid network resembling the RSRCNN, and Zhang et al. [68] introduced a deep residual UNet for road extraction. Despite these advances, many serious challenges remain in using DL for road segmentation. A good example involves widely spanning and narrow roads, which often results in subpar results with DL models. For instance, [3] have highlighted the high noise output of convolutional neural networks (CNNs) which they addressed with an iteratively applied post-processing measure called RoadTracer. Reference [71] leveraged an encoder-decoder structure and proposed D-LinkNet that makes use of a dilated convolution to amplify the overall receptive field of the model without compromising the resolution of the feature map. Reference [52] offered a point-based iterative graph exploration scheme to boost the model's perception of global information, thereby enhancing overall segmentation results. Reference [54] observed that conventional CNN-based methods fail to preserve global road connectivity because these



(a)

(b)

FIGURE 1. **Example of (a) an aerial image and (b) the corresponding mask from the Massachusetts Roads Dataset [37].**

networks typically utilize pixel-wise losses for optimization. Reference [34] developed a strip convolution module to capture long-range contextual information from various directions, which significantly improved the connectivity of segmentation results, as demonstrated by this method's high performance on the DeepGlobe dataset [34]. Reference [8] pointed out that the network needs to extract rich long-range dependencies for road segmentation in remote sensing images. Their solution was CSANet, a CNN-based method that utilizes a cross-scale axial attention mechanism to achieve superior results [8]. Further advancements include the work of [51], who offered a lightweight semantic segmentation model incorporating attention mechanisms, and the DeepWindow model developed by [27]. The latter, guided by a CNN-based decision function, uses a sliding window to directly track the road network from the images, bypassing the need for prior road segmentation. Recently a new benchmark dataset of optical remote sensing images has been published, providing the corresponding high-quality labels for road extraction.

Recently, SEG-Road [53] combines transformer and convolutional neural network structures along with a novel pixel connectivity structure, to enhance road segmentation by addressing issues related to road complexity, resulting in state-of-the-art performance on the DeepGlobe and Massachusetts datasets, thereby contributing to sustainable urban development. A combination of different models is also the approach that proposes conditional Generative Adversarial Network (GAN) architecture along with Attention U-Net and PatchGAN to improve road segmentation from aerial images [18]. Reference [1] introduces DeepRoadNet, a robust model offering a more intricate approach, which employs a pre-trained EfficientNetB7 architecture in the encoder and utilizes residual blocks in the decoder, mirroring the segmentation process of U-Net. Another new approach consists of UnetEdge, a transfer learning-based framework designed to enhance road feature extraction from high-resolution remote sensing imagery by effectively propagating topological information through the network [15].

A. LOSS FUNCTIONS

Seldom emphasized in studies on automatic road extraction is the significance of the selected loss function [64]. Only recently, a survey conducts a comprehensive analysis of 12 widely-used loss functions for road segmentation in remote sensing imagery, revealing significant performance differences in terms of overall model performance, precision, and recall, with region-based loss functions generally outperforming distribution-based ones, thus providing valuable insights for selecting appropriate loss functions in road segmentation tasks [58]. As can be noticed from the survey, standard loss functions are usually adopted in the realm of road segmentation and there exist a small number of approaches that try to design specific loss functions for this domain. The cross-entropy function has been predominantly utilized for semantic segmentations of images [61]. Other loss functions used in segmentation generally have been designed specifically for unbalanced datasets. These include focal loss [28], dice loss [36], Tversky loss [46], log-cosh dice loss [22], and generalized dice loss [13]. Despite their use, these loss functions have not been successful in ensuring accurate road continuity as they were not designed with linear object recognition in mind [64]. Reference [39] were one of the first to propose a topology-aware loss function. A problem they recognized with DL had to do with pixel-wise losses, such as binary cross-entropy, which negatively impact the topological structure of road distribution in road segmentation tasks. These researchers proposed a linear structure to enhance higher-order topological features that improved overall model accuracy. Rather than computing and comparing topology, their method leveraged selected filters from a pretrained VGG19 network to formulate the loss function. These filters, which favored elongated shapes, mitigated the continuous road issues. Despite its effectiveness, this methodology has difficulty adapting to more complex environments containing a range of arbitrary shapes. More importantly, the loss function was created only for the proposed segmentation network of Mosinska et al., thereby limiting its adaptability to other networks. Similarly, [5] approached roads as directional objects, employing network sharing to concurrently execute the tasks of segmentation and direction learning, thus improving the precision of road continuity. However, this methodology necessitated the segmentation of the label into line strings, making the preprocessing computations excessively laborious. As with Mosinska et al., the loss function they utilized was also specific to the network. Reference [21] introduced a topological loss function that was more generalizable. Due to the Betti number being a discrete value and not directly derivable, they used the persistent topology theory to construct the loss function. Although the method can be combined with any segmentation network, it demands significant computational resources. Additionally, at the onset of training, the network may not have sufficient prediction results for precise topology assessment, making this method only suitable for fine-tuning previously obtained results. While it enhanced road continuity, it did not improve road segmentation accuracy. Finally, [64] proposed a novel attention loss function they call Gaploss that was tested across three road segmentation datasets. Gaploss obtained similar results to the best methods, if not state-of-the-art. The Gaploss process begins with a DL network that generates a binary prediction mask. This process is followed by extracting a vector skeleton from the mask. Subsequently, eight neighboring pixels sharing the same value are computed for each pixel. A pixel is recognized as an endpoint if its value is 1. Next, based on the count of endpoints within a defined buffer range, every pixel in the prediction image receives a corresponding weight. The procedure culminates with the calculation of the final loss function value, which is determined as the weighted average of the cross-entropy across all pixels in the batch.

III. MATERIALS AND METHODS

In this section, we describe the different components of the proposed ensemble and we detail the loss functions tested/proposed in this study.

A. CONVOLUTIONAL NEURAL NETWORKS

Convolutional Neural Networks (CNNs) are a specific category of deep neural networks crafted for tasks like image classification, computer vision, and various applications such as face recognition, and object detection. They emulate the human brain's visual perception [25]. Using these models in ensembles has demonstrated advantages in terms of enhanced performance (see for instance, [12], [42], [43], [47]).

Convolution entails systematically moving a small filter (also referred to as a kernel) across the input data, commonly a two-dimensional grid of pixels in the context of images. The values of the filter (learnable weights) undergo element-wise multiplication with corresponding values in the input data, and the resultant products are summed to yield a singular value. This process iterates as the filter slides over the entire input, employing a specific stride, resulting in the creation of a new output matrix termed a feature map.

The convolution operation empowers the neural network to identify patterns or features within the input data, ranging from basic elements like edges or corners in the case of images. As the network undergoes training, it adapts the filter weights to capture progressively intricate and abstract features, encompassing textures, shapes, or components of objects. CNN architectures typically incorporate stacked convolutional layers, with each layer specializing in recognizing distinct levels of features. This hierarchical feature extraction endows CNNs with the capability to comprehend image content in a manner reminiscent of how the human visual system processes information [7].

CNNs are employed in many different computer vision tasks. For instance, in medical diagnostics, they are employed to improve the early detection of pathologies [41] or in different domains integrated with other methods to tackle the problem of face hallucination [23].

CNNs have been used to address the challenge of semantic gap reduction and information redundancy in multisource remote sensing image classification, effectively refining intrasource correlation features and generating nonredundant multisource representations [30].

In our experiments, we established the baseline performance by evaluating the results of the DeepLabV3+ model [10]. This model was trained end-to-end on each of the datasets used in our study.

For our experiments, we utilized a DeepLabV3+ model with ResNet101 as the backbone architecture. Instead of training the model from scratch, we initiated the training process using pre-trained weights on the Pascal VOC2012 segmentation dataset.[1] This dataset consists of RGB images of size 513 × 513, encompassing various categories such as airplanes, buses, cars, trains, persons, horses, and more.

B. TRAINING AND TEST PHASES

During the training phase, we employed specific hyperparameters. These included an initial learning rate of 0.01, training for a total of 15 epochs, a momentum value of 0.9, L2 regularization with a coefficient of 0.005, a learning rate (LR) drop period of 5 epochs, a learning rate drop factor of 0.2 (therefore 5 epochs with LR=0.01, 5 epochs

[1] https://github.com/matlab-deep-learning/pretrained-deeplabv3plus - Last access 18 December 2023.

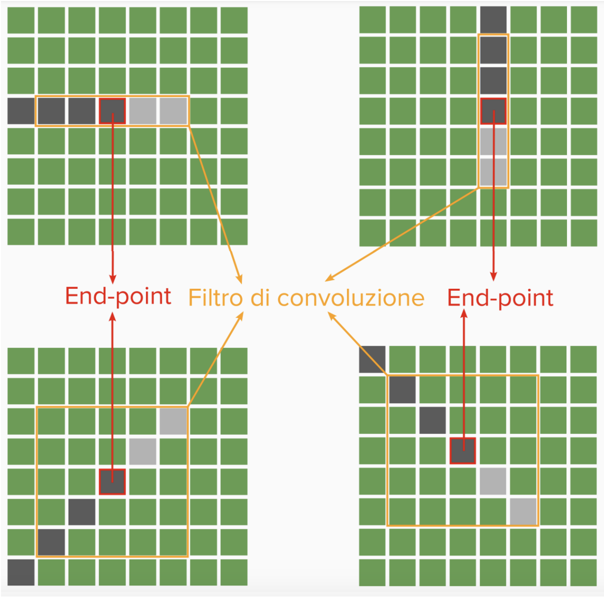


FIGURE 2. Graphical description of topological loss proposed in this work.

with LR=0.002 and 5 epochs with LR=0.0004), shuffling of training images at every epoch, and the use of the SGD (Stochastic Gradient Descent) optimizer.

The baseline performance provided by the DeepLabV3+ model, trained on each dataset, serves as a reference point for evaluating the effectiveness and enhancements achieved by our proposed methods.

C. TOPOLOGICAL LOSS

In this subsection, we detail the proposed loss, named Topological Loss (TL). This is a variant of the classic version of GapLoss. The standard version aims to connect interrupted road segments considering a neighborhood of 9 × 9 pixels (see line 12 of the pseudocode in [65]). Nevertheless, a challenge arises in this context: for numerous images, interruptions within certain segments may extend to larger sizes, ranging from dozens of pixels to varying dimensions depending on the specific case.

The main contribution of the proposed Topological Loss (TL) function lies in its departure from the traditional GapLoss approach, addressing the challenge of interruptions within road segments of varying sizes. Unlike GapLoss, which relies on proximity considerations between endpoints, TL intentionally disregards this proximity and instead leverages topological information within road networks for logical inference. By simplifying the directional trends of road segments and focusing on pixel classification as endpoints, TL identifies pixels that should logically represent a continuation of the road direction. Through a convolution-based weighting process, TL assigns greater weight to pixels indicative of road continuation, resulting in more continuous road segments with fewer interruptions.

Topological Loss intentionally disregards the proximity of two endpoints (extremities of the interruption), subsequently, it capitalizes on topological information within road networks for logical inference. While recognizing that this assertion may have some approximation, road segments typically exhibit fundamental directional trends, such as vertical, horizontal, increasing diagonal, or decreasing diagonal (among others), justifying this simplification. Moreover, given the often straight and lengthy nature of road segments, it is highly probable that their direction remains consistent. Leveraging this insight and employing the initial standard GapLoss process to obtain the binary schema, we endeavor to identify pixels classifiable as endpoints, commencing with straight segments.

For each of these endpoints, we attempt to infer the direction of the preceding road segment by observing whether the preceding pixels, categorized as road (depicted in dark gray in Figure 2), originate from one of the aforementioned four approximate directions. Subsequently, as highlighted in light gray in Figure 2, we consider the following pixels, which should logically represent a continuation of the road in that specific direction (we apply four different convolution filters for weighting each pixel).

Ultimately, the output of the Cross-Entropy function is multiplied (element-wise) by a weighted matrix obtained by a set of convolutions. By calculating the average of the value for all pixels, the final loss value is obtained.

With this approach, greater weight is given to pixels that should represent a continuation of the road direction, resulting in more continuous road segments with fewer interruptions.

Here's a brief description of the main components and steps involved in Topological Loss as reported in Algorithm 1:

- Softmax Prediction: Topological Loss starts with the softmax prediction of an image, often generated by a neural network. This prediction represents the likelihood of each pixel belonging to a certain class (e.g., road or background).
- Binarization (line 2): The softmax prediction is binarized using a threshold. Pixels with values above the threshold (in our experiments this is set to 0.5) are considered part of the segmented object (e.g., road), while those below the threshold are considered background. This binarization process converts the prediction into a binary mask.
- Skeletonization (line 3): The binary mask obtained in the previous step is subjected to a skeletonization process. Skeletonization reduces the binary mask to a thinner representation, typically consisting of lines or curves that trace the central axis of the segmented object (e.g., road). This helps in capturing the essential structure of the object.
- Weighted Loss (lines from 8 to 46): Topological Loss adopts a weighting mechanism to assign different weights to different pixels in the skeletonized representation. Pixels that are considered more critical for maintaining the continuity of the object are assigned

Algorithm 1 Pseudocode for TopologicalLoss

```
Require: $L$, Prediction of the image obtained from the network
Ensure: Loss function value
 1: $Cr \leftarrow$ cross-entropy of $L$
 2: $A \leftarrow$ binarization (threshold 0.5) of $L$
 3: $B \leftarrow$ Skeletonization of $A$
 4: $D \leftarrow$ zero matrix with the same size of $B$
 5: $E \leftarrow$ zero matrix with the same size of $B$
 6: $F \leftarrow$ zero matrix with the same size of $B$
 7: $G \leftarrow$ zero matrix with the same size of $B$
 8: HORIZONTAL DIRECTION
 9: $f = filter : [1; 1; 1; 1; 1]$;
10: $C = conv(B, f)$;
11: $C(C \sim= 2) = 0$;
12: $C(C == 2) = 1$;
13: $f = filter : [10; 10; 10; 10; 10]$; filter
14: $D = conv(C, f)$;
15: $D(D > 10) = 10$;
16: $D(D == 0) = 1$;
17: VERTICAL DIRECTION
18: $f = [1; 1; 1; 1; 1]'$;
19: $C = conv(B, f)$
20: $C(C \sim= 2) = 0$;
21: $C(C == 2) = 1$;
22: $f = [10; 10; 10; 10; 10]$;
23: $E = conv(C, f)$;
24: $E(E > 10) = 10$; lower the high values to 10
25: $E(E == 0) = 1$; set the 0s to 1
26: RIGHT DIAGONAL DIRECTION
27: $f = diag([1; 1; 1; 1; 1])$;
28: $C = conv(B, f)$;
29: $C(C \sim= 2) = 0$;
30: $C(C == 2) = 1$;
31: $f = diag([10; 10; 10; 10; 10])$
32: $F = conv(C, f)$;
33: $F(F > 10) = 10$;
34: $F(F == 0) = 1$;
35: LEFT DIAGONAL DIRECTION
36: $f = flip(diag([1; 1; 1; 1; 1]))$;
37: $C = conv(B, f)$
38: $C(C \sim= 2) = 0$
39: $C(C == 2) = 1$
40: $f = flip(diag([10; 10; 10; 10; 10]))$;
41: $G = conv(C, f)$;
42: $G(G > 10) = 10$
43: $G(G == 0) = 1$
44: $W = D + E + F + G$
45: $W(W >= 10) = 10$;
46: $J = W. \times Cr$;
47: return average($J$);
```

higher weights. This is typically done by analyzing the local context or connectivity of pixels in the skeleton.

TABLE 1. Description of the datasets used in this study.

Short Name	#Classes	#Samples	Format	Image Size	Ref
MA	2	1171	TIFF	1500x1500	[37]
DG	2	8570	JPG	1024x1024	[14]
TO	2	14374	RGB	512x512	[24]
EA	2	200	RGB	1920x886	[50]
BO	2	1	RGB	23104x23552	[29]
CO	2	62	RGB	512x512	[40]

- Loss Computation (lines 46 and 47): The final Topological Loss is computed by combining, by element-wise multiplication, the weighted matrix obtained in the above step with the original cross-entropy. The loss encourages the network to produce predictions that preserve the connectivity and coherence of the segmentation map.

In the Algorithm 1, $conv(X, y)$ refers to the convolutional operation between a matrix X and a filter y; $average(J)$ computes the average value of all the pixels.

IV. EMPIRICAL EVALUATION

We performed a thorough empirical evaluation to assess the performance of our proposal. In the following subsections, we describe the results of our analysis.

A. DATASETS AND TESTING PROTOCOLS

We have used different datasets for assessing the performance of the loss functions. For all the datasets, the sub-windows without foreground pixels are not used for training the net. Table 1 reports a brief description for each dataset: a short name, the number of classes and samples, the testing protocol, and the original reference.

The following datasets have been adopted in the empirical evaluation:

i) Massachusetts Roads Dataset (MA) [37], it is openly accessible.[2] This dataset was built using images of Massachusetts, USA. It consists of 1108 training images, 14 validation images, and 49 test images, along with their corresponding label images. For this particular experiment, no separate validation was conducted. Instead, the validation images and test images were combined to form the test dataset, resulting in a total of 63 images, as in [65]. Each image is split in non-overlapped sub-windows of size 500 × 500, overall, a total of 9 data samples for each image were obtained from this process.

ii) DeepGlobe Roads Dataset (DG), this dataset has been used in the 2018 DeepGlobe Road Extraction Challenge. These images were obtained from remote sensing and were captured in various locations including Thailand, India, and Indonesia. They encompassed diverse settings such as cities, villages, suburbs, seashores, and tropical rainforests.

Unfortunately, only the training dataset (of the original contest) contained available road masks, each image

[2] https://www.kaggle.com/datasets/balraj98/massachusetts-roads-dataset - Last access 18 December 2023.

(size 1024 × 1024) is split into non-overlapped sub-windows of size 512 × 512. We performed two different protocols:

- the one used in [65], of the original training images, the last 2000 were designated as the test dataset, resulting in a training dataset of 4226 images; this split is used in the Tables 2 and 3.
- (b) we apply the subdivision used in [34] in table 5, 4,696 images are used for training and 1,530 for testing, split is openly accessible.[3]

iii) Aerial Dataset (TO), this dataset [24] consisted of aerial imagery from six different regions: Berlin, Chicago, Paris, Potsdam, Zurich, and Tokyo. The images in the dataset were obtained from Google Maps and GroundTruth for Berlin, Chicago, Paris, Potsdam, and Zurich. We use the Tokyo image as a test set, feeding the networks using data from Chicago, Paris, and Zurich. Chicago contributed 457 aerial images with a resolution of 3328 × 2560 pixels and corresponding annotations. Paris included 625 aerial images with a resolution of 3328 × 3072 pixels and corresponding annotations. Zurich contained 364 aerial images with a resolution of 3072 × 2816 pixels and corresponding annotations. Lastly, Tokyo had a single aerial image with a resolution of 2500 × 2500 pixels, along with corresponding annotations. As in other datasets, the images are split to a set of subwindows of size of 512 × 512 pixels.

iv) Equatorial Asia (EA) [50], this dataset consists of 8904 satellite image tiles and corresponding road features across Equatorial Asia, including Indonesia, Malaysia, and Papua New Guinea. It was created for training AI models to identify road features in rural/remote tropical regions using true-color satellite imagery. The main dataset is derived from 200 input satellite images, each acquired at a resolution of 1920 × 886 pixels, although the actual image resolution appears coarser, estimated to be around 5 meters per pixel. These images were observed using the Elvis Elevation and Depth spatial data portal, akin to Google Earth. The 200 images represent screenshots of high-resolution imagery, with each image covering forest-agricultural mosaics or intact forest landscapes with limited human intervention. The images are already divided between training and test sets, but from visual inspection many masks are not accurate, affecting performance; we used only 100 images of the test set extracting new masks of these images, but using the original training set for feeding the nets.

v) Corniolo (CO) [40], this dataset[4] contains 62 images of size 512 × 512 extracted from 8 images extracted using Google Maps around the village of Corniolo (Appennino tosco-romagnolo, Italy).

TABLE 2. Comparison of the loss functions in terms of the different metrics (in %). Performance of one network for each loss function. Best performance in bold.

	Loss function	Prec.	Recall	Acc.	Dice	IoU
MA	DI	71.96	**67.85**	97.23	74.10	58.92
	CE	74.11	56.09	97.12	69.43	53.24
	GL	73.99	64.27	97.28	74.64	59.51
	TL	74.01	63.37	97.25	73.77	58.48
	GL+TL	75.07	60.17	97.35	74.86	59.88
	GL+TL+DI	**75.11**	65.92	**97.50**	**75.34**	**60.44**
	[65]	—	—	97.20	68.60	—
DG	DI	63.56	**62.67**	98.03	77.65	63.46
	CE	61.85	59.62	97.65	76.14	61.47
	GL	64.03	58.34	98.02	78.62	64.47
	TL	64.52	60.19	98.15	79.05	65.36
	GL+TL	65.34	59.77	98.12	79.57	66.08
	GL+TL+DI	**66.04**	61.23	**98.25**	**79.78**	**66.36**
	[65]	—	—	97.60	70.40	—
TO	DI	81.64	40.06	81.57	53.51	36.52
	CE	**84.02**	38.06	81.08	42.81	35.93
	GL	80.12	48.10	83.11	60.63	43.50
	TL	80.65	**49.91**	83.24	**61.46**	**44.36**
	GL+TL	81.26	48.87	**83.27**	61.39	44.29
	GL+TL+DI	83.86	47.01	83.03	60.00	42.90
EA	DI	59.52	59.49	96.65	75.30	56.85
	CE	58.29	58.11	96.14	73.89	56.04
	GL	61.42	58.36	96.84	76.09	57.74
	TL	60.36	58.55	97.03	77.76	59.33
	GL+TL	61.38	**60.77**	97.21	77.25	60.66
	GL+TL+DI	**62.99**	59.62	**97.82**	**78.80**	**62.23**

vi) Boston (BO) [29], [40], this dataset[5] contains large-scale images collected from Google Earth and accurately labeled for evaluation purposes. This dataset is collected and shared by the RSIDEA research group of Wuhan University, specifically designed for validating road detection tasks. The images in the dataset are 23104 × 23552 pixels with a resolution of 0.44 meters per pixel. As in other datasets, the image is split into a set of subwindows of size 512 × 512 pixels.

DG and MA are the most used datasets in literature for training models [48].

B. EXPERIMENTS

We employed standard accepted metrics, namely Precision, Recall, Dice score, and Intersection over Union (IoU), as performance indicators. In the following formulas, TP, TN, FP, and FN stand for true positives, true negatives, false positives, and false negatives, respectively. Here, *A* denotes the predicted mask, while *B* corresponds to the ground truth map.

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

$$Accuracy = \frac{TP + TN}{TP + FP + FP + FN} \tag{3}$$

$$F1_{Score} = Dice = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{4}$$

$$IoU = \frac{TP}{TP + FP + FN} \tag{5}$$

[3] https://github.com/mj129/CoANet/tree/main/data/spacenet

[4] https://zenodo.org/records/11107140

[5] https://zenodo.org/records/11107140

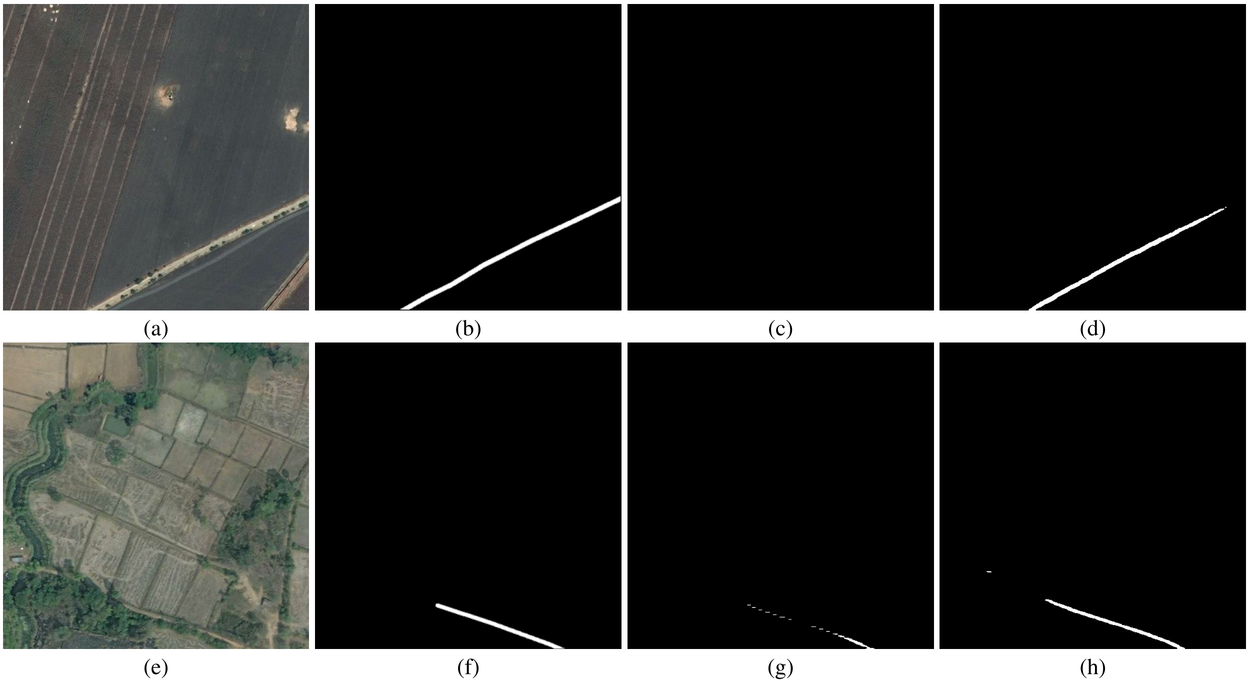


FIGURE 3. Examples of masks that are correctly computed by our proposal compared with CoaNet. (a) and (e) Input aerial images, (b) and (f) ground truths, (c) and (g) CoaNet predictions, (d) and (h) masks predicted by our proposal.

TABLE 3. Comparison of the loss functions in terms of the different metrics (in %). Performance of two networks, for each loss function, combined using sum rule. Best performance in bold.

	Loss function	Prec.	Recall	Acc.	Dice	IoU
MA	DI	72.39	**68.14**	97.42	74.52	59.35
	CE	74.55	56.30	97.19	69.73	53.57
	GL	74.53	64.77	97.56	75.08	60.17
	TL	74.48	63.95	97.50	74.28	59.19
	GL+TL	**75.54**	64.55	97.59	75.17	60.23
	GL+TL+DI	75.21	66.00	**97.60**	**75.53**	**60.69**
	[65]	—	—	97.20	68.60	—
DG	DI	64.17	63.13	98.12	78.24	64.26
	CE	61.92	59.74	97.72	76.36	61.76
	GL	64.99	59.23	98.25	79.16	65.51
	TL	64.91	60.69	98.30	79.49	65.96
	GL+TL	65.80	60.10	**98.33**	79.82	66.42
	GL+TL+DI	**66.20**	**61.40**	**98.33**	**79.93**	**66.58**
	[65]	—	—	97.60	70.40	—
TO	DI	**84.54**	43.80	82.38	57.70	40.55
	CE	83.15	34.07	80.02	48.33	31.87
	GL	78.84	46.60	81.92	58.58	41.42
	TL	81.45	**50.84**	**83.34**	**62.60**	**45.56**
	GL+TL	80.49	48.06	82.55	60.18	43.04
	GL+TL+DI	84.38	47.22	83.12	60.60	43.40
EA	DI	61.15	60.05	96.84	76.89	58.21
	CE	60.11	59.21	96.26	75.81	57.56
	GL	61.68	60.28	97.08	77.41	59.23
	TL	62.19	60.48	97.16	77.84	60.12
	GL+TL	62.65	61.09	97.56	78.95	61.98
	GL+TL+DI	**63.19**	**61.57**	**98.05**	**80.67**	**62.58**

In Tables 2 and 3 ,we report the performance of the tested loss functions. In particular, Table 2 reports the performance adopting a single network for each loss function, while Table 3 reports the performance employing two networks for each loss function. When networks are more than one for each loss function or when we combine more loss functions, then the output of the networks is aggregated using the sum rule. GL and TL are compared with the largely used Dice loss (DI) and cross-entropy loss (CE). As can be seen from Tables 2 and 3 and often happens in deep/machine learning, there is no specific method that achieves the best performance in all data sets. Among the stand-alone loss functions, in terms of Dice score and IoU, the highest average performance is obtained by the proposed TL, which outperforms the other functions in 3 out of 4 datasets, providing a piece of evidence of the improvements that this approach can give. Moreover, a simple combination of these loss functions provides an even better performance. From Tables 2 and 3, we can see that the best average performance is obtained by GL+TL+DI in all datasets but TO. GL+TL+DI is the sum rule between networks trained using GL, the nets trained using TL, and nets trained using DI. In particular, considering all the datasets, Dice, and IoU indicators, GL+TL+DI combination outperforms each stand-alone approach with a p-value of 0.05 (this is calculated by Colton approach [20]). In both MA and DG, the proposed ensemble strongly outperforms the segmentation approach proposed in [65], where GL was proposed. Notice that, to reduce the computation time, in the above tests we did not use a data augmentation approach for enlarging the size of the training set.

C. COMPARISON WITH SOTA

In this section, we describe the experiments that compare the proposed approach with the current state-of-the-art (SOTA). In particular, we utilized a DeepLabV3+ model with ResNet101 as the backbone architecture, combining data augmentation (where possible) with the proposed loss function. For MA dataset, we combine a data augmentation method with our approach, a standard solution adopted in the literature [6]. For training our networks, the following data augmentation approach has been applied: horizontal flip, vertical flip, 90° rotation, two different approaches for modifying the brightness, adding speckle noise, modifying contrast and blur, and modifying shadows (all the cited data augmentation approaches are as in [44]) and the contrast enhancement proposed in [67]. By augmenting the training set with these operations, we created additional variations of the training samples, ultimately improving the robustness and adaptability of the trained network. To reduce the computation time, we use only the first 10 epochs for training DeepLabV3+. In the DG dataset, due to the training set size, the data augmentation step is not performed.

Notice that typically, the threshold for binarization is set to a logit value of 0.5. In our experiments, we also tested a different threshold: a logit value of 0.375. In the following tables, the notation "x(y)" denotes that we use the method x with a threshold set to y, a pixel is classified as 'road' if its score is higher than the threshold. The tests in this section clearly show that a threshold of 0.375 allows for better performance in almost all tests, and it is our suggested threshold. To avoid overfitting we did not optimize that hyperparameter, but only increased the threshold since in many images 'road' pixels had borderline scores considering a threshold of 0.5.

In Table 4, we report the performance on the MA dataset. The performance of the method GI+TL+DI is compared with literature using IoU (since that performance indicator is largely used in the literature), in Table 5 our approach is compared with literature on the DG dataset. Both tables include only methods that share comparable dataset splitting strategies with ours.

Besides attention layer-based methods such as CoANet and SegRoad, the proposed method achieves a good performance. We achieve such results by modifying the loss function and using standard DeepLabV3+, while the works in the table propose ad hoc network topologies for road segmentation. The tested loss functions are not network topology-related, so they can also be used in other network topologies such as CoANet (code available on GitHub) and SegRoad (code not available on GitHub). Moreover, in Table 5, we have combined by sum rule our approach with CoANet (pre-trained net is available on GitHub), the method named CoaNet-GI+TL+DI() is given by sum rule among GI, TL, DI, and CoANet; since GI, TL, and DI are built by two nets, we assign a weight of two to the output of CoANet. CoaNet-GI+TL+DI obtains SOTA in the DG dataset. It is important to notice that in Table 5, we report two CoANet approaches: the first (named CoaNet [34]) shows the IoU reported in the original paper; the latter (named CoANet [here]) reports the IoU obtained by employing the pre-trained model available online[6] on the DG test images. The performance difference between CoANet and CoANet [here] could be due to several reasons, for example, we did not perform any image processing on the test images.

TABLE 4. Performance on MA dataset (in %). The performance of the compared method is from [53]. Best performance in bold.

# Approach	IoU
GI+TL+DI(0.5)	63.33
GI+TL+DI(0.375)	63.65
D-LinkNet [71]	61.45
DeepRoadMapper [32]	59.66
RoadCNN [2]	62.54
PSPNet [69]	58.91
LinkNet34 [60]	61.35
CoANet [34]	61.67
Seg-Road-s [53]	64.78
Seg-Road-m [53]	66.29
Seg-Road-l [53]	**68.38**

TABLE 5. Performance on DG dataset (in %). The performance of the compared method is from [34] but Seg-Road is obtained from [53]. Best performance in bold.

# Approach	IoU
GI+TL+DI(0.5)	66.27
GI+TL+DI(0.375)	66.49
CoaNet-GI+TL+DI(0.375)	**68.41**
D-LinkNet [71]	63.26
DeepRoadMapper [32]	63.98
RoadCNN [2]	65.40
LinkNet34 [60]	63.98
CoANet [34]	68.37
CoANet [here]	66.50
Seg-Road-s [53]	61.14
Seg-Road-m [53]	64.31
Seg-Road-l [53]	67.20

It is worth noticing that the proposed ensemble CoaNet-GI+TL+DI() outperforms CoatNet [here] with a p-value 0.05, which is calculated by Colton approach [20]. In Figure 3, we report some examples as a comparison between the output of our proposal and the output of CoaNet. As can be noticed, the output of CoaNet (i.e., Figure 3c and Figure 3g) could be less precise or incomplete compared to the prediction of our method that employs the proposed topological loss.

D. CROSS-DOMAIN ANALYSIS

In this section, we address the need for a comprehensive experimental evaluation of our proposed loss function by performing a cross-domain analysis with state-of-the-art road extraction methods. In the cross-domain approach, we leverage trained models and evaluate their performance on diverse datasets beyond their original training domain. This methodology allows us to assess the generalizability and adaptability of the models across varied data sources and scenarios, aiming to substantiate and showcase the effectiveness of our approach in comparison to existing methodologies. In Table 6 , we report the performance obtained by training the models on the DG dataset and performing the test on MA, CO, and BO datasets. From these results, we can notice that the proposed loss function is performing well in all the datasets: TL outperforms other single loss functions, namely Dice and GapLoss. Moreover, the combination of these loss functions provides the best performance in CO and it remains pretty stable in the other two. CoaNet performs nicely on MA but its performance drops on CO and BO, showing an unstable behaviour in

[6] https://github.com/mj129/CoANet/ - Last visited May 6, 2024.

TABLE 6. Comparison of the loss functions in terms of the different metrics (in %). Each model is trained on the DeepGlobe dataset and tested on different datasets reported in the table. Best performance in bold.

Dataset	Models	Dice	IoU
MA	DI(0.375)	58.27	41.12
	GL(0.375)	58.09	40.94
	TL(0.375)	59.26	42.11
	DI+GL+TL(0.5)	59.52	42.37
	DI+GL+TL(0.375)	59.21	42.05
	CoaNet	**63.54**	**46.56**
	CoaNet-GI+TL+DI(0.375)	61.34	44.26
	MSMDFF-Net [56]	—	—
CO	DI(0.375)	60.50	43.37
	GL(0.375)	60.54	43.41
	TL(0.375)	61.57	44.48
	DI+GL+TL(0.5)	59.72	42.57
	DI+GL+TL(0.375)	**61.95**	**44.88**
	CoaNet	23.73	13.46
	CoaNet-GI+TL+DI(0.375)	60.55	43.43
	MSMDFF-Net [56]	—	—
BO	DI(0.375)	76.19	61.54
	GL(0.375)	76.30	61.68
	TL(0.375)	76.07	61.38
	DI+GL+TL(0.5)	74.29	59.01
	DI+GL+TL(0.375)	77.27	62.96
	CoaNet	41.55	26.22
	CoaNet-GI+TL+DI(0.375)	76.49	61.92
	MSMDFF-Net [56]	**78.12**	**64.10**

cross-domain applications. We also launched a test by training on the MA dataset and adopting BO as the test set, [56] this test provides an IoU of 8.86, DI+GL+TL of 14.55, and the suggested method DI+GL+TL(0.375) of 21.77. From these experiments, we can infer that our method could be more stable in this difficult test than CoaNet.

E. DIFFERENT APPLICATION

To assess the versatility of our approach beyond road segmentation, we conducted experiments in a different domain, specifically focusing on wire segmentation. This distinct task shares similarities with road segmentation, as both involve the segmentation of linear features in images, following specific paths within photographs. This experiment serves as an exploratory step toward assessing the broader applicability of our methodology. Wires and powerlines often detract from the visual appeal of photographs, necessitating precise segmentation and removal, which is a labor-intensive process, especially on high-resolution images. This preliminary investigation paves the way for future research aimed at expanding the scope of our approach to diverse tasks and applications beyond traditional road segmentation. We employed the wire dataset described in [11]. A dataset of 420 copyright-free test images along with annotations for public use and evaluation is available, we split the datasets in 2-fold (50% training and 50% test sets). As in other datasets, the images are split into a set of subwindows of the size of 512 × 512 pixels. Table 7 reports the performance of the tested loss functions (for each loss one DeepLabV3+ is trained). Once again, we can see that among the stand-alone loss functions the highest average performance is obtained by the proposed TL. Moreover, the combination of these functions obtains good performance and provides an interesting yet initial starting point for future work. We are unable to directly compare with [11] because they utilize a significantly larger dataset that is not accessible.

TABLE 7. Comparison of the loss functions in terms of the different metrics (in %). Performance of a single network on the wire dataset [11]. Best performance in bold.

Model	Precision	Recall	Acc.	Dice	IoU
DI(0.375)	74.63	64.93	97.62	75.32	61.03
CE(0.375)	71.30	62.68	96.56	72.56	58.30
GL(0.375)	75.38	66.22	97.93	77.23	63.25
TL(0.375)	75.91	67.08	98.12	78.74	64.10
GL + TL(0.375)	76.27	67.72	98.65	**79.25**	64.95
GL+TL+DI(0.375)	**77.06**	**68.74**	**98.92**	79.15	**65.32**

V. CONCLUSION

In this study, we introduced a novel loss function designed to tackle discontinuity challenges encountered in road extraction from remote sensing images using deep learning networks. Our proposed loss function is model-agnostic, providing flexibility to seamlessly integrate with various segmentation models, both existing and future.

The effectiveness of our proposed loss function was evaluated across different segmentation datasets, demonstrating significant improvements across a range of evaluation metrics. This underscores its robustness and generalizability in diverse settings beyond our initial focus.

Additionally, to assess the versatility and transferability of our approach, we conducted experiments in a different domain-specifically focusing on wire segmentation. This exploratory study revealed promising results, suggesting that our methodology extends beyond road segmentation to other tasks involving the segmentation of linear features in images. This initial test into wire segmentation inspires future research directions aimed at expanding the applicability of our approach across diverse domains and tasks.

Looking ahead, we aim to extend our investigation beyond the DeepLabV3+ architecture explored in this work. Future research will explore applying these loss functions within attention-based network topologies and exploring ensemble strategies with varied network architectures. Moreover, our ongoing research will delve into utilizing specialized convolution filters, such as curved filters tailored for segmenting complex road features like roundabouts. This endeavor represents a promising direction for enhancing the capabilities of road extraction systems in real-world applications. The positive performance observed in a different domain further inspires and guides our future work in this exciting direction.

ACKNOWLEDGMENT

The authors would like to acknowledge the support that NVIDIA provided them through the GPU Grant Program. They used a donated TitanX GPU to train deep networks used in this work. They would like to thank Lorenzo Giannini who created the masks for the Corniolo dataset.

DECLARATION OF INTERESTS

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

REFERENCES

[1] M. I. Ahmed, M. Foysal, M. D. Chaity, and A. B. M. A. Hossain, ''DeepRoadNet: A deep residual based segmentation network for road map detection from remote aerial image,'' *IET Image Process.*, vol. 18, no. 1, pp. 265–279, Jan. 2024.

[2] F. Bastani, S. He, S. Abbar, M. Alizadeh, H. Balakrishnan, S. Chawla, D. J. DeWitt, and S. Madden, ''Unthule: An incremental graph construction process for robust road map extraction from aerial images,'' 2018, *arXiv:1802.03680*.

[3] F. Bastani, S. He, S. Abbar, M. Alizadeh, H. Balakrishnan, S. Chawla, S. Madden, and D. DeWitt, ''RoadTracer: Automatic extraction of road networks from aerial images,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4720–4728.

[4] F. Bastani and S. Madden, ''Beyond road extraction: A dataset for map update using aerial images,'' in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 11885–11894.

[5] A. Batra, S. Singh, G. Pang, S. Basu, C. V. Jawahar, and M. Paluri, ''Improved road connectivity by joint learning of orientation and segmentation,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 10377–10385.

[6] R. Bravin, L. Nanni, A. Loreggia, S. Brahnam, and M. Paci, ''Varied image data augmentation methods for building ensemble,'' *IEEE Access*, vol. 11, pp. 8810–8823, 2023.

[7] S. A. Cadena, G. H. Denfield, E. Y. Walker, L. A. Gatys, A. S. Tolias, M. Bethge, and A. S. Ecker, ''Deep convolutional models improve predictions of macaque V1 responses to natural images,'' *PLOS Comput. Biol.*, vol. 15, no. 4, Apr. 2019, Art. no. e1006897.

[8] X. Cao, K. Zhang, and L. Jiao, ''CSANet: Cross-scale axial attention network for road segmentation,'' *Remote Sens.*, vol. 15, no. 1, p. 3, Dec. 2022.

[9] A. Chaurasia and E. Culurciello, ''LinkNet: Exploiting encoder representations for efficient semantic segmentation,'' in *Proc. IEEE Vis. Commun. Image Process. (VCIP)*, Dec. 2017, pp. 1–4.

[10] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, ''Encoder–decoder with atrous separable convolution for semantic image segmentation,'' 2018, *arXiv:1802.02611*.

[11] M. Tik Chiu, X. Zhang, Z. Wei, Y. Zhou, E. Shechtman, C. Barnes, Z. Lin, F. Kainz, S. Amirghodsi, and H. Shi, ''Automatic high resolution wire segmentation and removal,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 2183–2192.

[12] C. Cornelio, M. Donini, A. Loreggia, M. S. Pini, and F. Rossi, ''Voting with random classifiers (VORACE): Theoretical and experimental analysis,'' *Auto. Agents Multi-Agent Syst.*, vol. 35, no. 2, p. 22, Oct. 2021.

[13] W. R. Crum, O. Camara, and D. L. G. Hill, ''Generalized overlap measures for evaluation and validation in medical image analysis,'' *IEEE Trans. Med. Imag.*, vol. 25, no. 11, pp. 1451–1461, Nov. 2006.

[14] I. Demir, K. Koperski, D. Lindenbaum, G. Pang, J. Huang, S. Basu, F. Hughes, D. Tuia, and R. Raskar, ''DeepGlobe 2018: A challenge to parse the Earth through satellite images,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 17200–17209.

[15] M. Dey, P. Prakash, and B. H. Aithal, ''UnetEdge: A transfer learning-based framework for road feature segmentation from high-resolution remote sensing images,'' *Remote Sens. Appl. Soc. Environ.*, vol. 34, Apr. 2024, Art. no. 101160.

[16] J. Dulac. (2013). *Global Land Transport Infrastructure Requirements*. Accessed: Apr. 4, 2024. [Online]. Available: https://www.iea.org/reports/global-land-transport-infrastructure-requirements

[17] X. Gao, X. Sun, Y. Zhang, M. Yan, G. Xu, H. Sun, J. Jiao, and K. Fu, ''An end-to-end neural network for road extraction from remote sensing imagery by multiple feature pyramid network,'' *IEEE Access*, vol. 6, pp. 39401–39414, 2018.

[18] G. V. George, M. S. Hussain, R. Hussain, and S. Jenicka, ''Efficient road segmentation techniques with attention-enhanced conditional GANs,'' *Social Netw. Comput. Sci.*, vol. 5, no. 1, pp. 1–12, Jan. 2024.

[19] M. Guo, H. Liu, Y. Xu, and Y. Huang, ''Building extraction based on U-Net with an attention block and multiple losses,'' *Remote Sens.*, vol. 12, no. 9, p. 1400, Apr. 2020.

[20] J. Hanley and B. Mcneil, ''The meaning and use of the area under a receiver operating characteristic (ROC) curve,'' *Radiology*, vol. 143, pp. 29–36, May 1982.

[21] X. Hu, F. Li, D. Samaras, and C. Chen, ''Topology-preserving deep image segmentation,'' in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 5658–5669.

[22] S. Jadon, ''A survey of loss functions for semantic segmentation,'' in *Proc. IEEE Conf. Comput. Intell. Bioinf. Comput. Biol. (CIBCB)*, Oct. 2020, pp. 1–7.

[23] K. Jiang, Z. Wang, P. Yi, G. Wang, K. Gu, and J. Jiang, ''ATMFN: Adaptive-threshold-based multi-model fusion network for compressed face hallucination,'' *IEEE Trans. Multimedia*, vol. 22, no. 10, pp. 2734–2747, Oct. 2020.

[24] P. Kaiser, J. D. Wegner, A. Lucchi, M. Jaggi, T. Hofmann, and K. Schindler, ''Learning aerial image segmentation from online maps,'' *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 11, pp. 6054–6068, Nov. 2017.

[25] A. Krizhevsky, I. Sutskever, and G. E. Hinton, ''ImageNet classification with deep convolutional neural networks,'' in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[26] I. Laptev, H. Mayer, T. Lindeberg, W. Eckstein, C. Steger, and A. Baumgartner, ''Automatic extraction of roads from aerial images based on scale space and snakes,'' *Mach. Vis. Appl.*, vol. 12, no. 1, pp. 23–31, Jul. 2000.

[27] R. Lian and L. Huang, ''DeepWindow: Sliding window based on deep learning for road extraction from remote sensing images,'' *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 1905–1916, 2020.

[28] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, ''Focal loss for dense object detection,'' in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2999–3007.

[29] X. Lu, Y. Zhong, Z. Zheng, and L. Zhang, ''GAMSNet: Globally aware road detection network with multi-scale residual learning,'' *ISPRS J. Photogramm. Remote Sens.*, vol. 175, pp. 340–352, May 2021.

[30] W. Ma, Y. Guo, H. Zhu, X. Yi, W. Zhao, Y. Wu, B. Hou, and L. Jiao, ''Intra- and intersource interactive representation learning network for remote sensing images classification,'' *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5401515.

[31] M. Maboudi, J. Amini, M. Hahn, and M. Saati, ''Object-based road extraction from satellite images using ant colony optimization,'' *Int. J. Remote Sens.*, vol. 38, no. 1, pp. 179–198, Jan. 2017.

[32] G. Máttyus, W. Luo, and R. Urtasun, ''DeepRoadMapper: Extracting road topology from aerial images,'' in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3458–3466.

[33] G. Máttyus, S. Wang, S. Fidler, and R. Urtasun, ''Enhancing road maps by parsing aerial images around the world,'' in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1689–1697.

[34] J. Mei, R.-J. Li, W. Gao, and M.-M. Cheng, ''CoANet: Connectivity attention network for road extraction from satellite imagery,'' *IEEE Trans. Image Process.*, vol. 30, pp. 8540–8552, 2021.

[35] J. R. Meijer, M. A. J. Huijbregts, K. C. G. J. Schotten, and A. M. Schipper, ''Global patterns of current and future road infrastructure,'' *Environ. Res. Lett.*, vol. 13, no. 6, Jun. 2018, Art. no. 064006.

[36] F. Milletari, N. Navab, and S.-A. Ahmadi, ''V-Net: Fully convolutional neural networks for volumetric medical image segmentation,'' in *Proc. 4th Int. Conf. 3D Vis. (3DV)*, Oct. 2016, pp. 565–571.

[37] V. Mnih, ''Machine learning for aerial image labeling,'' Ph.D. thesis, Dept. Comput. Sci., Univ. Toronto, Toronto, ON, Canada, 2013.

[38] V. Mnih, *Machine Learning for Aerial Image Labeling*. Toronto, ON, Canada: Univ. Toronto, 2013.

[39] A. Mosinska, P. Marquez-Neila, M. Kozinski, and P. Fua, ''Beyond the pixel-wise loss for topology-aware delineation,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3136–3145.

[40] L. Nanni, May 2024, ''Boston & Corniolo datasets—Road segmentation—Described in 'An enhanced loss function for semantic road segmentation in remote sensing images,''' *Zenodo*, doi: 10.5281/zenodo.11107139.

[41] L. Nanni, C. Fantozzi, A. Loreggia, and A. Lumini, ''Ensembles of convolutional neural networks and transformers for polyp segmentation,'' *Sensors*, vol. 23, no. 10, p. 4688, May 2023.

[42] L. Nanni, A. Loreggia, L. Barcellona, and S. Ghidoni, ''Building ensemble of deep networks: Convolutional networks and transformers,'' *IEEE Access*, vol. 11, pp. 124962–124974, 2023.

[43] L. Nanni, A. Lumini, A. Loreggia, S. Brahnam, and D. Cuza, ''Deep ensembles and data augmentation for semantic segmentation,'' in *Diagnostic Biomedical Signal and Image Processing Applications With Deep Learning Methods*. Amsterdam, The Netherlands: Elsevier, 2023, pp. 215–234.

[44] L. Nanni, A. Lumini, A. Loreggia, A. Formaggio, and D. Cuza, "An empirical study on ensemble of segmentation approaches," *Signals*, vol. 3, no. 2, pp. 341–358, Jun. 2022.

[45] T. Panboonyuen, P. Vateekul, K. Jitkajornwanich, and S. Lawawirojwong, "An enhanced deep convolutional encoder–decoder network for road segmentation on aerial imagery," in *Proc. 13th Int. Conf. Comput. Inf. Technol. (IC2IT)*. Switzerland: Springer, 2017, pp. 191–201.

[46] S. S. M. Salehi, D. Erdogmus, and A. Gholipour, "Tversky loss function for image segmentation using 3D fully convolutional deep networks," in *Proc. 8th Int. Workshop Mach. Learn. Med. Imag.* Switzerland: Springer, 2017, pp. 379–387.

[47] B. Savelli, A. Bria, M. Molinara, C. Marrocco, and F. Tortorella, "A multi-context CNN ensemble for small lesion detection," *Artif. Intell. Med.*, vol. 103, Mar. 2020, Art. no. 101749.

[48] P. Sharma, R. Kumar, M. Gupta, and A. Nayyar, "A critical analysis of road network extraction using remote sensing images with deep learning," *Spatial Inf. Res.*, pp. 1–11, Feb. 2024.

[49] S. Shashaani, M. Teshnehlab, A. Khodadadian, M. Parvizi, T. Wick, and N. Noii, "Using layer-wise training for road semantic segmentation in autonomous cars," *IEEE Access*, vol. 11, pp. 46320–46329, 2023.

[50] S. Sloan, R. R. Talkhani, T. Huang, J. Engert, and W. F. Laurance, "Mapping remote roads using artificial intelligence and satellite imagery," *Remote Sens.*, vol. 16, no. 5, p. 839, Feb. 2024.

[51] J. Sun and Y. Li, "Multi-feature fusion network for road scene semantic segmentation," *Comput. Electr. Eng.*, vol. 92, Jun. 2021, Art. no. 107155.

[52] Y.-Q. Tan, S.-H. Gao, X.-Y. Li, M.-M. Cheng, and B. Ren, "VecRoad: Point-based iterative graph exploration for road graphs extraction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8907–8915.

[53] J. Tao, Z. Chen, Z. Sun, H. Guo, B. Leng, Z. Yu, Y. Wang, Z. He, X. Lei, and J. Yang, "Seg-road: A segmentation network for road extraction based on transformer and CNN with connectivity structures," *Remote Sens.*, vol. 15, no. 6, p. 1602, Mar. 2023.

[54] S. Vasu, M. Kozinski, L. Citraro, and P. Fua, "TopoAL: An adversarial learning approach for topology-aware road segmentation," in *Proc. 16th Eur. Conf. Comput. Vis. (ECCV)*, Glasgow, U.K. Cham, Switzerland: Springer, Aug. 2020, pp. 224–240.

[55] J. Wang, J. Song, M. Chen, and Z. Yang, "Road network extraction: A neural-dynamic framework based on deep learning and a finite state machine," *Int. J. Remote Sens.*, vol. 36, no. 12, pp. 3144–3169, Jun. 2015.

[56] Y. Wang, L. Tong, S. Luo, F. Xiao, and J. Yang, "A multiscale and multidirection feature fusion network for road detection from satellite imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5615718.

[57] 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, May 2017.

[58] H. Xu, H. He, Y. Zhang, L. Ma, and J. Li, "A comparative study of loss functions for road segmentation in remotely sensed road datasets," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 116, Feb. 2023, Art. no. 103159.

[59] Z. Xu, Y. Sun, and M. Liu, "Topo-boundary: A benchmark dataset on topological road-boundary detection using aerial images for autonomous driving," *IEEE Robot. Autom. Lett.*, vol. 6, no. 4, pp. 7248–7255, Oct. 2021.

[60] H. Yan, C. Zhang, J. Yang, M. Wu, and J. Chen, "Did-linknet: Polishing D-block with dense connection and iterative fusion for road extraction," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2021, pp. 2186–2189.

[61] M. Yi-De, L. Qing, and Q. Zhi-Bai, "Automated image segmentation using improved PCNN model based on cross-entropy," in *Proc. Int. Symp. Intell. Multimedia, Video Speech Process.*, Oct. 2004, pp. 743–746.

[62] Z. Yu, R. Chang, and Z. Chen, "Automatic detection method for loess landslides based on GEE and an improved YOLOX algorithm," *Remote Sens.*, vol. 14, no. 18, p. 4599, Sep. 2022.

[63] Z. Yu, Z. Chen, Z. Sun, H. Guo, B. Leng, Z. He, J. Yang, and S. Xing, "SegDetector: A deep learning model for detecting small and overlapping damaged buildings in satellite images," *Remote Sens.*, vol. 14, no. 23, p. 6136, Dec. 2022.

[64] W. Yuan and W. Xu, "NeighborLoss: A loss function considering spatial correlation for semantic segmentation of remote sensing image," *IEEE Access*, vol. 9, pp. 75641–75649, 2021.

[65] W. Yuan and W. Xu, "GapLoss: A loss function for semantic segmentation of roads in remote sensing images," *Remote Sens.*, vol. 14, no. 10, p. 2422, May 2022.

[66] J. Zhang, Y. Dong, M. Kuang, B. Liu, B. Ouyang, J. Zhu, H. Wang, and Y. Meng, "The art of defense: Letting networks fool the attacker," *IEEE Trans. Inf. Forensics Security*, vol. 18, pp. 3267–3276, 2023.

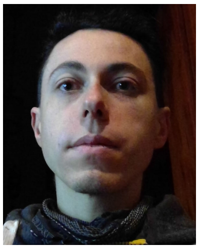
[67] W. Zhang, P. Zhuang, H.-H. Sun, G. Li, S. Kwong, and C. Li, "Underwater image enhancement via minimal color loss and locally adaptive contrast enhancement," *IEEE Trans. Image Process.*, vol. 31, pp. 3997–4010, 2022.

[68] Z. Zhang, Q. Liu, and Y. Wang, "Road extraction by deep residual U-Net," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 5, pp. 749–753, May 2018.

[69] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6230–6239.

[70] Z. Zhong, J. Li, W. Cui, and H. Jiang, "Fully convolutional networks for building and road extraction: Preliminary results," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2016, pp. 1591–1594.

[71] L. Zhou, C. Zhang, and M. Wu, "D-LinkNet: LinkNet with pretrained encoder and dilated convolution for high resolution satellite imagery road extraction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 182–186.



LORIS NANNI is currently an Associate Professor with the Department of Information Engineering, University of Padova. He carries out research with DEI, University of Padova, in the fields of biometric systems, pattern recognition, machine learning, image databases, and bioinformatics. He is the coauthor of more than 300 research articles. He has an H-index of 57 and 12 000 citations (Google Scholar). He has extensively served as a Referee for international journals, IEEE Transactions on Pattern Analysis and Machine Intelligence, *Pattern Recognition*, *Bioinformatics*, *BMC Bioinformatics*, and *Pattern Recognition Letters* and projects.

SHERYL BRAHNAM received the master's degree from the City College of New York, in 1997, and the Ph.D. degree in computer science from the Graduate Center, City University of New York, in 2002. She is currently a Professor with Missouri State University. Her research interests include pattern recognition, face recognition, bioinformatics, and medical image analysis.

ANDREA LOREGGIA received the master's degree (cum laude) from the University of Padova, in 2012, and the Ph.D. degree in computer science, in 2016. He is currently an Assistant Professor with the Department of Information Engineering, University of Brescia. His studies are dedicated to designing and providing tools for developing intelligent agents capable of representing and reasoning with preference and ethical-moral principles. His research interest includes artificial intelligence span from knowledge representation to deep learning. He is a member of the UN/CEFACT Group of Experts, he actively participates in the dissemination and sustainable development of technology.

Open Access funding provided by 'Università degli Studi di Padova' within the CRUI CARE Agreement