IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 7 April 2023, accepted 5 May 2023, date of publication 11 May 2023, date of current version 18 May 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3275435

APPLIED RESEARCH

A Comparative Evaluation of Deep Learning Techniques for Photovoltaic Panel Detection From Aerial Images

EDOARDO ARNAUDO[1,2], GIACOMO BLANCO[2], ANTONINO MONTI[1], GABRIELE BIANCO[3], CRISTINA MONACO[3], PAOLO PASQUALI[2], AND FABRIZIO DOMINICI[2]

[1]Dipartimento di Automatica e Informatica (DAUIN), Politecnico di Torino, 10129 Turin, Italy
[2]Fondazione LINKS, 10138 Turin, Italy
[3]ITHACA S.r.l., 10138 Turin, Italy

Corresponding author: Edoardo Arnaudo (edoardo.arnaudo@polito.it)

This work was supported in part by the Projects: DYDAS (CEF Programme) under Grant 2018-IT-IA-0101, in part by ATLANTIS (Horizon Europe) under Grant 101073909, and in part by the Project NODES through the MUR—M4C2 1.5 of PNRR under Grant ECS00000036.

ABSTRACT Solar energy production has significantly increased in recent years in the European Union (EU), accounting for 12% of the total in 2022. The growth in solar energy production can be attributed to the increasing adoption of solar photovoltaic (PV) panels, which have become cost-effective and efficient means of energy production, supported by government policies and incentives. The maturity of solar technologies has also led to a decrease in the cost of solar energy, making it more competitive with other energy sources. As a result, there is a growing need for efficient methods for detecting and mapping the locations of PV panels. Automated detection can in fact save time and resources compared to manual inspection. Moreover, the resulting information can also be used by governments, environmental agencies and other companies to track the adoption of renewable sources or to optimize energy distribution across the grid. However, building effective models to support the automated detection and mapping of solar photovoltaic (PV) panels presents several challenges, including the availability of high-resolution aerial imagery and high-quality, manually-verified labels and annotations. In this study, we address these challenges by first constructing a dataset of PV panels using very-high-resolution (VHR) aerial imagery, specifically focusing on the region of Piedmont in Italy. The dataset comprises 105 large-scale images, providing more than 9,000 accurate and detailed manual annotations, including additional attributes such as the PV panel category. We first conduct a comprehensive evaluation benchmark on the newly constructed dataset, adopting various well-established deep-learning techniques. Specifically, we experiment with instance and semantic segmentation approaches, such as Rotated Faster RCNN and Unet, comparing strengths and weaknesses on the task at hand. Second, we apply ad-hoc modifications to address the specific issues of this task, such as the wide range of scales of the installations and the sparsity of the annotations, considerably improving upon the baseline results. Last, we introduce a robust and efficient post-processing polygonization algorithm that is tailored to PV panels. This algorithm converts the rough raster predictions into cleaner and more precise polygons for practical use. Our benchmark evaluation shows that both semantic and instance segmentation techniques can be effective for detecting and mapping PV panels. Instance segmentation techniques are well-suited for estimating the localization of panels, while semantic solutions excel at surface delineation. We also demonstrate the effectiveness of our ad-hoc solutions and post-processing algorithm, which can provide an improvement up to +10% on the final scores, and can accurately convert coarse raster predictions into usable polygons.

INDEX TERMS Computer vision, deep learning, image processing, machine learning, semantic segmentation, instance segmentation, remote sensing.

I. INTRODUCTION

Considering the European soil, solar energy production has grown significantly in recent years. In 2022, solar energy accounted for 12% of total energy production in the European

The associate editor coordinating the review of this manuscript and approving it for publication was Mohammad Shorif Uddin.

Union (EU) [1], making it one of the largest sources of renewable energy in the continent, together with hydroelectrical and wind [2]. The recent growth of solar energy production in Europe can be attributed to several factors: first and foremost, the increasing adoption of solar photovoltaic (PV) panels in households, businesses, and facilities. Given the established technology, PV panels have slowly become a cost-effective and efficient means of energy production, and their deployment has been steadily supported by government policies and incentives in many EU countries [3]. The consequent market growth and the derived maturity of solar technologies also contribute to a decrease in the overall cost of solar energy, making it more competitive with other energy sources.

Last but not least, its inherently renewable nature represents a major contributing factor: while, on one hand the cost of fossil fuel resources has been steadily growing, on the other hand solar energy is a clean and renewable source of electricity that does not produce greenhouse gases or other pollutants, making it an attractive alternative to other sources. This transition trend into sustainable sources is expected to grow significantly in the coming years, making solar energy an increasingly important asset [2].

As a result, there is a growing need for efficient methods for detecting and mapping the locations of PV panels, exploiting resources such as satellite and aerial images to produce a detailed and automatically updated census. Accurate detection of solar installations can in fact be extremely beneficial for several reasons. First and foremost, automated detection allows for rapid identification of PV panels, saving time and resources compared to a manual inspection. This is true for registered large-scale farms, as well as smaller agricultural or domestic plants, where the extent and location are often unknown. Second, the information contained in updated solar maps can be exploited by governments, environmental agencies and other companies to track the adoption of renewable sources or to optimize energy distribution across the grid. This can also include the identification of malfunctioning installations to cut down maintenance and repair costs and maintain high efficiency for longer periods. Last but not least, automatically detecting areas where PV panels are already deployed can also be exploited to identify promising regions where an increased deployment could provide greater benefits.

However, effective and accurate automated detections present several unique challenges, often derived from typical issues of aerial computer vision. Image quality and resolution are essential factors in determining the accuracy and reliability of models, particularly with regard to the recall of PV panels. Very-High Resolution (VHR) acquisitions, especially considering extents lower than $1m$ per pixel, can greatly benefit the final detection performance, and most importantly represent the only viable option for smaller installations. In fact, PV panels appear in a wide range of scales, from small residential components, partially covering roofs, to industrial-grade plants, covering entire fields. Moreover, varying lighting conditions or occluding structures such as trees, buildings or other constructions can negatively impact the delineation or the classification of a given panel into a specific category. Lastly, machine learning approaches typically require large amounts of high-quality annotations to be effective, which are often hard to come by, especially considering the specific task.

In this work, we attempt to address these problems using several incremental steps. First, we construct our own dataset of PV panels using VHR aerial imagery to cope with the lack of data, especially considering the European soil. We focus our efforts in Piedmont, one of the regions with the highest count of photovoltaic plants in Italy [4]. The dataset comprises 105 large-scale images, each measuring on average $20,000 \times 16,000$ pixels, and provides more than 9,000 accurate manual annotations including every known industrial plant and a large portion of smaller agricultural and domestic installations. Each polygon label is also enriched with several attributes, such as its category (polycrystalline or monocrystalline), the installation type (domestic, agricultural or industrial), orientation, area, and power generation estimate.

Exploiting this dataset, we then provide a comprehensive evaluation of different segmentation techniques, comparing instance and semantic approaches. The former comprises the generation of a single output for each detected panel, consisting of the predicted class, a bounding box defining its boundaries, and a segmentation mask delineating the perimeter [5]. While, on the one hand, instance segmentation outputs are very well suited for this task thanks to their ability to separate each panel into a single component, on the other hand, they are often more difficult to train due to the complexity of their architecture. Conversely, semantic segmentation can be instead defined as a classification of each pixel in the image [6], which allows for a precise localization of PV panels and background information. However, generating a probability distribution for each point independently may include some noise in the final predictions, consisting of incomplete panels (false negatives), background noise (false positives), or panel surfaces with mixed categories (classification errors).

The task of delineating PV panels also presents several unique challenges, including the wide range of scales of the installations, and the sparsity of the annotations, which cover a relatively small surface of the overall dataset. For these reasons, we propose to further improve the performance of these baselines through ad-hoc modifications, that specifically address such issues. These include (i) a dataset extension process, that combines the annotated portions of the images with the remaining visual content lacking its own annotations, to generate artificially augmented samples. (ii) A multiscale training paradigm for semantic segmentation and adjustments to the region proposal for instance segmentation, to cope with the variety of shapes in input. (iii) A post-processing phase to polygonize the raster output and generate a cleaner result.

In summary, our contributions can be summarized as follows:

- We construct and release a new dataset aimed at the delineation and classification of PV panels using VHR aerial imagery. The dataset provides thousands of accurate manual annotations. To our knowledge, this is the first dataset to provide not only the boundaries of the panels but also their categorization and several useful metadata information.
- We provide a comprehensive benchmark including a comparison between instance and semantic segmentation techniques, highlighting their strengths and weaknesses in the context of this task.
- To further improve on the baselines, we provide a *bag of tricks* specifically aimed at this delineation task, including a post-processing algorithm to polygonize and regularize the final output.

To foster further work in this field, we intend to make the code and dataset publicly accessible for research purposes.[1]

The remainder of this document is organized in the following way. In Sec. II we discuss works related to instance and semantic segmentation tasks, as well as ad-hoc aerial techniques, including solutions for PV panels analysis. In Sec. III we introduce the VHR dataset, describing data acquisition, processing, and annotation procedure. In Sec. IV we formally introduce the problem setting, describing our semantic and instance segmentation approaches, and then providing a series of improvements for each modality. In Sec. V we present the experimental validation of the proposed methods, discussing the results on the benchmark and the resulting performance boost from our approaches. Last, in Sec. VI we draw final conclusions, as well as possible directions for future research in this field.

II. RELATED WORK

A. SEMANTIC SEGMENTATION

Deep learning-based approaches for semantic segmentation are mostly based on convolutional encoder-decoder architectures, where features from deeper layers of the network are expanded back to the pixel space by either using multiscale aggregation or by repeated upsampling and integration with shallower features.

Solutions using multiscale aggregation include models such as PSPNet [7], where the decoding step exploits Pyramid Pooling modules to capture context information at multiple scales. Similarly, DeepLab [8] introduces the *atrous convolution*, which adopts larger kernels with increasing dilations instead of pooling to increase the receptive field, while maintaining a higher output resolution, and the Conditional Random Fields (CRF) to further clean the segmentation output. DeepLabV2 [9] improves upon its predecessor by adopting ResNet101 as a deeper encoder, and by introducing the Atrous Spatial Pyramid Pooling (ASPP) layer, a multiscale module that encodes the input features with parallel filters with different dilations. Last, DeepLabV3 [10] incrementally improves upon DeepLabV2, introducing cascaded convolutions and a parallel ASPP with image-level features, and DeepLabV3+ [11] provides a deeper decoder, with a more efficient ASPP module, exploiting separable convolutions. Derivative works build onto these architectures in several ways, providing better decoding stages or more efficient parallel atrous feature extraction modules, such as AdaptNet [12] or AdaptNet++ [13].

Considering upsampling and feature integration, Fully Convolutional Networks (FCN) [14] and Segnets [15] represent one of the first effective solutions where fully-connected layers are substituted with convolutions to preserve spatial information, paving the way for several subsequent works. Another well-known example is U-Net [16], a fully convolutional network originally developed for medical image segmentation with few training samples. Its symmetrical architecture with skip connections allows information to flow from the encoder to the decoder, which helps to preserve the spatial resolution and improve the segmentation accuracy. Several works have improved upon this architecture with different variants [17], [18], [19], by either including attention mechanisms [19] or by introducing residual layers for an even faster convergence [18]. Inspired by the Transformer architecture and its effectiveness in language processing, several works proposed effective spatial variants of attention mechanism, such as Criss-Cross attention [20], or Dual Attention [21].

The recent breakthrough of Vision Transformers also allowed for new effective encoder architectures, such as ViT [22], Swin [23], as well as end-to-end transformer-based segmentation models such as Segmenter [24] or SegFormer [25]. Most of these solutions work with image patches, exploiting spatial attention mechanisms to improve feature maps with contextual information. Similar updates appear to be extremely effective for convolutional networks, with completely revised architectures such as ConvNext [26]. Despite their effectiveness on large-scale datasets, transformer-based architectures typically require more information at training time in order to reach robust performance. This is often not the case with older, albeit conceptually simpler, architectures. In this paper, we focus our efforts on more lightweight solutions such as U-Net, exploiting their ability to learn from few examples [16] to provide a robust feature extraction.

B. OBJECT DETECTION AND INSTANCE SEGMENTATION

The major limitation of semantic segmentation is that it can not distinguish between different instances as long as they belong to the same class. To this end, instance segmentation appears as a more refined option, detecting individual objects within an image and assigning a unique mask and label to each of them. Traditional approaches for object detection can be divided into single-stage or two-stage methods. Among the former family, we can find YOLO (You Only Look Once) and its derived approaches [27], [28], [29]. These solutions

[1] https://github.com/edornd/solar-panels

are often optimized for real-time usage and employ a single neural network to predict the bounding boxes and class probabilities of objects in an image.

Multi-stage approaches divide the extraction of Region of Interest (RoIs) from its analysis: in the first stage, the full image is processed to understand which areas may contain objects, while subsequent stages focus on these proposals to extract relevant information. Region Convolutional Neural Network (R-CNN) [30] represents one of the first approaches of this kind. It first generates a set of region proposals, or candidate objects, using a sliding window approach. Each region is then processed with a standard CNN classifier. Fast RCNN [31] tries to improve on the predecessor by extracting features from the entire image at once, rather than processing each region separately while maintaining the sliding window approach for the proposals. A further speed-up to the process is brought by Faster RCNN [32], which replaces the slow proposal generation algorithm with a Region Proposal Network (RPN), an additional CNN trained to predict object regions from the extracted features maps. These three approaches only provide object detection, that is, a class label and approximate bounding box for each object. One of the first instance segmentation proposals is represented by Mask RCNN [5], which improves on Faster RCNN by means of a mask prediction branch, which generates a pixel-level mask of the detected instance, on top of the box regression and classification.

Driven by the surge of Transformers, recent approaches are instead steering towards end-to-end architectures, such as DETR [33] for object detection, or MaskFormer [34] and Mask2Former [35] for instance segmentation, or even universal segmentation. Despite their effectiveness, these proposals suffer from the same problems as their semantic counterparts, namely the more challenging training setup and the higher hardware requirements, especially during training.

For what concerns the topic of aerial images, these differ from natural ground images in several aspects, including scale and orientation. They usually have a larger field of view than traditional images, capturing a wide range of scales, from large forests to smaller objects such as cars. Contrary to natural images, orientation is not limited to a single direction: objects in aerial images can appear in different orientations, without specific priors (e.g., if an object is standing on the ground, it should appear at the bottom of the image). Algorithms dealing with aerial imagery must be able to recognize objects in all possible orientations, preferring Oriented Bounding Boxes (OBB) to Axis-Aligned Bounding Boxes (AABB). Starting from aligned counterparts, several oriented variants have been proposed for aerial imagery. For instance, Rotated Faster RCNN [36] is an extension of Faster RCNN that allows the bounding boxes to be rotated to better align with the object orientation while maintaining aligned RoIs. To this end, RoI Transformer [37] substitutes the RPN with a transformer architecture to process the object proposals. The model handles a rotated RoI and generates a rotation-invariant feature map that is used to classify and localize the object. This method is particularly useful for handling the variability in object size and orientation that can be common in aerial images. Similarly, Oriented R-CNN [38] is proposed as a simpler alternative: in the first stage, an oriented Region Proposal Network (oriented RPN) directly generates high-quality oriented proposals, while the second stage exploits an oriented R-CNN head to refine oriented Regions of Interest (oriented ROIs) and recognizing them.

C. DEEP LEARNING FOR PANEL DETECTION

The increasingly large availability of remote sensing resources in recent years has driven research in several application fields, including energy and automated PV panel delineation. To this day, many datasets have been created for this task, including different areas and sources. However, the resources with the highest volume in terms of inspected surface and number of annotations mainly include the USA territory, with larger scale surveys such as the Californian dataset [40], or deep learning frameworks like DeepSolar [42]. Despite the undeniable usefulness of such resources, different geographical areas may show quite different visual features, from the landscapes and the vegetation types to the different building structures. This often undermines the effectiveness of automatic detection systems. Considering the European continent, covered areas appear quite limited in comparison, with different feasibility studies in the Netherlands [43], [44] or Switzerland [45], mainly using satellite data or aerial acquisitions. This is also probably due to the higher costs of VHR aerial and remote sensing imagery, which remains crucial for accurate detection and delineation of PV panels when compared to open data such as Sentinel feeds. In this work, we attempt to close this gap by creating a high-quality PV panel dataset, focusing on the study area of Piedmont. Similar to existing resources [40], this dataset provides more than a hundred VHR images over two large provinces of the region, together with several thousands manual annotations, as detailed in Sec. III. Table 1 provides a non-exhaustive list of similar works, comparing them with our data sources.

Focusing on the methodology, the vast majority of proposals available in literature extract PV panels through semantic segmentation [42], [46], [47], [48], carrying out a binary subdivision between background and panel surface. Given its proven effectiveness in remote sensing and low-resource datasets, the U-Net architecture [16] remains one of the most popular choices, even though other models such as DeepLab [10] have been successfully applied also in this field [48]. The most notable differences among the available approaches often lie in the input processing phase. For instance, given the challenges of VHR imagery and its high processing costs, several works apply a two-step system, where inputs are first inspected with a more coarse approach through a simpler classifier, processing only those actually containing PV panels with segmentation algorithms [42], [47], [48]. This can also be carried out through satellite feeds, such as Sentinel-2, to detect larger plants [43].

TABLE 1. Comparison between similar datasets. Our PV panel dataset is comparable with similar resources, such as the Californian set, especially considering the covered area. For comparison, we also include DOTA, a general-purpose dataset with comparable characteristics, albeit on a much larger scale.

Dataset	# Images	Image Size	Resolution (cm)	Bands	# Instances	# Categories	Area (Km^2)
Vaihingen [39]	33	$2,500 \times 2,000$	9	IRRG	None	6	1.33
Potsdam [39]	38	$6,000 \times 6,000$	5	RGBIR	None	6	11.08
California [40]	601	$5,000 \times 5,000$	30	RGB	$\geq 19,000$	1	1,352
DOTA v1.0 [41]	2,806	$4,000 \times 4,000$	50	RGB	188,282	15	11,224
Ours	105	$20,000 \times 16,000$	30	RGBIR	9,462	≤ 3	3,614

The adoption of object detection and instance segmentation in this task appears instead quite limited: only a few works have explored object detection models for PV panels detection, with a handful of proposals using standard approaches such as Faster RCNN [49] from aerial photography, or YOLO [50], using drone feeds.

In summary, related works present some limitations: in the first place, the lack of openly available data sources severely hinders the research efforts in this field. Considering instead the methodologies explored, semantic approaches exclusively focus on the distinction of PV panels from the background, without taking into account possible categorizations, such as the installation type or the panel category (e.g., monocrystalline or polycrystalline). Instance segmentation approaches remain instead almost unexplored, despite their effectiveness in many similar fields [41]. In this work, we propose to address these shortcomings by providing a custom dataset, together with a comprehensive evaluation of segmentation techniques, including more advanced approaches such as multiscale input [51] or instance delineation with oriented bounding boxes [36].

III. DATASET

The complete pipeline carried out for the construction of the PV panels dataset is illustrated in Fig. 1. This dataset focuses on the provinces of Asti and Alessandria in Piedmont, which is the fourth Italian region in terms of quantity of solar panels deployed, and the first in terms of energy production from photovoltaic in 2022 [52]. Inside this area, these two provinces present the highest count of plants. Moreover, it includes both urban and rural regions, with a focus on areas known to have large-scale industrial PV plants and a significant number of agricultural and domestic installations. This approach ensures that the dataset is representative of a wide range of panel types and locations.

The images used in the dataset were obtained from the *Terraitaly* catalogue of *Compagnia Generale di Riprese aeree* (GCR) s.p.a, an Italian provider of high-resolution aerial imagery, from the most recent orthophoto archive dating 2018. Similar to the California dataset [40], the available images provide a spatial resolution of 30*cm* per pixel. This resolution is considered to be high enough to accurately detect and delineate individual PV panels, while also providing a sufficient level of detail for smaller domestic installations. The dataset includes a total of 105 VHR images in RGBIR format (i.e., visible spectrum with infrared), including 60 acquisitions for Asti and 45 for Alessandria. Each orthophoto covers on average $20,000 \times 16,000$ pixels, roughly translating into a region of 28 square kilometres each. The images are provided as georeferenced TIFF files, using the UTM zone 32N coordinate reference system (CRS) to minimize the distortion from the projection.

The annotation process for the dataset involved not only identifying and outlining the boundaries of PV panels in the images but also assigning metadata information to each processed instance. The task was carried out by a team of trained annotators, covering every registered large-scale industrial plant in the area using their known addresses or approximate coordinates, while the agricultural and domestic PV installations were annotated on a random basis by manual lookup, ensuring a certain degree of uniform spatial distribution of the samples, in order not to introduce geographical biases in the set. Together with its geometry, each panel is assigned several attributes, specifically: a unique identifier for the PV panel, a unique ID for the plant in case of industrial or agricultural installations, the province it belongs to, its orientation (e.g., south or south-east), its estimated power, the installation type (*industrial*, *agricultural*, *domestic*), and the PV category (*monocrystalline*, *polycrystalline*). This last attribute is particularly important to accurately estimate the energy production of a plant, given the characteristics of the two types: while being less expensive, polycrystalline panels are less efficient and have a shorter lifespan, while monocrystalline have higher performance in terms of production, duration and also resistance to high temperature [53]. However, from an object detection perspective, the latter represents a more difficult target due to their completely black appearance, extremely influenced by reflections and the hit angle of the light on the surface, and lower absolute counts due to the relatively new technology. On the other hand, polycrystalline installations typically present the signature blue pattern on a white grid, much more recognizable from an aerial point of view. The final dataset includes 9,462 manual annotations, including 8,967 polycrystalline and 495 monocrystalline PV panels. The results were combined and stored in a single shapefile to provide a georeferenced output, more easily composable with the available images. A sample of the final dataset is available in Fig. 2, displaying the major installation types, namely industrial (left) and domestic plants (right).

IV. METHODOLOGY

In this work, we compare the performance of instance and semantic segmentation models for the extraction and delineation of PV panels from aerial images, focusing on the

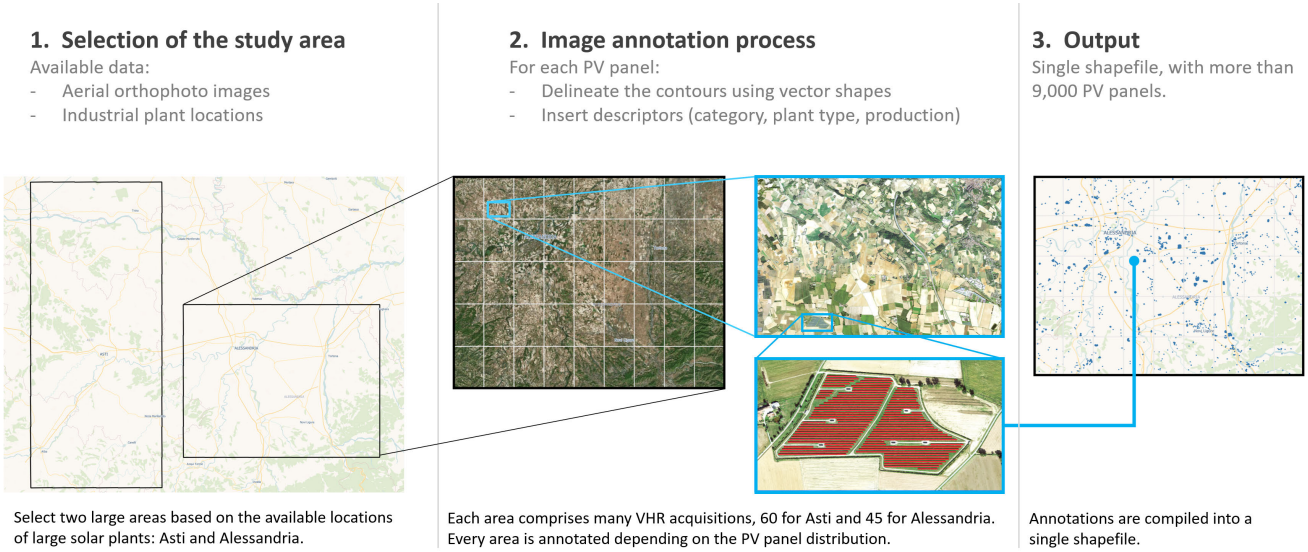

FIGURE 1. Visualization of the dataset construction pipeline: the selected areas are matched against the database of VHR images to construct the mosaic, then each image is processed individually to manually annotate panels, using approximate locations for industrial plants, and manual lookup for smaller installations. The final annotation output is stored in a single shapefile.

FIGURE 2. Examples tiles extracted from the dataset, together with their annotations. Starting from the left, the first two images show large industrial installations, employing both monocrystalline and polycrystalline panels, while the last two illustrate the smaller domestic installations.

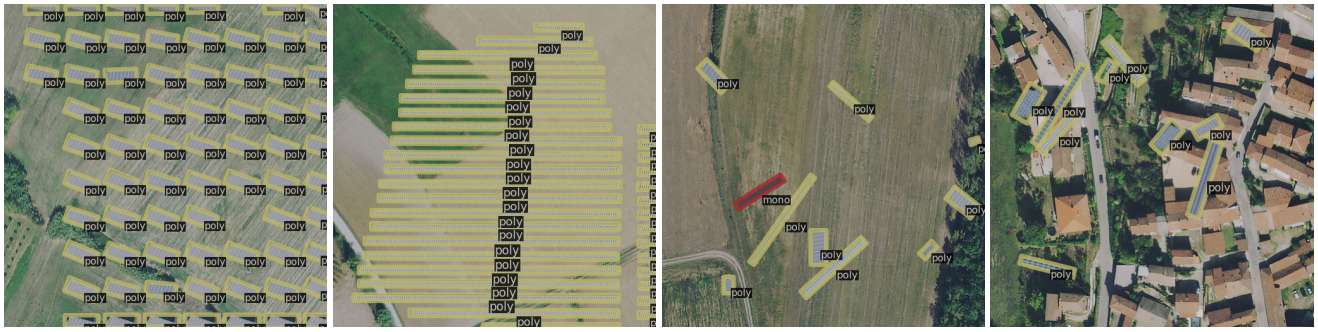

FIGURE 3. Examples of tiles obtained from the dataset extension process. The first two images shows the results of superimposing the whole plant from a labelled tile (*semantic-wise* mixing), while the last two images illustrate tiles obtained by the random copy-paste of several single panels (*instance-wise*) mixing.

classification of each panel into mono- or polycrystalline, being the most useful for practical applications such as energy production estimate. The semantic approach consists of a dense categorization, where each pixel is assigned its own label independently. Formally, let us define X as the input space, composed of a set of images $x \in X$ with a fixed amount of pixels I, and C as the set of semantic categories $c \in C$. The objective of semantic segmentation is the definition of a parametric model f_θ, mapping each pixel in the image to a probability distribution over the classes,

i.e. $f_\theta : X \rightarrow \mathbb{R}^{|I|\times|C|}$. The ground truth Y used in training also presents a dense categorization, where each image label y is composed of pixel-wise annotations y_i.

Instance segmentation involves instead a more complex pipeline, where the output is composed of a variable list of instances, each one associated with its own label, bounding box (either axis-aligned or oriented) and instance mask. In this case, the objective involves finding a set of object instances $O = o_1, o_2, \ldots, o_m$ such that each instance o_i is associated with a unique label $l_i \in C$ and a corresponding pixel-wise mask $M_i \subseteq I$, i.e. possibly smaller than the full input image. Similar to the semantic approach, we can define this task as the problem of finding a function $f_\theta : X \rightarrow (O, L, M)$, where O is a set of object instances, $L = l_1, l_2, \ldots, l_m$ is a set of labels corresponding to each instance, and M_i is a binary mask indicating the pixels of the image that belong to the i^{th} instance. Our desired output is still represented by a set of individual panels, including their geometry and category. While the output of the instance segmentation pipeline already fits this description, the semantic segmentation approach still requires some effort to process the raw predictions and obtain a comparable output. The next sections describe the methodologies devised to reach this goal, comprising (i) baseline solutions, (ii) ad-hoc modifications to both versions to improve results, and (iii) the post-processing refining steps.

A. EXTENDING THE DATASET

One of the main weaknesses of the available data is represented by the challenging annotation process: delineating every existing solar panel would in fact require knowing the position of each instance beforehand, which is not available for smaller installations. This consequently leaves a good portion of the images effectively unused, since only the portions of images intersecting at least one panel are selected during training to provide a fair balance between objects and background information. Given the extremely wide area analysed and the relatively limited number of annotated panels, we further attempt to artificially extend the available training dataset with an approach comparable to Copy-Paste augmentations [54] or ClassMix [55], constructing an alternate, handmade set of ground truth annotations.

The dataset extension process can be formalized as follows: given the current set of labelled images $x \in X$ and their respective labels $y \in Y$, we select a fixed number of unlabelled tiles from the sources images, each one represented as $x_u \in X_u$. For each unlabelled image x_u, a labelled pair (x, y) is randomly selected. In order to construct a plausible result, we carry out the mixing strategy between labelled and unlabelled images in one of two similar variants, with a given probability p_t, the first at label level (*semantic-wise*), while the second at instance level (*instance-wise*). Starting from the ground truth, each panel is selected with probability p, where p is equal to 1 in the first variant. This selection is further augmented to introduce more variation, and then a binary mask M is generated from the resulting output so that the unlabelled version and the augmented PV panels can be merged together. Considering the single image x_u, the output of the extension can be expressed as a tuple made of the selected panels as ground truth, and an input image $x_e = x_u * M + x * (1 - M)$. In case instead of *semantic-wise* extension version, each panel is selected from a randomly sampled labelled pair (x,y) with a probability p until a given number of panels N_p have been superimposed on the current unlabelled image x_u.

Simply put, we build the dataset extension in two steps. First, we exploit the portion of the dataset without panels by selecting a sample of image content from the training subset. This selection is carried out by randomly choosing a fixed amount of tiles during the preprocessing phase with uniform distribution, with the only requirement that no panel shall intersect their bounds. This does not guarantee the absence of panels in the crop, however, the possible presence of false negatives is heavily mitigated by the subsequent mixing phase. Second, we mix these empty samples with the annotated tiles, by selecting a portion of the available labels, cutting from the latter images along the borders of the annotation, and pasting on top of the unlabelled variants, as illustrated in Fig. 3. In one case, we simply copy and superimpose on the current tile all the panel annotations found in a training tile, further augmented at the image level (i.e., applying the transformation to the whole image). This can be beneficial to both keep a high degree of plausibility with respect to the source dataset and maintain the regular appearance of larger plants after the process. In the other case, we randomly select a predefined number of panels in the training dataset, and superimpose them on the unlabelled image in random poses and locations after applying geometrical transformations to each instance. This mixing approach is introduced to mimic the distribution of smaller installations, where size and orientation greatly vary among tiles.

Example tiles generated by the extension process are reported in Fig. 3.

Although these images could appear less realistic to the human eye, this simple approach may improve the ability of the model to distinguish panels against several types of backgrounds. We apply this extension to both semantic and instance approaches, to assess the performance gains in each context.

B. INSTANCE SEGMENTATION

In this study, we build upon Rotated Faster RCNN [36], an extension of Faster RCNN [32] able to detect arbitrarily oriented objects in aerial images. In this framework, the input image is first processed by a backbone composed of a ResNet50 [56] encoder and a Feature Pyramid Network (FPN) [57], capturing rich image features on five different scales. A standard Region Proposal Network (RPN) is then employed to generate potential candidates by applying convolutional filters separately for each level of the output features map, typically using three aspect ratios of anchors, namely 1:3, 1:1 and 3:1.

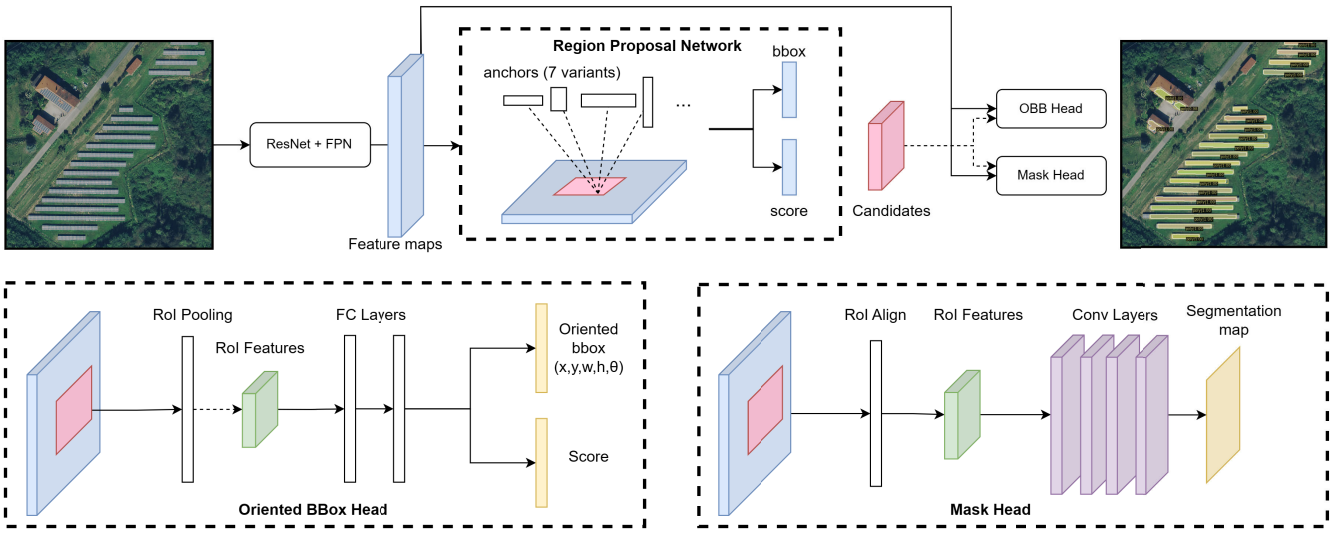

FIGURE 4. Architecture of the custom Rotated Faster RCNN model. An input image is first processed by a feature extraction network that balances high-level and low-level features. An RPN is then employed to produce axis-aligned region proposals that will be classified by both the *OBB head* and the *mask head*, that compute as output (i) an instance label with confidence in the interval [0, 1], (ii) an oriented bounding box, and (iii) a binary segmentation map respectively.

The standard Rotated RCNN processes the generated proposals through a bounding box head, in order to output both an OBB estimate and an instance label. It initially applies RoI pooling to the feature maps of the full image in order to extract RoI features related to candidates generated by RPN, to first map them into a feature vector thanks to two fully connected layers and then produce the final output as bounding box parameters and a label with its confidence score. The OBB, in both the initial ground truth and the final prediction, is expressed in the form of (x, y, w, h, θ) where x and y represent coordinates of the box centre, w and h its width and height, and θ the rotation angle of the box with respect to the main axis.

To improve the performance on this particular task, we first update the anchor aspect ratios for each sampled feature point, applying anchors with 7 different ratio variants, specifically: 1:10, 1:5, 1:2, 1:1, 2:1, 5:1, 10:1. This modification is introduced to better cope with the wide variety of shapes and orientations that PV panels could have. To further adapt the object detection model to this task, we include a separate branch to the proposed architecture to generate an instance mask, named *mask head*, obtaining three separate outputs, similar to Mask RCNN [5]. The mask head, just like the oriented bounding box head, applies the RoI pooling process to the entire image features. Following Mask RCNN [5], we use RoIAlign to perform a more refined interpolation to better align the cropped proposals. The features, extracted following the RPN proposals, are subsequently processed in four serial convolutional layers in order to produce a segmentation map that separates pixels of the underlying panel from background ones. Unlike other OBB approaches, such as Oriented RCNN [38] or RoI Transformer [37], the region proposals are still axis aligned for both the bounding box regression and the mask head. This is however extremely convenient for the instance-level binary segmentation task, which can receive a crop containing a good balance between foreground and background content.

Since PV panels have a regular rectangular shape, the OBB itself is often enough for precise localization and delineation; however, using this branch as an auxiliary task during training can be beneficial for the other object detection tasks as well. The architecture of our final approach is reported in Fig. 4.

C. SEMANTIC SEGMENTATION

Given the effectiveness in literature [40], [46], we also provide an assessment using semantic segmentation. In order to provide an initial benchmark, we evaluate different convolutional architectures to evaluate the performance of various combinations of encoders. In terms of backbone, we first test the standard ResNet [56], which represents the baseline in many computer vision tasks [18], in the variant with 50 layers (i.e., RN50). Second, given their robust feature extraction abilities, we include a ConvNext encoder [26], which provides several macro and micro architectural changes to close the gap between transformer-based models and convolutional networks. These include changes such as depth-wise convolutions, larger kernel sizes, and layer normalization, while also maintaining a throughput higher than Swin transformers [23], which is a crucial feature in semantic segmentation, where the encoder only comprises the first part of the whole architecture.

Among the decoders, we select U-Net [16], which is by far the most diffused and effective decoder architecture to this day [58]. This solution provides a reverse pyramid structure, bringing the computation back to the pixel level through a series of upscaling and feature concatenations with the

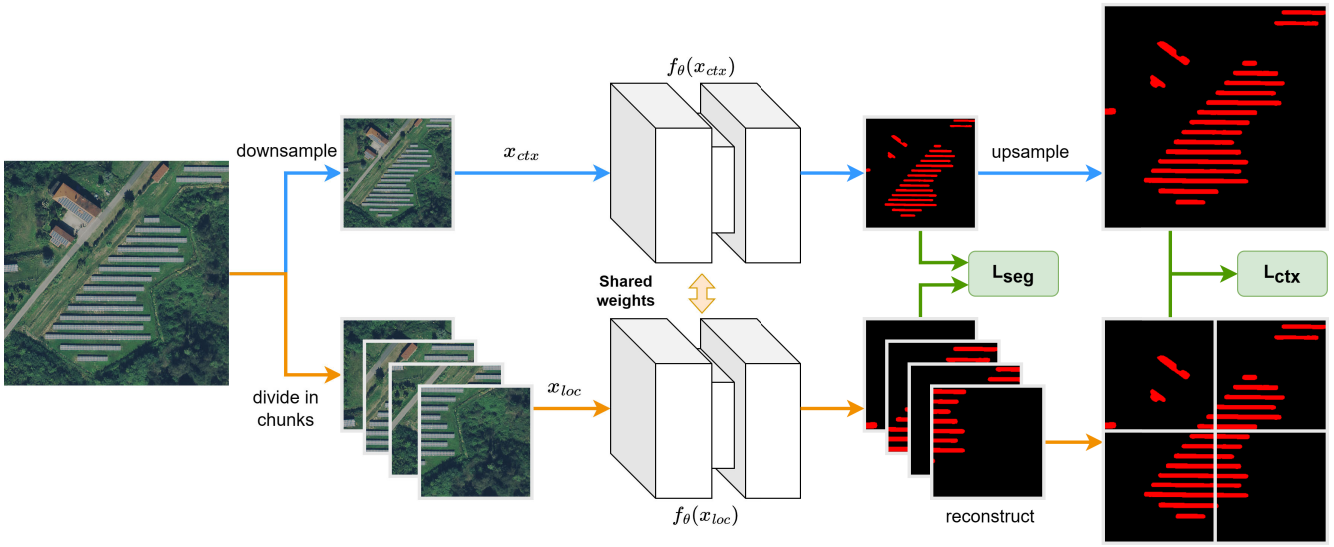

FIGURE 5. Multi-scale semantic segmentation framework adopted in this work. Each image is processed by two parallel branches, using shared weights. A *contextual* branch extracts features from a downscaled version of the full input, while a *local* branch extracts features from a batch made of four splits of the original image. A standard semantic loss is computed on both predictions, while a consistency loss is applied between local and global outputs.

corresponding encoder layers in a symmetrical way. This decoder structure has been applied successfully in countless applications, especially considering remote sensing and aerial imagery [18], [45]. For an improved convergence during training, we adopt the residual U-Net variant (ResUnet) [18], which includes residual skip connections also at the decoder layer, similar to the encoder [56].

A recurrent issue with PV panel delineation is the wide range of scales of the installations, from limited surfaces of domestic plants to entire fields in industrial contexts. To cope with this variety, we build upon these baselines by introducing a multiscale training similar to GLNet [51]. The prohibitive size of the available images does not allow the application of a complete *global-to-local* or *local-to-global* regularization [51], however we introduce a *context loss*, where a larger portion of the image is compared with smaller crops, extracted from the same region. Each tile is subdivided into four components, by splitting once both vertically and horizontally. The transformed output is processed independently and then compared with the overall context, represented by a downscaled version of the image. We can then guide the model toward multiscale consistency by forcing the local and contextual features to be as similar as possible. To carry out this step, we reconstruct the full-size output from the tiled features, and upscale the contextual features back to the original dimensions. We then apply a pixel-wise quadratic penalty, computed as Mean Squared Error (MSE):

$$L_{ctx} = \frac{1}{N}\sum_{1}^{N}(\phi_\theta(x_{loc}) - \phi_\theta(x_{ctx}))^2 \qquad (1)$$

where x_{loc} and x_{ctx} respectively represent the tiled and downscaled inputs, and ϕ_θ the model output before the final Softmax operation (i.e., *logits*). During training, this additional contextual loss is added to the standard semantic loss as a regularizing factor. An illustration of the final semantic framework is shown in Fig. 5.

D. POLYGONIZATION

In order to provide a more accessible result for further analysis and use, the raw model output often requires further processing. In the first place, raster information should be converted back into polygonal data, in such a way that the generated vectors maintain a regular shape, without noisy or overlapping borders. Furthermore, while instance segmentation already provides panel-level information, semantic segmentation necessitates additional steps to bring the pixel-wise classifications into a more meaningful list of instances.

Starting from this last point, we extract individual instances from the semantic predictions using Connected Components Labelling (CCL) [59]. As the name suggests, this algorithm separates distinct areas, assigning different numerical identifiers to each component. Since the prediction is done on a pixel level, the same panel may contain both monocrystalline and polycrystalline outputs: to prevent this, we first apply the CCL pass on a binarized version of the output, so that the panel category does not interfere with the instance extraction. Then, we iterate over each component, discarding those instances whose surface is below a minimum threshold t and assigning a single class to each panel, selected via majority voting on the pixels of the component itself.

Once a panel-level delineation is obtained, we convert the raster outputs into polygons, more useful for practical use, following a custom procedure adapted from building extraction [60]. Assuming that PV panels have a rectangular shape, or at most that the installations are made of rectangular components, the procedure attempts to regularize the edges

of each polygon so that its internal angles always measure 90 degrees, while minimizing the changes with respect to the original shape. First, the raster image is converted into an actual polygon using the standard Douglas-Peucker algorithm, keeping a low tolerance to avoid oversimplifications. From the coarse polygon, we extract the oriented Minimum Bounding Rectangle (MBR), defined as the oriented rectangle (.e., not constrained by the axes) with minimum area, fully containing the considered shape.

Given the typical shape of PV panels, we assume that the MBR is the best approximation of its orientation, therefore we exploit this to further regularize the edges of the underlying, coarse shape. Once the MBR is computed, we align each edge of the polygon with the direction of the MBR that is closer to the current rotation of the segment. This allows for maintaining the leading orientations of the shape, while at the same time making sure that every angle becomes a multiple of 90 degrees. Rotating the edges of the polygon independently generates discontinuities along the perimeter, which can be fixed by reconnecting adjacent segments. Before the restoration, we further simplify the polygon, removing consecutive edges that are now parallel (i.e., lie on the same side) and whose length of the reconnecting segment is below a dynamic threshold T, computed as $T = \alpha * L$, where α is a scalar factor in the range [0, 1] and L represents the length of the longest edge in the polygon. In practice, this step removes those edges that are the least important to define the object boundaries.

From this simplified, discontinuous shape, the final polygon is computed by reconnecting the segments. The reconstruction is done by computing the intersection points between the lines defined by the edges, without further processing, as shown in Fig. 6. The full procedure is detailed in Algorithm 1.

Algorithm 1 Polygonization Procedure

Input:
R, raster prediction to be polygonized
t_A, minimum area threshold
t_{DP}, tolerance for polygonization
α, length factor for edge filtering
Output:
P_r, a set of regularized polygons
Extract polygons from the raster
// binarize the input prediction
$B \leftarrow Binarize(R)$
// extract the set of connected components
$C \leftarrow CCL(B)$
// Discard components with area $< t_A$
$C \leftarrow MinSurface(C, t_A)$
// Apply Douglas-Peucker to each component
$P \leftarrow \{\}$
for c *in* C **do**
$p \leftarrow DouglasPeucker(c, t_{DP})$
$P \leftarrow P \cup p$
end
end
Regularize the extracted polygons
$P_r \leftarrow \{\}$
for p *in* P **do**
// Extract the MBR
$mbr \leftarrow MBR(p)$
// Align edges E with MBR directions
$E \leftarrow Align(p, mbr)$
// Remove useless edges
$E \leftarrow Filter(E, \alpha)$
// Rebuild the final polygon
$p_r \leftarrow Link(E)$
$P_r \leftarrow P_r \cup p_r$
end
end

V. EXPERIMENTS

In this section, we evaluate the presented models and frameworks against the constructed PV panel dataset. We first provide a thorough description of our experimental setup, including dataset subdivisions, combinations of experiments and hyperparameters employed in the tests. We then discuss the obtained results, comparing instance and semantic predictions down to the computed metrics, and highlighting the strengths and weaknesses of both techniques.

A. IMPLEMENTATION DETAILS

We assess the performance of the baseline solutions and our improved approaches on the two variants of the PV panels dataset: the standard version, where we keep as input image only those tiles intersecting an annotation, or the extended version, as described in the previous section. We perform the tiling offline, using a tile size of 512 × 512 pixels to ensure a good ratio between visual content and PV panel size, and using an overlap of 256 pixels to ensure that border information gets centred in the next tile. To ensure a fair and robust comparison, we split the two areas into separate sets, keeping the Alessandria province for training purposes, and the Asti province for testing. Among those containing annotations, we further select 20% of the training orthophoto for validation purposes. In total, we obtain 696 tiles, divided into 404 images in training, 123 for validation purposes, and 169 for testing, with a number of panels equal to 12, 495, 2, 260 and 3, 154 respectively, accounting for the repeated annotations introduced by the overlapped preprocessing. To reduce any possible overfitting, we further apply geometric and photometric augmentations, including flipping, rotation and scale, as well as random brightness, contrast or gamma adjustments, and weak Gaussian noise or blur applications. We test every solution on both the RGB and RGB-IR variants, leading to a total of four dataset versions, namely: *base RGB*, *extended RGB*, *base RGBIR* and extended *RGBIR*. These combinations contain the same ground truth annotations, they only differ in the number of channels or the number of tiles.

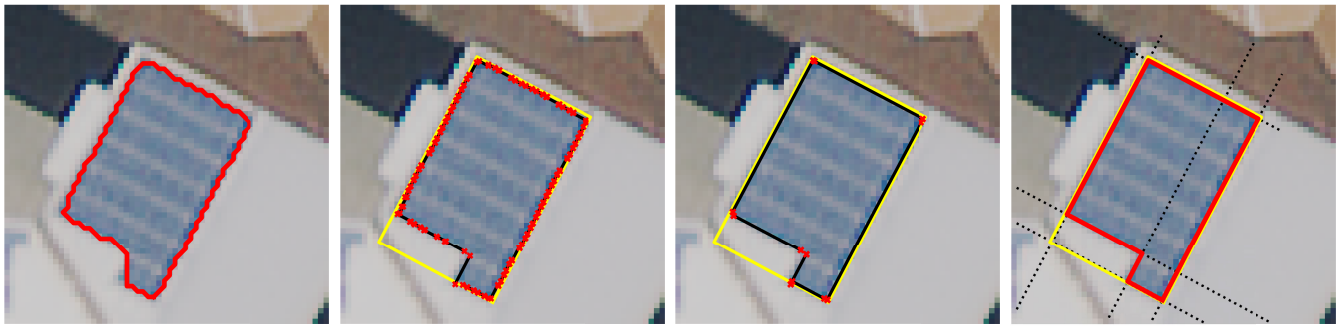

FIGURE 6. Visual representation of the steps of the polygonization algorithm, from left to right: (i) the raster prediction is converted into a coarse polygon using the standard Douglas-Peucker algorithm, (ii) each edge is aligned with its dominant MBR orientation, (iii) similar parallel edges are further removed to clean the final shape, (iv) the discontinuous segments are joined back computing the intersection points between the lines passing through the edges.

During the dataset extension process, we exploit an automatically annotated sample of 10, 000 tiles, retrieved from regions without tiles inside the Alessandria province. We further divide 70% of them into *panel-wise* augmentation, namely copying the entire label from another annotated tile, leaving for the remaining 30% to superimpose randomly chosen panels in random locations (i.e., *instance-wise*). For the experiments using the RGB-IR dataset variants, we adapt each model to support the additional infrared band by simply copying the weights from the red channel, including its mean and standard deviation for the input normalization step.

We adopt AdamW as optimizer to update model weights during training, with an initial learning rate set to 2.5×10^{-3}, momentum equals to 0.9 and weight decay of 1×10^{-4}. Following standard practice, every instance segmentation variant has been trained using a smooth L1 loss for the OBB regression head, and a cross-entropy loss for the class and mask heads. Similarly, without considering the additional regularizations, semantic segmentation models also adopt a standard, pixel-wise, cross-entropy loss. Every model has been trained on a workstation equipped with *NVIDIA GeForce RTX 2080 Ti* GPUs, using a batch size of 4 tiles for a total of 80 epochs for the base datasets and 12 epochs for the extended datasets.

All experiments and data processing have been conducted using the Python programming language, adopting PyTorch[2] as deep learning framework, together with *shapely* and *rasterio* to handle vector and raster data respectively. Instance segmentation models further exploit the *mmrotate* framework,[3] for training and evaluation purposes.

B. INSTANCE SEGMENTATION RESULTS

We assess the results of instance segmentation using the mean Average Precision (mAP), computed on the test set. The metric can be expressed as the average of the class-wise Average Precision (AP):

$$mAP = \frac{1}{N} \sum_{i=1}^{N} AP_i \tag{2}$$

where N is the number of classes in the dataset, equal to $N = 2$ in our case, and AP_i is the average precision observed for the class i. The AP of a specific class is evaluated as the area under the Precision-Recall (PR) curve, in turn obtained by plotting the values of precision and recall at different IoU values. We compare our custom architecture against state-of-the-art OBB detection solutions, namely Oriented RCNN [38] and RoI Transformer [37]. Since the former also uses a RPN and a separate head for final output, we evaluate the impact of the additional mask head, as reported in Fig. 4, exploiting the feature maps RoI candidates of the RCNN head to output a segmentation prediction. The same modification cannot be directly applied without further modifications to RoI Transformer, given this component directly substitutes the RPN network, thus completely changing the candidate extraction mechanism. However, for comparison, we present the results obtained considering the OBB performance only.

To provide an initial performance evaluation of the provided architecture against similar architectures, we first train all three models on the base RGB dataset under the same configuration settings described above. Given the similarity in terms of downstream task, we also assess the impact of initializing the model using weights pretrained on the DOTA dataset [41]. Baseline results are shown in Table 2. As expected, starting from a pre-trained model leads to better performances, regardless of the chosen architecture, also given the relatively low-resource dataset employed. Furthermore, the use of a mask head can provide an additional increment to the final score, albeit contained. The Rotated Mask RCNN reaches 24.64 mAP. with a +2 mAP increment with respect to the OBB variant.

In order to further assess the impact of pretraining, mask head and the dataset extension with visible and infrared bands, we perform a second set of experiments, studying the performance variations at each addition. Given the comparable performance with other approaches, for simplicity we only leverage on the custom Rotated RCNN architecture for this set, who results are detailed in Table 3. These scores highlight how the use of *extended* dataset version, regardless of the input bands, mask head or pretraining, is always beneficial in this configuration: this is most likely due to the region-based approach of the chosen architecture, which does not only

[2]https://pytorch.org
[3]https://github.com/open-mmlab/mmrotate

focus on the pixel-level information but mostly learns to discern instances from the background. At the same time, the limited resources exploited during the *base* training may not be enough to train this more complex architecture, in the opposite trend with semantic segmentation. Concerning the addition of the infrared channel, the results highlight a peculiar trend: the only use of the visible spectrum leads to better performances in the *base* configuration, while this behaviour completely reverses when considering the *extended* dataset, displaying a minimum of +1 mAP increment on each RGBIR variant with respect to its 3-channel equivalent. Again, this may be due to the lower amount of training data in the former case: the *base* dataset might not be enough for the model to take advantage of infrared data, especially considering configurations pretrained on RGB images. On the contrary, the use of the extended dataset allows the full exploitation of this modality, allowing the model to adapt and exploit the additional information over time. Finally, Table 3 again confirms the importance of the mask head and its contribution to the final results. We observe that such addition almost consistently outperforms the OBB-only variants in every scenario, up to obtaining a total 30.73 mAP from the combination of the extended dataset with RGBIR bands, DOTA pretraining and auxiliary mask head, surpassing the other models by at least +2 mAP. However, considering the class-wise outputs we note that, while polycrystalline PV panels are often well-defined and delineated, the monocrystalline category remains extremely challenging. As we can see in the qualitative results in Fig. 7, instance segmentation outputs tend to mistake dark, rectangular shadows and areas for the monocrystalline panels, placing however proposals in plausible locations, such as house roofs.

In summary, while instance segmentation may not produce accurate delineations, their proposal-based approach can be a good fit for a simpler detection estimate, where the focus remains on the localization of the installations. If an accurate segmentation is instead required, semantic approaches provide an edge over instance solutions, especially with smaller datasets, as shown in the next section.

C. SEMANTIC SEGMENTATION RESULTS

We evaluate the results obtained from the semantic segmentation models using the Intersection over Union (IoU) metric, or Jaccard index, computed as the ratio of overlapping pixels between predictions and ground truths and their union. We maintain the residual UNet (ResUnet) decoder across every experiment, and we test different combinations of encoders, bands, and dataset configurations. Given the effectiveness demonstrated in the previous experiments, we adopt pretrained weights for the backbones in all the semantic tests. Considering the unbalanced dataset, we compute both the macro-averaged and micro-averaged IoU to provide a more comprehensive overview of the performances, respectively referred to as $MIoU$ and $mIoU$.

The results, detailed in Table 4, highlight in the first place the challenging detection of the monocrystalline panels. For comparison, we report in the first row our best instance segmentation approach rasterizing its output and computing the same metrics on a pixel level. Considering the *base* dataset without additional improvements, i.e. only considering those training tiles with annotated PV panels, we observe that the semantic approach is already enough to surpass the delineation capabilities of the instance-based model in two out of three categories. The amount of input information is in fact enough to discern between background and polycrystalline pixels, which represent the vast majority of installations, while either struggling or completely failing to delineate monocrystalline plants. Training with the additional infrared band appears to be extremely beneficial for the latter category, bringing the class IoU to 45.90, as well as for the remaining ones, improving by at least one point each. Contrary to the expectations, the dataset extension is instead detrimental for the semantic use case, managing good results only on the easier polycrystalline class and reaching a maximum $MIoU$ of 57.07, which does not improve over the baselines. This is most likely due to the label contamination [55] generated by the copy-paste mechanism: while pixels completely inside or outside correctly represents their respective class, labels on the decision boundary may contain some noise.

TABLE 2. Results of instance segmentation experimented architectures, considering RGB base dataset. Combining pretraining and mask head leads to higher results for the simpler rotated RCNN, when compared to state-of-the-art solutions.

Model	Mask	Pretrain	mAP
Oriented RCNN			19.15
Oriented RCNN		✓	21.49
Oriented RCNN	✓		23.15
Oriented RCNN	✓	✓	22.03
ROI Transformer			23.85
ROI Transformer		✓	24.09
Rotated Faster RCNN			20.81
Rotated Faster RCNN		✓	22.25
Rotated Faster RCNN	✓		22.33
Rotated Faster RCNN	✓	✓	**24.64**

TABLE 3. Variation study on the instance segmentation framework, assessing the contribution of each component to the overall score.

Set	Bands	Pretrained	Mask	Mono	Poly	mAP
Base	RGB			7.58	34.04	20.81
	RGBIR			1.98	37.16	19.57
	RGB	✓		8.09	35.41	21.75
	RGBIR	✓		1.21	36.87	19.04
	RGB		✓	9.09	40.20	24.64
	RGBIR		✓	4.55	37.93	21.24
	RGB	✓	✓	9.09	35.57	22.33
	RGBIR	✓	✓	0.94	36.44	18.69
Ext.	RGB			2.31	47.47	24.89
	RGBIR			9.09	48.00	28.55
	RGB	✓		2.42	47.17	24.79
	RGBIR	✓		3.03	48.74	25.89
	RGB		✓	2.82	52.33	27.58
	RGBIR		✓	4.55	53.44	28.99
	RGB	✓	✓	3.06	53.51	28.29
	RGBIR	✓	✓	9.19	52.37	**30.73**

FIGURE 7. Qualitative results of our best solutions on four different installations, including industrial plants with different scales and domestic PV panels. From top to bottom: instance segmentation, semantic segmentation, semantic segmentation with post-processing polygonization and regularization algorithm, initial ground truth. Best viewed on screen.

Moreover, the copy-paste mechanism introduces, on these borders, a strong discontinuity in the visual patterns, working against the pixel-wise prediction mechanism of semantic segmentation approaches.

Our best results with this approach, as well as the best overall results across all the experiments, were obtained from the multiscale training, indicated by the *ms* entries. The introduction of the consistency regularization across scales allows in fact for a +6.2 macro IoU improvement over the best baseline, with the strongest contribution given by the monocrystalline class, reaching 62.49 without further processing. The additional infrared band appears once again crucial to correctly define this last category, while the polycrystalline and background classes report comparable results. As a final test, we assess the effectiveness of our post-processing algorithm by extracting regularized polygons from the raw prediction and subsequently revert them back to raster tiles for evaluation. With the only application of this procedure, we obtain a substantial improvement in the monocrystalline category, reaching 74.34 IoU with an increment of +11.63 from the highest raw value obtained. The improvements over the raw predictions can be observed in Fig. 7, where the regularized polygons are compared with the raw output, as well as the predictions from the instance segmentation framework. Concerning large-scale installations (left), we observe that both semantic and instance approaches provide excellent results. On smaller industrial and agricultural plants, both approaches suffer in terms of performance, with instance segmentation favouring a higher recall, although with misplaced proposals, and semantic segmentation providing a more precise localization, at the cost of losing parts of the PV panel surface. Similar behaviour can also be observed on domestic

TABLE 4. Semantic segmentation results using ResUnet with various combinations of encoders, bands, and datasets. The first row provides an assessment of the best instance segmentation model using semantic metrics.

Encoder	Bands	Dataset	Backgr.	Mono	Poly	MIoU	mIoU
resnet50	RGBIR	ext.	93.23	3.41	56.12	50.92	87.38
resnet50	RGB	base	94.97	0.00	82.71	59.23	91.25
convnext	RGB	base	95.62	0.00	85.08	60.24	92.17
resnet50	RGBIR	base	96.32	36.46	86.01	72.93	93.41
convnext	RGBIR	base	97.24	45.90	86.83	76.66	94.37
resnet50	RGB	ext.	93.19	0.00	57.59	50.26	87.24
convnext	RGB	ext.	94.12	5.14	58.34	52.53	88.76
resnet50	RGBIR	ext.	94.78	0.00	68.74	54.51	90.85
convnext	RGBIR	ext.	95.21	5.95	70.05	57.07	91.05
resnet50	RGB	base-ms	97.34	37.50	81.46	59.48	93.84
convnext	RGB	base-ms	97.32	24.74	83.03	53.88	94.06
resnet50	RGBIR	base-ms	98.25	58.88	85.69	80.94	96.57
convnext	RGBIR	base-ms	98.17	62.49	85.53	82.86	96.53
convnext	RGBIR	ms+post	**98.63**	**74.12**	**86.91**	**86.55**	**97.34**

installations, where the instance-based approach manages to correctly delineate even non-exposed panels, while semantic segmentation approaches only provide a partial definition. We also note that both approaches were able to identify additional panels whose annotations were not present in the ground truth, as visible in three out of four images in the last row of Fig. 7.

VI. CONCLUSION

In this work, we investigated the task of PV panel detection and delineation from aerial imagery. To this end, we proposed a novel custom dataset, providing hundreds of VHR images and more than 9000 annotated panels in the Piedmont region, in Italy. This dataset not only provides accurate and manually defined labels but also several additional attributes and metadata such as their orientation or category, which could be exploited in a wide range of contexts, from energy production estimates to monitoring and performance analysis.

To provide an initial benchmark over this dataset, we then focused on a comparative evaluation of instance and semantic segmentation methods, aimed at the delineation of PV panels and their classification in monocrystalline and polycrystalline installations, considering every available industrial, agricultural or domestic plant. Furthermore, we provided several ad-hoc adjustments to both approaches to improve their performance, including a simple dataset extension mechanism, very effective for instance segmentation approaches, a multiscale training for semantic-based training, and a post-processing algorithm to exploit prior knowledge about the shape of PV panels and provide a more accurate and clean vector output.

Despite the promising results, we note a number of limitations that could be addressed in future iterations. First, the dataset only contains annotations for all the known industrial plants, but only contains a small percentage of the overall PV panel coverage in this area, due to both the absence of a local or national census and the difficulty in manually identifying such installations from aerial photography. Consequently, this provides further limitations on the effectiveness of downstream tasks, especially considering more complex approaches: this is particularly noticeable in the instance segmentation variants, where the only addition of a dataset extension through *copy-paste* augmentations was enough to greatly improve the results on this task. Second, despite the large area with a high number of PV panels and the VHR imagery, we acknowledge that the covered surface remains quite limited for applications on a larger scale, due to inherent differences among geographical areas in terms of both land cover and land use.

Future works will focus on addressing these limitations, considering both the underlying data and the methodologies devised to solve the task at hand. On one hand, the dataset can be expanded in several ways: in the first place, by providing manual annotations for every PV panel, also exploiting the very same models derived from this work with an iterative, human-in-the-loop (HITL), approach. Additional option would be adding new VHR imagery to the set, including new areas, more recent time periods, or even new modalities such as Digital Surface Maps (DSM). On the other hand, the proposed methods could be improved to better exploit the available information, exploiting for instance semi-supervised training mechanisms by generating pseudo-labels [61], or more recent state-of-the-art architectures, such as Segformer [25].

ACKNOWLEDGMENT

The authors would like to thank CGR s.p.a. for providing the aerial imagery that was instrumental in the completion of this research.

REFERENCES

[1] *EU Market Outlook for Solar Power, 2022—2026*, European Photovoltaic Industry Association (EPIA), EPIA, Brussels, Belgium, Jan. 2022.
[2] *State of the Energy Union 2022*, European Union, Brussels, Luxembourg, Accessed: Jan. 10, 2022.
[3] G. Di Francia, "The effect of technological innovations on the cost of the photovoltaic electricity," in *Proc. Int. Conf. Renew. Energy Res. Appl. (ICRERA)*, Nov. 2015, pp. 542–546.
[4] *Statistical Report—Solar and Photovoltaic*, Gestore Servizi Energetici, Rome, Italy, Accessed: Jan. 10, 2022.
[5] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.
[6] G. Stockman and L. G. Shapiro, *Computer Vision*. Upper Saddle River, NJ, USA: Prentice-Hall, 2001.
[7] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6230–6239.
[8] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and L. Alan Yuille, "Semantic image segmentation with deep convolutional nets and fully connected CRFS," in *Proc. 3rd Int. Conf. Learn. Represent. (ICLR)*, Y. Bengio and Y. LeCun, Eds. San Diego, CA, USA, May 2015. [Online]. Available: https://dblp.org/rec/journals/corr/ChenPKMY14.html?view=bibtex
[9] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*.
[10] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.
[11] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder–decoder with atrous separable convolution for semantic image segmentation," in *Proc. Comput. Vis. (ECCV)*, in Lecture Notes in Computer Science, V. Ferrari, M. Hebert, C. Sminchisescu, Y. Weiss, Eds. Munich, Germany: Springer, vol. 11211, Sep. 2018, pp. 833–851.

[12] A. Valada, J. Vertens, A. Dhall, and W. Burgard, "AdapNet: Adaptive semantic segmentation in adverse environmental conditions," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, Singapore, May 2017, pp. 4644–4651.

[13] A. Valada, R. Mohan, and W. Burgard, "Self-supervised model adaptation for multimodal semantic segmentation," *Int. J. Comput. Vis.*, vol. 128, no. 5, pp. 1239–1285, May 2020.

[14] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 3431–3440.

[15] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder–decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

[16] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Med. Image Comput. Comput.-Assist. Intervent (MICCAI)*, in Lecture Notes in Computer Science, N. Navab, J. Hornegger, W. M. Wells III, A. F. Frangi, Eds. Munich, Germany: Springer, vol. 9351, Oct. 2015, pp. 234–241.

[17] Q. Jin, Z. Meng, T. D. Pham, Q. Chen, L. Wei, and R. Su, "DUNet: A deformable network for retinal vessel segmentation," *Knowl.-Based Syst.*, vol. 178, pp. 149–162, Aug. 2019.

[18] F. I. Diakogiannis, F. Waldner, P. Caccetta, and C. Wu, "ResUNet—A: A deep learning framework for semantic segmentation of remotely sensed data," *ISPRS J. Photogramm. Remote Sens.*, vol. 162, pp. 94–114, Apr. 2020.

[19] Y. Sun, F. Bi, Y. Gao, L. Chen, and S. Feng, "A multi-attention UNet for semantic segmentation in remote sensing images," *Symmetry*, vol. 14, no. 5, p. 906, Apr. 2022.

[20] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "CCNet: Criss-cross attention for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Seoul, South Korea, Oct. 2019, pp. 603–612.

[21] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 3146–3154.

[22] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[23] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10012–10022.

[24] R. Strudel, R. Garcia, I. Laptev, and C. Schmid, "Segmenter: Transformer for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 7262–7272.

[25] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and efficient design for semantic segmentation with transformers," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 12077–12090.

[26] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 11976–11986.

[27] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 779–788.

[28] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 6517–6525.

[29] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," 2018, *arXiv:1804.02767*.

[30] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587.

[31] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.

[32] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[33] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 213–229.

[34] B. Cheng, A. Schwing, and A. Kirillov, "Per-pixel classification is not all you need for semantic segmentation," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 17864–17875.

[35] B. Cheng, A. Choudhuri, I. Misra, A. Kirillov, R. Girdhar, and A. G. Schwing, "Mask2Former for video instance segmentation," 2021, *arXiv:2112.10764*.

[36] S. Yang, Z. Pei, F. Zhou, and G. Wang, "Rotated faster R-CNN for oriented object detection in aerial images," in *Proc. 3rd Int. Conf. Robot Syst. Appl.* New York, NY, USA: Association for Computing Machinery, Jun. 2020, pp. 35–39.

[37] J. Ding, N. Xue, Y. Long, G. Xia, and Q. Lu, "Learning RoI transformer for oriented object detection in aerial images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2844–2853.

[38] X. Xie, G. Cheng, J. Wang, X. Yao, and J. Han, "Oriented R-CNN for object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 3500–3509.

[39] *2D Semantic Labelling Challenge, 2016*, ISPRS, Inst. Photogramm. GeoInf., Leibniz Univ. Hannover, Hanover, Germany. Accessed: Jan. 10, 2022.

[40] K. Bradbury, R. Saboo, T. L. Johnson, J. M. Malof, A. Devarajan, W. Zhang, L. M. Collins, and R. G. Newell, "Distributed solar photovoltaic array location and extent dataset for remote sensing object identification," *Scientific Data*, vol. 3, no. 1, Dec. 2016, Art. no. 160106.

[41] G. Xia, X. Bai, J. Ding, Z. Zhu, S. Belongie, J. Luo, M. Datcu, M. Pelillo, and L. Zhang, "DOTA: A large-scale dataset for object detection in aerial images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 3974–3983.

[42] J. Yu, Z. Wang, A. Majumdar, and R. Rajagopal, "DeepSolar: A machine learning framework to efficiently construct a solar deployment database in the United States," *Joule*, vol. 2, no. 12, pp. 2605–2617, Dec. 2018.

[43] V. Plakman, J. Rosier, and J. V. Vliet, "Solar park detection from publicly available satellite imagery," *GIScience Remote Sens.*, vol. 59, no. 1, pp. 461–480, 2022.

[44] B. B. Kausika, D. Nijmeijer, I. Reimerink, P. Brouwer, and V. Liem, "GeoAI for detection of solar photovoltaic installations in The Netherlands," *Energy AI*, vol. 6, Dec. 2021, Art. no. 100111.

[45] R. Castello, S. Roquette, M. Esguerra, A. Guerra, and J.-L. Scartezzini, "Deep learning in the built environment: Automatic detection of rooftop solar panels using convolutional neural networks," *J. Phys., Conf. Ser.*, vol. 1343, no. 1, Nov. 2019, Art. no. 012034.

[46] J. Camilo, R. Wang, L. M. Collins, K. Bradbury, and J. M. Malof, "Application of a semantic segmentation convolutional neural network for accurate automatic detection and mapping of solar photovoltaic arrays in aerial imagery," 2018, *arXiv:1801.04018*.

[47] F. Ge, G. Wang, G. He, D. Zhou, R. Yin, and L. Tong, "A hierarchical information extraction method for large-scale centralized photovoltaic power plants based on multi-source remote sensing images," *Remote Sens.*, vol. 14, no. 17, p. 4211, Aug. 2022.

[48] G. Kasmi, L. Dubus, P. Blanc, and Y.-M. Saint-Drenan, "Towards unsupervised assessment with open-source data of the accuracy of deep learning-based distributed PV mapping," 2022, *arXiv:2207.07466*.

[49] V. Golovko, A. Kroshchanka, S. Bezobrazov, A. Sachenko, M. Komar, and O. Novosad, "Development of solar panels detector," in *Proc. Int. Sci.-Practical Conf. Problems Infocommunications. Sci. Technol. (PIC ST)*, Oct. 2018, pp. 761–764.

[50] A. Greco, C. Pironti, A. Saggese, M. Vento, and V. Vigilante, "A deep learning based approach for detecting panels in photovoltaic plants," in *Proc. 3rd Int. Conf. Appl. Intell. Syst.* New York, NY, USA: Association for Computing Machinery, Jan. 2020, pp. 1–7.

[51] W. Chen, Z. Jiang, Z. Wang, K. Cui, and X. Qian, "Collaborative global-local networks for memory-efficient segmentation of ultra-high resolution images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 8916–8925.

[52] *Rapporto Statistico Sull'Energia in Piemonte*, Regione Piemonte, France, Switzerland, Accessed: Jan. 10, 2022.

[53] M. Mirzaei and M. Z. Mohiabadi, "A comparative analysis of long-term field test of monocrystalline and polycrystalline PV power generation in semi-arid climate conditions," *Energy Sustain. Develop.*, vol. 38, pp. 93–101, Jun. 2017.

[54] G. Ghiasi, Y. Cui, A. Srinivas, R. Qian, T.-Y. Lin, E. D. Cubuk, Q. V. Le, and B. Zoph, "Simple copy-paste is a strong data augmentation method for instance segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 2918–2928.

[55] V. Olsson, W. Tranheden, J. Pinto, and L. Svensson, "ClassMix: Segmentation-based data augmentation for semi-supervised learning," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 1369–1378.
[56] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 770–778.
[57] T. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 936–944.
[58] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 10684–10695.
[59] L. Di Stefano and A. Bulgarelli, "A simple and efficient connected components labeling algorithm," in *Proc. 10th Int. Conf. Image Anal. Process.*, 1999, pp. 322–327.
[60] S. Wei, S. Ji, and M. Lu, "Toward automatic building footprint delineation from aerial images using CNN and regularization," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 3, pp. 2178–2189, Mar. 2020.
[61] W. Tranheden, V. Olsson, J. Pinto, and L. Svensson, "DACS: Domain adaptation via cross-domain mixed sampling," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 1379–1389.

EDOARDO ARNAUDO received the M.S. degree in computer science with specialization in machine learning and artificial intelligence from UniversitÃ degli Studi di Torino (UniTo), in 2019. He is currently pursuing the Ph.D. degree in computer vision focusing on semantic segmentation applied to aerial and satellite imagery, which comprise his main research interests. In 2019, he was with LINKS Foundation, Turin, as an Applied Researcher in AI, data, and space domain in several multidisciplinary projects, with a focus on aerial and remote sensing applications.

GIACOMO BLANCO received the M.Sc. degree in computer engineering with a specialization in machine learning and artificial intelligence from Politecnico di Torino, Italy, in 2020. He is currently an Applied Researcher in AI, data, and space domain with LINKS Foundation on several research projects in image classification, image segmentation, text classification, and time-series analysis domains. His research interests include computer vision and natural language processing applications.

ANTONINO MONTI received the M.Sc. degree in computer engineering from Politecnico di Torino (PoliTo), with a specialization in software development, in 2022. He collaborated with LINKS Foundation to work on his master's thesis "Semi-supervised techniques for solar panel segmentation in aerial images," which provided part of the groundwork for the dataset used in this paper. His research interests include software and game development, data science, and machine learning.

GABRIELE BIANCO joined ITHACA to support international project focused on emergency mapping and satellite image processing for the extraction of thematic information, in January 2019. He is currently in charge of financial management with ITHACA S.r.l. and especially of the internal accounting of the rapid mapping service. He is also involved in the development and organization of the production team of the rapid mapping service.

CRISTINA MONACO received the Ph.D. degree in architectural heritage and landscape from the Polytechnic of Turin, in 2011. She is currently the Program Manager of the Earth Observation Business Unit, ITHACA S.r.l. She is also the Internal Project Manager of Copernicus Emergency Management Service (Rapid Mapping and Risk and Recovery Service). She acts as a Coordinator on the Copernicus In-Situ component of EEA on geospatial reference data. She is specialized in processing and interpretation of satellite imagery, production of thematic maps with GIS applications, assessment of reference spatial data, support to early impact analyses, and project management. Before joining ITHACA, in 2013, she was with the Higher Institute on Territorial Systems for Innovation, Turin (2011–2015) for national and international research projects relating to valuation and classification of natural and cultural heritage, preservation of environmental assets, UNESCO nomination, database development, historical maps, and planning and land management.

PAOLO PASQUALI received the master's degree in architecture from Politecnico di Milano, in 1998. He was with the De Agostini Group, Cartographic Department, as the Location-Based Services Manager (2000–2008) and a technical coordinator focusing on portable car navigation systems, web, and mobile applications. Then, he was with ITHACA, a non-profit association, from 2008 to 2021, as a Web GIS Developer and the Product Manager of several world bank, UN, and EU funded projects. He was a Core Developer of GeoNode (OSGeo Project) (2013–2019) and the Chair of the 2018 GeoNode Summit. He is currently a Senior Analyst with the LINKS Foundation Earth Observation Unit.

FABRIZIO DOMINICI received the M.S. degree in telecommunications engineering from Politecnico di Torino, in 2005. In 2005, he began with Istituto Superiore Mario Boella (ISMB) as a Researcher, focusing on satellite and navigation systems. In 2013, he became the Head of Research of the Mobile Solutions Area with ISMB, now LINKS Foundation, and the Director of the Microsoft Innovation Center of Torino (MIC), between 2012 and 2019. He is currently the Head of AI, Data, and Space Research Area with LINKS Foundation. He also manages and leads a large research group working in domains, such as IA, big data, GNSS technologies, and geospatial services. He has valuable experience managing multidisciplinary innovation projects at national and European levels. He supports the European commission as an expert, and he has been a coordinator of several EU-funded projects.

Open Access funding provided by 'Politecnico di Torino' within the CRUI CARE Agreement