WTA/TLA: A UAV-captured Dataset for Semantic Segmentation of Energy Infrastructure

Georgios Zampokas[*1,2†], Evangelos Skartados[*1†], Dimitrios Alexiou[1], Kosmas Tsiakas[1], Ioannis Tzanakis[3], Nikolaos Roussos[3], Dimitrios Giakoumis[1], Ioannis Kostavelis[1], Christos-Savvas Bouganis[2], Dimitrios Tzovaras[1]

***Abstract*— Automated inspection of energy infrastructure with Unmanned Aerial Vehicles (UAVs) is becoming increasingly important, exhibiting significant advantages over manual inspection, including improved scalability, cost/time effectiveness, and risks reduction. Although recent technological advancements enabled the collection of an abundance of vision data from UAVs' sensors, significant efforts are still required from experts to interpret manually the collected data and assess the condition of energy infrastructure. Thus, semantic understanding of vision data collected from UAVs during inspection is a critical prerequisite for performing autonomous robotic tasks. However, the lack of labeled data introduces challenges and limitations in evaluating the performance of semantic prediction algorithms. To this end, we release two novel semantic datasets (WTA and TLA) of aerial images captured from power transmission networks and wind turbine farms, collected during real inspection scenarios with UAVs. We also propose modifications to existing state-of-the-art semantic segmentation CNNs to achieve improved trade-off between accuracy and computational complexity. Qualitative and quantitative experiments demonstrate both the challenging properties of the provided dataset and the effectiveness of the proposed networks in this domain.**

The dataset is available at: `https://github.com/gzamps/wta_tla_dataset`.

I. INTRODUCTION

The high voltage power cable network is responsible for transferring electricity and spans across the ground globally. Meanwhile, considering the fact that wind energy met 16% of European electricity demand in 2020, and is projected to increase to 20% by 2022 and 25% by 2025, according to the European Commission [1], the number of wind turbines and their size are increasing. Thus, there is an immense need for the inspection automation of such facilities, to allow the smooth and efficient operation of infrastructure such as wind turbine farms, power transmission networks and solar panel installations, which usually extend to large areas and demonstrate a sparse structure. Moreover, such facilities are often located in distant areas, hard to be reached, with adverse weather and environmental conditions. Thus, they introduce challenges in maintenance and inspection, especially when it is performed by human workers. Aerial supervision and surveillance with UAVs is becoming more popular, due to its flexibility, safety and cost-effectiveness, compared with human supervision [25]. There is set of tasks that often included in the inspections of large targets with UAVs, such as object detection, navigation over specified target and close range inspection. This requires some sort of semantic knowledge, which can be provided a-priori or extracted during flight. In order to facilitate a methodology and serve as a reference point for similar problems, this work focuses on the delivery of a dataset with aerial visual instances captured during real inspection of transmission lines and wind turbines to enable the automated extraction of useful semantic information from visual feedback. More specifically, in the transmission towers domain, the inspection tasks are 1) navigation over power cables, 2) detection of transmission towers as points of interest, 3) detection of vegetation around transmission towers which may be hazardous for their regular operation and 4) detection of insulators and acquisition of close-up views for further inspection from experts or base station.

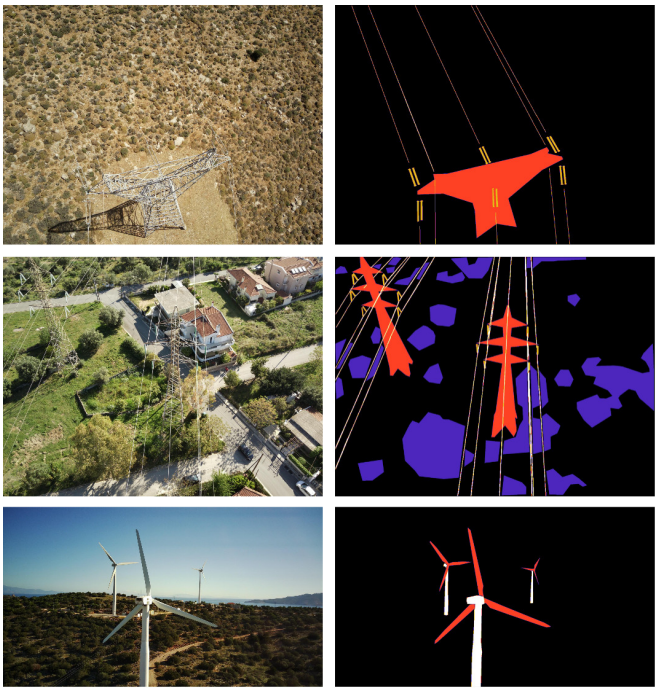

Fig. 1. Sample images from TLA and WTA datasets in the left column along with corresponding semantic ground truth maps in the right column.

∗Denotes equal contribution to this work
† Emails: {gzampokas, eskartad}@iti.gr
[1]Information Technologies Institute, The Centre for Research and Technology Hellas, Thessaloniki, Greece
[2]Imperial College London, London, United Kingdom
[3]Public Power Corporation S.A., Innovation Hub Athens, Greece

Similar tasks are defined for wind turbines including: 1) detection of wind turbine towers from range and 2) detection of wind turbine blades and acquisition of close-up views for further inspection.

In order to deal with the above-mentioned inspection tasks, it is essential for UAVs to be integrated with software modules that will facilitate their autonomous navigation through context-aware path planning on information-rich semantic maps. Integration of UAVs with RGB cameras and LiDAR sensors, and their joint calibration with semantic annotations allows a 3D semantic representation [26]. This enhanced perception system enables UAVs to conduct automated close proximity inspection in wind turbines, power towers and their components, register defects and components in a 3D map and finally transform it into a semantic map that can be apprehended by human operators. The resulting reckoning of the 3D position of targets leads towards optimal path planning and automated inspection in energy infrastructures. To create and process visual feedback with respect to the discussed scenarios, we gather video sequences from flights of manually operated drones that were performing the above tasks.

Extracting semantic information from visual feedback is important for ensuring safe and correct navigation by providing the basic knowledge behind the mapped environment, defining targets, and avoiding collisions and dangerous locations. Semantic segmentation is recent computer vision task with the goal of assigning a class to every pixel in the input image. Within the context of energy facility inspection, pixel locations of critical objects such as transmission towers, cables, insulators and wind turbine blades can be predicted yielding semantically rich visual scene representations.

We summarize our contributions within the context of visual inspection of wind turbine farms and power transmission networks:

- A new semantic segmentation dataset with images from various locations of installed transmission towers and wind turbines, along with carefully annotated relevant classes.
- Adaptation of state-of-the-art semantic segmentation network for real-time optimal operation in this semantic domain using embedded GPU hardware.

II. RELATED WORK

A. Visual Inspection of Energy Facilities

Recent work has mainly focused on visual inspection of defects from close range images. Qiu [2] proposed an autonomous visual inspection system for wind turbine blade inspection, which combines YOLO model [3] with an additional CNN into a deep learning pipeline. Shihavuddin [4] used the R-CNN object detection framework, with an Inception-ResNet-V2 architecture backbone, to classify locations on turbine blades in four defect types which often appear in blades. Shihavuddin [5] suggested a holistic approach for inspection of various targets including wind turbine blades and solar panels, using a YOLO and EfficientDet variants as detectors.

Fig. 2. Different types of wind turbine models included in the Wind Turbine dataset.

Fig. 3. Transmission towers and power cable views from the 4 different locations included in the Power Cable dataset.

The previous approaches consider the problem of processing images captured from UAVs, however efficient scene understanding requires semantic information about the depicted scene. Towards that direction, Moolan-Feroze [6] studied the problem of simultaneous drone localisation and model fitting for accurately mapping the target wind turbine object. A skeleton parameterisation model of the wind turbine is used and optimized, along with a pose graph representation of the drone's 3-D trajectory. Abdelfattah [7] proposed an instance segmentation approach for aerial images containing transmission towers and power lines, also arguing the importance of semantic knowledge in such scenarios.

B. Semantic Segmentation

The effectiveness of most current semantic segmentation algorithms depends on capturing long-range dependencies between picture pixels and contextual interactions. [8],[9] achieve that using dilated convolutions, effectively increasing the receptive field of the network. The latter employs a decoder to recover high resolution features lost in downsampling. Spatial Pyramid Pooling [10] is another technique to expand the receptive field of a network by fusing features from different strides. Aforementioned methods are

TABLE I
ENERGY FACILITY RELEVANT DATASET SPECIFICATIONS AND ATTRIBUTES

	DTU-Nordtank	TTPLA	WTA	TLA
Image Resolution	5280x2970	3840×2160	1920x1080	4000x3000
Training Samples	459	830	187	306
Validation Samples	100	270	37	43
Total Samples	559	1100	224	349
Classes	Blade, Tower	Cable, Lattice Tower Concrete Tower, Wooden Tower	Blade, Tower	Tower, Insulator Cable, Vegetation
Capture Environments	2	80	3	4
Capture Distance (m)	5-40	N/A	5-100	40-60
Semantic Ground Truth	No	Yes	Yes	Yes

performance oriented. However, when targeting real-time inference, a balance between accuracy and computational complexity needs to be established. SegNet [11] proposed an encoder-decoder architecture with skip connections and transposed convolutions to handle upsampling, while GUNet [12] fuses information from multi-scale input images using a guided upsampling module. More recently, dual branch approaches have gained ground due to their ability to split information flow in two complementary paths. BiSeNetV1 [13] and BiSeNetV2 [14] increase the receptive field by introducing a deep path with strided convolutions, while a second parallel path extracts high-resolution information. DDRNets [15] propose an efficient architecture with two branches and introduce information flow between them by using multiple bilateral fusions. Multi-scale and global context extraction is handled by a Deep Aggregation Pyramid Pooling Module (DAPPM). In this work, we use a lightweight DDRNet variant as baseline for our experimental evaluation.

C. Traditional Semantic Segmentation Datasets

With the advent of deep learning technologies for the semantic segmentation task the need for densely annotated data has emerged. Several datasets have been introduced that enabled research in various traditional computer and robotic vision fields ranging from generic scene understanding (PASCAL VOC 2012) to autonomous driving (Cityscapes, CamVid).

PASCAL VOC 2012 [16] contains images of varying resolutions, up to 500 × 500, representing 21 classes (including a class for the background). This set originally contained 1,464 training, 1,449 validation, and 1,456 test images. Its training set was later extended by others to a total of 10,582 images.

Cityscapes [17] is one of the most popular urban street scene segmentation datasets, with 19 classes relevant to urban driving, such as roads, sidewalk, pedestrians etc. It consists of 2975 training, 500 validation and 1525 test samples.

CamVid [18] is a similar street scene dataset to Cityscapes, containing densely annotated frames of 960x720 resolution. More specifically, it includes 367 photos for training, 101 images for validation, and 233 images for testing.

D. Aerial Semantic Segmentation Datasets

Proliferation of autonomous aerial vehicles introduced a new branch of computer vision that focuses on automated processing of aerial images. Traditional computer vision tasks, including semantic segmentation, play an integral role within this new branch. However, they need to be applied on a domain substantially shifted compared to the operational domain of terrain-based agents. Visual appearance of same classes of interest (e.g. human, car, street, vegetation, building) in aerial and traditional images significantly changes. Additionally, it is common for objects to appear at multiple scales within the same scene. Aforementioned challenges are attributed to the combination of the following factors: New viewpoints of objects of interest (top-down views in contrast to side views in traditional computer vision), variations in the altitude of the camera (from a few meters to a few hundred meters) and a wide field of view. What is more, modern UAVs are equipped with HD cameras able to capture high resolution images. As a result new performance to efficiency trade-offs need to be sought, especially for agents that require real-time visual feedback. Following datasets have been lately introduced targeting the semantic segmentation task within the aerial image domain:

Aeroscapes [27] consists of 3269 images at the 1280x720 resolution, extracted from 149 separate video sequences. They were captured by a fleet of drones operating at altitudes varying from 5 to 50 meters. Annotated labels are separated in the ”stuff” (vegetation, roads, sky, construction) and the ”things” (person, bike, car, drone, boat, obstacle, animal) groups.

UAVid [8] contains 300 images at the 4096x2160 and the 3840x2160 resolutions, from 30 separate video sequences. Camera viewpoint is set to oblique and flying height is at around 50 meters. Captured scenes depict complex urban environments and 8 labels are annotated (building, tree, clutter, road, vegetation, static car, moving car, human).

Semantic Drone Dataset [6] focuses on safe drone autonomous flying and landing procedures. It provides scenes of domestic/urban environments with 20 annotated labels. 400 train set images are captured at more than 20 different locations, with a nadir view, at varying (3-30 meters) heights, at the 6000x4000 resolution.

TABLE II

PERCENTAGE OF PIXELS OF BELONGING IN CLASSES FOR TLA DATASET

Label	location 1	location 2	location 3	location 4
background	83.43%	59.34%	94.28%	29.88%
vegetation	12.55%	36.79%	-	63.74%
transmission tower	2.03%	2.41 %	4.79%	5.45 %
insulator	0.14%	0.04%	0.13%	0.05%
power line cable	1.86%	1.41%	0.79%	0.86%

TABLE III

PERCENTAGE OF PIXELS FOR BELONGING IN CLASSES OF WTA DATASET

Label	location 1	location 2	location 3
background	90.79%	93.48%	96.73%
blade	6.51%	4.51%	1.65%
tower	2.69%	2.04 %	1.62%

E. Energy Facility Inspection Datasets

The aforementioned datasets are aimed for semantic segmentation problems, however their content is quite different from our target domain. Focusing on data captured at energy facilities, the **DTU-Nordtank** [22] dataset contains images from Nordtank wind turbines, captured by a UAV at wind facilities in Roskilde, Denmark. It contains data captured at two subsequent years, 2017 (161 samples) and 2018 (398 samples), resulting in 559 total samples. The image content is mostly close-up views of wind turbine towers and blades with image resolution is 5280x2970 pixels. This dataset is aimed for wind turbine inspection for faults or defects, however it does not include any sort of ground truth. Since the dataset provides multiple views of wind turbines, we use it for our semantic segmentation task. To achieve that, within the context of this work, we manually annotate the relevant classes for all the samples in the dataset and divide it into 458 train and 100 test samples.

TTPLA [7] is a public image dataset for instance segmentation, which includes power lines and transmission towers from USA. It consists of 1,100 aerial images of 3840×2160 resolution, captured with a Parrot-ANAFI UAV, ensuring variety in terms of view angles, scales, backgrounds, lighting conditions and zooming levels.

III. DATASETS

To harness the power of end-to-end trainable CNNs, ground truth data is essential to train and evaluate their performance. When focusing on real applications, it is important to train on data with similar properties and conditions with the actual application. This reduces the domain-shift when transferring knowledge from trained models to deployment scenarios, ensuring more robust performance. Thus, we introduce two new datasets for semantic segmentation, **W**ind **T**urbine **A**erial-image dataset **(WTA)** and **T**ransmision **L**ine **A**erial-image **(TLA)** dataset relevant to inspection of the respective energy infrastructure facilities.

A. Data Capture

Our data recording and annotation methodology was designed to capture variability of environment conditions and actual targets. Hundreds of frames are captured from a moving UAV in three different wind turbine farms in the region of Attica, Greece. A DJI Matrice 600 Pro drone was used for the data acquisition, recording videos at 1920x1080 pixel resolution. The minimum flight altitude is 3 meters and can reach up to 40 meters, which is the highest part of the wind turbines. In WTA dataset, the drone starts from a long distance and gradually moves towards the wind turbine targets, ensuring capturing of various scales.

In TLA dataset, we provide images extracted from video sequences at 4000x3000 resolution, recorded from manual drone flights at 4 different locations in the region of Attica, Greece. Flight scheduling was designed to resemble an automated electrical grid inspection. Consecutive transmission towers were defined as target waypoints and navigation was performed over the power line cables. The camera view was set to oblique ($\sim 45°$) and the flight altitude was about 50 meters (keeping a minimum 12 meters distance from the towers for safety reasons). For one of the locations we provide dense annotation of the whole flight while for the rest of them we extract frames around the transmission towers.

We attempt to increase the descriptiveness of the datasets while keeping the count of the samples relatively small, by minimizing redundancy of samples and keeping the most diverse ones. We manually annotate the contents of image samples by hand, using VIA Annotator tool [24] with the guidance of field experts, generating class ground truth maps of pixel granularity for each image sample.

B. Technical details/structure

Both datasets contain a set of RGB images accompanied by their corresponding semantic ground truth maps, which are generated by pixel-level annotations from humans. Images and ground truth maps of the WTA datasets have a resolution of 1920x1080 pixels and those of the TLA dataset

4000x3000 pixels. Each dataset contains classes relevant to the energy facility infrastructure it includes. We calculate the frequency of the classes within the samples, and Table II compares the distribution of annotations across individual classes for both datasets. In TLA dataset, background classes and often vegetation (location 2 and 4) usually occupy the almost 95% of the images, whereas the rest 5% contains the relatively small objects of interest (transmission towers, insulators and power line cables). Similarly in WTA, over 90% of the pixels belong to background class, while blade class is more frequent than tower, since the dataset contains frames from close-range inspection of blades. Overall, both datasets are background heavy, with small percentage of pixels accounting for objects of interest. Even though percentages are low for relatively small objects, there are some basic structural differences between blades and cables/insulators. The former are generally thicker objects which occupy a specific spatial location in an image, whereas the latter are much thinner and can extend to multiple locations. Due to costly annotations, we omit including a validation test and simply divide the datasets in train and test splits, using 85:15 train to test ratio.

IV. Semantic Segmentation

Recent deep learning approaches have demonstrated successful results in popular benchmark datasets presented in Section II-C. We attempt to provide a similar baseline for evaluating semantic segmentation performance in the challenging scenarios of energy facility infrastructure inspection. Therefore, we evaluate the performance of the recently proposed DDRNet [15] semantic segmentation approach on our datasets, and explore if its success is transferred to this narrower semantic domain.

There are two important issues which need to be considered in actual deployment of semantic segmentation in real scenarios using hardware mounted on UAVs. First, our target data contains structures with different properties and most importantly various sizes and scales. For example, power cables are very thin structures, which typically resemble lines, while insulators and wind turbines are also small sized when depicted from distance, which introduces the need for high resolution in both input and output of the semantic segmentation process. This is typically solved by using more complex segmentation CNNs, able to operate at higher resolutions at the cost of increased complexity. However, when aiming at real-time execution, the complexity of the network is limited by the capabilities of the target device. Targeting deployment on UAVs, introduces an additional limitation on the target device, from an energy and weight perspective, resulting in the use of embedded GPUs. Therefore, the need for accurate segmentation and real-time execution on relatively small hardware, casts this task highly challenging and introduces the need to strike a balance between accuracy and computational complexity.

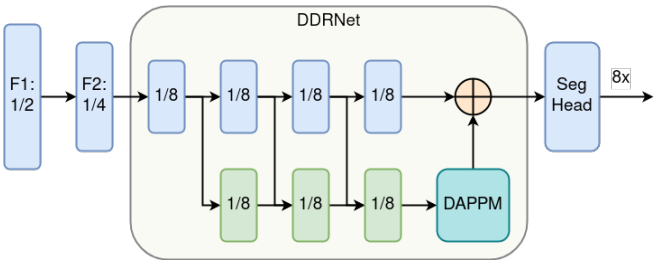

Fig. 4. Deep Dual-resolution Network (DDRNet) Architecture. Fractions denote tensor size with respect to input resolution.

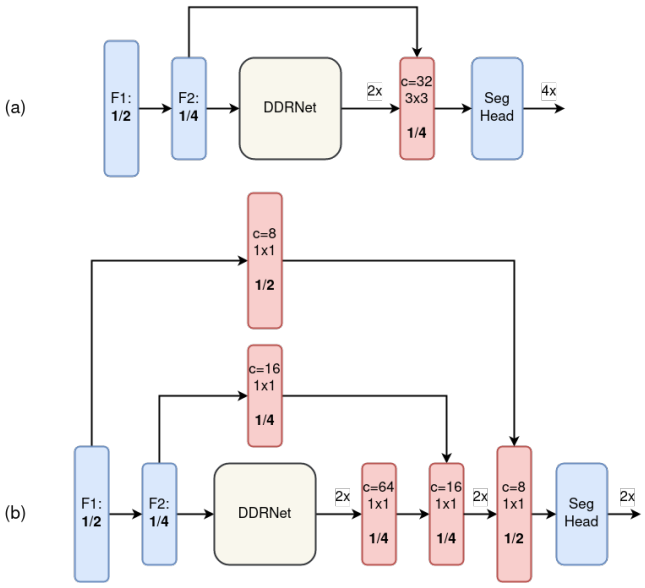

Fig. 5. Decoders on top of DDRNet architecture: (a) hix2 decoder upsampling to 1/4 of input resolution and (b) hix4 decoding to 1/2 input resolution. Bold numbers indicate feature size with respect to input resolution.

A. Baseline Model

We utilize DDRNet, depicted in 4, as our baseline model. Its building block is a simple Residual Block that applies the Convolution-Batch Normalization-ReLU activation sequence twice to the input feature maps. DDRNet architecture follows the multiple pathway approach. Specifically, the main trunk applies two fast convolutions to the input image to downsample it by 4. Subsequently, two Residual Blocks reduce the output stride to 1/8. At this point the main trunk branches out to two separate paths. The low-resolution path applies 3 Residual Blocks to the 1/8 feature maps. At each stage strided convolutions are utilized and as a result feature maps are downsampled by 2, while at the same time their dimensionality is doubled. The final output of the low-resolution branch has an 1/64 output stride. The high-resolution path is built with a one to one correspondence to low-resolution's blocks. However, it preserves the 1/8 resolution of the feature maps since no striding is involved in its convolutions. High resolution path remains shallow producing feature maps with a constant number of channels. Features from the high and the low resolution branches are fused after every processing stage with connections

that involve downsampling/upsampling and reprojection operations. At the end of the low-resolution path a Deep Aggregation Pyramid Pooling Module (DAPPM) is applied that performs multiple parallel convolutions with large kernel sizes and global pooling. Output is upsampled to the 1/8 resolution of the high-resolution and a segmentation head is responsible for transforming the sum of the two branches and producing the final probability map for each class of the segmentation task.

Low-resolution branch with the subsequent DAPPM operate as the context path, increasing the receptive field of the final output and guaranteeing global reasoning. High-resolution branch preserves all the spatial information necessary to produce accurate results. Finally, bilateral fusion achieves mixing spatial with semantic information.

B. DDRNet-Hi Model

DDRnets have demonstrated good performance when it comes to popular benchmark datasets, like Cityscapes and CamVid which are targeted to urban driving scenarios. Considering the domain of the proposed dataset, there are two important differences from the urban driving scenarios. First, according to the data from II and III our dataset has a large sparsity factor since the largest part of the images is occupied by the background class, while the important contents of images are thin details which require large resolutions for good visibility. This leads to contradictory requirements that need to be met by suitable design choices when applying semantic segmentation CNNs on the specific dataset. The sparsity of the images along with the hardware capabilities of target devices, point towards using efficient networks with multiple downsamplings, to avoid redundant calculations in the unimportant background regions. However, accurately classifying small image contents requires processing in high resolutions, since such detail is lost in lower resolutions.

Towards adapting DDRNets for our target tasks, we propose the addition of a decoder module on top of the output of DDRNet, which is responsible for recovering the semantic details lost in the backbone due to multiple downsampling layers. We propose two decoder architectures, one for each dataset, which can be seen in Figure 5. The first decoder hix2, contains simply a skip connection from features of $1/4$ resolution to the upsampled output of DDRNet, which is then processed by a 3x3 Conv-BN-ReLU layer of 32 channels before entering the segmentation head module. The final output is at the $1/4$ of image input resolution. The second decoder hix4, introduces 2 skip connections with redirection layers from 1/2 and 1/4 initial feature layers, which gradually upsample and refine DDRNet output, resulting in a final prediction with 1/2 of input resolution. We focus on keeping a small number of channels while increasing resolution to avoid increasing computational complexity compared to the base network.

TABLE IV
DECODER ABLATION FOR WTA AND TLA DATASETS, FOR DDRNET23-SLIM BASE ARCHITECTURE, NOTING MIOU% FOR THE RELATIVELY SMALL OBJECT CLASSES

TLA Dataset				
Decoder	GFLOPs	mIOU%		
		total	insulator	cable
-	**14.06**	72.27	53.76	47.01
RegSeg	19.10	**77.38**	**54.06**	60.52
hix2	15.88	75.52	45.66	59.3
hix4	15.59	76.80	62.50	**67.28**
WTA Dataset				
Decoder	GFLOPs	mIOU%		
		total	blade	tower
-	**9.07**	91.91	87.1	89.13
RegSeg	12.79	91.82	87.34	88.60
hix2	10.28	**92.17**	**88.09**	**88.90**

TABLE V
RESOLUTION ABLATION FOR WTA AND TLA DATASETS, USING THE DDRNET23-SLIM-HI ARCHITECTURE WITH VARIOUS INPUT RESOLUTIONS

TLA Dataset				
Input Resolution	GFLOPs	mIOU%		
		total	insulator	cable
384x512	3.85	66.53	28.07	34.85
768x1024	15.88	**75.52**	**45.66**	**59.3**
WTA Dataset				
Input Resolution	GFLOPs	mIOU%		
		total	blade	tower
256x512	**2.57**	89.27	84.04	84.38
512x1024	10.28	**92.17**	**88.09**	**88.90**

V. EXPERIMENTAL RESULTS

A. Training Setting

We choose the lightest model from the family of DDRNets, namely the DDRNet23-slim model for our experimental base. Our model is trained with the WTA and DTU-Nordtank datasets for the wind turbine segmentation task and the TLA dataset for the tower and cable segmentation task. We train our model minimizing the Cross Entropy Loss for 300 epochs and batch size of 8 and use the Adam optimizer an initial learning rate of 0.01 which decreases every epoch according to poly learning policy as with 0.9 exponent. Finally for the TLA dataset, we propose a class loss weighting scheme of 0.8, 1, 1, 3, 3 for background, vegetation, tower, insulator and cable classes. Experiments have proven that it helps the network focus on relevant small classes (insulator, cable) which offer important semantic content.

B. Decoder Ablation

The decoder module is evaluated to measure it's impact in both segmentation accuracy and computational complexity. We examine the two decoder modules presented in Section IV-B plus the decoder proposed in [23]. Adding an extra

TABLE VI

QUANTITATIVE RESULTS OF OF DDRNET23-SLIM BASELINE AND PROPOSED MODEL FOR WIND TURBINE AERIAL (WTA) AND DTU-NORDTANK DATASETS

Model	Input Resolution	GFLOPs	Params (M)	Wind Turbine Aerial mIOU%				DTU-Nordtank mIOU%			
				total	bg	blade	tower	total	bg	blade	tower
DDRNet23-slim	512x1024	**9.07**	5.67	91.91	99.51	87.10	**89.13**	94.04	99.21	94.55	88.37
DDRNet23-slim-hix2	512x1024	10.27	5.67	**92.17**	**99.52**	**88.09**	88.90	**94.32**	**99.25**	**94.62**	**89.09**

TABLE VII

QUANTITATIVE RESULTS OF DDRNET23-SLIM BASELINE AND PROPOSED MODEL FOR TRANSMISSION LINE AERIAL (TLA) DATASET

Model	Input Resolution	GFLOPs	Params (M)	Transmission Line Aerial mIOU%					
				total	bg	vegetation	tower	insulator	cable
DDRNet23-slim	768x1024	**14.06**	5.67	68.48	**93.91**	76.18	82.52	52.32	33.98
DDRNet23-slim-hix4	768x1024	15.59	5.67	**80.35**	93.76	**79.22**	**90.72**	**62.50**	**67.80**

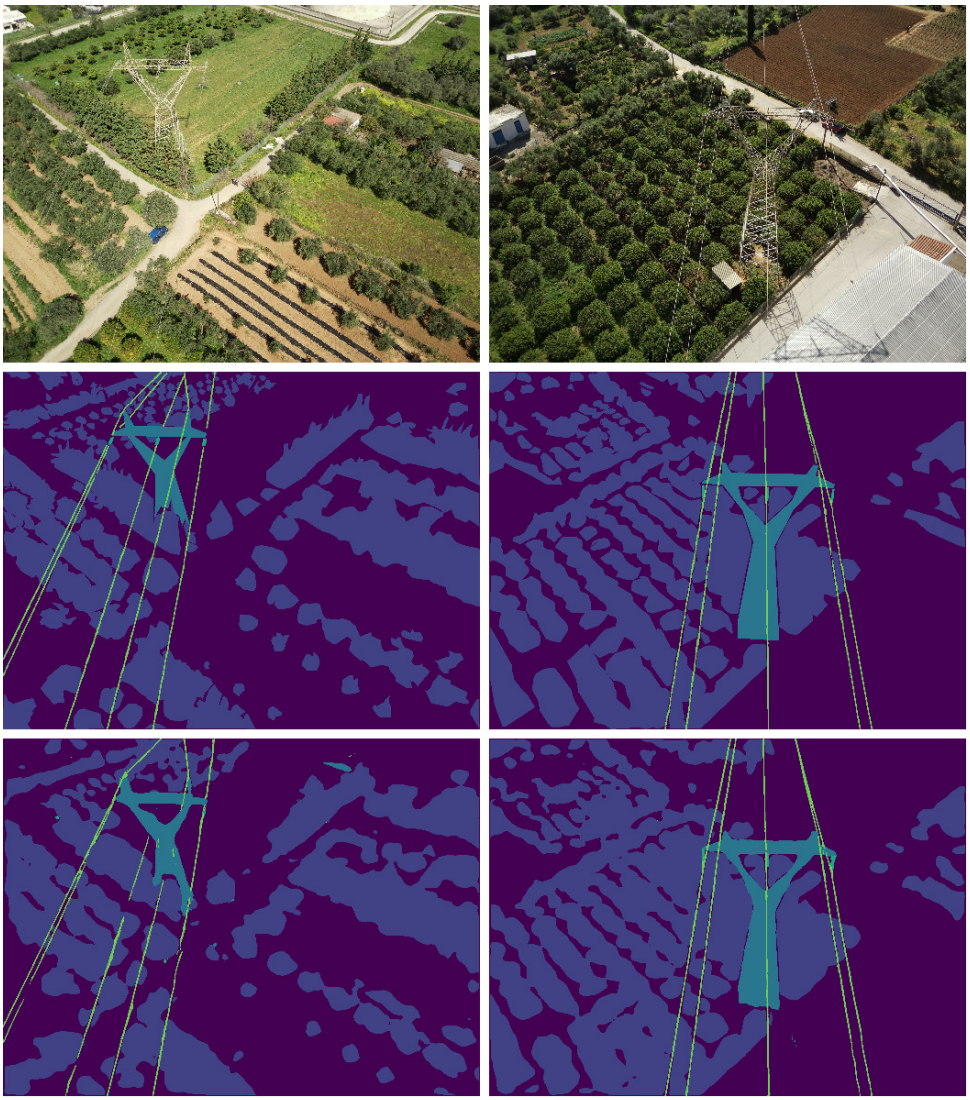

Fig. 6. TLA Line risky vegetation detection scenario: The UAV navigates over the power lines, and gathers semantic information on vegetation around towers, which can be risky for their operation. Top row contains RGB images, middle row contains ground truth data and bottom row contains DDRNet23-slim-hix4 segmentation maps

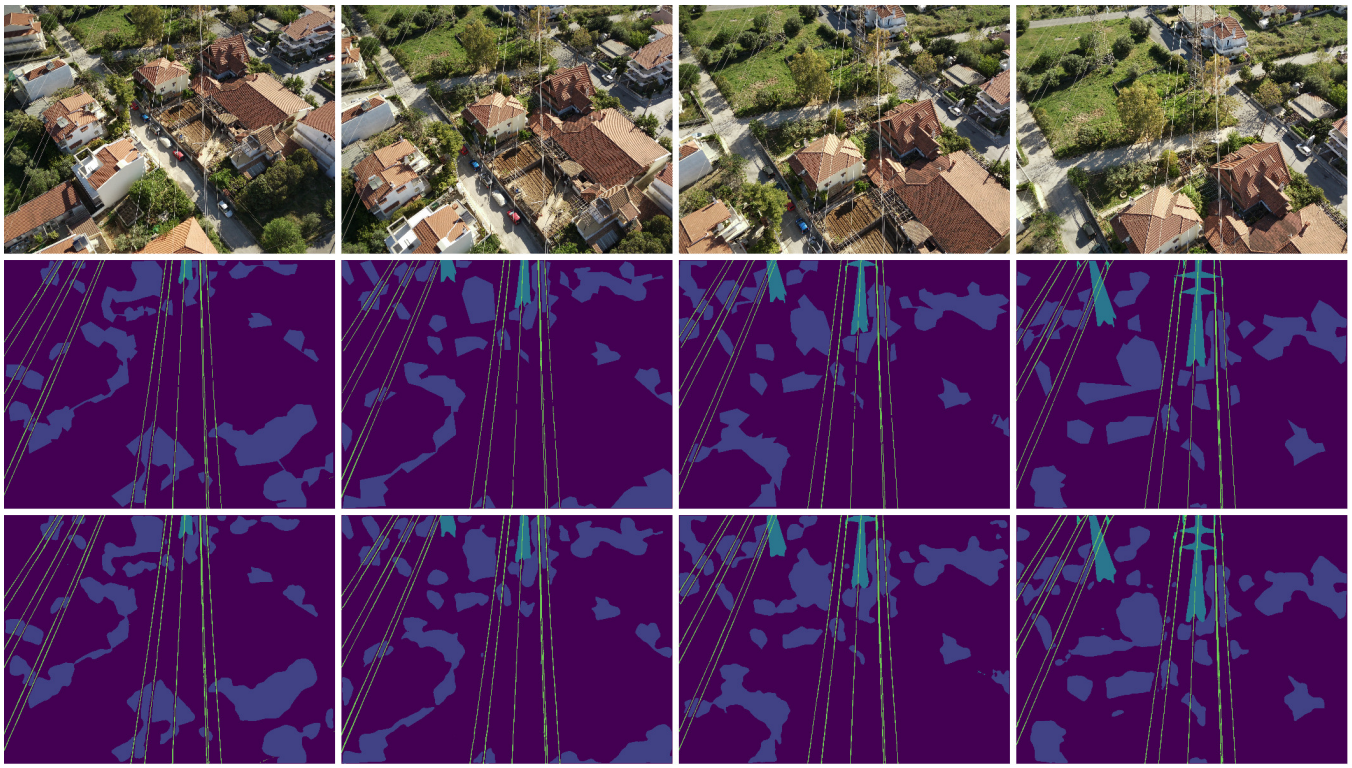

Fig. 7. TLA Line navigation scenario: The UAV navigates over the power lines. Top row contains RGB images, middle row contains ground truth data and bottom row contains DDRNet23-slim-hix4 segmentation maps.

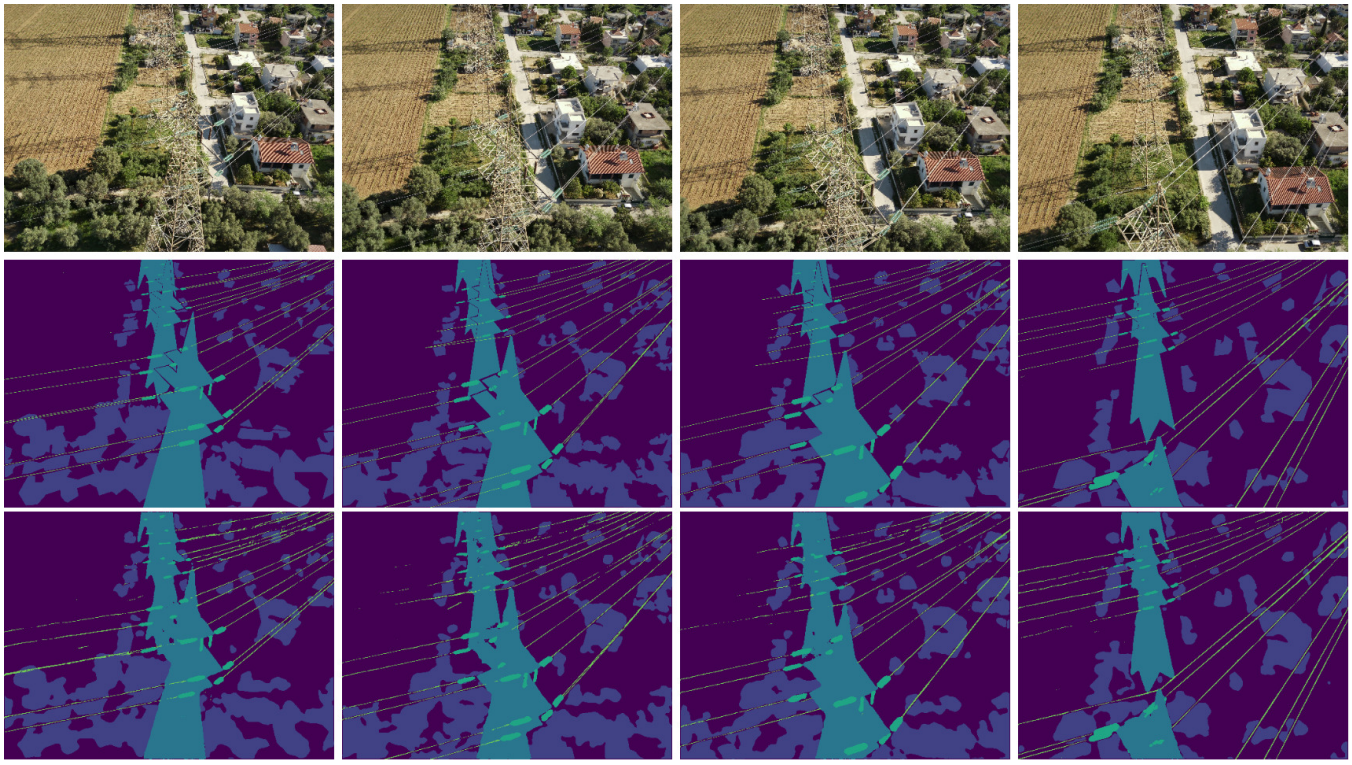

Fig. 8. TLA insulator inspection scenario: The UAV detects insulators in transmission towers and moves closer to get a better view of them. Top row contains RGB images, middle row contains ground truth data and bottom row contains DDRNet23-slim-hix4 segmentation maps.

decoder module on top of the existing DDRNet architecture yields important accuracy gains in all classes and datasets, however it is especially impactful in small classes such as cables (Table IV).

Fig. 9. WTA Wind turbine navigation scenario: The UAV detects a wind turbine target and moves closer to it. Top row contains RGB images, middle row contains ground truth data and bottom row contains DDRNet23-slim-hix2 segmentation maps.

Fig. 10. WTA Wind turbine inspection scenario: The UAV detects a wind turbine blade and reaches for closer inspection. Top row contains RGB images, middle row contains ground truth data and bottom row contains DDRNet23-slim-hix2 segmentation maps.

C. Resolution Ablation

Since our semantic targets involve thin structures and shapes, operating at high resolution is key for accurately recovering them. We perform experiments at different resolutions to measure the impact of increasing or decreasing the base image resolution, as seen in Table V. Smaller resolution in TLA dataset demonstrate a large performance drop in both total miou% and especially in small object classes, as expected. However, the case is different in WTA dataset where the model with reduced input resolution performs comparably to its higher input resolution counterpart. This is an important finding regarding, since it also comes with mobile-level computational complexity, making it suitable for applications with very limited available computational resources.

D. Discussion

We perform a full quantitative evaluation of the baseline DDRNet23-slim and our variants with the added decoder modules. Overall performance of DDRNet23-slim-hix4 is much improved from the baseline in TLA dataset VII, showing important accuracy gains across all classes, due to the larger output resolution of the model. This increases the achieved mIOU around 17% while adding 1.50 extra GFLOP in computational complexity. The addition of the decoding module is less impactful in WTA and DTU-Nordtank datasets, demonstrating a mere performance gain

of around 0.27% for both datasets.

Qualitative evaluation is also performed by providing RGB images, semantic predictions and ground truth for frame sequences resembling actual robotic tasks performed by the drone. The model retrieves most of the critical semantic content of the images, allowing satisfying perception of the environment during flight.

Overall, the contributed datasets pose interesting challenges for the semantic segmentation task, such as distinction between grass and taller vegetation and the introduction of small object views (wind turbines viewed from a large distance and insulators and thin cables). Additionally, semantic ground truth is provided for scenes which are rarely found annotated, providing a starting point for developing scene understanding aimed at facility inspection applications.

VI. CONCLUSIONS

In this work, a new semantic segmentation dataset with images from real inspection of wind turbine farms and transmission towers and lines has been introduced. Data are gathered from multiple locations and under various conditions and challenging views, aiming to serve as a training and evaluation benchmark towards automating inspection application. A set of modifications to adapt a state-of-the-art semantic segmentation CNN to these datasets is also proposed, accompanied with quantitative and qualitative analysis and experiments.

ACKNOWLEDGMENT

This research has been co-financed by the European Union and Greek national funds through the Operational Program Competitiveness, Entrepreneurship and Innovation, under the call RESEARCH – CREATE – INNOVATE (project name AUTOPSY, code: T1EDK- 04005).

REFERENCES

[1] https://windeurope.org/intelligence-platform/product/wind-energy-in-europe-in-2020-trends-and-statistics/

[2] Z. Qiu, S. Wang, Z. Zheng and D. Yu (2019) Automatic visual defects inspection of wind turbine blades via YOLO-based small object detection approach. Journal of Electronic Imaging 28, p 1-11

[3] Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 779-788).

[4] Shihavuddin, A. S. M., Chen, X., Fedorov, V., Nymark Christensen, A., Andre Brogaard Riis, N., Branner, K., ... & Reinhold Paulsen, R. (2019). Wind turbine surface damage detection by deep learning aided drone inspection analysis. Energies, 12(4), 676.

[5] ASM Shihavuddin, Mohammad Rifat Ahmmad Rashid, Md Hasan Maruf, Muhammad Abul Hasan, Mohammad Asif ul Haq, Ratil H. Ashique, Ahmed Al Mansur (2021) Image based surface damage detection of renewable energy installations using a unified deep learning approach, Energy Reports, Volume 7, Pages 4566-4576

[6] O. Moolan-Feroze, K. Karachalios, D. N. Nikolaidis and A. Calway, "Simultaneous Drone Localisation and Wind Turbine Model Fitting During Autonomous Surface Inspection," 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2019, pp. 2014-2021, doi: 10.1109/IROS40897.2019.8968247.

[7] Abdelfattah, R., Wang, X., & Wang, S. (2020). Ttpla: An aerial-image dataset for detection and segmentation of transmission towers and power lines. In Proceedings of the Asian Conference on Computer Vision.

[8] Chen, L. C., Papandreou, G., Schroff, F., & Adam, H. (2017). Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587.

[9] Chen, L. C., Zhu, Y., Papandreou, G., Schroff, F., & Adam, H. (2018). Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European conference on computer vision (ECCV) (pp. 801-818).

[10] He, K., Zhang, X., Ren, S., & Sun, J. (2015). Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE transactions on pattern analysis and machine intelligence, 37(9), 1904-1916.

[11] Badrinarayanan, V., Kendall, A., & Cipolla, R. (2017). Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE transactions on pattern analysis and machine intelligence, 39(12), 2481-2495.

[12] Hu, P., Caba, F., Wang, O., Lin, Z., Sclaroff, S., & Perazzi, F. (2020). Temporally distributed networks for fast video semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 8818-8827).

[13] Yu, C., Wang, J., Peng, C., Gao, C., Yu, G., & Sang, N. (2018). Bisenet: Bilateral segmentation network for real-time semantic segmentation. In Proceedings of the European conference on computer vision (ECCV) (pp. 325-341).

[14] Yu, C., Gao, C., Wang, J., Yu, G., Shen, C., & Sang, N. (2021). Bisenet v2: Bilateral network with guided aggregation for real-time semantic segmentation. International Journal of Computer Vision, 129(11), 3051-3068.

[15] Hong, Y., Pan, H., Sun, W., & Jia, Y. (2021). Deep dual-resolution networks for real-time and accurate semantic segmentation of road scenes. arXiv preprint arXiv:2101.06085.

[16] Everingham, M. and Van Gool, L. and Williams, C. K. I. and Winn, J. and Zisserman, A. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results", (2012)

[17] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, The Cityscapes Dataset for Semantic Urban Scene Understanding, (2016)in Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)

[18] Brostow, G. J., Fauqueur, J., & Cipolla, R. (2009). Semantic object classes in video: A high-definition ground truth database. Pattern Recognition Letters, 30(2), 88-97.

[19] http://dronedataset.icg.tugraz.at

[20] Nigam, I., Huang, C., & Ramanan, D. (2018, March). Ensemble knowledge transfer for semantic segmentation. In 2018 IEEE Winter Conference on Applications of Computer Vision (WACV) (pp. 1499-1508). IEEE.

[21] Lyu, Y., Vosselman, G., Xia, G. S., Yilmaz, A., & Yang, M. Y. (2020). UAVid: A semantic segmentation dataset for UAV imagery. ISPRS Journal of Photogrammetry and Remote Sensing, 165, 108-119.

[22] https://data.mendeley.com/datasets/hd96prn3nc/2

[23] Gao, R. (2021). Rethink Dilated Convolution for Real-time Semantic Segmentation. arXiv preprint arXiv:2111.09957.

[24] https://www.robots.ox.ac.uk/~vgg/software/via/app/via_image_annotator.html

[25] A., Savva, A., Zacharia, R., Makrigiorgis, A., Anastasiou, C., Kyrkou, P., Kolios, ... T., Theocharides, T. ICARUS: Automatic Autonomous Power Infrastructure Inspection with UAVs. In 2021 International Conference on Unmanned Aircraft Systems (ICUAS) (pp. 918-926). IEEE.

[26] Zhong, H., Wang, H., Wu, Z., Zhang, C., Zheng, Y., Tang, T. (2021). A survey of LiDAR and camera fusion enhancement. Procedia Computer Science, 183, 579-588.

[27] I. Nigam, C. Huang and D. Ramanan, (2018) Ensemble Knowledge Transfer for Semantic Segmentation, 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), 2018, pp. 1499-1508