Computer Vision for International Border Legibility

Trevor Ortega[1], Thomas Nelson[1], Skyler Crane[1], Josh Myers-Dean[2] and Scott Wehrwein[1]

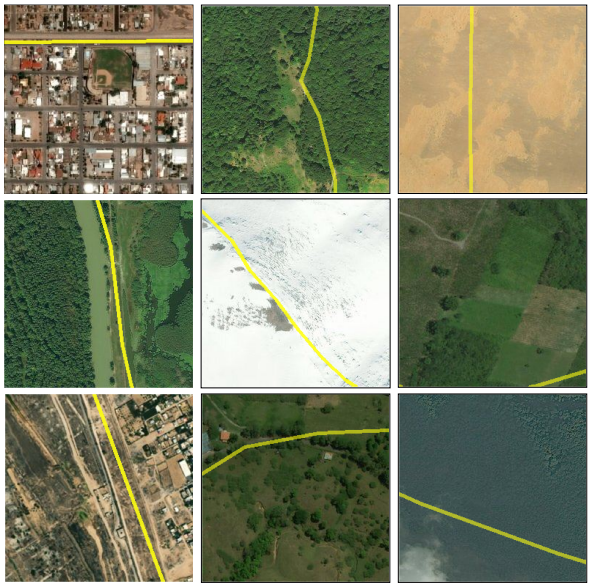
[1]Western Washington University {ortegat, nelso343, cranes2, wehrwes}@wwu.edu
[2]University of Colorado Boulder josh.myers-dean@colorado.edu

Abstract

Key aspects of international policy, such as those pertaining to migration and trade, manifest in the physical world at international political borders; for this reason, borders are of interest to political science studying the impacts and implications of these policies. While some prior efforts have worked to characterize features of borders using trained human coders and crowdsourcing, these are limited in scale by the need for manual annotations. In this paper, we present a new task, dataset, and baseline approaches for estimating the legibility of international political borders automatically and on a global scale. Our contributions are to (1) define the border legibility estimation task; (2) collect a dataset of overhead (aerial) imagery for the entire world's international borders, (3) propose several classical and deep-learning-based approaches to establish a baseline for the task, and (4) evaluate our algorithms against a validation dataset of crowdsourced legibility comparisons. Our results on this challenging task confirm that while low-level features can often explain border legibility, mid- and high-level features are also important. Finally, we show preliminary results of a global analysis of legibility, confirming some of the political and geographic influences of legibility.

1. Introduction

Recently in political science, Simmons and Kenwick proposed the concept of border orientation, defined as "the extent to which the state is committed to the public, authoritative, and spatial display of control over territorial entry and exit at its national borders" [28]. Border orientation, which ranges from permissive to controlling, is not directly observable. However, initial efforts using extensive manual coding of border control structures demonstrated that indirect measurements can yield interesting political insights [28, 19].



(a) (b) (c)

Figure 1: Examples of the variety in border legibility in our dataset with columns (a) highly legible, (b) moderately legible, and (c) not legible. The location of the border is drawn in yellow.

Prior work relies on manual coding of a few indicators including built structures at border crossings, a relatively small number of well-documented border walls, and population-weighted density of police stations. In this paper, we propose a new computer vision task with the high-level goal of providing a more general and scalable indicator of border orientation. Casting the measurement of border orientation—a policy-level concept—as a visual task, we propose to estimate the legibility of international political borders; legibility is defined as the extent to which a border can be visually detected. Our intuition is that the visible manifestation of border-related infrastructure should provide a good proxy for the extent to which a state has invested

in the spatial display of territorial control over its borders. While legibility could be assessed from a ground-level perspective, ground-level imagery at borders is not universally available; instead, we focus on estimating border legibility in overhead imagery. The ability to estimate a state's underlying strength near its frontier by a visual characterization of the environment would enable exciting research directions in political science; measuring legibility on a global scale would enable research on topics central to political science and international relations, such as state consolidation, border conflicts, and human development.

Meanwhile, the border legibility estimation problem also presents an interesting computer vision challenge. Readily available overhead imagery and well-documented political border locations mean that image data with ground-truth border locations is plentiful, opening the door to self-supervised approaches. However, we find that these approaches require careful consideration because legibility is very low in a large fraction of many of the world's borders (e.g., relatively featureless forests, deserts, or agricultural land in remote areas, such as the tiles in Figure 1(c)) are highly illegible. Furthermore, we find that though ground truth border locations are easily obtained, legibility itself is a subjective concept and requires human judgements to supply ground truth.

Another interesting technical feature of the legibility estimation problem is that the relevant visual features range widely from low-level to high-level. For example, in Figure 1(a) middle, the border is legible due to low-level edge features due to the river. In other locations, legibil-ity might be due to mid-level features like textural contrast caused by differences in land use, vegetation, or building style. Finally, high-level semantic reasoning may be required around border crossings where traffic control patterns or buildings suggest the presence of the border.

In this work, we define the border legibility estimation task, collect a global dataset of overhead image tiles covering the world's international land borders, and develop a suite of methods for the border legibility estimation task. In line with our intuition about the variety of visual features relevant to the task, our methods range from classical pixel-based image statistics to a self-supervised deep network trained using contrastive learning. To validate our methods, we used crowdsourcing (i.e., Amazon Mechanical Turk) to collect a dataset of pairwise human legibility comparisons, and benchmark our baselines on various metrics, considering both raw pairwise comparison agreement as well as metrics comparing to a ranking built from pairwise human judgements.

2. Related Work

Overhead imagery. Many large scale overhead image datasets have been proposed in recent years taking advantage of the increasing availability of such imagery from around the world [31, 34, 33, 40, 23, 15, 22, 38, 7, 27, 30, 39, 2]. These datasets have helped further progress in key areas such as satellite image classification [18, 26]. They have also helped pioneer the benchmarking of novel tasks such as canopy height estimation [21], parking lot detection [35], and parcel segmentation [1]. Similar to these works, we benchmark a novel overhead image task making use of a newly collected dataset. We are also inspired by previous works for the use of classical Computer Vision methods, such as K-means clustering, as a baseline for satellite image tasks [16, 18].

Self-Supervised training. Unlike many other satellite image task frameworks, we do not have sufficient ground truth data to directly learn the legibility task from our ground truth. Therefore, we look toward self-supervised methodologies capable of learning useful data properties with imperfect or missing labels. In particular, we draw inspiration from two approaches to self-supervised learning that have recently demonstrated impressive results: Contrastive Siamese networks [37, 5, 6, 17, 11] and Cut-and-Mix based training augmentation [36, 25, 24]. In our work, we propose a novel combination of both approaches to learn a model that performs pairwise the border legibility judgements. As in SimCLR [5], we are interested in representation learning based on two differing views of the same image; however rather than enforcing similarity between the representations, we introduce legibility-specific CutMix [36]-style augmentations that stand in for ground truth pairwise legibility comparisons.

Pairwise Evaluation. Our contrastive learning model, validation dataset, and evaluation metrics all rely on pairwise legibility judgements rather than prediction of an absolute legibility "score". Previous work in both computer vision [3] and political science [4] has demonstrated the usefulness of pairwise comparisons for quantities where absolute scales are inappropriate due to subjectivity (e.g., sentiment) or humans' inability to make absolute judgements (e.g., surface albedo). Given sufficient pairwise comparisons among a set of example tiles, a total order can be generated using various methods such as Elo [29]. In our work, we leverage the framework developed by Carlson and Montgomery [4], which uses a random utility model to jointly estimate a score for each example and a worker reliability score for each human annotator.

3. Border Legibility

We define border legibility as the extent to which the border is visible to the naked eye. Figure 1 gives examples of aerial imagery depicting highly legible (a), moderately legible (b), and illegible (c) borders, with the true border location overlaid in yellow. Borders can be legible for a variety of reasons, some of which relate to direct hu-

man influence on the landscape, while others relate to geographic features. Human-influenced features might include differences in land use on each side of the border; markers, roads, fences, walls, or other structures running along or parallel to the border; and differences in the built environment. Meanwhile, highly legible geographic features like rivers and mountain ridges often coincide with borders.

In large part, we leave the precise interpretation of "legibility" up to human visual judgement with one key exception related to human influence. Motivated by our eventual goal of measuring causes and consequences of legibility, we let human-influenced features be a tie breaker in comparisons where visual legibility is otherwise equal. For example, if one border has a road along it and another follows an equally visually distinctive river, the border with the road should be considered more legible since the visible evidence for the border is human-made. Further details on annotation data collection are provided in section 5 and the supplemental material.

3.1. Defining the Legibility Estimation Task

The legibility estimation task can be posed in absolute or relative terms. An absolute prediction problem requires a model to output some kind of legibility score given a single tile. While such scores could be normalized to fit on some numerical scale (e.g., 0 to 1), legibility does not naturally follow any specific scale, so even for such models we do not impose this constraint in model development or evaluation.

Because scores may be on an arbitrary scale and human judgements on a absolute scale are likely to be unreliable, we can also develop models that make relative legibility judgments: given a pair of images, a model may decide which is more legible. Using the same pairwise ranking technique as we use for our ground truth human annotations [4] (see section 2), we can use pairwise judgements over a set of inputs to generate a total order over the whole set. While most of our more classical baseline methods output absolute scores given a single tile, our strongest method is based on a siamese architecture that performs pairwise judgements.

The second important determination in posing the legibility problem is what border information the task is conditioned on. Since the border legibility of a non-border image tile is not readily defined, our task is conditioned on the presence of a border: given an image with a border, how legible is that border? However, a human annotator or a vision model may also be provided with the image-space location of the border. In our early experiments judging legibility with and without the border location, we found that both had drawbacks. Knowing the border location opens the door for confirmation bias, but not knowing the border location allows for non-border features to be mistaken for evidence of legibility (e.g., the boundary of an agricultural field that does not follow the border). In informal annotation experiments, we found that the latter effect was much more common, and chose to pose the task as conditioned on the location of the border.

In summary, we work with two definitions of the border legibility estimation task:

1. Absolute task: Given an image containing a border and the location of the border, output a single real number legibility score, with higher being more legible.

2. Relative task: Given two images containing borders and the location of each border, determine which of the images has the more legible border.

The pairwise ranking framework [4] allows methods solving the relative task to generate results on the absolute task for a given set of images.

4. Legibility Estimation Methods

While machine learning is well-suited for data-driven approaches to understanding high-level concepts like border legibility, training a model to directly predict legibility was not feasible because a large-scale dataset of legibility labels was too expensive to collect. Specifically, our 1000-image validation set cost around $500 to collect, while the whole world has over 600,000 tiles. Self-supervised approaches seem promising, but require careful design to work well with the particulars of the border legibility task. This section begins by presenting some relatively simple statistical baselines that leverage raw pixel values or features from pretrained deep neural networks. We then describe a pairwise self-supervised siamese model trained using contrastive learning that outperforms the classical baselines on some metrics.

4.1. "Classical" Baselines

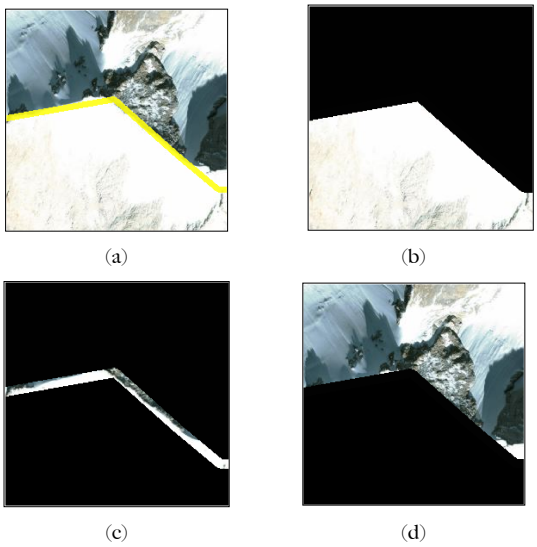
To establish reasonable baselines that do not rely on custom-trained neural networks, we tried a number of variations on a general feature analysis framework that compares collections of per-pixel features from different regions of the image. Our guiding intuition is that in a suitable feature space, the differences among pixel-wise features in different locations with respect to the border should correlate with the legibility of the border. Consider as a simple example the tile in Figure 2a, where the distribution of RGB pixel colors clearly differs from one side of the border to the other due to differences in land use. Mid- and high-level features might also differ, with similar colors but different textures due to vegetation or the built environment.

Another key intuition is that there are two general reasons that borders may be legible. The two sides of the border may be distinguishable as in Figure 2a; in this case it makes sense to compare features on opposite sides of the

(a) (b)

Figure 2: Examples of images with discrepancies in border segments primarily due to (a) features on either side of the border and (b) features along the border.

border. However, Figure 2b shows an example of a border that is legible only because of a feature running along the border. For this reason, we consider not only two sides of the border, but three segments including a buffered area along the border. An example of the three segments under consideration is shown for another example tile in Figure 3.



(a) (b) (c) (d)

Figure 3: (a) An input image I with the border drawn in yellow. Three collections of features are created by masking the pixels (b) on one side of the border, (c) in a margin around the border, and (d) on the other side.

Our general approach for baselines is to extract a collection of features from each of these three segments of the border, then use some measure to compare the similarity of the feature collections. Formally, let I be an input tile, φ be a feature extractor, and define three masks that select the pixels on one side (M_A), a margin surrounding and including the border (M_B), and the other side (M_C) as shown in Figure 3. We compute corresponding feature collections F_A, F_B, F_C as:

$$\begin{aligned} F_A &= \varphi(I)[M_A] \\ F_B &= \varphi(I)[M_B] \\ F_C &= \varphi(I)[M_C] \end{aligned} \tag{1}$$

where the Numpy-like [12] notation $A[M]$ extracts the masked locations of A specified by a binary mask M. This results in an $n \times d$ feature matrix, where n is the number of ones in M and d is the number of channels in A.

We experimented with three different feature extractors (φ). Considering color only, φ^P simply treats the RGB pixel values as feature vectors. $\varphi^{R(L)}$ takes the output features from layer L of a pretrained convolutional neural network model; we used ResNext-101 [32], taking the output of the `conv1` layer as $\varphi^{R(1)}$ and similarly for `conv2` and `conv3`. Finally, φ^T extracts the per-patch features produced by the encoder of a large Transformer model (we used Masked AutoEncoder [13]). We denote the corresponding feature collections F^P, $F^{R(L)}$, and F^T, respectively.

Given these three collections of features, we use some dissimilarity measure $D(\cdot,\cdot)$ to compare them, usually in a pairwise or one-vs-all manner, taking the maximum of individual dissimilarities as the legibility score prediction. Using the maximum dissimilarity enables identification of legibility due to difference in sides (Figure 2a) or due to features along the border (Figure 2b).

We experimented with a variety of dissimilarity measures; we present two of the most successful ones here, and include a few others in the supplemental material. A dissimilarity measure D compares two collections of features; while the feature dimensionality is the same, the number of features may differ based on the number of pixels in the corresponding segment's mask.

Average Pairwise Feature Distance. Our simplest measure of dissimilarity among feature collections is a simple average distance among pairs of individual features. We experimented with L^2 distance and found that, somewhat surprisingly, the cosine distance $d(f_1, f_2) = \frac{f_1}{||f_1||} \cdot \frac{f_2}{||f_2||}$ worked best even for RGB pixels.

Formally, we define the dissimilarity between two feature collections as the average distance between a pair of features, with one chosen from each collection:

$$D(F_1, F_2) = \frac{1}{|F_1||F_2|} \sum_{f_1 \in F_1, f_2 \in F_2} d(f_1, f_2). \tag{2}$$

The legibility score is then calculated as the maximum dissimilarity among each segment and the other two:

$$L_{Cos} = \max(D(F_A, F_{BC}), D(F_B, F_{AC}), D(F_C, F_{AB})) \tag{3}$$

where F_{AC} represents the concatenation of F_A and F_C.

Cluster Assignment Distributions. We use K-means clustering to cluster the whole tile's features into k (we set k = 3) clusters. Given the cluster assignment of each feature in the tile, we calculate a normalized discrete distribution of cluster assignments p_{ABC} for the whole tile, and for only the features in each tile p_A, p_B, p_C. Legibility is then measured as the maximum disagreement between the distribution of cluster assignments for each segment and the overall full-tile distribution:

$$L_{Cluster} = \max_{S \in A,B,C} D_{KL}(p_S || p_{ABC}) \quad (4)$$

where D_{KL} denotes Kullback–Leibler divergence.

4.2. Pairwise Legibility Prediction using a Self-Supervised Siamese Network

While ground truth legibility labels are not readily available, image tiles containing known borders are plentiful. With this in mind, we sought to train a legibility estimation model using self-supervised learning. We designed a contrastive Siamese network, BorderCut, to make relative legibility predictions, i.e., to predict which of two images has a more legible border.

Contrastive Training Approach Inspired by recent success in contrastive learning methods (e.g., [5]), we devised an augmentation scheme that produces pairs of synthetic training examples where the ground truth label (i.e., the more legible of the pair) is known. Our key idea is that, while we cannot know how legible a single tile is, in most cases we can augment it to become more legible with high probability. We accomplish this using a CutMix-style augmentation, replacing one or more segments with pixels from a random other image.

Let x represent an unedited border image tile and select two other random tiles z_1, z_2. We then construct a synthetic training pair (x'_1, x'_2) using one of three augmentation strategies:

1. x'_1 is set to x, while one side of the border (M_A or M_C) from x_2 is randomly chosen to be replaced with pixels from z_1, e.g., $x'_2[M_A] = z_1[M_A]$. An example is given in Figure 4(a).
2. x'_1 is set to x, while the border (M_B) in x'_2 is replaced with pixels from z_1, as shown in Figure 4(b).
3. x'_1 has one side of the border replaced (M_A or M_C), while x'_2 has the same replacement but also has the border replaced with images from a different image z_2. In other words, $x'_1[M_A] = z_1[M_A]$ and $x'_2[M_A] = z_1[M_A]$ and $x'_2[M_B] = z_2[M_B]$, as shown in Figure 4(c).

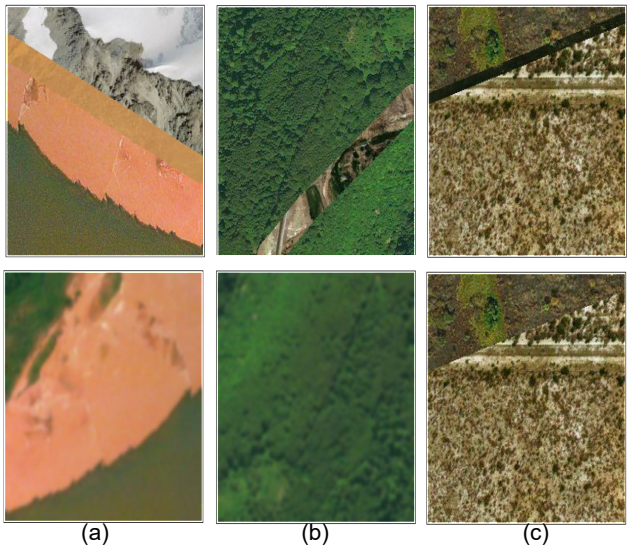


Figure 4: Examples of augmented training pairs where the bottom image represents x'_1 and the top represents x'_2.

By mixing border segments from randomly sampled images, we take advantage of the natural diversity of features from the world's borders: barring highly unlikely coincidences, these mixed features introduce border segments in x'_2 that are artificially distinct and do not exist in x, regardless of its initial legibility. While we can't know the absolute legibility of a mixed tile, most augmented tiles will be more legible than their non-mixed counterpart, so the ground truth "more legible" tile is set to be x'_2. Finally, we randomly swap x'_1 and x'_2 at training time to ensure our updated ground truth label $\hat{y}$ is not always the same.

Network Architecture. The BorderCut model takes two images x_0 and x_1 and makes a binary prediction, outputting a two-class softmax probability vector $\hat{y}$ using a siamese architecture as shown in Figure 5. We consider BorderCut as a composition of two separate functions: a shared backbone feature extractor $\varphi(\cdot)$, and a combined classification head $\phi(\cdot)$. The network can be described by:

$$\hat{y}(x_0, x_1) = \phi(\varphi(x_0) \oplus \varphi(x_1)) \quad (5)$$

where $\oplus$ denotes concatenation.

For $\varphi(\cdot)$ we use the fully convolutional backbone of Resnet18 [14], while $\phi(\cdot)$ is a 2-layer MLP with ReLU activations [9]. After two inputs are passed to $\varphi(\cdot)$, both returned feature representations are flattened to 512 dimensional vectors and concatenated to create a 1024 dimensional input to $\phi(\cdot)$. The classification head (ϕ) uses one linear layer to transform the input to a 512 dimensional vector and a second layer to reduce the representation into a 2 dimensional classification result, which is then softmaxed

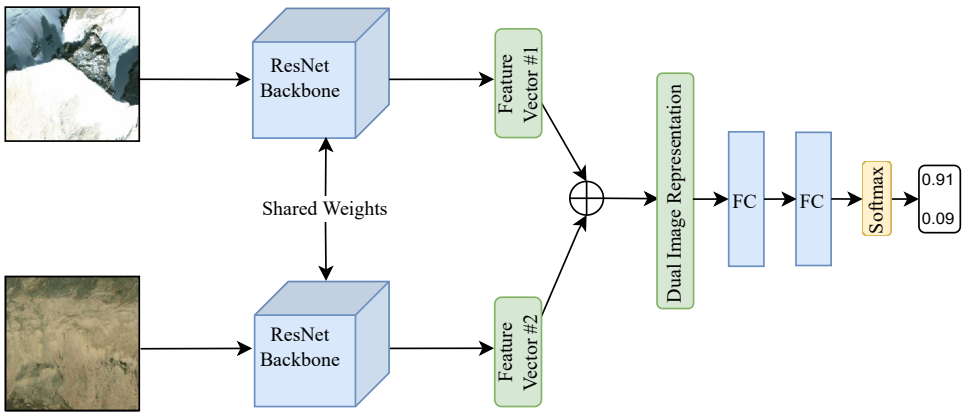


Figure 5: The BorderCut architecture. ResNet denotes φ from Equation 5, while everything after concatentation represents ϕ.

to get corresponding probabilities (i.e., higher corresponds to the more legible sample).

Training Details We train BorderCut on 40,000 randomly sampled images from our dataset and run validation on an additional 10,000 images. The model is trained using cross-entropy loss for 100 epochs with a batch size of 8 and a fixed learning rate of $1e^{-5}$. In section 6, we report the best validation performance achieved throughout training.

At test time, as shown in Table 2, our model generalizes from our training task to the border legibility task without any supervised labels.

5. Evaluation

This section discusses the data, metrics, and experiments we used to evaluate our legibility estimation methods. We collected a global dataset of aerial imagery and a crowdsourced validation set of pairwise legibility judgements. We evaluate our methods using raw accuracy on pairwise comparisons and two metrics comparing rankings of the 1000-tile validation set.

5.1. Data Collection

Using the International-Borders-2 dataset from Simmons and Kenwick [28] as the source of truth for the location of political borders, we collected aerial image tiles from the Bing Maps Imagery API[1] at densely sampled locations along all international land borders. The dataset contains a total of 612,374 aerial image tiles. Each tile has 256 × 256 pixels and covers a land area of 400 × 400 meters. We use Shapely [10] to derive image-space coordinates of the border linestring. Due to API terms of use, we are not able to release the full dataset, but it can be reproduced using the code and tile query locations, which are available on our project webapge[2].

To evaluate our methods against a "ground truth" notion of border legibility, we collected a validation set of legibility annotations on a set of 1000 random tiles from the global dataset. We used Amazon Mechanical Turk to collect about 12,000 pairwise judgements by asking annotators to decide which of two images has the more legible border. Annotators were shown both tiles with the border overlaid, with the ability to toggle the border off to see any features obscured by the border line. We used the framework from Carlson and Montgomery [4] to aggregate the pairwise annotations into a ranking that also yields worker confidence scores. The worker confidence scores, as well as manual inspection, suggest that, while worker reliability was variable, the ground truth data is not corrupted by large numbers of objectively incorrect annotations. Further details of the crowdsourcing data collection process, including the instructions given to annotators, are provided in the supplemental material.

5.2. Evaluation Metrics

We evaluate our method predictions on raw accuracy of pairwise comparisons and two ranking metrics: Tau (Kendall's Tau Rank Correlation [20]) and Footrule (Spearman's Footrule [8]). For all ranking metrics, we create absolute image rankings by applying the statistical model from [4] to both ground truth and predicted pairwise comparisons. Applying metrics on the rankings has two benefits: first, the metrics are less influenced by random sampling noise due to which pairs were annotated; and secondly, the ranking algorithm models the reliability of each individual annotator, helping to smooth over noise and

[1] https://docs.microsoft.com/en-us/bingmaps/rest-services/imagery/

[2] https://fw.cs.wwu.edu/~wehrwes/BorderLegibility/

Method	φ^P			$\varphi^{R(1)}$			$\varphi^{R(2)}$			$\varphi^{R(3)}$			φ^T		
	Acc.	τ	Footrule	Acc.	τ	Footrule	Acc.	τ	Footrule	Acc.	τ	Footrule	Acc.	τ	Footrule
Distance	62.40	0.151	277.91	60.28	0.084	301.38	58.17	0.059	313.26	56.19	0.261	244.82	51.63	-0.027	332.82
Clustering	61.28	0.075	305.33	63.42	0.116	290.92	60.19	0.209	262.53	58.80	0.449	186.68	49.86	0.032	314.49

Table 1: Results of baseline methods (subsection 4.1) on the annotated validation set, for different choices of input features φ. Higher is better for accuracy and τ, while lower is better for Footrule.

achieve greater global consistency by down-weighting contributions by workers whose annotations were less reliable.

Accuracy We compute the discrete number of times a given method prediction agrees with the annotated truth, then divide the total agreements by the total number of comparisons. In cases where a pair of images has been annotated multiple times, we take the majority vote winner as the most legible image. In cases where the annotations result in a tie, a random choice is made.

Kendall's Tau measures the normalized rate of "inversions" in the ranking. For a given pair of images, an inversion occurs if the images appear in the opposite order in the predicted ranking versus the ground truth ranking. For a set of images, $X = \{x_1, \ldots, x_N\}$, a predicted ranking $\hat{r}$, and a ground truth ranking r that yield the ordinal position of an image in the ranking, $d_\tau(\cdot)$ can be written as:

$$d_k(r, \hat{r}) = |\{(x_i, x_j) : i = j \wedge r(x_i) < r(x_j) \wedge \hat{r}(x_i) > \hat{r}(x_j)\}| \quad (6)$$

$$d_\tau(r, \hat{r}) = 1 - \frac{4 \cdot d_k(r, \hat{r})}{N \cdot (N-1)} \quad (7)$$

Equation 7 normalizes the number of inversions to the range of a correlation coefficient such that $d_\tau(r, \hat{r}) \in [-1, 1]$, with 1 representing perfect agreement (0 inversions) between the two rankings.

Spearman's Footrule is a slightly more interpretable metric that measures the total displacement: the sum of all absolute differences in rank positions between r, and $\hat{r}$. To aid with interpretability, we divide by N to give the average displacement; in other words, this metric measures, on average, how far an image's predicted rank is from its ground truth ranking position. Formally, with X, r, and $\hat{r}$ as defined in Equation 6, we define $d_F(\cdot)$:

$$d_F(r, \hat{r}) = \frac{1}{N} \sum_{i=1}^{N} |r(x_i) - \hat{r}(x_i)| \quad (8)$$

6. Results and Discussion

Baselines We benchmark the performance of our two baselines - Average Pairwise Feature Distance (Distance) and Cluster Assignment Distribution (Cluster) using five different feature extractors for each: pixels (φ^P), three intermediate ResNext-101 [32] convolutional feature layers ($\varphi^{R(1)}$, $\varphi^{R(2)}$, and $\varphi^{R(3)}$), and features from the encoder output of a Masked Autoencoder [13] (φ^T).

The results of our baseline methods are given in Table 1. Although the best score for each metric was achieved by the Cluster method, the best method depends on the choice of input features. Shallower CNN features win on raw pairwise accuracy, while the conv3 features are the clear winner on the ranking metrics. However, it is worth noting that even the simplest baseline, using average pairwise distance among RGB features φ^P achieves an accuracy only about 1% below the best performer on that metric. This is consistent with our observation that borders are often legible due to low-level visual features such as color differences.

The conventional wisdom that "deeper is better" is supported by our experiments using the Cluster baseline with CNN features, although the even deeper Transformer features perform significantly worse. For example, the τ performance of the Cluster baseline increases monotonically from 0.075 to 0.449 as the input features vary from raw pixels φ^P to conv3 ResNext features $\varphi^{R(3)}$. The Transformer features perform worse; we hypothesize that this is due to a lack of interpretable spatial reasoning in the encoded space for Cluster, and the fact that images must be processed in large chunks for Distance.

BorderCut Table 2 shows the performance of our self-supervised BorderCut model. Although the model is not trained on any ground truth labels, it achieves better raw accuracy than any of our baseline methods, while its performance on the ranking metrics remains worse than the best clustering baselines. This close competition between pretrained features (not even trained on overhead imagery) and a custom-trained self-supervised approach highlights the challenge of working without ground truth labels, and suggests that there remains potential for better performance, e.g., through additional experimentation with augmentation methods.

Method	Accuracy↑	τ ↑	Footrule↓
BorderCut	65.85 ± 1.6	0.145 ± 0.02	283.18 ± 8.8
$\varphi^{R(3)}$	58.80	0.449	186.68

Table 2: Comparison of BorderCut with the best performing "classical" method (i.e., $\varphi^{R(3)}$).

Discussion Overall, the performance metrics of all our methods on this task appear low: the best accuracy is around 63% (50% is random chance); the best τ correlation is below 0.5, and the best average displacement is around 187. While this does suggest that there is room for significant improvment over our baselines and BorderCut methods, we believe that due to ambiguity and human disagreement, perfect accuracy is not reasonable to expect or necessary for our applications. For example, in our judgement, at least the bottom 10% of tiles in the 1000-image validation set ground truth ranking appear equally illegible. Future work could investigate the extent of this ambiguity by looking at levels of human agreement in the annotations to quantify an upper bound on these metrics.

Supervised Machine Learning with Proxy Tasks For our purposes, direct supervised learning is prohibitively expensive as discussed in section 4. We attempted to train models to predict the position or angle of the border, then use their accuracy as a proxy for legibility. However, we found that such models were difficult to train because too many training examples have illegible borders, making the signal-to-noise ratio quite low. We also considered comparing border images to images from non-border locations, but this remains problematic because non-border tiles may still contain features that would be evidence for legibility if they were along a border (e.g., a river that does not coincide with a border).

Global Results As a preliminary experiment towards using our methods to understand global legibility trends, we ran the Distance method with φ^{P} features on the entire global Overhead-Borders dataset. Table 3 shows the top 10 most legible borders, computed by averaging per-tile legibility along each border. We find the most legible country borders with this method tend to be relatively short with distinctive, usually natural, features. For instance, short mountain borders such as France-Andorra and Russia-Georgia or river borders (e.g., Suriname-France, Zimbabwe-South Africa, Tanzania-Mozambique, Liechtenstein-Switzerland) are prominently featured. However, we also observe the effect of policy and human influence in North Macedonia-Greece, where a border fence began construction in 2015, and Zimbabwe-Botswana, where Hunter's Road covers much of the length of the border.

Rank	Border	Score
1	French Guiana–Suriname	0.535
2	North Macedonia-Greece	0.500
3	France-Andorra	0.397
4	Zimbabwe-South Africa	0.389
5	Liechtenstein-Switzerland	0.381
6	Armenia-Iran	0.354
7	Tanzania-Mozambique	0.328
8	Hungary-Yugoslavia	0.322
9	Zimbabwe-Botswana	0.307
10	Russia-Georgia	0.298

Table 3: Top 10 most legible borders according to Average Feature Cosine Distance with φ^{P} on the global Overhead-Borders data set. Border-level legibility is computed by averaging scores for all tiles in a border.

Limitations Our border legibility estimates show promise, but remain limited. In particular, our methods all rely, directly or indirectly, on the comparison among the three segments A, B, and C. Future work is needed to devise more general approaches that can learn these distinctions alongside even higher-level reasoning, such as the ability to identify border control structures. Our methods are also evaluated only on aerial imagery at a single resolution and with a fixed amount of spatial context; due to the imperfect aerial data source, in rare cases images are blurry, or the ground is obscured by clouds. Further examination of the effect of these parameters might yield improved performance and interesting insights about the spatial extent of features that give rise to legibility.

7. Conclusion

This paper introduced the novel computer vision task of border legibility estimation. We defined the task, collected a dataset, established baselines and benchmarks, introduced a self-supervised model for legibility prediction, and evaluated our methods against a small crowdsourced validation dataset of ground truth legibility annotations. While further research is needed for improved performance, our results already show promise in elucidating global legibility trends and their implications on the geography and policies of the world's countries.

Acknowledgements

This work was supported in part by the National Science Foundation under Grant No. 1917573. The authors gratefully thank Andrew Dunn, Nate Maassen, and Vivian White for their help with data collection.

References

[1] Han Lin Aung, Burak Uzkent, Marshall Burke, David Lobell, and Stefano Ermon. Farm parcel delineation using spatio-temporal convolutional networks. In CVPR Workshops, June 2020.

[2] Saikat Basu, Sangram Ganguly, Supratik Mukhopadhyay, Robert DiBiano, Manohar Karki, and Ramakrishna Nemani. Deepsat: A learning framework for satellite imagery. In Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems, SIGSPATIAL '15, New York, NY, USA, 2015. Association for Computing Machinery.

[3] Sean Bell, Kavita Bala, and Noah Snavely. Intrinsic images in the wild. ACM Trans. on Graphics (SIGGRAPH), 33(4), 2014.

[4] David Carlson and Jacob M Montgomery. A pairwise comparison framework for fast, flexible, and reliable human coding of political texts. American Political Science Review, 111(4):835–843, 2017.

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. arXiv preprint arXiv:2002.05709, 2020.

[6] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. arXiv preprint arXiv:2011.10566, 2020.

[7] Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote sensing image scene classification: Benchmark and state of the art. Proceedings of the IEEE, 105(10):1865–1883, oct 2017.

[8] Persi Diaconis and R. L. Graham. Spearman's footrule as a measure of disarray. Journal of the Royal Statistical Society: Series B (Methodological), 39(2):262–268, 1977.

[9] Kunihiko Fukushima. Cognitron: A self-organizing multilayered neural network. Biological Cybernetics, 20:121–136, 2004.

[10] Sean Gillies et al. Shapely: Manipulation and analysis of geometric objects. https://github.com/Toblerity/Shapely, 2007–.

[11] Mark Hamilton, Zhoutong Zhang, Bharath Hariharan, Noah Snavely, and William T. Freeman. Unsupervised semantic segmentation by distilling feature correspondences. In International Conference on Learning Representations, 2022.

[12] Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. Nature, 585(7825):357–362, Sept. 2020.

[13] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 16000–16009, 2022.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770–778, 2016.

[15] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Introducing eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. In IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium, pages 204–207. IEEE, 2018.

[16] Xu Ji, João F Henriques, and Andrea Vedaldi. Invariant information clustering for unsupervised image classification and segmentation. In Proceedings of the IEEE International Conference on Computer Vision, pages 9865–9874, 2019.

[17] Li Jing, Jiachen Zhu, and Yann LeCun. Masked siamese convnets, 2022.

[18] Mohammed Kadhim and Mohammed Abed. Convolutional Neural Network for Satellite Image Classification, pages 165–178. Springer International Publishing, 01 2020.

[19] Michael R. Kenwick and Beth A. Simmons. Pandemic response as border politics. International Organization, 74(S1):E36–E58, 2020.

[20] William R. Knight. A computer method for calculating kendall's tau with ungrouped data. Journal of the American Statistical Association, 61(314):436–439, 1966.

[21] Nico Lang, Walter Jetz, Konrad Schindler, and Jan Dirk Wegner. A high-resolution canopy height model of the earth. arXiv preprint arXiv:2204.08322, 2022.

[22] Haifeng Li, Xin Dou, Chao Tao, Zhixiang Wu, Jie Chen, Jian Peng, Min Deng, and Ling Zhao. Rsi-cb: A large-scale remote sensing image classification benchmark using crowdsourced data. Sensors, 20(6):1594, 2020.

[23] Haifeng Li, Hao Jiang, Xin Gu, Jian Peng, Wenbo Li, Liang Hong, and Chao Tao. Clrs: Continual learning benchmark for remote sensing image scene classification. Sensors, 20(4), 2020.

[24] Siyuan Li, Zedong Wang, Zicheng Liu, Di Wu, and Stan Z. Li. Openmixup: Open mixup toolbox and benchmark for visual representation learning, 2022.

[25] Jihao Liu, Boxiao Liu, Hang Zhou, Yu Liu, and Hongsheng Li. Tokenmix: Rethinking image mixing for data augmentation in vision transformers. In Proceedings of the European Conference on Computer Vision (ECCV), 2022.

[26] Mark D. Pritt and Gary Chern. Satellite image classification with deep learning. 2017 IEEE Applied Imagery Pattern Recognition Workshop (AIPR), pages 1–7, 2017.

[27] Xiaoman Qi, Panpan Zhu, Yuebin Wang, Liqiang Zhang, Junhuan Peng, Mengfan Wu, Jialong Chen, Xudong Zhao, Ning Zang, and P. Takis Mathiopoulos. Mlrsnet: A multi-label high spatial resolution remote sensing dataset for semantic scene understanding. ISPRS Journal of Photogrammetry and Remote Sensing, 169:337–350, 2020.

[28] Beth A. Simmons and Michael R. Kenwick. Border orientation in a globalizing world. American Journal of Political Science, n/a(n/a), 2022.

[29] Steven S. Skiena. The Data Science Design Manual. Springer Publishing Company, Incorporated, 1st edition, 2017.

[30] Gencer Sumbul, Marcela Charfuelan, Begum Demir, and Volker Markl. Bigearthnet: A large-scale benchmark archive for remote sensing image understanding. In IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium. IEEE, jul 2019.

[31] Gui-Song Xia, Xiang Bai, Jian Ding, Zhen Zhu, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang. Dota: A large-scale dataset for object detection in aerial images. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2018.

[32] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1492–1500, 2017.

[33] Lu Xu, Yiyun Chen, Jiawei Pan, and Gao Aji. Multi-structure joint decision-making approach for land use classification of high-resolution remote sensing images based on cnns. IEEE Access, PP:1–1, 02 2020.

[34] Yi Yang and Shawn Newsam. Bag-of-visual-words and spatial extensions for land-use classification. In Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems, GIS '10, page 270–279, New York, NY, USA, 2010. Association for Computing Machinery.

[35] Yifang Yin, Wenmiao Hu, An Tran, Hannes Kruppa, Roger Zimmermann, and See-Kiong Ng. A context-enriched satellite imagery dataset and an approach for parking lot detection. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pages 1371–1380, January 2022.

[36] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In International Conference on Computer Vision (ICCV), 2019.

[37] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. arXiv preprint arXiv:2103.03230, 2021.

[38] Weixun Zhou, Shawn Newsam, Congmin Li, and Zhenfeng Shao. PatternNet: A benchmark dataset for performance evaluation of remote sensing image retrieval. ISPRS Journal of Photogrammetry and Remote Sensing, 145:197–209, nov 2018.

[39] Xiao Xiang Zhu, Jingliang Hu, Chunping Qiu, Yilei Shi, Jian Kang, Lichao Mou, Hossein Bagheri, Matthias Haberle, Yuansheng Hua, Rong Huang, Lloyd Hughes, Hao Li, Yao Sun, Guichen Zhang, Shiyao Han, Michael Schmitt, and Yuanyuan Wang. So2sat lcz42: A benchmark data set for the classification of global local climate zones [software and data sets]. IEEE Geoscience and Remote Sensing Magazine, 8(3):76–89, 2020.

[40] Qin Zou, Lihao Ni, Tong Zhang, and Qian Wang. Deep learning based feature selection for remote sensing scene classification. IEEE Geoscience and Remote Sensing Letters, 12(11):2321–2325, 2015.