

Received 11 October 2022, accepted 29 October 2022, date of publication 4 November 2022, date of current version 10 November 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3219879

RESEARCH ARTICLE

Federated Onboard-Ground Station Computing With Weakly Supervised Cascading Pyramid Attention Network for Satellite Image Analysis

TAEWOO KIM^{ID}, MINSU JEON^{ID}, CHANGHA LEE^{ID},
JUNSOO KIM^{ID}, (Graduate Student Member, IEEE), GEONWOO KO, (Member, IEEE),
JOO-YOUNG KIM^{ID}, (Senior Member, IEEE), AND CHAN-HYUN YOUN^{ID}, (Senior Member, IEEE)

School of Electrical Engineering, Korea Advanced Institute of Science and Technology, Daejeon 34141, South Korea

Corresponding author: Chan-Hyun Youn (chyoun@kaist.ac.kr)

This work was supported in part by Samsung Electronics Co., Ltd (IO201210-07976-01), in part by Electronics and Telecommunications Research Institute(ETRI) grant funded by the Korean government.[22ZS1100, Core Technology Research for Self-Improving Integrated Artificial Intelligence System], and in part by Institute for Information & Communications Technology Promotion (IITP) grant funded by the Korea government (MSIT) (No.2020-0-00537, Development of 5G based low latency device – edge cloud interaction technology).

ABSTRACT With advances in NanoSat (CubeSat) and high-resolution sensors, the amount of raw data to be analyzed by human supervisors has been explosively increasing for satellite image analysis. To reduce the raw data, the satellite onboard AI processing with low-power COTS (Commercial, Off-The-Shelf) HW has emerged from a real satellite mission. It filters the useless data (e.g. cloudy images) that is worthless to supervisors, achieving efficient satellite-ground station communication. In the application for complex object recognition, however, additional explanation is required for the reliability of the AI prediction due to its low performance. Although various explainable AI (XAI) methods for providing human-interpretable explanation have been studied, the pyramid architecture in a deep network leads to the background bias problem which visual explanation only focuses on the background context around the object. Missing the small objects in a tiny region leads to poor explainability although the AI model corrects the object class. To resolve the problems, we propose a novel federated onboard-ground station (FOGS) computing with Cascading Pyramid Attention Network (CPANet) for reliable onboard XAI in object recognition. We present an XAI architecture with a cascading attention mechanism for mitigating the background bias for the onboard processing. By exploiting the localization ability in pyramid feature blocks, we can extract high-quality visual explanation covering the both semantic and small contexts of an object. For enhancing visual explainability of complex satellite images, we also describe a novel computing federation with the ground station and supervisors. In the ground station, active learning-based sample selection and attention refinement scheme with a simple feedback method are conducted to achieve the robustness of explanation and efficient supervisor's annotation cost, simultaneously. Experiments on various datasets show that the proposed system improves task accuracy in object recognition and accurate visual explanation detecting small contexts of objects even in a peripheral region. Then, our attention refinement mechanism demonstrates that the inconsistent explanation can be efficiently resolved only with very simple selection-based feedback.

INDEX TERMS XAI, visual explanation, satellite image analysis, human-in-the-loop.

I. INTRODUCTION

The modern small satellites including CubeSat are becoming interesting technologies in the space industry. As the

The associate editor coordinating the review of this manuscript and approving it for publication was Yang Li^{ID}.

revisit period of the acquired raw images is getting shorter, complex applications (e.g. object tracking, detection, etc.) in object recognition are interested in a research field. However, the massive amount of raw data makes it difficult for the limited supervisors to analyze an inspection of all image patches over a broad area (covering >km in a raw image). To assist the

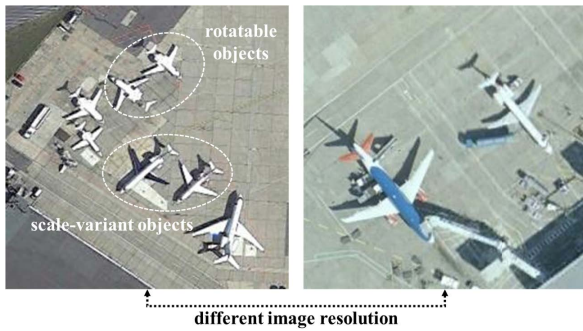


FIGURE 1. Characteristics of satellite images make it difficult for a DL model to infer accurate prediction and visual explanation. Each image may contain scale-variant and rotation-equivariant objects with the same category (airplane). And, training dataset includes multiple image resolutions due to the optical sensor, altitude, etc.

supervisors, the DL-based image analysis has emerged [1], [2], [3], [4], [5], [6]. By providing the prediction results for object recognition, the DL model enables to efficiently analyze the large-scale data with high accuracy. Furthermore, due to the advent of low-power AI computing HW such as visual processing unit (VPU) and field programmable gate array (FPGA), a DL-based satellite onboard system has recently been introduced in a satellite image analysis. The onboard computing is important for efficient satellite-ground station computing by filtering unnecessary images in advance. Typically, the onboard DL system, CloudScout [7], filters cloudy images having no information to analyze using simple binary classification. In complex applications, however, the prediction results of a DL model still remain ambiguous, especially in object recognition. The false negative error is a challenging issue for the reliability of the onboard AI system. Explainable AI (XAI) technique has the key to a reliable AI-based system by providing *visual explanation* of the prediction for a black-box DL model. It represents the form of saliency maps highlighting the pixels that are important for the DL model to predict the class of a target object. These human-interpretable results enable humans to expect the model behavior for other samples and sometimes retrain (refine) the model for improving performance.

Satellite images have low resolution (e.g., $> 0.5m^2$ per one pixel) compared with other object recognition images due to the long distance between the optical sensors and target regions. Fig. 1 shows the characteristics of top-view images: rotation-equivariant and scale-variant objects. It also contains different quality images due to the sensor resolution and environmental reasons (e.g. cloudiness, sunlight, etc.). That makes it difficult for DL models to predict the class of a target object as well as visual explanation accurately. In the satellite images, the background has a large portion of the entire image. Accordingly, if a certain class is highly involved with the background (e.g., ship – ocean), the model could be trained to recognize objects for the class based on the only background rather than the object's own characteristics, called the *background bias problem* [8], [9]. In particular, this problem is critical due to small object sizes and a

high correlation between background and object. The wrong visual explanation highlighting only on background context reduces the reliability of the mission-critical application in the satellite system. In addition, due to the high computation overhead of the XAI model and no supervision for output visual explanation, the onboard computing system itself cannot refine the model for explainability, which supervisors and rich computing resources in the ground station are required.

To resolve these problems, we propose a novel federated onboard-ground station (FOGS) computing system for satellite image analysis. Our system deals with the reliability of complex object recognition applications in satellite images, which has not been addressed in the conventional onboard AI system [7]. We are attempting to handle this problem through the development of a novel satellite image analysis system that cooperates with onboard-ground station and supervisors. Different from the conventional onboard AI system [7], we build a sustainable onboard XAI-based analysis system by iteratively updating the refinement of the analysis model between the onboard and ground station. We describe a novel XAI model, *cascading pyramid attention network (CPANet)*, for visual explanation of satellite image analysis. We consider the attention mechanism from the multiple layers in CNN to detect small objects in visual explanation. To propagate useful context to the following layers, we connect the attention branch of each layer in a cascading manner. Then, to correct inconsistent explanation from the onboard prediction, we describe an attention refinement scheme with supervisors in the ground station. The ground system conducts the data sampling to select inconsistent explanation in the intermediate layers with different representation, in advance. Then, with the selection-based feedback mechanism, we describe the refinement scheme of the our XAI model.

II. RELATED WORK AND PROBLEM DESCRIPTION

A. ONBOARD COMPUTING SYSTEM FOR SATELLITE IMAGE ANALYSIS

Modern satellite system contains low-power commercial off-the-shelf (COTS) accelerators (e.g FPGA, embedded GPU, and VPU) for onboard AI processing. In [10], the authors introduce low-power NVIDIA-TX1 for target recognition and segmentation using CNN. In [11], the authors claim that the DL model can be a promising solution in terms of communication cost in the satellite and facilitating navigation. They conduct a detailed analysis of deploying the DL model on COTS HW used in the satellite, and a case study of space-related applications such as cloud detection and object tracking.

As the on-orbit AI processing, CloudScout [7] is used for onboard image filtering for useless data as shown in Fig. 2. From the hyperspectral images, the onboard AI model performs cloud segmentation via the convolutional encoder-decoder network. The result is binary classification (cloudiness or not) for each pixel. If the captured image contains more than 70% of cloud pixels, the system drops the image

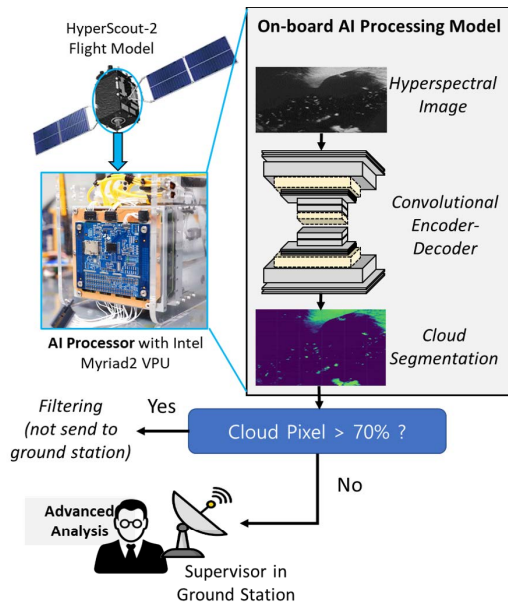


FIGURE 2. Onboard AI processing in CloudScout [7]. With low-power Intel Myriad 2 HW, it filters unnecessary cloudy images for reducing the communication bottleneck and supervisor's annotation cost.

since it has no information to analyze. By filtering the useless images on the onboard side, it can reduce the communication and supervisor's annotation costs between the satellite and the ground station. Equipped with HyperScout2 flight model, they evaluated the AI-based system in the real satellite environment, Φ -Sat1 mission. In addition, NASA and Qualcomm [12], [13] cooperate the development of onboard AI HW with Qualcomm Snapdragon and Intel Movidius Myriad X Processor for NASA jet propulsion laboratory (JPL) application. They evaluated SAR image processing with the U-Net model [14] and mars image analysis with AlexNet [15] and DeepLabv3 architecture [16]. They showed promising task performance, especially % of missed pixels (less than 10%). However, in the case of complex application such as multi-label classification and object detection, an AI model still remains in question due to its low performance compared to binary classification, resulting in the critical error of false negatives (about 4% FN error of binary classification in CloudScout).

B. XAI TECHNIQUES FOR PROVIDING VISUAL EXPLANATION

To interpret a black-box DL model in image processing, there are several methods for extracting a saliency map explaining the basis of the model prediction. Among them, we do not handle the perturbation-based methods [17], [18], [19] which need to process randomly perturbed images repeatedly, it takes more than a few times processing than the prediction process. Therefore, it is not suitable in an onboard environment with limited computing power.

As widely used a class activation map method, CAM [20] was developed as an ancestor of the class activation mapping family; it produces visual explanation results by weighting the feature maps (after the top convolution layer) from global averaging pooling (GAP) [21] and the fully-connected layer, for a target class. To avoid architectural restriction about GAP, gradient-based approaches have been introduced [22], [23]. These approaches can interpret the single-layer representation (normally the top convolution layer). LayerCAM [24] fuses the global explanation from multiple local explanations of the intermediate layers of CNN, which take advantage of localization ability in the low-level layers. However, this approach still has limitations due to the simple fusion method, which just performs an element-wise maximization of local explanations. In this case, the redundant information may contain producing the ambiguous result.

These post-hoc explanation methods except CAM still require additional operations (back-propagation for gradient computation) to generate visual explanation. And they just interpret the output feature maps from the prediction result, not any improvement in task performance (i.e. accuracy or precision). Meanwhile, instead of considering post-hoc explanation, some of the very recent studies such as ABN [25] and LFI-CAM [26] take response-based approaches similar to CAM [20]. They extend CAM by introducing an attention mechanism that is sub-branch from the CNN backbone, which improves visual explainability and allows end-to-end training (i.e., no need for any network architecture modification or fine-tuning). By doing so, *these attention branch methods* not only enable to generate visual explanation within feed-forward passes but also achieve to overcome the drawbacks of CAM, mentioned above. On the contrary, they are limited to the top convolution layer to generate visual explanation while gradient-based methods can extract any layer in the backbone CNN. In the satellite image, the top layer interpretation leads to the ambiguous explanation due to the *background bias problem* in object recognition. This problem makes visual explanation focus on the background pixels around a target object. We describe describing the details of this phenomenon in Section II-D.

On the other hand, several studies about the pyramidal attention networks [27], [28], [29] have been conducted to utilize the rich context of multiscale feature maps, but they only handle the feature attention to improve the task performance. In [30], the authors consider the various episodes from the multi-layer attention modules to generate reliable visual explanation in satellite images.

C. EXPLAINABILITY ENHANCEMENT WITH HUMAN-IN-THE-LOOP

The concept of training a model by incorporating human knowledge and experience has received attention to overcome the lack of training data and the high cost of annotation. This is called human-in-the-loop (HITL), which human experts provide feedback for achieving better performance. Our XAI-based satellite image analysis, on the other hand,

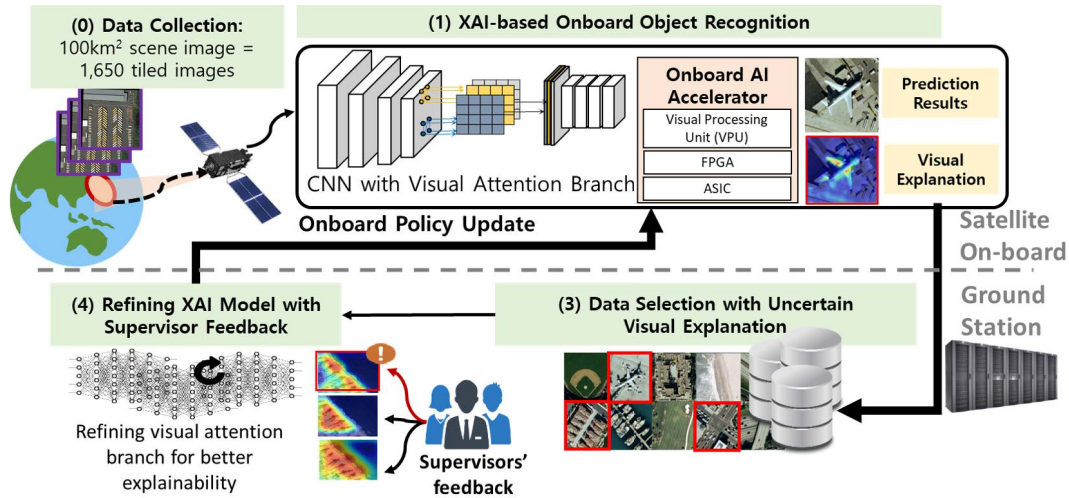


FIGURE 3. Scenario of the proposed federated XAI computing of onboard-ground station in satellite image analysis.

considers human cognition on improving model explainability rather than task performance. There are some studies that utilize HITL techniques to improve the feature explanation ability in computer vision [30], [31]. Their goal is to obtain not only accurate predictions but also proper explanations for the predictions. They collect human annotation which highlights “important regions” for decision-making. By doing so, the model trained with human knowledge in their parameters. For example, human importance-aware network tuning (HINT) [32] proposes a ranking loss between human-based importance scores [8] and gradient-based sensitivities. In self-critical reasoning (SCR) [33], the model penalizes itself for the wrong answers on the important region that most influences the prediction of the correct answer. However, these approaches take a huge time to generate the human attention map, scoring the importance of all pixels. In the case of satellite image analysis, it is difficult to annotate an attention map over all patches in a raw image ($> \text{km}^2$ per an image). To adapt HITL to satellite image analysis efficiently, our approach is to use weak supervision (i.e. simple feedback for an attention map) rather than full supervision (i.e. humanly create a ground-truth attention map for every pixel) for visual explainability.

D. PROBLEMS ON ADAPTING XAI METHODS TO ONBOARD

In this section, we argue the technical issues of existing XAI methods adapting satellite images, especially in terms of visual explainability, and sustainable system issues exploiting supervisors and computing resources in the ground station.

In practice, the layered architecture of CNN consists of the pyramid feature blocks, group of convolutional layer. Passing the pyramid feature block, the spatial dimension (i.e. width and heights) of the output feature map is reduced and the number of channels is increased due to the computation efficiency and extraction of semantic contexts along with

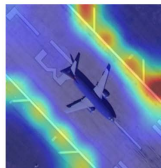
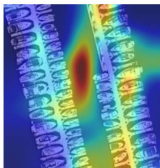
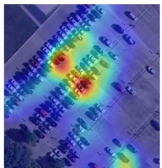
	image x_1	image x_2	image x_3
ground-truth label y_{gt}	“airplane”	“harbor”	“parkinglot”
visual explanation of top convolution layer			

FIGURE 4. Background bias problem of conventional visual explanation based on top convolution layer. The trained DL model infers the ground-truth label by focusing background context around the target objects.

the channel. As mentioned in Section II-B, existing attention branch and post-hoc explanation methods only consider the feature map of the top convolution layer to generate visual explanation due to its rich semantic context. However, they may miss the contexts in the low-level features by passing through pooling layers. Due to this *spatial information loss*, visual explanation only from the high-level feature map often fails to detect the small context or boundary of an object. It is critical in terms of explainability, especially in satellite images. As a result, generated visual explanation focuses on the background pixels, not on the target object. We refer this phenomenon as the *background bias*, which already mentioned in object recognition researches [8], [9].

To verify the background bias in a satellite image, we conducted the experiment training CNN with the pyramid feature blocks [34] into a satellite image dataset [35]. We resized the RGB input image to $224 \times 224 \times 3$ and observed the visual explanation by the post-hoc explanation method [22] widely used in an XAI field. Fig. 4 shows the background bias problem of the visual explanation method using the top convolution layer. We let an input image as x and its ground-truth

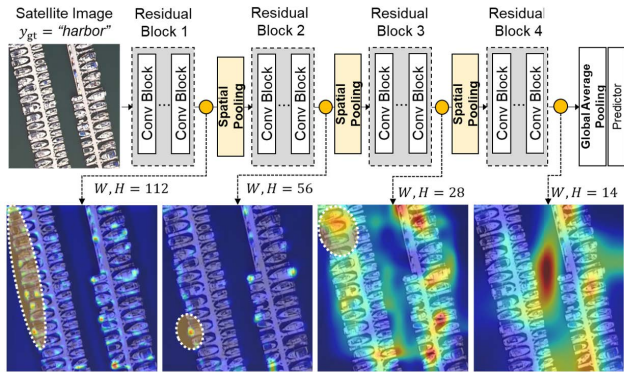


FIGURE 5. Spatial information loss caused by spatial pooling operations in a pyramid network. The highlighted region shows valuable context which is missed in visual explanation of top convolution layer.

label (i.e. category) as y_{gt} . The trained model corrects the category of all input images with high confidence. Form its visual explanation of y_{gt} , however, the model determines the category, not focusing on semantic context (i.e. airplane, ship, and car) but on background information (i.e. airstrip, ocean, and asphalt). The result implies that the trained model could fail to correct the object's category and visual explanation if the background of an input image is not commonly appeared with the target objects (e.g. the airplane passing by the ocean).

To identify the reason why this background bias occurs, we extract visual explanation of intermediate convolution layers in the pyramid feature blocks. The result is shown in Fig. 5. In the pyramid feature blocks in CNN, there are spatial pooling operations between the blocks, which reduces the width and height of the feature map. The lower convolution layers seem to only focus on the small contexts over a local region (see Residual Block 1 and 2). As mentioned in Layer-CAM [24], these layers can highlight the accurate boundary information although they cannot explain the entire context of the input image. In the contrast, visual explanation of Residual Block 3 seems to highlight the semantic contexts over the whole region of the input image while containing some redundant background areas. In the top convolution layer, however, the model fails to explain the valuable contexts highly biased to the background. This comes from the spatial pooling operations missing features about "boat". In this example, it seems to be the proper choice to select a visual explanation of residual block 3 for the highest explainability. In summary, visual explanation only using the top convolution layer cannot guarantee explainability to human supervisors. Our approach is to strengthen visual explanation of the top convolution layer by combining useful context for accurate boundaries and small objects in the lower convolution layers. To this end, we present a novel attention branch method with multiple attention blocks connected in a cascading manner.

Furthermore, to ensure the reliability of complex object recognition tasks, continuous updates of the model (i.e. trainable weights) according to newly captured images in the satellite are required. The refinement (retraining) of the trained

model is not suitable for processing in the onboard computing in terms of no supervision for explanation and computing capability for retraining. The onboard system cannot acquire the ground truth visual explanation of the captured images. In addition, retraining an XAI model needs powerful computing resources (e.g. GPUs) and a huge time to complete the (re)training. This is not suitable in the onboard computing environment with a limited computing constraint. Furthermore, the manual correction on each pixel of visual explanation [32], [33] also needs a huge amount of labeling cost. To handle these issues, we propose a concept of federated XAI computing framework for the onboard-ground station. Fig. 3 is overall scenario of the proposed concept. In the onboard, an XAI-based object recognition model is performed for the captured images from the satellite. Note that the filtering criteria in the onboard is determined by users, so we do not handle this. The prediction results including the object's class and its visual explanation transmit to the ground station. In the ground station, the samples with ambiguous visual explanation are automatically classified by the active learning-based sampling. Then, an XAI model is retrained based on the supervisor's feedback to enhance the visual explainability. The updated weights transmit the onboard HW to process new incoming images, and then repeat the entire procedure.

III. THE PROPOSED METHOD

In this section, we propose the FOGS computing system for an XAI-based satellite image analysis. Through the system, we are going to improve the visual explainability of the XAI model to provide reliable object recognition on the satellite onboard. Our approach is to consider the computing interaction mechanisms between a satellite onboard HW and the ground station, and ensure that the XAI model is trained with the supervisor's knowledge using a simple feedback mechanism, simultaneously.

A. OVERALL PROCEDURE OF FEDERATED ONBOARD-GROUND STATION COMPUTING

Fig. 6 shows the overall architecture of the proposed onboard-ground station federated computing framework. In the proposed framework, the onboard system directly executes the inference of an XAI model according to the captured raw images from the satellite. It outputs the prediction result of target objects (e.g. "airplane", "ship", etc.) and its visual reason (i.e. explanation) in the form of a saliency map. From the input image \mathbf{x} and the trained black-box CNN model, we denote visual explanation of the predicted class as $VE_{ij}(\mathbf{x})$, feature importance factor of (i, j) pixel in the input image. It can be said that a (i, j) pixel with a high $VE_{ij}(\mathbf{x})$ value contributes to the prediction result, significantly. Based on the prediction result of the XAI model, the onboard system determines the raw images to be dropped (i.e. not informative images related to the mission). Different from CloudScout [7], our system can handle more complex image analysis tasks (e.g. object recognition, scene classification,

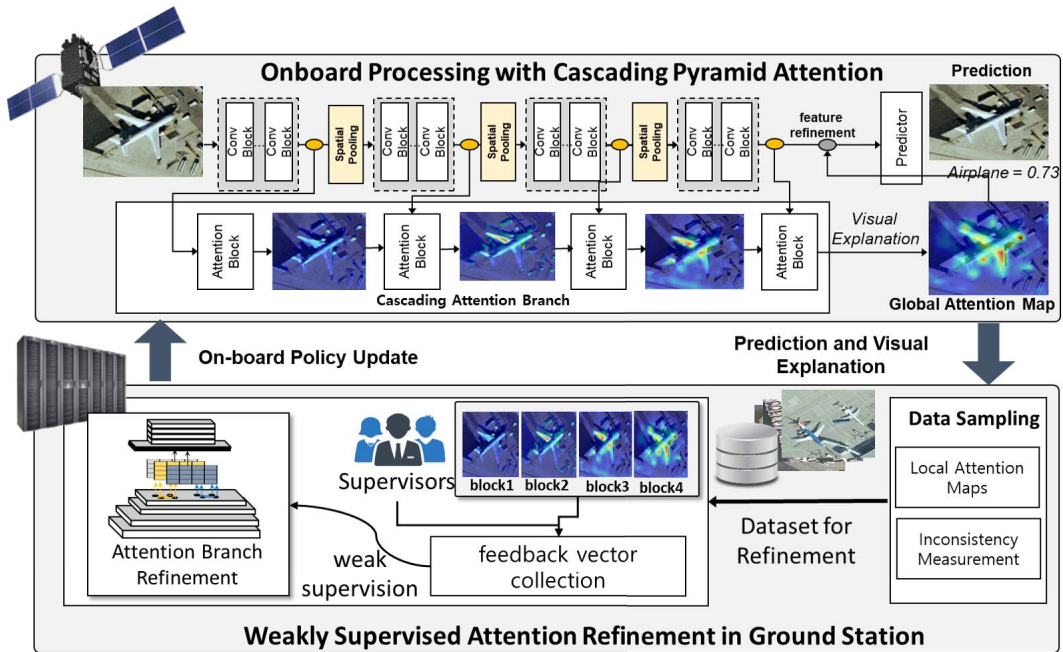


FIGURE 6. Overall architecture of the proposed onboard-ground station federated XAI computing with cascading pyramid attention and weakly supervised refinement.

etc.) via the XAI model. Once images and the analysis results are transmitted to the ground station, the supervisors analyze the correction of visual explanation. Since the onboard XAI processing should be reliable, especially preventing critical error (in the case of FN error in image selection), the ground station refines the trained XAI model by inducing the supervisor's knowledge about the target prediction. We consider the annotation cost of the supervisor when correcting visual explanation while enhancing visual explainability in terms of consistency. We describe the following methods based on the onboard-ground station federation with the supervisors.

1) ONBOARD PROCESSING WITH CPANet

First, we present a novel attention branch method, cascading pyramid attention network (CPANet), to mitigate a quality degradation of visual explanation due to the background bias problem [8], [9]. As presented in Section II-B, we identify a critical failure in visual explanation of the conventional attention branch methods [25], [26] which exploit the top convolution layer. Our approach is to exploit multiscale feature maps from various layers in pyramid feature blocks of CNN. We denote visual explanation aggregating valuable context over multiple feature maps as *global explanation*. On the other hand, *local explanation* represents visual explanation about a single feature map. They contain not only semantic information of objects but localization ability to detect the small context (e.g. ship head, wings of an airplane) or objects (e.g. car, boat). To extract elaborate global explanation from feature maps with different spatial resolutions, we design a cascading attention branch, subpath of pyramid feature blocks to propagate the valuable context from the bottom

to the top convolution layer. In a cascading manner, local explanation (i.e. local attention map) of the previous pyramid feature block becomes a guide for extracting local explanation of the following block while amplifying feature values of the region where the previous one highlights. The global explanation (i.e. global attention map) of the top block is utilized for the refinement of the output feature map in the feature pyramid network and visual explanation providing to a supervisor.

2) WEAKLY SUPERVISED ATTENTION REFINEMENT IN GROUND STATION

Next, we discuss an attention refinement method adapting the supervisor's knowledge only using a simple feedback mechanism to improve visual explainability. In this step, the parameters of the onboard XAI model are refined by supervisors in the ground station. In the classical approaches [32], [33] using full supervision, supervisors corrects all (i, j) pixel values in visual explanation, which is very time-consuming and highly dependent on the supervisor's ability. In our method, we split two steps for refinement of the attention branch; choosing the set of images showing the inconsistent explanation, and weakly supervised attention refinement with selection-based feedback in a feature pyramid network. The refined attention branch containing the supervisor's knowledge is uploaded to the onboard system for more reliable prediction and explanation. By updating policy according to this attention branch iteratively, the proposed framework can achieve a sustainable and reliable system for satellite image analysis.

In the following section, we describe an XAI model and training strategies in the proposed system, in detail.

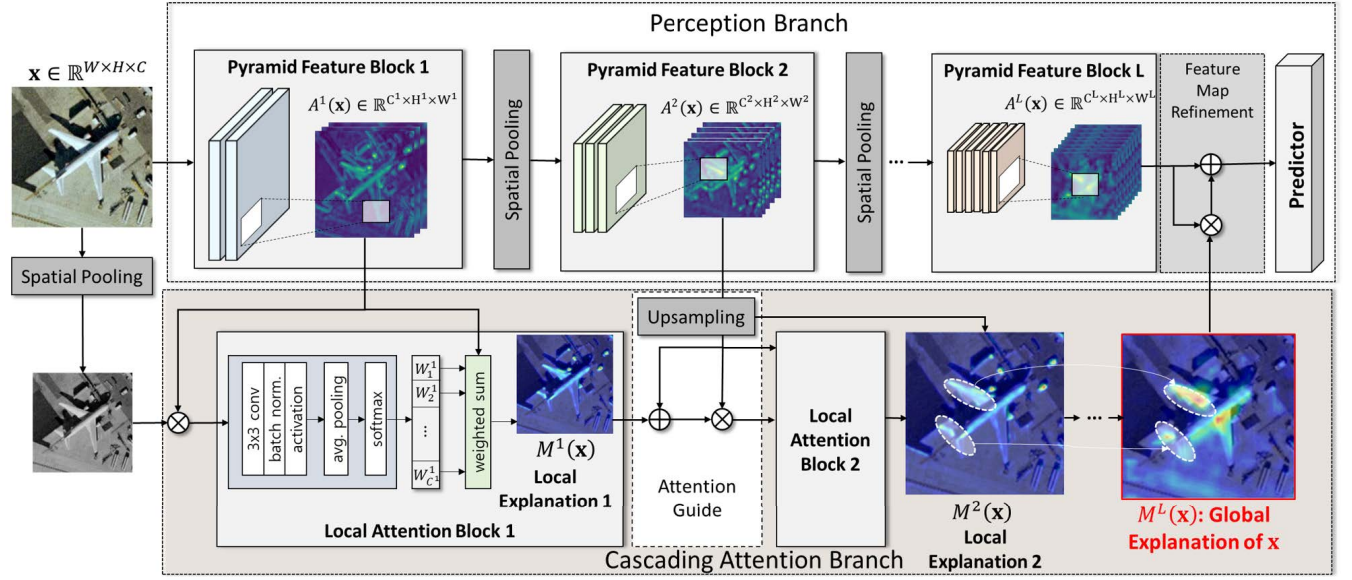


FIGURE 7. Proposed CPANet for generating the global context to explain the satellite images. It consists of the perception branch and cascading attention branch to transmit useful context from local explanations to global explanation via the bottom-up pathway. Detailed formulations are shown in Section III-B.

B. CASCADING PYRAMID ATTENTION NETWORK FOR ONBOARD IMAGES PROCESSING

In this section, we describe CPANet architecture for the onboard processing of the captured images by the satellite, as shown in Fig. 7. In advance, let \mathcal{D} be a training dataset with N pairs of (\mathbf{x}, \mathbf{y}) , an image $\mathbf{x} \in \mathbb{R}^{C \times W \times H}$ (channel, width, height) and its ground-truth label $\mathbf{y} \in \{1, 2, \dots, K\}$, where K is the number of classes in \mathcal{D} . The proposed CPANet $\Theta = \{\mathbf{u}, \mathbf{v}\}$ consists of the perception branch \mathbf{u} , and the cascading attention branch \mathbf{v} . Building upon the CNN with the pyramid feature blocks, we first denote a *pyramid feature block* as a group of consecutive layers in which output feature maps have the *same spatial dimension of width and height*. At the end of each pyramid block, there is the spatial pooling (e.g. average pooling) layer which compresses the spatial dimension for computational efficiency. In the pyramid feature blocks, we consider L pyramid feature blocks from different convolution layers. Given the input image \mathbf{x} , we denote L feature maps of the top convolution layer in each pyramid feature block described in

$$\mathbf{A}(\Theta; \mathbf{x}) = \{\mathcal{A}^i(\Theta; \mathbf{x})\}_{i=1}^L. \quad (1)$$

Each feature map $\mathcal{A}^i(\Theta; \mathbf{x}) \in \mathbb{R}^{C^i \times W^i \times H^i}$ has its own dimension. Because CPANet architecture is fixed in the following, we simply use a term $\mathcal{A}^i(\mathbf{x})$ instead of $\mathcal{A}^i(\Theta; \mathbf{x})$ for mathematical simplicity, except in the theoretical analysis.

We present the cascading attention branch with subpath from multiple pyramid feature blocks to extract the global attention map over the multiscale feature maps $\{\mathcal{A}^i(\mathbf{x})\}_{i=1}^L$. To propagate the valuable contexts in a local feature map to the following feature map, the previous attention map becomes a guide when generating the following attention map. As shown in Fig. 7, the feature map $\mathcal{A}^1(\mathbf{x})$ is passed

into the following formula:

$$\mathbf{W} = \text{AvgPool}(\tilde{\mathbf{x}}) \odot \frac{\mathcal{A}^1(\mathbf{x}) - \min \mathcal{A}^1(\mathbf{x})}{\max \mathcal{A}^1(\mathbf{x})}, \quad (2)$$

$$\mathbf{W}' = \text{AvgPool}(\text{AttB}^1(\mathbf{W})), \quad (3)$$

$$W_i^1(\mathbf{x}) = \frac{\exp(W_i')}{\sum_{j=1}^{C^1} \exp(W_j')}, \quad \forall i = \{0, 1, \dots, C^1 - 1\}. \quad (4)$$

In the first term of Eq. (2), the input image transformed to grey-scale $\tilde{\mathbf{x}}$ and down sampling matching width and height to \mathcal{A}^1 via average pooling (AvgPool), which multiplies by the normalized feature map of the bottom pyramid feature block. We denote as \odot a element-wise multiplication. In Eq. (3), the output \mathbf{W} is passing the local attention block and average pooling to reduce the output to the C^1 -dimensional vector. We design the local attention block with stacked convolutional layers, and the input \mathbf{W} in each convolution block is forwarded as $\sigma(\text{BN}(\omega * \mathbf{W} + \beta))$, where σ is an activation function of ReLU, BN is batch normalization, and (ω, β) is a pair of weight and bias for convolution operation $*$. After passing the local attention block, the output $\mathbf{W}' \in \mathbb{R}^{C^1}$ of Eq. (3) becomes confidence for channel importance of the feature map \mathcal{A}^1 . It means that the particular channel with high value has the valuable context for explaining the objects. To transform it to relative importance among the C^1 channels, the softmax function is passed in Eq. (4). The output vector $\mathbf{W}^1(\mathbf{x}) = \{W_i^1(\mathbf{x})\}_{i=1}^{C^1}$ weighs the feature map $\mathcal{A}^1(\mathbf{x})$ to generate the local attention map, as follows.

$$\mathbf{M}(\mathbf{x}) = \sigma \left(\sum_{i=1}^{C^1} W_i^1(\mathbf{x}) \mathcal{A}_i^1(\mathbf{x}) \right),$$

$$M^1(\mathbf{x}) = \frac{\mathbf{M}(\mathbf{x}) - \min \mathbf{M}(\mathbf{x})}{\max \mathbf{M}(\mathbf{x})}, \quad (5)$$

where we denote $\mathcal{A}_i^1(\mathbf{x})$ as the i -th channel in the feature map. The local attention map $M^1(\mathbf{x}) \in \mathbb{R}^{W^1 \times H^1}$ is a heatmap where each value $M_{ij}^1 \in [0, 1]$ is an (i, j) pixel importance for visual explanation. Through the weighted sum of channels in the feature map $\mathcal{A}^1(\mathbf{x})$, the informative channels are emphasized in the local attention map $M^1(\mathbf{x})$.

Once the local attention map $M^i(\mathbf{x})$ is generated, our purpose of the cascading attention is to enforce the following feature map \mathcal{A}^{i+1} focusing on the region where the attention map $M^i(\mathbf{x})$ is highlighting. To this end, we conduct the pre-processing on the feature map $\mathcal{A}^i(\mathbf{x})$ guided by the previous attention map $M^{i-1}(\mathbf{x})$, which is obtained by

$$\mathbf{W} = (1 + M^{i-1}(\mathbf{x})) \odot \frac{\mathcal{A}^i(\mathbf{x}) - \min \mathcal{A}^i(\mathbf{x})}{\max \mathcal{A}^i(\mathbf{x})}, \quad \forall i = \{2, 3, \dots, L\}. \quad (6)$$

And, the following calculation is same as Eq. (3) and Eq. (4). After passing the $L - 1$ local attention blocks, the global attention map $M^L(\mathbf{x})$ creates containing the global context over the pyramid feature blocks. We regard it as global visual explanation $VE(\mathbf{x})$ (i.e. $VE(\mathbf{x}) = M^L(\mathbf{x})$). Moreover, the global attention map $M^L(\mathbf{x})$ is applied to refine the feature map $\mathcal{A}^L(\mathbf{x})$ to improve the task performance. The refined feature map $\tilde{\mathcal{A}}^L(\mathbf{x})$ is derived as

$$\tilde{\mathcal{A}}^L(\mathbf{x}) = (1 + M^L(\mathbf{x})) \odot \mathcal{A}^L(\mathbf{x}). \quad (7)$$

From Eq. (6), the regions where the global attention map $M^L(\mathbf{x})$ are emphasized into the refined feature map $\mathcal{A}^L(\mathbf{x})$, whereas other regions are maintained. Finally, the refined feature map passes to the predictor for correcting the object's class. The predictor outputs the c -class confidence $p^c(\Theta; \mathbf{x})$ of objects from the refined feature map. To train the model, let the mini-batch of the training dataset be $\zeta = (\mathcal{X}, \mathcal{Y})$ with image set \mathcal{X} and corresponding labels \mathcal{Y} , the loss term is derived as

$$\mathcal{L}(\Theta; \zeta) = -\frac{1}{|\zeta|} \sum_{(\mathbf{x}, y) \in \zeta} \sum_{j=1}^K \mathbf{1}[y=j] \log(p^j(\Theta; \mathbf{x})), \quad (8)$$

where (\mathbf{x}, y) is a pair of image and label in mini-batch and the indicator function $\mathbf{1}[y=j]$ is 1 only if $y=j$. Based on Eq. (8), we update the trainable parameters Θ^t in t -th iterations using gradient descent described in Eq. (9).

$$\Theta^{t+1} = \Theta^t - \eta \nabla_{\Theta^t} \mathbb{E}_{\zeta \sim \mathcal{D}} [\mathcal{L}(\Theta^t; \zeta^t)], \quad (9)$$

where η is learning rate and ζ^t is mini-batch in t -th training iterations.

Now we provide a detail analysis for relationship between local attention maps via the proposed CPANet. First, we define the pyramid feature blocks and the local attention blocks as follows:

Definition 1 (Relationship Between the Pyramid Feature and Local Attention Block): In the perception branch \mathbf{u} , we denote the trainable parameters of the i -th pyramid feature block and the local attention block as \mathbf{u}_i and \mathbf{v}_i , respectively.

In addition, the trainable parameters from i -th block to j -th block is denoted as \mathbf{u}_i^j and \mathbf{v}_i^j .

Training with the parameter update rule denoted in Eq. (9), we suppose the following assumption:

Assumption 1: For training dataset \mathcal{D} and parameter update function Eq. (9) with proper hyperparameter setting, we assume that the gradient of loss term $\mathbb{E}_{\zeta \sim \mathcal{D}} [\mathcal{L}(\Theta^t; \zeta)]$ is converged. And, let local attention map with the cascading attention branch be $M^i(\Theta; \mathbf{x}) = M^i(\mathbf{u}_i, \mathbf{v}_i; \zeta)$. Then we assume that the local gradients $\|\nabla_{\mathbf{u}_i} M^i(\mathbf{u}_i^1, \mathbf{v}_i^1; \zeta)\|, \|\nabla_{\mathbf{v}_i} M^i(\mathbf{u}_i, \mathbf{v}_i; \zeta)\| \forall i = \{1, 2, \dots, L\}$ have upper-bound Z .

Note that the i -th local attention map is derived by corresponding i -th pyramid feature block and local attention block (i.e. the local attention map calculated from the path of the pyramid feature block and local attention block, see Fig. 7). From Assumption 1, we can construct the effectiveness of the cascaded attention in the following Lemma 1.

Lemma 1: In the i, j -th the pyramid feature blocks with $\forall i, j = \{1, 2, \dots, L\}, i \geq j$, the gradients of attention maps M^i and M^j are satisfied as

$$\|\nabla_{\Theta} M^i(\Theta; \zeta) - \nabla_{\Theta} M^j(\Theta; \zeta)\| \leq (2Z)^j (1 + (2Z)^{i-j}). \quad (10)$$

Proof: From the Assumption 1 and from the gradient chain rule, we can have

$$\begin{aligned} & \|\nabla_{\Theta} M^i(\Theta; \zeta) - \nabla_{\Theta} M^j(\Theta; \zeta)\| \\ &= \|\nabla_{(\mathbf{u}_i^1, \mathbf{v}_i^1)} M^i(\mathbf{u}_i^1, \mathbf{v}_i^1; \zeta) - \nabla_{(\mathbf{u}_j^1, \mathbf{v}_j^1)} M^j(\mathbf{u}_j^1, \mathbf{v}_j^1; \zeta)\| \\ &\leq \|\nabla_{(\mathbf{u}_i^1, \mathbf{v}_i^1)} M^i(\mathbf{u}_i^1, \mathbf{v}_i^1; \zeta)\| + \|\nabla_{(\mathbf{u}_j^1, \mathbf{v}_j^1)} M^j(\mathbf{u}_j^1, \mathbf{v}_j^1; \zeta)\| \\ &= \|\nabla_{M^{i-1}} M^i(\mathbf{u}_i^1, \mathbf{v}_i^1; \zeta) \cdot \nabla_{(\mathbf{u}_i^{i-1}, \mathbf{v}_i^{i-1})} M^{i-1}(\mathbf{u}_i^{i-1}, \mathbf{v}_i^{i-1}; \zeta)\| \\ &\quad + \|\nabla_{M^{j-1}} M^j(\mathbf{u}_j^1, \mathbf{v}_j^1; \zeta) \cdot \nabla_{(\mathbf{u}_j^{j-1}, \mathbf{v}_j^{j-1})} M^{j-1}(\mathbf{u}_j^{j-1}, \mathbf{v}_j^{j-1}; \zeta)\| \\ &= \|\nabla_{\mathbf{u}_i} M^i(\mathbf{u}_i^1, \mathbf{v}_i^1; \zeta) + \nabla_{\mathbf{v}_i} M^i(\mathbf{u}_i^1, \mathbf{v}_i^1; \zeta)\| \\ &\quad \cdot \nabla_{(\mathbf{u}_i^{i-1}, \mathbf{v}_i^{i-1})} M^{i-1}(\mathbf{u}_i^{i-1}, \mathbf{v}_i^{i-1}; \zeta)\| \\ &\quad + \|\nabla_{\mathbf{u}_j} M^j(\mathbf{u}_j^1, \mathbf{v}_j^1; \zeta) + \nabla_{\mathbf{v}_j} M^j(\mathbf{u}_j^1, \mathbf{v}_j^1; \zeta)\| \\ &\quad \cdot \nabla_{(\mathbf{u}_j^{j-1}, \mathbf{v}_j^{j-1})} M^{j-1}(\mathbf{u}_j^{j-1}, \mathbf{v}_j^{j-1}; \zeta)\| \\ &= 2Z \cdot \{\|\nabla_{(\mathbf{u}_i^{i-1}, \mathbf{v}_i^{i-1})} M^{i-1}(\mathbf{u}_i^{i-1}, \mathbf{v}_i^{i-1}; \zeta)\| \\ &\quad + \|\nabla_{(\mathbf{u}_j^{j-1}, \mathbf{v}_j^{j-1})} M^{j-1}(\mathbf{u}_j^{j-1}, \mathbf{v}_j^{j-1}; \zeta)\|\} \\ &\leq (2Z)^j (1 + (2Z)^{i-j}) = A. \end{aligned} \quad (11)$$

where the first inequality is obtained from $\|\mathbf{z}_1 - \mathbf{z}_2\| \leq \|\mathbf{z}_1\| + \|\mathbf{z}_2\|, \forall \mathbf{z}_1, \mathbf{z}_2 \in \mathbb{R}^d$; the second inequality is obtained from $\|\mathbf{z}_1 \mathbf{z}_2\| \leq \|\mathbf{z}_1\| \|\mathbf{z}_2\|, \forall \mathbf{z}_1, \mathbf{z}_2 \in \mathbb{R}^d$. Note that the partial gradients of M^1 is also bounded to Z because the perception branch and attention branch is independent in the first block (i.e. $M^1(\mathbf{u}^1, \mathbf{v}^1; \zeta) = \text{Per}(\mathbf{u}^1; \zeta) + \text{Att}(\mathbf{v}^1; \zeta)$, where Per and Att are the function of the perception branch and attention branch. \square

We denote $(2Z)^j (1 + (2Z)^{i-j})$ to A that is used in the Section III-D. From the Lemma 1, we can assert that the object-related region from the previous attention map propagates the next attention map. Through the L cascading

attention blocks, visual explanation of the top convolution layer can have the highlighted region in the previous attention maps by compensating for the spatial information loss.

C. EXPLANATION INCONSISTENCY-BASED DATA SAMPLING

CPANet deployed to the onboard system can provide elaborate visual explanation catching the small context in a satellite image while mitigating the background bias in a feature pyramid network. However, the problem of ambiguous visual explanation may still occur due to the environmental change and newly captured images out of the distribution of the training dataset. Therefore, we consider the refinement mechanisms of the proposed CPANet between the onboard and ground station. In this section, we describe the active learning-based data sampling for finding valuable samples to improve visual explainability of CPANet. To this end, it needs to resolve the problem of filtering the samples showing ambiguous explanation automatically. As shown in Fig. 5, visual explanations which inconsistent in different blocks may indicate the information loss about the target object. Inspired by this phenomenon, we introduce the criteria about how *inconsistent visual explanations* are with respect to the pyramid feature blocks.

In advance, we provide a method to compare the inconsistency of different visual explanations. It is measured by comparing the similarity of two visual explanations. Previous similarity metrics [36], [37] for visual maps is based on pixel-wise comparison, given the image \mathbf{x} . However, in the case of visual explanation, this measurement include redundant similarity of the region where the model is not focused on. In this paper, we define a similarity metric of two visual explanation in Definition 2.

Definition 2 (Similarity of Two Visual Explanations): Given the input \mathbf{x} the similarity of two visual explanations generated by the cascading attention branch, $VE^1(\mathbf{x})$ and $VE^2(\mathbf{x})$ is defined as “similarity of the spatial region where the two explanations are commonly highlighting”.

From Definition 2, we ignore the common area where the both explanations are not focused on (i.e. blue area in Fig. 4). To quantify the similarity, we describe a simple method as follows: eliminating pixels (i.e. value to 0) with low values of the region in visual explanation. To remain informative pixels we use a threshold as 15% of the maximum value in visual explanation similar with Grad-CAM [22]. We denote transformed visual explanations as $\overline{VE}^1(\mathbf{x})$, $\overline{VE}^2(\mathbf{x})$, respectively. Note that each pixel value in explanation has a range of [0, 1]. And then, the similarity $SIM(VE^1(\mathbf{x}), VE^2(\mathbf{x}))$ is derived as:

$$SIM(VE^1(\mathbf{x}), VE^2(\mathbf{x})) = \frac{\sum_{(i,j) \in \mathcal{S}} \{1 - |\overline{VE}_{ij}^1(\mathbf{x}) - \overline{VE}_{ij}^2(\mathbf{x})|\}}{\text{area}(\mathcal{S})}, \quad (12)$$

where \mathcal{S} is a non-zero pixels in explanations,

$$\mathcal{S} = \{(i, j) | \overline{VE}_{ij}^1(\mathbf{x}) \neq 0 \vee \overline{VE}_{ij}^2(\mathbf{x}) \neq 0\}. \quad (13)$$

Note that $\text{area}(\mathcal{S})$ is total number of pixels in \mathcal{S} and \overline{VE}_{ij}^1 is (i, j) pixel value. Based on the similarity in Eq. (12), we can define the inconsistency of $\overline{VE}^1(\mathbf{x})$ and $\overline{VE}^2(\mathbf{x})$ as

$$\mathcal{U}(VE^1(\mathbf{x}), VE^2(\mathbf{x})) = 1 - SIM(VE^1(\mathbf{x}), VE^2(\mathbf{x})). \quad (14)$$

From the inconsistency measurement in Eq. (14) the system conducts the data sampling. Note that the proposed CPANet is to pass the valuable local contexts to the top convolution layer. Therefore, our data selection is based on the inconsistency between the local attention maps $\{M^i(\mathbf{x})\}_{i=1}^L$. Over training dataset $\mathcal{D} = (\mathbf{x}, \mathbf{y})_{i=1}^N$, we evaluate the inconsistency of $VE^i(\mathbf{x})$ and $VE^j(\mathbf{x})$ of the i -th and j -th local attention maps with the trained model Θ . Based on the maximum inconsistency $\max_{(i,j)} \mathcal{U}(VE^i(\mathbf{x}), VE^j(\mathbf{x}))$ we filter the sample \mathbf{x} to be used for attention refinement using threshold γ . We refer $\mathcal{D}_U = (\mathbf{x}, \mathbf{y})_{i=1}^N$ be the retraining dataset.

D. ATTENTION REFINEMENT USING WEAK SUPERVISION IN GROUND STATION

In this section, we handle to improve the explanation fidelity of the saliency map to supervisors by fine-tuning the cascading attention branch. To retrain CPANet for consistent explanation, we propose a novel weakly-supervised learning mechanism with selection-based simple feedback from supervisors. In conventional fully-supervised approach [32], a supervisor should manually create the ground-truth attention map. In our approach, we concentrate on inconsistent local attention maps of ambiguous samples, which are filtered the active learning. Using this characteristics of various explainability over the pyramid feature blocks, we just provide a selection among the local attention maps with high interpretability. Based on the selected attention map as weak supervision, CPANet is retrained with the attention regularization loss term. Fig. 8 shows the overall procedure of the proposed refinement method.

To provide useful explanations to the supervisor we consider the training the self-attention weights with supervisor intervention to visual explanations. From the retraining dataset \mathcal{D}_U we define the supervisor's feedback in the L pyramid feature blocks, denoted as \mathcal{G} , for feedback interface to L local attention maps.

Definition 3 (Selection-Based Feedback): Supervisor feedback \mathcal{G} over L local attention maps is indicator for selecting the visual explanation with the human interpretability, which is

$$\mathcal{G}(\Theta; \mathbf{x}_i \sim \mathcal{D}_U) : \{M^j(\mathbf{x}_i)\}_{j=1}^L \rightarrow \mathbb{R}^{1 \times L}, \quad (15)$$

where each $\mathcal{G}^j(\Theta; \mathbf{x}_i \sim \mathcal{D}_U) \in [0, 1]$ of j -th local attention map has the following states $\mathcal{ST} = \{\text{“wrong”}, \text{“correct”}\}$.

We simply denote $\mathcal{G}(\Theta; \mathbf{x}_i \sim \mathcal{D}_U)$ as $\mathcal{G}^j(\mathbf{x}_i)$. For “wrong” attention map, we add the penalty function. In the ground station, there are several supervisors for analyzing satellite images. We assume that each supervisor has own private knowledge about analyzing the satellite images (e.g. domain, class, etc.). In this situation, it is possible to mitigate the

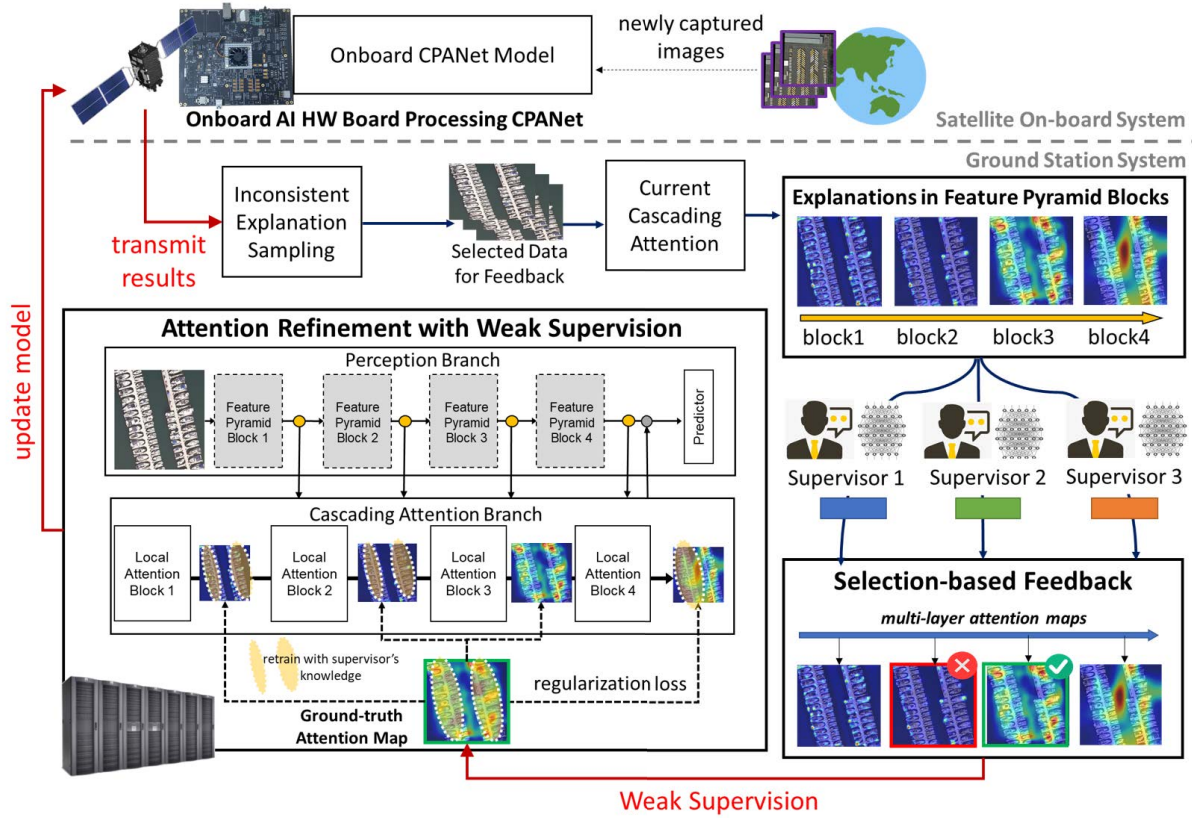


FIGURE 8. Update for onboard attention policy using weak supervision by selection-based annotation in pyramid feature blocks.

refined attention branch to be biased on a particular supervisor by conducting the aggregated labeling of feedback from multiple supervisors. Assume that there are \mathcal{T} supervisors with different domain knowledge in the ground station. We denote $\mathcal{G}_k^j(\mathbf{x}_i) \in \mathbb{R}^{1 \times L}$ the feedback (i.e. binary vector) of j -th pyramid block of supervisor k according to the input image \mathbf{x}_i . Through the majority voting to maximize consensus among supervisors we derive the collective supervisor feedback as $\sum_{k=1}^{\mathcal{T}} \mathcal{G}_k^j(\mathbf{x}_i)$. If this value exceeds consensus criteria δ , we set $\bar{\mathcal{G}}^j(\mathbf{x}_i)$ to 1 (*correct*), else to 0 (*wrong*). In this paper, we assume that there is only 1 supervisor for simplicity (i.e. $\mathcal{G} = \bar{\mathcal{G}}$). We set “correct” attention map as the weak ground-truth for retraining.

When the image \mathbf{x} and weak supervision $\{\bar{\mathcal{G}}^j(\mathbf{x})\}_{j=1}^L$ are given, we obtain a regularization term w.r.t the attention map as Eq. (16).

$$\frac{1}{Q} \sum_{j=1}^L \sum_{k=1}^L \|\mathbf{1}[\bar{\mathcal{G}}^j(\mathbf{x})=0]M^j(\mathbf{x}) - \mathbf{1}[\bar{\mathcal{G}}^k(\mathbf{x})=1]M^k(\mathbf{x})\|_2^2, \quad (16)$$

where Q are the number of the pyramid feature blocks with “wrong” labeling. From Eq. (16), the attention block and connected pyramid feature block producing the wrong attention map can refine the weights guided by the supervisor’s

knowledge. The loss function can be derived from the regularization term according to the explanation distance in a weakly supervised attention map. As a result, the loss function \mathcal{L}_{ref} for refining CPANet is derived as

$$\mathcal{L}_{\text{ref}}(\mathcal{D}_U) = -\frac{1}{N} \sum_{i=1}^{\hat{N}} \mathcal{L}(\mathbf{x}_i, \mathbf{y}_i) + \alpha \left\{ \frac{1}{Q} \sum_{j=1}^L \sum_{k=1}^L \|\mathbf{1}[\bar{\mathcal{G}}^j(\mathbf{x})=0]M^j(\mathbf{x}) - \mathbf{1}[\bar{\mathcal{G}}^k(\mathbf{x})=1]M^k(\mathbf{x})\|_2^2 \right\}, \quad (17)$$

where α is a control variable for refinement. The α value means how the refined model reflects the supervisor’s knowledge for explainability. The update function at t -th refinement iterations is denoted as Eq. (18).

$$\Theta^{t+1} = \Theta^t - \eta \nabla_{\Theta^t} \mathbb{E}_{\zeta \sim \mathcal{D}_U} [\mathcal{L}_{\text{ref}}(\Theta^t; \zeta^t)]. \quad (18)$$

After training all weights in the attention branch, the inconsistency of L local attention maps can be reduced. Our approach has a intuitive interface for supervisor intervention reducing the annotation costs (correction of the attention map) of large satellite images.

Now, we provide the theoretical analysis for explainability boundary of the local attention maps.

Theorem 1 (Explainability Consistency for Attention Maps): Let Lemma 1 be satisfied. We assume that there are only single correct explanation $M^{GT}(\Theta; \mathbf{x})$ among L local

Algorithm 1 Procedure of Attention Refinement for Visual Explainability in the Ground Station

Input: Trained CPANet Θ , training dataset \mathcal{D} , inconsistency threshold of visual explanation γ , consensus criteria δ

```

1:  $\mathcal{D}_U \leftarrow \{\}$   $\triangleright$  sampling data for refinement
2: for all  $\mathbf{x}$  in  $\mathcal{D}$  do
3:    $SIM_{min} \leftarrow 1$ 
4:    $U_{max} \leftarrow 0$ 
5:   for all  $i$  in  $\{1, 2, \dots, L\}$  do
6:     for all  $j$  in  $\{1, 2, \dots, L\}$  do
7:        $VE^i \leftarrow M^i(\Theta; \mathbf{x})$  and  $VE^j \leftarrow M^j(\Theta; \mathbf{x})$ 
8:        $\overline{VE}^i, \overline{VE}^j \leftarrow$  remove non-informative pixels
9:        $\mathcal{S} \leftarrow \{(i, j) | \overline{VE}_{ij}^1(\mathbf{x}) \neq 0 \vee \overline{VE}_{ij}^2(\mathbf{x}) \neq 0\}$ 
10:       $AD \leftarrow |\overline{VE}^1 - \overline{VE}^2|$ 
11:       $SIM(VE^1, VE^2) \leftarrow \frac{\sum_{(i,j) \in \mathcal{S}} 1 - AD_{ij}}{\text{area}(\mathcal{S})}$ 
12:      if  $SIM(VE^1, VE^2) \leq SIM_{min}$  then
13:         $SIM_{min} \leftarrow 1 - SIM(VE^1, VE^2)$ 
14:      end if
15:    end for
16:  end for
17:   $U_{max} \leftarrow 1 - SIM_{min}$ 
18:  if  $U_{max} \geq \gamma$  then
19:     $\mathcal{D}_U \leftarrow \mathcal{D}_U \cup \mathbf{x}$ 
20:  end if
21: end for
22: for all  $\mathbf{x}$  in  $\mathcal{D}_U$  do  $\triangleright$  attention refinement
23:   Calculate  $\{M^j(\mathbf{x})\}_{j=1}^L$ 
24:    $\mathcal{G}_i^j(\mathbf{x}) \leftarrow \mathbb{R}^{T \times L} \forall i, j$  from  $\mathcal{T}$  supervisors
25:   if  $\sum_{k=1}^T \mathcal{G}_k^j(\mathbf{x}_i) \geq \delta$  then
26:      $\overline{\mathcal{G}}^j(\mathbf{x}_i) \leftarrow 1$ 
27:   else
28:      $\overline{\mathcal{G}}^j(\mathbf{x}_i) \leftarrow 0$ 
29:   end if
30: end for
31:  $\Theta \leftarrow$  Retraining CPANet by Eq. (17)
32: Update policy  $\Theta$  to onboard

```

attention maps. From the initial trained model Θ , under the loss function Eq. (17), the difference of the output of the two pyramid feature blocks $M^{GT}(\mathbf{x})$ and $M^j(\mathbf{x})$ is bounded as

$$\|M^{GT}(\Theta^t; \zeta^t) - M^j(\Theta^t; \zeta^t)\| \leq A^t(1 - 2\alpha\eta)^t \|M^{GT}(\Theta^0; \zeta^0) - M^j(\Theta^0; \zeta^0)\|, \quad (19)$$

where Θ^0 denotes the initial trained ones for Θ and ζ^t is mini-batch for t -iterations during refinement.

Proof: From the loss function and update rule given in Eq. (17) and Eq. (18) and Lemma 1, we can obtain

$$\begin{aligned} & \|M^{GT}(\Theta^t; \zeta^t) - M^j(\Theta^t; \zeta^t)\| \\ &= \|M^{GT}(\Theta^{t-1}; \zeta^{t-1}) - \eta \nabla_{\Theta^{t-1}} \mathcal{L}_{\text{ref}}(\Theta^{t-1}; \zeta^{t-1}) \\ &\quad - M^j(\Theta^{t-1}; \zeta^{t-1}) + \eta \nabla_{\Theta^{t-1}} \mathcal{L}_{\text{ref}}(\Theta^{t-1}; \zeta^{t-1}) \\ &\quad - 2\alpha\eta \{M^{GT}(\Theta^{t-1}; \zeta^{t-1}) - M^j(\Theta^{t-1}; \zeta^{t-1})\} \end{aligned}$$

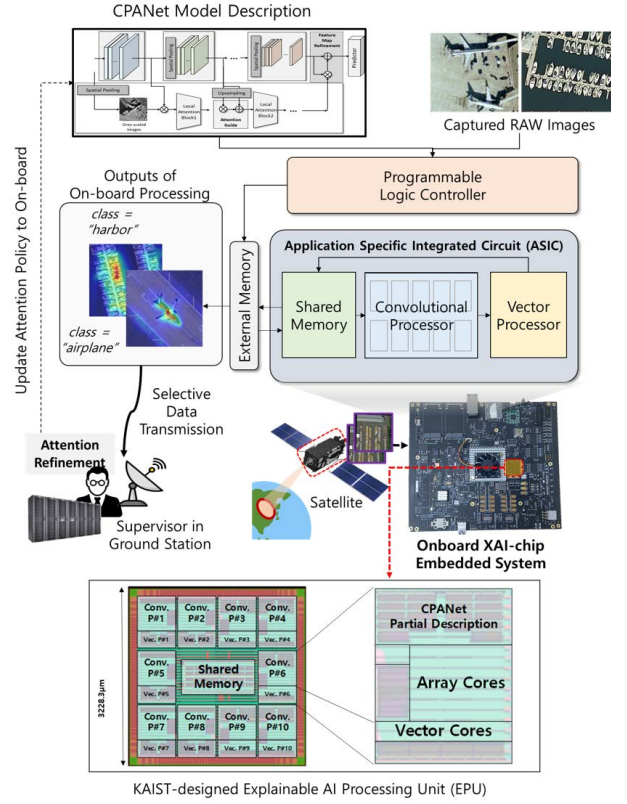


FIGURE 9. Onboard XAI-chip embedded system with CPANet prototype in a AI processor connected to the COTS Xilinx FPGA. Especially, we prototyped explainable AI processing unit (EPU) under Samsung Foundry 28-nm CMOS Process with 200mW power consumption and 7.5W in the entire onboard system.

$$\begin{aligned} & \cdot \{\nabla_{\Theta^{t-1}} M^{GT}(\Theta^{t-1}; \zeta^{t-1}) - \nabla_{\Theta^{t-1}} M^j(\Theta^{t-1}; \zeta^{t-1})\} \| \\ &= \|M^{GT}(\Theta^{t-1}; \zeta^{t-1}) - M^j(\Theta^{t-1}; \zeta^{t-1}) \\ &\quad - 2\alpha\eta \{M^{GT}(\Theta^{t-1}; \zeta^{t-1}) - \nabla_{\Theta^{t-1}} M^j(\Theta^{t-1}; \zeta^{t-1})\} \\ &\quad \cdot \{\nabla_{\Theta^{t-1}} M^{GT}(\Theta^{t-1}; \zeta^{t-1}) - \nabla_{\Theta^{t-1}} M^j(\Theta^{t-1}; \zeta^{t-1})\} \| \\ &= \| \{M^{GT}(\Theta^{t-1}; \zeta^{t-1}) - M^j(\Theta^{t-1}; \zeta^{t-1})\} \\ &\quad \cdot \{(1 - 2\alpha\eta) \{\nabla_{\Theta^{t-1}} M^{GT}(\Theta^{t-1}; \zeta^{t-1}) \\ &\quad - \nabla_{\Theta^{t-1}} M^j(\Theta^{t-1}; \zeta^{t-1})\}\} \| \\ &\leq \|M^{GT}(\Theta^{t-1}; \zeta^{t-1}) - M^j(\Theta^{t-1}; \zeta^{t-1})\| \\ &\quad \cdot \|(1 - 2\alpha\eta) \{\nabla_{\Theta^{t-1}} M^{GT}(\Theta^{t-1}; \zeta^{t-1}) \\ &\quad - \nabla_{\Theta^{t-1}} M^j(\Theta^{t-1}; \zeta^{t-1})\}\| \\ &= A(1 - 2\alpha\eta) \|M^{GT}(\Theta^{t-1}; \zeta^{t-1}) - M^j(\Theta^{t-1}; \zeta^{t-1})\|. \end{aligned} \quad (20)$$

Therefore, we can prove that explanation consistency for the pyramid feature blocks is bounded with Eq. (19). \square

Theorem 1 shows that “wrong” visual explanation can be corrected by Eq. (17). Overall procedure of the attention refinement in the ground station is described in Algorithm 1.

E. ONBOARD XAI-CHIP EMBEDDED SYSTEM IMPLEMENTATION

For explainability in the onboard XAI computing, CPANet contains huge parameter space and computation

requirements. Conventional onboard AI system [7] equipped with the low-power HW (i.e. VPU) is not suitable for visual explainability model deployed in satellite computing federation. That's why we prototyped a newly designed explainable AI processor, as shown in Fig. 9. In this section, we describe our prototype of application specific integrated circuit (ASIC) based the onboard XAI system for low-power and high computation ability. We implement the proposed CPANet on a low-power ASIC with COTS components and the embedded system. In Fig. 9, it consists of programmable logic controller, external/shared memory, convolutional/vector processor. In ASIC, the intrinsic resources are limited, therefore we perform onboard processing adaptation for CPANet weights and operators to utilize low-power ASIC. To accelerate XAI methods, we design the explainable AI processing unit (EPU) chip which is used in the onboard HW prototype.

1) PROGRAMMABLE LOGIC CONTROLLER FOR EPU

On low-power peripheral, the task partitioning control of CPANet is necessary because the input data x and CPANet information (M^L, Θ) that can be processed are limited. Since the input data size may change, the partitioned task $T_k = (\mathbf{x}_k, M_k^L, \Theta_k)$ where input data $\{x_k | x = \bigcup_k x_k\}$, $\Theta_k = (\mathbf{u}_k, \mathbf{v}_k)$ and CPANet graph operators with parameters $\{(M_k^L, \Theta_k) | (M^L, \Theta) = \bigcup_k (M_k^L, \Theta_k)\}$, is generated in the Programmable Logic Controller. To guarantee each task can be performed in the ASIC, the task size $Mem(T_k)$ should satisfy $Mem(T_k) + Mem(M_k^L(x)) \leq S$ where shared memory size has S MB, $Mem(T_k)$ equals $Mem(\mathbf{x}_k) + Mem(\Theta_k)$ because operators do not occupy the storage, and $Mem(M_k^L)$ is the k -th partial output. Finally, the partial tasks to run are stored in external memory. After processing a partitioned task of CPANet, Programmable Logic Controller loads the results from the shared memory in ASIC.

2) CONVOLUTIONAL/VECTOR PROCESSOR IN EPU

Entire operators to run CPANet are consisting of convolutional and vector computations. In convolutional processor, there are multiple array processing units. Array processing units process the set of kernel sizes 1×1 , 3×3 , and 7×7 with various zero-padding and stride sizes in parallel. Vector Processor is designed to treat MaxPool, AvgPool, Batch Normalization, ReLU activation, and GEMM. From the vector processor, the partial output $M_k^L(\mathbf{x}_k)$ is stored to shared memory. The power consumption of our onboard processing is commonly 7.25 W . The designed EPU shows that the power consumption is about 200 mW in die area 10.24 mm^2 .

IV. EXPERIMENTS AND DISCUSSION

In this section, we show the performance comparison of the proposed methods with other approaches and discuss the results.

A. EXPERIMENT SETTING AND BASELINE METHODS

We conducted experiments on UC Merced land use (UCMerced) [35] and NWPU-RESISC45 [38] dataset.

UCMerced dataset contains 21,000 images with 256×256 resolution with 21 land use classes. 90% of total images is randomly split into training, and the other 10% is used for validation. NWPU-RESISC45 contains 31,500 images with 256×256 including 45 classes with from 0.2 to 30m pixel resolution. There are 700 images per class, captured from Google Earth. We compare our methods with other attention branch methods using only top-level feature map such as ABN [25] and LFI-CAM [26]. In addition, we compared the visual explainability with class activation methods such as CAM [20], Grad-CAM [22], Grad-CAM++ [23], and LayerCAM [24] with the proposed CPANet. Note that these post-hoc explanation methods cannot influence task performance and require additional backpropagation operations except CAM. Note that our training CPANet and refinement are conducted on a server platform with NVIDIA GPUs, these stages are irrelevant to the onboard HW that is only for inference after training. We use 4 NVIDIA RTX 3080 GPUs for training and we use python 3.6.13, CUDA 11.3, and pytorch 10.2 as DL framework.

B. MODEL ARCHITECTURE AND EVALUATION METRICS

We use ResNet-18 [34] architecture as a common perception branch in the baselines and the proposed CPANet. In ResNet, we set each residual block as each pyramid feature block (i.e. $L = 4$). The image augmentation and optimizer settings are similar to ABN [25] and LFI-CAM [26]. Training images in datasets are cropped with a random ratio and resized to 224×224 and randomly adapted to horizontal flips. We use stochastic gradient descent (SGD) with momentum as the optimizer for all models. The initial learning rate is set to 0.1 with momentum to 0.9. In our experiments, the total training epochs is 200, and the learning rate decaying to 0.1 and 0.01 in 100 and 150 epochs, respectively. We use a weight decaying set to $1e-4$ for all cases.

Basically, we adapt the top-1 error (%) and the number of trainable parameters as the metrics for comparing the task performance. In addition to that, we evaluate the qualitative analysis of visual explainability on the proposed model and baselines. As a metric for evaluating explainability, we measure *maximum sensitivity* about input perturbation. In [39], *maximum sensitivity* of the explanation is defined as the maximum norm of differences of the explanation VE for a black-box model f in the input \mathbf{x} and r -perturbation input \mathbf{y} , which is defined as

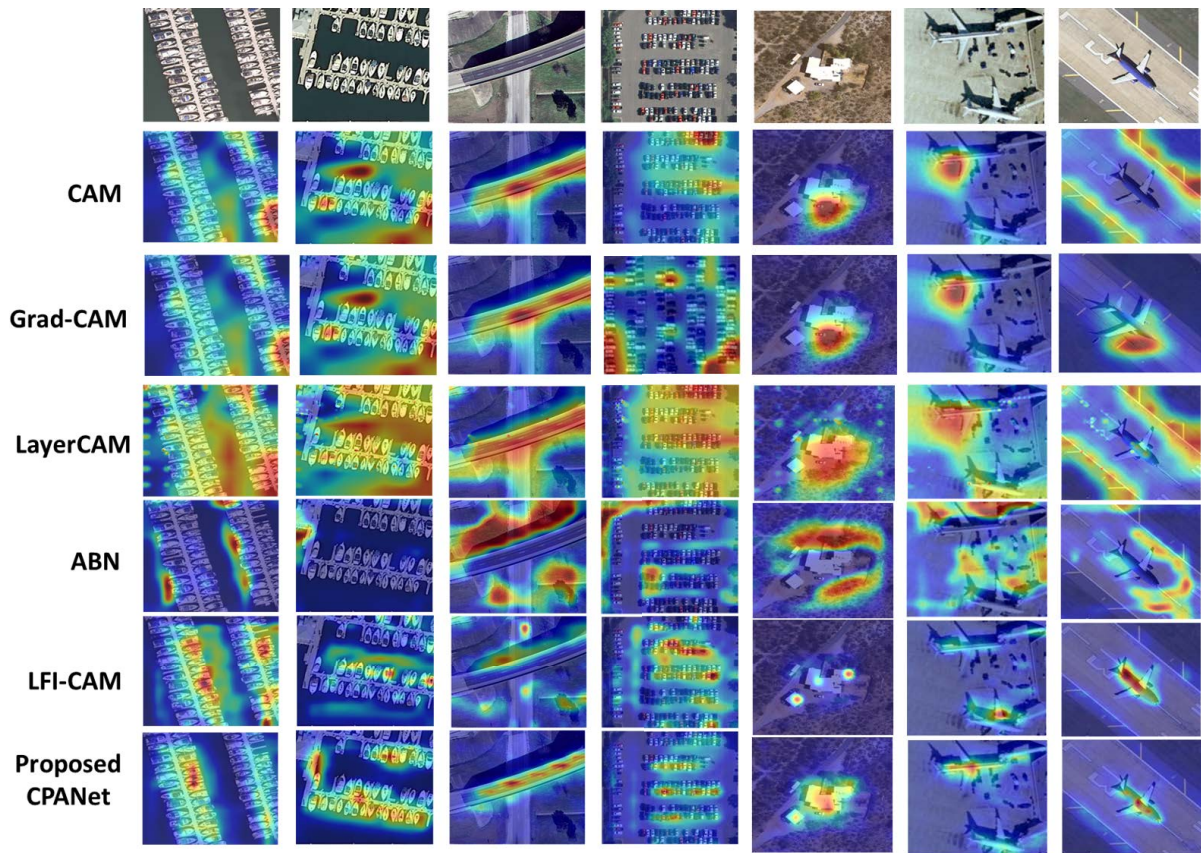
$$S_{\max}(VE, \Theta, \mathbf{x}, \mathbf{y}) = \max_{\|\mathbf{y}-\mathbf{x}\| \leq r} \|VE(\Theta; \mathbf{y}) - VE(\Theta; \mathbf{x})\|. \quad (21)$$

In every experiments, we use Gaussian noise with $(-r, +r)$ for total 100 perturbed inputs \mathbf{y} , and measure the maximum difference by Eq. (21). In satellite images, there are many environmental variances such as sunlight, noise, etc., resulting in the change in prediction and visual explanation. Therefore, this maximum sensitivity can be an indicator of how the XAI model can be robust in satellite image analysis. For validating ambiguous explanation, we use *inconsistency*

TABLE 1. Performance comparison of visual explanation methods on UCMerced and NWPU-RESISC45: The proposed CPANet vs. baseline CNN with CAM [20] and post-hoc explanation methods, and attention branch methods (ABN [25] and LFI-CAM [26]).

	UCMerced			NWPU-RESISC45		
	Params.(M)	Top-1 Err.	S_{\max} $r = 0.2$	Params.(M)	Top-1 Err.	S_{\max} $r = 0.2$
Base+CAM	11.19	6.67	0.048	11.23	9.3	0.49
+GC*			0.05			0.5
+GCpp*			0.05			0.57
+LC*			0.04			0.49
ABN	19.59	6.19	0.024	19.62	9.07	0.48
LFI-CAM	20.63	7.62	0.139	20.64	8.41	0.41
Proposed CPANet	20.59	3.81	0.11	20.61	8.08	0.4

*The post-hoc explanation methods. GC: GradCAM [22], GCpp: GradCAM++ [23], and LC: LayerCAM [24].

**FIGURE 10.** Comparison of visual explanation results on UCMerced images with the proposed CPANet and baselines.

of visual explanation between the pyramid feature blocks, described in Eq. (14). Note that inconsistency measures how much the trained XAI method focuses on the exclusive region compared to the common region between the feature maps over pyramid feature blocks. As discussed in Section II-D, it seems that the phenomenon of background bias results in a rapid change in the region where the pyramid feature block is focused, especially in the satellite images including lots of small objects. The inconsistency metric can be used to measure the changing in visual explanation. In addition, we also evaluate the explainability metrics of *average % drop* and *% of increase in confidence*, widely used in XAI

methods [23], [40], [41]. The both metrics is based on the comparison of original image \mathbf{x} with ground-truth label c and explanation map $\mathbf{x}_{exp} = \mathbf{x} \odot VE(\Theta; \mathbf{x})$. The explanation map is the synthetic image with highlighting regions where visual explanation focuses and the other regions are removed (remind that each pixel of visual explanation has $[0,1]$ floating-point value). Average % drop (denote as “average drop”) represents average drop ratio of c -confidence $p^c(\Theta; \mathbf{x})$ (see Eq. (8)) over validation dataset when the explanation map is fed into an XAI model. It means that if visual explanation misses the contexts representing the objects or highlights the uncorrelated region the average drop could

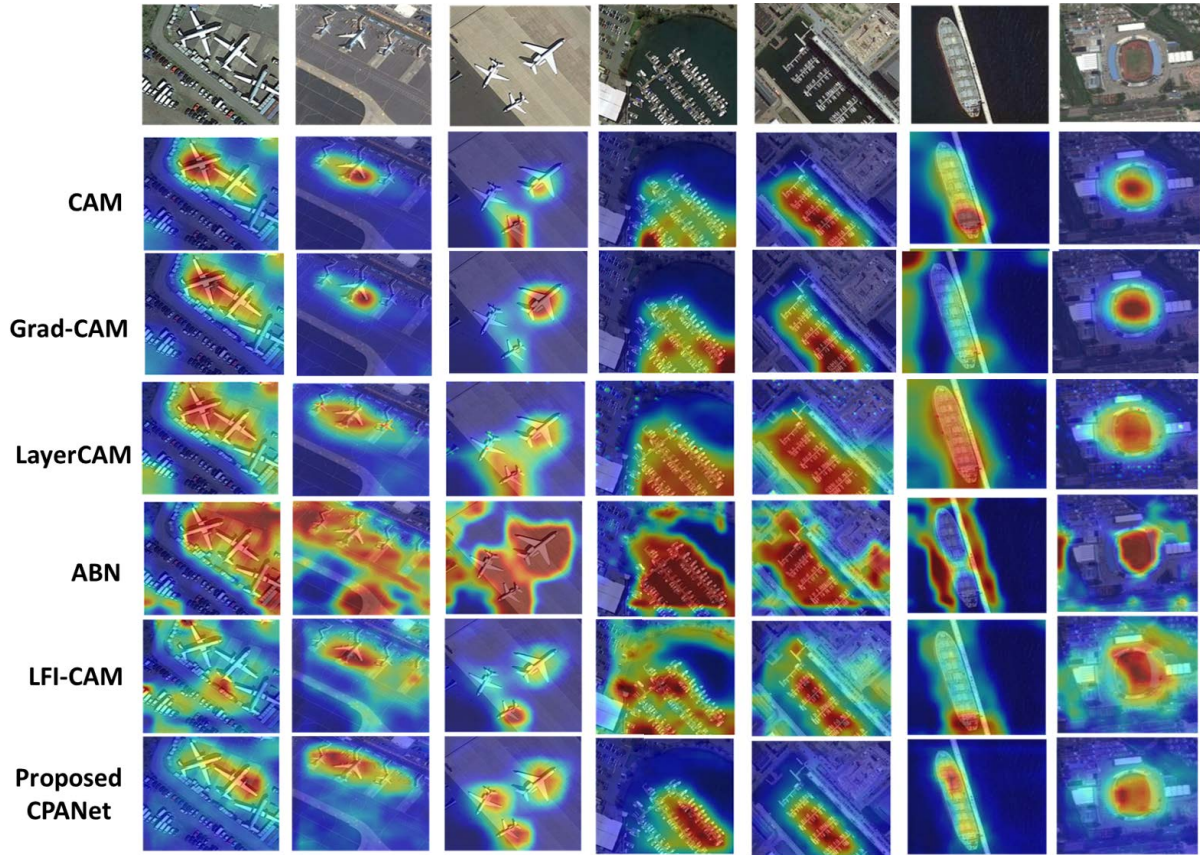


FIGURE 11. Comparison of visual explanation results on NWPU-RESISC45 images with the proposed CPANet and baselines.

be high. On the other hand, increase % in confidence is the ratio of the number of validation images with increasing c -class confidence $p^c(\Theta; \mathbf{x})$ when feeding the explanation map. It means that the explanation map strongly highlights the most discriminative region of the objects. Similar to [23], we remove 50% of lower pixels of visual explanation when generating the explanation map.

C. EVALUATION ON SATELLITE IMAGE DATASET

Table 1 shows the results of top-1 accuracy and explainability on UCMerced and NWPU-RESISC45 datasets. Comparing with conventional ResNet without any attention (see *Base*), the proposed model can achieve 2.86% and 1.22% lower top-1 error in UCMerced and NWPU-RESISC45, respectively. Compared to the existing attention branches the proposed CPANet reduces the top-1 error up to 2.4% and in UCMerced. Likewise, in NWPU-RESISC45 dataset, the proposed CPANet shows better task performance compared to conventional methods while achieving about 1% lower top-1 error. In terms of trainable parameters, all attention branch methods have a large number of parameters due to the sub-path built on the pyramid feature blocks (i.e. ResNet backbone). As mentioned above, ABN and LFI-CAM consider only generating a visual attention map from the top-level convolution layer. In the proposed model, however,

TABLE 2. The evaluation of quantitative explainability in terms of average drop (“Avg. Drop”) and increase in confidence (“Inc. in Conf.”) introduced in [23] in the proposed CPANet and baselines of attention branch methods (ABN, LFI-CAM). We used the UCMerced dataset in this experiment.

	Avg. Drop	Inc. in Conf.
ABN	79.28	4.76
LFI-CAM	74.29	7.62
Proposed CPANet	63.11	4.76

we provide visual explanation based on 4 pyramid feature blocks (i.e. last layer of each residual block) for better explainability. Although the attention branch of the proposed model covers more feature maps comparing to ABN and LFI-CAM, it only requires $\leq 1\%$ additional parameters, which shows the structural efficiency of our method. In terms of explainability measurement, CPANet is quietly worse in UCMerced dataset and better in NWPU-RESISC45 dataset. Although ABN shows the higher maximum sensitivity, we confirmed that attention was highly concentrated in the peripheral regions, even in the background. In addition, we show the enhancement of the maximum sensitivity through the attention refinement in the following section.

The qualitative results of the visual explanation of UCMerced dataset are depicted in Fig. 10. Note that we extract Grad-CAM results from the top convolution layer,

as mentioned in [22]. In terms of small objects (see “*harbor*” images), visual explanations derived from the conventional methods (CAM, ABN, LFI-CAM) that only considers the top-level convolution layer cannot accurately capture the object (ship), where the trained model rather focus on the background context (ocean) when predicting the category. Especially, ABN severely fails to capture the objects while only highlighting the background. Even in the multiple layers fusion-based explanation method (LayerCAM), the model cannot filter the background context because it just aggregates visual explanation with pixel-wise maximization from the multiple layers. This aggregation method may include unnecessary information about feature maps in visual explanation. However, the proposed method can efficiently remove these redundant feature maps biased on the background by weighting the different importance on each layer to generate visual explanation. Especially, in the results of “*parkinglot*” (fourth column), the proposed CPANet completely excludes the background, and it achieves high-quality visual explanation comparing to the other methods which highlight the background and miss some objects. Moreover, such background bias is also mitigated on the larger constructions such as freeway (third column) and buildings (fifth column) by allocating small weights on the background biased feature map with our cascading attention method. Likewise, in the “*airplane*” images (last two columns), the critical background bias occurs in the other comparing methods focusing on the airstrip together. In summary, in terms of capturing the exact boundary of the object, the proposed model shows better quality than the top-layer explanation methods (CAM, ABN, LFI-CAM). The spatial information loss in the pyramid feature blocks carries unclear object location in visual explanation, resulting in the background focusing. And, the method of multiscale feature maps (LayerCAM) cannot distinguish explainability of local explanations by highlighting both object and background. On the other hand, CPANet can distinguish between objects and backgrounds accurately in complex images. Fig. 11 also shows visual explanations about the NWPU-RESISC45 dataset. In the results, CPANet can also identify the target objects from the background and redundant objects (see the first column), while the other methods raised the background bias in the results. In the case of multiple object in a single image (see from first to third column), CPANet concentrates on all objects as fairly as possible while separating the background and unnecessary context (e.g. “*car*” in first column). By mitigating such background bias on visual explanation efficiently, our method can provide the higher explainability.

Table 2 demonstrates explainability of average drop and increase in confidence metrics for various attention branch methods in the UCMerced dataset. Note that the higher average drop and lower increase in confidence mean better explainability. ABN shows low performance compared with other methods. CPANet achieves >11.18% better average drop, which implies that the proposed CPANet generates visual explanation covering the entire objects with the small

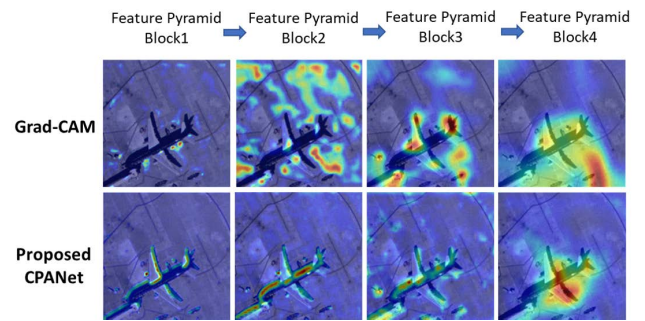


FIGURE 12. Comparison of visual explanation results in the feature pyramid blocks on a UCMerced image. The proposed CPANet can propagate the useful context in the low layer (boundary of “*airplane*”) in a cascading manner.

context while ABN and LFI-CAM can miss the discriminative parts of the objects and highlight the unnecessary region disturbing the model prediction. In contrast, increase in confidence of CPANet is worse than LFI-CAM (about 2.86%). It implies that the number of explanations highlighting the *most discriminative* region is large in LFI-CAM. Summarizing the results, LFI-CAM can highlight the most distinctive parts of the objects, but it can also miss small contexts and highlight unnecessary backgrounds. On the other hand, we can argue that the proposed CPANet can stably capture the overall contexts of the objects while showing better task accuracy and maximum sensitivity than that of LFI-CAM.

D. ABLATION STUDIES OF THE PROPOSED CPANet

In this section, we provide a detailed analysis of the advantages of the proposed CPANet. For the ablation studies, we verify visual explanation of the intermediate pyramid feature blocks. We measure the inconsistency of visual explanations of the different blocks. We compare the results with Grad-CAM which can extract visual explanation in the intermediate feature maps. We get the inconsistency w.r.t all possible combination cases of the pyramid feature blocks. For the validation dataset in UCMerced, the mean and variance of the proposed CPANet are 0.212 and 0.004 while Grad-cam shows 0.31 and 0.019, respectively. It seems that the result is due to the feature map guided by the previous attention map described in Eq. (6). Through the cascading attention blocks, the spatial information loss of the higher convolution layer can be compensated, resulting in the relaxation of the background bias. Fig. 12 illustrate visual explanation in CPANet and Grad-CAM. We can observe that the proposed model can propagate the boundary of the airplane in pyramid feature block 1 to the following blocks while Grad-CAM only focuses on the regions of the local pyramid feature block.

Table 3 demonstrates explainability results of the average drop and increase in confidence. In terms of increase in confidence, CPANet shows higher visual explainability over all the pyramid feature blocks which implies visual explanation can capture the most discriminative region compared with Grad-CAM. In the average drop, however, Grad-CAM seems

TABLE 3. The evaluation of average drop and increase in confidence of visual explanation for Grad-CAM and CPANet over the pyramid feature blocks (Block) in ResNet-18. We also use the UCMerced dataset in this experiment.

	Grad-CAM		Proposed CPANet	
	Avg. Drop	Inc. in Conf.	Avg. Drop	Inc. in Conf.
Block 1	-	3.81	68.88	5.71
Block 2	83.16	1.43	65.5	5.71
Block 3	69.88	2.86	71.46	3.81
Block 4	62.8	4.76	63.11	4.76

to be slightly high performance (i.e. lower average drop) in Block 4 and Block 3. We discuss the two perspectives from this result. First, Grad-CAM shows inconsistency of explanations in the pyramid feature blocks. Explanation in block 1 shows *NaN* value because the model fails to predict *c*-class probability in the explanation map (values to *NaN*). They show unstable explainability among the pyramid feature blocks. Second, in terms of the average drop, it can be able to argue that the proposed explanation mechanism is slightly worse than Grad-CAM (about 0.31% in Block4). Grad-CAM is the class activation map method which means that visual explanation is only about the specific class via the gradient of *c*-class confidence. Whereas, that of the attention branch methods (CPANet, ABN, LFI-CAM) is about the salient objects independent of the target classes. Therefore, if visual explanation of CPANet only highlights the object unrelated to the ground-truth class *c*, then *c*-class confidence of the explanation map may be near 0 because all regions about the target objects are removed. In Grad-CAM, however, even if the model predicts the different target class, visual explanation of *c*-class may contains the tiny regions about the target class because it is generated by only the gradient of *c*-class confidence. In this reason, the class activation map based Grad-CAM outperforms ABN and LFI-CAM in terms of the average drop (see Table 2).

E. EVALUATION OF ATTENTION REFINEMENT WITH WEAK SUPERVISION

In this section, we evaluate the attention refinement scheme containing the active learning-based sampling with inconsistent explanation and weak supervision in the ground station. From the ResNet model initially trained with the UCMerced training dataset, we measure the maximum inconsistency among the local attention maps (i.e. visual explanation of the local pyramid feature block), then sample the images with a higher value than the threshold. In this experiment, we set the threshold γ to 0.3 and weighting α to 1.0. Through the data sampling method based on the explanation inconsistency, we sample almost 700 images among total 2, 100 training samples. An example of sampled data is illustrated in Fig. 13. We can notice that the background bias occurs in the pyramid feature block 4. It results in inconsistent explanation of the pyramid feature blocks. In this case, the supervisor can resolve the background bias by setting the weak supervision

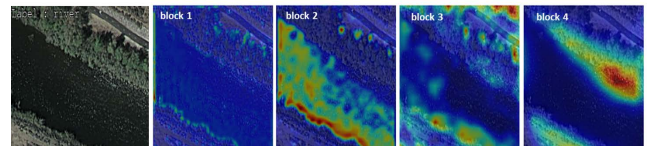


FIGURE 13. Example of the sample (ground-truth label is “river”) with inconsistent explanation of pyramid feature blocks. Supervisor can select proper visual explanation (block 2) as weak supervision.

TABLE 4. The evaluation of attention refinement based on weakly supervision for the validation dataset: *Init. Train* is the trained CPANet with training dataset and *Att. Ref* is retrained CPANet with supervisor's feedback for the images with inconsistent explanation. Note that value of inconsistency represents *mean(variance)*.

	Init. Train	Att. Ref
Top-1 error	3.81	3.81
Inconsistency \mathcal{U}	0.212(4e-3)	0.19 (2.7e-3)
Max. Sensitivity S_{max}	0.11	0.06

(i.e. select *block2* as weak ground-truth) and retraining it with loss function Eq. (17).

As the hyperparameter of the refinement, we set the learning rate to 0.01 and weight decay to $1e-4$ same as the initial training stage. The results is presented in Table 4. The results show that the top-1 error is the same after the attention refinement. It means that the DL model corrects the target class of the sampled data, which the cross-entropy loss Eq. (8) is already saturated before the refinement. However, from the perspective of explainability, the proposed attention refinement reduces the inconsistency \mathcal{U} and maximum sensitivity S_{max} . The results show that the refinement scheme with simple supervisor feedback can make CPANet generate robust explanations on the onboard.

F. COMPARISON OF COMPUTATIONAL COST FOR THE PROPOSED FEDERATED COMPUTING WITH EPU EMBEDDED SYSTEM

In this section, we discuss the computational costs of the proposed FOGS computing compared to the conventional approach (i.e. all captured images are transmitted via satellite downlink and processed on the ground station computing). In the proposed method, the interesting objects are selectively transmitted to the ground station, and the area occupied by an object may be minor in the entire satellite image. Due to the limitation of the satellite electrical power system (EPS), we define the *computational cost metric C* as the expected energy consumption for processing all captured images \mathcal{D} from a satellite. We denote $|\mathcal{D}|$ and $data(\mathcal{D})$ is the number of image patches and data volume (GB) to be transmitted \mathcal{D} . The computation cost C_{GS} of the conventional approach, *ground station (GS) only*, is given to Eq. (22),

$$C_{GS} = P_{comm} \cdot \frac{data(\mathcal{D})}{R_{comm}} + P_{GS} \cdot (|\mathcal{D}| \cdot t_{GS}). \quad (22)$$

P_{GS} is the active power consumption of ground station HW, t_{GS} is the processing time per an image, P_{comm} and R_{comm} are the

TABLE 5. Comparison of computational costs of the conventional approach (ground station only) C_{GS} and federated onboard-ground station (FOGS) computing C_{FOGS} with various onboard HW based on Eqs. 22 and 23.

C_{GS} (KJ)		C_{FOGS} (KJ)	
NVIDIA RTX 3080		Xilinx XCZU7EV embedded system [45]	EPU embedded system
18.06	$\rho=1.0$	14.26	1.43
	$\rho=0.5$	13.7	1.07
	$\rho=0.3$	13.42	0.59
	$\rho=0.1$	13.2	0.35

transmission power and transmission rate from the satellite to the ground station, respectively. Similarly, the computational cost C_{FOGS} of FOGS is given to Eq. (23),

$$C_{FOGS} = P_{OBD} \cdot (|\mathcal{D}| \cdot t_{OBD}) + P_{comm} \cdot \frac{data(\rho \cdot \mathcal{D})}{R_{comm}}. \quad (23)$$

P_{OBD} and t_{OBD} are the active computation power and processing time of onboard (OBD) HW, and ρ is the ratio of the selected image patches (i.e. interesting targets from the ground station).

Then, we evaluate the computational cost in the case of processing $|\mathcal{D}| = 100,000$ captured image patches (each image has 224×224 spatial dimension). Following the SANSA parameters referred from [42], we set $P_{comm} = 1W$ and $R_{comm} = 200Mbps$. We assume that the processing time calculates the total number of operations divided by the maximum TOPS (Trillion Operation per Second) of the HW accelerator for the simulation. And, we use the image compression ratio for captured image patches followed by consultative committee for space data systems (CCSDS) 123.0-B-2 standard [43] (in the case of the lossless compression). Table 5 shows the result of computational costs of energy consumption (Kilo Joule; KJ) in various settings. From the result, it seems to be that the filtering patches in FOGS computing are effective in terms of the computational cost C , providing a lower energy consumption for processing the given images. Xilinx XCZU7EV based onboard system, the COTS HW evaluating the onboard CloudScout model by [44], shows the higher cost ($> 9.9\times$) than the EPU embedded system (in Section III-E) due to its low power consumption. In the low-power HW setting, the effect of the proposed onboard-ground station computing is more pronounced. The result implies that the low-power engineering of the onboard HW is mandatory for processing for the onboard XAI computing.

V. CONCLUSION

In this paper, we proposed a federated onboard-ground station computing framework for satellite image analysis. For reliable analysis in complex space-related applications, especially in object recognition, we introduce a novel XAI method with CPANet in the onboard processing. By utilizing rich information for explainability in the multiple pyramid feature blocks, the proposed model improves not only visual explainability in terms of robustness in data perturbation but the task

performance. In addition, we propose the onboard refinement scheme with the supervisor's feedback. Using weak supervision, the proposed refinement mechanism can reduce the cost of supervisor annotation, and improve visual explainability. In future work, we are going to extend the architecture to the object detection task and develop the prototype system with a low-power AI accelerator. Due to the limited electrical power system (EPS) in the satellite, an onboard AI model should be light-weight. Though the proposed CPANet can improve accuracy and visual explainability effectively, it requires additional computation. Therefore, we additionally consider co-design of the network compression (pruning and weight quantization) for the implementation. Then, we will validate the feasibility of our onboard system in terms of processing time and power consumption.

ACKNOWLEDGMENT

(Taewoo Kim and Minsu Jeon contributed equally to this work.)

REFERENCES

- [1] X. Zhou, K. Shen, Z. Liu, C. Gong, J. Zhang, and C. Yan, "Edge-aware multiscale feature integration network for salient object detection in optical remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–15, 2021.
- [2] Q. Ming, Z. Zhou, L. Miao, H. Zhang, and L. Li, "Dynamic anchor learning for arbitrary-oriented object detection," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, no. 3, 2021, pp. 2355–2363.
- [3] Z. Lv, T. Liu, and J. A. Benediktsson, "Object-oriented key point vector distance for binary land cover change detection using VHR remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 9, pp. 6524–6533, Sep. 2020.
- [4] H. Kim, K. Lee, C. Lee, S. Hwang, and C.-H. Youn, "An alternating training method of attention-based adapters for visual explanation of multi-domain satellite images," *IEEE Access*, vol. 9, pp. 62332–62346, Apr. 2021.
- [5] W.-J. Kim and C.-H. Youn, "Cooperative scheduling schemes for explainable DNN acceleration in satellite image analysis and retraining," *IEEE Trans. Parallel Distrib. Syst.*, vol. 33, no. 7, pp. 1605–1618, Jul. 2022.
- [6] Q. Wang, W. Huang, Z. Xiong, and X. Li, "Looking closer at the scene: Multiscale representation learning for remote sensing image scene classification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 4, pp. 1414–1428, Apr. 2020.
- [7] G. Giuffrida, L. Fanucci, G. Meoni, M. Batic, L. Buckley, A. Dunne, C. Van Dijk, M. Esposito, J. Hefele, N. Vercruyssen, G. Furano, M. Pastena, and J. Aschbacher, "The Θ -Sat-1 mission: The first on-board deep neural network demonstrator for satellite Earth observation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–14, 2022.
- [8] K. Xiao, L. Engstrom, A. Ilyas, and A. Madry, "Noise or signal: The role of image backgrounds in object recognition," 2020, *arXiv:2006.09994*.
- [9] S. Mo, H. Kang, K. Sohn, C.-L. Li, and J. Shin, "Object-aware contrastive learning for debiased scene representation," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 12251–12264.
- [10] N. Buonaiuto, M. Louie, J. Aarestad, R. Mital, D. Mateik, R. Sivilli, A. Bhopale, C. Kief, and B. Zufelt, "Satellite identification imaging for small satellites using NVIDIA," in *Proc. Small Satell. Conf.*, 2017, pp. 1–12.
- [11] V. Kothari, E. Liberis, and N. D. Lane, "The final frontier: Deep learning in space," in *Proc. 21st Int. Workshop Mobile Comput. Syst. Appl.*, 2020, pp. 45–49.
- [12] E. Dunkel, J. Swope, Z. Towfic, S. Chien, D. Russell, J. Sauvageau, D. Sheldon, J. Romero-Canas, J. L. Espinosa-Aranda, L. Buckley, E. Hervas-Martin, M. Fernandez, and C. Knox, "Benchmarking deep learning inference of remote sensing imagery on the Qualcomm snapdragon and Intel Movidius Myriad X processors onboard the international space station," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2022, pp. 5301–5304.

- [13] Z. Towfic, D. Ogbe, J. Sauvageau, D. Sheldon, A. Jongeling, S. Chien, F. Mirza, E. Dunkel, J. Swope, M. Ogut, V. Cretu, and C. Pagnotta, "Benchmarking and testing of Qualcomm snapdragon system-on-chip for JPL space applications and missions," in *Proc. IEEE Aerosp. Conf. (AERO)*, Mar. 2022, pp. 1–12.
- [14] 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, 2015, pp. 234–241.
- [15] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 25, F. Pereira, C. J. Burges, L. Bottou, K. Q. Weinberger, Eds. Red Hook, NY, USA: Curran Associates, 2012, pp. 84–90.
- [16] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*.
- [17] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should i trust you? Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2016, pp. 1135–1144.
- [18] M. Scott Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. NeurIPS*, 2017, pp. 1–10.
- [19] V. Petsiuk, A. Das, and K. Saenko, "RISE: Randomized input sampling for explanation of black-box models," 2018, *arXiv:1806.07421*.
- [20] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016.
- [21] M. Lin, Q. Chen, and S. Yan, "Network in network," in *Proc. ICLR*, 2014, pp. 1–10.
- [22] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, vol. 128, Oct. 2017, pp. 618–626.
- [23] A. Chattopadhyay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2018, pp. 839–847.
- [24] P.-T. Jiang, C.-B. Zhang, Q. Hou, M.-M. Cheng, and Y. Wei, "LayerCAM: Exploring hierarchical class activation maps for localization," *IEEE Trans. Image Process.*, vol. 30, pp. 5875–5888, 2021.
- [25] H. Fukui, T. Hirakawa, T. Yamashita, and H. Fujiyoshi, "Attention branch network: Learning of attention mechanism for visual explanation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 10705–10714.
- [26] K. H. Lee, C. Park, J. Oh, and N. Kwak, "LFI-CAM: Learning feature importance for better visual explanation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 1355–1363.
- [27] H. Li, P. Xiong, J. An, and L. Wang, "Pyramid attention network for semantic segmentation," in *Proc. BMVC*, 2018, pp. 1–13.
- [28] W. Wang, S. Zhao, J. Shen, C. H. Steven Hoi, and A. Borji, "Salient object detection with pyramid attention and salient edges," in *Proc. CVPR*, 2019, pp. 1–10.
- [29] T. Zhao and X. Wu, "Pyramid feature attention network for saliency detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3085–3094.
- [30] M. Jiang, J. Xu, and Q. Zhao, "Saliency in crowd," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 17–32.
- [31] A. Das, H. Agrawal, L. Zitnick, D. Parikh, and D. Batra, "Human attention in visual question answering: Do humans and deep networks look at the same regions?" *Comput. Vis. Image Understand.*, vol. 163, pp. 90–100, Oct. 2017.
- [32] R. R. Selvaraju, S. Lee, Y. Shen, H. Jin, S. Ghosh, L. Heck, D. Batra, and D. Parikh, "Taking a HINT: Leveraging explanations to make vision and language models more grounded," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 2591–2600.
- [33] J. Wu and R. Mooney, "Self-critical reasoning for robust visual question answering," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–11.
- [34] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
- [35] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proc. 18th SIGSPATIAL Int. Conf. Adv. Geographic Inf. Syst. (GIS)*, 2010, pp. 270–279.
- [36] N. Komodakis and S. Zagoruyko, "Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer," in *Proc. ICLR*, 2017, pp. 1–13.
- [37] Z. Huang, Y. Zou, B. Kumar, and D. Huang, "Comprehensive attention self-distillation for weakly-supervised object detection," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 16797–16807.
- [38] G. Cheng, J. Han, and X. Lu, "Remote sensing image scene classification: Benchmark and state of the art," *Proc. IEEE*, vol. 105, no. 10, pp. 1865–1883, Oct. 2017.
- [39] C.-K. Yeh, C.-Y. Heish, and A. S. Suggala, "On the (in)fidelity and sensitivity of explanations," in *Proc. NeurIPS*, 2019, pp. 1–12.
- [40] H. Wang, Z. Wang, M. Du, F. Yang, Z. Zhang, S. Ding, P. Mardziel, and X. Hu, "Score-CAM: Score-weighted visual explanations for convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 24–25.
- [41] S. Desai and H. G. Ramaswamy, "Ablation-CAM: Visual explanations for deep convolutional network via gradient-free localization," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 983–991.
- [42] Y. Wang, J. Zhang, X. Zhang, P. Wang, and L. Liu, "A computation offloading strategy in satellite terrestrial networks with double edge computing," in *Proc. IEEE Int. Conf. Commun. Syst. (ICCS)*, Dec. 2018, pp. 450–455.
- [43] M. Hernández-Cabrero, A. B. Kiely, M. Klimesh, I. Blanes, J. Ligo, E. Magli, and J. Serra-Sagrà, "The CCSDS 123.0-B-2 'low-complexity lossless and near-lossless multispectral and hyperspectral image compression' standard: A comprehensive review," *IEEE Geosci. Remote Sens. Mag.*, vol. 9, no. 4, pp. 102–119, Feb. 2021.
- [44] E. Rapuano, G. Meoni, T. Pacini, G. Dinelli, G. Furano, G. Giuffrida, and L. Fanucci, "An FPGA-based hardware accelerator for CNNs inference on board satellites: Benchmarking with myriad 2-Based solution for the CloudScout case study," *Remote Sens.*, vol. 13, no. 8, p. 1518, Apr. 2021.



TAEWOO KIM received the B.S. degree in electrical engineering from Kyungpook National University, Daegu, South Korea, in 2015, and the M.S. degree in electrical engineering from the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea, in 2017, where he is currently pursuing the Ph.D. degree in electrical engineering. His research interests include deep learning (DL) framework, GPU computing for DL, and explainable AI (XAI).



MINSU JEON received the B.S. degree in electronic engineering from Sogang University, in 2016, and the M.S. degree in electrical engineering from the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea, in 2017, where he is currently pursuing the Ph.D. degree in electrical engineering. His research interests include deep learning (DL) application/model, DL serving, and high performance computing systems.



CHANGHA LEE received the B.S. degree in electronic engineering from Hanyang University, Seoul, South Korea, in 2018, and the M.S. degree in electronic engineering from the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea, in 2020, where he is currently pursuing the Ph.D. degree. Since 2018, he has been a member of Network and Computing Laboratory at KAIST. His current research interests include deep learning acceleration platform, continual learning, and integrated system for explainable AI.



network inference/training accelerators.

JUNSOO KIM (Graduate Student Member, IEEE) received the B.S. degree from the School of Electrical and Computer Engineering, Ulsan National Institute of Science and Technology (UNIST), Ulsan, South Korea, in 2021. He is currently pursuing the M.S. degree in electrical engineering with the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea. His research interests include low-power system on-chip design and energy-efficient deep-neural

Azure Data Lake. He was also one of the initial members of the Catapult project at Microsoft Research, Redmond, where he deployed a fabric of field-programmable gate arrays (FPGAs) in data centers to accelerate critical cloud services, such as machine learning, data storage, and networking. His research interests include various aspects of hardware design, including VLSI design, computer architecture, FPGA, domain-specific accelerators, hardware/software co-design, and agile hardware development. He was a recipient of the 2016 IEEE Micro Top Picks Award, the 2014 IEEE Micro Top Picks Award, the 2010 DAC/ISSCC Student Design Contest Award, the 2008 DAC/ISSCC Student Design Contest Award, and the 2006 A-SSCC Student Design Contest Award. He currently serves as an Associate Editor for the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS—I: REGULAR PAPERS.



GEONWOO KO (Member, IEEE) received the B.S. degree in biomedical engineering and electrical engineering from Korea University, Seoul, South Korea, in 2022. He is currently pursuing the M.S. degree in electrical engineering with the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea. His research interests include near-data processing, domain-specific accelerators, and low-power system-on-chip design.



CHAN-HYUN YOUN (Senior Member, IEEE) received the B.Sc. and M.Sc. degrees in electronics engineering from Kyungpook National University, Daegu, South Korea, in 1981 and 1985, respectively, and the Ph.D. degree in electrical and communications engineering from Tohoku University, Japan, in 1994. Before joining the university, from 1986 to 1997, he was the Head of the High-Speed Networking Team, KT Telecommunications Network Research Laboratories. Since

1997, he has been a Professor at the School of Electrical Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea. He was an Associate Vice-President of office of planning and budgets at KAIST, from 2013 to 2017. He is the Director of the Grid Middleware Research Center and the XAI Acceleration Technology Research Center, KAIST, where he is developing core technologies that are in the areas of high performance computing, explainable AI systems, satellite imagery analysis, and satellite onboard computing with deep learning acceleration systems. He wrote a book on *Cloud Broker and Cloudlet for Workflow Scheduling* (Springer, in 2017). He served as a TPC member for many international conferences. He was selected to the inaugural class of IEEE Computer Society Distinguished Contributor, in 2021. He was the General Chair of the 6th EAI International Conference on Cloud Computing (Cloud Comp 2015), KAIST, in 2015. He was a Guest Editor of IEEE WIRELESS COMMUNICATIONS, in 2016.



engineering Lead at Microsoft Azure, Redmond, WA, USA, working on hardware acceleration for its hyper-scale big data analytics platform named

JOO-YOUNG KIM (Senior Member, IEEE) received the B.S., M.S., and Ph.D. degrees in electrical engineering from the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea, in 2005, 2007, and 2010, respectively. He is currently an Assistant Professor with the School of Electrical Engineering, KAIST. He is also the Director of the AI Semiconductor Systems Research Center. Before joining KAIST, he was a Senior Hardware Engi-

...