Towards Reliable Evaluation of Algorithms for Road Network Reconstruction from Aerial Images

Leonardo Citraro, Mateusz Koziński, and Pascal Fua

Computer Vision Laboratory, École Polytechnique Fédérale de Lausanne (EPFL)
{firstname.lastname}@epfl.ch

Abstract. Existing connectivity-oriented performance measures rank road delineation algorithms inconsistently, which makes it difficult to decide which one is best for a given application. We show that these inconsistencies stem from design flaws that make the metrics insensitive to whole classes of errors. This insensitivity is undesirable in metrics intended for capturing overall general quality of road reconstructions. In particular, the scores do not reflect the time needed for a human to fix the errors, because each one has to be fixed individually. To provide more reliable evaluation, we design three new metrics that are sensitive to all classes of errors. This sensitivity makes them more consistent even though they use very different approaches to comparing ground-truth and reconstructed road networks. We use both synthetic and real data to demonstrate this and advocate the use of these corrected metrics as a tool to gauge future progress.

1 Introduction

Reconstruction of road networks from aerial images is an old computer vision problem. It has been tackled almost since the inception of the field in the 1970's [4,28,26,13]. Yet, it is still open and is addressed by many recent papers [23,22,9,21,19,5,25,10,6,24,33]. One pitfall however, is that the metrics used to gauge performance often prove to be inconsistent. It is not unusual for a method to perform well according to one popular metric and poorly according to another. Trusting to one single metric can therefore be misleading and can hamper progress.

This situation arises from the fact that assessing the quality of a road graph is hard. The quality assessment should not only depend on the spatial accuracy of the reconstructed road centerline but also on the topology of the network these centerlines form. The first is relatively easy to quantify while the second is much more difficult and there is no generally accepted way of doing it. This is because comparing the predicted topology to the ground truth one amounts to solving a complex graph-matching problem for which no efficient algorithm exists. Current topology-aware metrics are therefore hand-crafted to perform a simplified comparison, with some constraints of the full graph-matching problem

relaxed. They fall into three main categories: the metrics that compare the junctions, or road intersections, in the two graphs, ones that compare the lengths of the paths connecting random pairs of junctions, and ones that match small subgraphs. Unfortunately, as we will show, all these metrics correlate poorly with the number of topological discrepancies—missed road branches, unwarranted or missing connections—between the two graphs. This makes it hard to use them to reason about the number of mistakes present in a reconstruction and the cost of fixing them.

In this paper, we show that these inconsistencies arise from design flaws in currently popular metrics and propose ways to fix them. We incorporate these fixes into three topologically-aware metrics that capture a large range of errors and balance their contributions in the final score. We show that this makes them more consistent both with each other and with the number of topological discrepancies. Our contributions are therefore:

- An in-depth analysis of existing metrics that exposes their lack of sensitivity to certain types of errors and the resulting lack of consistency when using them to compare different algorithms.
- Three new measures free from this problem and that we advocate for future algorithm evaluation.

We make the code for computing our measures publically available. In the remainder of the paper, we first describe existing metrics and their shortcomings. We then introduce our new metrics and test them on synthetic and real data.

2 Existing Metrics

Road networks are often represented by graphs whose nodes have a double function. They serve as control points that enable modeling potentially curvy road segments and represent road intersections, and end points. Let us assume we are given a *predicted* graph and a *ground truth* graph, whose similarity we want to assess. Comparing these two networks that are similar, but not identical, is non-trivial. Doing this in a graph-theoretic way can be viewed as an NP-complete graph matching problem [30]. In this section, we review strategies commonly used to circumvent this problem.

One such approach is to first project the graphs to point clouds, by representing graph edges as sequences of closely spaced points, and then to evaluate the spatial overlap of these point clouds. Chamfer and Hausdorff measures may be employed to evaluate the sum and maximum of distances from each point in one set to its closest point in the second set. However, Chamfer and Hausdorff do not measure connectivity and are not sensitive to connectivity-oriented errors, like small gaps in roads, or predicting two closely spaced roads in place of a single real road. Such errors are common in road network reconstructions from aerial images, and that is why these metrics are typically not used to evaluate them.

Task-specific measures are used in some problems involving predicting graphs from images, but they do not generalize to the case of road networks. The DIADEM score [15], used for comparing reconstructions of neuronal morphologies,

relies on the assumption that both graphs have tree-like topologies. The Rand index [2,14], used for comparing segmentations of neuronal cells in microscopy images, compares the presence or absence of connections between pixel pairs in the prediction and annotation. In road lane reconstruction [3,16,17,20], the evaluation involves comparing the number of predicted lanes to the number of the annotated ones. These comparisons are not well suited for city-scale road networks, which form a single connected component with thousands of loops.

In this paper, we focus specifically on metrics for evaluating connectivity of road networks reconstructed from aerial images. Several metrics have been developed for this purpose, and they can be classified into four main categories, depending on whether they are pixel-based, junction-based, path-based, or subgraph-based. We now review these four classes of existing metrics and argue that they *all* ignore particular type of errors.

2.1 Pixel-Based Metrics

Road delineation can be understood as foreground/background segmentation problem. The quality of the segmentation can be evaluated in terms of the *precision* $P = \frac{|\mathrm{TP}|}{|\mathrm{PP}|}$ and *recall* $R = \frac{|\mathrm{TP}|}{|\mathrm{AP}|}$, where PP is the set of pixels predicted to be foreground, AP is the set of pixel labeled as foreground, and $\mathrm{TP} = \mathrm{PP} \cap \mathrm{AP}$. When a single number is preferred, either the *intersection-over-union* $\mathrm{IoU} = \frac{\mathrm{TP}}{\mathrm{PP} \cup \mathrm{AP}}$, or the *f1-score* $\mathrm{F1} = 2/(\frac{1}{P} + \frac{1}{R})$ is used.

*Correctness/Completeness/Quality (***CCQ***).* To account for the fact that the position of the pixels estimated to be foreground might be slightly off, the definitions of precision and recall were relaxed in [32,29] to allow small shifts in pixel location. *correctness* is the relaxed precision, *completeness* the relaxed recall, and *quality* is the equivalent of intersection over union.

Discussion. **CCQ** is adequate to gauge segmentation quality but does not capture connectivity of the predicted road maps. This makes it insensitive to topological errors, for example, road breaks smaller than the allowed shift between the annotation and the prediction. We demonstrate this insensitivity experimentally in section 4.1. The indifference of the pixel-based metrics to connectivity variations inspired the creation of the path- and junction-based metrics, described below.

2.2 Path-Based Metrics

The idea behind path-based metrics is that if two graphs are similar, so should paths connecting any pair of their nodes via a sequence of edges. Edges that appear in one graph and not in the other result in measurably different paths. There are two main ways to measure such differences.

*Too Long / Too Short (***TLTS***).* In [31], it was proposed to compare the length of the shortest path between randomly-chosen but corresponding pairs of nodes in the predicted and ground-truth networks. Here, and in the metrics that we describe below, the correspondences are found simply by randomly selecting a

point in one graph and taking its closest point from the other graph. A path in the predicted graph is classified as *correct* if its length is within 5% of that of the path in the ground-truth graph, and as *too-long*, or *too-short* otherwise. A path is marked *infeasible* if its end points are not connected in the other network. The percentage of *correct* paths is used to assess the quality of a delineation and the other percentages serve to characterize the errors.

*Average Path Length Similarity (***APLS***).* The alternative to counting too long/short paths is aggregating path differences. This has been proposed first for evaluating road network reconstructions from GPS tracks [18,1] and, more recently, for image-based reconstructions [12], in the form of the Average Path Length Similarity score

$$1 - \frac{1}{|\mathcal{P}|} \sum_{(p^{\text{gt}}, p^{\text{est}}) \in \mathcal{P}} \min \left\{ 1, \frac{|l(p^{\text{gt}}) - l(p^{\text{est}})|}{l(p^{\text{gt}})} \right\} , \tag{1}$$

where p^{gt} is a ground-truth path, p^{est} the corresponding predicted one and $l(.)$ denotes the path length. The set $\mathcal{P}$ is obtained by sampling pairs of points in one graph, retrieving the corresponding pairs in the other graph, and computing the shortest paths between them.

Discussion. **TLTS** and **APLS** are better at capturing topological differences than pixel-based scores, but they suffer from major flaws. Since both metrics rely on the comparison of the length of shortest paths, false positive roads that do not alter the length of these are completely neglected. Moreover, since paths are sampled independently, multiple paths from a graph can be compared to a single path in the other graph. This makes the scores insensitive to the errors made by predicting one road where many closely spaced roads exist and predicting more than one road where there is just one, as shown in Fig. 1.

2.3 Junction-Based Metric (JUNCT)

The path-based metrics capture the topological similarity indirectly. A more direct approach [5], is to compare the degree of corresponding nodes with at least three incident edges, called junctions. The correspondences are established greedily by matching closest nodes. For each ground-truth junction v that is matched to a predicted junction u, the per-junction recall $f_{v,\text{correct}}$ is taken to be the fraction of edges incident on v that are also captured around u. Similarly, the false discovery rate — one minus precision — $f_{u,\text{error}}$ is taken to be the fraction of edges incident on u that do not appear around v. For unmatched junctions, $f_{v,\text{correct}} = 0$ and $f_{u,\text{error}} = 1$, respectively. These per-junction scores are then aggregated

$$n_{\text{correct}} = \sum_{v \in \mathcal{V}} f_{v,\text{correct}} , \qquad n_{\text{error}} = \sum_{u \in \mathcal{U}} f_{u,\text{error}} ,$$

$$F_{\text{correct}} = \frac{n_{\text{correct}}}{|\mathcal{V}|} , \qquad \text{and} \qquad F_{\text{error}} = \frac{n_{\text{error}}}{n_{\text{error}} + n_{\text{correct}}} ,$$

where $|\mathcal{V}|$ is the number of ground-truth junctions.

Discussion The main issue with **JUNCT** is that it only accounts for nodes with three or more incident edges. This disregards what happens at road end points and makes the metric insensitive to interruptions in predicted networks. In other words, a predicted network where all the roads are broken in the middle still receives a perfect score. Moreover, a node that lacks $k-2$ out of its k incident edges is penalized more than any other, because it is no longer considered a junction. This amounts to saying that a road junction with only two correctly predicted incident roads is completely misclassified, a conclusion that is hard to justify. The top of Fig. 2 illustrates this problem: An edge is missing from a junction with three incident edges, which results in $n_{\text{correct}} = \frac{0}{3}$ instead of $\frac{2}{3}$.

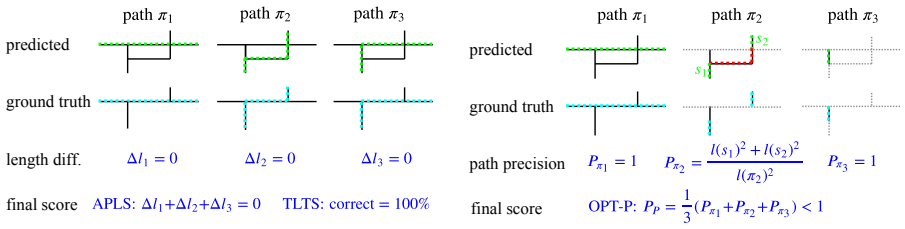
2.4 Subgraph-Based Metric (SUBG)

In [7], it is suggested to compare the sets of locations accessible by traveling a predefined distance away from corresponding points in two graphs. To this end, a starting location is randomly selected in the ground truth network, and its closest point in the predicted network is identified. Then, local subgraphs are extracted by a breadth-first exploration of the graphs away from the starting locations. The computation of the score is based on spatial coincidence of 'control points' inserted at regular intervals to the subgraphs. A control point is considered to be matched, or a true positive, if it lies sufficiently close to a control point in the other network. Unmatched control points in the predicted, and annotated subgraphs are treated as false positives and false negatives, respectively. Sampling and matching of local subgraphs is repeated many times, and precision and recall are computed from the total counts of matched and unmatched control points.

Discussion As the starting point is always sampled from the ground truth network, the false positive roads that are sufficiently far from any ground truth road are not covered by control points. In consequence, **SUBG** is not sensitive to such errors. Moreover, since multiple control points of the ground truth network can be matched to the same control point in the prediction, errors consisting in predicting just one instead of two closely spaced roads go unnoticed.

2.5 Summary

One might imagine that the metrics were deliberately designed to expose some errors and suppress the others in the interest of some specific application. However, no trace of such intentions can be found in the original publications. All the studied metrics were proposed for general-purpose evaluation of road network reconstructions, their insensitivity was never reported before, and seems to be an artifact of relaxing the underlying graph-matching problem to independent comparison of junctions, paths, and subgraphs. In many cases, the insensitivity is not immediately obvious from the design of the metric alone, and in section 4.1 we propose a benchmark dataset for exposing it. In section 3, we show that these flaws can be removed by careful metric design.



(a) The existing path length statistics **TLTS**, **APLS**. Both sampled paths (green) and their matches (cyan) overlap. As a result, the scores do not capture the difference between the networks.

(b) Our new path-based score (**OPT-P**). Paths do not overlap and the score captures the difference between the networks.

Fig. 1. A comparison of (a) the existing path-based statistics and (b) our new path score. Three paths are sampled from the predicted network (overlayed in green), and matched to the ground truth network (the matching chains are highlighted in cyan). The unmatched parts of the paths are highlighted in red. Removed parts of the networks are shown in dotted gray. See section 2.2 for the definition of the **APLS** and **TLTS** and section 3.1 for **OPT-P**.

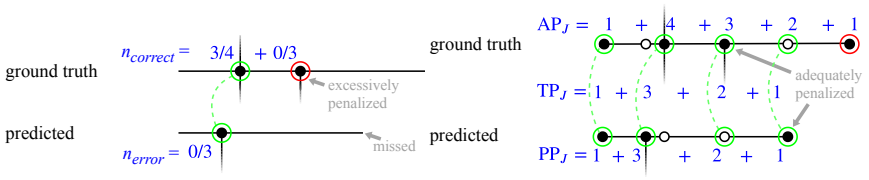
3 New Metrics

In Section 2, we identified weaknesses of existing metrics that make them insensitive to whole classes of errors, such as producing unwarranted breaks and spurious roads, or merging parallel but distinct roads. Here, we introduce new metrics that are sensitive to *all* these errors.

3.1 Path-Based Metric (OPT-P)

In section 2.2 we argued that **TLTS** [31] and **APLS** [12] are insensitive to false positive predictions that do not affect the length of the shortest paths, for example, ones that run close to other predicted roads. We illustrate such a case in the left part of Fig. 1. We therefore introduce a new path-based metric **OPT-P**, not affected by this insensitivity. It involves computing R_P, which can loosely be interpreted as path recall, and P_P, which plays the role of path precision. In contrast to earlier metrics, we do not sample or match paths independently. Instead, we ensure that no two paths sampled from a graph share the same edges, and that any two sampled paths are matched to two disjoint sets of edges in the other graph. Moreover, when matching a pair of paths, we not only compare their lengths, as done in existing metrics, but also ensure that their trajectories run close in the image. This makes P_P sensitive to the types of false positive road predictions that the **TLTS** and **APLS** miss.

More precisely, we developed an iterative path sampling and matching scheme. To compute recall, we iteratively sample a path from the ground truth network and match it to the predicted network. Using the match, we compute our measure of connectivity as described in the next paragraph. We then remove the sampled path from the ground-truth network to ensure that no two paths share the same edges. We also remove the matched edges from the predicted network



(a) the existing junction score **JUNCT** (b) the new junction score **OPT-J**

Fig. 2. A comparison of the existing junction score (a) to our junction score (b). Feature points are marked as black dots, matches in green and unmatched features in red. For readability, we only consider the features on the horizontal lines, and assume the vertical lines continue indefinitely. Candidate and actual edge matches are depicted by hollow nodes. Unmatched hollow nodes are not penalized. See section 2.3 for the definition of **JUNCT** and section 3.2 for the definition **OPT-J**.

in order to guarantee that no edge from the predicted network is matched to two different paths. We iterate until one of the networks has no more edges. Fig. 1 illustrates this process. Precision is computed similarly, but the roles of the networks are exchanged.

Matching a path π to a graph is performed using the Viterbi algorithm, and more details of this procedure can be found in the appendix. If possible, the path is projected to a chain of connected graph nodes. If the whole path cannot be matched to a single chain due to disconnections in the predicted graph, its sub-paths are still matched to connected chains whenever possible. This matching induces a partitioning of the path π into a set of segments $\mathcal{S}(\pi)$, such that each $s \in \mathcal{S}(\pi)$ maps to a different chain. If the path π exists in the graph and has no disconnections, $\mathcal{S}(\pi)$ contains only one segment. In case of disconnections $|\mathcal{S}(\pi)| > 1$ and $|\mathcal{S}(\pi)| = 0$ if π does not exist in the graph.

To compute P_P and R_P, we use the matched segments $\mathcal{S}(\pi)$ to estimate the probability that a sub-path of π, with end points selected randomly and with uniform probability along π, is connected in the target network. The sub-path is connected in the target network only if its both end-points lie on the same path segment s. The probability of such event is $\mathrm{P}_\pi = \frac{\sum_{s \in \mathcal{S}} l(s)^2}{l(\pi)^2}$, where $l(.)$ denotes path length. Note that, if the matched path is entirely connected, then $|\mathcal{S}(\pi)| = 1$, and this probability is $\mathrm{P}_\pi = 1$. When the matched path has disconnections, $|\mathcal{S}(\pi)| > 1$ and $\mathrm{P}_\pi < 1$. We define path recall as the average of these probabilities over all paths $\pi \in \Pi$ sampled from the ground truth network $R_P = \frac{1}{|\Pi|} \sum_{\pi \in \Pi} \mathrm{P}_\pi$. The path precision P_P is computed according to the same formula, but with paths sampled from the predicted network and matched to the ground truth one.

3.2 Junction-Based Metric (OPT-J)

As discussed in section 2.3, the junction score **JUNCT** [5] is insensitive to road interruptions and excessively penalizes junctions that lack $k-2$ out of k incident edges. To address these shortcomings, we propose a new junction score

OPT − J, including the junction precision P_J and recall R_J. As for **JUNCT**, computing **OPT − J** involves matching feature nodes in the two networks and comparing the numbers of edges incident on them. Unlike in **JUNCT**, where the set of features comprises only junctions – nodes with at least three incident edges – we use both junctions and endpoints as features. Moreover, we enable matching a feature of one graph not only to a feature of the other graph, but also to any point on its edge. This gives our metric the desired sensitivity to unwarranted road breaks and prevents excessively penalizing specific patterns of missing edges. Fig. 2 illustrates the differences between **JUNCT** and **OPT − J**.

We denote a match by (i, j), where i belongs to the ground truth and j to the predicted graph, and both are features, or one of them is a feature, and the other is its closest point on an edge. We perform greedy matching with the cost of a match $c_{ij} = \alpha d_{ij} + |o_i - o_j|$, where o_i is the number of edges incident on i if i is a feature and by convention $o_i = 2$ if i is a point on an edge. d_{ij} is the distance between i and j. α is a parameter of the score. We only allow a feature to be matched once, but we do not constrain the number of features matched to a single edge. We only consider matches (i, j) such that i and j are within a predefined small distance $d^{\max}$.

We denote the set of matches M, the sets of unmatched ground truth features by F^-_{gt} and the set of unmatched predicted features by F^-_{est}. We estimate the number of true positive incident edges as $\text{TP}_J = \sum_{(i,j)\in M} \min\{o_i, o_j\}$, the number of predicted edges as $\text{PP}_J = \sum_{(i,j)\in M} o_j + \sum_{j\in F^-_{\text{est}}} o_j$, and the number of ground truth edges as $\text{AP}_J = \sum_{(i,j)\in M} o_i + \sum_{i\in F^-_{\text{gt}}} o_i$. We compute the precision and recall as $P_J = \frac{\text{TP}_J}{\text{PP}_J}$ and $R_J = \frac{\text{TP}_J}{\text{AP}_J}$.

3.3 Subgraph-Based Metric (OPT-G)

In section 2.4 we have exposed the lack of sensitivity of the local graph comparison **SUBG** [7] to false positive roads that are far away from ground truth roads and to errors involving missing one of several closely spaced roads. To remove this lack of sensitivity, we propose a new score **OPT-G**. Like **SUBG**, **OPT-G** is based on comparing sets of graph locations accessible by traveling a short distance in the graph from a randomly selected starting point. To prevent distinct roads that parallel each other closely from being matched to a single one in the other graph, we force the matching to be one-to-one. In addition, unlike in the old score, we sample the starting points both in the ground-truth and predicted graphs, which makes the score sensitive to false positive, as shown in Fig. 3.

To compute the score, we iteratively sample a starting point in one of the graphs. We then find its closest point in the other graph. Using breadth-first graph traversal, we crop out subgraphs accessible by traveling a predefined distance from the starting points. Control points are inserted at equal intervals during the traversal. We then perform a one-to-one matching of control points from the two graphs by the Hungarian algorithm, with the cost of matching two points equal to the Euclidean distance between them. Only points within a predefined distance are matched. Calculation of the score is based on the number

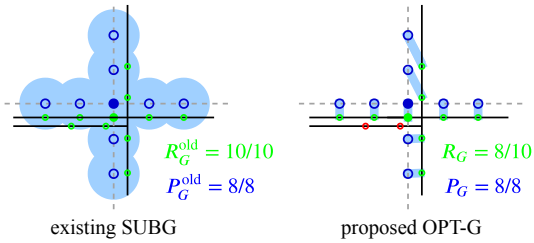


Fig. 3. The difference between the existing subgraph-based score **SUBG** and our subgraph-based score **OPT-G**. A single starting point is shown in both the predicted network (as filled blue circle) and the ground truth networks (as a filled green circle). The networks are drawn with dashed gray and solid black lines, respectively. All the control points in the ground truth network (hollow green circles) are within a matching distance (visualized with light blue disks) from the control points in the predicted network (hollow blue circles). This makes the existing score insensitive to the missing road. In our score, the matching is one-to-one (visualized by light blue lines). In result, some of the control points remain unmatched (marked in red), which gives the score sensitivity to the missing road.

of matched and unmatched control points. We define subgraph-based precision as $P_G = \frac{\text{TP}_G}{\text{PP}_G}$ and subgraph-based recall as $R_G = \frac{\text{TP}_G}{\text{AP}_G}$, where TP_G is the total number of matched control points, PP_G is the number of control points in the predicted graph and AP_G is the number of control points in the ground truth graph.

4 Experiments

In this section, we first use synthetic data to compare the behavior of the current and new performance metrics. We will show that current ones are insensitive to certain types of errors, while ours capture all of them. We then evaluate all the metrics on real data and use them to compare state-of-the-art road reconstruction methods.

4.1 Synthetic Data

We created a synthetic benchmark dataset from crops of road networks from [5]. Its purpose is to enable the analysis of the responses of the metrics to varying numbers of errors of a single type, for different types of errors.

We formed the dataset by duplicating the selected crops to emulate pairs of 'ground truth' and 'predicted' networks. We then perturbed the graphs by introducing a controlled number of errors that are representative of those encountered in practice.

- Interruptions: Unwarranted breaks in roads.
- Overconnections: Spurious additional roads connecting randomly selected pairs of points.

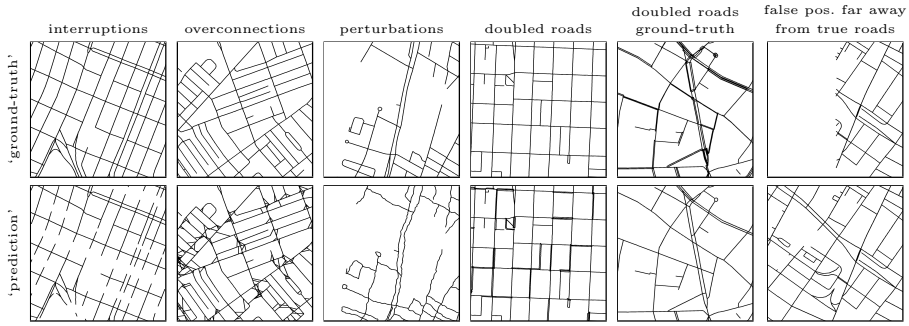


Fig. 4. Example pairs of road networks from the benchmark dataset.

- Perturbations: displacing graph nodes from their true locations without disconnecting the roads.
- Doubled roads: Spurious copies of road segments shifted slightly and connected to the originals.
- Doubled roads-ground truth: Same as above, but the copies are added to the ground-truth, to emulate roads missing from the prediction.
- False positives far away from true roads: To simulate them, we removed part of the ground truth while keeping the prediction unchanged.

Fig. 4 depicts example graphs from our dataset. As shown in Fig. 7, similar errors appear in real reconstructions. In Fig. 5, we plot the behavior of all the metrics as a function of the number of perturbations. If a metric is sensitive to a particular kind of error, the curve will exhibit a large slope. By contrast, if the metric is insensitive to that kind of error, the curve will be flat. Note that the curves for our new metrics are never flat, which indicates that are adequately sensitive to all the kinds of errors listed above. Unfortunately, this cannot be said of the existing metrics.

4.2 Real Data

We now turn to the recent road delineation algorithms, and analyze their predictions for the publicly available Roadtracer [5] and DeepGlobe [11] datasets. To do so we used publicly available algorithms implementations, and ones whose authors kindly shared with us the delineation results:

- *Segmentation*. Segmentation-based approach where the output probability map is thresholded and skeletonized. We use the prediction provided in [5] for the Roadtracer dataset and our own implementation of UNet [27] for DeepGlobe.
- *RoadTracer*. Iterative graph construction where node locations are selected by a CNN [5].
- *Seg-Path*. Unified approach to segmenting linear structures and classifying potential connections. [24]
- *DeepRoad*. Image segmentation followed by post-processing focused at fixing missing connections [21].

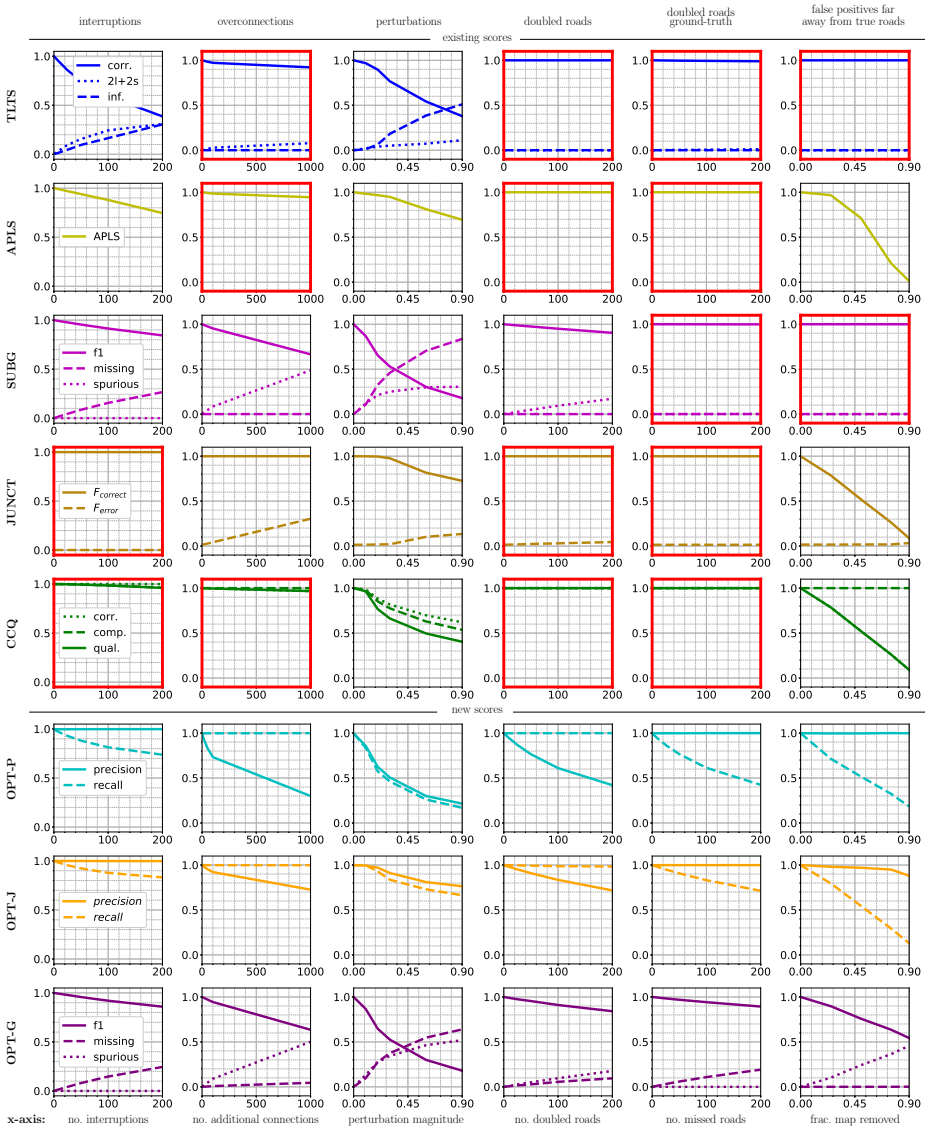


Fig. 5. Sensitivity of the existing and the new scores to different types of errors. The plots that exhibit lack of sensitivity are outlined in red. Our proposed metrics do not exhibit any of such insensitivity. See section 4.1 for details. Best viewed in color.

- *RCNNUNet*. Recursive image segmentation with post-processing for graph extraction [33].
- *MultiBranch*. A recursive architecture co-trained in road segmentation and orientation estimation [6].
- *LinkNet*. An encoder-decoder architecture [8] co-trained in segmentation and orientation estimation [6].

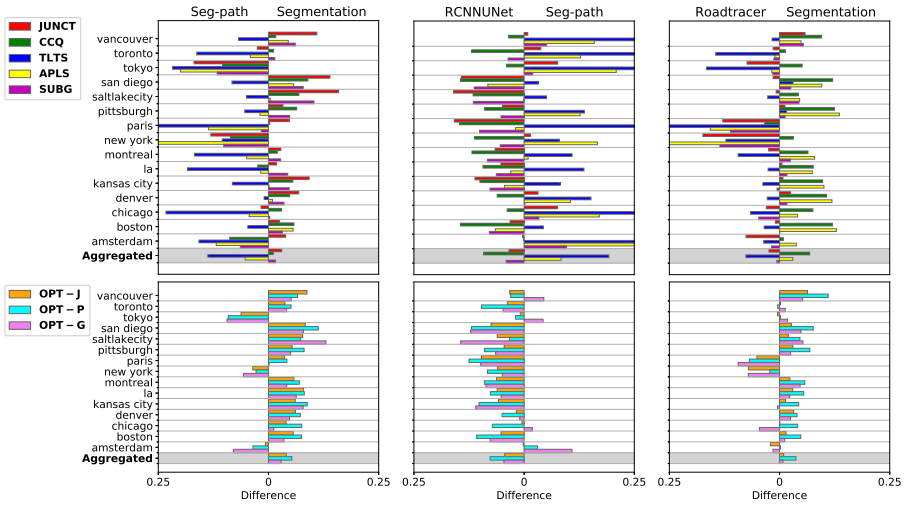


Fig. 6. This work was inspired by the observation that existing metrics are inconsistent. We compared road reconstruction algorithms in pairs, by visualizing the differences of their scores for each city of the *RoadTracer* test set. Bars extending to the right express preference for the first delineation method and ones extending to the left for the second. The results produced by our metrics (bottom) are far more consistent than ones produced by existing metrics (top). While, for a single pair of methods, it is possible to pick a triplet of existing metrics that give roughly consistent results, we show in Fig. 8 that their correlation is much weaker than for our scores.

Comparing Reconstruction Methods. A performance measure is meant to be used to compare methods and, ultimately, to decide which one is best. In this section, we show that this is difficult to do using existing metrics because they tend to return different rankings. By contrast, ours are far more consistent. To show this, we performed the following experiment.

For each individual metric, we computed the differences of the scores it returns for two different delineation methods in a specific city of the RoadTracer dataset. These differences are depicted in Fig. 6 by colored bars that extend to the left when the score of the first method is higher than that of the second and to the right otherwise. It is almost impossible to discern a clear pattern at the top of the figure where we plotted the bars corresponding to the existing metrics. By contrast, our metrics deliver a far clearer picture.

A possible interpretation of this result is that some type of problems are encountered more often in specific cities. As the existing metrics are more sensitive to some kinds of errors than others, that would explain the discrepancies. This would not necessarily be an issue if the metrics were designed to measure different, possibly uncorrelated, qualities of interest. However, they are typically used as overall quality measures and to demonstrate the advantage of one method over the others. In this context their inconsistency *is* an issue, and the greater consistency of our proposed measures is a distinct advantage.

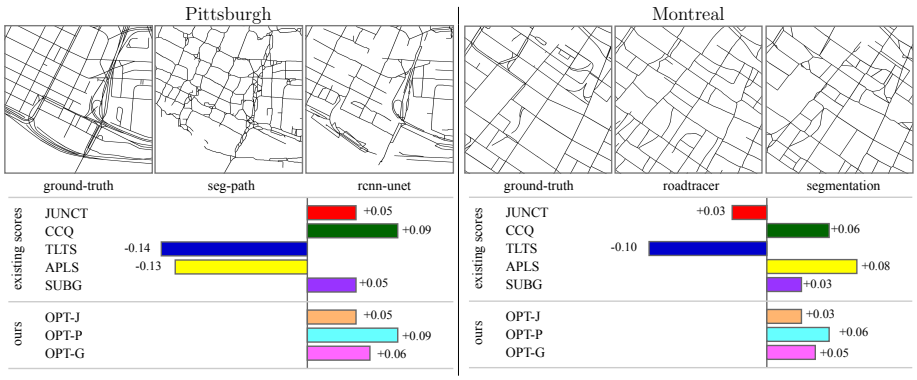


Fig. 7. Crops of a road network of (*left*) Pittsburgh and (*right*) Montreal and their reconstructions from aerial images. *Left*: Predictions by *seg-path* [24] and *rcnn-unet* [33]. *Right*: Predictions by *roadtracer* and *segmentation* both provided by [5] *Bottom*: Differences of the metrics for the two reconstructions.

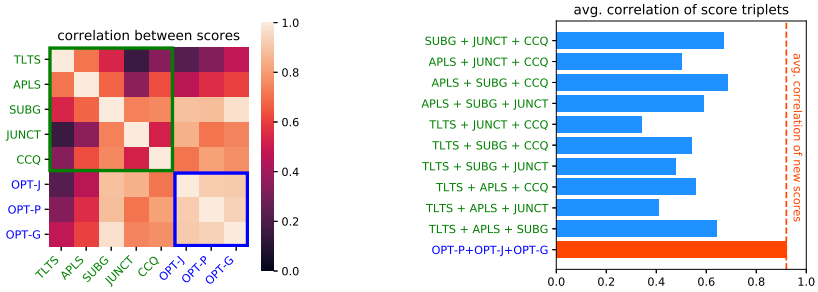


Fig. 8. Analysis of the correlations. *Left*: A matrix of correlations of the scores computed for the maps reconstructed by different methods on the roadtracer dataset. The correlation coefficients of the old scores are outlined in green, the correlation coefficients of the new scores in blue. *Right*: The average correlation of all possible existing score triplets (blue bars) against the average correlation of the three new scores (dashed red line). Our metrics show better correlation than the existing ones.

In Fig. 7, we visualize fragments of two predicted networks. The scores on the bottom of the figure show the comparison by the existing scores is inconclusive. As we have seen in the benchmark experiment, **APLS** and **TLTS** are insensitive to overconnections therefore they positively react to overconnected graphs as *seg-path* predictions. On the other hand, **OPT-P**, which is based on the same principle, takes spurious roads into account and agree with the other metrics. In term of cost of fixing the errors, an operator would spend more time removing spurious roads from *seg-path* than by adding the missing connection in *rcnn-unet*. More visual comparisons are provided in the supplementary material.

Correlation Analysis. In contrast to the consistency analysis we just discussed we now turn to compare correlations between the various metrics. On the left side of Fig. 8 we show the correlations between pairs of scores on the RoadTracer dataset

		existing scores										new scores						
		CCQ			**TLTS**		**APLS**	**JUNCT**			**SUBG**	**OPT-P**			**OPT-J**			**OPT-G**
	dataset	corr.	comp.	qual.	corr.	2l+2s		F_{cor}	F_{err}	f1	f1	pre.	rec.	f1	pre.	rec.	f1	f1
RoadTracer	Roadtracer [5]	0.681	0.615	0.478	0.422	0.174	0.595	0.763	0.129	**0.812**	0.710	0.597	0.458	0.519	0.801	0.758	0.779	0.683
	Segmentation [5]	0.775	0.649	0.546	0.346	0.190	0.625	0.728	0.137	0.782	0.704	0.645	0.488	0.556	0.820	0.760	0.788	0.690
	Seg-path [24]	0.647	**0.755**	0.535	**0.483**	**0.137**	**0.678**	**0.925**	0.355	0.754	0.688	0.443	**0.581**	0.503	0.649	**0.882**	0.748	0.662
	DeepRoadMapper [21]	**0.842**	0.474	0.435	0.071	0.235	0.243	0.422	0.203	0.514	0.477	0.651	0.271	0.383	0.822	0.524	0.640	0.482
	RCNN-Unet [33]	0.830	0.719	**0.626**	0.291	0.353	0.594	0.723	**0.120**	0.790	**0.729**	**0.672**	0.510	**0.580**	**0.827**	0.759	**0.792**	**0.707**
DGlobe	*LinkNet* [6]	0.778	0.803	0.653	0.632	0.107	0.660	0.682	0.234	0.722	0.819	0.727	0.761	0.744	0.782	0.793	0.787	0.789
	MultiBranch [6]	**0.804**	0.826	**0.687**	0.684	**0.101**	**0.699**	0.734	**0.185**	**0.773**	**0.843**	**0.740**	0.792	**0.765**	**0.805**	0.810	**0.807**	**0.813**
	Segmentation [27]	0.545	**0.841**	0.495	**0.720**	0.138	0.618	**0.941**	0.542	0.616	0.787	0.520	**0.859**	0.648	0.582	**0.872**	0.698	0.724

Table 1. Values of the existing and the new scores computed for road networks reconstructions by different methods on the RoadTracer and DeepGlobe datasets. Our scores rank the methods much more consistently.

and on the right side we compare the average correlation of all possible triplets of existing metrics to that of our three new scores. To evaluate the consistency of a score triplet, we average correlations for all pairs within the triplet. Either way, it can be seen that our metrics are far more correlated among themselves than any pair of triplet of the other ones.

Comparing State-of-the-Art Methods. We now use both existing and new metrics to compare state-of-the-art road reconstruction methods. As can be seen in Table 1, on the RoadTracer dataset, existing metrics favor *RoadTracer*, *RCNNUNet*, or *Seg-Path*. By contrast, the proposed metrics consistently point to *RCNNUNet*. Moreover, all of them rank *Segmentation* second and *Seg-Path* and *RoadTracer* compete for the third place with very similar scores in all our metrics. As seen in the bottom part of Table 1 this also holds for the DeepGlobe data. The existing scores are less inconsistent than for the RoadTracer dataset, with **TLTS** favoring segmentation while other scores prefer *MultiBranch*, but our metrics all agree on *MultiBranch*. Note also that precision- and recall-related scores for **OPT-J** and **OPT-P** show recurring patterns whereas **CCQ** and **JUNCT** do not.

5 Conclusion

We were surprised to discover that *all* the existing scores for evaluation of road network reconstructions suffer from design faults that make them insensitive to particular types of errors. Our experiments show that the concerns this rises about the reliability of evaluation by means of these scores are justifiable – one could overturn the results of a study by carefully selecting the score used for evaluation. We have demonstrated that correcting the flaws of the existing metrics leads to improved consistency – our three new metrics are much more coherent than the old ones, despite the fact that each of them is computed in a different way.

We have focused on road network reconstructions, but the proposed scores can be used for comparing any curvilinear networks.

Acknowledgments. This work was supported in part by the Swiss National Science Foundation.

References

1. Ahmed, M., Fasy, B., Hickmann, K., Wenk, C.: A Path-Based Distance for Street Map Comparison. ACM Trans. Spatial Algorithms Syst. **1**(1), 31–328 (July 2015)
2. Arganda-Carreras, I., Turaga, S., Berger, D., Cireşan, D., Giusti, A., Gambardella, L., Schmidhuber, J., Laptev, D., Dwivedi, S., Buhmann, J., Liu, T., Seyedhosseini, M., Tasdizen, T., Kamentsky, L., Burget, R., Uher, V., Tan, X., Sun, C., Pham, T., Bas, E., Uzunbas, M., Cardona, A., Schindelin, J., Seung, S.: Crowdsourcing the Creation of Image Segmentation Algorithms for Connectomics. Frontiers in Neuroanatomy p. 142 (2015)
3. Bai, M., Máttyus, G., Homayounfar, N., S, W., Lakshmikanth, S., Urtasun, R.: Deep Multi-Sensor Lane Detection. CoRR **abs/1905.01555** (2019)
4. Bajcsy, R., Tavakoli, M.: Computer Recognition of Roads from Satellite Pictures. IEEE Transactions on Systems, Man, and Cybernetics **SMC-6**(9), 623–637 (1976)
5. Bastani, F., He, S., Alizadeh, M., Balakrishnan, H., Madden, S., Chawla, S., Abbar, S., Dewitt, D.: Roadtracer: Automatic Extraction of Road Networks from Aerial Images. In: Conference on Computer Vision and Pattern Recognition (2018)
6. Batra, A., Singh, S., Pang, G., Basu, S., Jawahar, C., Paluri, M.: Improved Road Connectivity by Joint Learning of Orientation and Segmentation. In: Conference on Computer Vision and Pattern Recognition (June 2019)
7. Biagioni, J., Eriksson, J.: Inferring Road Maps from Global Positioning System Traces: Survey and Comparative Evaluation. Transportation Research Record: Journal of the Transportation Research Board **2291**, 61–71 (12 2012)
8. Chaurasia, A., Culurciello, E.: Linknet: Exploiting Encoder Representations for Efficient Semantic Segmentation. CoRR **abs/1707.03718** (2017)
9. Cheng, G., Wang, Y., Xu, S., Wang, H., Xiang, S., Pan, C.: Automatic Road Detection and Centerline Extraction via Cascaded End-To-End Convolutional Neural Network. IEEE Trans. Geoscience and Remote Sensing **55**(6), 3322–3337 (2017)
10. Chu, H., Li, D., Acuna, D., Kar, A., Shugrina, M., Wei, X., Liu, M., Torralba, A., Fidler, S.: Neural Turtle Graphics for Modeling City Road Layouts. In: International Conference on Computer Vision (2019)
11. Demir, I., Koperski, K., Lindenbaum, D., Pang, G., Huang, J., Basu, S., Hughes, F., Tuia, D., Raskar, R.: Deepglobe 2018: A Challenge to Parse the Earth through Satellite Images. In: Conference on Computer Vision and Pattern Recognition (June 2018)
12. Etten, A.V., Lindenbaum, D., Bacastow, T.: Spacenet: A Remote Sensing Dataset and Challenge Series. CoRR **abs/1807.01232** (2018)
13. Fischler, M., Tenenbaum, J., Wolf, H.: Detection of Roads and Linear Structures in Low-Resolution Aerial Imagery Using a Multisource Knowledge Integration Technique. Computer Vision, Graphics, and Image Processing **15**(3), 201–223 (March 1981)
14. Funke, J., Tschopp, F.D., Grisaitis, W., Sheridan, A., Singh, C., Saalfeld, S., Turaga, S.C.: Large Scale Image Segmentation with Structured Loss Based Deep Learning for Connectome Reconstruction. IEEE Transactions on Pattern Analysis and Machine Intelligence **41**(7), 1669–1680 (2018)
15. Gillette, T., Brown, K., Ascoli, G.: The Diadem Metric: Comparing Multiple Reconstructions of the Same Neuron. Neuroinformatics **9**, 233–245 (May 2011)
16. Homayounfar, N., Ma, W., Lakshmikanth, S., Urtasun, R.: Hierarchical Recurrent Attention Networks for Structured Online Maps. In: Conference on Computer Vision and Pattern Recognition. pp. 3417–3426 (2018)

17. Homayounfar, N., Ma, W., Liang, J., Wu, X., Fan, J., Urtasun, R.: Dagmapper: Learning to Map by Discovering Lane Topology. In: International Conference on Computer Vision (October 2019)
18. Karagiorgou, S., Pfoser, D.: On Vehicle Tracking Data-Based Road Network Generation. In: Proceedings of the 20th International Conference on Advances in Geographic Information Systems. pp. 89–98 (2012)
19. Li, Y., Zhang, X., Chen, D.: CSRNet: Dilated Convolutional Neural Networks for Understanding the Highly Congested Scenes. In: Conference on Computer Vision and Pattern Recognition (2018)
20. Liang, J., Homayounfar, N., Ma, W., and1 Urtasun, S.W.R.: Convolutional Recurrent Network for Road Boundary Extraction. In: Conference on Computer Vision and Pattern Recognition. pp. 9512–9521 (2019)
21. Máttyus, G., Luo, W., Urtasun, R.: Deeproadmapper: Extracting Road Topology from Aerial Images. In: International Conference on Computer Vision. pp. 3458–3466 (2017)
22. Mnih, V.: Machine Learning for Aerial Image Labeling. Ph.D. thesis, University of Toronto (2013)
23. Mnih, V., Hinton, G.: Learning to Detect Roads in High-Resolution Aerial Images. In: European Conference on Computer Vision. pp. 210–223 (2010)
24. Mosińska, A., Kozinski, M., Fua, P.: Joint Segmentation and Path Classification of Curvilinear Structures. IEEE Transactions on Pattern Analysis and Machine Intelligence **42**(6), 1515–1521 (2020)
25. Mosińska, A., Marquez-Neila, P., Kozinski, M., Fua, P.: Beyond the Pixel-Wise Loss for Topology-Aware Delineation. In: Conference on Computer Vision and Pattern Recognition. pp. 3136–3145 (2018)
26. Quam, L.: Road Tracking and Anomaly Detection. In: DARPA Image Understanding Workshop. pp. 51–55 (May 1978)
27. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Conference on Medical Image Computing and Computer Assisted Intervention. pp. 234–241 (2015)
28. Vanderbrug, G.: Line Detection in Satellite Imagery. IEEE Transactions on Geoscience Electronics **14**(1), 37–44 (January 1976)
29. Wang, S., Bai, M., Máttyus, G., Chu, H., Luo, W., Yang, B., Liang, J., Cheverie, J., Fidler, S., Urtasun, R.: Torontocity: Seeing the World with a Million Eyes. In: International Conference on Computer Vision. pp. 3028–3036 (2017)
30. Wegener, I., Pruim, R.: Complexity Theory: Exploring the Limits of Efficient Algorithms. Springer-Verlag (2005)
31. Wegner, J., Montoya-Zegarra, J., Schindler, K.: A Higher-Order CRF Model for Road Network Extraction. In: Conference on Computer Vision and Pattern Recognition. pp. 1698–1705 (2013)
32. Wiedemann, C., Heipke, C., Mayer, H., Jamet, O.: Empirical Evaluation of Automatically Extracted Road Axes. In: Empirical Evaluation Techniques in Computer Vision. pp. 172–187 (1998)
33. Yang, X., Li, X., Ye, Y., Lau, R.Y.K., Zhang, X., Huang, X.: Road Detection and Centerline Extraction via Deep Recurrent Convolutional Neural Network U-Net. IEEE Transactions on Geoscience and Remote Sensing pp. 1–12 (2019)