Improved Road Connectivity by Joint Learning of Orientation and Segmentation

Anil Batra *[1] Suriya Singh *† [2] Guan Pang[3] Saikat Basu[3] C.V. Jawahar[1] Manohar Paluri[3]

[1]IIIT Hyderabad [2]MILA / Polytechnique Montréal [3]Facebook

Abstract

Road network extraction from satellite images often produce fragmented road segments leading to road maps unfit for real applications. Pixel-wise classification fails to predict topologically correct and connected road masks due to the absence of connectivity supervision and difficulty in enforcing topological constraints. In this paper, we propose a connectivity task called Orientation Learning, motivated by the human behavior of annotating roads by tracing it at a specific orientation. We also develop a stacked multi-branch convolutional module to effectively utilize the mutual information between orientation learning and segmentation tasks. These contributions ensure that the model predicts topologically correct and connected road masks. We also propose Connectivity Refinement approach to further enhance the estimated road networks. The refinement model is pre-trained to connect and refine the corrupted ground-truth masks and later fine-tuned to enhance the predicted road masks. We demonstrate the advantages of our approach on two diverse road extraction datasets SpaceNet [30] and DeepGlobe [11]. Our approach improves over the state-of-the-art techniques by 9% and 7.5% in road topology metric on SpaceNet and DeepGlobe, respectively.

1. Introduction

A mapped road network provides routing information to find the traversable paths, which are important for planning in various applications such as navigation and disaster management. Example of a connected road network is shown in Figure 1a. Manual mapping of a complex road network is time consuming and requires intensive human effort. Automatic extraction of road networks from satellite imagery has been proposed [2, 6, 18, 29, 33], where recently, deep learning based techniques have shown high quality mapping results in diverse scenarios [3, 8, 10, 19, 21–23, 28, 31, 35]. However, the extracted road networks often produce fragmented road segments, and therefore, are unfit for real applications (Figure 1b). Satellite images pose difficulties in the extraction of roads due to (a) shadows of clouds and trees, (b) diverse appearance and illumination condition due to terrain, weather, geography, etc., and, (c) similarity of road texture with other materials. Label scarcity [28] as well as omission and registration noise in road ground-truths [22] also inhibit the accurate estimation of road maps.

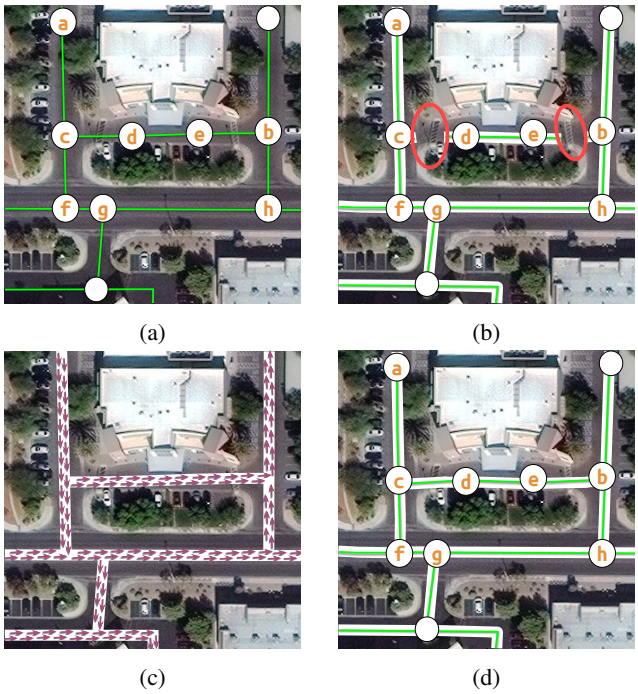

Figure 1: Road network extraction formulated as binary segmentation fails to produce topologically correct road map due to change in road appearance. (a) Annotators trace lines (highlighted nodes) along the center of roads with a traversable shortest path (a, c, d, e, b) for $a \rightarrow b$. (b) Fragmented road network estimated using segmentation resulting in path (a, c, f, g, h, b) for $a \rightarrow b$. (c) Tracing roads with orientation to achieve connectivity. (d) We extract connected and topologically correct road networks using segmentation and orientation.

Road network extraction is explored in [8, 10, 19, 21, 22], where the problem is posed as segmentation followed by post-processing steps to refine and couple the missing

*Equal Contribution
†Work partially done as Research Fellow at IIIT Hyderabad
Code available at: https://github.com/anil-2185/road-connectivity

connections. The pixel-wise classification supervision does not constrain the model to learn representations for connected road segments [23], leading to poor estimation of road topology. Predicting masks with accurate topology is a challenging task due to difficulty in enforcing topological constraints via a loss function [20, 23] or during post-processing [19]. To measure deviations in topology, Mosinska *et al*. [23] rely on higher-level abstract features of ground-truth and predicted road masks whereas Máttyus *et al*. [20] employ an adversarial matching paradigm. To improve road connectivity, Máttyus *et al*. [19] proposed post-processing steps to reason for missing connection hypotheses while Bastani *et al*. [3] and Ventura *et al*. [31] iteratively connect road segments in the neighbouring image patches.

Our focus is on improving connectivity in road network extraction from binary segmentation of overhead imagery. Characterizing connectivity supervision in the way human annotates road maps requires topological and structural information of roads. We build our approach on the intuition that to annotate road maps human trace lines along the road orientation to connect the fragmented road segments. Consider Figure 1b, tracing lines $c \rightarrow b$ via d and e can connect the broken roads. This motivates us to design a connectivity task using available road labels to predict road orientation angle along with the road segmentation (Figure 1c).

In this paper, we propose to learn a road orientations jointly with per-pixel road segmentation in multi-branch CNN model (Figure 2). We also propose connectivity refinement which connect small gaps and reduces false positives in the prediction. The connectivity refinement model is pre-trained to restore the corrupted road ground-truth masks (Figure 2 and 4). This allows the model to effectively correct diverse failure scenarios. Similar to Mosinska *et al*. [23], our connectivity refinement model can be employed in an iterative manner, however, our refinement approach focuses on improving connectivity with the help of pre-training in addition to segmentation improvement. Lastly, we design a joint learning module by stacking multi-branch encoder-decoder structure (Figure 5 and 6). This module is a variant of stacked hourglass network [24], however our motivation is different i.e., flow of information between the related tasks to improve the performance of individual task in a multi-task learning framework. In contrast to [3, 19, 22, 28], our segmentation model inherently captures the information of connected road segments in the intermediate representation, leading to an accurate topology in road network estimation (Figure 1d).

Contributions:

1. We design an orientation learning task and demonstrate that the joint learning of orientation and segmentation improves the connectivity of road network.
2. We propose a connectivity refinement approach pre-trained with corrupted road ground-truth masks and fine-tuned with segmentation outputs to iteratively enhance the topology of the estimated road networks.
3. We design a stacked multi-branch module to effectively utilize the dual supervision. We show that the proposed module enables the flow of information between the tasks and helps in boosting the connectivity.

2. Related Work

Road Network Extraction: Numerous techniques have been developed in literature to extract road networks from satellite images. Traditional methods impose connectivity by incorporating contextual priors such as road geometry [18], higher order CRF formulation [33], marked point processes [6, 29], and solving integer programming on road graphs [2]. These methods utilized hand designed features and optimized for complex objectives. In recent deep learning based techniques, road extraction is formulated as segmentation problem [19, 21–23, 28] using convolutional encoder-decoder structured models, which are able to capture large spatial context. Different from segmentation based approaches, Bastani *et al*. [3] introduced graph based methodology to predict road line strings. In the current scope, we focus on segmentation based approaches. Mnih *et al*. [21] learn road classification by CNN model in multiple stages (to reduce false negative rate due to label noise), operating on the image patches. Máttyus *et al*. [19] propose encoder-decoder structure model and pose it as multi-class (`roads`, `building` and `background`) segmentation. The model performs well in segmentation, however, fails to predict connected roads, and missing roads are connected using shortest path algorithms in the post processing steps to improve the connectivity. Máttyus *et al*. [19] further use a binary decision classifier to predict the correctness of connections. We found that [19] face difficulty in correctly adding and classifying the missing road connections in regions with high road density, ambiguous road appearance, occlusions, and complex road topology present in the datasets (SpaceNet [30] and DeepGlobe [11]) we validate our methods on.

The other well admired encoder-decoder structure to learn thin curvilinear road structures are U-Net [27] and LinkNet [7]. Their variants are proposed to learn the road segmentation in [8, 10]. LinkNet34 [7] has been primarily utilized to segment the roads in DeepGlobe challenge [11]. Nevertheless, connectivity is achieved with more heuristic based post-processing in these methods. In contrast, we propose joint learning of connectivity task and road segmentation with a stacked encoder-decoder structure. The most recent work of Mosinska *et al*. [23] combine pixel-wise classification and perceptual losses [12] to learn road topology in U-Net [27]. Mosinska *et al*. [23] also proposed

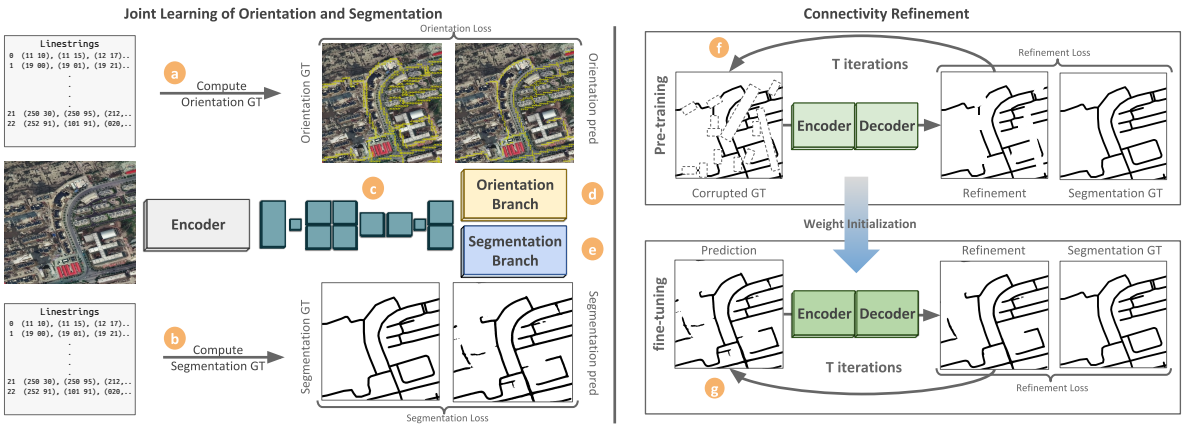

Figure 2: Our approach for extracting connected road topology from satellite images. Annotations in the form of line strings, are converted to (a) orientation groundtruth and (b) road groundtruth masks. We use encoder-decoder structure with (c) *stacked multi-branch module* to jointly learn (d) orientation and (e) segmentation, providing dual supervision to the model. The orientation task is designed to improve the road connectivity. Finally, a connectivity refinement network, (f) pre-trained with corrupted groundtruths to remove false roads and further improve the road connectivity, is (g) fine-tuned with road segmentation output to iteratively enhance the estimated road networks.

iterative refinement to fill the small gaps in road segments. The introduced loss term favors the road like structures but is inefficient in connecting the road segments.

Multi-Task Learning (MTL): It is a learning mechanism [4], inspired from human beings to acquire knowledge of complex tasks by performing different shared sub-tasks simultaneously. Multi-task learning improves the performance by inducing mutual information of the tasks in the learning process. MTL has been applied successfully in various domains such as speech recognition, natural language processing [9] and computer vision [17]. Readers are suggested to read survey [34] on multi-task learning.

Humans perform two related tasks while annotating the roads i.e. identify the road pixels and trace lines to connect them. In our work, we use multi-task learning to incorporate road annotation as two tasks i.e. while labeling the satellite images, humans recognize roads and connect them by tracing lines, inherently identifying the orientation. We show that these related tasks improve the connectivity with improved encoded representation in the encoder.

3. Method

Road extraction from overhead images via segmentation based methods produce disconnected road segments. To address this, we develop an orientation task from the road line strings (Section 3.1) and use it as an auxiliary loss along with pixel-wise segmentation loss. The motivation of orientation loss is to capture the relational information between the neighboring pixels through explicit learning of orientations between them. We formulate the problem as a two stage process: (a) joint learning of road orientation and segmentation in multi-task fashion, and (b) a connectivity refinement using a pre-trained CNN model (Section 3.2). We first present our novel inductive task followed by a connectivity refinement technique. Finally, we outline the proposed end-to-end joint learning pipeline with two stacks of multi-branch encoder-decoder which can flow the information across the tasks (Section 3.3).

3.1. Orientation Learning

The pixel level annotation of roads is a computationally costly and time consuming task. To reduce the human effort, roads are preferably annotated with line strings connecting 2D points. We visualize each road line string as a directional vector between two consecutive points in 2D image plane (see Figure 3). The directional vector provides the orientation (tracing angle) of each road segment.

The orientation learning task is partly inspired from Part Affinity Fields [36] and bears resemblance with the deep watershed technique for instance segmentation [1]. Intuitively, representations learned for instance (road segments) segmentation would lead to improved connectivity in the estimated road network. However, road segments, unlike object instances or human body parts, do not have defined boundary between them and are rather interconnected. Therefore, instead of predicting orientation from the object boundary towards its centroid [36], we encode and predict the unit vector pointing towards the next pixel in the same or the connected adjacent road segment. Learning orientation with a pixel based cross-entropy loss poses a connectivity constraint in the encoded representation as learning of road orientations favors the connected road segments and joint learning of related tasks often leads to more generalizable

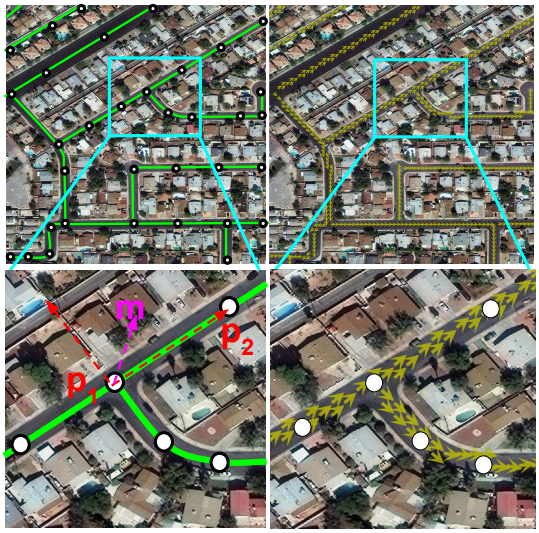

Figure 3: Road Orientations. Top left: road line strings annotations. Bottom left: two consecutive points to compute the orientation angle. Top right: Ground-truth road orientation vectors. Bottom right: Road orientation ground-truth in an image patch.

features [4, 17]. Orientation learning can be extended to applications like automatic segmentation along the object boundary [3, 5], connect the occluded lanes in lane detection, connect broken alphabets in OCR, etc.

We now describe the process to generate the orientation ground-truths from line strings. Consider an image shown in Figure 3 with road line strings $\{l_1, l_2, \ldots, l_m\}$ and each line string l_k consists of 2D points $\{p_1, p_2, \ldots, p_n\}$. We assume undirected road network, ignoring driving direction of the roads. We sort the coordinates of the points of each line string such that most of the directional vectors point from left to right and top to bottom, which we find to be appropriate for the neural network to learn and focus on connected road representation. We compute a unit directional vector $|\vec{v}(x,y)| \in [-1,1]$ between two consecutive point pairs $\{(p_1,p_2),\ldots,(p_{n-1},p_n)\}$ of l_k using (1) and convert it into polar domain to obtain orientation angle o_r using (2). For each point pair (p_i, p_j) using (3), the pixels lying within the threshold width λ_{orient} along the perpendicular direction of l_k, are assigned the same orientation value; for all other pixels non-road orientation angle o_b is assigned.

$$\vec{v}_{ij}(x,y) = \frac{p_i(x,y) - p_j(x,y)}{||p_i(x,y) - p_j(x,y)||_2^2} \tag{1}$$

$$\vec{v}_{ij}(x,y) \equiv \langle 1 \quad \measuredangle o_r \rangle \tag{2}$$

$$o_{l_k}(m) = \begin{cases} o_r & \text{if } |\vec{v}_\perp \cdot \overrightarrow{(m - p_1)}| < \lambda_{orient} \\ o_b & \text{otherwise.} \end{cases} \tag{3}$$

where $||p_i - p_j||_2^2$ is the total length between the consecutive points, $v_\perp$ is a vector perpendicular to unit directional vector, (x,y) are the coordinates of points and o is ground truth for orientations. We ignore non-road orientation angle during plotting of the vectors in Figure 2 and 3.

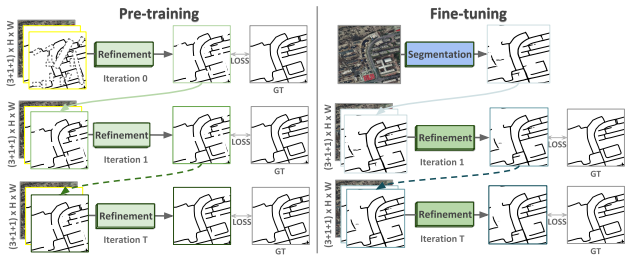

Figure 4: Connectivity Refinement. We pre-train the encoder-decoder CNN to remove false roads with pre-text task of correcting the corrupted road ground-truth masks. The model is later fine-tuned to refine the road segmentation outputs.

3.2. Connectivity Refinement

The orientation supervision improves the connectivity in the estimated road network. However, complex and dense road topology such as bridges and parking lots leads to failure in orientation prediction. The model also hallucinates roads in regions with similar textures e.g. road like patterns in farms. To further improve the prediction topology and suppress false positives, we employ the connectivity refinement (see Figure 4). Motivated by the success of restoring the images from corruption [25, 28], we interpret missing and spurious road segments as corrupted road ground-truth mask. We first pre-train the refinement network to restore the corrupted masks allowing the model to learn connectivity pattern as well as remove false roads. Note that, we opt for weight initialization and do not train the connectivity refinement using segmentation outputs and corrupted GT simultaneously to avoid overfitting to a single distribution of corruptions [14]. In pre-training stage, we concatenate satellite image X, corrupted ground-truth y' along with previous road prediction $\bar{y}_{t-1}$ (where $\bar{y}_0 = y'$) and feed it as input to the refinement model $g(\cdot)$.

$$\bar{y}_t = g\Big([X, y', \bar{y}_{t-1}]\Big) \quad t = 1, \ldots, T \tag{4}$$

At the end of pre-training stage the neural network learns to effectively encode the available contexts and fills the missing road segments. The pre-trained model is further fine-tuned to improve the road segmentation. In fine tuning stage, we replace the manually corrupted ground truth mask with the output of segmentation network.

$$\hat{y}_t = g\Big([X, \hat{y}, \hat{y}_{t-1}]\Big) \quad t = 1, \ldots, T \tag{5}$$

where $\hat{y} = f_{seg}(X)$, $\hat{y}_0 = \hat{y}$, and $[\cdot]$ denotes concatenation along channel axis. We use $T = 3$ and identical encoder-decoder architectures for $g(\cdot)$ and $f_{seg}(\cdot)$.

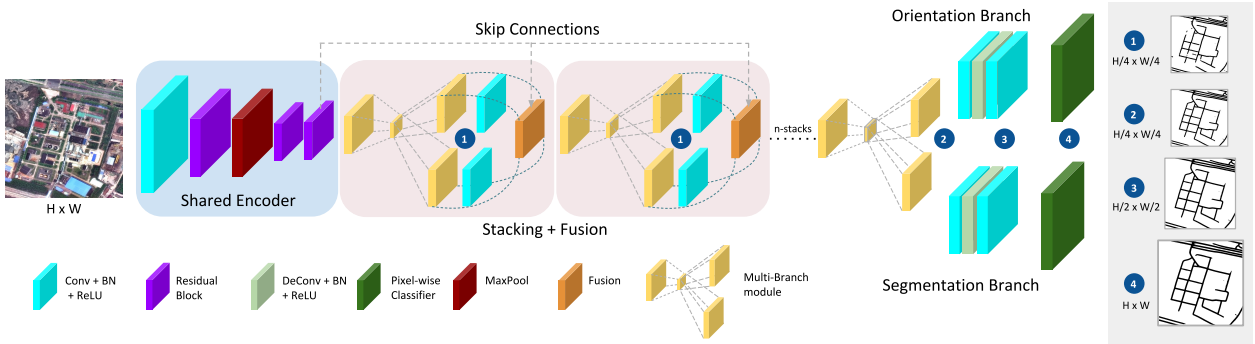

Figure 5: Architecture of *n-stacked multi-branch* CNN to learn road orientation and segmentation simultaneously. The stacked module is capable to calculate losses L_{seg} & L_{orient} at different scales $(\{\frac{1}{4}, \frac{1}{4} \ldots \text{n times}\}, \frac{1}{2}$ and $1)$ to optimize the CNN. We use two stacks of *multi-branch module* (Figure 6) with features fusion in first stack only. Refer to supplementary material for additional architectural details.

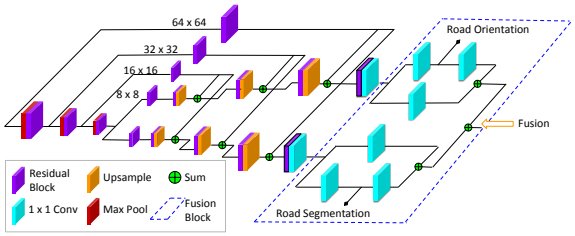

Figure 6: A multi-branch module. The intermediate output is extracted from each branch using 1×1 convolution and are merged using a fusion block.

3.3. Stacked Multi-branch Module

The stacked multi-branch module as shown, in Figure 5 is composed of three blocks: (a) shared encoder, (b) iterative fusion with multi-branch, and (c) prediction branches for orientation and segmentation. The proposed CNN model performs the following tasks simultaneously: (a) learn a robust common representation for connected road segments in the shared encoder, (b) predicts road orientation and road segmentation, and (c) allows the information flow between the tasks to encourage road connectivity.

The shared encoder takes the input image X and learns a mapping function E, which projects the input to an encoded representation for both tasks. The encoding $z = E(X)$ is fed to the stacked multi-branch module to learn the coarse predictions. The motivation for n-stack multi-branch module is three fold: (a) large receptive field to capture the spatial context, (b) mini encoder-decoder structure learns to recalibrate features and coarse predictions in a repetitive fashion, and (c) it allows the information to flow from previous stack to the subsequent stack and refine the coarse predictions. We denote the stacking with a function H_n, where n is number of stacked multi-branch modules and coarse predictions with $\bar{o}$ for orientation and $\bar{y}$ for roads in (6).

To learn refine predictions $\hat{o}$ and $\hat{y}$ from the coarse predictions $\bar{o}_n$ and $\bar{y}_n$, we create two symmetric branches for each task. Each branch learns to up-sample the predictions using decoder networks consisting of two transposed convolutions followed by a pixel-wise convolutional classifier.

$$\bar{o}_n, \bar{y}_n = \begin{cases} H_n(\bar{o}_{n-1} + \bar{y}_{n-1} + z) & \text{if } n > 1 \\ H(z) & \text{if } n = 1 \end{cases} \tag{6}$$

Loss Function: The proposed network is capable of yielding the intermediate outputs at different scales, n outputs from each stack of multi-branch module at $\frac{1}{4}$ scale and two from successive transposed convolution at $\frac{1}{2}$ and 1. Hence, this allows to use multi-scale loss to guide the network while training. Let (X, y, o) be a given labeled sample from the dataset and $f(\cdot)$ denotes the prediction function using our model. We optimize the following loss functions:

$$L_{seg}(\hat{y}, y) = -\texttt{SoftIoU}\Big(f_{seg}(X), y\Big) \tag{7}$$

$$L_{orient}(\hat{o}, o) = -\sum_{c=0}^{o_l} o_c \log\Big(f_{orient}(X)\Big) \tag{8}$$

$$Loss = \sum_{s} \big(L^s_{seg} + L^s_{orient}\big) \tag{9}$$

where SoftIoU is differentiable IoU loss function [19], o_l is the number of bins in the quantized orientation, and s is scale having values $\{\frac{1}{4}, \frac{1}{4}, \ldots \text{n times}\}, \frac{1}{2}$ and 1.

4. Evaluation Metrics

Pixel Based Metrics: We evaluate the performance of our approach for road segmentation using intersection over union (IoU) and $F1$-score metrics. The groundtruth for road segmentation is obtained by rasterizing the road line strings with constant width in SpaceNet dataset [30]. The constant road mask for varying road widths can adversely affect the pixel based metrics. Thus, we use the relaxed

metrics, suggested by Mnih *et al*. [22] with a buffer of 4 pixels in our evaluations.

Graph Based Metric: To measure the estimated topology and road connectivity, we use the Average Path Length Similarity ($APLS$) [30] as the evaluation metric. The metric captures the deviations in shortest path distances between all pair of nodes in a graph. The ground-truth and predicted road network graphs are obtained from y and $\hat{y}$, respectively. $S_{P\rightarrow T}$ (10) measures the sum of difference of shortest path for each node pair in groundtruth graph $G = (V, E)$ and estimated graph $\hat{G} = (\hat{V}, \hat{E})$. To penalize the false positives, symmetric term $S_{T\rightarrow P}$ is added to $APLS$ metric which considers predicted graph as groundtruth and true graph as prediction.

$$S_{P\rightarrow T} = 1 - \frac{1}{|V|}\sum min\left(1, \frac{|L(a,b) - L(\hat{a},\hat{b})|}{L(a,b)}\right) \quad (10)$$

$$APLS = \frac{1}{N}\sum_{(y,\hat{y})}\left(\frac{1}{\frac{1}{S_{P\rightarrow T}(G,\hat{G})} + \frac{1}{S_{T\rightarrow P}(\hat{G},G)}}\right) \quad (11)$$

where $a, b \in V$, $\hat{a}, \hat{b} \in \hat{V}$, $|V|$ is number of nodes in groundtruth graph, and N is number of images. $L(a, b)$ and $L(\hat{a}, \hat{b})$ are the path length of $a \rightarrow b$ and $\hat{a} \rightarrow \hat{b}$, respectively.

5. Experiments and Results

5.1. Dataset

We perform our experiments on SpaceNet [30] and DeepGlobe [11] datasets using only 3-band RGB images. We follow the experimental protocols and dataset splits of [28]. We evaluate and report the road connectivity metrics on full resolution images at inference time for each dataset.

SpaceNet [30]: This dataset provides imagery from four different cities with ground resolution of 30cm/pixel and pixel resolution of 1300×1300. Annotations are provided in the form of line strings, representing centerline of the roads. The dataset consists of 2780 images and, following [28], we split the dataset into 2213 images for training and 567 for testing. To augment the training dataset we create crops of 650×650 with overlapping region of 215 pixels, thus providing $\sim$32K images. For validation we use the crops of same size without overlap.

DeepGlobe [11]: This dataset includes imagery from three different regions with pixel level annotations. The ground resolution is 50cm/pixel and pixel resolution is 1024×1024. Following [28], we create splits of 4696 images for training and 1530 for validation. We augment it by creating crops of size 512×512 with overlapping region of 256 pixels, yielding $\sim$42K images for training phase. We compute the road line string ground-truths by skeletonizing the pixel level annotations and smoothing it using Ramer-Douglas-Peucker algorithm [13, 26].

Method	SpaceNet		DeepGlobe	
	road IoU^a	$APLS$	road IoU^a	$APLS$
ResNet18	59.04	52.65	62.12	63.31
ResNet18 + Orientation	61.90	59.06	**64.77**	**68.93**
ResNet18 + Junctions	58.41	52.76	63.54	66.20
LinkNet34	60.33	55.69	62.75	65.33
LinkNet34 + Orientation	**62.45**	**60.77**	64.72	68.71
LinkNet34 + Junctions	60.72	55.91	63.79	67.42

Table 1: Comparison of orientation and junction learning auxiliary tasks for road connectivity. It shows that improvement in the road connectivity is due to orientation learning. road IoU^a: accurate pixel based intersection over union. $APLS$: average path length similarity on the extracted graph from road segmentation.

5.2. Implementation Details

Dataset Preprocessing: Similar to [28], we generate road heatmaps using Euclidean distance transform along the center line of roads and create binary masks with threshold of 0.76. We use narrower road ground-truth masks, as compared to threshold of 0.4 in [28], to avoid merging of lanes and nearby roads. This step is crucial to obtain maps with high connectivity and accurate topology (ablation studies in the supplementary material). We set $\lambda_{orient} = 12$ pixels in (3) as orientation width along the roads which is approximately equal to the width of road masks.

Training Details: We use random crops of size 256×256 from the image followed by mean subtraction. To improve the generalization of network, random horizontal flip, mirroring and rotation is employed as data augmentation. We train the joint network with a batch size of 32 for 120 epochs. We use `SGD` optimizer with momentum = 0.9, weight decay = 0.0005 and initial learning rate of 10^{-2} with step scheduler having drop factor of 10 at epochs $\{60, 90, 110\}$. We perform simple graph processing to remove small hanging road segments and graph smoothing. Following [32], we formulate the regression for road orientations as classification task as direct regression tends to smoothen predictions to the mean [15, 32]. We quantize road orientation angles into bins of 10° (refer to the supplementary material for ablation studies on quantization levels).

5.3. Results

Orientation Learning: We choose two architectures ResNet18 [16] and LinkNet34 [7] to study the performance of orientation learning. We modify both architectures with dual and identical decoders having shared encoder. The results in Table 1 shows that our proposed task for road connectivity generalizes to different architectures. Incorporating the orientation learning as an auxiliary loss improves the $APLS$ for both CNN architectures by 6.41% and 5.08% for SpaceNet [30], respectively. This suggests that multi-task learning of two related task improves the intermediate representation, leading to better generalization. To study

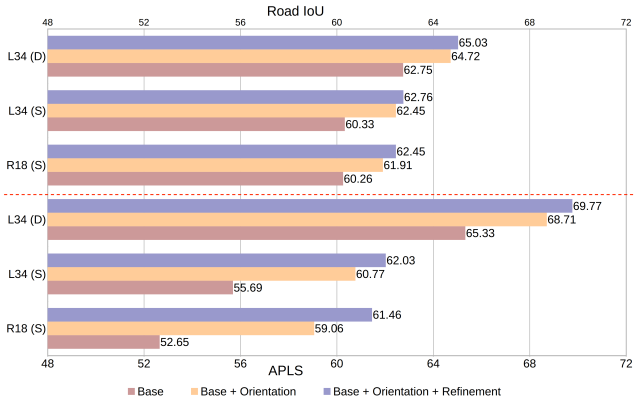

Figure 7: Quantitative improvement with orientation learning and connectivity refinement. **R18**, **L34**: ResNet18 and LinkNet34 based encoder-decoder as joint learning model. **S** and **D**: denote the SpaceNet and DeepGlobe dataset. Horizontal axes show road IoU (top) and $APLS$ improvements (bottom).

the significance of orientation task in road connectivity as an auxiliary loss, we compare it with another shared task of predicting road junctions in multi-task learning framework. In the interest of space, we defer details of *Junction Learning* to the supplementary material. The results in Table 1 shows that the connectivity metric $APLS$ improve with orientation task and not due to the multi-task learning. This validates the efficacy of the orientation task in predicting the connected road topology.

Connectivity Refinement: In contrast to [23], we pretrain the connectivity refinement model with corrupted road ground-truth masks. We analyze different manipulations of road masks such as erasing regions with block structures of different sizes and linear structures of different lengths and thickness. We also add false roads randomly to the road masks with the same structure. We found that manipulations with linear structures appear similar to the real segmentation outputs, as it blends in with the linear road structures thus, we report results for only such manipulations. Figure 7 shows the improvement with connectivity refinement and marginal improvement in road IoU. This shows that the proposed refinement is able to connect gaps and remove the false roads, rather than enhancing road width.

Stacked multi-branch module: We perform experiments to compare our proposed stacked multi-branch module as joint learning module with the state-of-art CNN models commonly used for segmentation of thin structures. We compare the number (n) of multi-branch modules in the model and found that performance stabilizes with two modules. Hence, we employ two stacks of multi-branch modules in our final pipeline. We hypothesize that with more training data, it would be beneficial to add more multi-branch modules which also makes the network deeper without overfitting. The results in Table 2 shows that stacking

Method	**SpaceNet**		**DeepGlobe**	
	road IoU^a	$APLS$	road IoU^a	$APLS$
ResNet18 [16] + Orientation	61.90	59.06	64.77	68.93
LinkNet34 [7] + Orientation	62.45	60.77	64.72	68.71
Unet [27] + Orientation	60.12	58.59	65.21	67.81
Multi-branch(1 Stack) + Orientation	63.26	60.92	65.60	70.23
Multi-branch(2 Stack) + Orientation	**63.75**	**63.65**	**67.21**	**73.21**
Multi-branch(3 Stack) + Orientation	63.73	62.89	66.61	72.48

Table 2: Comparison of joint learning modules with orientation learning employed for road segmentation. It shows that our stacked multi-branch module improves the $APLS$ by 2.7%.

Multi-Scale	**Orientation Learning**	**Feature Fusion**	**Connectivity Refine**	**SpaceNet**		**DeepGlobe**	
				IoU^a	$APLS$	IoU^a	$APLS$
				61.51	58.70	64.23	67.98
✓				61.80	58.49	64.44	67.92
✓	✓			63.44	61.78	66.81	72.03
✓	✓	✓		63.75	63.65	**67.21**	**73.21**
✓	✓	✓	✓	**63.76**	**63.79**	67.02	73.20

Table 3: Step-wise improvement with multi-scale loss, orientation learning, and cross task information flow by feature fusion.

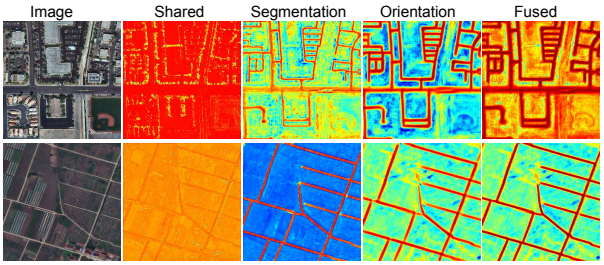

Figure 8: Feature maps for different stages in the proposed model. Shared: feature map from the shared encoder and before the first stack. Segmentation / Orientation: feature map from segmentation / orientation branch of the first stack of multi-branch module before fusion. Fused: additive fusion of all feature maps which is fed to the second stack of multi-branch module.

of multi-branch modules improve the road connectivity over the single encoder-decoder modules by $\sim 2.5\%$.

We study the incremental improvement as a result of our contributions and show the results in Table 3. Initially, we hypothesize that knowledge of road orientation helps in tracing lines to connect the broken road segments, which we achieve by cross information flow between the tasks in stacked multi-branch module. We discover that adding the orientation features with segmentation performs better. This confirms that the neural network utilize the orientation information to connect the broken road segments and improve $APLS$ by 1.87% and 1.18% on respective datasets.

Performing connectivity refinement on the segmentation output of stacked multi-branch model improves $APLS$ marginally. We hypothesize that the n-stack multi branch

Method	SpaceNet						DeepGlobe					
	Precision	Recall	F1	IoU^r	IoU^a	$APLS$	Precision	Recall	F1	IoU^r	IoU^a	$APLS$
DeepRoadMapper (segmentation) [19]	60.61	60.80	60.71	43.58	59.99	54.25	79.82	80.31	80.07	66.76	62.58	65.56
DeepRoadMapper (full) [19]	57.57	58.29	57.93	40.77	N/A	50.59	77.15	77.48	77.32	63.02	N/A	61.66
Topology Loss (with `BCE`) [23]	50.35	50.32	50.34	33.63	56.29	49.00	76.69	75.76	76.22	61.58	64.95	56.91
Topology Loss (with `SoftIoU`) [23]	52.94	52.86	52.90	35.96	57.69	51.99	79.63	79.88	79.75	66.32	64.94	65.96
LinkNet34 [7]	61.30	61.45	61.39	44.27	60.33	55.69	78.34	78.85	78.59	64.73	62.75	65.33
LinkNet34 [7] + Orientation (Ours)	63.82	63.96	63.89	46.94	62.45	60.76	81.24	81.73	81.48	68.75	64.71	68.71
MatAN [20]	49.84	50.16	50.01	33.34	52.86	46.44	57.59	56.96	57.28	40.13	46.88	47.15
RoadCNN (segmentation) [3]	62.82	63.09	62.95	45.94	62.34	58.41	82.85	83.73	83.29	71.36	**67.61**	69.65
Ours (full)	**64.65**	**64.77**	**64.71**	**47.83**	**63.75**	**63.65**	**83.79**	**84.14**	**83.97**	**72.37**	67.21	**73.12**

Table 4: Comparison of our technique with the state-of-the-art road network extraction techniques. IoU^r and IoU^a refers to relaxed and accurate road IoU. Ours (full) include the proposed stacked multi-branch module with orientation learning. We use implementation from [3] for DeepRoadMapper [19] and our own implementation for [23].

modules enhance the representation during fusion (Figure 8) in a similar way as connectivity refinement iteratively enhance the predictions. The second multi-branch module inherently refine the road connectivity, which functions upon the fused feature space. In the end, joint learning and fusion improve the road IoU by $\sim 2.5\%$ and $APLS$ by $\sim 5\%$ on both datasets over the pixel-wise classification supervision.

Effect of fusion: We perform ablation study on fusion strategies to enable the information flow and report the results in Table 5. We discover that feature fusion by adding the orientation features with segmentation performs better. It shows that the simple feature addition improve the $APLS$ by 1.87% and 1.18% over the no fusion on both datasets.

Fusion	SpaceNet		DeepGlobe	
	IoU^a	$APLS$	IoU^a	$APLS$
No Fusion	63.44	61.78	66.81	72.03
Sum	**63.75**	**63.65**	**67.21**	**73.21**
Concatenate	63.53	63.01	66.59	72.23

Table 5: Effect of different fusion strategies in our proposed module to allow the information flow between orientation learning and segmentation tasks in the first stack.

Comparisons with state-of-the-art results: We compare the effectiveness of the proposed methods with state-of-art segmentation based methods [19], [20] and [23] (see Table 4 and Figure 9). Máttyus *et al*. [19] hypothesize the connections with shortest path algorithms between the nodes of road graph and validates the connection with a classifier. We found that the classifier is unable to detect the false connections in cases with densely connected roads which leads to a decrease in $APLS$ after post-processing. Mosinka *et al*. [23] introduce the topology loss term with recursive refinement. However, it also face challenges in predicting the roads in densely connected areas, and unpaved roads. Inspite of large diversity in both datasets, our approach significantly improves the connectivity in the extracted road graph against the baselines. However, the proposed technique faces challenges to accurately connect roads under the bridges as well as in the presence of large occlusion (see row #4 in Figure 9). We also observe the false road detection in farm outlines due to it's visual similarity with unpaved roads and parking lots on top of buildings due to the absence of relative depth cues. We show additional qualitative results in the supplementary material.

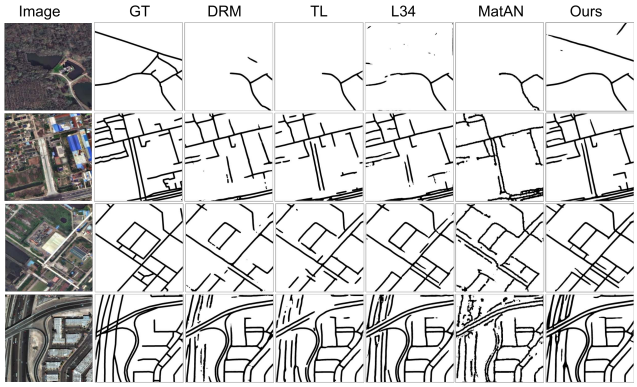

Figure 9: Qualitative Comparisons with state-of-the-art methods — DRM [19], TL [23], L34 [7], and MatAN [20].

6. Conclusion

In this paper, we propose a novel task of orientation learning that constrain the model to produce connected and topologically accurate road networks. We show that pixel-wise classification supervision leads to road networks with fragmented road segments and poor connectivity. Our experiments show that the joint learning of orientation and segmentation followed by connectivity refinement leads to a significant improvement in the road connectivity. We also show the effectiveness of the stacked encoder-decoder structure model as a joint learning module, which can efficiently utilize the information from related tasks.

Acknowledgement We thank NSERC for partially supporting the work.

References

[1] Min Bai and Raquel Urtasun. Deep watershed transform for instance segmentation. In *CVPR*, 2017. 3

[2] Meir Barzohar and David B Cooper. Automatic finding of main roads in aerial images by using geometric-stochastic models and estimation. *PAMI*, 1996. 1, 2

[3] Favyen Bastani, Songtao He, Sofiane Abbar, Mohammad Alizadeh, Hari Balakrishnan, Sanjay Chawla, Sam Madden, and David DeWitt. Roadtracer: Automatic extraction of road networks from aerial images. In *CVPR*, 2018. 1, 2, 4, 8

[4] R Caruna. Multitask learning: A knowledge-based source of inductive bias. In *ICML*, 1993. 3, 4

[5] Lluıs Castrejón, Kaustav Kundu, Raquel Urtasun, and Sanja Fidler. Annotating object instances with a polygon-rnn. In *CVPR*, 2017. 4

[6] Dengfeng Chai, Wolfgang Forstner, and Florent Lafarge. Recovering line-networks in images by junction-point processes. In *CVPR*, 2013. 1, 2

[7] Abhishek Chaurasia and Eugenio Culurciello. Linknet: Exploiting encoder representations for efficient semantic segmentation. In *VCIP*, 2017. 2, 6, 7, 8

[8] Guangliang Cheng, Ying Wang, Shibiao Xu, Hongzhen Wang, Shiming Xiang, and Chunhong Pan. Automatic road detection and centerline extraction via cascaded end-to-end convolutional neural network. *Transactions on Geoscience and Remote Sensing*, 2017. 1, 2

[9] Ronan Collobert and Jason Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *ICML*, 2008. 3

[10] Dragos Costea and Marius Leordeanu. Aerial image geolocalization from recognition and matching of roads and intersections. *arXiv preprint arXiv:1605.08323*, 2016. 1, 2

[11] Ilke Demir, Krzysztof Koperski, David Lindenbaum, Guan Pang, Jing Huang, Saikat Basu, Forest Hughes, Devis Tuia, and Ramesh Raskar. Deepglobe 2018: A challenge to parse the earth through satellite images. In *CVPRW*, 2018. 1, 2, 6

[12] Alexey Dosovitskiy and Thomas Brox. Generating images with perceptual similarity metrics based on deep networks. In *NIPS*, 2016. 2

[13] David H Douglas and Thomas K Peucker. Algorithms for the reduction of the number of points required to represent a digitized line or its caricature. *Cartographica: The International Journal for Geographic Information and Geovisualization*, 1973. 6

[14] Ruohan Gao and Kristen Grauman. On-demand learning for deep image restoration. In *CVPR*, 2017. 4

[15] Ruohan Gao, Bo Xiong, and Kristen Grauman. Im2flow: Motion hallucination from static images for action recognition. In *CVPR*, 2018. 6

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 6, 7

[17] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *CVPR*, 2018. 3, 4

[18] Ivan Laptev, Helmut Mayer, Tony Lindeberg, Wolfgang Eckstein, Carsten Steger, and Albert Baumgartner. Automatic extraction of roads from aerial images based on scale space and snakes. *MVA*, 2000. 1, 2

[19] Gellért Máttyus, Wenjie Luo, and Raquel Urtasun. Deeproadmapper: Extracting road topology from aerial images. In *ICCV*, 2017. 1, 2, 5, 8

[20] Gellért Máttyus and Raquel Urtasun. Matching adversarial networks. In *CVPR*, 2018. 2, 8

[21] Volodymyr Mnih and Geoffrey E Hinton. Learning to detect roads in high-resolution aerial images. In *ECCV*, 2010. 1, 2

[22] Volodymyr Mnih and Geoffrey E Hinton. Learning to label aerial images from noisy data. In *ICML*, 2012. 1, 2, 6

[23] Agata Mosinska, Pablo Márquez-Neila, Mateusz Kozinski, and Pascal Fua. Beyond the pixel-wise loss for topology-aware delineation. In *CVPR*, 2018. 1, 2, 7, 8

[24] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, 2016. 2

[25] Deepak Pathak, Philipp Krähenbühl, Jeff Donahue, Trevor Darrell, and Alexei Efros. Context encoders: Feature learning by inpainting. In *CVPR*, 2016. 4

[26] Urs Ramer. An iterative procedure for the polygonal approximation of plane curves. *Computer graphics and image processing*, 1972. 6

[27] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015. 2, 7

[28] Suriya Singh, Anil Batra, Guan Pang, Lorenzo Torresani, Saikat Basu, Manohar Paluri, and C. V. Jawahar. Self-supervised feature learning for semantic segmentation of overhead imagery. In *BMVC*, 2018. 1, 2, 4, 6

[29] Radu Stoica, Xavier Descombes, and Josiane Zerubia. A gibbs point process for road extraction from remotely sensed images. *IJCV*, 2004. 1, 2

[30] Adam Van Etten, Dave Lindenbaum, and Todd M Bacastow. Spacenet: A remote sensing dataset and challenge series. *arXiv preprint arXiv:1807.01232*, 2018. 1, 2, 5, 6

[31] Carles Ventura, Jordi Pont-Tuset, Sergi Caelles, Kevis-Kokitsi Maninis, and Luc Van Gool. Iterative deep learning for road topology extraction. In *BMVC*, 2018. 1, 2

[32] Jacob Walker, Abhinav Gupta, and Martial Hebert. Dense optical flow prediction from a static image. In *ICCV*, 2015. 6

[33] Jan D Wegner, Javier A Montoya-Zegarra, and Konrad Schindler. A higher-order crf model for road network extraction. In *CVPR*, 2013. 1, 2

[34] Yu Zhang and Qiang Yang. A survey on multi-task learning. *CoRR*, abs/1707.08114, 2017. 3

[35] Zhengxin Zhang, Qingjie Liu, and Yunhong Wang. Road extraction by deep residual u-net. *IGRSL*, 2018. 1

[36] Cao Zhe, Simon Tomas, Wei Shih-En, and Sheikh Yaser. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017. 3