Using Semantic Relationships among Objects for Geospatial Land Use Classification

Gilbert Rotich[1], Sathyanarayanan Aakur[1], Rodrigo Minetto[2],
Mauricio Pamplona Segundo[3] and Sudeep Sarkar[1]
[1] University of South Florida (USF)
[2] The Federal University of Technology - Paraná (UTFPR)
[3] Federal University of Bahia (UFBA)
grotich@mail.usf.edu

***Abstract*—The geospatial land recognition is often cast as a local-region based classification problem. We show in this work, that prior knowledge, in terms of global semantic relationships among detected regions, allows us to leverage semantics and visual features to enhance land use classification in aerial imagery. To this end, we first estimate the top-k labels for each region using an ensemble of CNNs called Hydra. Twelve different models based on two state-of-the-art CNN architectures, ResNet and DenseNet, compose this ensemble. Then, we use Grenander's canonical pattern theory formalism coupled with the common-sense knowledge base, ConceptNet, to impose context constraints on the labels obtained by deep learning algorithms. These constraints are captured in a multi-graph representation involving generators and bonds with a flexible topology, unlike an MRF or Bayesian networks, which have fixed structures. Minimizing the energy of this graph representation results in a graphical representation of the semantics in the given image. We show our results on the recent fMoW challenge dataset. It consists of 1,047,691 images with 62 different classes of land use, plus a false detection category. The biggest improvement in performance with the use of semantics was for false detections. Other categories with significantly improved performance were: zoo, nuclear power plant, park, police station, and space facility. For the subset of fMow images with multiple bounding boxes the accuracy is 72.79% without semantics and 74.06% with semantics. Overall, without semantic context, the classification performance was 77.04%. With semantics, it reached 77.98%. Considering that less than 20% of the dataset contained more than one ROI for context, this is a significant improvement that shows the promise of the proposed approach.**

***Index Terms*—Remote sensing, convolutional neural networks, pattern theory**

I. Introduction

After a natural disaster such as an earthquake or a hurricane, there is a need for damage assessment for humanitarian assistance purposes. Leveraging aerial imagery can assist in informed search-rescue and rebuilding effort. Other emerging uses of satellite images are in urban design and transportation management in smart cities. These new application areas, coupled with ready availability of high quality satellite images is leading a resurgence of interest in computer algorithms for geospatial datasets, consisting of satellite images, metadata, and temporal views. Computer vision challenges include the ability to overcome occlusions, varying perspectives, fine-grained discrimination, and class imbalance. We briefly discuss the latest datasets, their objectives and mention the best performing solutions for conventional understanding of remote sensing applications.

SpaceNet [1] is a dataset used for a series of challenges with the purpose of developing the algorithms to extract buildings footprints and also road networks automatically. Top performing solutions implemented different image segmentation algorithms. For example, an ensemble of three U-Net models [2] provided the best performance on extracting buildings footprints. U-Net is a convolutional neural network (CNN) architecture originally proposed for biomedical image segmentation that was later applied to many other domains [3].

In the xView challenge dataset [4], the goal was to detect and classify objects in satellite imagery. This dataset covers the following remote sensing problems: large image size, low interclass variability, occlusion from clouds, constrained computational resources and unbalanced classes. A top performing method [5] utilizes five Single Shot MultiBox Detector (SSD) [6] models, whose detections are merged by an adapted version of the non-maximum suppression (NMS) algorithm. SSD is an end-to-end CNN architecture for general object detection. It generates region proposals with labels from multi-scale feature maps and filter out duplicates using NMS.

In the IARPA Functional Map of the World (fMoW) [7] competition, participants had to develop machine learning algorithms to classify the land use of given regions of interest (ROI) in satellite images. Figure 1 illustrates the fMoW task. The top five participants employed ensembles of CNN leveraging visual and metadata information.

In this work, our goal is to use semantic relationships among multiple ROIs and investigate if this information can be used to enhance land use classification in satellite images. To this end, fMoW emerges as an appropriate benchmark, as it provides a straightforward scenario to assess the relevance of semantics. Almost 20% of its 53,473 testing images have more than one ROI that allow analyzing the semantic associations between them.

As our main contribution, we present an approach that combines Grenander's canonical pattern theory formalism with the common-sense knowledge base, ConceptNet, to impose context constraints on the ROI labels obtained by one of the

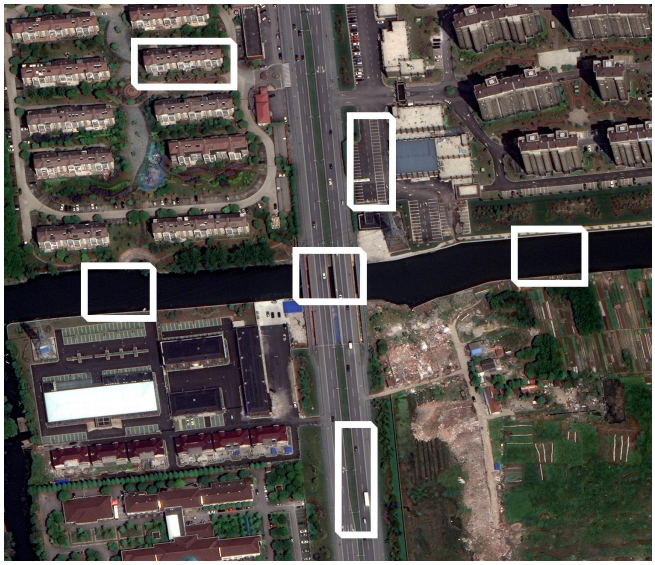


Fig. 1: Image sample from fMoW dataset. The fMoW task consists in, given a satellite image with one or more regions of interest, classify these regions as one of the 62 given classes (*e.g.* parking lot, shopping mall, crop field) or as a false detection.

fMoW winners' solutions, Hydra [8]. In this approach, the energy of a multi-graph representation involving generators and bonds with a flexible topology is minimized taking these constraints into account, resulting in a graphical representation of the semantics in a given image and ROI labels with the highest semantic agreement.

II. Related Work

The ability to use context among categories and or objects have shown to improve accuracies when dealing with object detection and classification problems because items generally appear together regularly in their natural environments [9]. Relationships among objects are based on scale, semantics or spatially. However, current deep learning paradigms focus on visual features and seldom take into account the environment or surrounding objects. The reason is that incorporating or modeling relations into learning algorithms is difficult.

Contextual modeling is either done at the feature level from feature maps or be formulated graphically using predicted labels. Examples of feature level strategies include CoupleNet [10], attention to context convolution neural network [11], spatial memory network [12] and most recently with state-of-art results, relations networks [13]. Alternatively, structure inference networks [14], Conditional random fields (CRF) for object recognition [15] and Graph-RCNN [16] are graphical-based.

Relation networks [13] propose an object relations module that leverages appearance features and geometry. It is easily integrable into many state-of-the-art object detectors today and is lightweight. Therefore there are no additional computational needs. Relation networks improve object detection and reduce the number of similar proposals.

In order to model context graphically to make use of spatial and visual features, usually, an object detector generates ROIs. A following non-maximum supression (NMS) algorithm reduces the number of duplicate ROIs. Each remaining ROI represents a node and the edges the relationship of object pairs. Yang *et al.* [16] follow this path using a region proposal network, Faster-RCNN, to localize ROIs in an image. Since the number of edges produced would be notably large, a relatedness score learned from the distribution of object co-occurrence is used to prune edges. The result is a sparse graph that becomes the input of an attentional graph convolutional network, leveraging contextual information from the entire scene.

Traditionally graphical topologies used to illustrate relations, such as CRFs or Bayesian Networks, are generated based on predefined set of rules and are not interchangeable. Hence the subtle variability of scenes or objects may not get captured or even observed. In video captioning, activity and action recognition are so challenging that attempting to generate a graph to cover every instance is not feasible. Therefore, de Souza *et al.* [17] propose a flexible framework that is discussed further in Section III while Aakur *et al.* [18] introduce the use of external information to define relationships.

Employing external knowledge to determine associations is advantageous because it is freely accessible and rich in the observable principles of the natural environment. A source of external knowledge is usually a graph representation of generally agreed upon facts and how they are expressed in language naturally and logically. ConceptNet, BabelNet, and WordNet are a few of such semantic graphs that tend to be very large.

III. Proposed Approach

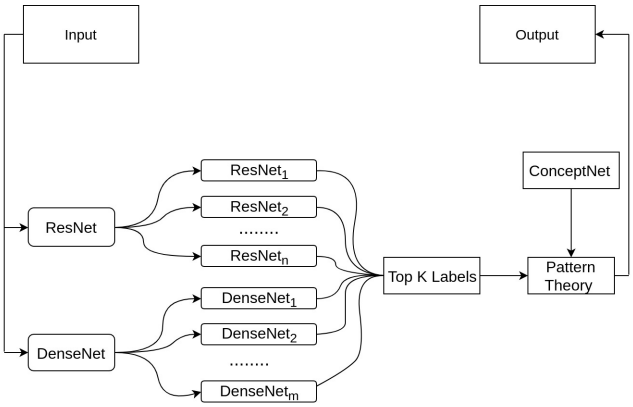


Fig. 2: Diagram of the proposed framework.

We propose an approach to process satellite imagery from the fMoW dataset, with the goal of categorizing land use in ROIs from satellite images. As illustrated in Figure 2, it consists of an ensemble of CNNs – Hydra [8] – and Grenander's

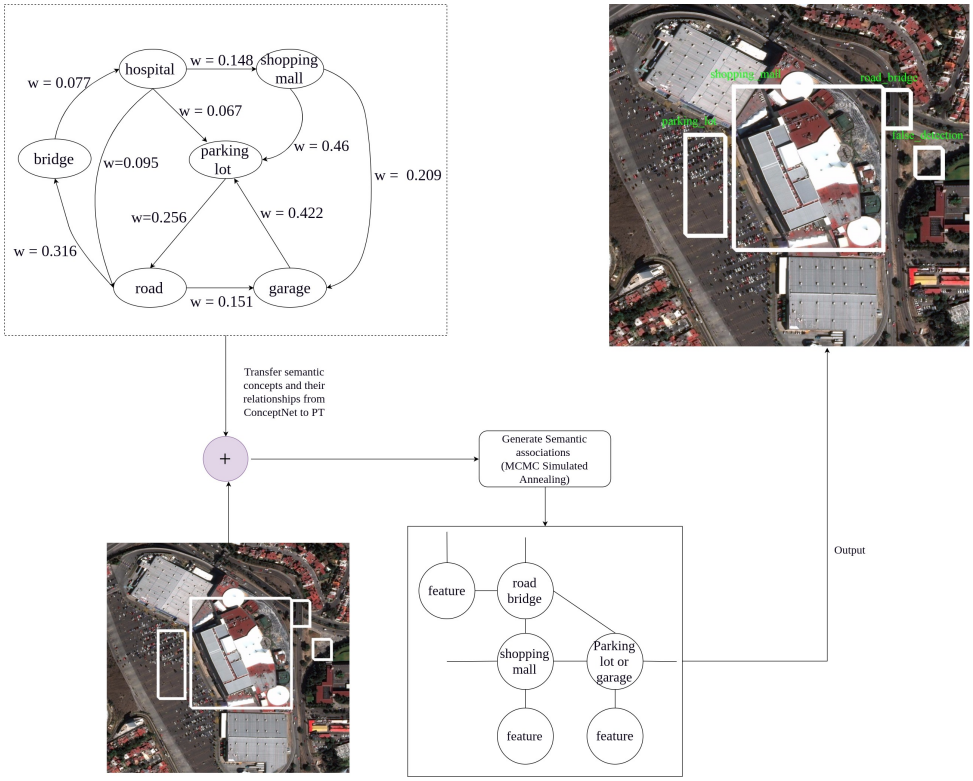


Fig. 3: Diagram of the pattern theory module. A graph topology representing semantic relationships is created using variations of top-10 labels for each bounding box and association weights from ConceptNet.

general pattern theory framework. Hydra produces the top-k labels for each ROI in an image (in this work, $k = 10$). The pattern theory module uses ConceptNet to generate a graphical structure representing semantic relationships among ROIs which is then used to select the most adequate labels for them. In the following sections, we will describe this framework in further details.

A. Dataset

The fMow challenge released a remote sensing dataset consisting of 1,047,691 satellite images, metadata, and temporal views. Each sample has a high-resolution pan-sharpened image and a low-resolution multispectral image. ROIs are represented by bounding boxes, and every bounding box in an image belongs to one of the 62 possible classes or is a false detection. A description of the training data provided is in Table I.

TABLE I: Number of images and ROIs in the fMow dataset.

	# of images	# of boxes	# of distinct boxes
Train	363,572	363,572	83,412
Validation	53,041	63,422	12,006
Test	53,473	76,998	16,948
Total	470,086	503,982	112,366

The test set contains 53,473 images, with 82.60% of them having a single ROI, 4.96% two ROIs, 5.66% three ROIs and 6.78% with four or more ROIs. Is important to mention that we are only able to extract semantic relationships from images with more than one ROI.

B. Ensemble of CNNs for ROI classification

Hydra is a framework to create ensembles of CNNs using two popular architectures, Residual Networks (ResNet) [19] and Densely Connected Convolutional Networks (DenseNet) [20]. Both networks are initialized with ImageNet [21] weights, then partially optimized for a few epochs using the fMoW training set. The updated weights are used to initialize several copies of these networks, which are referred to as the heads of the Hydra. Each head is then optimized for a few more epochs with a different configuration regarding image format, class weights, and data augmentation technique. During test, each head generates a vector of score values that show the likelihood of a ROI belonging to all existing classes. The results of all heads are fused by a simple sum, and then the top-k classes and their respective scores are passed to the pattern theory module.

C. Modeling Semantic Relationships

Discovering latent patterns to understand the basic concepts of this patterns is how we model context. The patterns include basic independent states, co-occurrence and inference functions. We utilized Grenander's canonical pattern theory formalism together with ConceptNet to represent semantic relationships among the categories of the fMoW dataset, as

shown in Fig 3. Below we will briefly discuss the ConceptNet and Grenander's pattern theory structure.

1) ConceptNet: ConceptNet, a semantic network relational graph, is a graphical abstraction of knowledge of the real world that is possessed by all people. This information mostly relies on conventional human experience that involves social, physical, temporal, psychological, spatial characteristics of everyday life. The knowledge is crowd-sourced and data mined from Wikipedia, Wiktionary, DBpedia, Freebase and WordNet. This contextual graph consists of concepts that are texts which will represent categories in our solution and assertions, relationships between the classes. It is intuitively a hypergraph where the nodes are objects classes, and edges are relations. There are 12.5 million edges, signifying 8.7 million statements connecting 3.9 million concepts. Associations between two concepts include *HasProperty*, *IsA*, and *RelatedTo* and are quantifiable by weight values. Figure 4 illustrates some of these associations in the context of the fMoW dataset.

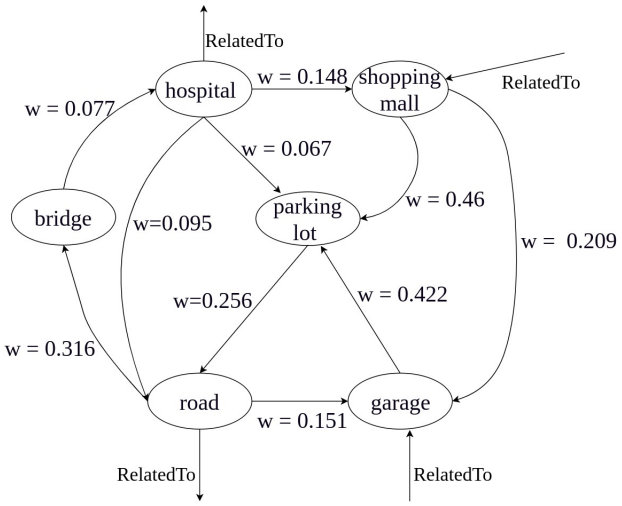


Fig. 4: A small subset of the semantic graph from ConceptNet in which nodes represent fMoW classes and are directionally connected by the *RelatedTo* association weight values.

2) Pattern Theory: We used the Grenander's pattern theory framework to encounter semantic relationships for image classification purposes. In this framework, a generator $g \in G$ is the basic unit of information. It may represent a concept (*i.e.* one of the fMoW classes) or a feature (*i.e.* the top-k classes and the respective scores for a ROI in the input image). The generator space G is the finite set of all the possible generators that can exist.

A generator g has a finite set of connectors, called bonds, and each bond is expressed as $\beta^l(g)$ where l is a unique identifier. Figure 5 illustrates these concepts.

There are different types of bonds, namely semantic and support. A support bond connects a generator g_i that represents an fMoW class to another generator g_j that represents a feature for a specific ROI. The bond value α_{supp} in this case is given by:

$$\alpha_{supp}(\beta'(g_i), \beta''(g_j)) = f(g_i, g_j) \tag{1}$$

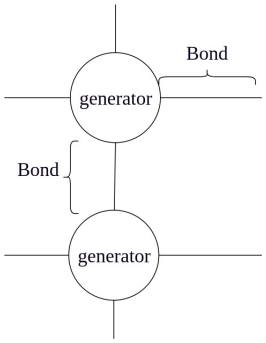


Fig. 5: Generators and bonds. A bond that is not connecting two generators is considered open, otherwise it is closed.

where $f(.)$ is the Hydra's confidence score for g_i's class in g_j's ROI. Semantic bonds represent the relationships between concept generators and are directional. This type of bond connects two generators g_i and g_j that represent two different fMoW classes. Its value α_{sem} is given by:

$$\alpha_{sem}(\beta'(g_i), \beta''j(g_j)) = tanh(\phi(g_i, g_j)) \tag{2}$$

where $\phi(.)$ is the *RelatedTo* weight value from ConceptNet. The $tanh$ function normalizes the output of ϕ to the range from -1 to 1.

A set of generators joined together using bonds to represent their semantic relationships form a graph $\sigma \in \Sigma$, with Σ being the collection of all finitely possible graph configurations. More specifically, a graph configuration c with n generators can be represented as:

$$c = \sigma(g_1, g_2, g_3, ..., g_n) \tag{3}$$

A visual representation of one configuration c is shown in Figure 6.

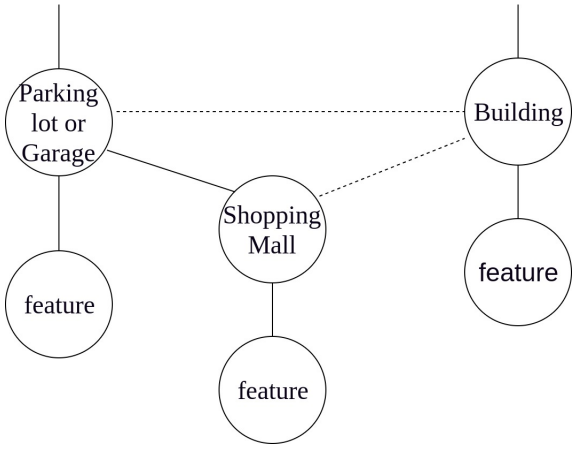


Fig. 6: Example of a configuration c with generators for fMoW categories in an image with multiple ROIs.

Finding the right configuration c for a set of ROIs in the input image can be seen as a search problem in which we maximize the probability of the graph topology:

$$P(c) \propto e^{-E(c)} \tag{4}$$

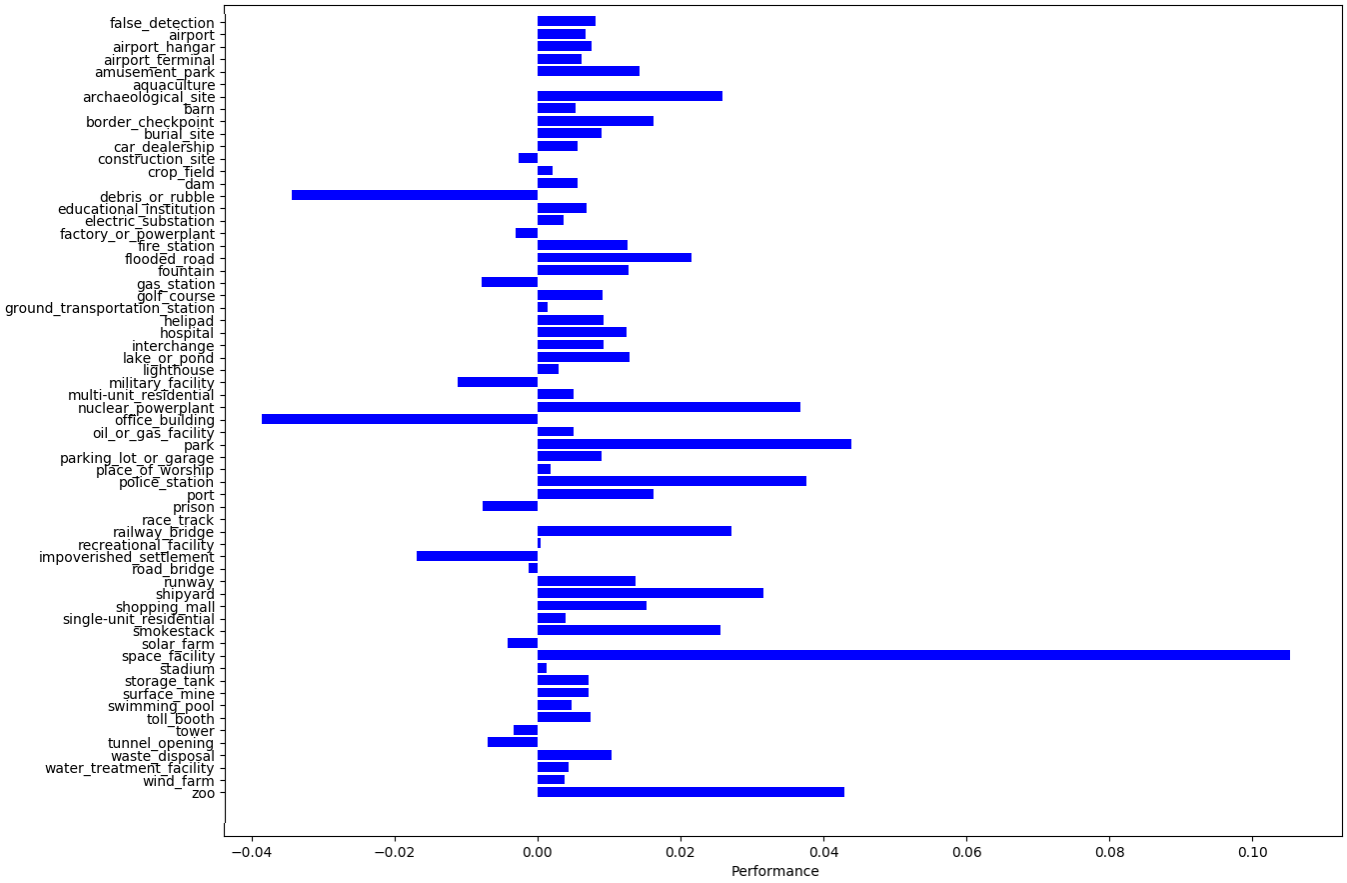


Fig. 7: Region classification performance gain by using semantical relationships: the classification of more than 75% of the categories were improved by using the region context.

To do so, we have to minimize the energy function $E(c)$. Low energy suggests that the concepts in the configuration have high associativity with each other based on their semantic relations. Given a configuration with bonds connecting a collection of generators, the energy function is the sum of all the bond values:

$$E(c) = -\sum \alpha_{supp}(\beta'(g_i), \beta''(g_j)) + \alpha_{sem}(\beta'(g_i), \beta''(g_j)) + Q(c) \quad (5)$$

where $Q(c)$ is the cost function of the structure. It is used to ensure that degenerate cases, like configurations with unconnected generators, do not happen and is calculated as follows:

$$Q(c) = \gamma \sum_{g \in G} \sum_{\beta' \in g} [D(\beta'(g))] \quad (6)$$

with $D(.)$ returning 1 if β' is open, and 0 otherwise. The parameter γ controls the importance degenerate cases have on the quality of the interpretation.

Since the number of generators and bonds is mutable, the search space can be exponential. We use a simulated annealing process as inference with an efficient Markov Chain Monte Carlo process for guiding the search process. It is an efficient sampling technique, and sampled configurations are either rejected or accepted based on their respective energy scores. In the end, the configuration with the lowest energy is chosen. Its generators are then used to obtain the final labels for the ROIs of the input image.

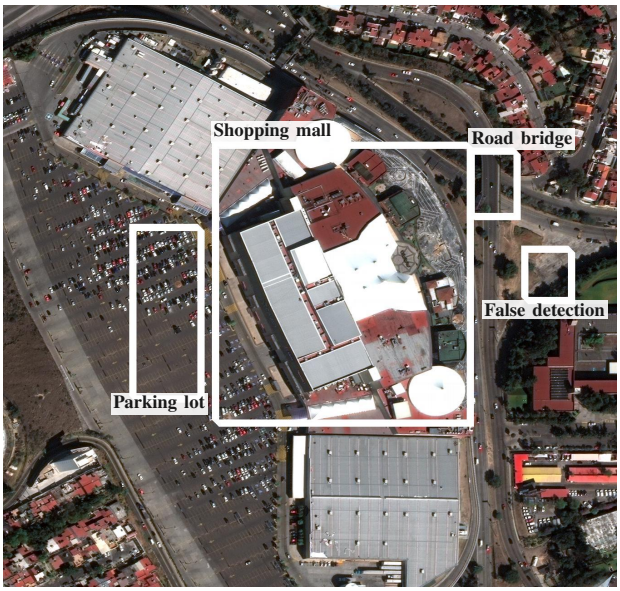
IV. Results and Discussion

The quantitative criteria used by the Functional Map of the World challenge to rank its competitors was a weighted average of the F-measure [7] for each class. As shown in Table II, the semantical relationships help to increased the classification accuracy in nearly 1%, from 77.04% to 77.98%. Although the gain does not appear to be significant in absolute numbers, is important to note that 49 out of 62 classes had an increase of performance, as can be seen in Figure 7. We need to remember that approximately 80% of the fMoW test images have a single ROI and cannot be analyzed through semantic relationships, making this 1% gain much more noteworthy.

Furthermore when we computed the effect of using semantics on the subset of images with multiple ROIs, there were improvements in accuracies of approximately 1.25% i.e., weighted F-measure score without semantics was 72.79% and with Semantics 74.06% while for unweighted score was, without semantics 73.59 % and with semantics 74.78%

The correction of mis-classifications primarily occurred in images with multiple bounding boxes. Hydra with semantics discriminated well between the background class, and the other

62 classes. This is attributed to the fact that in the test set there were 6000 distinct bounding boxes and 2800 were false detections (see Figure 8).



(a) Groundtruth.

(b) Hydra without and with semantics relationships.

Fig. 8: Land classification: (a) image sample from Functional Map of the World Challenge with groundtruth regions and labels; (b) region classification without semantical context and with semantical context.

V. Conclusion and Future Work

Leveraging semantic relationships among the object classes from external knowledge leads to an improvement in classification accuracies, particularly in images with multiple bounding boxes. Our approach in this instance reduces the false detections. In the future, we propose incorporating pattern

TABLE II: Classification accuracies for each of the 62 categories in the IARPA Functional Map of the World dataset, as captured by the category weighted F1 scores.

Categories	w/o Semantics	with semantics
airport	0.9130	0.9197
airport hangar	0.7453	0.7529
airport terminal	0.7913	0.7975
amusement park	0.8939	0.9082
aquaculture	0.8462	0.8462
archaeological site	0.7054	0.7313
barn	0.7715	0.7768
border checkpoint	0.5814	0.5977
burial site	0.9036	0.9126
car dealership	0.8475	0.8531
construction site	0.4841	0.4815
crop field	0.9529	0.9551
dam	0.9162	0.9218
debris or rubble	0.6735	0.6392
educational institution	0.6105	0.6174
electric substation	0.8983	0.9019
factory or powerplant	0.6346	0.6316
fire station	0.6223	0.6349
flooded road	0.6835	0.7051
fountain	0.8564	0.8691
gas station	0.8810	0.8732
golf course	0.9378	0.9469
ground transportation station	0.8022	0.8036
helipad	0.8604	0.8696
hospital	0.5018	0.5143
interchange	0.9146	0.9239
lake or pond	0.7358	0.7487
lighthouse	0.8197	0.8227
military facility	0.7107	0.6995
multi-unit residential	0.5353	0.5403
nuclear powerplant	0.5882	0.6250
office building	0.3252	0.2866
oil or gas facility	0.8692	0.8743
park	0.7364	0.7803
parking lot or garage	0.7856	0.7946
place of worship	0.7679	0.7698
police station	0.4040	0.4416
port	0.7294	0.7456
prison	0.7500	0.7423
race track	0.9449	0.9449
railway bridge	0.8062	0.8333
recreational facility	0.9303	0.9308
impoverished settlement	0.8056	0.7887
road bridge	0.7713	0.7701
runway	0.9099	0.9237
shipyard	0.6000	0.6313
shopping mall	0.6910	0.7063
single-unit residential	0.7598	0.7637
smokestack	0.7970	0.8226
solar farm	0.9453	0.9412
space facility	0.8421	0.9474
stadium	0.8598	0.8611
storage tank	0.9368	0.9440
surface mine	0.8993	0.9064
swimming pool	0.9238	0.9286
toll booth	0.9474	0.9549
tower	0.7940	0.7907
tunnel opening	0.9722	0.9652
waste disposal	0.6880	0.6984
water treatment facility	0.9122	0.9165
wind farm	0.9740	0.9778
zoo	0.7097	0.7527
Overall	**0.7705**	**0.7798**

theory into the training process as well as increasing the number of contextual parameters.

REFERENCES

[1] A. Van Etten, D. Lindenbaum, and T. M. Bacastow, "Spacenet: A remote sensing dataset and challenge series," *arXiv preprint arXiv:1807.01232*, 2018.

[2] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[3] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[4] D. Lam, R. Kuzma, K. McGee, S. Dooley, M. Laielli, M. Klaric, Y. Bulatov, and B. McCord, "xview: Objects in context in overhead imagery," *arXiv preprint arXiv:1802.07856*, 2018.

[5] G. Rotich, R. Minetto, and S. Sarkar, "Resource-constrained simultaneous detection and labeling of objects in high-resolution satellite images," *arXiv preprint arXiv:1810.10110*, 2018.

[6] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *European conference on computer vision*. Springer, 2016, pp. 21–37.

[7] G. Christie, N. Fendley, J. Wilson, and R. Mukherjee, "Functional map of the world."

[8] R. Minetto, M. P. Segundo, and S. Sarkar, "Hydra: an ensemble of convolutional neural networks for geospatial land classification," *arXiv preprint arXiv:1802.03518*, 2018.

[9] M. Pei, Y. Jia, and S. Zhu, "Parsing video events with goal inference and intent prediction," in *2011 International Conference on Computer Vision*, 2011, pp. 487–494.

[10] Y. Zhu, C. Zhao, J. Wang, X. Zhao, Y. Wu, H. Lu *et al.*, "Couplenet: Coupling global structure with local parts for object detection."

[11] J. Li, Y. Wei, X. Liang, J. Dong, T. Xu, J. Feng, and S. Yan, "Attentive contexts for object detection," *IEEE Transactions on Multimedia*, vol. 19, no. 5, pp. 944–954, 2017.

[12] X. Chen and A. Gupta, "Spatial memory for context reasoning in object detection."

[13] H. Hu, J. Gu, Z. Zhang, J. Dai, and Y. Wei, "Relation networks for object detection," in *Computer Vision and Pattern Recognition (CVPR)*, vol. 2, no. 3, 2018.

[14] Y. Liu, R. Wang, S. Shan, and X. Chen, "Structure inference net: Object detection using scene-level context and instance-level relationships."

[15] A. Quattoni, M. Collins, and T. Darrell, "Conditional random fields for object recognition," in *Advances in neural information processing systems*, 2005, pp. 1097–1104.

[16] J. Yang, J. Lu, S. Lee, D. Batra, and D. Parikh, "Graph r-cnn for scene graph generation," *arXiv preprint arXiv:1808.00191*, 2018.

[17] F. D. de Souza, S. Sarkar, A. Srivastava, and J. Su, "Spatially coherent interpretations of videos using pattern theory," *International Journal of Computer Vision*, vol. 121, no. 1, pp. 5–25, 2017.

[18] S. Aakur, F. D. de Souza, and S. Sarkar, "Towards a knowledge-based approach for generating video descriptions," in *2017 14th Conference on Computer and Robot Vision (CRV)*. IEEE, 2017, pp. 24–31.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE CVPR*, 2016.

[20] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *IEEE CVPR*, 2017, pp. 4700–4708.

[21] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *IEEE CVPR*, 2009.