IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received June 13, 2019, accepted July 16, 2019, date of publication July 25, 2019, date of current version August 9, 2019.

*Digital Object Identifier 10.1109/ACCESS.2019.2931144*

# Deep Learning-Based System for Automatic Recognition and Diagnosis of Electrical Insulator Strings

**CARLOS SAMPEDRO, JAVIER RODRIGUEZ-VAZQUEZ, ALEJANDRO RODRIGUEZ-RAMOS, ADRIAN CARRIO, (Member, IEEE), AND PASCUAL CAMPOY, (Member, IEEE)**

Centre for Automation and Robotics, Computer Vision and Aerial Robotics Group, Universidad Politécnica de Madrid, 28006 Madrid, Spain

Corresponding author: Carlos Sampedro (carlos.sampedro@upm.es)

This work was supported in part by the Ministry of Economy and Competitiveness through the National Research and Development Program under Grant INNPACTO IPT-2012-0491-120000 and in part by the Spanish Ministry of Science under Grant MCYT DPI2010-20751- C02-01. The work of C. Sampedro was supported in part by the Universidad Politécnica de Madrid and in part by the MONCLOA Campus of International Excellence.

**ABSTRACT** This paper presents a complete system for automatic recognition and the diagnosis of electrical insulator strings which efficiently combines different deep learning-based components to build a versatile solution to the automation problem of the power line inspection process. To this aim, the proposed system integrates one component responsible for insulator string segmentation and two components in charge of its diagnosis. The insulator string segmentation component consists of a novel fully convolutional network (FCN) architecture, termed Up-Net, which enhances the capabilities of the state-of-the-art U-Net network by introducing new skip connections at certain levels of the architecture. Furthermore, we propose a second variant of the Up-Net network by training it within a generative adversarial network (GAN) framework. The capabilities of the proposed Up-Net variants are incremented by the application of data augmentation and transfer learning techniques, achieving accurate segmentation of the insulator string elements (i.e., discs and caps). Regarding the insulator string diagnosis, we design a convolutional neural network (CNN) which takes as input the mask generated by the insulator string segmentation component and is capable of identifying the absence of a variable number of discs. The second diagnosis component consists of a novel strategy which integrates a Siamese convolutional neural network (SCNN) designed for modeling the similarity between adjacent discs and allowing the detection of several types of disc defects using the same model. The proposed system has been extensively evaluated in several video sequences from real aerial inspections of high-voltage insulators, showing robust insulator recognition and diagnosis capabilities.

**INDEX TERMS** Convolutional networks, deep learning, generative adversarial networks, intelligent fault diagnosis, power line inspection, semantic segmentation, Siamese networks, transfer learning.

## I. INTRODUCTION

Insulator strings are an important component in high voltage power transmission lines owing to their role in electrical insulation as well as mechanical support. Damage in these elements can cause serious outages in the power transmission lines. Thus, prevention of possible failures in these elements is a priority for electric companies. Insulators are exposed to wildlife and meteorological conditions such as rain, wind, or snowfall. As a consequence, these components are vulnerable to the appearance of defects such as ruptures, pollution, and contamination, which can even cause explosions. This forces electrical companies to carry out inspections on a regular basis to prevent the appearance of the aforementioned defects.

In recent years, diverse inspection strategies have been developed using manned and unmanned aerial solutions in order to inspect electrical transmission lines efficiently. Nowadays, one of the most common approaches is the use of manned helicopters flying along the power line corridor and equipped with several sensors for recording the inspection data (RGB cameras, IR cameras, LiDAR, etc.).

The associate editor coordinating the review of this manuscript and approving it for publication was Hui Xie.

However, this kind of inspection is usually very demanding in terms of human and energy resources. In addition, inspections with manned helicopters may present some risk because the helicopter must fly close to the electric power transmission line in order to obtain a better capture of the inspection data.

In order to address these limitations, several alternative solutions from the field of mobile robotics have emerged: the use of Unmanned Aerial Vehicles (UAVs) [1], [2] and Rolling on Wire (ROW) robots or climbing robots [3], [4]. UAVs have the main advantage over ROW robots in the sense that their design does not need to be adapted to a different scenario (e.g. high versus medium voltage power lines). As a counterpart, UAVs and particularly multirotors, have an important limitation in terms of flight endurance, which makes it very difficult to inspect large power transmission corridors during the same flight.

Regardless of the type of strategy used to capture the data, the huge amount of data acquired during the inspection are analyzed by an experienced crew of human inspectors. This process is extremely time-consuming since the specialized inspectors are required to exhaustively review the video sequences searching for potential defects in the elements of the electric power transmission lines. In order to reduce the considerable workload derived from this inspection process, research efforts are currently focused on automating the diagnostic process of these elements. Building such an automatic system requires robust insulator detection and diagnosis algorithms capable of dealing with the enormous variability in the captured data. In this context, images captured from helicopters, UAVs, or other platforms used for power line inspection, usually have very complex and heterogeneous backgrounds. In addition, insulators can be made of different materials (glass, ceramic, composite, etc.) which drives to a huge variability of the appearance of the insulator in the image. Glass and ceramic insulators surfaces can have reflections due to sunshine, which greatly hinder the diagnosis of possible defects. Furthermore, depending on the platform used to capture the images, insulators can appear with very different points of view in the image and even present occlusions, which significantly complicates the tasks of inspection and diagnosis.

Added to the above difficulties is the enormous variability of defects that can appear in the insulator strings, varying from burned, rusted, polluted, or cracked elements to the presence of bird excrements on their surface. The latter has proven to be an important cause in the appearance of defects in electrical power supply facilities, such as flashovers [5]. Furthermore, some defects are rare and only appear under particular conditions. Thus, there is a significant difficulty in obtaining enough data for each type of defect in all its variants. Under these constraints, where only a small number of examples per defect type are available, and when only some types of defects are known in advance at the time of training, traditional classification methods have some limitations as highlighted in [6]. In these situations, Siamese neural networks have shown to be an effective solution, providing remarkable results in other fields of computer vision such as face verification [6] and character recognition [7]. On the other hand, one of the requirements of electric companies is the detection of defects in early stages which reduces the risk of power cuts with consequent savings for electric companies.

In order to accurately detect and segment the insulator string in RGB images under the aforementioned conditions, and to be able to precisely detect the defects and diagnose them even in early stages, a versatile system with advanced capabilities is required. For this purpose, strategies based on traditional computer vision algorithms can provide appropriate results in structured images under controlled illumination and background conditions. However, as stated in [8], most of these approaches are based on heuristics which require several assumptions and handcrafted thresholds [9], [10] which need to be manually tuned and re-adjusted to properly operate under previously unseen conditions. In contrast, strategies based on machine learning techniques, when trained in meaningful datasets, can provide more flexible solutions, making the insulator recognition and segmentation process more robust to changes in illumination, background, type of insulator, etc. As a drawback, machine learning models and especially deep learning ones, usually require a large amount of data to be trained properly.

Despite this limitation, recent advances in deep learning, and more concretely in semantic segmentation, have demonstrated outstanding capabilities for solving computer vision problems in a wide range of industrial applications, facilitating their automation [11]. However, fully autonomous solutions for power line inspection, that is, without the intervention of human inspectors in the inspection process, are far from being a reality. This means that in production systems, the presence of specialized human operators in the last stages of the inspection process is inevitable. Taking this argument into consideration, this paper focuses on the development of a reliable system which can be integrated into an actual production system for reducing the workload derived from completely manual inspections. For this purpose, we do not focus specifically on the accurate detection of defects in the image but instead, we focus on providing accurate predictions of the frames that contain defective insulator strings within a video sequence. As a result, only the frames where our system has automatically found the insulator string to be defective will be further reviewed by a specialized human inspector, resulting in shortened inspection times.

The main contributions of this work, which are aimed to achieve the aforementioned objectives are:

- A fully automatic system for insulator string recognition and diagnosis. The proposed system is based on two main subsystems for insulator string segmentation and fault diagnosis and has been validated over several video sequences captured in real aerial inspections, containing different types of insulators, backgrounds, defects, etc.
- A novel Fully Convolutional Network (FCN) architecture for insulator string segmentation. The proposed approach is able to accurately segment the elements of

the insulator string (caps and discs) facilitating the subsequent stages for fault diagnosis. Furthermore, we validate the proposed network using an adversarial training procedure based on a conditional Generative Adversarial Network (GAN) framework [12].

- A new strategy for detecting the absence of disc insulator units within the insulator string. The proposed approach uses a Convolutional Neural Network (CNN) trained directly on the masks generated by the insulator segmentation component and is capable of identifying the absence of a variable number of disc units within the insulator string.
- A novel strategy for intelligent diagnosis of damaged discs based on a Siamese Convolutional Neural Network (SCNN), which is capable of diagnosing several types of disc units defects using the same network. The proposed method is mainly trained using synthetic images of defective disc units and is evaluated in different sequences from real aerial inspections of high voltage insulators.

The remainder of the paper is organized as follows: Section II provides an overview of the related work. Section III introduces the proposed system for insulator segmentation and fault diagnosis. In Section IV, the experiments conducted and the results obtained during the evaluation of the proposed system are described before being discussed in Section V. Finally, Section VI provides the conclusions and future research directions.

## II. RELATED WORK

In the past decades, most of the approaches for detecting the electric tower and the insulators in RGB images were based on traditional computer vision algorithms using color, shape, and texture information. As an example, direct template matching was used in [13]. In [14] a corner detector was combined with a corner matcher in order to track the movement of pole tops of medium voltage towers. In [15], a median filter was applied in the HSL color space for segmenting medium voltage towers, which was combined with a watershed segmentation for refining the detection of the towers when no cross arms were present. The Graph-cut algorithm was utilized in [16] for power pole segmentation. More recently, authors in [9], [17] used color information as the main feature for segmenting the insulator string. In both works, several thresholds in the RGB color space were identified from a set of 100 insulator images.

Recent advances in power line inspection make use of machine learning techniques in order to identify the elements that comprise the electric power transmission infrastructure. Support Vector Machine (SVM) classifiers were used in [18]–[20] for insulator identification and analysis. In [18] Gabor features were used to train an SVM classifier for insulator identification. Qualitative results were only reported for one type of insulator and two aerial images. In [19] Local Binary Pattern (LBP) and wavelet features were combined for training an SVM classifier in order to diagnose medium voltage insulators among three categories (good, marginal or risky). Results were only reported on a test set composed of 30 images. SVMs were also used for electric tower detection in images of several spectra (visible and infrared) in [21].

In the last five years, machine learning strategies have evolved towards the usage of deep learning techniques, which have proven to provide outstanding results for object recognition in multiple applications. For this reason, several authors have addressed the power line inspection problem by using deep learning models: in [20], insulator diagnosis was performed by means of a pre-trained CNN based on the AlexNet architecture [22] for feature extraction followed by an SVM with RBF kernel classifier. In [23], two models were trained and compared for electric wire detection: a pre-trained CNN based on the GoogLeNet architecture [24], and a CNN model with 4 convolutional layers where HOG features [25] were used as input to the CNN. In the approach presented in [26], GoogLeNet pre-trained network was also used for addressing the detection of several power line components such as wires, pylons, and insulators, where the output of the CNN model was post-processed using a spectral clustering algorithm. More recently, several authors have addressed the problem of insulator detection by using consolidated object detectors such as Faster R-CNN [27], YOLOv2 [28] or Single Shot multibox Detector (SSD) [29]. The former was used in [30] and [31], YOLOv2 detector was utilized in [32] for detecting different type of insulators (polymer and ceramic insulators). SSD pre-trained using the COCO dataset was used in [33] to detect composite and ceramic insulators. SSD in combination with deep residual networks was also utilized in [2] for detecting different components and faults in medium voltage electric towers.

Within the field of deep learning, most recent approaches for insulator string segmentation make use of Semantic Segmentation algorithms [8], [30], [31]. In [30] and [31] Faster R-CNN was used for insulator localization in RGB images. After locating the insulator, in [30] an FCN-8s [34] allowed the extraction of each disc of the insulator string for posterior diagnosis tasks. Again, only qualitative results were presented in a very limited set of test images, showing inaccurate disc segmentation results. In [31] U-Net [35] was used for segmenting the pixels of the absent disc unit within the insulator string. A conditional GAN framework was adopted in [8] for insulator segmentation providing good segmentation results for a large variety of insulator types with different material, point of view, and orientation. However, a posterior stage for fault diagnosis was not addressed in their solution.

Regarding the fault diagnosis problem, most of the approaches in the state-of-the-art are focused only on the identification of absent disc units within the insulator string [9], [17], [30], [31], [36]. Moreover, most of these approaches based their diagnosis capacity on the computation of the distance (in pixels) between adjacent disc units, making them prone to unpredictable results when more than one disc is missing in the insulator string. In this paper, we use a

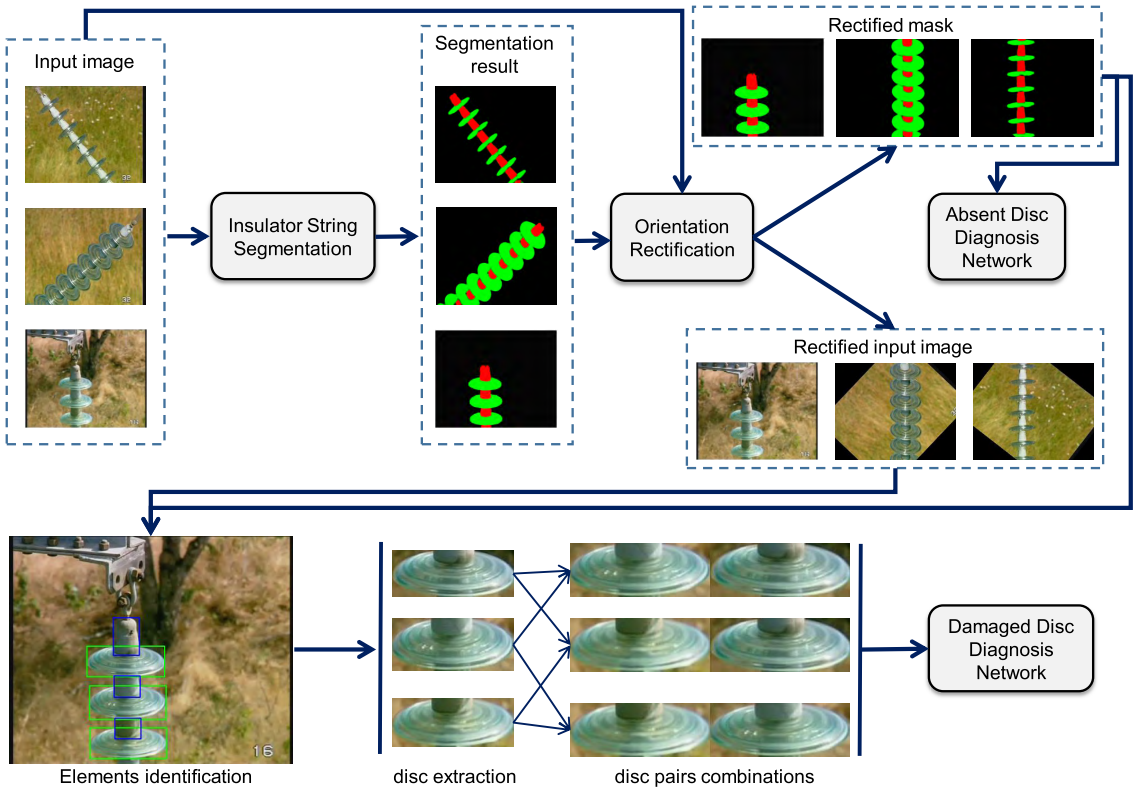


**FIGURE 1.** System description (best seen in color). The proposed system is mainly based on the combination of an insulator string segmentation component, a post-processing component for rectifying the orientation of the insulator string, and two components responsible for fault diagnosis.

CNN for the identification of absent disc units regardless of the number of disc units missing in the insulator string. Despite the problem of absent disc units is very relevant due to its implications on possible interruptions of the power system, other types of defects, such as dirtiness or pollution in insulator discs, may also produce severe damages to the infrastructure such as flashovers [5]. Only a few works in the literature address the identification of multiple defects. In [37], five types of defects were addressed by using a Nearest-Neighbor algorithm on handcrafted Local Directional Pattern (LDP) descriptors with Chi-square similarity measure. In [20], authors addressed the diagnosis of cracked and dirty insulator discs using a CNN for feature extraction in combination with an SVM with RBF kernel. However, the results presented in these works were only validated in a few number of images of defective components. In addition, most of the works in the state-of-the-art for insulator recognition and diagnosis do not provide any result in video sequences captured during real flights, which would greatly help in assessing their results in a qualitative way. In this paper, we conduct an extensive evaluation of the proposed system over several video sequences captured during real aerial inspections, each of them composed of thousands of frames.

As compared to most of the previous approaches in the literature for insulator recognition and diagnosis, in this paper we do not use an insulator detection algorithm prior to its segmentation. Alternatively, we explore the effectiveness of directly using semantic segmentation on the entire input image with the purpose of saving computation resources. In addition, for diagnosing multiple types of defects, we adopt an SCNN architecture responsible for computing the similarity between adjacent disc units. A similar concept about measuring the similarity between adjacent disc units was also presented in [38], with the difference that in our approach the weights of the SCNN are learned for maximizing the similarity/dissimilarity between pairs of non-defective or defective discs respectively, whereas in [38] these descriptors were fixed and handcrafted.

In our previous works [21], [39], [40] we presented an automatic electric tower detection and tracking system for operating in fast speed helicopter inspections (non-intensive inspections). In this paper, we extend our previous works for providing a complete solution in intensive inspections, where the higher quality of the captured images allows the application of systems for fault diagnosis.

## III. SYSTEM DESCRIPTION

The main objective of the proposed system is to automate the process of fault diagnosis in the RGB images captured after the inspection flights. For this purpose, the proposed system is based on the combination of different deep

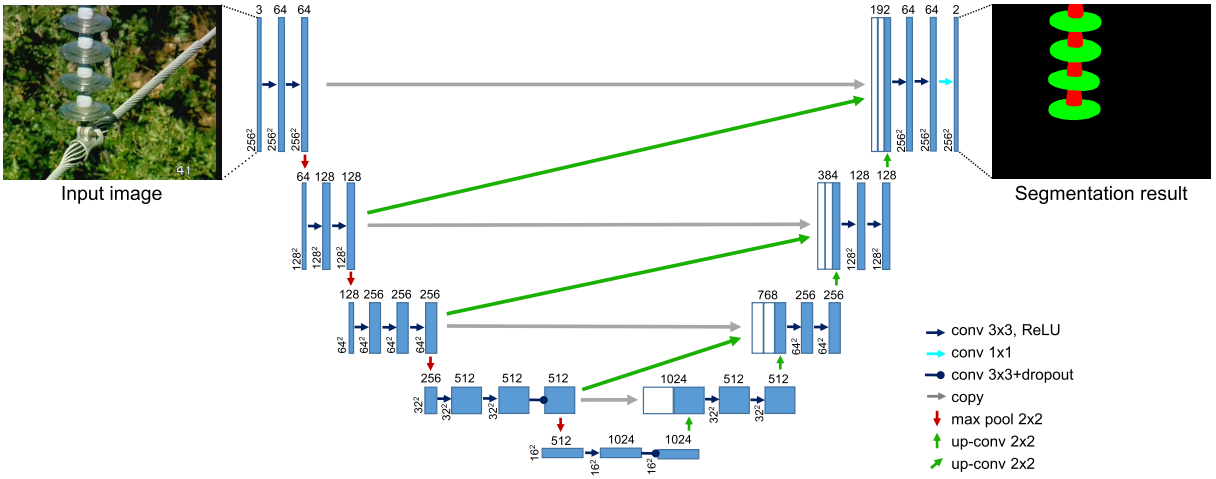


**FIGURE 2.** Up-Net architecture designed for insulator string segmentation (best seen in color). The U-Net architecture presented in [35] is modified in this work by adding new skip connections (green oblique arrows). In addition, when transfer learning is conducted from VGG16, new convolutional layers are added to the encoding path, yielding Up-Net_vgg16 (in the figure) which has 35.45M parameters. Up-Net takes an RGB image as input and outputs a 2-channel image with pixels belonging to *cap* and *disc* classes. This illustration follows the structure presented in [35] for better comprehension and comparison with their seminal work.

learning-based components for insulator string segmentation and fault diagnosis (see Fig. 1). The input to the system consists of an RGB image which is subsequently processed by the insulator string segmentation component. As a result, a 2-channel mask is obtained where the different elements in the insulator string (i.e., caps and discs) are classified at pixel level. Using the resulting mask, we compute the orientation of the insulator string and rectify it for further processing. The rectified mask is used by the absent discs diagnosis component (see Section III-C.1) in order to predict if the current frame contains a defective insulator string in these terms (i.e., absent disc units). At the same time, the rectified mask and input image are used to extract the Region Of Interest (ROI) of every element in the insulator string using image processing algorithms based on contours extraction and their subsequent filtering by size. Using these ROIs, the cropped images corresponding to the different discs are combined in pairs and passed to the SCNN, which works as the core of the damaged disc diagnosis component (see Section III-C.2).

### A. INSULATOR STRING SEGMENTATION

This component is based on a fully convolutional neural network used for semantic segmentation purposes. The proposed FCN network uses the base architecture of the U-Net [35] network, which is modified by adding several ''up-skip'' connections at certain levels of the architecture. The proposed network, termed Up-Net (see Fig. 2), is inspired by the FCN networks design proposed in [34], where the skip connections are used in order to fuse more local information extracted by shallower layers, which have smaller receptive fields, with the semantic information extracted from deeper layers. The ''up-skip'' connections are added in consecutive levels of the U-Net architecture (see oblique green connections in Fig. 2), leveraging the benefits of the FCN networks described in [34] and the U-Net network proposed in [35].

The proposed Up-Net is composed of an encoding and a decoding paths. When transfer learning is conducted from VGG16 (Up-Net_vgg16), the encoding path follows the VGG16 architecture proposed in [41], which consists of a combination of $3 \times 3$ convolutional layers (padded convolutions) with a stride of 1 and ReLU activation functions, followed by $2 \times 2$ max-pooling layers with a stride of 2. After each pooling operation, the number of channels is increased by a factor of 2 until reaching 512 channels in the fourth convolutional block of the encoding path. The decoding path is nearly symmetric to the encoding one, and consists of a combination of upsampling and convolutional layers. In the upsampling layers, a nearest-neighbor interpolation with an upsampling factor of 2 is applied to double the resolution of the feature maps in the previous level. This upsampling operation is always followed by $2 \times 2$ convolutional layers (''up-convolution'') in order to halve the number of channels, which allows the application of the skip connections to fuse the information coming from the encoding path. Each ''up-convolution'' is followed by $3 \times 3$ convolutional layers (padded convolutions) with a stride of 1 and ReLU activation functions, mimicking the design of the encoding path. The last layer of the decoding path consists of a $1 \times 1$ convolutional layer in order to reduce the number of feature channels to the number of output labels in the model.

In the specific case of the proposed system, we have considered two labels for classifying the pixels as belonging to the *cap* and *disc* classes. The activation function of the last layer implements a sigmoid activation function which maps the output predictions of the network to the interval $[0, 1]$, where 0 encodes the pixels that belong to the *background* class and 1 represents the labels associated to the *cap* or *disc* class. As compared to previous state-of-the-art works in

which the proposed segmentation algorithms consider the insulator string as a whole, we believe that the proposed strategy of segmenting the different elements of the insulator string in separate classes can greatly facilitate posterior stages for component identification and fault diagnosis.

### B. INSULATOR STRING ORIENTATION RECTIFICATION

This component of the system is responsible for computing the orientation of the segmented insulator string in the image and rectify its orientation for the subsequent operation of the fault diagnosis components. In order to identify the current orientation of the insulator string, this component uses a pixel-wise OR operation in order to fuse the information of the different channels from the output mask provided by the insulator string segmentation component. Once the information is fused, giving as a result a single-channel image, computer vision techniques are applied in order to detect the most prominent contour in the image associated with the insulator string. A rotated bounding box enclosing the insulator string contour is then extracted, which allows computing the angle of the insulator string with respect to the horizontal image axis. Having identified the actual orientation angle, a *warping* operation consisting of a 2D rotation is applied to the pixels of the original RGB and segmentation mask images in order to rotate vertically the insulator string. Some examples of the output provided by this component can be reviewed in Fig 1.

### C. FAULT DIAGNOSIS

In this work, we divide the diagnosis of defects in the power line insulator strings into two separate problems. On the one hand, we leverage the output provided by the insulator string segmentation component and use it as input to a CNN model trained for recognizing if there are disc units absent within the insulator string. This approach is described in detail in Section III-C.1. Second, we address the diagnosis of defects in the disc units of an insulator string by designing a Siamese network which takes as input pairs of disc images extracted from the insulator string. This approach is explained in detail in Section III-C.2.

#### 1) DIAGNOSIS OF ABSENT DISC UNITS

This component consists of a 10-layer CNN (see Fig. 3) which takes as input the segmentation mask (extended to 3 channels) provided by the insulator string segmentation component, and outputs a probability of the occurrence of the disc absence defect. Thus, the training procedure of this network has been conducted using only RGB masks such as the one shown as the input image in Fig. 3. This design provides important capabilities to our system:

- A large number of training samples can be synthetically generated by randomly subtracting disc units within the insulator string mask.
- A variable number of disc units can be removed from the mask, making the proposed network robust to a large number of possible defect configurations where several discs can be absent from the insulator string.

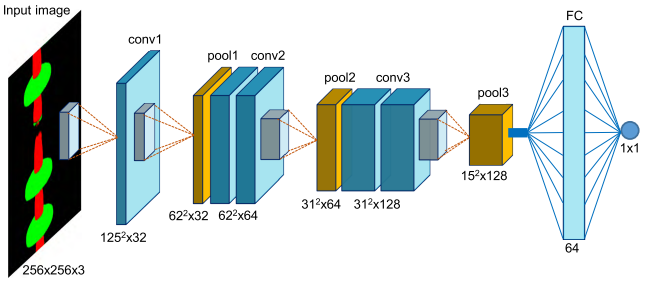


**FIGURE 3.** Proposed Convolutional Neural Network architecture for diagnosing the absence of disc units within an insulator string. The proposed CNN takes as input a segmented insulator string image and outputs the probability of discs absence. "FC" stands for fully-connected.

This property makes this component more flexible and robust as compared to other state-of-the-art works trained directly on the original RGB images, where there is a limited amount of defect configurations.

The architecture of the proposed CNN for disc absence diagnosis is shown in Fig. 3. As shown in Fig. 3, the proposed CNN consists of 5 convolutional layers, 3 max-pooling layers, 1 fully-connected layer, and a single-unit output layer. The first layer in the architecture applies a $7 \times 7$ unpadded convolution with a stride of 2 and ReLU activation function, followed by a $3 \times 3$ max-pooling layer with a stride of 2. The following layers alternate $7 \times 7$ convolutional layers (padded convolutions) with a stride of 1 and ReLU activation function, with $2 \times 2$ max-pooling layers with a stride of 2. In each level of the architecture, we increase the number of feature channels by a factor of 2 until reaching 128 channels in the last convolutional and pooling layers. The resulting $15 \times 15 \times 128$ volume is flattened and introduced to the fully-connected part of the CNN, which contains 64 hidden units, featuring ReLU non-linearity as the activation function, connected to a single-unit output layer. The final output unit consists of a sigmoid unit which allows for providing predictions in the range [0, 1], where 1 indicates a fault in the insulator string, and 0 reveals that the insulator string is not defective.

Several CNN architectures have been studied, conducting a detailed analysis about the number and size of the convolutional layers and the size of their receptive field, with the intuition that larger receptive fields can be beneficial for detecting the absence of disc units within the insulator string. For this purpose, several CNN architectures with varying kernel sizes have been evaluated. The results of this evaluation are presented in Section IV-C. The CNN architecture adopted in the final system is the one shown in Fig. 3.

#### 2) DIAGNOSIS OF DAMAGED DISC UNITS

In this section, we describe the strategy designed for diagnosing different types of defects in the disc units of the insulator string (e.g. disc polluted, rusted, burned, etc.). For convenience, we will name any type of disc unit defect as ''damaged'' from now on in the document. Insulator strings can appear under a wide range of conditions within the

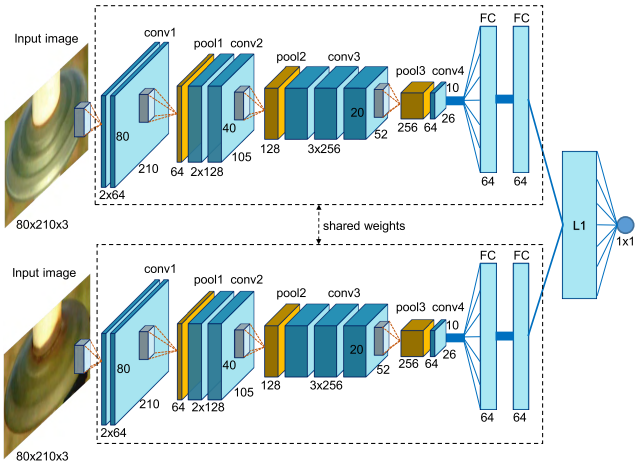


**FIGURE 4.** **Proposed Siamese Convolutional Neural Network architecture for diagnosing damaged disc units. The SCNN takes as input a pair of disc images within the insulator string and outputs the probability of dissimilarity between the input images. "FC" stands for fully-connected.**

image (e.g. different color, texture, point of view, lighting conditions). Moreover, the great majority of insulators disposed along the power transmission lines are ceramic or made of glass, which usually causes the presence of reflections in the discs (see input image in Fig. 2). This constraint can significantly complicate the diagnostic process of the insulator string elements due to visual similarities with respect to certain types of disc surface pollution. In order to build a robust diagnosis system able to tackle the aforementioned constraints, in this work we propose a novel strategy based on an SCNN architecture (see Fig. 4). As shown in Fig. 4, the proposed SCNN consists of two twin CNNs with shared weights which are joined after the fully-connected layers by an energy function, which computes a distance metric between the hidden states of the twin CNNs. In this work, we have evaluated different SCNN architectures and energy functions whose classification results are shown in Section IV-D. The best performing SCNN is the one shown in Fig. 4, where each twin CNN follows a VGG16 architecture until the third convolutional block (conv3). After this, a max-pooling layer with kernel size and stride of 2 is followed by a $1 \times 1$ convolutional layer, which plays an important role in reducing the number of channels before the fully-connected network. The resulting volume is flattened and introduced to the fully-connected part of each twin CNN, consisting of two hidden layers of 64 units each, using ReLU activation function. Dropout with 50% probability is introduced between the two hidden layers. The two hidden states from the twin CNNs (denoted by $h_1$ and $h_2$) are then introduced to the L1 layer followed by a single logistic output unit. Similar to the work in [42] and [7], this final block of the SCNN computes a weighted L1 distance between the hidden states of the form $d(h_1, h_2) = \sigma(\sum_i \omega_i |h_1^{(i)} - h_2^{(i)}|)$, where $\sigma$ is the sigmoid activation function, and $\omega_i$ are the weights learned in the output layer. The output of the SCNN computed by the sigmoid output unit is within the range [0, 1], where 0 encodes the condition where the two input images are the "same", which means no defect in the discs, and 1 encodes the condition where the two input images are "different", thus, indicating the presence of a defect in one or both discs.

In order to identify damaged disc units and locate them within the insulator string, our strategy uses Algorithm 1, which is based on computing all the combinations between pairs of discs within the insulator string in the current frame using ${}^nC_2 = n!/(2!(n-2)!)$, where $n$ is the number of detected discs in the current frame. Each of these combinations of cropped images (see Fig. 1) is introduced to the SCNN model for classifying them as "damaged" or not. From these combinations, we extract a list of the corresponding disc pairs that have been classified as "damaged" by the proposed SCNN and compute the most frequent elements in the list. If the number of occurrences of these elements is greater or equal to $n-1$ it means that all remaining discs within the insulator string have voted for the specific disc as being "damaged", and thus the defective disc is located. It should be noted that the proposed strategy needs at least two disc units in order to start diagnosing the insulator string, and at least three disc units in order to precisely locate the defective disc in the image. In the former case, when only two disc units are visible at the current frame, we mark the corresponding pair of discs as "damaged", in case there is a defect.

It should be noted that despite the proposed strategy can increase the computational cost of the whole system since

**Algorithm 1** Damaged Discs Diagnosis Strategy

**Input:** $I_{rgb}$ (current RGB frame), $I_{mask}$ (current segmented frame), model (SCNN model)

**Output:** `damaged_disc_flag`, `L_defect_ids` (ids of damaged disc units)

```
damaged_disc_flag ← false
n ← ComputeNumDiscs(I_mask)                    ▷ e.g. n = 3
C ← nC2 = (n 2)                                ▷ compute disc combinations
for each c in C do
    img_a ← I_rgb[c[0].roi]
    img_b ← I_rgb[c[1].roi]
    P(damaged) ← model.Predict([img_a, img_b])
    if P(damaged) > threshold then
        L_ids.Add(c[0].index)
        L_ids.Add(c[1].index)
    end if
end for                                        ▷ e.g. L_ids={0,1,0,2}
count ← CountVotes(L_ids)                      ▷ e.g. count={0:2, 1:1, 2:1}
if max(count.values) ≥ n − 1 then
    damaged_disc_flag ← true
    for k ← 0 to n − 1 do
        if count[k] == n − 1 then
            L_defect_ids.Add(k)
        end if
    end for
end if                                         ▷ e.g. L_defect_ids={0}
```

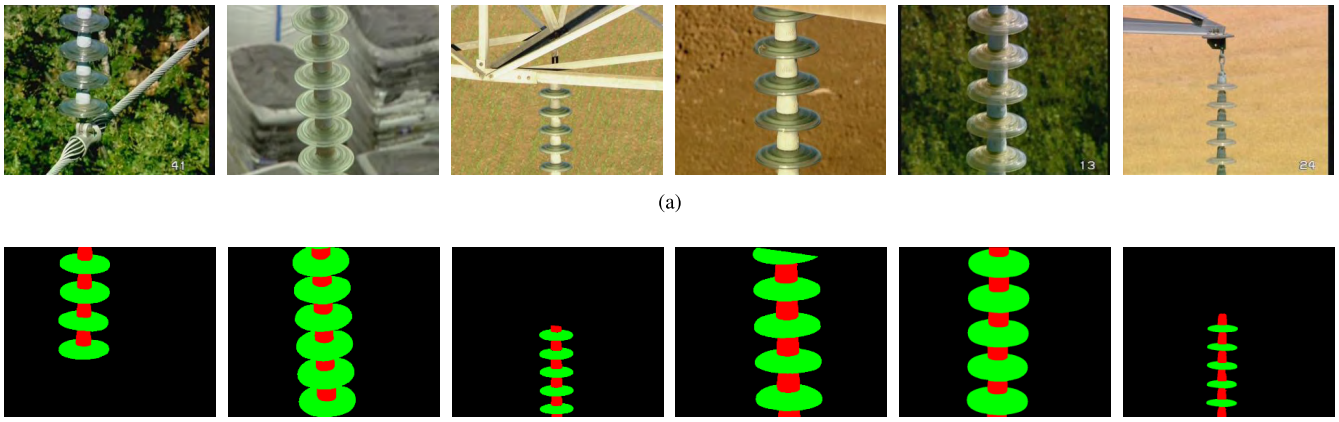

**FIGURE 5.** Examples of images used for training the insulator string segmentation models. a) Original RGB images. b) Ground truth annotations. The pixels of the insulator string are assigned to the *disc* (green) or *cap* (red) class.

all the disc pairs combinations are introduced to the SCNN, it provides an additional filtering stage which allows for the removal of possible false positives.

## IV. EXPERIMENTAL RESULTS

The aim of this section is to describe the experiments and present the results obtained during the evaluation of the proposed system. In Sections IV-B to IV-D we provide a detailed evaluation of the models selected for each component of the system. In Section IV-E the experiments and results regarding the evaluation of the final system are presented.

All the videos relative to the results presented in Section IV-E for the evaluation of the proposed final system can be reviewed in: https://vimeo.com/album/5782445.

### A. EXPERIMENTAL SETUP

All the networks described in Section III have been implemented in Python and trained using the Keras[1] and PyTorch[2] deep learning frameworks. The latter has been used for working with all the models that involve GANs [43], while the former has been used for working with the remaining models. All the experiments have been conducted on an Nvidia GeForce GTX 1080Ti GPU under Ubuntu 16.04 LTS. OpenCV 4.0.0[3] is used for image processing tasks.

### B. INSULATOR STRING SEGMENTATION

This section presents the experiments that have been conducted in order to select the most appropriate model for the insulator string segmentation task. For this purpose, several state-of-the-art semantic segmentation networks have been evaluated and compared to the proposed Up-Net architecture. In addition, a thorough analysis has been conducted for studying the effects of three main aspects:

- Data augmentation: all the networks have been trained using the original dataset composed of 160 images, and an augmented dataset containing 640 images.
- Transfer learning: most of the networks evaluated in this section have been trained after transferring the weights of the VGG16 network previously trained on the ImageNet database [44].
- Generative adversarial networks: we evaluate the performance of state-of-the-art conditional GANs applied to the insulator string segmentation task. In addition, we integrate the U-Net and Up-Net networks within a conditional GAN framework (GAN_U-Net and GAN_Up-Net) in order to study its influence. In this case, no transfer learning is performed in order to properly compare to other baseline networks involving GANs.

#### 1) DATASET

In order to train and validate all the models implemented for the task of insulator string segmentation, a total of 160 images of $720 \times 576$ pixels were manually annotated, where the pixels of the insulator string were assigned the *disc* or *cap* class. Some examples of the original images and the ground truth annotations can be reviewed in Fig. 5. In order to analyze the effect of the number of training samples in each evaluated model, and taking into account the characteristics of the images that make up the original dataset, in this work two main data augmentation techniques have been considered. First, all the images have been augmented by a horizontal flip, thus doubling the number of images in the dataset. Subsequently, a gamma correction technique is applied to the original image by applying a *lookup table* (LUT), where each value (pixel intensity) in the table is computed using the expression: $LUT[i] = (i/255)^{\gamma} \times 255$, where $i$ ranges from 0 to 255 and we clip the previous expression to be within the range [0, 255]. For each image in the original dataset, two random values of gamma were picked from the range [0.4, 0.8] for the first value and [1.2, 2.5] for the second. Using the aforementioned data augmentation techniques, the original dataset was augmented by a factor of 4, obtaining a total of 640 images.

[1] https://keras.io/
[2] https://pytorch.org/
[3] https://docs.opencv.org/4.0.0/

In order to perform a detailed evaluation and comparison of all the networks presented in this section, a test set of 240 images was created and manually annotated. This test set includes images from a similar domain as the ones presented in the training set, and 75 images (i.e., 31%) with a completely different domain, which were downloaded from the Internet.

#### 2) TRAINING METHODOLOGY

All the networks have been trained by considering a binary classification problem for each of the two classes (i.e., disc or cap), where each of the two channels of the output layer of the corresponding network computes a sigmoid function which models the probability of the corresponding pixel as belonging to *disc* or *cap* class versus the *background* class. Thus, a *binary cross-entropy* is used as the loss function, using the Adam optimizer [45] for its minimization during a training process of 60 epochs and a minibatch size of 4 images. The learning rate used in the Adam optimizer has been empirically found for each network using a grid search procedure, being $10^{-4}$ for the proposed Up-Net and GAN_Up-Net networks. The rest of the hyperparameters are set as described in the original work [45]. We apply the weight initialization technique proposed in [46] where the initial values of the weights are sampled from a truncated normal distribution centered on 0 and standard deviation of $\sqrt{2/fan\_in}$, where *fan_in* is the number of input units in the weight tensor.

Inspired by the results obtained in [47] we also train the best performing models under the conditional GAN framework proposed in pix2pix [47]. GANs are generative models which learn a mapping function from a random noise vector to an output image ($G : z \mapsto y$). Conversely, in the conditional GAN framework, the mapping is learned from a random noise vector and an input image ($G : \{x, z\} \mapsto y$).

The loss functions of the generator and discriminator used for training the conditional GAN models are computed using (1) and (2) respectively. As can be seen in (1), the loss function encompasses two terms. The first term implements an L1 operation, aiming at modeling the low-frequency part of the desired output. The second term is intended to model the high-frequency part of the signal (edges and complex details) and is learned using a second network (discriminator). This new network is trained using (2) in order to distinguish ground truth labels from the ones generated by the generator network. During training, the two networks (generator and discriminator) take turns to update their parameters. We have found that performing a training step every time a batch is sampled is the most appropriate method for obtaining a correct convergence in the specific task of insulator string segmentation presented in this work.

$$\mathcal{L}_{GEN}(G, D) = \mathbb{E}_{x,y,z}\left[||y - G(x, z)||_1\right] + \alpha\left(\mathbb{E}_{x,z}\left[\log(D(x, G(x, z))\right]\right) \quad (1)$$

$$\mathcal{L}_{DIS}(G, D) = \mathbb{E}_{x,y}\left[\log D(x, y)\right] + \mathbb{E}_{x,z}\left[\log(1 - D(x, G(x, z))\right] \quad (2)$$

where $G$ and $D$ denote the generator and discriminator networks respectively, $x$ and $y$ represent the input RGB image and the desired output mask respectively, and $z$ is a random noise vector. The parameter $\alpha$ in (1) models the influence of the conditional GAN framework. This parameter has been selected by performing a grid search over a wide range of values ($\alpha \in [0.01, 50]$), obtaining an optimal value of $\alpha = 0.1$.

In all the experiments, the images of the corresponding dataset (i.e., original or augmented) are resized to $256 \times 256$ pixels and randomly divided into train and validation sets with a percentage of samples of 80% and 20%, respectively.

#### 3) BASELINE METHODS

A detailed evaluation of the proposed method for segmenting insulator strings has been conducted by comparing its performance to several state-of-the-art algorithms which have proven to provide outstanding results in different semantic segmentation problems. In the next paragraphs we provide a brief description of the implementation details of the different baseline approaches considered:

- Fully Convolutional Networks (FCNs): we implement the FCN-8s model presented in [34] with the base architecture of the VGG16 network and adding the skip connections described in the original work which allow fusing the information from the last convolutional layer (conv7) with the information of lower levels in the architecture (pool4 and pool3). This is done by applying $4\times$ and $2\times$ upsamplings using deconvolution layers. In addition, we experimented with several versions of the FCN-8s model with and without L2 regularization with a decay of $2^{-4}$. The resulting model has 134.27M parameters. The Adam optimizer with a base learning rate of $10^{-4}$ is used for training the FCN-8s network.
- SegNet: this network follows the implementation described in [48] which consists of an encoder-decoder architecture, in which the encoder is based on the VGG16 architecture with 13 convolutional layers, each of one having its analogous convolutional layer in the 13-layer decoding network, yielding a model of 29.46M parameters. Each convolutional layer in the encoder and decoder networks is followed by a batch normalization layer after which a ReLU non-linearity is applied. As described in the original work, the upsampling operations are performed by using the memorized pooling indices from the max-pooling layers of the encoder network. In the case of the SegNet model, the most suitable learning rate for the Adam optimizer has been empirically obtained, being $5^{-5}$.
- Convolutional Autoencoder (CAE): the topology of this model follows the encoder-decoder architecture of the SegNet network with the only difference being that in this case we do not use the pooling indices from the encoder network but instead the upsampling operations use a nearest-neighbor interpolation. Thus, the number of parameters of this model is the same as the SegNet

architecture. This model is trained using a base learning rate of $10^{-3}$ for the Adam optimizer.

- U-Net: in this case, two types of architectures have been implemented depending on whether transfer learning is carried out or not. The first architecture in which transfer learning is not applied, named U-Net as the original work, follows the architecture described in [35], having about 31M parameters. The only difference with the original U-Net architecture is the application of padded convolutions, and two 50% dropout layers in the fourth and fifth convolutional blocks. In the second architecture, termed U-Net_vgg16, we perform transfer learning from the VGG16 network trained on the ImageNet database. In this case, the original encoder network from the U-Net architecture is extended from 10 to 12 convolutional layers, yielding an architecture with 33.98M parameters, where the first 10 convolutional layers are identical to those of the VGG16 network. All the networks consisting of a U-Net-based architecture are trained considering a base learning rate of $10^{-4}$ for the Adam optimizer.

The implementation details of the baseline models trained within a GAN configuration are summarized in the next paragraphs.

- ISNet: we implement the conditional GAN network described in [8], having a generator with 9.65M parameters. The implementation follows closely the one explained in the original paper with two main differences. First, we only perform a single-stage training instead of the two-stage training finally suggested by the authors, and second, we perform a grid search of all the parameters (not provided by the authors) as described in Section IV-B.2.
- pix2pix: this model follows the generator-discriminator architecture presented in [47], where the generator consists of a U-Net architecture and the discriminator is based on an $N \times N$ "PatchGAN", in which the discriminator is responsible for distinguishing between real and fake image patches. In this case, we use directly the released code[4] provided by the authors, in which they include several improvements as compared to the original paper. First, a least squares GAN framework [49] is implemented, which substitutes the loss function in (1), and second, a deeper version of the U-Net generator is included, yielding a network with 54.41M parameters. A minimal change was made for training the pix2pix model, consisting of deactivating the data augmentation per batch in order to properly compare to the other semantic segmentation models.

#### 4) EVALUATION METRICS

The metrics used to evaluate the different insulator string segmentation networks are taken from [34], and are summarized here for a better understanding of the results obtained:

- Pixel accuracy: $\sum_i n_{ii} / \sum_i t_i$
- Mean accuracy: $(1/n_{cl}) \sum_i (n_{ii}/t_i)$
- Mean Intersection over Union (Mean IoU): $(1/n_{cl}) \sum_i n_{ii} / \left(t_i + \sum_j n_{ij} - n_{ii}\right)$
- Frequency weighted IoU (f.w. IoU): $\left(\sum_k t_k\right)^{-1} \sum_i t_i n_{ii} / \left(t_i + \sum_j n_{ij} - n_{ii}\right)$

where $n_{cl}$ represents the number of classes included in the ground truth segmentation, $n_{ij}$ is the number of pixels of class $i$ predicted to belong to class $j$, and $t_i$ is the total number of pixels of class $i$ in the ground truth segmentation. In this work, we consider $n_{cl} = 3$, including the *disc*, *cap* and *background* classes.

#### 5) RESULTS IN INSULATOR STRING SEGMENTATION

The results obtained during the evaluation of the insulator string segmentation networks are shown in Table 1. Several trainings (from 5 to 10) have been conducted for each network in order to select the most appropriate set of hyperparameters. The results obtained by the best models are shown in Table 1.

As can be noticed in Table 1, the networks that integrate into their architecture the topology proposed in this paper (i.e., Up-Net_vgg16 and GAN_Up-Net), are those that obtain the best performance in the task of insulator string segmentation when training on the augmented dataset. Table 1 also reveals the huge impact in performance produced by the application of data augmentation and transfer learning techniques. Data augmentation clearly benefits networks without pre-training (e.g. FCN-8s and SegNet where the mean accuracy is improved by 4.7% and 3.6%, and the mean IoU is improved by 6.3% and 4.2%, respectively), and is remarkable in the networks GAN_U-Net and GAN_Up-Net in which the mean accuracy is improved by 4.3% and 3.6%, and the mean IoU is improved by 7.3% and 5.2%, respectively. It is interesting to note that although in most of the networks the increase in performance is clear after the application of data augmentation, in others the improvement is very small, as in the case of U-Net_vgg16 and pix2pix.

On the other hand, the application of transfer learning produces the most significant improvement in performance across all the networks considered. As shown in Table 1, the mean accuracy of the networks FCN-8s, SegNet, and U-Net is increased by 7%, 10.5% and 6.2%, and the mean IoU is improved by 10.8%, 12.3% and 8.6% respectively when the networks are pre-trained using the weights of the VGG16 previously trained on ImageNet.

Qualitative segmentation results on a subset of images from the test set are illustrated in Fig. 6 and 7. This subset of images is quite representative as it encompasses insulators made of different materials (i.e., glass and ceramic), different points of view, backgrounds, presence of shadows and reflections on the surface of the insulator, etc. As shown in Fig. 6 and 7, most of the networks provide accurate semantic segmentation results, being the most challenging images the second (i.e., ceramic insulator with shadows) and sixth (i.e., a tilted insulator with shadows). Fig. 7 shows the results

[4]https://github.com/junyanz/pytorch-CycleGAN-and-pix2pix

**TABLE 1.** Semantic segmentation results on the test set composed of 240 images, resized to 256 × 256 pixels, when training on the original and augmented datasets. Metrics are expressed in percentage. Networks whose name ends in "vgg16" are pre-trained with the weights of the VGG16 network trained on ImageNet. Best results are indicated in bold.

| Model | Original dataset | | | | Augmented dataset | | | | Inference time (ms/image) |
|---|---|---|---|---|---|---|---|---|---|
| | Pixel Acc. | Mean Acc. | Mean IoU | f.w. IoU | Pixel Acc. | Mean Acc. | Mean IoU | f.w. IoU | |
| FCN-8s | 94.60 | 80.51 | 72.65 | 90.64 | 96.04 | 85.19 | 78.98 | 92.81 | 52.07 |
| FCN-8s_vgg16 | 96.84 | 87.56 | 83.47 | 94.08 | 97.81 | 90.36 | 86.87 | 95.84 | 51.75 |
| SegNet | 94.52 | 81.59 | 76.26 | 90.19 | 95.49 | 85.23 | 80.48 | 91.89 | 29.12 |
| SegNet_vgg16 | 97.99 | 92.09 | 88.53 | 96.21 | 98.07 | 93.64 | 90.19 | 96.41 | 29.37 |
| CAE | 96.34 | 85.82 | 79.98 | 93.33 | 96.83 | 88.28 | 83.07 | 94.23 | 14.42 |
| CAE_vgg16 | 97.04 | 90.02 | 83.95 | 94.57 | 97.45 | 91.07 | 85.85 | 95.30 | 14.41 |
| U-Net | 96.25 | 87.98 | 81.80 | 93.26 | 96.77 | 89.24 | 85.44 | 94.00 | 14.91 |
| U-Net_vgg16 | 98.19 | **94.14** | **90.41** | 96.58 | 98.20 | 93.98 | 90.94 | 96.60 | 15.67 |
| Up-Net_vgg16 (ours) | **98.21** | 93.65 | 90.31 | **96.60** | **98.58** | **95.16** | **92.06** | **97.30** | 19.28 |
| ISNet | 95.91 | 87.67 | 79.10 | 92.73 | 97.35 | 91.56 | 85.77 | 95.11 | **4.52** |
| pix2pix | 97.32 | 90.00 | 84.71 | 95.05 | 97.20 | 91.20 | 84.60 | 95.00 | 10.95 |
| GAN_U-Net | 96.27 | 89.45 | 81.42 | 93.38 | 97.99 | 93.78 | 88.70 | 96.32 | 10.81 |
| GAN_Up-Net (ours) | 97.01 | 91.05 | 84.84 | 94.60 | 98.24 | 94.60 | 90.03 | 96.73 | 14.72 |

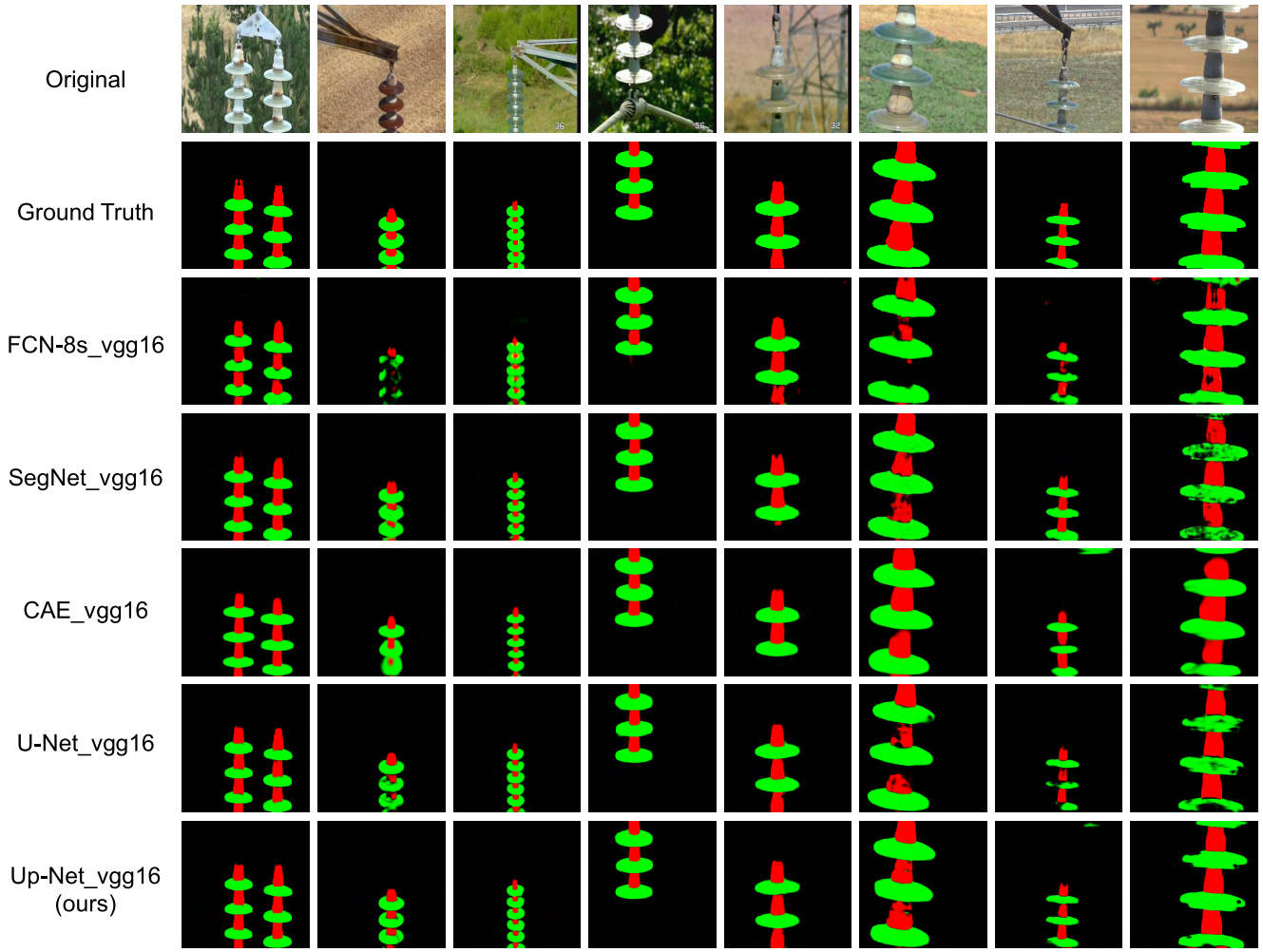


**FIGURE 6.** Semantic segmentation results on a subset of images from the test set obtained by the networks trained without a conditional GAN framework.

of the networks trained within a conditional GAN framework, where it can be clearly appreciated how these networks tend to produce more noisy segmentation results, in which false positives, related to the structure of the electric tower, appear for most of the networks.

Taking into account the results presented in Table 1, the network Up-Net_vgg16 is selected as the most appropriate for the insulator string segmentation task and is the one used for the evaluation of the complete system presented in Section IV-E. For this purpose, and in order to be robust

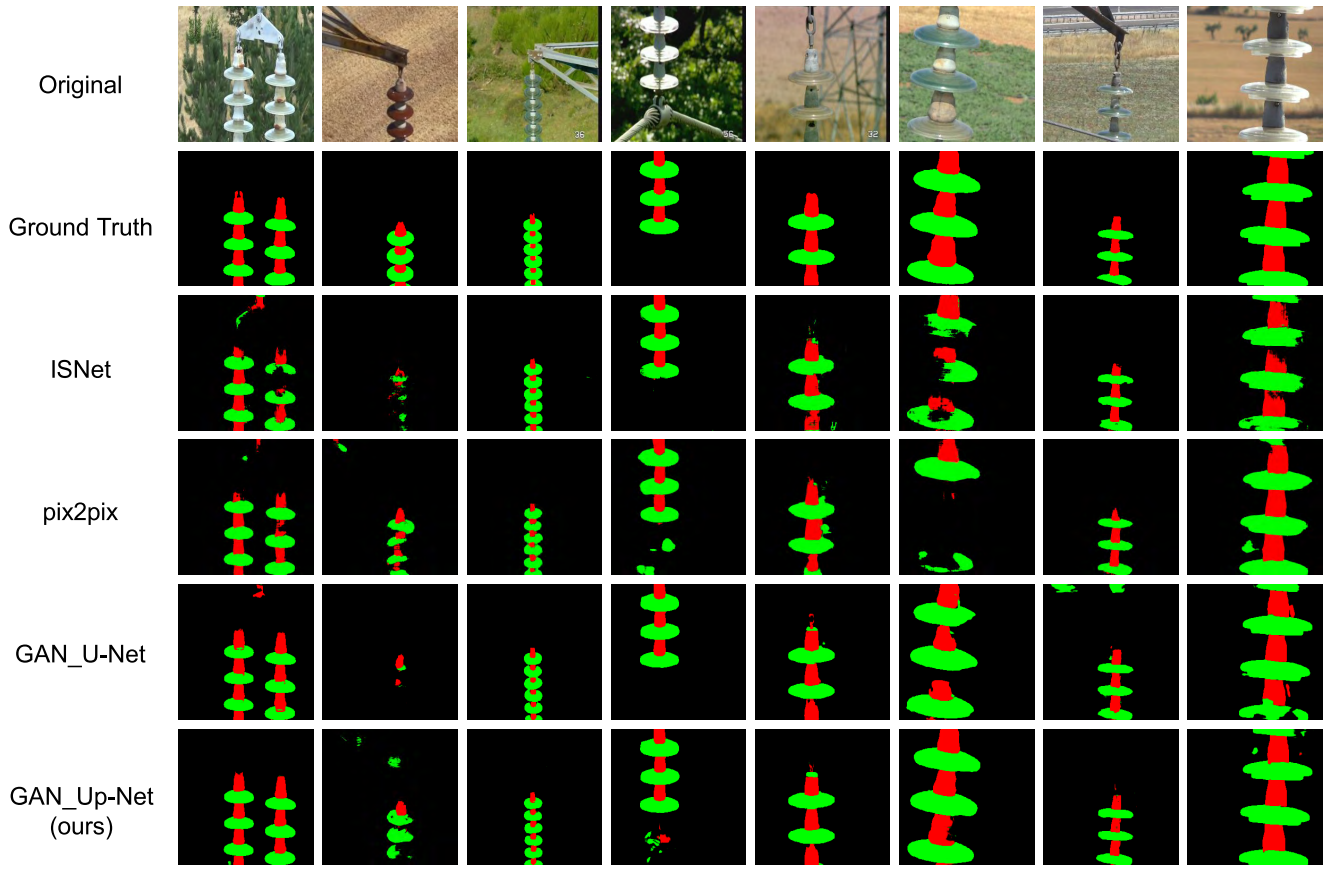


**FIGURE 7.** Semantic segmentation results on a subset of images from the test set obtained by the networks trained within a conditional GAN framework.

to different insulator string orientations, this network is re-trained on a second augmented dataset. This new augmentation involves the application of two random rotations between the range $[30°, 60°]$ and $[-60°, -30°]$ respectively for each of the 640 images of the augmented dataset, yielding a total of 1920 images. We have observed that training the Up-Net_vgg16 network using this second augmented dataset does not worsen the ability to segment vertical insulator strings and provides the appropriate capabilities for segmenting insulator strings in various orientations.

### C. DIAGNOSIS OF ABSENT DISC UNITS

In this work, we address the problem of diagnosing the absence of disc insulator units within an insulator string as a classification problem, where the objective is to identify the occurrence of this fault at frame-level. For this purpose, the network in charge of diagnosing this defect takes as input the output mask generated by the insulator string segmentation component and outputs the probability of fault occurrence.

#### 1) DATASET

Owing to the relatively small amount of defective images in our dataset, we have left the real images containing defective samples only for evaluation purposes and create a synthetic dataset for training and validating the component responsible for diagnosing the absence of disc insulator units. The dataset is composed of 1200 negative images (non-defective) and 1200 positive images (i.e., with one or more absent disc units) of $256 \times 256$ pixels (see Fig. 8 for an example). Inside this dataset, the negative samples have been obtained by doing inference over some of the networks presented in Table 1 for the RGB training images used for semantic segmentation. We use several networks at different epochs of the training process for obtaining a heterogeneous dataset of negative samples. In addition, 10% of these images have been manually altered by removing some pixels of the *disc* and *cap* classes in order to simulate errors in the semantic segmentation stage (see the last image in Fig. 8a). Using the negative samples, we generate the dataset of positive samples, where 60% of the 1200 images have been automatically generated using computer vision techniques by alternatively removing some of the disc units within an insulator string. In addition, random noise is added to some of these images. The remaining 40% of the positive samples have been manually generated by removing the pixels belonging to some of the caps and disc units within the insulator string.

The evaluation of the different CNNs designed for diagnosing the absence of disc insulator units is conducted on a test set composed of 400 images, where the samples are

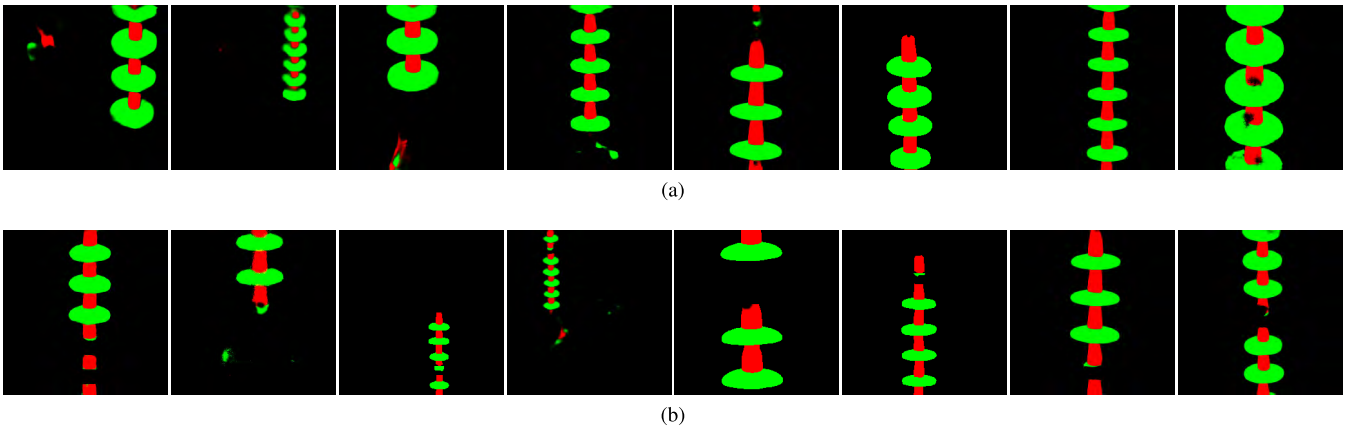

**FIGURE 8.** Examples of images used for training the component responsible for diagnosing the absence of disc insulator units. Images containing false positives and false negatives are included in the training set to provide robustness to possible errors that might appear in the insulator segmentation stage. a) Negative samples (non-defective images). b) Positive samples (defective images) synthetically generated to reproduce different defect configurations relative to the absence of disc units.

**TABLE 2.** CNN architectures designed for diagnosing the absence of disc insulator units. *k* represents the kernel size, and *s* is the stride. The number before the @ represents the number of kernels in the corresponding convolutional layer. "FC" refers to the fully-connected part of the CNN.

| Layer | CNN1 | CNN2 | CNN3 |
|---|---|---|---|
| conv1 | $32@k \times k \times 3, s = 2$ | $32@k \times k \times 3, s = 2$ | $32@k \times k \times 3, s = 2$ |
| pool1 | $3 \times 3, s = 2$ | $3 \times 3, s = 2$ | $3 \times 3, s = 2$ |
| conv2 | $64@k \times k \times 32, s = 1$ | $64@k \times k \times 32, s = 2$<br>$64@k \times k \times 64, s = 1$ | $64@k \times k \times 32, s = 1$<br>$64@k \times k \times 64, s = 1$ |
| pool2 | $2 \times 2, s = 2$ | $2 \times 2, s = 2$ | $2 \times 2, s = 2$ |
| conv3 | $128@k \times k \times 64, s = 1$ | $64@k \times k \times 64, s = 1$<br>$64@k \times k \times 64, s = 1$ | $128@k \times k \times 64, s = 1$<br>$128@k \times k \times 128, s = 1$ |
| pool3 | $2 \times 2, s = 2$ | $2 \times 2, s = 2$ | $2 \times 2, s = 2$ |
| FC | $64 \times 1$ | $64 \times 1$ | $64 \times 1$ |

equally distributed between the negative and positive classes (i.e., 200 images for each class).

### 2) NEURAL NETWORK VARIANTS

We have experimented with different CNN architectures, conducting a detailed analysis of the number and size of the convolutional layers and the size of the convolution kernels. Three main CNN architectures have been evaluated, whose architecture is depicted in Table 2. As shown in Table 2, CNN1 is composed of 3 convolutional and 3 pooling layers. CNN2 architecture consists of a 10-layer CNN, in which we add one more convolutional layer in the second and third convolutional blocks (conv2 and conv3). Another representative characteristic of CNN2 is that we increase the stride of the first convolutional layer in the second convolutional block (conv2) with the purpose of reducing the number of parameters in the network. Finally, CNN3 has a similar architecture as compared to CNN2, where the stride of the first convolutional layer in the second convolutional block is restored to 1, and we double the number of kernels in the third convolutional block (conv3). All the pooling layers are identical in the three CNN architectures and consist of a $3 \times 3$ max-pooling operation with a stride of 2 in the first pooling layer, and a $2 \times 2$ max-pooling operation with a stride of 2 in the second and third pooling layers. The fully-connected network at the end of each CNN is identical for the three architectures and consists of one hidden layer composed of 64 units followed by a single-unit output layer.

### 3) TRAINING METHODOLOGY

In this work, we consider the problem of diagnosing the absence of disc insulator units as a binary classification problem, in which the current frame analyzed can be defective, which means that one or several disc units are absent from the insulator string, or correct, which implies that the insulator string is in good condition. As a binary classification problem, we use the *binary cross-entropy* loss function, which is minimized during a training process of 60 epochs using the Adam optimizer with a base learning rate of $2^{-4}$, and the rest of hyperparameters as described in the original work [45]. The minibatch size utilized in each update of the optimization process is 128 images.

In each experiment, we perform a randomly stratified split of the training dataset (i.e., 2400 negative and positive samples) into training and validation sets, where the percentage of samples in each set is 80% and 20%, respectively.

### 4) EVALUATION METRICS

For the sake of obtaining an appropriate comparison of the CNN architectures considered, we use standard metrics widely applied in the research community for evaluating binary classifiers. These metrics are summarized here for a

**TABLE 3.** Mean and standard deviation obtained by averaging the results of 5 different models (5 trainings) for each CNN configuration while diagnosing the absence of disc insulator units on a test set of 400 images. The number of parameters of the network is expressed in millions. Best results are indicated in bold.

| CNN arch. | kernel size | Accuracy (%) | Precision (%) | Recall (%) | $F_1 score$ (%) | ROC-AUC (%) | #param (M) |
|---|---|---|---|---|---|---|---|
| CNN1 | $3 \times 3$ | $90.75 \pm 0.91$ | $91.29 \pm 1.03$ | $90.13 \pm 2.56$ | $90.68 \pm 1.03$ | $97.07 \pm 0.36$ | 1.94 |
| | $5 \times 5$ | $94.63 \pm 0.32$ | $93.76 \pm 0.56$ | $95.75 \pm 0.65$ | $94.74 \pm 0.24$ | $98.59 \pm 0.19$ | 2.10 |
| | $7 \times 7$ | $96.38 \pm 0.14$ | $96.05 \pm 1.01$ | $96.75 \pm 0.96$ | $96.39 \pm 0.13$ | $99.01 \pm 0.26$ | 2.35 |
| CNN2 | $3 \times 3$ | $94.44 \pm 0.85$ | $92.09 \pm 1.22$ | $97.38 \pm 0.75$ | $94.66 \pm 0.85$ | $98.72 \pm 0.20$ | 0.39 |
| | $5 \times 5$ | $96.72 \pm 0.70$ | $95.58 \pm 1.45$ | $98.06 \pm 0.78$ | $96.80 \pm 0.71$ | $99.40 \pm 0.26$ | 0.56 |
| | $7 \times 7$ | $97.85 \pm 0.45$ | $97.44 \pm 1.02$ | $98.30 \pm 0.84$ | $97.86 \pm 0.44$ | $99.36 \pm 0.07$ | 0.91 |
| CNN3 | $3 \times 3$ | $95.06 \pm 0.38$ | $95.12 \pm 0.26$ | $95.00 \pm 0.58$ | $95.06 \pm 0.38$ | $98.65 \pm 0.05$ | 2.12 |
| | $5 \times 5$ | $96.84 \pm 0.35$ | $96.54 \pm 1.01$ | $97.19 \pm 0.59$ | $96.86 \pm 0.33$ | $99.15 \pm 0.25$ | 2.61 |
| | $7 \times 7$ | $\mathbf{98.88 \pm 0.48}$ | $\mathbf{98.40 \pm 0.72}$ | $\mathbf{99.38 \pm 0.95}$ | $\mathbf{98.88 \pm 0.48}$ | $\mathbf{99.89 \pm 0.10}$ | 3.35 |

better understanding of the results:

- Accuracy: $(tp + tn)$ / $(tp + tn + fp + fn)$
- Precision: $tp/\ (tp + fp)$
- Recall: $tp/\ (tp + fn)$
- $F_1 score$: $2tp/\ (2tp + fp + fn)$
- Area under the Receiver Operating Characteristic (ROC) curve (ROC-AUC): this curve depicts the True Positive Rate (TPR) versus the False Positive Rate (FPR), showing how the correctly classified positive examples (TPR) vary with respect to the incorrectly classified negative examples (FPR).

where $tp$, $tn$, $fp$, and $fn$ stand for true positive, true negative, false positive, and false negative, respectively.

All the metrics utilized for the evaluation of the different models have been computed using *scikit-learn* [50].

#### 5) RESULTS IN THE DIAGNOSIS OF ABSENT DISCS

The results obtained after the evaluation of the different CNN architectures considered are depicted in Table 3, which shows the mean and standard deviation after conducting 5 trainings for each of the CNN configurations. According to these results, the network CNN3 with a kernel size of $7 \times 7$ outperforms the other CNN configurations, having an average ROC-AUC of 99.89%. The results presented in Table 3 also reveal the considerable effect of the number of convolutional layers and the size of the convolution kernels in the performance of the different CNN architectures. The former is evidenced when moving from CNN1 to CNN3, in which the $F_1 score$ is improved by 4.4%, 1.9%, and 2.5% using kernel sizes of $3 \times 3$, $5 \times 5$, and $7 \times 7$ respectively. The effect of the size of the convolution kernels is noted when increasing the size from $3 \times 3$ to $7 \times 7$ which leads to an increment in the $F_1 score$ of 5.7%, 3.2%, and 3.8% for the CNN1, CNN2, and CNN3 architectures respectively. This effect is also evidenced when analyzing the ROC-AUC, which experiments an increment of 1.9%, 0.6%, and 1.2% for the CNN1, CNN2, and CNN3 respectively. These results are consistent with the intuition that larger receptive fields can benefit the diagnosis of this type of defect.

In order to provide better insight into each CNN architecture studied in this section, in Fig. 9 we show the ROC curves generated by the best models selected from the 5 training experiments conducted on each CNN configuration. The results presented in Fig. 9 provide additional information about the capabilities of each CNN architecture, showing the high AUC obtained for the CNN2 and CNN3 architectures with $5 \times 5$ and $7 \times 7$ kernel sizes. Taking into account the results presented in Table 3 and Fig. 9, CNN3_$7 \times 7$ is selected as the network utilized in the final system for diagnosing the absence of disc insulator units.

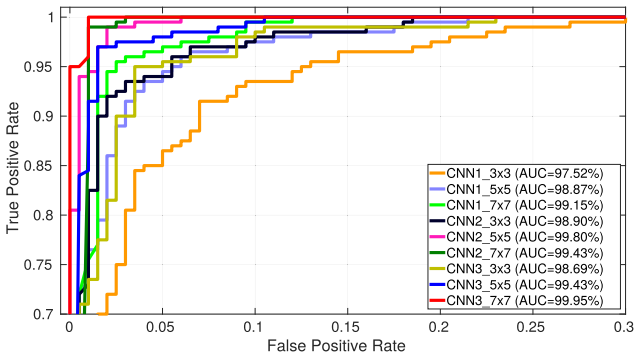


**FIGURE 9.** ROC curves generated by the best model for each CNN configuration tested for the disc absence defect diagnosis (best seen in color).

Some of the prediction results generated by the CNN3_$7 \times 7$ model on a subset of representative images from the test set are shown in Fig. 10. In this figure, we also show the real RGB images (see Fig. 10a) from which the masks are computed using the Up-Net_vgg16 model (see Fig. 10b). Furthermore, Fig. 10b provides a complementary validation of the performance obtained by the insulator string segmentation component (Up-Net_vgg16).

### D. DIAGNOSIS OF DAMAGED DISC UNITS

In this section, we describe the experiments that have been conducted in order to select the most appropriate network for the task of diagnosing the presence of damaged discs within the insulator string. As explained in Section III-C.2, we address the problem of diagnosing this type of defect by designing a siamese architecture, which receives as input a pair of disc images and outputs a probability which encodes the dissimilarity between the input images, where 0 represents that the two input images are the "same" (non-defective) and 1 when they are "different" (defective).

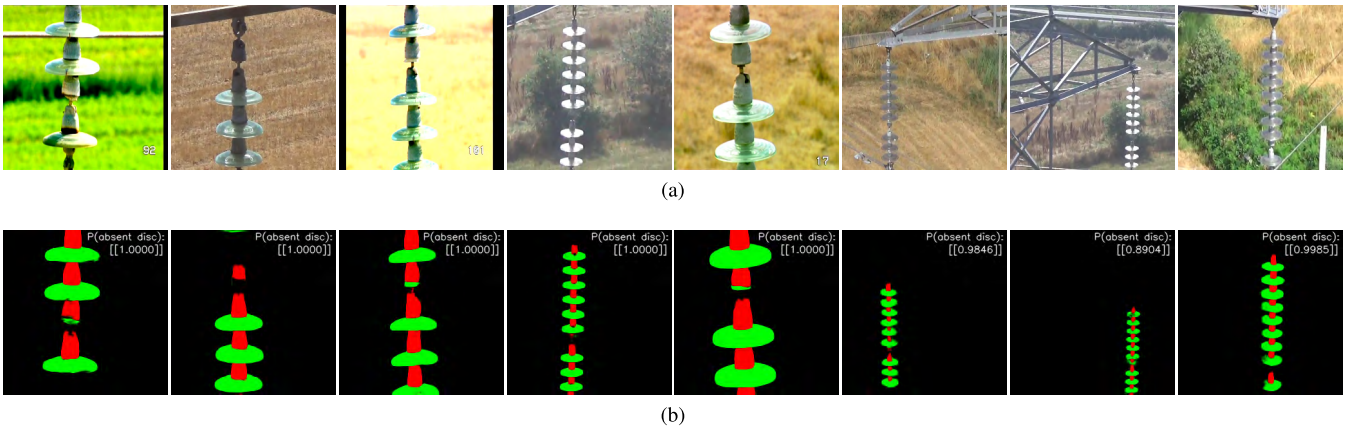


**FIGURE 10.** Examples of images from the test set used for evaluating the component responsible for diagnosing the absence of disc units. a) Original RGB images from which the masks for evaluation purposes are extracted. b) Masks generated by the insulator string segmentation component (Up-Net_vgg16) and probabilities predicted by the selected model for diagnosing the absence of disc units (CNN3_7 × 7). As shown in the two right-most images in Fig. 10b, the proposed CNN3_7 × 7 is able to correctly handle false positives and false negatives that might appear in the insulator segmentation stage.

#### 1) DATASET

In order to build the dataset for diagnosing damaged discs, we leverage from the annotations performed for training the insulator string segmentation networks (see Section IV-B.1). Thus, from the 160 images annotated for semantic segmentation purposes, we extract all the discs within the insulator strings of these images, yielding 579 and 18 non-damaged and damaged discs, respectively. After that, we generate all possible combinations of disc pairs within an insulator string for the corresponding image. This process allows obtaining a total of 802 images of negative (non-damaged) pairs and 95 positive (damaged) pairs (see Fig. Fig. 11a and 11b). In order to deal with this highly imbalanced dataset, several strategies have been applied and evaluated.

- Synthetic samples generation: we manually generate synthetic defects within the disc surface by mimicking the appearance of real defective discs. For this purpose, 125 synthetic damaged disc images were created by placing synthetic defects on top of healthy insulator discs. After this, the 125 images generated were combined with non-defective disc images obtaining a total of 762 images of defective pairs (see Fig. 11c), which are merged with the 95 real defective pairs in order to obtain a total of 857 defective pairs for training purposes.
- Training using only real or only synthetic data: the second strategy studied in this section consists of training the SCNNs taking into account only real or only synthetic data. In the first case, the training is performed using the dataset composed of 802 non-damaged pairs and 95 damaged pairs. This strategy implies the modification of the loss function used for training the SCNNs for classification (i.e., binary cross-entropy loss) in order to tackle the problems derived from a highly imbalanced dataset. This new loss function is explained in Section IV-D.3. In the second case, the training is performed using the dataset composed of 802 non-damaged pairs and 762 damaged pairs.

In order to evaluate all the networks considered for diagnosing damaged discs, a test set of 193 images provided by an electric company has been used, which is composed of 133 non-damaged and 60 damaged discs. These images are combined in pairs as in the training set, yielding 245 image pairs. As in the case of the training set, this test set is also imbalanced, being composed of 162 images of non-damaged disc pairs and 83 damaged pairs, containing only real defective and non-defective images.

#### 2) NEURAL NETWORK VARIANTS

Within the SCNN framework, we experiment with different variants of the main architecture presented in Fig. 4. The main analysis is based on the evaluation of different distance layers. In this work, we evaluate the performance of the SCNNs using the L1 and the Chi-squared distances. The latter, once connected to the final output layer of the SCNN, computes the similarity between hidden states using the expression: $d(h_1, h_2) = \sigma(\sum_i \omega_i (h_1^{(i)} - h_2^{(i)})^2 / (h_1^{(i)} + h_2^{(i)} + \epsilon))$, where $\sigma$ is the sigmoid activation function, $\omega_i$ are the weights learned in the output layer, and $\epsilon = 10^{-6}$ is introduced to obtain numeric stability. Within the SCNN framework, three main neural network variants have been designed, depending on the architecture of the twin CNNs within the SCNN. The first SCNN architecture, named SCNN-L1/$\chi^2$, has fewer parameters as compared to other SCNN variants and consists of 3 convolutional blocks (conv1 to conv3 in Fig. 4) of $1 \times 32$, $2 \times 64$, and $2 \times 64$ respectively. The second SCNN architecture studied, named SCNN-c3-L1/$\chi^2$, follows the VGG16 architecture until the third convolutional block (conv3), consisting of $2 \times 64$, $2 \times 128$, and $3 \times 256$ convolutional blocks. The final SCNN architecture analyzed, named SCNN-c4-L1/$\chi^2$, introduces a fourth $1 \times 1$ convolutional block and is explained in detail in Section III-C.2. All the CNNs within the different SCNN architectures finish with a fully-connected part composed of 2 hidden layers with 64

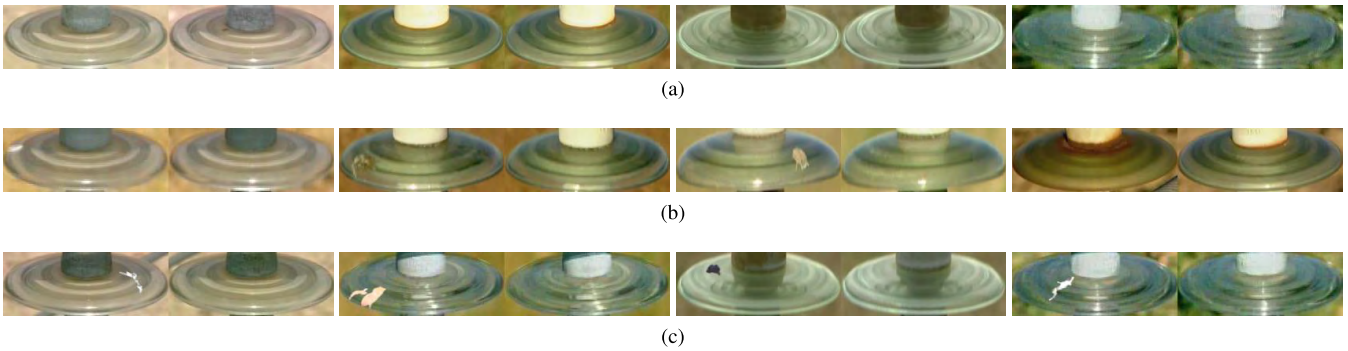


**FIGURE 11.** Examples of images used for training the proposed SCNN model. a) Real examples of non-damaged pairs. b) Real examples of damaged pairs. c) Examples of damaged pairs which contain synthetic defects. Images of disc units showing a small degree of pollution due to bird excrements and with rust on their surface are used for training the proposed SCNN.

units each. Finally, we analyze the effect of transfer learning in the SCNN-c3-L1/$\chi^2$ and the SCNN-c4-L1/$\chi^2$ network variants, which are evaluated with and without pre-training using the VGG16 network previously trained on ImageNet.

Additionally, in order to compare the performance of the proposed SCNN framework to traditional classification methods such as CNNs, we evaluate the best performing SCNN model (SCNN-c4-L1_vgg16) against its analogous CNN (CNN-c4_vgg16). The latter is built by extracting one of the twin CNNs from the SCNN-c4-L1_vgg16 and connecting it directly to the output unit (i.e., omitting the L1 layer), yielding the CNN-c4_vgg16.

#### 3) TRAINING METHODOLOGY

The SCNN training procedure is almost identical to the one described previously in Section IV-C.3 for the CNNs used for diagnosing the absence of disc insulator units. We also model the diagnosis of damaged discs as a binary classification problem where the pair of images introduced to the siamese network ($x_1$ and $x_2$) can be classified as damaged or non-damaged. Thus, we use a binary cross-entropy loss function, which is modified according to (3) for training using an imbalanced dataset. In all cases, we minimize the corresponding loss using the Adam optimizer during a training process of 60 epochs, using a minibatch size of 128 images, and considering a base learning rate of $10^{-4}$.

$$\mathcal{L}(x_1^{(i)}, x_2^{(i)}) = -y(x_1^{(i)}, x_2^{(i)}) \log \hat{y}(x_1^{(i)}, x_2^{(i)}) - \alpha \left(1 - y(x_1^{(i)}, x_2^{(i)})\right) \log \left(1 - \hat{y}(x_1^{(i)}, x_2^{(i)})\right) \quad (3)$$

where $y$ represents the ground truth label which is 0 whenever $x_1$ and $x_2$ are the "same" (non-damaged) and 1 otherwise, $\hat{y}$ is the output prediction of the SCNN, and $\alpha = 1$ for all the network variants except when training the SCNN-c4-L1_vgg16 model using only real damaged disc defects, where $\alpha = 0.002$ has been empirically found.

The CNN-c4_vgg16 used in the comparative analysis has been trained using the 579 and 143 (18 real, plus 125 synthetic) images of non-damaged and damaged discs, respectively. The set of damaged discs is augmented by a horizontal flip procedure, doubling the number of defective images. Regarding the training methodology, the CNN-c4_vgg16 is trained using a binary cross-entropy loss, which is minimized using the same optimization method with the same hyperparameters as in the training process of the SCNN.

#### 4) EVALUATION METRICS

As we deal with a slightly imbalanced dataset for evaluation purposes (test set), in which the number of negative samples doubles the number of the positive ones, in this section we consider several metrics, computed using *scikit-learn*, that are usually utilized in the state-of-the-art for imbalanced datasets [51]:

- True Negative Rate (TNR): also called specificity is computed by: $tn/\,(tn + fp)$
- True Positive Rate (TPR): also called sensitivity or recall is given by: $tp/\,(tp + fn)$
- Geometric mean (G-mean) [52]: $\sqrt{TPR \times TNR}$
- Area under the ROC and Precision-Recall curves: In the specific case of this section, we also consider the Precision-Recall (PR) curve which analyzes the precision versus the recall (see Section IV-C.4). The PR curve can provide additional information in the case of imbalanced datasets as the precision is affected by the class imbalance owing to the computation of the false positives.

where $tp$, $tn$, $fp$, and $fn$ stand for true positive, true negative, false positive, and false negative, respectively.

#### 5) RESULTS IN THE DIAGNOSIS OF DAMAGED DISCS

The results obtained during the evaluation of the different SCNN models on the test set composed of 245 images are illustrated in Table 4. The results presented in Table 4 show the mean and standard deviation after conducting 10 experiments on each SCNN configuration and thresholding at 0.5 the output of the sigmoid activation function of the final layer for computing the TNR, TPR and G-mean metrics. The ROC-AUC and PR-AUC are computed in the standard way by considering several classification thresholds. As can be seen in Table 4, both distance layers (i.e., L1 and Chi-squared) provide high performance for diagnosing damaged

**TABLE 4.** Mean and standard deviation obtained by averaging the results of 10 different models (10 trainings) for each SCNN configuration while diagnosing damaged disc units on a test set of 245 images. Networks whose name ends in "vgg16" are pre-trained with the weights of the VGG16 network trained on ImageNet. The number of parameters of the networks is expressed in millions. Best results are indicated in bold.

| SCNN conf. | TNR (%) | TPR (%) | G-mean (%) | ROC-AUC (%) | PR-AUC (%) | #param (M) |
|---|---|---|---|---|---|---|
| SCNN-L1 | $86.67 \pm 4.95$ | $92.77 \pm 2.72$ | $89.61 \pm 2.41$ | $93.58 \pm 1.05$ | $92.53 \pm 1.22$ | 1.29 |
| SCNN-$\chi^2$ | $87.53 \pm 4.20$ | $91.93 \pm 3.01$ | $89.64 \pm 1.44$ | $92.08 \pm 1.85$ | $91.50 \pm 2.11$ | 1.29 |
| SCNN-c3-L1 | $88.64 \pm 4.10$ | $89.52 \pm 4.33$ | $89.02 \pm 2.69$ | $91.48 \pm 3.57$ | $91.27 \pm 3.06$ | 6.00 |
| SCNN-c3-$\chi^2$ | $89.32 \pm 5.13$ | $87.47 \pm 2.74$ | $88.32 \pm 2.03$ | $89.71 \pm 2.32$ | $89.62 \pm 2.53$ | 6.00 |
| SCNN-c3-L1_vgg16 | $86.23 \pm 4.87$ | $93.14 \pm 3.17$ | $89.54 \pm 1.67$ | $92.41 \pm 1.24$ | $88.64 \pm 1.53$ | 6.00 |
| SCNN-c3-$\chi^2$_vgg16 | $81.98 \pm 2.17$ | $93.62 \pm 2.34$ | $87.59 \pm 1.64$ | $92.50 \pm 2.22$ | $90.57 \pm 2.65$ | 6.00 |
| SCNN-c4-L1 | $88.83 \pm 5.75$ | $92.41 \pm 1.61$ | $90.55 \pm 2.80$ | $93.92 \pm 1.24$ | $93.98 \pm 1.28$ | 2.82 |
| SCNN-c4-$\chi^2$ | $82.29 \pm 5.35$ | $92.17 \pm 2.97$ | $87.00 \pm 2.04$ | $93.11 \pm 1.80$ | $92.88 \pm 2.22$ | 2.82 |
| SCNN-c4-L1_vgg16 | $\mathbf{90.49 \pm 4.09}$ | $\mathbf{94.10 \pm 1.55}$ | $\mathbf{92.25 \pm 1.93}$ | $\mathbf{96.59 \pm 1.60}$ | $\mathbf{96.36 \pm 1.11}$ | 2.82 |
| SCNN-c4-$\chi^2$_vgg16 | $84.32 \pm 3.26$ | $93.50 \pm 2.14$ | $88.76 \pm 1.30$ | $94.89 \pm 0.83$ | $94.45 \pm 1.00$ | 2.82 |
| SCNN-c4-L1_vgg16_SYN | $83.33 \pm 4.83$ | $89.88 \pm 3.69$ | $86.47 \pm 2.03$ | $92.37 \pm 2.68$ | $89.89 \pm 3.54$ | 2.82 |
| SCNN-c4-L1_vgg16_REAL | $87.41 \pm 5.41$ | $60.12 \pm 5.66$ | $72.36 \pm 3.30$ | $71.90 \pm 3.99$ | $73.67 \pm 4.30$ | 2.82 |

disc units, being the performance of the networks using the L1 layer greater in most of the configurations analyzed. This can be clearly observed when analyzing the configurations SCNN-L1/$\chi^2$, SCNN-c3-L1/$\chi^2$, SCNN-c4-L1/$\chi^2$, and SCNN-c4-L1/$\chi^2$_vgg16, where the relative improvement considering the ROC-AUC and PR-AUC scores respectively is 1.5% and 1.0% for the SCNN-L1/$\chi^2$, 1.8% and 1.7% for the SCNN-c3-L1/$\chi^2$, 0.8% and 1.1% for the SCNN-c4-L1/$\chi^2$, and 1.7% and 1.9% for the SCNN-c4-L1/$\chi^2$_vgg16. Another interesting result derived from the results presented in Table 4, is the considerable influence of transfer learning. In all the architectures analyzed, the usage of pre-trained weights from the VGG16 model produced an important increase of the TPR. This can be noticed in Table 4 when analyzing the results of the configurations SCNN-c3-L1 and SCNN-c4-L1, where the usage of their pre-trained counterparts produced an increment of the TPR of 3.6% and 1.7% respectively. This effect is also appreciated in the case of the SCNN-c3-$\chi^2$ and SCNN-c4-$\chi^2$ variants, where the increment in the TPR when using a pre-trained model is 6.2% and 1.3% respectively. Finally, it should be remarked the effect of adding the $1 \times 1$ convolutional layer (c4 configurations). According to the results presented in Table 4, this new architecture has beneficial effects when considering the L1 variant. The transition from SCNN-c3-L1 to SCNN-c4-L1 produces an increment of 1.5%, 2.4%, and 2.7% when analyzing the G-mean, ROC-AUC, and PR-AUC respectively. Similarly, The transition from SCNN-c3-L1_vgg16 to SCNN-c4-L1_vgg16 causes an increment in G-mean, ROC-AUC, and PR-AUC of 2.7%, 4.2%, and 7.7% respectively.

The two final rows of Table 4 show the results obtained when training the SCNN-c4-L1_vgg16 model using only synthetic or only real defects. Interestingly, training the SCNN-c4-L1_vgg16 model using only synthetic defects provides a G-mean score of 86.47% which is very close to the performance of other variants of the model when trained on the complete dataset (i.e., real plus synthetic defects). On the other hand, training the SCNN-c4-L1_vgg16 model using only real defects allows obtaining a decent performance (G-mean score of 72.36%) taking into account that only 18 damaged discs images (95 image pairs) containing real defects were available for training purposes.

Fig. 12 shows the ROC and PR curves generated by the best models among the ten trainings conducted on each SCNN configuration. In this case, the best model is selected by taking the one with the highest ROC-AUC score. According to the results presented in Table 4 and Fig. 12, the SCNN-c4-L1_vgg16 architecture exhibits the best performance in the diagnosis of damaged discs, and is further evaluated in Table 5 against its CNN counterpart. In order to properly compare both models, since they take different types of inputs (i.e., the CNN takes as input a single image, whereas the SCNN takes as input a pair of images), for each image evaluated on the CNN we extract the corresponding pairs in which this image appears from the SCNN's test set. Then, the probabilities generated by the SCNN are averaged using the corresponding pairs and compared to the probability predicted by the CNN. The result of this comparison is summarized in Table 5, which shows how the SCNN-c4-L1_vgg16 outperforms the CNN-c4_vgg16 by 1.8% and 3.5% in ROC-AUC and PR-AUC, respectively. Furthermore, the SCNN-c4-L1_vgg16 exhibits an improvement of 3.6% when analyzing the TPR, which is of significant interest for fault diagnosis purposes.

Taking into account the results presented in Table 4 and 5, the SCNN-c4-L1_vgg16 architecture is selected as the final model for diagnosing damaged disc units in the final system. Fig. 13 shows the predictions generated by the selected SCNN-c4-L1_vgg16 in some representative examples extracted from the test set. As compared to the training set, the test set contains a more heterogeneous set of defective images in terms of the type of material and the type of defect. As shown in Fig. 13b, the test set encompasses images of disc units made of different material such as glass or ceramic, which can be burned, polluted (with a small to severe degree of pollution), or present rust on their surface. Furthermore, within each defect type, there exists considerable variability in the appearance of the defect on the surface of the disc, and even some discs can show more than one type of defect. The high performance of the proposed SCNN-c4-L1_vgg16 for

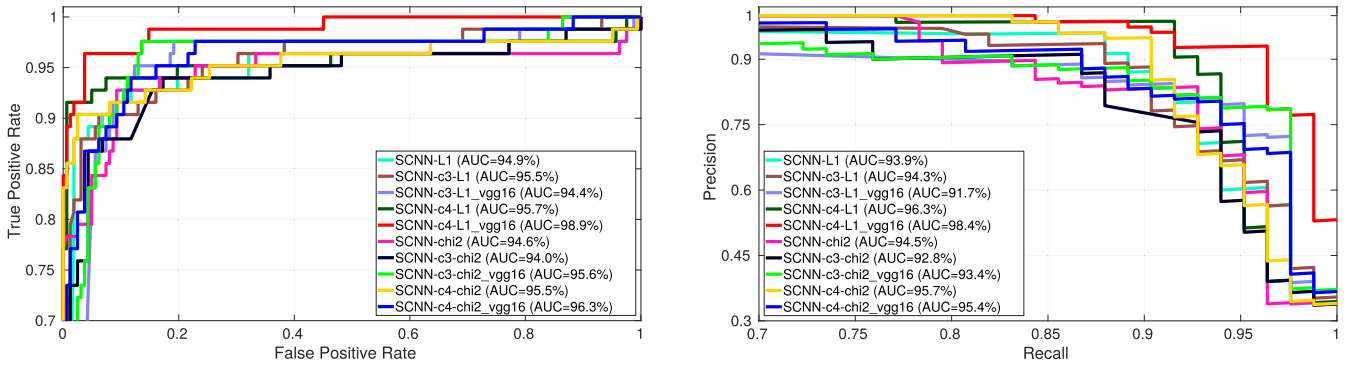


**FIGURE 12.** ROC and Precision-Recall curves generated by the best model for each SCNN configuration tested for the diagnosis of damaged disc units (best seen in color). a) ROC curve depicted using the best model out of 10 for each SCNN configuration. b) Precision-Recall curve generated by the selected model from Fig. 12a. Axes in both figures have been adjusted for visualization purposes.

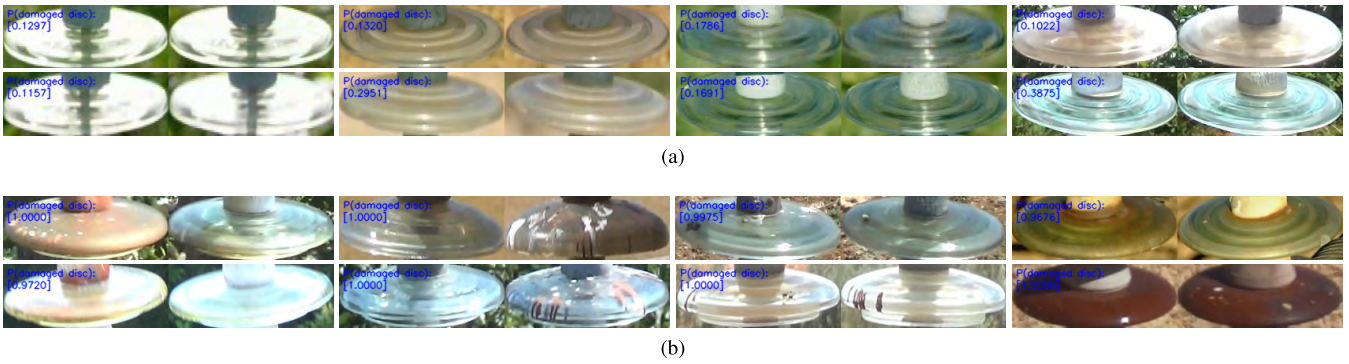


**FIGURE 13.** Examples of disc images captured in different inspections of high voltage power lines, extracted from the test set used for evaluating the SCNN models. a) Non-damaged examples and probabilities predicted by the SCNN-c4-L1_vgg16. b) Damaged examples and probabilities predicted by the SCNN-c4-L1_vgg16. Defects of varying nature such as discs burned, polluted (with different degrees of pollution), and rusted are efficiently detected by the proposed SCNN.

the diagnosis of the aforementioned types of defects reveals the good generalization capabilities of this model, which are confirmed in the results presented in Table 6 for several video sequences related to inspections of high voltage power transmission lines.

### E. SYSTEM EVALUATION RESULTS

The objective of this section is to perform a thorough evaluation of the complete system proposed in this work. To this aim, we integrate all the previously evaluated components and evaluate them on several video sequences captured during real aerial inspections of overhead power line infrastructure. Our proposed system has been evaluated in 5 video sequences corresponding to inspections of high voltage power lines (see Table 6), comprising a total of 5806 frames of $720 \times 576$ pixels, which have not been used in the training process of any of the components of the system. Sequences 1 to 4 contain high voltage insulator strings with different types of defects, sequence 1 being the most representative as it comprises defects of varying nature such as discs with bird excrements, with rust, etc. As stated previously in this document, we include all these types of defects within the "damaged" category for convenience. Sequence 5 contains four high voltage insulator strings, which do not present any defects. However, this sequence is of particular interest since two insulator strings are in a vertical position, and the other two are tilted, thus allowing the validation of the proposed system to changes in the orientation of the insulator string. All the sequences have been manually annotated where each frame has been classified as defective or not by a human inspector.

Regarding the absent disc defect, none of the five test sequences present any defect of this type. Thus, in order to further evaluate the disc absent diagnosis component, two additional sequences have been generated (see Table 7) where we simulate the absence of disc insulator units in each insulator string by alternatively removing disc units in the mask generated by the insulator string segmentation component. In order to automate the process for removing disc units within the masks, a computer program has been implemented, which takes as input a correct segmentation image, and uses image processing algorithms (contour extraction, contour approximation, and drawing functions) to alternatively remove disc units. The disc or discs to be removed for a given image are indicated to the program by manually clicking on the designated disc.

The results obtained for each diagnosis component on the video sequences of real inspections are shown in Table 6.

**TABLE 5.** Comparison between the best performing SCNN and its CNN counterpart. The results are obtained by averaging the performance of 10 different models (10 trainings) for each type of architecture using a test set of 193 disc images. Best results are indicated in bold.

| Model arch. | TNR (%) | TPR (%) | G-mean (%) | ROC-AUC (%) | PR-AUC (%) | #param (M) |
|---|---|---|---|---|---|---|
| CNN-c4_vgg16 | $\mathbf{86.69 \pm 2.38}$ | $86.83 \pm 2.77$ | $86.74 \pm 1.73$ | $91.84 \pm 0.83$ | $89.28 \pm 2.16$ | 2.82 |
| SCNN-c4-L1_vgg16 | $86.56 \pm 3.93$ | $\mathbf{90.42 \pm 3.18}$ | $\mathbf{88.42 \pm 1.64}$ | $\mathbf{93.67 \pm 2.03}$ | $\mathbf{92.75 \pm 1.70}$ | 2.82 |

**TABLE 6.** Results obtained in the evaluation of the proposed system using five video sequences of real aerial inspections. The processing time represents the time taken by the proposed system for processing each video sequence. The average diagnosis time considers each diagnosis component independently.

| Sequence | #frames | #insulator strings | processing time (s) | diagnosis component | #diagnosed frames | accuracy(%) | avg. diagnosis time (ms/frame) |
|---|---|---|---|---|---|---|---|
| Seq1 | 1700 | 3 | 50.57 | damaged disc | 565 | 96.64 | 15.93 |
| | | | | absent disc | 732 | 99.45 | 2.87 |
| Seq2 | 1376 | 3 | 41.90 | damaged disc | 386 | 93.26 | 19.09 |
| | | | | absent disc | 537 | 99.81 | 2.88 |
| Seq3 | 1213 | 3 | 38.16 | damaged disc | 400 | 98.75 | 18.13 |
| | | | | absent disc | 537 | 100 | 2.85 |
| Seq4 | 487 | 2 | 20.04 | damaged disc | 279 | 88.89 | 23.17 |
| | | | | absent disc | 348 | 100 | 2.87 |
| Seq5 | 1030 | 4 | 31.95 | damaged disc | 205 | 100 | 32.88 |
| | | | | absent disc | 205 | 97.56 | 2.88 |

**TABLE 7.** Results of the disc absent fault diagnosis on the simulated sequences shown in Fig. 15. The processing time represents the time taken by the proposed system for processing each video sequence, disabling the damaged disc diagnosis component.

| Sequence | #frames | #insulator strings | processing time (s) | diagnosis component | #diagnosed frames | accuracy(%) | avg. diagnosis time (ms/frame) |
|---|---|---|---|---|---|---|---|
| Seq1 | 1700 | 3 | 44.58 | absent disc | 804 | 99.25 | 2.81 |
| Seq2 | 1213 | 3 | 33.42 | absent disc | 600 | 98.33 | 2.83 |

Similarly, the results obtained for the disc absent diagnosis component on the simulated sequences are shown in Table 7. In both tables, the accuracy represents the performance of each diagnosis component for classifying each frame as defective or not, taking as ground truth the labels annotated by a human inspector. As can be seen in Table 6, the accuracy of the system is above 88.9% and 97.6% for the damaged-disc and absent-disc diagnosis components respectively in all the sequences.

In order to properly understand the results presented in Table 6, in Fig. 14 we show the average probability generated by the damaged disc diagnosis component on each frame of the sequences. This probability is calculated by averaging the probabilities of all the disc pairs combinations when the criteria for considering a disc damaged are not satisfied, and averaging the probabilities of the positive disc pairs (those combinations with a probability given by the SCNN-c4-L1_vgg16 model greater than a threshold) when the defect criteria are satisfied. As can be noticed in Fig. 14a and 14c, sequences 1 and 3 contain prominent defects which are correctly identified by our proposed system with a high probability in almost every frame in which they appear. This leads to a very low false negative rate which is translated into the high accuracy obtained for sequences 1 and 3 (96.6% and 98.8% respectively).

On the other hand, Fig. 14b and 14d show the average probabilities given by the damaged disc diagnosis component on sequences 2 and 4 respectively. As can be seen in these figures, sequences 2 and 4 present small defects mainly due to bird excrements which are correctly detected by our system. The second insulator in sequence 2 (frames 548 to 709 in Fig.14b) has a very small defect annotated by the human inspector which is detected by our system with a probability greater than 0.5 in only two frames of the sequence. Similarly, the second insulator in sequence 4 (frames 266 to 458 in Fig.14d) has a small defect which in addition is blurred due to the camera movement. Despite this fact, our system is capable of detecting this defect with a probability greater than 0.5 in four frames of the sequence. The fact of detecting the defect in only a few frames increases the false negative rate in detriment of the accuracy, which is 93.3% and 88.9% for sequence 2 and 4 respectively. Despite obtaining few detections in some sequences, due to the aforementioned constraints, we want to remark that the overall performance of the damaged disc diagnosis component, when considering the number of detected defects, is greater than the accuracy presented in Table 6 as only one frame is enough to consider one defect as detected.

Another important aspect to take into account when analyzing the results presented in Table 6 and Fig.14 are the criteria adopted for considering the current frame as defective in terms of damaged discs. The criteria followed by the experiments presented in this section are explained in Algorithm 1 and consider a frame as defective if there is unanimity of all

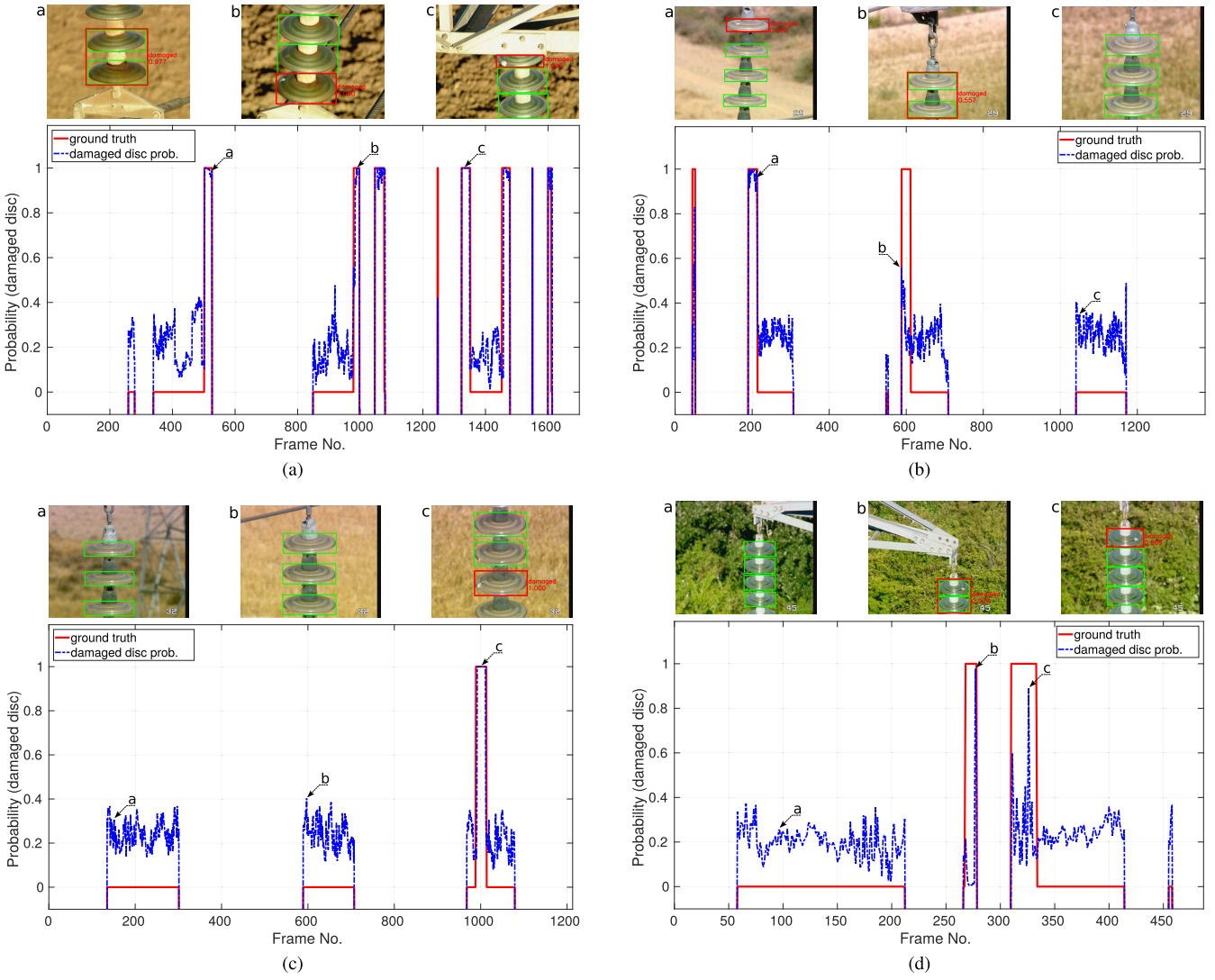


**FIGURE 14.** Results obtained by the damaged disc diagnosis component in 4 test video sequences corresponding to real aerial inspections of high voltage power transmission lines (best seen in color). a) to d) correspond to sequences 1 to 4 in Table 6. Taking into account the output provided by the insulator string segmentation component, some of the frames are automatically discarded (not diagnosed) by our system when the insulator string does not contain the sufficient number of discs.

the disc image pairs in the current frame. As an example, if we consider the first insulator string in sequence 2 (frames 45 to 307 in Fig.14b), frame 208 (see frame *a* in Fig.14b) is marked as defective, and the defect on the first disc is located in the image, since three out of the four disc units are voting for having a defect when combined with the first disc of the insulator string.

Regarding the evaluation of the absent disc diagnosis component on the simulated sequences shown in Table 7, in Fig.15 we illustrate the probabilities generated by the absent disc diagnosis component on each frame of the sequences. As shown in Fig.15a, different scenarios for the disc absent defect have been simulated, where we alternately remove one or more disc units. The results presented in Fig.15 show the low false positive and false negative rates of the system which leads to a classification accuracy of 99.3% for sequence 1. Similarly, Fig.15b depicts the probabilities generated by the absent disc diagnosis component in sequence 2. As shown in Fig. 15b, the system is able to correctly detect the absence of disc insulator units even when the segmented insulator string occupies a small area in the image (see frame *c* in Fig.15b) which leads to an accuracy of 98.3%.

## V. DISCUSSION

The process of inspecting electric power transmission lines is a hugely expensive process for electric companies because of both the capture process and the analysis of the captured data. Due to the considerable variety of defects that may appear in these facilities, the analysis of the data is carried out by a team of specialized inspectors who review the collected images frame by frame. In order to automate the inspection process, machine learning approaches are gaining significant importance as they can provide versatile solutions able to

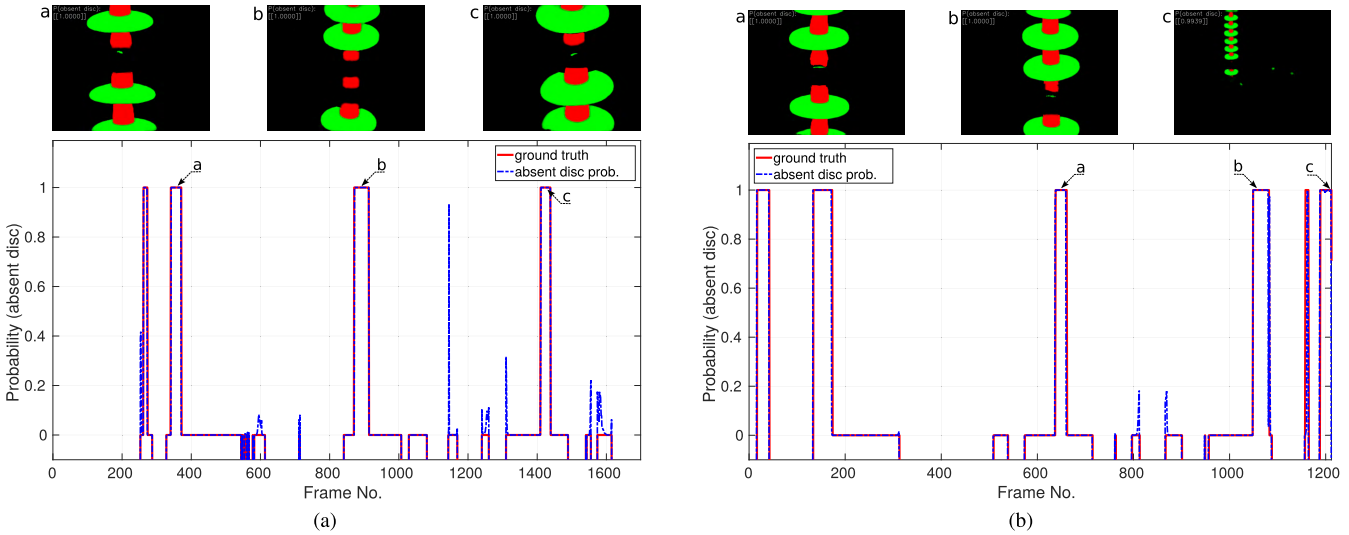


**FIGURE 15.** Results obtained by the absent disc diagnosis component in two simulated sequences (best seen in color). a), b) correspond to sequences 1 and 2 in Table 7 respectively. The absence of disc units is simulated by alternatively removing discs from the mask generated by the insulator string segmentation component. Taking into account this mask, our system automatically discards some of the frames when they do not contain an insulator string for being diagnosed.

work under a wide range of conditions. The results presented throughout this document demonstrate that the proposed system for insulator string recognition and diagnosis represents a breakthrough in this direction, providing an effective solution for automating the inspection process.

Regarding insulator string recognition and segmentation, throughout the results presented in Section IV-B.5, it is shown that semantic segmentation techniques are able to provide accurate segmentation results for the insulator recognition task without the need for a prior stage using an object detection algorithm. Avoiding a previous stage of object detection can lead to considerable savings in computational resources, which can become of utmost importance in systems with hard computational constraints. In this work, we have conducted an extensive evaluation and comparison of several semantic segmentation techniques, ranging from classical state-of-the-art models to more sophisticated approaches involving GANs. In addition, our proposed architecture for semantic segmentation, termed Up-Net, has demonstrated outstanding capabilities obtaining a mean accuracy and mean IoU of 95.16% and 92.06% respectively. Furthermore, we have explored the integration of the proposed Up-Net within a GAN configuration (GAN_Up-Net). The results obtained when using the GAN_Up-Net correspond to the second best performing model in terms of pixel accuracy, mean accuracy and frequency weighted IoU (98.24%, 94.60%, and 96.73% respectively), which consolidates the proposed architecture.

The proposed Up-Net architecture is an adaptation of the original U-Net network [35] in which we add new "up-skip" connections from deeper to shallower levels of the architecture. The modifications added to the base U-Net network provide an improvement with respect to the original network of 1.2%, 1.1%, and 0.7% in mean accuracy, mean IoU, and frequency weighted IoU respectively when training in the final dataset utilized (augmented dataset). These results reveal how the semantic information coming from deeper layers can be efficiently incorporated by these new connections, refining the segmentation results.

Results presented in Table 1 also reveal the influence that the data augmentation and transfer learning techniques have in the insulator string segmentation task. The former consists of using a horizontal flip and two gamma correction operations which allow multiplying by 4 the number of training images in the original dataset. Transfer learning is explored in this paper by using the VGG16 network pre-trained on the ImageNet dataset. The analysis of the results presented in Table 1 evidences the positive impact of the data augmentation and transfer learning techniques on the performance of the semantic segmentation networks evaluated in this work. In most of the networks, the usage of the proposed data augmentation strategy produces an increment between 2% to 5% and 3% to 7% in mean accuracy and mean IoU respectively. The networks more influenced by the data augmentation strategy are the FCN-8s with an improvement in mean accuracy and mean IoU of 4.7% and 6.3% respectively, and the GAN_U-Net which improves the mean accuracy and mean IoU by 4.3% and 7.3% respectively. Using transfer learning, most of the networks experiment an increment in mean accuracy and mean IoU between 4% to 11% and 4% to 12% respectively, being the SegNet network the one which exhibits highest improvements with an increment in mean accuracy and mean IoU of 10.5% and 12.3% respectively.

Regarding the fault diagnosis problem, in this work we leverage from the accurate results obtained by the insulator string segmentation component and adopt the strategy of splitting the problem of fault identification and diagnosis into two separate components. On the one hand, absent discs within the insulator string are diagnosed by a convolutional

network trained directly on the masks generated by the insulator string segmentation component. On the other hand, we address the problem of diagnosing damaged disc units within the insulator string by designing a novel strategy which uses a siamese convolutional network architecture, trained for modeling the similarity between adjacent disc units within the insulator string.

The final selected CNN responsible for diagnosing the absence of disc units within the insulator string (CNN3_7×7) consists of a 10-layer convolutional network, with 3 convolutional blocks (1 × 32, 2 × 64, and 2 × 128) and convolution kernels of 7 × 7. This network has been trained using only synthetic masks and is able to perform remarkably well on real masks generated by the semantic segmentation component from real defective RGB images. This can be reviewed in Fig. 10, where the accurate results produced by the insulator string segmentation component are again validated on images with very different backgrounds, points of view, zoom, and lighting conditions. However, some errors in the insulator segmentation stage are inevitable, especially in extremely challenging conditions such as the ones shown in the two right-most images in Fig. 10b. The false positive generated in the penultimate image (disc partially generated between caps in the region where a disc is absent) and the false negative shown in the last image (missing cap) maintain the alternating order between caps and discs, which is usually presented in healthy insulator strings. Despite this challenging scenario, where traditional image processing techniques would require a set of non-trivial heuristics to identify the absent disc defect, the proposed CNN3_7 × 7 model is able to identify the absence of disc units with high confidence, revealing its good generalization capabilities. This high performance is further demonstrated in the rest of the test images in Fig. 10b for a wide range of scenarios with a variable number and size of the discs within the insulator string. This yields to an $F_1score$ and ROC-AUC of 98.9% and 99.9% respectively in a test set of 400 images as shown in Table 3. The results presented in Table 3 also reveal the good performance obtained by the CNN2 network variants and concretely the CNN2_5 × 5 which obtains a ROC-AUC of 99.4% having 0.56 millions of parameters. This balance between performance and computational resources makes the CNN2_5 × 5 an appropriate alternative candidate in systems with hard computational constraints.

The results presented in Section IV-C.5 also show that the proposed strategy based on training using the masks generated by the insulator string segmentation component can provide several advantages over other state-of-the-art methods. First, it has been demonstrated to be an effective strategy when no training data containing real defects are available. Second, contrary to many state-of-the-art approaches, the proposed method is capable of detecting defects that involve the absence of a variable number of disc units within the insulator string, as can be seen in the results presented in Fig.15a. As a limitation of the proposed method for diagnosing the absence of disc insulator units, the predictions of the network are completely dependent on the segmentation results of previous stages.

With respect to the diagnosis of damaged disc units, in this paper we have analyzed the effect of different distance layers (i.e., L1 and Chi-squared) for computing the similarity between the hidden states of the twin CNNs inside the SCNN. According to the results presented in Table 4, despite both distance layers provide similar results, the SCNN architectures integrating the L1 layer obtain better performance in most of the cases with improvements between 1% to 2% in both ROC-AUC and PR-AUC as compared to their Chi-squared counterparts. The highest difference in performance is produced when analyzing the SCNN-c4-L1/$\chi^2$ with pre-training, in which the network SCNN-c4-L1_vgg16 outperforms by 1.7% and 1.9% in ROC-AUC and PR-AUC respectively its Chi-squared counterpart (SCNN-c4-$\chi^2$_vgg16). The final SCNN model selected for diagnosing damaged discs (SCNN-c4-L1_vgg16) is composed of 4 convolutional blocks (2 × 64, 2 × 128, 3 × 256, and 1 × 64) and two fully-connected layers of 64 units each, after which an L1 layer is applied. The previous architecture when removing the L1 layer, and directly connecting the fully-connected part to the final output unit, allows obtaining an analogous CNN (CNN-c4_vgg16), which is also pre-trained with the weights of the VGG16 network trained on ImageNet and has the same number of parameters as the SCNN-c4-L1_vgg16. The latter has proven to provide better generalization capabilities as compared to its analogous CNN, highlighting the improvement of 3.6% and 3.5% in TPR and PR-AUC, respectively, as shown in Table 5. Despite this improvement, it should be noted that the computational cost derived from the operation of the SCNN-c4-L1_vgg16 is higher than the one of the CNN-c4_vgg16, since the former operates on all the disc pairs combinations at the current frame, and the computational cost of the CNN-c4_vgg16 is directly proportional to the number of discs at the current frame. Thus, in detriment of the performance, the CNN-c4_vgg16 can be used as an alternative solution in systems with hard computational constraints.

As derived form Table 4, the SCNN-c4-L1_vgg16 is benefited by its fourth convolutional block (conv4), which implements a 1 × 1 convolutional layer as a similar concept of the ''Network in Network'' proposed in [53], with a double purpose of shrinking the number of channels before reaching the fully-connected part, providing at the same time another complex level of abstraction to the SCNN. The performance of the SCNN-c4-L1_vgg16 is also benefited by the training strategy proposed in this work (see Section IV-D.3). Training the SCNN-c4-L1_vgg16 using a combined dataset which includes synthetic and real defects (89% and 11% respectively) provides much better performance (PR-AUC of 96.4%) than training the proposed model using only synthetic defects (PR-AUC of 89.9%) or using only the small available dataset of real defects (PR-AUC of 73.7%).

Relevant qualitative and quantitative results showing the performance of the SCNN-c4-L1_vgg16 are also illustrated in Fig. 13. These results on a representative subset of images

from the test set of 245 images, validate the performance of the SCNN-c4-L1_vgg16 under a wide range of real-world conditions such as type of insulator (e.g. glass or ceramic), type of defect (disc burned, painted, with bird excrements on the surface, etc.), lighting conditions, backgrounds, etc. Furthermore, we want to emphasize at this point, the capability of adaptation of the proposed model based on SCNNs. As in other ''one-shot'' learning applications (e.g. face verification), where the system must deal with new samples that can be incorporated into the system, the proposed SCNN can be easily adapted to new defects that may arise in the power line inspection process, without needing to change the structure of the model.

A limitation of the proposed strategy for diagnosing damaged discs is that it requires a minimum of two disc units for providing a diagnosis. Thus, it can not diagnose an insulator disc unit in isolation. Although this could be a problem in another type of applications, the great majority of high voltage power transmission lines contain insulator strings which are made up of dozens of disc units. Therefore, the inspection of these facilities with a predefined flight strategy to correctly capture the data would be beneficial to ensure the proper functioning of the proposed system.

After integrating the different components for insulator string segmentation and fault diagnosis, the proposed system has been thoroughly evaluated using several video sequences corresponding to real aerial inspections of high voltage power transmission lines. The results after evaluating the proposed system are shown in Table 6 and 7, and Fig. 14 and 15. The performance of the proposed system is always above 88.9% in both the damaged-disc and absent-disc diagnosis. We want to remark here that this performance takes into account the classification accuracy computed for each frame of the sequence as containing a defective or non-defective insulator string. Overall, the recall of the system when considering the number of defects is 100%, meaning that all the defects that are present in the video sequences are detected by the proposed system. In this case, we consider that detecting the corresponding defect in one frame can be sufficient as we assume that a specialized human inspector is in charge of reviewing the positive frames diagnosed by our system.

The fact that in some sequences (e.g. sequences 2 and 4 in Table 6, Fig. 14b and 14d) the proposed system only detects the corresponding defect in few frames of the sequence is in part due to the conservative voting scheme implemented inside the damaged disc diagnosis component. Based on this, the system diagnoses a damaged disc only when there is unanimity of all the remaining discs. That is, if an insulator string is composed of several disc units with one defective disc, the system would detect the defect if all the combinations which include the defective disc are giving positive predictions. In other conditions where the low false positive rate is not a requirement, a more relaxed condition can be used where the defect can be identified taking into account only one positive combination. This would lead to a considerable increment in the number of positive detections, but also in the false positive rate.

The results presented in Table 6 also reveal the real-time diagnosis capabilities of the proposed system. Diagnosing absent disc units involves the computation of a feed-forward pass to the proposed 10-layer CNN, giving very fast diagnosis results (2.9 ms per image of $256 \times 256$ pixels on average). The diagnosis of damaged discs using the proposed strategy based on SCNNs involves the computation of a feed-forward pass to the SCNN for each disc pair combination. This strategy makes the computation depends on the number of disc units that are present in the current frame. For this reason, the average diagnosis time varies according to the sequence, being the diagnosis time for sequence 5 higher as compared to other sequences as the number of discs to be diagnosed per frame can be up to 6, which is translated into the diagnosis of ${}^{6}C_{2} = 15$ disc pair combinations. The current strategy can be greatly optimized by splitting the computations of the SCNN into two steps for feature extraction (considering the CNN of the SCNN) and distance calculation (considering the L1 layer and the final output layer). On the one hand, the feature extraction step for computing the hidden states of the disc images can be performed in batch for all the disc images at the current frame. Subsequently, the resulting hidden states can be arranged in pairs for the corresponding combinations and passed to the L1 layer for computing the final prediction of the SCNN. This would avoid repetitive calculations for the extraction of the hidden states of the disc images, with the corresponding savings in computation time. This optimization procedure will be considered for future work.

## VI. CONCLUSIONS AND FUTURE WORK

In this paper, a fully automatic system which efficiently combines different deep learning-based components for insulator string recognition and fault diagnosis has been presented. The proposed system has been extensively evaluated on a large number of images and video sequences captured during real aerial inspections of overhead power line infrastructure, containing more than 5800 frames. The results obtained in these sequences reveal that the proposed system brings an effective solution for automating the power line inspection process and can greatly help in reducing the workload required by the manual inspection process which is nowadays conducted by specialized human inspectors for diagnosing defects in these facilities.

Regarding the insulator string recognition and segmentation, the proposed Fully Convolutional Network (FCN) architecture, termed Up-Net, and its variant GAN-Up-Net (when training Up-Net within a Generative Adversarial Network framework) have shown accurate segmentation capabilities as compared to several state-of-the-art semantic segmentation networks under a wide range of conditions with varying insulator type (e.g. ceramic and glass), lighting conditions, point of view and orientation of the insulator string, among others. The proposed Up-Net architecture is an adaptation

of the state-of-the-art U-Net network in which new ''up-skip'' connections have been integrated at certain levels of the architecture. These connections have proven to provide beneficial effects for incorporating the semantic information from deeper layers of the architecture, obtaining accurate results for segmenting the different elements in the insulator string (i.e., caps and discs), which greatly facilitates the posterior stages for fault diagnosis. These results are positively influenced by the use of data augmentation and transfer learning techniques (using the VGG16 trained on ImageNet) which allow increasing by a large margin the performance of the proposed model and state-of-the-art networks.

With respect to the fault identification and diagnosis, two main components have been implemented for diagnosing the absence of disc insulator units and the presence of damaged disc units within an insulator string. The detection of absent disc units has been addressed in this paper by a 10-layer Convolutional Neural Network (CNN) which takes as input the segmented image generated by the insulator string segmentation component. This strategy permitted to train the proposed CNN using only synthetic masks of insulator strings, allowing to create a large number of training samples in order to reproduce a wide variety of defect configurations (e.g. absence of a variable number of disc units). With these capabilities, the proposed 10-layer CNN obtained diagnostic accuracies above 97% in all the evaluated sequences, reaching 100% in some cases.

On the other hand, the diagnosis of damaged disc units has been addressed in this paper by a novel strategy which integrates a Siamese Convolutional Neural Network (SCNN) responsible for modeling the similarity between adjacent disc units. The proposed diagnosis strategy has proven to be an effective approach for diagnosing a wide range of defects (e.g. discs polluted, burned, rusted, etc.) when a small amount of training data of each type of defect are available. The results obtained in the evaluation of different SCNN architectures reveal the beneficial effect of integrating a $1 \times 1$ convolutional layer before the fully-connected part of each twin CNN of the SCNN, which helps in reducing the number of parameters of the network while providing an additional level of abstraction. In addition, the integration of an L1 distance layer has shown to provide better diagnosis capabilities as compared to other distance layers (e.g. Chi-squared). Added to this is the positive effect shown when conducting transfer learning using the weights of the VGG16 model previously trained on the ImageNet dataset. The selected SCNN architecture composed of two 13-layer twin CNNs joined by a weighted L1 energy function obtained diagnostic accuracies above 89% in all the evaluated sequences, reaching 100% in some cases.

According to the results presented in this work, the authors are greatly satisfied with the results obtained to date, however, there is still room for improvement due to the enormous variability of elements and defects that may arise in the electric power transmission infrastructures. To this aim, some future work lines have been considered, which are summarized next.

To provide a higher level of robustness to the current system, we believe that the disc extraction stage used before the operation of the SCNN can be improved by designing an end-to-end learning-based solution. To this aim, a regression stage can be integrated after the insulator string segmentation component for providing the location of the discs (i.e., the coordinates of their ROIs) within the segmented image and filtering at the same time possible errors that might occur in the segmentation stage. In addition, according to the outstanding results obtained for diagnosing damaged disc units by training the SCNN models using manually generated synthetic defects, we believe that an automatic synthetic defect generation method using generative models such as cycle GANs can greatly benefit the proposed diagnosis system. Furthermore, the versatility shown by the proposed SCNN while modeling the similarity between inputs has led to the consideration of alternative solutions for addressing the absent disc and damaged disc defects using the same SCNN framework. To this aim, one alternative solution for diagnosing the absence of disc units is to design an SCNN which takes as input a pair of segmented images corresponding to insulator strings extracted from the output provided by the insulator string segmentation component. This alternative SCNN can operate in video sequences taking as input a pair of segmented insulator strings corresponding to two consecutive frames of the sequence, providing a positive detection when one or both images of the pair present an absent disc.

Finally, the extension of the proposed automatic inspection strategy to other components in the power line infrastructure will be considered for future works.

## ACKNOWLEDGMENT

The authors would like to thank the companies Gas Natural Unión Fenosa and Prysma for the aerial inspection data supplied within the INNPACTO IPT-2012-0491-120000 project. In addition, we would like to acknowledge Eduardo Temprano for the exhaustive annotations of the images used for evaluating the semantic segmentation networks. Finally, the authors would like to thank Dr. Miguel Fernandez for the proofreading of the final version of the manuscript.

## REFERENCES

[1] F. Mirallès, P. Hamelin, G. Lambert, S. Lavoie, N. Pouliot, M. Montfrond, and S. Montambault, ''LineDrone technology: Landing an unmanned aerial vehicle on a power line,'' in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2018, pp. 6545–6552.

[2] V. N. Nguyen, R. Jenssen, and D. Roverso, ''Intelligent monitoring and inspection of power line components powered by UAVs and deep learning,'' *IEEE Power Energy Technol. Syst. J.*, vol. 6, no. 1, pp. 11–21, Mar. 2019.

[3] P. Debenest, M. Guarnieri, K. Takita, E. F. Fukushima, S. Hirose, K. Tamura, A. Kimura, H. Kubokawa, N. Iwama, and F. Shiga, ''Expliner—Robot for inspection of transmission lines,'' in *Proc. IEEE Int. Conf. Robot. Automat.*, May 2008, pp. 3978–3984.

[4] N. Morozovsky and T. Bewley, ''SkySweeper: A low DOF, dynamic high wire robot,'' in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Nov. 2013, pp. 2339–2344.

[5] P. Taklaja, R. Oidram, J. Niitsoo, and I. Palu, ''Main bird excrement contamination type causing insulator flashovers in 110 kv overhead power lines in estonia,'' *Oil Shale*, vol. 30, no. 2S, pp. 211–224, 2013.

[6] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 1, Jun. 2005, pp. 539–546.

[7] G. Koch, R. Zemel, and R. Salakhutdinov, "Siamese neural networks for one-shot image recognition," in *Proc. ICML Deep Learn. Workshop*, vol. 2, 2015, pp. 1–30.

[8] W. Chang, G. Yang, J. Yu, and Z. Liang, "Real-time segmentation of various insulators using generative adversarial networks," *IET Comput. Vis.*, vol. 12, no. 5, pp. 596–602, Aug. 2018.

[9] Y. Zhai, R. Chen, Q. Yang, X. Li, and Z. Zhao, "Insulator fault detection based on spatial morphological features of aerial images," *IEEE Access*, vol. 6, pp. 35316–35326, 2018.

[10] Y. Han, Z. Liu, D. Lee, W. Liu, J. Chen, and Z. Han, "Computer vision–based automatic rod-insulator defect detection in high-speed railway catenary system," *Int. J. Adv. Robotic Syst.*, vol. 15, no. 3, pp. 1–15, 2018.

[11] A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, V. Villena-Martinez, and J. Garcia-Rodriguez, "A review on deep learning techniques applied to semantic segmentation," 2017, *arXiv:1704.06857*. [Online]. Available: https://arxiv.org/abs/1704.06857

[12] M. Mirza and S. Osindero, "Conditional generative adversarial nets," 2014, *arXiv:1411.1784*. [Online]. Available: https://arxiv.org/abs/1411.1784

[13] C. C. Whitworth, A. W. G. Duller, D. I. Jones, and G. K. Earp, "Aerial video inspection of overhead power lines," *Power Eng. J.*, vol. 15, no. 1, pp. 25–32, Feb. 2001.

[14] I. Golightly and D. Jones, "Corner detection and matching for visual tracking during power line inspection," *Image Vis. Comput.*, vol. 21, no. 9, pp. 827–840, Sep. 2003.

[15] C. Sun, R. Jones, H. Talbot, X. Wu, K. Cheong, R. Beare, M. Buckley, and M. Berman, "Measuring the distance of vegetation from powerlines using stereo vision," *ISPRS J. Photogramm. Remote Sens.*, vol. 60, no. 4, pp. 269–283, 2006.

[16] W. Cheng and Z. Song, "Power pole detection based on graph cut," in *Proc. IEEE Congr. Image Signal Process.*, vol. 3, May 2008, pp. 720–724.

[17] Y. Zhai, D. Wang, M. Zhang, J. Wang, and F. Guo, "Fault detection of insulator based on saliency and adaptive morphology," *Multimedia Tools Appl.*, vol. 76, no. 9, pp. 12051–12064, 2017.

[18] X. Wang and Y. Zhang, "Insulator identification from aerial images using support vector machine with background suppression," in *Proc. IEEE Int. Conf. Unmanned Aircr. Syst. (ICUAS)*, Jun. 2016, pp. 892–897.

[19] P. S. Prasad and B. P. Rao, "Condition monitoring of 11 kV overhead power distribution line insulators using combined wavelet and LBP-HF features," *IET Gener., Transmiss. Distrib.*, vol. 11, no. 5, pp. 1144–1153, 2016.

[20] Z. Zhao, G. Xu, Y. Qi, N. Liu, and T. Zhang, "Multi-patch deep features for power line insulator status classification from aerial images," in *Proc. IEEE Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2016, pp. 3187–3194.

[21] C. Martinez, C. Sampedro, A. Chauhan, J. F. Collumeau, and P. Campoy, "The power line inspection software (PoLIS): A versatile system for automating power line inspection," *Eng. Appl. Artif. Intell.*, vol. 71, pp. 293–314, May 2018.

[22] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[23] J. Gubbi, A. Varghese, and P. Balamuralidhar, "A new deep learning architecture for detection of long linear infrastructure," in *Proc. IEEE 14th IAPR Int. Conf. Mach. Vis. Appl. (MVA)*, May 2017, pp. 207–210.

[24] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1–9.

[25] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2005, vol. 1, no. 1, pp. 886–893.

[26] A. Varghese, J. Gubbi, H. Sharma, and P. Balamuralidhar, "Power infrastructure monitoring and damage detection using drone captured images," in *Proc. IEEE Int. Joint Conf. Neural Netw. (IJCNN)*, May 2017, pp. 1681–1687.

[27] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.

[28] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 7263–7271.

[29] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, Oct. 2016, pp. 21–37.

[30] F. Gao, J. Wang, Z. Kong, J. Wu, N. Feng, S. Wang, P. Hu, Z. Li, H. Huang, and J. Li, "Recognition of insulator explosion based on deep learning," in *Proc. IEEE 14th Int. Comput. Conf. Wavelet Act. Media Technol. Inf. Process. (ICCWAMTIP)*, Dec. 2017, pp. 79–82.

[31] Z. Ling, R. C. Qiu, Z. Jin, Y. Zhang, X. He, H. Liu, and L. Chu, "An accurate and real-time self-blast glass insulator location method based on faster R-CNN and U-net with aerial images," 2018, *arXiv:1801.05143*. [Online]. Available: https://arxiv.org/abs/1801.05143

[32] Z. A. Siddiqui, U. Park, S.-W. Lee, N.-J. Jung, M. Choi, C. Lim, and J.-H. Seo, "Robust powerline equipment inspection system based on a convolutional neural network," *Sensors*, vol. 18, no. 11, p. 3837, 2018.

[33] X. Miao, X. Liu, J. Chen, S. Zhuang, J. Fan, and H. Jiang, "Insulator detection in aerial images for transmission line inspection using single shot multibox detector," *IEEE Access*, vol. 7, pp. 9945–9956, 2019.

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

[35] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, Oct. 2015, pp. 234–241.

[36] X. Tao, D. Zhang, Z. Wang, X. Liu, H. Zhang, and D. Xu, "Detection of power line insulator defects using aerial images analyzed with convolutional neural networks," *IEEE Trans. Syst., Man, Cybern. Syst.*, to be published.

[37] T. Jabid and T. Ahsan, "Insulator detection and defect classification using rotation invariant local directional pattern," *Int. J. Adv. Comput. Sci. Appl.*, vol. 9, no. 2, pp. 265–272, 2018.

[38] M. Oberweger, A. Wendel, and H. Bischof, "Visual recognition and fault detection for power line insulators," in *Proc. 19th Comput. Vis. Winter Workshop*, Krtiny, Czech Republic, 2014, pp. 1–8.

[39] C. Sampedro, C. Martinez, A. Chauhan, and P. Campoy, "A supervised approach to electric tower detection and classification for power line inspection," in *Proc. IEEE Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2014, pp. 1970–1977.

[40] C. Martinez, C. Sampedro, A. Chauhan, and P. Campoy, "Towards autonomous detection and tracking of electric towers for aerial power line inspection," in *Proc. IEEE Int. Conf. Unmanned Aircr. Syst. (ICUAS)*, May 2014, pp. 284–295.

[41] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: https://arxiv.org/abs/1409.1556

[42] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "DeepFace: Closing the gap to human-level performance in face verification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1701–1708.

[43] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[44] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[45] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: https://arxiv.org/abs/1412.6980

[46] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1026–1034.

[47] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 1125–1134.

[48] 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.

[49] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. P. Smolley, "Least squares generative adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2794–2802.

[50] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Oct. 2011.

[51] G. H. Nguyen, A. Bouzerdoum, and S. L. Phung, "Learning pattern classification tasks with imbalanced data sets," in *Pattern Recognition*. Rijeka, Croatia: InTech, 2009.
[52] M. Kubat and S. Matwin, "Addressing the curse of imbalanced training sets: One-sided selection," in *Proc. 14th Int. Conf. Mach. Learn.*, 1997, pp. 179–186.
[53] M. Lin, Q. Chen, and S. Yan, "Network in network," 2013, *arXiv:1312.4400*. [Online]. Available: https://arxiv.org/abs/1312.4400

**CARLOS SAMPEDRO** received the B.Sc. degree in industrial engineering (major in industrial electronics), obtaining the best marks degree award, and the master's degree in automation and robotics from the Universidad Politécnica de Madrid (UPM), Madrid, Spain, in July 2011 and 2014, respectively. He is currently pursuing the Ph.D. degree with the Computer Vision and Aerial Robotics (CVAR) Group, Centre for Automation and Robotics, Universidad Politécnica de Madrid. To this aim, he received a predoctoral grant from the Universidad Politécnica de Madrid, in January 2017. In addition, he was a Visiting Researcher for three months with Arizona State University, AZ, USA, from September 2015 to December 2015. His research interest includes the applications of learning-based techniques for solving computer vision problems, with a special interest in object detection and recognition using deep learning techniques. Besides, he is actively involved in the development of deep reinforcement learning algorithms for autonomous navigation and control of unmanned aerial vehicles (UAVs).

**JAVIER RODRIGUEZ-VAZQUEZ** received the B.Sc. degree in computer engineering (double major in hardware engineering and computer science) and the M.Sc. degree in systems engineering and computing research from the Universidad de Cádiz (UCA), Cádiz, Spain, in July 2015 and March 2017, respectively. He is currently pursuing the Ph.D. degree in artificial intelligence with the Universidad Politécnica de Madrid (UPM). His research interest includes deep learning methods for solving computer vision tasks, with a special interest in object detection and image segmentation.

**ALEJANDRO RODRIGUEZ-RAMOS** received the B.Sc. degree in telecommunication engineering (major in electronics and micro-electronics) from the Universidad Politécnica de Madrid (UPM), Madrid, Spain, in 2015. He is currently pursuing the Ph.D. degree with the Computer Vision and Aerial Robotics (CVAR) Group, Centre for Automation and Robotics, Universidad Politécnica de Madrid, where he is currently a Researcher. Previously, he worked for more than a year in the aerospace sector, contributing to projects from the European Space Agency. His research interests include deep reinforcement learning techniques applied to aerial robotics, deep learning, aerial robotics, and image processing. Besides, he was a Visiting Researcher with the Aerospace Controls Laboratory (ACL), Massachusetts Institute of Technology (MIT), from October to December 2018, for three months.

**ADRIAN CARRIO** (M'18) received the B.Sc. degree in industrial engineering from the University of Oviedo, Asturias, Spain, in 2012. He has served as a Research Scholar with the Autonomous System Technologies Research and Integration Laboratory (ASTRIL), Arizona State University, and the Aerospace Controls Laboratory (ACL), Massachusetts Institute of Technology (MIT). He is currently pursuing the Ph.D. degree with the Computer Vision and Aerial Robotics Group (CVAR), Universidad Politécnica de Madrid, where he is currently a Research Associate. His research interest includes the development of vision-based collision avoidance systems for unmanned aerial vehicles (UAVs). His research interests include perception in challenging and dynamic environments, applied machine learning for object detection and recognition, and autonomous UAV navigation.

**PASCUAL CAMPOY** (M'95) is a Full Professor on automation and robotics with the Universidad Politécnica de Madrid (UPM), Madrid, Spain, and a Visiting Professor with TUDelft, The Netherlands. He has also been Visiting Professor with Tongji University, Shanghai, China, and QUT, Australia. He is currently a Lecturer on control, machine learning, and computer vision. He is leading the Computer Vision and Aerial Robotics (CVAR) Research Group, Centre for Automation and Robotics (CAR). He has been the Head Director of over 40 research and development projects, including research and development European projects, national research and development projects, and over 25 technological transfer projects directly contracted with the industry. He is the author of over 200 international scientific publications. He holds nine patents, three of them registered internationally. He has received several international prizes in UAV competitions such as IMAV 2012, IMAV 2013, IARC 2014, IMAV 2016, and IMAV 2017.