Gencer Sumbul, Sonali Nayak, Begüm Demir

SD-RSIC: Summarization-Driven Deep Remote Sensing Image Captioning

Sumbul, G., Nayak, S., & Demir, B. (2020). SD-RSIC: Summarization-Driven Deep Remote Sensing Image Captioning. IEEE Transactions on Geoscience and Remote Sensing, 1–13. https://doi.org/10.1109/tgrs.2020.3031111

SD-RSIC: Summarization Driven Deep Remote Sensing Image Captioning

Gencer Sumbul, *Graduate Student Member, IEEE*, Sonali Nayak, and Begüm Demir *Senior Member, IEEE*

***Abstract*—Deep neural networks (DNNs) have been recently found popular for image captioning problems in remote sensing (RS). Existing DNN based approaches rely on the availability of a training set made up of a high number of RS images with their captions. However, captions of training images may contain redundant information (they can be repetitive or semantically similar to each other), resulting in information deficiency while learning a mapping from the image domain to the language domain. To overcome this limitation, in this paper, we present a novel Summarization Driven Remote Sensing Image Captioning (SD-RSIC) approach. The proposed approach consists of three main steps. The first step obtains the standard image captions by jointly exploiting convolutional neural networks (CNNs) with long short-term memory (LSTM) networks. The second step, unlike the existing RS image captioning methods, summarizes the ground-truth captions of each training image into a single caption by exploiting sequence to sequence neural networks and eliminates the redundancy present in the training set. The third step automatically defines the adaptive weights associated to each RS image to combine the standard captions with the summarized captions based on the semantic content of the image. This is achieved by a novel adaptive weighting strategy defined in the context of LSTM networks. Experimental results obtained on the RSCID, UCM-Captions and Sydney-Captions datasets show the effectiveness of the proposed approach compared to the state-of-the-art RS image captioning approaches. The code of the proposed approach is publicly available at https://gitlab.tubit.tu-berlin.de/rsim/SD-RSIC.**

***Index Terms*—Image captioning, caption summarization, deep learning, remote sensing.**

I. Introduction

THE new generation of remote sensing (RS) sensors characterized by very high geometrical resolution can acquire images with sub-metric spatial resolution. Thus, the significant amount of geometrical details can be presented in very high resolution RS image scenes. Accordingly, one of the most important applications is the RS image captioning, which aims at automatically assigning descriptive sentences (i.e., captions) to RS image scenes by accurately characterizing their semantic content. Recent studies in RS have shown that deep neural networks (DNNs) are capable of generating accurate image captions for RS images due to their ability to model a mapping from the high-level semantic content of RS images in image domain into the descriptive captions in language domain [1]. DNN based encoder-decoder framework is one of the most effective methods for RS image captioning. Within this framework, image captioning is achieved based on two steps. In the first step, convolutional neural networks (CNNs) are used to extract image features, while in the second step recurrent neural network (RNN) based sequential approaches are used as a natural language model to generate a caption for each image based on the image features. The overall framework is considered as an encoder-decoder neural network where the encoder (CNN) takes an image as input and generates the corresponding encoded features, whereas the decoder generates a caption for the image based on the features. Then, the neural network trained on image-caption pairs can automatically generate a caption for a new image. Accordingly, in [2], CNNs and RNNs are employed to generate captions by combining image features of very high resolution RS images with the associated captions. In detail, pre-trained CNN models on a widely used computer vision dataset (i.e., ImageNet) are used to extract image features, while long-short term memory (LSTM) networks are utilized to sequentially characterize the image captions. In this study, two image captioning datasets are introduced as the first time in RS to evaluate the success of RS image captioning approaches. In [3], a conventional template-based method is presented in the context of RS image captioning for the cases where the number of RS images annotated with captions is not sufficient. This method represents RS images with a combination of ground elements, their attributes and relations that derive a language template. In detail, a fully convolutional network is introduced for the detection of multi-level ground elements, while captions are generated based on the predefined templates. In [4], the largest RS image captioning dataset, which is called RSICD, is introduced. In this study, traditional hand crafted features are compared with the features extracted through different CNN models in the context of RS image captioning, while the caption generation strategy introduced in [2] is used. A Collective Semantic Metric Framework (CSMLF) that models the common semantic space of RS images and their captions is recently introduced in [5]. In detail, CSMLF maps the GloVe based representations of image captions and the image features from a pre-trained CNN model into a common semantic space with a metric learning strategy. Then, the distance between a new image and all captions in the common space is computed to generate a new caption. In [6], an attribute attention strategy that exploits the correlation between image regions and generated caption words is integrated into the standard encoder-decoder approach to further improve the semantic content characterization of images. In this approach, fully

This work is funded by the European Research Council (ERC) through the ERC-2017-STG BigEarth Project under Grant 759764. *(Corresponding author: Begüm Demir).*

Gencer Sumbul, Sonali Nayak and Begüm Demir are with the Faculty of Electrical Engineering and Computer Science, Technische Universität Berlin, 10623 Berlin, Germany. Email: gencer.suembuel@tu-berlin.de, sonalirosy91@gmail.com, demir@tu-berlin.de.

connected (FC) layers of a CNN are considered to characterize the image attributes, while convolutional layers are employed to obtain image features. The caption generation is achieved by using LSTMs (where the log likelihood of generating a caption word by word is maximized given the previous words), the image feature and corresponding image attributes. We would like note that although only a few DNN based RS image captioning approaches are proposed in RS literature, this research field has been extensively studied in computer vision. As an example, the above-mentioned encoder-decoder framework that jointly employs CNNs and RNNs for image captioning is initially introduced in [7] as the first time. In [8], an attention mechanism is employed to characterize where or what to look in images to generate their captions. In [9], topic embeddings are first extracted from a CNN-based multi-label classifier and then used with image features in an LSTM-based language model to generate topic-oriented image captions. We refer the readers to [10] for a detailed review of DNN based image captioning approaches introduced in computer vision.

Most of the existing DNN based approaches in the context of RS image captioning rely on the availability of a training set, which consists of very high resolution RS images with their captions (which accurately describe the semantic content of images). Due to the complexity of learning in RS image and language domains, multiple captions are usually assigned to each training image to effectively and efficiently learn an image captioning model. Although each RS image is expected to be ideally described with different captions, each of which embodies different information of the image, a training set may contain redundant information through multiple captions. As an example, in the existing benchmark image captioning datasets (e.g., RSICD, Sydney-Captions and UCM-Captions), most of the RS images are associated with repetitive captions or similar captions with small differences. This can cause the information deficiency while learning a mapping from the image domain to the language domain. Redundant information in training sets may also lead to over-fitting in training, which reduces the generalization capability of image captioning models and thus causes poor image captioning performance. None of the existing DNN based approaches in RS take into account the above-mentioned problems. Thus, if a DNN model is trained on image caption pairs that include redundant information, existing captioning methods in RS may provide insufficient captioning performance.

To overcome this limitation, in this paper, we introduce a novel Summarization Driven Remote Sensing Image Captioning (SD-RSIC) approach. The SD-RSIC aims at: i) learning to summarize image captions learned on large text corpora; and then ii) integrating it with the learning procedure of the captioning task to guide the whole training process. To this end, the proposed approach is made up of three main steps: 1) generation of standard captions; 2) summarization of ground-truth captions; and 3) integration of summarized captions with standard captions. In the first step, CNNs and LSTMs are jointly used as in the literature works for learning of standard image captions based on image features. In the second step, unlike the existing methods, we propose to exploit a sequence-to-sequence DNN model to summarize ground-truth captions of each image into a single caption. Due to this step, the proposed SD-RSIC approach is capable of eliminating redundant information present in captions, while enhancing the word vocabulary that provides more detailed captions for semantically complex RS images. In the third step, to integrate the summarized captions with the standard captions, the vocabulary word probabilities of standard captions are combined with those of the summarized captions based on the image features by a novel adaptive weighting strategy in the framework of LSTMs. This step reduces the risk of over-fitting during training, and thus improves the generalization capability of the whole approach. The novelty of the proposed approach consists in: 1) summarization of ground-truth captions into single caption per RS image to eliminate the redundancy present in the ground-truth captions; 2) integration of the summarized captions with standard captions by an adaptive weighting strategy; and 3) exploiting the summarization approach that guides whole training procedure.

The rest of the paper is organized as follows: Section II provides the formulation of the image captioning task and introduces the proposed SD-RSIC approach. Section III describes the considered datasets, while Section IV provides the experimental results. Section V concludes our paper.

II. Proposed Summarization Driven Remote Sensing Image Captioning (SD-RSIC) Approach

In this section, we first formulate the RS image captioning task, and then explain our Summarization Driven Remote Sensing Image Captioning (SD-RSIC) approach. Let $\mathcal{I} = \{I_1, \ldots, I_M\}$ be an archive that consists of M images, where I_i is the i^{th} image. We assume that a training set $\mathcal{T} \subset \mathcal{I}$ of images, each of which is annotated with one or more captions, is initially available. Let $C_i = \{c_{i,j}\}_{j=1}^{N_i}$ be the caption set associated with the i^{th} image I_i, where $c_{i,j}$ is the j^{th} caption of the set C_i and N_i is the number of considered captions. Each caption of the set C_i can be formulated as the set of ordered words $c_{i,j} = \{w_k\}_{k=1}^{L_{i,j}}$, where w_k is the k^{th} word in the caption and $L_{i,j}$ is the length of the caption $c_{i,j}$. The image captioning task aims to learn a function $F(I^*;\theta)$ that assigns a descriptive caption to a new image I^*. To this end, the parameters of the function can be learned by maximizing the log probability of the ground-truth captions for each (I_i, C_i) training instance pair as follows:

$$\theta^* = \arg\max_{\theta} \left(\sum_{i=1}^{|\mathcal{T}|} \sum_{j=1}^{N_i} \sum_{k=1}^{L_{i,j}} \log P(w_k | w_{1:k-1}, I_i; \theta) \right) \quad (1)$$

where θ is the whole parameter set of the function and $P(w_k | w_{1:k-1}, I_i; \theta)$ is the probability of the k^{th} word w_k, which is conditioned on the previous words of the caption $c_{i,j}$ and the image I_i. Then, the caption of the image I^* can be obtained by estimating the probabilities of corresponding words $P(w_k^* | w_{1:k-1}^*, I^*; \theta^*)$ with learned parameters. Conventional image captioning approaches in deep learning are based on encoder-decoder architectures for which the semantic content of RS images is encoded to facilitate the caption generation.

Learning image-caption mapping generally requires describing each image with many captions in the training set since

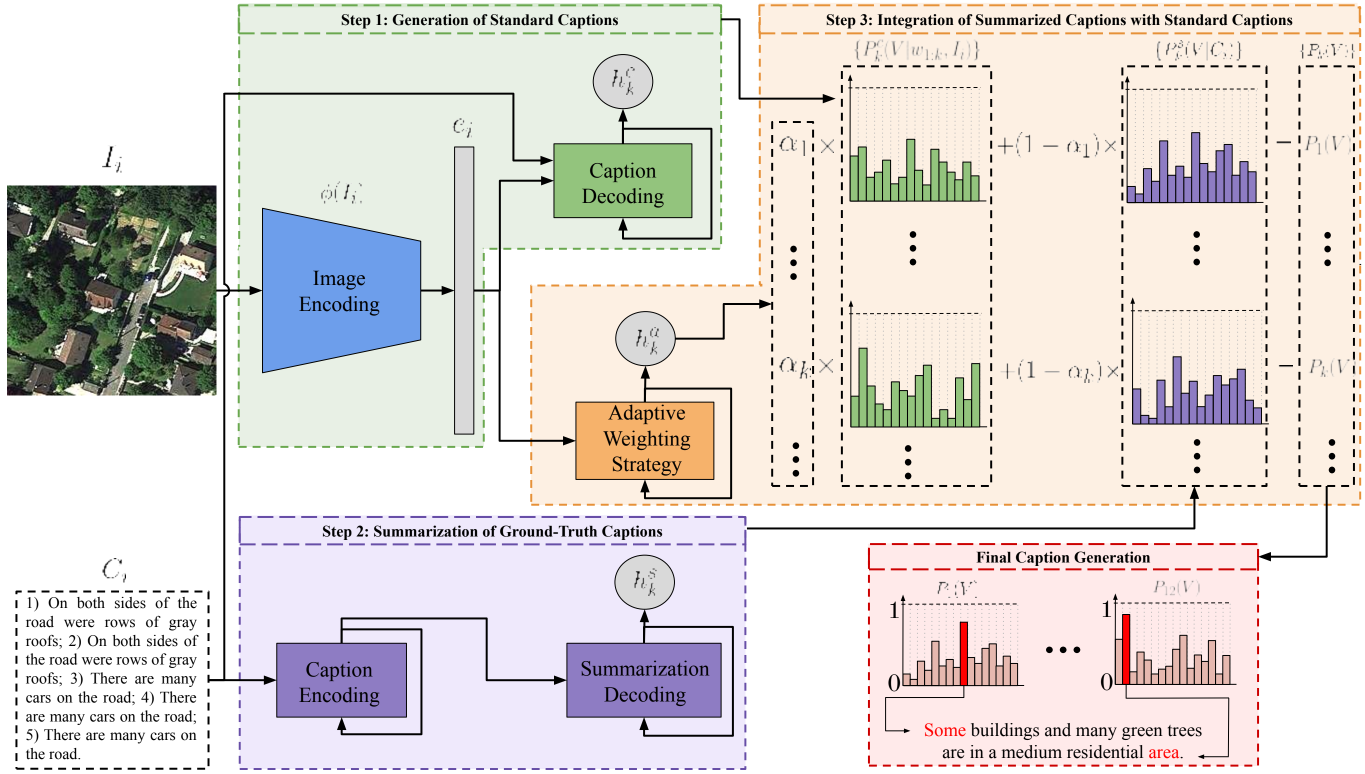

I_i
Step 1: Generation of Standard Captions
$\phi(I_i)$
Image Encoding
e_i
Caption Decoding
h_k^c
Step 3: Integration of Summarized Captions with Standard Captions
$\{P_k^c(V|w_{1:k}, I_i)\}$
$\{P_k^s(V|C_i)\}$
$\{P_k(V)\}$
$\alpha_1 \times$
$+(1-\alpha_1)\times$
$- P_1(V)$
$\alpha_k \times$
$+(1-\alpha_k)\times$
$- P_k(V)$
h_k^a
Adaptive Weighting Strategy
C_i
1) On both sides of the road were rows of gray roofs; 2) On both sides of the road were rows of gray roofs; 3) There are many cars on the road; 4) There are many cars on the road; 5) There are many cars on the road.
Step 2: Summarization of Ground-Truth Captions
Caption Encoding
Summarization Decoding
h_k^s
Final Caption Generation
$P_1(V)$
$P_{12}(V)$
1
0
Some buildings and many green trees are in a medium residential area.

Fig. 1. The proposed Summarization Driven Remote Sensing Image Captioning (SD-RSIC) approach.

by this way caption and image semantics can be accurately associated. However, the captions can share very similar semantics or include a large number of same words with similar orders. The disadvantages of redundant information present in ground-truth captions are twofold. First, this can cause the information deficiency during the learning process. Second, redundancy present in the captions can lead to overfitting in training, which reduces the generalization capability of captioning models and thus causes poor image captioning performance. To address these problems, the proposed SD-RSIC approach is characterized by three main steps: 1) generation of standard captions; 2) summarization of ground-truth captions; and 3) integration of summarized captions with standard captions. The first step is based on the widely used learning method that jointly exploits CNNs and LSTMs for the image captioning problems. The novelty of the proposed SD-RSIC approach relies on the last two steps. In the second step, we propose to exploit sequence-to-sequence DNN models for the summarization of ground-truth image captions to eliminate the redundant information. In the third step, we introduce a novel adaptive weighting strategy to accurately define the weights for integrating the summarized captions with the standard captions according to the image features. Fig. 1 presents a general overview of the proposed SD-RSIC approach and each step is explained in the following sections.

A. Step 1: Generation of Standard Captions

This step aims at generating consecutive words in a meaningful order that characterizes the standard image captions based on the image features. To this end, similar to the literature works in RS (e.g., [4]), we utilize: i) CNNs to capture the high level semantic content of RS images; and ii) LSTMs to learn a mapping between the image features and consecutive word embeddings by sequentially modeling the language semantics. Let ϕ be any type of CNN. For a given image I_i, $\phi(I_i)$ provides a feature vector (i.e., image descriptor) to model the content of the image. In order to map the extracted feature vector to a common space with image captions, the extracted feature vector is given as input to a FC layer, which provides the final image embedding e_i having the dimension of W. After the characterization of image features, an LSTM network produces a word at each time step based on the previous LSTM states and the word predictions to sequentially capture word semantics, while relying on the image features. At the beginning of the sequence, the image embedding e_i is fed into the LSTM network that performs as the initial input of the sequence to affect the following word predictions. To start the caption sequence, we employ the special start token w_0 for all captions. Word generation is repeated until the special end token w_e reaches to the network. To this end, we represent each word as a one-hot vector of dimension $|V|$, where V is the vocabulary set including all unique words. In order to encode semantic similarity in words, we apply mapping from the one-hot vector representation into a real-valued embedding of words with the dimension of W as follows:

$$u_k = \mathbf{E}w_k, \quad w_k \in V \tag{2}$$

where $\mathbf{E}$ is the word embedding matrix with the size of $W \times |V|$. The LSTM network of this step exploits word embedding

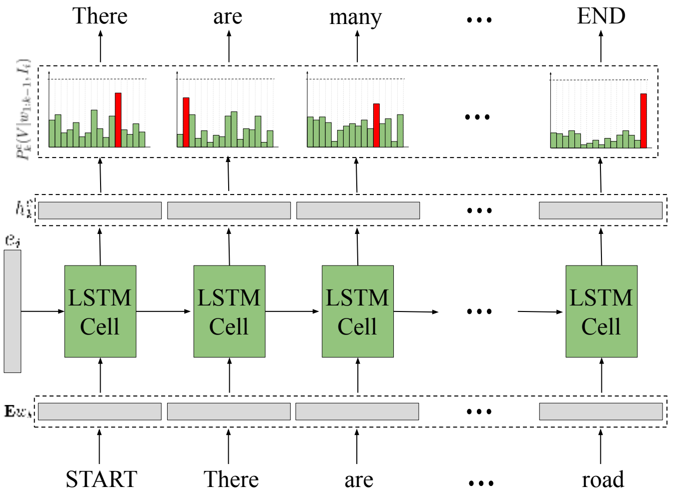


Fig. 2. The first step of the SD-RSIC approach. The LSTM network used for this step is represented as unrolled, showing the input and output of a time step in the sequence.

and previous information of the sequence at each time step as follows:

$$\begin{aligned} f_k &= \delta(\mathbf{W}_{f,u} u_k + \mathbf{U}_{f,h} h_{k-1}^c + b_f) \\ i_k &= \delta(\mathbf{W}_{i,u} u_k + \mathbf{U}_{i,h} h_{k-1}^c + b_i) \\ o_k &= \delta(\mathbf{W}_{o,u} u_k + \mathbf{U}_{o,h} h_{k-1}^c + b_o) \\ c_k^c &= f_k \odot c_{k-1}^c + i_k \odot \tanh(\mathbf{W}_{c,u} u_k \\ &\qquad\qquad + \mathbf{U}_{c,h} h_{k-1}^c + b_c) \\ h_k^c &= o_k \odot \tanh(c_k^c) \end{aligned} \tag{3}$$

where $\mathbf{W}_.$ and $b_.$ are the weight and bias parameters, respectively. $\tanh$ and δ are the hyperbolic tangent and sigmoid functions, and i, f, o and c are input gate, forget gate, output gate and cell state, respectively (for a detailed explanation, see [11], [12]). At the beginning of the sequence, c_0^c and h_0^c are randomly initialized. Then, we obtain word probabilities at each time step with softmax function following to a classification layer as follows:

$$P_k^c(V|w_{1:k-1}, I_i; \theta) = \sigma(\mathbf{W}_{p,h} h_k^c + b_p) \tag{4}$$

where σ is the softmax function and $\mathbf{W}_{p,h}$ and b_p are the weight and bias parameters of a FC layer. $P_k^c(V|w_{1:k-1}, I_i; \theta)$ denotes the probability distribution of all vocabulary words produced at the k^{th} time step of the corresponding LSTM network. This step is illustrated in Fig. 2.

B. Step 2: Summarization of Ground-Truth Captions

This step aims to summarize the ground-truth captions of RS images. The summarized captions guide the whole training process of the proposed SD-RSIC approach. To this end, we propose to adapt the automatic summarization task of natural language processing literature into the image captioning problem. The summarization task is defined as condensing a text to a shorter version that contains the most important information. In our approach, we exploit pointer-generator DNNs [13] as a special type of sequence-to-sequence neural networks. To this end, we consider to train the pointer-generator model on news articles to automatically extract headlines. Then, we exploit the model for summarizing ground-truth captions in our approach. To this end, we stack all corresponding captions of each RS image as a single text to summarize them into single caption. Then, all words of stacked captions are embedded as in (2) and fed into the pre-trained model. Two recurrent neural networks sequentially encode the stacked captions and decode them to generate a summarized caption in order. In addition, pointer-generator structure decides the probability of generating words from the vocabulary versus copying from all captions. This allows an accurate reproduction of information, while retaining the ability to produce novel words through the generator (for a detailed explanation, see [13]). Let ψ be the pre-trained summarization network, $\psi(\{c_{i,j}\}_{j=1}^{N_i})$ produces the word probabilities of the vocabulary $P_k^s(V|\{c_{i,j}\}_{j=1}^{N_i})$ at the k^{th} time step.

Due to the summarization of ground-truth captions, the proposed SD-RSIC approach is capable of eliminating redundant information present in the multiple captions associated with each training image by condensing all captions into a single caption that captures the most significant information content. In addition, the summarization model is pre-trained on a dataset whose vocabulary is excessively larger than any RS image captioning dataset. In this way, our approach uses significantly bigger vocabulary (which is also used in all steps of the SD-RSIC) compared to existing approaches. Using enriched vocabulary increases the capability of our approach to generate more detailed captions for semantically complex RS images.

C. Step 3: Integration of Summarized Captions with Standard Captions

This step aims to define a final caption for each image by reducing the limitations of redundant information in ground-truth captions, while providing the detailed language semantics. To this end, we propose to integrate the standard caption of each image with its summarized caption based on a novel adaptive weighting strategy. The proposed strategy employs an LSTM network, which automatically characterizes the weights for combining the vocabulary word probabilities of standard captions with those of the summarized captions at each time step. Initial cell state c_0^a and hidden state h_0^a of the LSTM network are randomly initialized, and then the LSTM takes the final image embedding e_i as input at each time step. Then, a single weight score h_t^a is produced as in (3) at each time step based on the previous cell states and the image embedding. To normalize the scores to the range of $[0, 1]$, we apply sigmoid function to obtain the final weights $\{\alpha_k\}_{k=1}^{N_i}$ for the RS image I_i. Then, final word probability distribution at time step k is obtained by the weighted combination of the word probabilities of standard captions (which is obtained in the first step) and those obtained in the second step as follows:

$$P_k(V) = \alpha_k \times P_k^c(V|w_{1:k-1}, I_i) + (1-\alpha_k) \times P_k^s(V|C_i). \tag{5}$$

If there is no corresponding output in the first or second step at the k^{th} time step, we apply zero-padding to the shorter output. After obtaining the probabilities for all time steps, we achieve the final caption by selecting the words leading to the highest probabilities. Since the learning of the weights is achieved based on the image features, weights are adaptive depending on the content of the images, i.e., different weights are assigned to different images. Due to the proposed adaptive weighting strategy, the proposed SD-RSIC approach is capable of exploiting the summarized captions to guide the training of the whole neural network. With this guidance, the training procedure is less affected by the redundancy present in the ground-truth captions. This process: i) reduces the risk of over-fitting and thus increases the generalization capability of the SD-RSIC; and ii) thus leads to a more effective learning procedure and more accurate RS image captions.

For the training of the proposed SD-RSIC approach, we use the stochastic gradient descent based optimization to maximize the log probability of the ground-truth captions for each (I_i, C_i) training instance using (1). After learning model parameters, the proposed approach automatically generates a caption for a new RS image. This process does not require any ground-truth caption since the summarization of ground-truth captions is only applied in the training stage. It is worth noting that finding the optimal word sequence is computationally expensive during the inference due to a large number of possible output sequences. Thus, we utilize the beam search algorithm with a beam size of four to acquire the best word sequence. This algorithm iteratively considers the set of best captions up to k^{th} time step to produce the captions for the time step of $k + 1$. However, it keeps only some of them depending on the beam size parameter value.

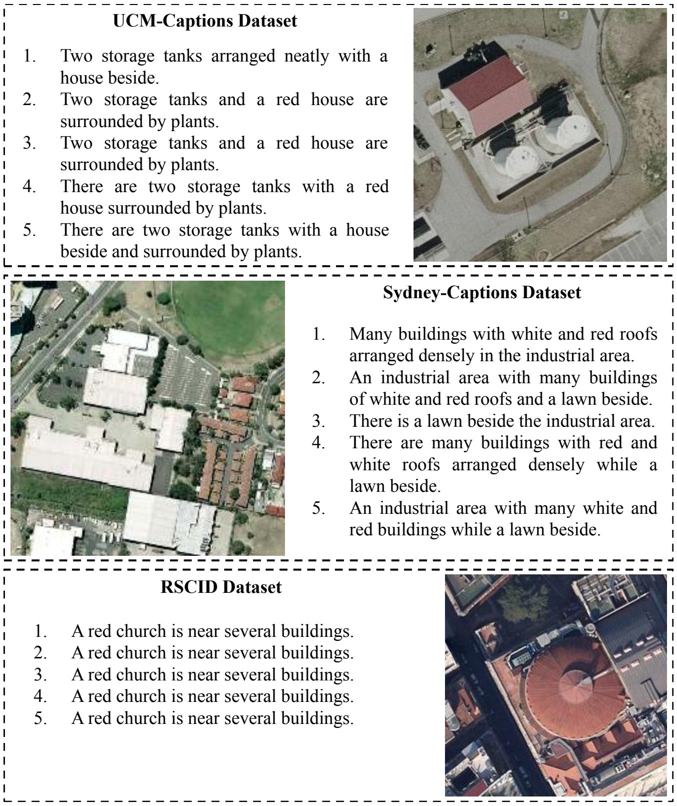
We would like to note that the summarized and standard captions can be semantically different (mainly due to possible differences between the lengths of the captions). However, since the adaptive weights of the words are iteratively learned, the proposed approach is not significantly affected by the possible semantic differences between the summarized and the standard captions. In detail, when the optimization process converges, the weights become more adapted to compensate the semantic differences between the summarized and the standard captions. In addition, iteratively learning the weights also forces generated weights and the standard captions to be in the same semantic order.

III. Datasets and Experimental Setup

In this section, we first describe the datasets used in the experiments and then present the experimental setup with the description of the baseline approaches.

A. Dataset Description

To evaluate our approach, we performed experiments on the Sydney-Captions [2], UCM-Captions [2] and RSICD [4] datasets. In addition, we utilized the Annotated Gigaword dataset [14], [15] for the second step of the proposed SD-RSIC approach.



UCM-Captions Dataset

1. Two storage tanks arranged neatly with a house beside.
2. Two storage tanks and a red house are surrounded by plants.
3. Two storage tanks and a red house are surrounded by plants.
4. There are two storage tanks with a red house surrounded by plants.
5. There are two storage tanks with a house beside and surrounded by plants.

Sydney-Captions Dataset

1. Many buildings with white and red roofs arranged densely in the industrial area.
2. An industrial area with many buildings of white and red roofs and a lawn beside.
3. There is a lawn beside the industrial area.
4. There are many buildings with red and white roofs arranged densely while a lawn beside.
5. An industrial area with many white and red buildings while a lawn beside.

RSCID Dataset

1. A red church is near several buildings.
2. A red church is near several buildings.
3. A red church is near several buildings.
4. A red church is near several buildings.
5. A red church is near several buildings.

Fig. 3. An example of RS images with their ground-truth captions selected from the UCM-Captions (top), the Sydney-Captions (middle) and the RSICD (bottom) datasets.

TABLE I
An Example of Article-Headline Pairs in the Annotated Gigaword dataset

Article	**Headline**
A fire on a freight shuttle in the channel tunnel on thursday forced an emergency rescue operation and the closure of the tunnel, officials said.	Fire closes channel tunnel
World oil prices rose in asian trade thursday as hurricane ike headed towards key energy facilities on the southern us coast, dealers said.	Oil prices up in asia on hurricane fears

The Sydney-Captions dataset includes 613 images, each of which has the size of 500×500 pixels with a spatial resolution of 0.5 meters. This dataset was built based on the Sydney scene classification dataset [16], which includes RS images annotated with one of the seven land-use classes. Each image in the Sydney-Captions dataset was annotated by the five captions, providing 3065 captions in total. The UCM-Captions dataset includes 2100 aerial images, each of which has a size of 256×256 pixels with a spatial resolution of one foot. This dataset is defined based on the UC Merced Land Use dataset [17], in which each image is associated with one of 21 land-use classes. Each image in the UCM-Captions dataset was annotated with five captions, resulting in 10500 captions in total. Although five captions per image are considered, captions belonging to the same classes are very similar in both datasets. Both the Sydney-Captions and the UCM-Captions datasets were initially built for scene classification problems with a small number of images. The RSICD is currently the largest RS image captioning dataset, including 10921 images in total with the size of 224×224 pixels with varying spatial resolutions. In this dataset, each image is described with a different number of captions [4]. In detail, 724 images have five different captions, 1495 images have four different captions, 2182 images have three different captions, 1667 images have two different captions and 5853 images have only one caption. As mentioned in [4], the number of captions was augmented in cases where images are described with less than five captions by randomly duplicating the existing captions. This leads to 54605 captions in the dataset. Fig. 3 shows an example of images and their captions for all considered RS image captioning datasets. The Annotated Gigaword dataset is a corpus of article-headline pairs that consists of nearly 10 million documents with a total of more than 4 billion words sourced from various news services. Instead of using the whole corpus, we follow the same removal and pre-processing steps presented in [18] that results in around 4 million articles. Table I shows an example of article-headline pairs in this dataset.

B. Experimental Setup

To perform the experiments, we split each considered dataset into training (80%), validation (10%) and test (10%) sets as suggested in the papers that the datasets were introduced ([2], [4]). All hyper-parameters were obtained based on the RS image captioning performance on the validation set. In the training sets of all datasets, there are five captions per image. Thus, we replicated each image five times to compose image-caption pairs of training. For the Annotated Gigaword dataset, we initially used the same training set splitting with [18] that results in 110,000 unique words, which is significantly higher than any vocabulary size within the RS captioning datasets. Then, we changed the vocabulary set of captioning datasets, since they do not contain all the words from the summarization vocabulary, and thus might miss several words when we summarize the five captions to one using the summarization model. Accordingly, we constructed a new common vocabulary set, which is used in all the steps of our approach. To this end, we selected 50000 words that include all the words from the Sydney-Captions, UCM-Captions and RSICD datasets and the list of most appearing words in the Annotated Gigaword dataset.

Before training our approach, we trained the pointer-generator network for summarization by following the same hyper-parameters presented in [13]. Then, we combine the pre-trained model with our approach. In addition, we also utilized the existing CNN models, which are pre-trained on the ImageNet for the feature extractor ϕ in the first step of the SD-RSIC. To select the CNN model for each dataset, ϕ is tested among the CNNs of the VGG [19], GoogleNet [20], InceptionV3 [21], ResNet [22] and DenseNet [23] models. We would like to note that we did not apply fine-tuning to the parameters of pre-trained models during the training of our approach. The extracted image features are mapped to the embedding space, whose dimension is the same as the word embedding dimension. In the experiments, the value of the embedding size W is varied as W = 128, 256, 512, 1024. However, for the selection of ϕ, the value of the embedding size is fixed to 512. In the first and third steps of our approach, we exploited the LSTM networks with W and 1 dimensional hidden states, respectively. We trained our approach with the learning of 10^{-3}, which decays by 20% if there are eight consecutive epochs without any improvement on the validation set performance. The training was conducted on NVIDIA Tesla V100 GPUs. To assess the effectiveness of the second and the third steps of the proposed approach, we considered a scenario for which these steps are neglected and only the first step of the proposed approach is applied. For this scenario, we randomly selected a single caption for each image in the training sets of all the considered datasets. It is denoted as Step 1 (Single Caption) in the experiments. To assess the effectiveness of the different steps of the proposed approach in terms of computational complexity, we provided the total number of parameters and floating-point operations associated to the different steps of the proposed approach.

In the experiments, we compared our approach with: 1) the cosine distance matching between the bag-of-words representation of image captions and the CNN features of images (which is denoted as BoW+CNN); 2) the cosine distance matching between the Deep Visual-Semantic Embedding (DeViSE) [24] of image captions and the CNN features of images (which is denoted as DeViSE+CNN); 3) the Collective Semantic Metric Learning Framework (CSMLF) [5]; and 4) the Neural Image Caption (NIC) [7]. RS image captioning accuracies of the BoW+CNN, DeViSE+CNN and CSMLF on each dataset were obtained in [5] by utilizing the ResNet model at the depth of 50 (ResNet50) as the feature extractor for RS images. Since the results were obtained by using the same sets with our approach, we did not repeat the corresponding experiments. For the NIC, which is one of the widely used state-of-the-art RS image captioning approaches, we applied the same CNN and caption generation procedure as the first step of our approach for each experiment of the NIC to fairly compare it with the proposed SD-RSIC approach.

Results of each experiment are provided in terms of four performance evaluation metrics: 1) the Bilingual Evaluation Understudy (BLEU) [25], 2) the Meteor Universal (METEOR) [26], 3) the Longest Common Subsequence-Based F-Measure of Recall-Oriented Understudy for Gisting Evaluation (ROUGE-L) [27] and 4) the Consensus-Based Image Description Evaluation (CIDEr) [28].

BLEU is not only the oldest but also the most well-known metric used for sentence similarity measurement. It measures the closeness of machine translation with one or more reference human translation according to numerical metrics that is proposed in [25]. It compares n-grams of machine generated captions with the n-grams of ground-truth captions and then counts the number of matches. Thus, the score is better if machine translation is closer to human translation. It is calculated by finding the geometric mean of n-gram precision scores as follows:

$$\text{BLEU-}n = \text{BP} \times e^{(\sum_{n=1}^{NB} w_n^B \log P_n^B)} \tag{6}$$

where P_n^B and w_n^B are the precision and weights of n-grams. It further applies brevity penalty BP for short sentences as follows:

$$\text{BP} = \begin{cases} 1 & if \quad l_c > l_r \\ e^{(1-l_r/l_c)} & if \quad l_c \leq l_r \end{cases} \tag{7}$$

TABLE II
IMAGE CAPTIONING PERFORMANCE ON THE SYDNEY-CAPTIONS DATASET WHEN USING DIFFERENT CNN MODELS FOR THE PROPOSED SD-RSIC APPROACH

Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr
VGG16	72.4	62.1	53.2	45.1	34.2	63.6	139.5
VGG19	73.4	63.1	55.2	48.7	34.8	64.1	160.3
GoogleNet	71.5	60.5	51.1	42.2	33.3	62.8	130.6
InceptionV3	73.3	62.6	54.5	47.7	35.1	62.9	143.9
ResNet34	73.0	62.9	54.4	46.8	34.3	63.7	137.6
ResNet50	71.6	59.2	49.1	39.8	32.0	61.6	108.7
ResNet101	76.1	66.6	58.6	51.7	36.6	65.7	169.0
ResNet152	73.3	61.9	51.7	42.5	31.8	62.0	114.6
DenseNet121	73.6	63.4	55.2	47.8	34.9	63.8	138.9
DenseNet169	73.0	63.2	54.6	46.7	34.1	62.9	140.2
DenseNet201	71.8	61.6	53.2	45.3	33.3	62.4	137.8

where l_c and l_r are the lengths of the candidate and ground-truth captions, respectively.

METEOR is based on word-to-word matching scores. For the multiple ground-truth captions, the score is calculated with respect to each caption and the best score is considered only. First, an F-Score (F^M) is calculated based on the word-to-word matching precision (P^M) and recall (R^M) scores as follows:

$$F^M = \frac{10 \times P^M \times R^M}{R^M + 9 \times P^M}. \tag{8}$$

Then, METEOR is calculated as follows:

$$\text{METEOR} = F^M \times (1 - \frac{0.5 \times |\text{Chunks}|}{|\text{Matched Words}|}) \tag{9}$$

where chunk is defined as a series of contiguous and identically ordered matches among the candidate and ground-truth captions.

ROUGE-L considers the longest common sub-sequence (LCS) between a pair of candidate and ground-truth captions. It is a type of F-Score based on the precision (P^L) and recall (R^L) scores of LCS results as follows:

$$\begin{aligned} R^L &= \frac{|LCS|}{l_r} \\ P^L &= \frac{|LCS|}{l_c} \\ \text{ROUGE-L} &= \frac{(1+\beta^2) \times R^L \times P^L}{R^L + \beta^2 \times P^L}. \end{aligned} \tag{10}$$

CIDEr considers a consensus of how often the n-grams in a candidate caption is present in ground-truth captions. It also considers the n-grams, which are not present in the ground-truth captions and should not be presented in the candidate caption [28]. To this end, it is calculated based on the Term Frequency Inverse Document Frequency (TF-IDF) weighting for each n-gram as follows:

$$\begin{aligned} \text{CIDEr}_n &= \frac{1}{m} \sum_j \frac{g^n(c_i^*) \cdot g^n(c_{i,j})}{||g^n(c_i^*)|| \, ||g^n(c_{i,j})||} \\ \text{CIDEr} &= \sum_{n=1}^{N} w_n^B \text{CIDEr}_n \end{aligned} \tag{11}$$

where c_i^* and $c_{i,j}$ are the candidate and ground-truth captions, respectively and g_n is a function that provides the vector of all n-grams of length n.

IV. EXPERIMENTAL RESULTS

We carried out different kinds of experiments in order to: 1) perform a sensitivity analysis; and 2) compare the effectiveness of the proposed SD-RSIC approach with the state-of-the-art image captioning approaches.

A. Sensitivity Analysis of the Proposed Approach

In this sub-section, we perform the sensitivity analysis of the proposed SD-RSIC approach in terms of: i) different CNN models utilized in the first step; ii) different embedding size used for image features and captions; iii) the effectiveness of the second and third steps; iv) the computational complexity associated to the different steps; and v) the sensitivity to zero-padding operation applied in the third step.

In the first set of trials, we analyzed the effect of different CNN models (the VGG model at the depths of 16 and 19 layers [VGG16, VGG19], the GoogleNet model, the InceptionV3 model, the ResNet model at the depths of 34, 50, 101 and 152 layers [ResNet34, ResNet50, ResNet101, ResNet152] and the DenseNet model at the depths of 121, 169 and 201 layers [DenseNet121, DenseNet169, DenseNet201]) in the first step of the proposed approach in terms of the image captioning performance. Table II shows the results for the Sydney-Captions dataset. By assessing the table, one can observe that the ResNet model at the depth of 101 layers leads to the highest scores under all metrics compared to the other CNNs. As an example, the ResNet101 provides almost 5% higher BLEU-1, more than 6% higher BLEU-2, almost 8% higher BLEU-3, more than 9% higher BLEU-4 and almost 3% higher ROUGE-L scores compared to the GoogleNet model. In detail, most of the CNN models (except ResNet101) achieve similar scores on the Sydney-Captions dataset under all metrics regardless of their depth. As an example, the VGG model at the lowest depth in considered CNNs (VGG16) provides less than 1% higher BLEU-1 and almost the same BLEU-4 scores compared to

TABLE III
Image Captioning Performance on the UCM-Captions Dataset When Using Different CNN Models for the Proposed SD-RSIC Approach

Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr
VGG16	74.8	66.4	59.8	53.8	39.0	69.5	213.2
VGG19	73.4	65.2	58.3	52.2	37.0	67.9	208.0
GoogleNet	74.6	65.3	58.3	52.6	37.8	67.5	214.9
InceptionV3	69.4	59.1	51.6	45.6	33.4	62.2	173.4
ResNet34	73.3	63.6	56.0	49.6	36.2	66.2	197.1
ResNet50	74.3	65.4	58.2	51.5	35.8	66.7	205.7
ResNet101	72.2	63.3	56.1	49.9	36.3	66.7	199.8
ResNet152	71.4	62.5	55.3	49.2	36.3	65.8	197.8
DenseNet121	72.6	63.1	55.6	49.1	35.7	65.8	196.7
DenseNet169	74.7	65.3	58.1	51.8	37.5	68.1	202.8
DenseNet201	73.1	63.5	56.2	49.8	35.3	65.3	195.5

TABLE IV
Image Captioning Performance on the RSICD Dataset When Using Different CNN Models for the Proposed SD-RSIC Approach

Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr
VGG16	64.5	47.1	36.4	29.4	24.9	51.9	77.5
VGG19	64.8	47.3	36.5	29.3	25.1	51.8	76.5
GoogleNet	63.7	45.7	34.9	28.0	24.4	51.0	73.6
InceptionV3	62.8	45.0	34.3	27.3	23.8	50.7	71.8
ResNet34	63.5	46.0	35.2	28.2	24.2	51.1	73.8
ResNet50	64.9	47.2	36.5	29.5	24.9	52.0	77.3
ResNet101	63.1	45.8	35.4	28.7	24.1	51.2	75.4
ResNet152	64.4	47.4	36.9	30.0	24.9	52.3	79.4
DenseNet121	63.2	46.2	35.8	28.8	24.5	51.4	75.7
DenseNet169	64.3	46.5	35.7	28.5	24.4	51.2	75.9
DenseNet201	62.5	45.7	35.1	28.1	24.0	51.3	74.2

the DenseNet model at the highest depth among all CNNs (DenseNet201). The image captioning results for the UCM-Captions dataset is given in Table III. By analyzing the table, one can see that the VGG model at the depth of 16 layers (VGG16) provides the highest scores under all metrics except CIDEr. As an example, the VGG16 provides more than 5% higher BLEU-1, more than 8% higher BLEU-4 and more than 7% higher ROUGE-L scores compared to the InceptionV3. However, only under CIDEr metric, the VGG16 leads to less than 2% lower score compared to the highest score obtained by the GoogleNet model. In detail, the InceptionV3 provides the lowest scores under all metrics. As an example, it provides more than 5% lower BLEU-1 and almost 6 lower ROUGE-L scores compared to the DenseNet169. These results show that almost all CNN models (except the InceptionV3) achieve similar scores on the UCM-Captions dataset. This supports our conclusion on the Sydney-Captions dataset. In greater details, increasing the depths of the ResNet and DenseNet models up to some extent achieves slightly higher metric scores compared to those at the lowest depth. However, further increasing their depths do not provide the highest scores. As an example, the ResNet model at the depth of 152 leads to the lowest score under most of the metrics compared to the other ResNet CNNs. Table IV shows the results for the RSICD dataset. By analyzing the table, one can observe that the ResNet model at the depth of 152 layers provides the highest scores under most of the metrics compared to the other CNNs. As an example, the ResNet152 achieves more than 2% higher BLEU-3 and BLUE-4 scores and almost 8% higher CIDEr score compared to the InceptionV3. It also achieves almost the same BLEU-1 and METEOR scores with the VGG19 and the ResNet50, which provide the highest score in BLEU-1 and METEOR metrics, respectively. In detail, the VGG model (which has the shallowest CNNs compared to the others) leads to higher scores under most of the metrics compared to the DenseNet model. As an example, the VGG model at the depth of 19 layers achieves more than 2% BLEU-1 and CIDEr scores compared to the DenseNet201, which has the highest depth in considered CNNs. These results show that accuracies obtained by most of the CNNs are, again, similar to each other.

The sensitivity analysis for different CNN models used in the first step shows that utilizing different models does not significantly affect the RS image captioning performance of our approach. However, the proper selection of a CNN model and its depth can improve the performance of the SD-RSIC. Accordingly, we utilized the ResNet101, VGG16 and

TABLE V
RESULTS OBTAINED BY THE PROPOSED SD-RSIC WHEN USING DIFFERENT EMBEDDING SIZES

Dataset	Embedding Size (W)	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr
Sydney-Captions	128	72.1	61.5	52.1	42.9	32.4	61.9	128.7
	256	73.0	62.7	54.0	45.6	33.5	62.7	140.8
	512	**76.1**	**66.6**	**58.6**	**51.7**	**36.6**	**65.7**	**169.0**
	1024	70.9	59.5	50.6	42.9	35.1	63.1	126.7
UCM-Captions	128	71.6	63.1	56.0	50.0	36.2	66.0	199.9
	256	74.2	65.7	58.7	52.3	38.1	68.4	202.8
	512	**74.8**	**66.4**	**59.8**	**53.8**	39.0	**69.5**	213.2
	1024	74.6	66.2	59.4	53.7	**39.2**	69.1	**213.9**
RSICD	128	61.0	43.2	33.1	26.5	23.0	49.4	66.0
	256	63.4	45.8	35.3	28.3	24.4	51.0	73.6
	512	64.4	**47.4**	**36.9**	**30.0**	24.9	**52.3**	**79.4**
	1024	**64.7**	46.8	35.9	28.8	**25.0**	51.5	78.7

TABLE VI
RESULTS OBTAINED BY THE PROPOSED SD-RSIC ON THE COMPLETE SET OF CAPTIONS AND ITS FIRST STEP ON A SINGLE CAPTION FOR EACH IMAGE

Dataset	Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr
Sydney-Captions	Step 1 (Single Caption)	66.5	55.3	47.0	39.9	31.2	60.6	109.9
	SD-RSIC	**76.1**	**66.6**	**58.6**	**51.7**	**36.6**	**65.7**	**169.0**
UCM-Captions	Step 1 (Single Caption)	70.0	60.2	52.6	46.0	33.2	63.6	177.4
	SD-RSIC	**74.8**	**66.4**	**59.8**	**53.8**	**39.0**	**69.5**	**213.2**
RSICD	Step 1 (Single Caption)	62.9	45.5	35.2	28.5	24.4	51.0	74.0
	SD-RSIC	**64.4**	**47.4**	**36.9**	**30.0**	**24.9**	**52.3**	**79.4**

ResNet152 for the rest of the experiments on Sydney-Captions, UCM-Captions and RSICD datasets, respectively.

In the second set of trials, we assessed the effect of the embedding size W used in the proposed approach. Table V shows the image captioning performances under the different sizes of embedding space for all the considered datasets. By assessing the table, one can observe that increasing the value of W up to some extent provides significantly higher scores under all the metrics compared to those obtained by using the lowest value of W. As an example, the proposed approach with the $W = 512$ provides more than 6% higher BLEU-3, more than 4% higher METEOR and almost 4% higher ROUGE-L scores compared to that of the $W = 128$ for the Sydney-Captions dataset. This is due to the fact that increasing the value of W allows to preserve more detailed information than the lowest dimensional embeddings for both image features and image captions. However, selecting a very high value of W (e.g., $W = 1024$) does not further improve the information preserving capability of the proposed SD-RSIC. As an example, the proposed approach with the $W = 512$ leads to almost 1% higher BLEU-2, BLEU-3, BLUE-4, ROUGE-L, CIDEr scores and almost the same BLEU-1 and METEOR scores compared to that of the $W = 1024$ for the RSICD dataset. It is worth noting that increasing the value of W also increases the computational complexity of the proposed approach. Accordingly, we selected the value of W as 512 for the rest of the experiments on the considered datasets.

In the third set of trials, we analyzed the effectiveness of the second and third steps of the proposed approach.

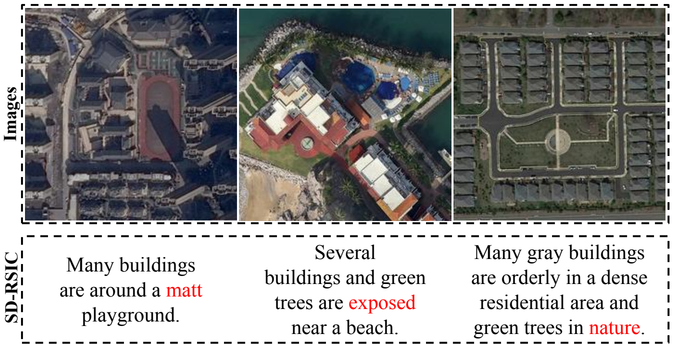

Images
SD-RSIC
Many buildings are around a matt playground.
Several buildings and green trees are exposed near a beach.
Many gray buildings are orderly in a dense residential area and green trees in nature.

Fig. 4. An example of the RSICD images with the generated captions by the SD-RSIC. The words, which are only from Annotated Gigaword dataset, are in red.

Table VI shows the image captioning performances obtained when: i) the first step of the proposed approach is applied by considering only a single caption for each image (i.e., Step 1 (Single Caption)); and ii) the proposed SD-RSIC approach is applied by considering the complete set of captions for all the considered datasets. By analyzing the table, one can see that our proposed approach results in significantly better performances with respect to the Step 1 (Single Caption) for all datasets. As an example, the proposed approach provides almost 6% higher BLUE-1 and more than 5% higher ROUGE-L scores for the Sydney-Captions dataset, while providing more than 7% higher BLEU-3 and BLEU-4 scores for the UCM-Captions dataset compared to the Step 1 (Single Cap-

TABLE VII
NUMBER OF REQUIRED MODEL PARAMETERS (NP) AND FLOATING-POINT OPERATIONS (FLOPs) ASSOCIATED TO THE DIFFERENT STEPS OF THE PROPOSED APPROACH (THE SYDNEY-CAPTIONS DATASET)

Steps of the Proposed Approach			NP	FLOPs
1^{st}	2^{nd}	3^{rd}	$(\times 10^6)$	$(\times 10^9)$
✓	✗	✗	43.55	7.83
✓	✓	✗	77.93	8.62
✓	✓	✓	77.94	8.63

TABLE VIII
NUMBER OF REQUIRED MODEL PARAMETERS (NP) AND FLOATING-POINT OPERATIONS (FLOPs) ASSOCIATED TO THE DIFFERENT STEPS OF THE PROPOSED APPROACH (THE UCM-CAPTIONS DATASET)

Steps of the Proposed Approach			NP	FLOPs
1^{st}	2^{nd}	3^{rd}	$(\times 10^6)$	$(\times 10^9)$
✓	✗	✗	119.57	15.46
✓	✓	✗	153.96	16.26
✓	✓	✓	153.97	16.27

tion). This is due to the fact that the second and the third steps of the SD-RSIC significantly addresses the problems related to redundancy present in ground-truth captions, and thus improves the image captioning performances. Fig. 4 shows an example of RSICD images with the generated captions by the SD-RSIC. By assessing the figure, one can observe that the SD-RSIC provides the enriched vocabulary compared to the original vocabulary of captioning datasets. As an example, the words for describing the objects on the ground (e.g., matt, nature) are from the Annotated Gigaword dataset and not included in the original vocabulary of the RSICD dataset. The enriched vocabulary of the proposed approach leads to more detailed captions for semantically complex RS images. These results show that the SD-RSIC overcomes the limitations of redundant information in ground-truth captions, while providing the detailed language semantics due to its second and third steps.

In the fourth set of trials, we assessed the computational complexity associated to the different steps of the proposed approach. Table VII, VIII and IX show the number of model parameters and the floating-point operations (FLOPs) for the Sydney-Captions, the UCM-Captions and the RSICD datasets, respectively. By analyzing the tables, one can observe that the selection of a CNN model for the first step at the proposed approach is one of the most important factors affecting the overall computational complexity. As an example, the total number of FLOPs for the UCM-Captions dataset is twice as large as that of the Sydney-Captions dataset due to the different CNNs used for these datasets. It is worth noting that this can affect almost all deep learning based image captioning approaches. In addition, the third step of the proposed approach does not significantly affect the computational complexity. As an example, when the third step is included within the proposed approach, the amount of increase in the total number of parameters is less than 1%. In greater details, the amount of increase in the FLOPs is significantly less than that in the number of parameters when the second step is included within the proposed approach. These results show that the second step of the proposed approach does not significantly increase the computational time during training.

TABLE IX
NUMBER OF REQUIRED MODEL PARAMETERS (NP) AND FLOATING-POINT OPERATIONS (FLOPs) ASSOCIATED TO THE DIFFERENT STEPS OF THE PROPOSED APPROACH (THE RSICD DATASET)

Steps of the Proposed Approach			NP	FLOPs
1^{st}	2^{nd}	3^{rd}	$(\times 10^6)$	$(\times 10^9)$
✓	✗	✗	59.18	11.55
✓	✓	✗	93.57	12.35
✓	✓	✓	93.58	12.35

TABLE X
THE AVERAGE RATE OF ZERO-PADDING OPERATION APPLIED IN THE THIRD STEP OF THE PROPOSED SD-RSIC

Sydney-Captions	UCM-Captions	RSICD
42.5%	32.5%	33.9%

In the fifth set of trials, we analyzed the effect of zero-padding operation applied to the summarized captions in the third step of the proposed approach. Table X shows the average rate of zero-padding operation during training for the considered datasets. By assessing the table, one can observe that the zero-padding operation is not often applied to the summarized captions. As an example, for the RSICD dataset, it is applied once in three times on average. To this end, integration of summarized captions with standard captions is not dominated by standard captions. If zero-padding operation is applied frequently, final caption generation may mostly relies on the standard captions. This condition can be eliminated by: 1) using other summarization approaches, which are capable of producing longer sentences, in the second step of the proposed approach; or 2) changing the pre-training of the pointer-generator model (which is utilized in the second step) with different datasets to produce longer sentences.

B. Comparison of the Proposed Approach with the State-of-the-Art Approaches

In the sixth set of trials, we assessed the effectiveness of the proposed SD-RSIC approach compared to the state-of-the art RS image captioning approaches, which are: the BoW+CNN [5], the DeViSE+CNN [5], the CCSMLF [5] and the NIC [7]. Table XI, XII and XIII show the corresponding image captioning performances on the Sydney-Captions, UCM-Captions and RSICD datasets, respectively. By analyzing the tables, one can observe that the proposed SD-RSIC approach leads to the highest scores under all metrics for all datasets. As an example, the SD-RSIC outperforms the CSMLF by almost 32% in BLEU-1 and more than 30% in BLEU-3 for the Sydney-Captions dataset, almost 45% in BLEU-2 and more than 44% in BLEU-4 for the UCM-Captions dataset, and almost 8% ROUGE-L and more than 26% in CIDEr for the RSICD dataset. Similar behaviors

TABLE XI
Results Obtained by the BoW+CNN, DeViSE+CNN, CCSMLF, NIC and the Proposed SD-RSIC (The Sydney-Captions Dataset)

Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr
BoW+CNN [5]	62.3	47.9	39.0	32.9	24.5	51.7	128.3
DeViSE+CNN [5]	64.2	51.5	43.5	38.1	27.0	56.6	139.2
CSMLF [5]	44.4	33.7	28.2	24.1	15.8	40.2	93.8
NIC [7]	70.7	59.1	50.3	42.5	32.0	60.6	127.7
SD-RSIC	**76.1**	**66.6**	**58.6**	**51.7**	**36.6**	**65.7**	**169.0**

TABLE XII
Results Obtained by the BoW+CNN, DeViSE+CNN, CCSMLF, NIC and the Proposed SD-RSIC (The UCM-Captions Dataset)

Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr
BoW+CNN [5]	40.6	25.5	18.4	14.4	14.4	36.6	41.6
DeViSE+CNN [5]	37.0	17.4	9.8	6.0	9.8	29.7	9.7
CSMLF [5]	38.7	21.5	12.5	9.2	9.5	36.0	37.0
NIC [7]	72.6	64.1	57.5	51.7	37.4	67.3	200.6
SD-RSIC	**74.8**	**66.4**	**59.8**	**53.8**	**39.0**	**69.5**	**213.2**

TABLE XIII
Results Obtained by the BoW+CNN, DeViSE+CNN, CCSMLF, NIC and the Proposed SD-RSIC (The RSICD Dataset)

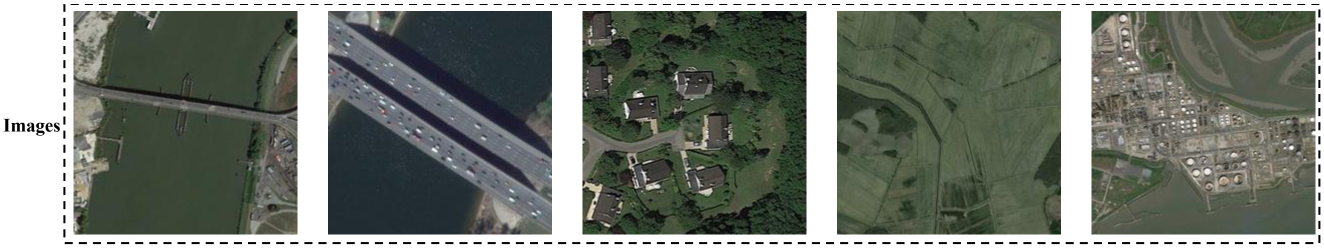
Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr
BoW+CNN [5]	29.7	11.3	5.8	3.4	9.6	25.1	12.9
DeViSE+CNN [5]	30.7	11.4	5.6	3.1	9.7	25.6	12.4
CSMLF [5]	57.6	38.6	28.3	22.2	21.3	44.6	53.0
NIC [7]	62.9	46.0	35.8	29.1	24.3	51.5	76.0
SD-RSIC	**64.4**	**47.4**	**36.9**	**30.0**	**24.9**	**52.3**	**79.4**

are also observed while comparing the BoW+CNN and DeViSE+CNN with our approach under different metrics. This shows that modeling image captions based on the joint characterization of language and RS image semantics significantly improves the RS image captioning performance compared to separately describing their semantics and applying matching. In addition, the proposed SD-RSIC approach outperforms the well-known automatic image captioning approach (the NIC) by almost 6% in BLEU-1, more than 9% in BLEU-4 and more than 5% in ROUGE-L for the Sydney-Captions dataset, more than 2% in BLEU-2 and BLUE-3 for the UCM-Captions dataset, and more than 3% in CIDEr and almost 2% in BLEU-1 for the RSICD dataset. This is due to the second and the third steps of the SD-RSIC that integrate the summarization of ground-truth image captions into the widely used CNN and LSTM based encoder-decoder strategy. This shows that the SD-RSIC is capable of: i) eliminating the redundant information in the training set; ii) increasing the generalization capability of the whole neural network; and iii) improving the vocabulary of training sets compared to the existing approaches.

Fig. 5 shows an example of RSICD images with their ground-truth captions and the generated captions by the NIC and the SD-RSIC. By assessing the figure, one can observe that the SD-RSIC provides more accurate image captions to describe the complex semantic content of RS images in the grammatically correct form compared to the NIC. As an example, in the first image, the SD-RSIC is able to describe the green trees near the bridge while this information is not captured by the NIC. In addition to the first image, the SD-RSIC is capable of describing the type of the residential area in the third image that is not characterized in the caption of the NIC. In greater details, for the first and last images, the SD-RSIC is capable of generating the single caption, which accurately describes most of the information associated with the semantic content of the image in a grammatically correct form. However, the NIC provides grammatically incorrect sentences, wrong information in the captions and phrases instead of sentences for the same images. This shows that the SD-RSIC can accurately describe the complex semantic content of RS images with single grammatically correct caption. We observed the similar behaviours for the other approaches and datasets. Thus, qualitative results further confirm that the proposed SD-RSIC approach achieves promising RS image captioning performance.

V. Conclusion

In this paper, we have introduced a novel Summarization Driven Remote Sensing Image Captioning (SD-RSIC) approach. The proposed SD-RSIC approach consists of three



SD-RSIC	A bridge is on a river with some green trees in two sides.	Many cars are on a bridge over a river with many green trees in two sides of it.	Some buildings and many green trees in a medium residential area.	It is a piece of green meadow.	Many storage tanks are in a factory near a river.
NIC	A bridge is over a river in a bridge over it.	Many cars are on a bridge over a parking lots.	Many green trees and a swimming pool are in a resort.	It is a large piece of green mountain.	Many green trees and green and parking.
Ground-truth Captions	1. On either side of the river there are many grey roofed houses. 2. On either side of the river there are many grey roofed houses. 3. On either side of the river there are many grey roofed houses. 4. There is a magnificent bridge over the river. 5. There is a magnificent bridge over the river.	1. There are many cars running on the road. 2. There are many cars running on the road. 3. There are many cars running on the road. 4. There are many tall trees planted on both sides of the river. 5. There are many tall trees planted on both sides of the river.	1. The residential with black villages is in the center of the forest. 2. The residential with black villages is in the center of the forest. 3. The residential with black villages is in the center of the forest. 4. This lush woods is surrounding the peaceful neighborhood with roads passes by. 5. Several buildings and many green trees are in a residential area.	1. A furcate road separates the grass green farmland. 2. A furcate road separates the grass green farmland. 3. The green farmland is divided by a furcate road. 4. It is a green farmland with several curved roads through it . 5. Many pieces of green farmlands are together.	1. There is a factory beside the river. 2. There is a factory beside the river. 3. There is a factory beside the river. 4. There are many storage tanks in the factory. 5. There are many storage tanks in the factory.

Fig. 5. An example of the RSICD images with their five ground-truth captions and the generated captions by the NIC and the SD-RSIC.

main steps. The first step generates the standard RS image captions by jointly exploiting CNNs and LSTMs. The second step summarizes all ground-truth captions into a single caption by using a sequence-to-sequence deep learning model. Third step automatically computes the adaptive weights for combining the standard captions with summarized captions, relying on the semantic content of RS images based on their image level features. Experimental results obtained on the existing RS image captioning datasets show the effectiveness of the proposed SD-RSIC approach over the state-of-the-art approaches. The main reasons for the success of our proposed SD-RSIC approach are summarized as follows:

1) Due to the summarization of ground-truth captions in the second step, the SD-RSIC eliminates the redundant information (occured because of the repetitive as well as highly similar captions) present in the RS image captioning datasets.
2) Due to the use of the summarization model, which is trained on large text corpora in the second step, the SD-RSIC significantly enriches the image captioning vocabulary in terms of the number and variety of words, resulting in more accurate image captions for complex scenarios.
3) Due to the adaptive weights among the standard and summarized captions provided in the third step, which allows effective integration of the condensed (summarized) information of ground-truth captions with standard captions, the SD-RSIC reduces the risk of overfitting during training and increases the generalization capability of the proposed DNN.

It is worth noting that an attention strategy that finds the most informative regions of RS images in terms of both the generation of standard captions and the integration of summarized captions can further improve the performance of the proposed approach. To this end, any attention strategy presented in the literature can be directly integrated within the proposed approach. We would like to point out that the existing image captioning metrics evaluate the accuracy of the automatically generated image captions by computing the word similarities of these captions with those of the ground truth captions (generated by human experts). These metrics do not compare the actual meaning of the generated and ground truth captions. As a future development of this work we plan to study on defining a new image captioning metric that can intrinsically address this issue. In addition, we also plan to improve the second step of the SD-RSIC by including different: i) summarization approaches (e.g., [29]); and ii) summarization datasets (e.g., the DUC 2004).

References

[1] G. Hoxha, F. Melgani, and B. Demir, "Toward remote sensing image retrieval under a deep image captioning perspective," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 13, pp. 4462–4475, 2020.

[2] B. Qu, X. Li, D. Tao, and X. Lu, "Deep semantic understanding of high resolution remote sensing image," in *Proc. Intl. Conf. Comput. Inf. Telecommunication Syst.*, Jul. 2016, pp. 1–5.

[3] Z. Shi and Z. Zou, "Can a machine generate humanlike language descriptions for a remote sensing image?" *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 6, pp. 3623–3634, Jun. 2017.

[4] X. Lu, B. Wang, X. Zheng, and X. Li, "Exploring models and data for remote sensing image caption generation," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 4, pp. 2183–2195, Dec. 2017.

[5] B. Wang, X. Lu, X. Zheng, and X. Li, "Semantic descriptions of high-resolution remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 8, pp. 1–5, Aug. 2019.

[6] X. Zhang, X. Wang, X. Tang, H. Zhou, and C. Li, "Description generation for remote sensing images using attribute attention mechanism," *Remote Sens.*, vol. 11, no. 6, Mar. 2019.

[7] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, June 2015, pp. 3156–3164.

[8] K. Xu, J. L. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. Intl. Conf. Mach. Learn.*, Jul. 2015, pp. 2048–2057.

[9] N. Yu, X. Hu, B. Song, J. Yang, and J. Zhang, "Topic-oriented image captioning based on order-embedding," *IEEE Trans. Image Process.*, vol. 28, no. 6, pp. 2743–2754, Jun. 2019.

[10] M. Z. Hossain, F. Sohel, M. F. Shiratuddin, and H. Laga, "A comprehensive survey of deep learning for image captioning," *ACM Comput. Surv.*, vol. 51, no. 6, pp. 118:1–118:36, Feb. 2019.

[11] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

[12] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: Continual prediction with LSTM," *Neural Comput.*, vol. 12, no. 10, pp. 2451–2471, Oct. 2000.

[13] A. See, P. Liu, and C. Manning, "Get to the point: Summarization with pointer-generator networks," in *Proc. Annu. Meet. Assoc. Comput. Linguistics*, 2017, pp. 1073–1083.

[14] D. Graff, J. Kong, K. Chen, and K. Maeda, *English gigaword*. Linguistic Data Consortium, Philadelphia, 2003.

[15] C. Napoles, M. Gormley, and B. Van, Durme, "Annotated gigaword," in *Proc. Joint Workshop Automatic Knowledge Base Construction and Web-scale Knowledge Extraction*, Jun. 2012, pp. 95–100.

[16] F. Zhang, B. Du, and L. Zhang, "Saliency-guided unsupervised feature learning for scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 4, pp. 2175–2184, Apr. 2015.

[17] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proc. Intl. Conf. Adv. Geogr. Inf. Syst.*, Nov. 2010, pp. 270–279.

[18] A. M. Rush, S. Chopra, and J. Weston, "A neural attention model for abstractive sentence summarization," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Sep. 2015, pp. 379–389.

[19] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Intl. Conf. Learn. Represent.*, May 2015, pp. 1–14.

[20] C. Szegedy, Wei Liu, Yangqing 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 Recog.*, Jun. 2015, pp. 1–9.

[21] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2016, pp. 2818–2826.

[22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2016, pp. 770–778.

[23] G. Huang, Z. Liu, L. v. d. Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jul. 2017, pp. 2261–2269.

[24] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, M. Ranzato, and T. Mikolov, "DeViSE: A deep visual-semantic embedding model," in *Proc. Adv. Neural Inf. Process. Syst.*, Dec. 2013, pp. 2121–2129.

[25] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. Annu. Meet. Assoc. Comput. Linguistics*, Jul. 2002, pp. 311–318.

[26] M. "Denkowski and A. Lavie, "Meteor universal: Language specific translation evaluation for any target language," in *Proc. Workshop Statistical Mach. Translation*, Jun. 2014, pp. 376–380.

[27] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Workshop on Text Summarization Branches Out: Proc. Annu. Meet. Assoc. Comput. Linguistics*, Jul. 2004, pp. 74–81.

[28] R. Vedantam, C. L. Zitnick, and D. Parikh, "CIDEr: Consensus-based image description evaluation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2015, pp. 4566–4575.

[29] K. Yao, L. Zhang, D. Du, T. Luo, L. Tao, and Y. Wu, "Dual encoding for abstractive text summarization," *IEEE Trans. Cybernetics*, vol. 50, no. 3, pp. 985–996, 2020.

Gencer Sumbul received his B.S. degree in Computer Engineering from Bilkent University, Ankara, Turkey in 2015 and the M.S. degree in Computer Engineering from Bilkent University in 2018. He is currently a research associate in the Remote Sensing Image Analysis (RSiM) group and pursuing the Ph.D. degree at the Faculty of Electrical Engineering and Computer Science, Technische Universität Berlin, Germany since 2019. His research interests include computer vision, pattern recognition and machine learning, with special interest in deep learning, large-scale image understanding and remote sensing. He is a referee for journals such as the IEEE Transactions on Geoscience and Remote Sensing, the IEEE Access, the IEEE Geoscience and Remote Sensing Letters, the ISPRS Journal of Photogrammetry and Remote Sensing and international conferences such as European Conference on Computer Vision and IEEE International Geoscience and Remote Sensing Symposium.

Sonali Nayak received the B.S. degree in Computer Science from International Institute of Information Technology Bhubaneswar, India in 2013 and the M.S. degree also in Computer Science from TU Berlin, Germany in 2019. Her research areas include remote sensing image analysis, natural language processing and machine learning.

Begüm Demir (S'06-M'11-SM'16) received the B.S., M.Sc., and Ph.D. degrees in electronic and telecommunication engineering from Kocaeli University, Kocaeli, Turkey, in 2005, 2007, and 2010, respectively.

She is currently a Full Professor and head of the Remote Sensing Image Analysis (RSiM) group at the Faculty of Electrical Engineering and Computer Science, Technische Universität Berlin, Germany since 2018. Before starting at TU Berlin, she was an Associate Professor at the Department of Computer Science and Information Engineering, University of Trento, Italy. Her research activities lie at the intersection of machine learning, remote sensing and signal processing. Specifically, she performs research on developing innovative methods for addressing a wide range of scientific problems in the area of remote sensing for Earth observation. She was a recipient of a Starting Grant from the European Research Council (ERC) with the project "BigEarth-Accurate and Scalable Processing of Big Data in Earth Observation" in 2017, and the "2018 Early Career Award" presented by the IEEE Geoscience and Remote Sensing Society. Dr. Demir is a senior member of IEEE since 2016.

Dr. Demir is a Scientific Committee member of several international conferences and workshops, such as: Conference on Content-Based Multimedia Indexing, Conference on Big Data from Space, Living Planet Symposium, International Joint Urban Remote Sensing Event, SPIE International Conference on Signal and Image Processing for Remote Sensing, Machine Learning for Earth Observation Workshop organized within the ECML/PKDD. She is a referee for several journals such as the PROCEEDINGS OF THE IEEE, the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, the IEEE GEOSCIENCE AND REMOTE SENSING LETTERS, the IEEE TRANSACTIONS ON IMAGE PROCESSING, Pattern Recognition, the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, the IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING, the International Journal of Remote Sensing), and several international conferences. Currently she is an Associate Editor for the IEEE GEOSCIENCE AND REMOTE SENSING LETTERS and MDPI Remote Sensing.