remote sensing

MDPI

Article

Credible Remote Sensing Scene Classification Using Evidential Fusion on Aerial-Ground Dual-View Images

Kun Zhao, Qian Gao, Siyuan Hao, Jie Sun and Lijian Zhou *

School of Information and Control Engineering, Qingdao University of Technology, Qingdao 266520, China
* Correspondence: zhoulijian@qut.edu.cn

Abstract: Due to their ability to offer more comprehensive information than data from a single view, multi-view (e.g., multi-source, multi-modal, multi-perspective) data are being used more frequently in remote sensing tasks. However, as the number of views grows, the issue of data quality is becoming more apparent, limiting the potential benefits of multi-view data. Although recent deep neural network (DNN)-based models can learn the weight of data adaptively, a lack of research on explicitly quantifying the data quality of each view when fusing them renders these models inexplicable, performing unsatisfactorily and inflexibly in downstream remote sensing tasks. To fill this gap, in this paper, evidential deep learning is introduced to the task of aerial-ground dual-view remote sensing scene classification to model the credibility of each view. Specifically, the theory of evidence is used to calculate an uncertainty value which describes the decision-making risk of each view. Based on this uncertainty, a novel decision-level fusion strategy is proposed to ensure that the view with lower risk obtains more weight, making the classification more credible. On two well-known, publicly available datasets of aerial-ground dual-view remote sensing images, the proposed approach achieves state-of-the-art results, demonstrating its effectiveness.

Keywords: multi-view data fusion; remote sensing scene classification; uncertainty estimation; evidential learning; Dirichlet distribution

Citation: Zhao, K.; Gao, Q.; Hao, S.; Sun, J.; Zhou, L. Credible Remote Sensing Scene Classification Using Evidential Fusion on Aerial-Ground Dual-View Images. *Remote Sens.* **2023**, *15*, 1546. https://doi.org/10.3390/rs15061546

Academic Editor: Salah Bourennane

Received: 26 December 2022
Revised: 2 March 2023
Accepted: 9 March 2023
Published: 11 March 2023

1. Introduction

Remote sensing scene classification is one of the fundamental tasks and research hotspots in the field of remote sensing information analysis, which is of great significance to the management of natural resources and urban activities [1,2]. Over the last few decades, significant progress has been made in designing efficient models for data from a single source, such as hyperspectral [3], synthetic aperture radar [4], very high-resolution images [5,6], and so forth. However, remote sensing scene classification is still regarded as a challenging task [7] when using only overhead images due to their lack of diverse detailed information.

Fortunately, with the rise of various social media platforms (e.g., Flickr) and mapping software (e.g., Google Maps, Google Street View), it is becoming easier to collect geo-tagged data from various sources, modals and perspectives. From a data processing standpoint, the various types of data mentioned above that describe the same object are commonly referred to as "multi-view" data. The fusion approaches for multi-view data can be implemented at three levels [8], namely the data-level, the feature-level and the decision-level. The data-level fusion strategy usually fuses raw or pre-processed data from multi-resolution images [9] or multi-spectral images [10], and so forth, usually appearing in early works. Feature-level fusion [11–22], on the other hand, combines multiple intermediate features extracted from multi-view data. The fused features were then used in downstream tasks. After the emergence of DNN, feature-level fusion often adopts the network structure of multi-input-single-output [14,17–19,21,22], in which some layers concatenate the intermediate features of multi-views into subsequent layers. Decision-level fusion adopts different

fusion rules to aggregate predictions from multiple classifiers, each of which is obtained from a separate model. It works best for data from various modals [23–26] because the features of each view are learned separately and independently of one another. When fusing data from the same modal [27–29] the decision-level fusion also performs well.

Among many multi-view data for remote sensing scene classification, aerial-ground dual-view images are frequently used due to their widespread geographic availability and easy of access [30]. Unlike other commonly used multi-view remote sensing images (e.g., hyperspectral and multi-spectral [20], hyperspectral and LiDAR [21]), the difference between the matched aerial image and ground image are too huge to be fused at data-level directly. Thus, fusion methods at feature-level and decision-level are often used. The first application of aerial-ground dual-view image fusion was for image geo-localization. Lin et al. [11] matched high resolution orthophoto (HRO) with street view images from Panoramio using four handcrafted features. To extend their method, they used a two-stream CNN to learn deep features between Google Street View (GSV) images and 45° aerial images [12]. Workman et al. [14] fused aerial images and GSV images by an DNN outputting a pixel-wise annotation for three different classification tasks: land use, building function and building age. Zhang et al. [15] fused thirteen parcel features from airborne light detection and ranging (LiDAR) data, HRO and GSV images for land use classification using a random forest classifier. Cao et al. [17] fused images from Bing Maps and GSV for land use segmentation using a two-stream encoder and one decoder architecture. Hoffmann et al. [18] also used a two-stream CNN to fuse overhead images and street view images for building functions classification. They significantly improved the performance using a decision-level fusion approach. Srivastava et al. [19] used a multi-input-single-output CNN to learn the complementarity of aerial-ground dual-view, which could deal with the situation of missing aerial images using canonical correlation analysis (CCA). Machado et al. [31] published two open datasets for scene classification based on aerial-ground dual-view images. A thorough evaluation of the commonly used feature-level and decision-level fusion methods is carried out on the datasets.

However, the use of multi-source data also brings about an increase in sample noise, namely the incompatibility between the visual content of the image and its semantic label. In the field of uncertainty estimation, sample noise is usually described by the term sample uncertainty [32]. Higher sample uncertainty means more serious sample noise and greater prediction risk. For example, Figure 1a shows different kinds of sample uncertainty in aerial images. In Figure 1b, new types of sample uncertainty are introduced after the addition of the ground view, which increases the risk of final prediction. Unfortunately, so far there is not enough research on sample uncertainty in multi-view remote sensing data fusion.

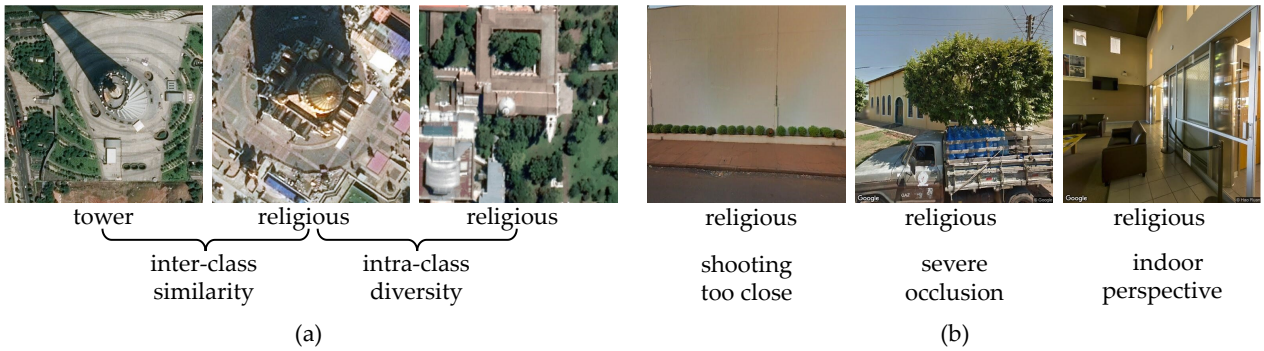


Figure 1. Different kinds of sample uncertainty in (**a**) aerial images and (**b**) ground images.

From the idea of explicitly modeling the sample uncertainty, a novel fusion approach is proposed in this paper based on evidential deep learning [33] for remote sensing scence classification on aerial-ground dual-view images. The primary contributions of this paper are as follows.

- A Theory of Evidence Based Sample Uncertainty Quantification (TEB-SUQ) approach is used in both views of aerial and ground images to measure the decision-making risk during their fusion.
- An Evidential Fusion strategy is proposed to fuse aerial-ground dual-view images at decision-level for remote sensing scene classification. Unlike other existing decision-level fusion methods, the proposed strategy focuses the results not only on the classification probability but also on the decision-making risk of each view. Thus, the final result will depend more on the view with lower decision-making risk.
- A more concise loss function, namely Reciprocal Loss is designed to simultaneously constrain the uncertainty of individual view and of their fusion. It can be used not only to train an end-to-end aerial-ground dual-view remote sensing scene classification model, but also to train a fusion classifier without feature extraction.

2. Data and Methodology

In this section, the Evidential Fusion Network (EFN) for remote sensing scene classification on aerial-ground dual-view images is introduced. First, two public aerial-ground dual-view images datasets used in this paper are described. Second, overall network architecture is introduced briefly. Third, we detail how to obtain sample uncertainty with evidential deep learning and how to perform the evidential fusion. Finally, the proposed Reciprocal Loss used to train the network is described.

2.1. Datasets Description

In this paper, performance of the proposed approach has been verified through experiments on two public aerial-ground dual-view images datasets [31].

The AiRound dataset is a collection of landmarks from all over the world. It consists of images from three different views: the Sentinel-2 images, the high-resolution RGB aerial images, and the ground images. Sentinel-2 images have a size of 224 × 224 pixels. Aerial images have a size of 500 × 500 pixels. Ground images are obtained from two different ways, namley, Google Places' database and Google Images. Thus they have different sizes. Their labels are obtained from the publicly available data of the OpenStreetMap. As shown in Figure 2a, there are 11 different land use classes in AiRound with a total of 11,753 groups of images. The aerial and ground images are used in our experiments.

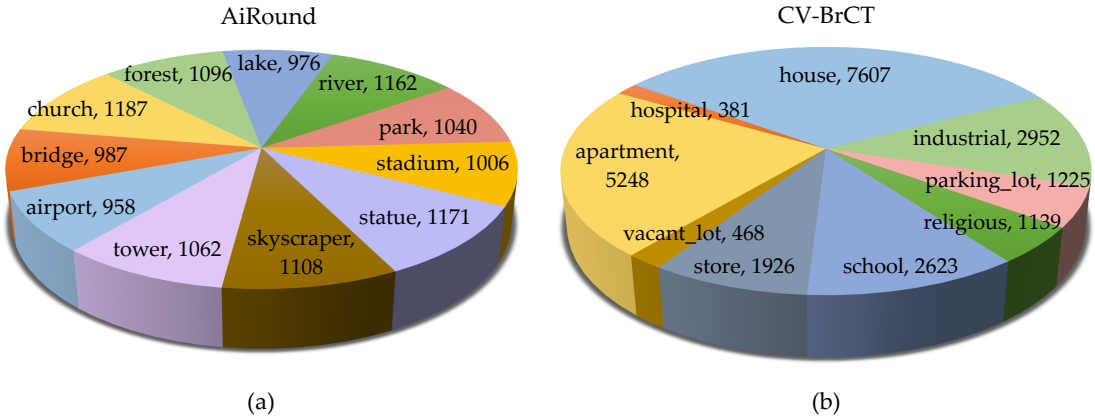


Figure 2. Class distribution of the (**a**) AiRound and (**b**) CV-BrCT dataset.

The CV-BrCT dataset contains 23,569 pairs of images in 9 classes: apartment, hospital, house, industrial, parking lot, religious, school, store, and vacant lot. Each pairs are composed of an aerial view image and a ground view image, both of which are 500 × 500 RGB images. The class distribution is shown in Figure 2b.

2.2. Overview

In the task of remote sensing scene classification on aerial-ground dual-view images, methods of data-level fusion and feature-level fusion are not satisfactory. This is because both the raw image and the visual feature of the aerial view differ too greatly from those of the ground view to meet the data requirements of data-level fusion and feature-level fusion: data-level fusion requires strict alignment of raw data from different views, well feature-level fusion requires that the dimensions and value ranges of intermediate features extracted from different views should not be too different. Recent works [18,31] show that decision-level fusion using a deep learning framework usually performs best. However, due to the lack of measurement of their credibility, multi-view decisions play equal roles during the fusion, despite the fact that their samples have unequal uncertainty. The proposed evidential fusion network (EFN) can assess the decision-making risk of each view by quantifying the sample uncertainty explicitly, and then assign different weight to the decisions base on their risks when fusing them. As a result, the final classification relies more on the view with lower decision-making risk therefore become more credible. The overall framework of EFN is shown in Figure 3. It can be further divided into three modules: the feature extraction module (FEM), the uncertainty quantification module (UQM) and the evidential fusion module (EFM).

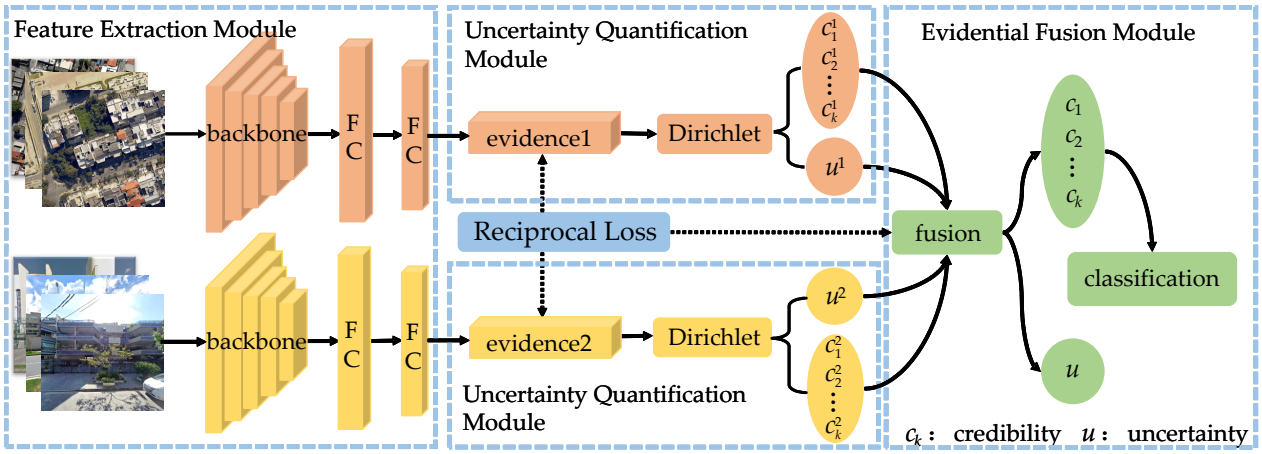


Figure 3. The overall framework of proposed Evidential Fusion Network (EFN) on aerial-ground dual-view images.

The trained backbone extracts the features in each view and feeds them into a fully connected (FC) layer with a non-negative activation function to generate the "evidence vector" for uncertainty computation. Then, the evidence vector of each view is individually mapped to the concentration parameters of a Dirichlet distribution to estimate its credibility and uncertainty (see Section 2.3). Finally, they are fused by a weighted decision-level fusion (see Section 2.4).

The proposed FEM is divided into two subnets, each with a backbone and two extra FC layers. The backbones of two the subnets could be the same or different because the proposed approach is independent of them. It is this flexibility that makes the approach applicable to any network structure. Finally, the last softmax layer is replaced with a non-negative activation layer, such as softplus. This simple replacement operation makes the proposed approach extremely portable.

2.3. Uncertainty Estimation Based on Evidential Learning

The use of the softmax operator to transform the output continuous activations into discrete class probabilities is widely recognized for classification networks. However, recent studies show that softmax operator has many shortcomings. Firstly, it may cause the issue of "over-confidence" [34]. For example, the scores of the three categories in Figure 4a are very close, indicating that the model is unsure about the input sample. However, after the

softmax mapping in Figure 4b, the difference between the probabilities is greatly increased, giving the model the false impression of being very certain. Secondly, softmax can only provide a point estimate for the category probability of a sample without providing the associated uncertainty, which often leads to unreliable conclusions [35]. Figure 5 shows the difference between the softmax and an uncertainty estimation operation on a case of a binary classification problem. In Figure 5a, a class probability distribution is generated by softmax, where p_i is the probability of class i, and we have $p_1 + p_2 = 1$ and $p_1 < p_2$, which means that the classifier is more inclined to classify the sample into class 2. However, we have no idea whether this prediction is credible.

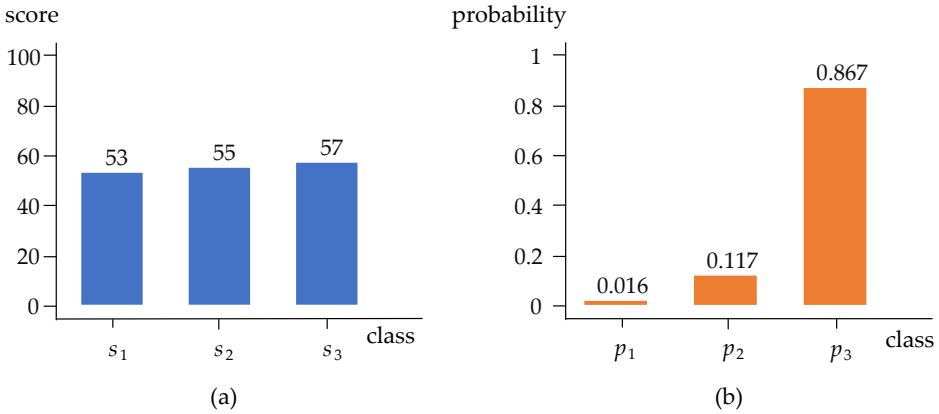


Figure 4. The "over-confidence" caused by softmax: (**a**) are the probable scores of three categories after feature extraction and (**b**) are their corresponding probabilities mapped by softmax.

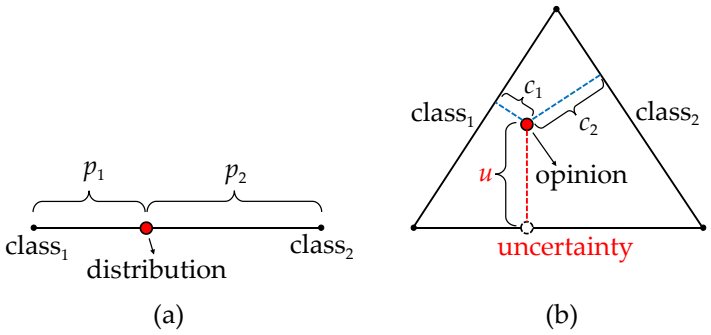


Figure 5. The case of a binary classification problem: (**a**) is the class probability distribution of a sample generated by softmax; (**b**) is the "opinion" of the sample generated by the an uncertainty estimation operation where u is the uncertainty and c_i is the credibility of class i.

To address the shortcomings of softmax, a non-negative activation function is used in place of the softmax operator to output the "evidence" for each class. Based on these evidences, the sample uncertainty can be described by the theory of evidence (TE) [36], which allows to explicitly express "ignorance" by taking the "base plane" (see Figure 6b) into consideration when building the "opinion space". In Figure 5b, an "opinion" (the red point in the equilateral triangle) of sample is generated, where c_i is the credibility of class i which is equal to the distance from the "opinion" to the side representing class i, and u is the sample uncertainty which is equal to the distance from the "opinion" to the base. It is well known that the sum of the distances from any point in an equilateral triangle to its three sides is equal to the height of the equilateral triangle. When setting the height of the equilateral triangle to 1, We have $c_1 + c_2 + u = 1$ and $p_1 < p_2 < u$, which means that the classifier is more likely to find the sample unreliable than to make a binary decision. Figure 6 shows examples with three classes where the opinion equilateral triangle in Figure 5 rises its dimension to become a opinion tetrahedron [37]. In a more general

sense, in a space with K classes, the K credibility values c_k and the uncertainty u are both non-negative and sum to one, that is,

$$u + \sum_{k=1}^{K} c_k = 1, \tag{1}$$

where $u \geq 0$ and $c_k \geq 0$ for $k = 1, 2, \cdots, K$, denote the overall uncertainty and the credibility of k-th class respectively. For each view, the credibility of each class can be calculated using the evidence of each class. The non-negative evidence set of a sample can be denoted as $\mathbf{e} = [e_1, e_2, \cdots, e_K]$. Let e_k be the evidence of k-th class of the sample, the credibility c_k and the uncertainty u can be calculated as

$$c_k = \frac{e_k}{\sum_{k=1}^{K}(e_k + 1)}, \tag{2}$$

$$u = \frac{K}{\sum_{k=1}^{K}(e_k + 1)}. \tag{3}$$

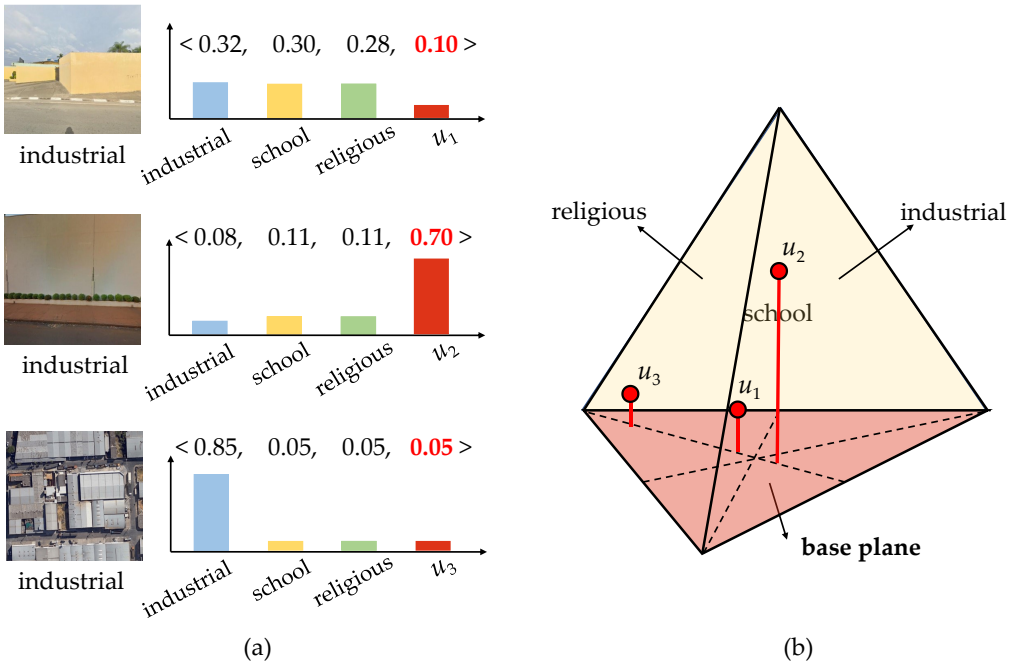


Figure 6. A multi-classification case of the proposed uncertainty estimation operation: (**a**) shows the "opinions" of three samples (two ground view image and one aerial image); (**b**) is the Opinion tetrahedron with example opinions.

The notion of TE could be further formalized as a Dirichlet distribution [37] with concentration parameter

$$\alpha_k = e_k + 1. \tag{4}$$

That is, the credibility and uncertainty can be easily obtained from the corresponding Dirichlet distribution using the following equations respectively:

$$c_k = \frac{\alpha_k - 1}{\alpha_0}, \tag{5}$$

$$u = \frac{K}{\alpha_0}, \tag{6}$$

$$\alpha_0 = \sum_{k=1}^{K} \alpha_k. \tag{7}$$

For a K-classification problem, the concentration parameter $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \cdots, \alpha_K]$, and the Dirichlet distribution is given by

$$D(\mathbf{p}|\boldsymbol{\alpha}) = \begin{cases} \frac{1}{B(\boldsymbol{\alpha})} \prod_{k=1}^{K} p_k^{\alpha_k - 1} & \mathbf{p} \in S_K \\ 0 & otherwise \end{cases}, \tag{8}$$

where $B(\boldsymbol{\alpha})$ is a K-dimensional multivariate beta function and $\mathcal{S}_K$ is the K-dimensional unit simplex,

$$\mathcal{S}_K = \{\mathbf{p} | \sum_{k=1}^{K} p_k = 1 \quad and \quad 0 \leq p_1, p_2, \cdots, p_K \leq 1\}. \tag{9}$$

Therefore, the concentration parameter of the Dirichlet distribution is thus associated with the evidence of each class. Up to now, the TEB-SUQ is ready to be used on both views of aerial and ground images to measure the decision-making risk during their fusion.

2.4. Evidential Fusion

As mentioned in Section 1, directly fusing the softmax outputs of the views will ignore their decision-making risks, resulting in unreliable results. In order to reduce the weight of the view with higher risk, a novel decision-level fusion strategy, namely Evidential Fusion is proposed based on Equations (2) and (3). Evidential fusion uses the sample uncertainty of each view as its decision-making risk to allocate the weight in the final prediction.

To be more precise, the opinions $\mathcal{O}^1 = \{c_1^1, c_2^1, \ldots, c_K^1, u^1\}$ and $\mathcal{O}^2 = \{c_1^2, c_2^2, \ldots, c_K^2, u^2\}$ of the two views are obtained after the UQM, where the superscripts denote the number of different views. The final decision opinions $\mathcal{O} = \{c_1, c_2, \ldots, c_K, u\}$ is then calculated as follows:

$$c_k = \frac{1}{\lambda}[c_k^1 c_k^2 + (1 - u^1)c_k^1 + (1 - u^2)c_k^2], \tag{10}$$

$$u = \frac{1}{\lambda} u^1 u^2, \tag{11}$$

$$\lambda = u^1 u^2 + (1 - u^1)^2 + (1 - u^2)^2 + \sum_{k=1}^{K} c_k^1 c_k^2, \tag{12}$$

where λ is scale factor to perform the normalization and to ensures that Equation (1) still holds after fusion.

After obtaining the final decision opinion $\mathcal{O}$ from the fusion of two views, the fused evidence e_k of the k-th class can be calculated according to the following equation:

$$e_k = \frac{K \cdot c_k}{u}, \tag{13}$$

and the concentration parameters α_k of the Dirichlet distribution of the k-th class can be updated by Equation (4) to calculate the loss during training step and the category with the largest e_k is the final predicted label during test step.

An example is given to demonstrate how the the proposed evidential fusion works in Figure 7. A multiplication fusion on the softmax outputs of aerial-ground dual-view images is shown in Figure 7a and the proposed evidential fusion on the UQM outputting

opinions is shown in Figure 7b, which obtains a more credible classification result based on the sample uncertainty.

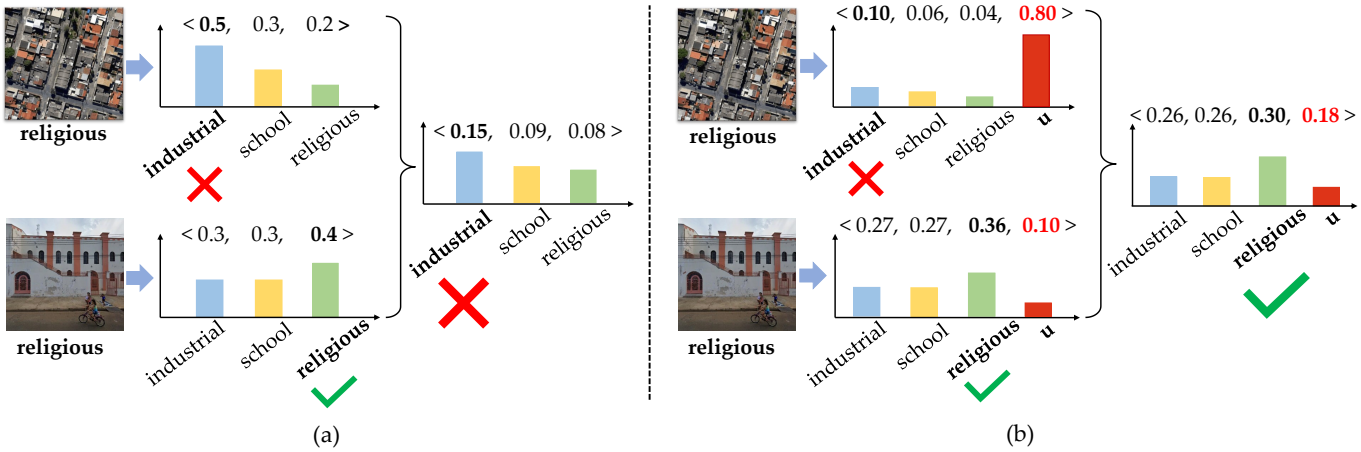


Figure 7. Different predictions are made when using different fusion strategies on aerial-ground dual-view decisions: (**a**) using a multiplication fusion on the softmax outputs of both views; and (**b**) using the proposed evidential fusion on the UQM outputting opinions. All predictions are the class label with the highest score, whether in a single view or after fusion.

2.5. Reciprocal Loss

To train the proposed EFN in Figure 3, the following loss function is designed and then adopted on each view:

$$L(\boldsymbol{\alpha}) = \sum_{i=1}^{N} (L_{pc}(\boldsymbol{\alpha}_i) + L_{nc}(\boldsymbol{\alpha}_i)), \tag{14}$$

$$L_{pc}(\boldsymbol{\alpha}_i) = \sum_{k=1}^{K} y_{ik} [\psi(\alpha_{i0}) - \psi(\alpha_{ik})], \tag{15}$$

$$L_{nc}(\boldsymbol{\alpha}_i) = \sum_{k=1}^{K} (1 - y_{ik}) \frac{1}{\psi(\alpha_{i0}) - \psi(\alpha_{ik})}, \tag{16}$$

where $L_{pc}(\boldsymbol{\alpha}_i)$ and $L_{nc}(\boldsymbol{\alpha}_i)$ are positive-class loss and negative-class loss of the i-th sample respectively, $\psi(\cdot)$ is the digamma function which is monotonically increasing in $(0, +\infty)$, α_{i0} is consistent with Equation (7), K is the number of classes and N is the number of samples.

As shown in Equation (14), the proposed loss function clearly separates the penalty terms of the positive and negative classes, making it easy to interpret. Furthermore, Equations (15) and (16) show that the positive-class loss and the negative-class loss are reciprocal. More specifically, since $\alpha_{ik} = e_{ik} + 1$ and $e_{ik} > 0$, so $\psi(\alpha_{ik})$ is monotonically increasing with α_{ik}, while $L_{pc}(\boldsymbol{\alpha}_i)$ is monotonically decreasing with α_{ik}, which ensures that the positive class of each sample generates more evidence. On the contrary, $L_{nc}(\boldsymbol{\alpha}_i)$ is monotonically increasing with α_{ik} to ensure that negative classes of each sample generate less evidence. Next, taking $L_{pc}(\boldsymbol{\alpha}_i)$ as an example, how the proposed Reciprocal Loss is related to the Dirichlet distribution will be derived.

For multi-class classification tasks, the cross-entropy loss function (CE Loss) is most commonly used. For a particular sample, its CE Loss can be calculated by

$$L_{ce} = -\sum_{k=1}^{K} y_k \log p_k, \tag{17}$$

where p_k is the predicted probability of the k-th class, y_k is its class label and $y_k = 1$ for positive classes and $y_k = 0$ for negative classes. In the proposed EFN, the evidence vector of

a sample $\boldsymbol{e}$ obtained from the FEM can be mapped to the concentration parameters $\boldsymbol{\alpha}$ of the Dirichlet distribution $D(\mathbf{p}|\boldsymbol{\alpha})$ by using Equation (4). Thus the Bayes risk of cross-entropy loss $L_{cer}(\boldsymbol{\alpha})$ can be calculated as

$$\begin{aligned} L_{cer}(\boldsymbol{\alpha}) &= \int L_{ce} D(\mathbf{p}|\boldsymbol{\alpha}) d\mathbf{p} \\ &= \int (-\sum_{k=1}^{K} y_k \log p_k) D(\mathbf{p}|\boldsymbol{\alpha}) d\mathbf{p}. \end{aligned} \tag{18}$$

Since p_k is a $D(\mathbf{p}|\boldsymbol{\alpha})$ random variable, the functions $\log p_k$ are the sufficient statistics of the Dirichlet distribution. Thus, the exponential family differential identities can be used to get an analytic expression for the expectation of $\log p_k$ [38]:

$$\begin{aligned} \mathbb{E}_{D(\mathbf{p}|\boldsymbol{\alpha})}(\log p_k) &= \int (\log p_k) D(\mathbf{p}|\boldsymbol{\alpha}) d\mathbf{p} \\ &= \psi(\alpha_k) - \psi(\alpha_0). \end{aligned} \tag{19}$$

In this regard, Equation (18) can be expanded further as

$$\begin{aligned} L_{cer}(\boldsymbol{\alpha}) &= \int (-\sum_{k=1}^{K} y_k \log p_k) D(\mathbf{p}|\boldsymbol{\alpha}) d\mathbf{p} \\ &= -\sum_{k=1}^{K} y_k \int (\log p_k) D(\mathbf{p}|\boldsymbol{\alpha}) d\mathbf{p} \\ &= -\sum_{k=1}^{K} y_k [\psi(\alpha_k) - \psi(\alpha_0)] \\ &= \sum_{k=1}^{K} y_k [\psi(\alpha_0) - \psi(\alpha_k)]. \end{aligned} \tag{20}$$

Given that the Equation (20) can only penalize the positive class, L_{cer} can be written as L_{pc}. We finally have $L_{pc}(\boldsymbol{\alpha}_i)$ in Equation (15) for the case of all samples.

In order to ensure that both views can provide reasonable opinions for scene classification and thus improve the overall opinion after fusion, the final multi-view global Reciprocal Loss is used:

$$L_{global} = L^1 + L^2 + L_{fused}, \tag{21}$$

where L^1, L^2 are the Reciprocal Loss (Equation (14)) for the first view and second view, respectively, and L_{fused} is obtained using Equation (14) on the fused parameters by Equations (4), (10), (11) and (13).

3. Results

3.1. Experimental Setup

For both datasets of AiRound and CV-BrCT, we randomly selected 80% of the samples from each class as the training/validation set and the remaining 20% as the test set to form one data split. All the test results except for Section 3.2.1 are the mean results of 10 splits. The training/validation set was randomly divided into the training set and validation set according to the ratio of 9:1.

All models were simulated by PyTorch on a computer with a GTX 1080Ti graphics card. The details during training are as follows. Batch size: 128; the number of epochs: 200 for feature extraction and credible fusion, respectively, and all the best models on validation data are saved, learning rate schedules: Cosine decay with the initial value of 0.01, optimizer: SGD for feature extraction and Adam for credible fusion, weight decay: 0.1, and momentum of SGD: 0.9. All models of feature extraction are trained by fine-tuning the officially published pre-trained ones.

To quantitatively evaluate the performance of each model, we used the classification accuracy (Acc) and F1-score (F1) as metrics. Acc is calculated by the ratio of the number of

correctly predicted test samples to the total number of test samples. F1 is calculated by the ratio of twice the product of precision and recall to the sum of them.

3.2. Ablation Study

3.2.1. Validation for the Effectiveness of Uncertainty Estimation

In order to validate the effectiveness of the TEB-SUQ in measuring sample uncertainty, a subjective evaluation of sample credibility was performed. All test samples from the two datasets (2347 pairs of AiRound and 4830 pairs of CV-BrCT) were evaluated by nine urban planning experts from the Qingdao Research Institute of Urban and Rural Construction based on whether their image content matched the label. A vote of nine experts determines the final conclusion (credible or uncertain) of each test sample. For the TEB-SUQ, an appropriate threshold u_{th} was set to distinguish between credible and uncertain samples. The number of credible samples calculated using $u \leq 0.4$ (n_k for the k-th class) was compared to the one generated by experts' votes (m_k for the k-th class) in each class of each view of the two datasets in Figure 8. Table 1 shows the average error $\bar{\sigma}$ of the TEB-SUQ relative to subjective evaluation of each view using

$$\bar{\sigma} = \frac{1}{K} \sum_{k=1}^{K} \frac{|n_k - m_k|}{m_k}, \tag{22}$$

where K is the total number of classes.

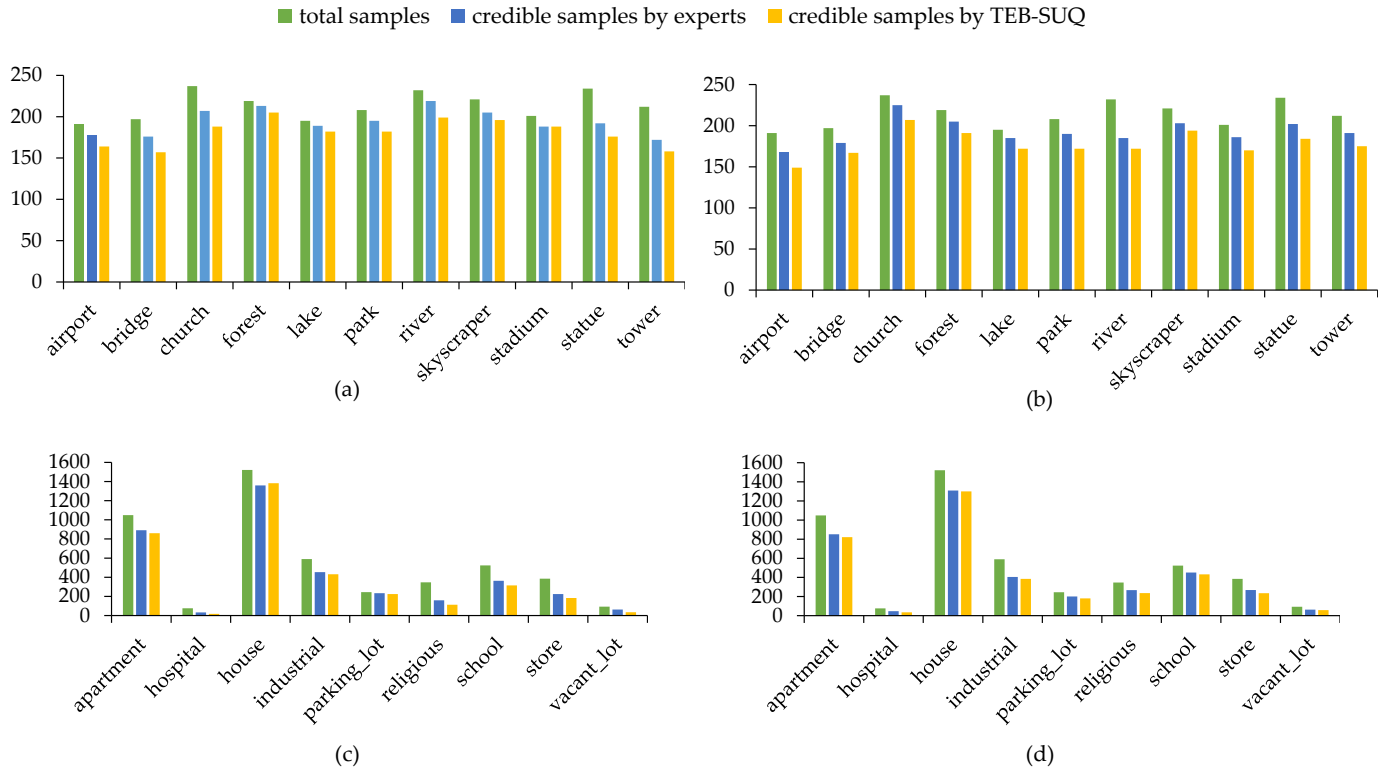


Figure 8. The number of credible samples in each class generated by experts' votes and the TEB-SUQ using $u \leq 0.4$ in the (**a**) aerial and (**b**) ground view of AiRound; (**c**) aerial and (**d**) ground view of CV-BrCT.

As can be observed from Table 1, except for the aerial view of CV-BrCT, the relative error between the predicted number of credible samples and the subjective evaluation in other views is less than 0.1, proving that the proposed sample uncertainty estimation approach is effective.

Table 1. The average error of the TEB-SUQ relative to subjective evaluation.

Views *	A-*AiRound*	G-*AiRound*	A-*CV-BrCT*	G-*CV-BrCT*
$\bar{\sigma}$	0.065	0.078	0.183	0.090

* **A-** for the aerial view and **G-** for the ground view.

3.2.2. Validation for the Effectiveness of the Reciprocal Loss

The widely used loss function for evidential deep learning (EDL Loss for short) consists of a term of general classification loss and a term of Kullback–Leibler (KL) divergence of two Dirichlet distributions with an extra annealing coefficient [33]. It is formally too complex and difficult to train, in contrast to the proposed Reciprocal Loss (Equation (14) to Equation (16)). Scene classification performances of EDL Loss and the proposed Reciprocal Loss are shown in Table 2, where in EDL-CE Loss, the first term of general classification loss is CE-like (see Equation (4) in the Ref. [33]), whereas in EDL-MSE Loss, it is MSE-like (see Equation (5) in the Ref. [33]). All results were obtained using VGG-11 [39] as the backbone with the same training setting and fusion strategy. The annealing coefficient in EDL was calculated as $\min(1.0, t/10) \in [0, 1]$, where t is the index of the current training epoch. It is clear that the use of Reciprocal Loss resulted in improved performances.

Table 2. Performances (%) $\pm$ STD of different loss functions.

Loss Functions	AiRound		CV-BrCT	
	Acc	F1	Acc	F1
EDL-CE Loss [33]	90.23 ± 0.32	90.64 ± 0.29	86.98 ± 0.17	81.56 ± 0.34
EDL-MSE Loss [33]	90.44 ± 0.03	91.02 ± 0.02	86.24 ± 0.02	80.89 ± 0.01
Reciprocal Loss	$\mathbf{92.16 \pm 0.31}$	$\mathbf{92.49 \pm 0.25}$	$\mathbf{88.21 \pm 0.26}$	$\mathbf{83.57 \pm 0.29}$

3.2.3. Validation for the Effectiveness of the Evidential Fusion Strategy

As described in Section 2.2, the proposed approach is backbone-independent, allowing it to be flexibly matched with different deep feature extraction networks. In this experiment, three of the most common backbones are used to compare the performance on scene classification task before and after using different fusion strategies. In the columns of single views in Tables 3 and 4, the "-s" represents the traditional deep learning (that is, softmax deep learning) approach where the softmax layer and CE Loss are used during the training phase, and the class corresponding to the maximum probability calculated by the softmax operator is used as the prediction result during the test phase. Additionally, "-e" denotes the evidential deep learning approach proposed in Sections 2.3 and 2.5 where the softmax layer is replaced by the softplus layer and the proposed Reciprocal Loss are used for training. The class corresponding to the maximum evidence value is used as the prediction result during the test. In the columns of fusion strategies, two common decision level fusion strategies (sum and product) [31] are used as baseline approaches to compare the proposed evidential fusion. The best results are highlighted in bold.

The following observations can be made from Tables 3 and 4. First, all of the fusion results are superior to any single view result. This demonstrates that information from multi-views can significantly improve classification accuracy. Second, the performance of the two single views obtained through evidential deep learning is marginally worse than that of the corresponding single view obtained through softmax deep learning. This is due to the fact that when the uncertainty of some samples is high, evidential deep learning focuses more on estimating the uncertainty values as accurately as possible, which may result in a loss of classification accuracy for these samples. However, in terms of sample quality, the category labels of these high-uncertainty samples lack actual semantics. It also does not matter whether their predictions are correct or incorrect. It makes more sense to quantify their uncertainty. Last but not least, the evidential fusion strategy proposed in

this paper outperforms the other two baseline fusion approaches, which demonstrates the efficacy of evidential fusion in the task of multi-view remote sensing scene classification.

Table 3. Classification accuracy (%) ± STD of single views and after decision-level fusion on AiRound.

Views		AlexNet [40]	VGG-11 [39]	ResNet-18 [41]
Single Views (softmax)	Aerial-s	76.96 ± 0.52	82.75 ± 0.61	80.93 ± 0.49
	Ground-s	71.35 ± 0.24	77.10 ± 0.28	76.68 ± 0.19
Single Views (evidential)	Aerial-e	76.04 ± 0.46	82.64 ± 0.49	80.83 ± 0.52
	Ground-e	70.96 ± 0.25	76.99 ± 0.17	76.36 ± 0.22
Decision-Level Fusion Strategies	Sum [31]	84.02 ± 0.47	87.75 ± 0.38	88.02 ± 0.25
	Product [31]	86.74 ± 0.25	90.41 ± 0.27	89.56 ± 0.24
	Proposed	**88.12 ± 0.23**	**92.16 ± 0.31**	**91.02 ± 0.35**

Table 4. Classification accuracy (%) ± STD of single views and after decision-level fusion on CV-BrCT.

Views		AlexNet [40]	VGG-11 [39]	ResNet-18 [41]
Single Views (softmax)	Aerial-s	84.63 ± 0.24	87.11 ± 0.42	86.74 ± 0.38
	Ground-s	68.01 ± 0.12	71.43 ± 0.22	70.96 ± 0.25
Single Views (evidential)	Aerial-e	84.37 ± 0.10	87.06 ± 0.38	86.18 ± 0.29
	Ground-e	66.36 ± 0.25	70.15 ± 0.29	70.86 ± 0.24
Decision-Level Fusion Strategies	Sum [31]	85.26 ± 0.45	86.70 ± 0.58	85.59 ± 0.62
	Product [31]	86.52 ± 0.25	87.21 ± 0.22	86.83 ± 0.18
	Proposed	**88.02 ± 0.28**	**88.21 ± 0.26**	**87.95 ± 0.19**

3.3. Comparison Experiment with Different Fusion Approaches at Data-Level, Feature-Level and Decision-Level

In Section 3.2, the effectiveness of three innovative contributions in this paper (namely TEB-SUQ, Evidential Fusion, and Reciprocal Loss) was validated, respectively. In this section, more multi-view fusion methods are compared with the proposed approach to assess its overall performance on the task of aerial-ground dual-view remote sensing scene classification. As mentioned in Section 1, existing multi-view fusion methods can be roughly classified as data-level, feature-level, and decision-level. In this experiment, one data-level fusion method (six-channel [42]), two feature-level fusion methods (feature concatenation [31] and CILM [43]), and five decision-level fusion methods (maximum [31], minimum [31], sum [31], product [31] and SFWS [44]) were chosen to compare with the proposed evidential fusion. These methods are briefly described below.

- Six-channel [42]: This method concatenates the RGB channels of the paired aerial view and ground view images into a six-channel image as the input of a CNN.
- Feature concatenation [31]: A Siamese-like CNN is used to concatenate the intermediate feature tensors before the first convolution layer that doubles its amount of kernels.
- CILM [43]: The loss function of contrast learning is combined with CE Loss in this method, allowing the features extracted by the two subnetworks to be fused without sharing any weight.
- Maximum [31]: Each view employs an independent DNN to obtain its prediction result, which consists of a class label and its probability. The final prediction is the class label corresponding to the maximum of the class probabilities predicted by each view.
- Minimum [31]: Each view employs an independent DNN to obtain its prediction result, which consists of a class label and its probability. The final prediction is the

class label corresponding to the minimum of the class probabilities predicted by each view.

- Sum [31]: Each view employs an independent DNN to generate a vector containing probabilities for each class. The fused vector is the sum of single view vectors. The final prediction result is the class label corresponding to the largest element in the fused vector.
- Product [31]: Each view employs an independent DNN to generate a vector containing probabilities for each class. An elementwise multiplication is performed between single view vectors to obtain the fused vector. The final prediction result is the class label corresponding to the largest element in the fused vector.
- SFWS (Softmax Feature Weighted Strategy) [44]: Each view employs an independent DNN to obtain a vector containing probabilities for each class. Then, the matrix nuclear norm of the vector is computed as the weight of the fusion. The final prediction result is the class label corresponding to the largest element in the fused vector.

Tables 5 and 6 show the performance of the fusion methods discussed above on AiRound and CV-BrCT, respectively, when different backbones are used. The following observations can be made. First, methods of the data-level fusion class performed the worst, while methods of the decision-level fusion class won across the board. This observation is consistent with the discussion of data-level fusion and feature-level fusion in Section 2.2. The performance of the fusion improves as the features involved in it become more abstract. This also explains why CILM outperforms feature concatenation among the two feature-level fusion methods: CILM lacks a shared weight structure, making it more similar to decision-level fusion in form. Second, among the decision-level fusion methods, sum and maximum perform nearly identically, and both perform slightly worse than minimum. This result may seem counter-intuitive at first. In fact, it confirms the overconfidence issue caused by the softmax mentioned in Section 2.3 (see Figures 4 and 5): an overestimated prediction is more likely to be incorrect. Last but not least, product and the proposed evidential fusion stand out among all the fusion methods, and the latter outperforms the former. In fact, Equation (10) can be seen as an enhancement of the product method. The inclusion of sample uncertainty breaks down the equality of views in the fusion: views with lower u values are given more weight adaptively.

Table 5. Classification accuracy (%) ± STD using different fusion methods on AiRound.

Methods	**AlexNet** [40]	**VGG-11** [39]	**Inception** [45]	**ResNet-18** [41]	**DenseNet** [46]
Six-Ch. [42]	70.19 ± 0.23	72.34 ± 0.21	71.76 ± 0.24	71.29 ± 0.26	71.57 ± 0.25
Concat. [31]	82.52 ± 0.32	84.69 ± 0.41	83.91 ± 0.45	83.56 ± 0.39	83.72 ± 0.42
CILM [43]	83.49 ± 0.17	85.72 ± 0.19	85.05 ± 0.15	84.72 ± 0.21	84.91 ± 0.20
Max. [31]	84.86 ± 0.36	88.17 ± 0.34	88.39 ± 0.33	88.21 ± 0.38	89.96 ± 0.35
Min. [31]	85.52 ± 0.23	89.56 ± 0.25	89.12 ± 0.27	88.42 ± 0.22	90.41 ± 0.27
Sum [31]	84.02 ± 0.47	87.75 ± 0.38	88.05 ± 0.30	88.02 ± 0.25	89.88 ± 0.31
Product [31]	86.74 ± 0.25	90.41 ± 0.24	90.02 ± 0.14	89.56 ± 0.24	91.16 ± 0.17
SFWS [44]	85.94 ± 0.17	89.61 ± 0.34	89.18 ± 0.12	88.95 ± 0.26	90.05 ± 0.37
Proposed	**88.12 ± 0.23**	**92.16 ± 0.31**	**91.41 ± 0.18**	**91.02 ± 0.35**	**92.16 ± 0.19**

Table 7 shows the training time of the fusion methods discussed above using VGG-11 as the backbone. The proposed evidential fusion approach takes 2.72% more time than the most time-efficient model (Six-Ch.) and only 0.51% more time than other decision-level fusion approaches (Dec.-Lev.). During inference, the added time of the proposed method is negligible and therefore does not affect the actual use.

Table 6. Classification accuracy (%) ± STD using different fusion methods on CV-BrCT.

Methods	AlexNet [40]	VGG-11 [39]	Inception [45]	ResNet-18 [41]	DenseNet [46]
Six-Ch. [42]	71.92 ± 0.26	73.46 ± 0.24	75.26 ± 0.25	73.25 ± 0.28	74.19 ± 0.27
Concat. [31]	81.86 ± 0.42	83.25 ± 0.39	84.65 ± 0.38	83.28 ± 0.41	84.24 ± 0.40
CILM [43]	83.10 ± 0.20	84.32 ± 0.19	85.22 ± 0.16	84.31 ± 0.17	85.19 ± 0.22
Max. [31]	85.52 ± 0.24	86.74 ± 0.39	86.70 ± 0.41	85.84 ± 0.43	86.95 ± 0.38
Min. [31]	86.02 ± 0.21	86.95 ± 0.27	86.95 ± 0.28	86.24 ± 0.35	87.02 ± 0.25
Sum [31]	85.26 ± 0.45	86.70 ± 0.58	86.24 ± 0.35	85.59 ± 0.62	86.85 ± 0.39
Product [31]	86.52 ± 0.25	87.21 ± 0.22	87.02 ± 0.21	86.83 ± 0.18	87.54 ± 0.17
SFWS [44]	86.21 ± 0.14	86.95 ± 0.22	86.73 ± 0.21	86.52 ± 0.16	87.21 ± 0.25
Proposed	**88.02 ± 0.28**	**88.21 ± 0.26**	**88.21 ± 0.23**	**87.95 ± 0.19**	**88.34±0.20**

Table 7. Training time (ms per sample) ± STD using different fusion methods.

Datasets	Training Time				
	Six-Ch. [42]	Concat. [31]	CILM [43]	Dec.-Lev. [31,44]	Proposed
AiRound	**9.16 ± 0.01**	9.21 ± 0.02	9.38 ± 0.02	9.37 ± 0.01	9.42 ± 0.01
Cv-BrCT	**8.01 ± 0.01**	8.06 ± 0.02	8.20 ± 0.02	8.19 ± 0.01	8.23 ± 0.01

Finally, on all backbones of both datasets, the proposed evidential fusion approach outperforms the best decision-level fusion method by 1.26%, the best feature-level fusion method by 4.96% and the data-level fusion method by 17.04%. Examples of predictions are shown in Figure 9.

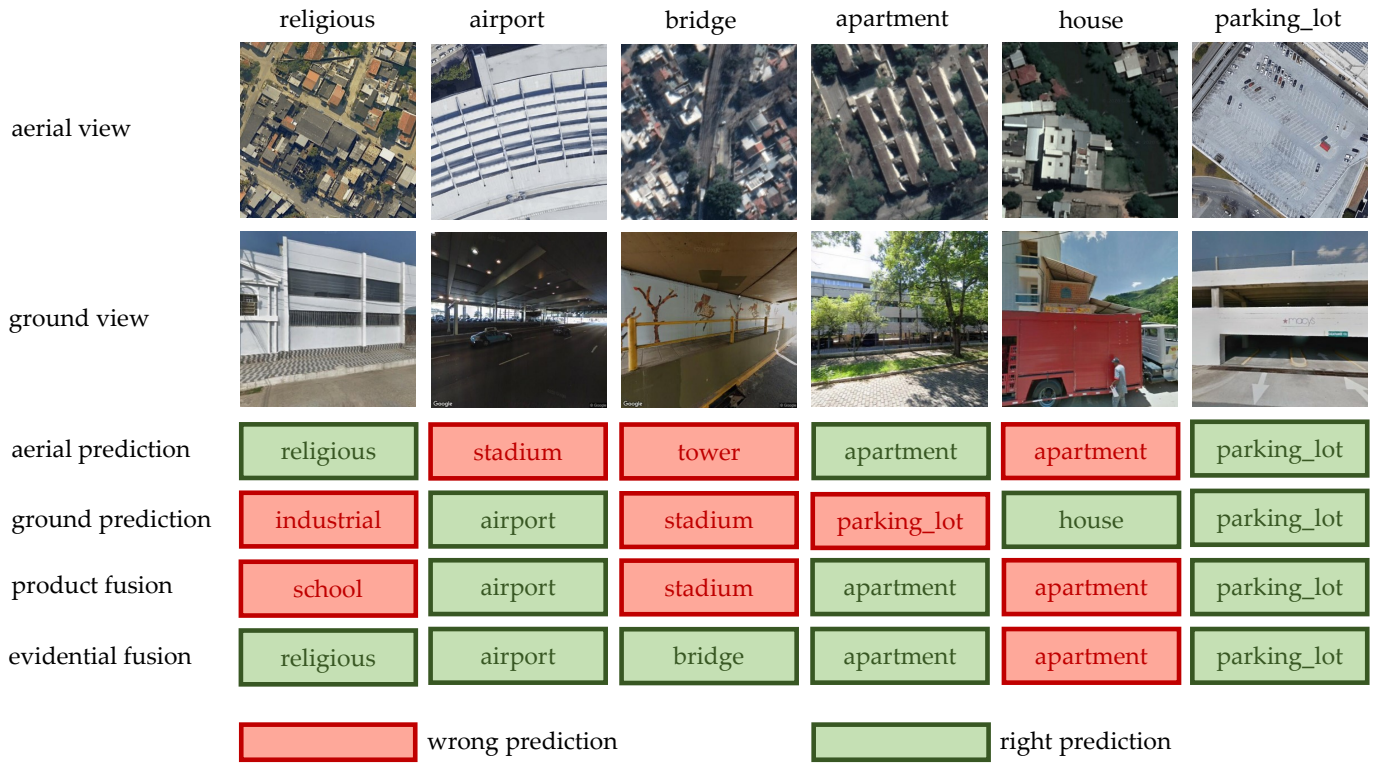


Figure 9. Examples of predictions by single views, the product fusion and the proposed evidential fusion.

4. Discussion

4.1. Discussion on Uncertainty Estimation

Figure 10 shows one case study of the class "stadium" in AiRound. By the TEB-SUQ, the uncertainties and evidence of two images in different views were obtained. The evidence values bigger than 1.00 of the aerial image are 31.59 (stadium) and 6.13 (statue), showing a good concentration. Accordingly, its uncertainty is only about 0.23. However, the evidence values bigger than 1.00 of the ground image are 3.24 (bridge), 3.01 (river), 1.68 (church) and 1.12 (stadium), whose distribution is more dispersed. Accordingly, its uncertainty is about 0.54, suggesting that the model is less than half as confident about its predictions. As can be seen from the images, the above conclusions are intuitive. More cases of the uncertainty of samples in the CV-BrCT datasets are shown in Figure 11, whose classes are random selected.

aerial view

$u = 0.23$

evidences

airport	bridge	church	forest	lake	park
9.78×10^{-12}	2.77×10^{-18}	1.16×10^{-6}	9.93×10^{-15}	2.84×10^{-7}	5.96×10^{-12}
river	skyscraper	stadium	statue	tower	
1.48×10^{-12}	2.66×10^{-21}	**31.59**	6.13	4.59×10^{-6}	

ground view

$u = 0.54$

evidences

airport	bridge	church	forest	lake	park
1.68×10^{-1}	**3.24**	1.68	4.09×10^{-3}	6.23×10^{-3}	7.35×10^{-5}
river	skyscraper	stadium	statue	tower	
3.01	9.74×10^{-4}	1.12	1.13×10^{-2}	2.83×10^{-2}	

Figure 10. The uncertainties and evidences of sample No. 9360855 of class "stadium" in AiRound computed by the TEB-SUQ. The evidence values in bold correspond to the predicted results.

Ybui2Cx84FwOb ioPZDaL0Q.png

RyuRbwI8ZIN1T duSMDO79A.png

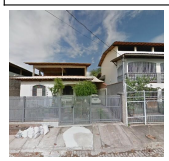
RKb6tT11_IFwo Xl19OPB6w.png

ab6kyiLxrh3Ubp EUDPAOEQ.png

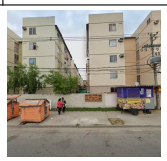
F2FxERCkELT-1x4u3ip0jw.png

hSTEGNpaEzzLf Aw2DL07hA.png

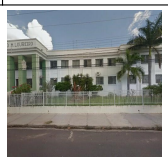
parking_lot	industrial	religious	hospital	hospital	store
0.1223	0.1497	0.4501	0.4738	0.7572	0.9881



_Oig9wjtfTGxWI y6LmEU-Q.jpg



1O_2zBm0_kid9 XOvePYhlA.jpg



CmxaN_MGo7yUL mC6B_CHWg.jpg

fuGqUbeqBWrHb NfgMVOuOg.jpg

WyNqGldabWvT 6OeQ6RjCVg.jpg

6slSgE_5eYsYx5r G6IqhrA.jpg

houes	apartment	school	religious	industrial	hospital
0.0790	0.0854	0.4133	0.4403	0.7013	0.9402

Figure 11. The uncertainty of samples in CV-BrCT computed by the TEB-SUQ.

More statistically, Figure 12 shows the uncertainty distributions of each view samples in the test sets of the AiRound and CV-BrCT. The figures clearly show that the samples of AiRound are distributed more densely in parts with lower uncertainty and have higher peak values. This advantage is even more visible in the ground view (Figure 12b). In other words, after the calculation by the TEB-SUQ, the quality of AiRound is higher than that of CV-BrCT, especially in the ground view. This conclusion is supported by the following facts. Firstly, CV-BrCT has more than twice the number of samples as AiRound. Secondly, unlike the average distribution of AiRound, the class distribution of CV-BrCT is a typical long-tail distribution (Figure 2b). It is well-understood that increasing the number of

samples and unbalanced category distribution will result in a decrease in data quality. Last but not least, the methodologies used to collect the ground view images for the two datasets differ [31]. The ground view images of AiRound are largely derived from the Google Places' database, which is a well-known high-quality dataset. The ground view images of CV-BrCT, on the other hand, were all obtained by the Google Images search engine, meaning that image quality cannot be guaranteed.

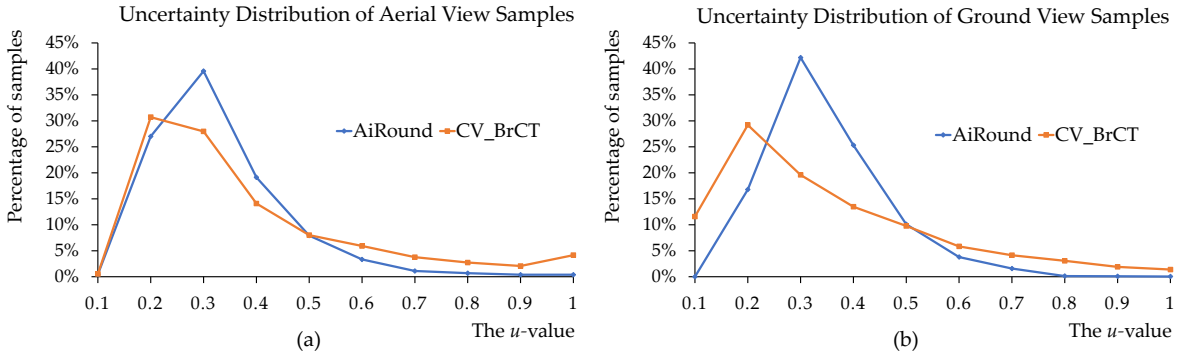


Figure 12. The uncertainty distributions of the (**a**) aerial and (**b**) ground view samples in the test sets of AiRound and CV BrCT.

4.2. Discussion on Loss Functions

Figure 13 shows the training and validation loss using EDL Loss and Reciprocal Loss (Equation (21)) with the same setting. Significant overfitting occurred during training using EDL Loss. It has been significantly improved since switching to Reciprocal Loss.

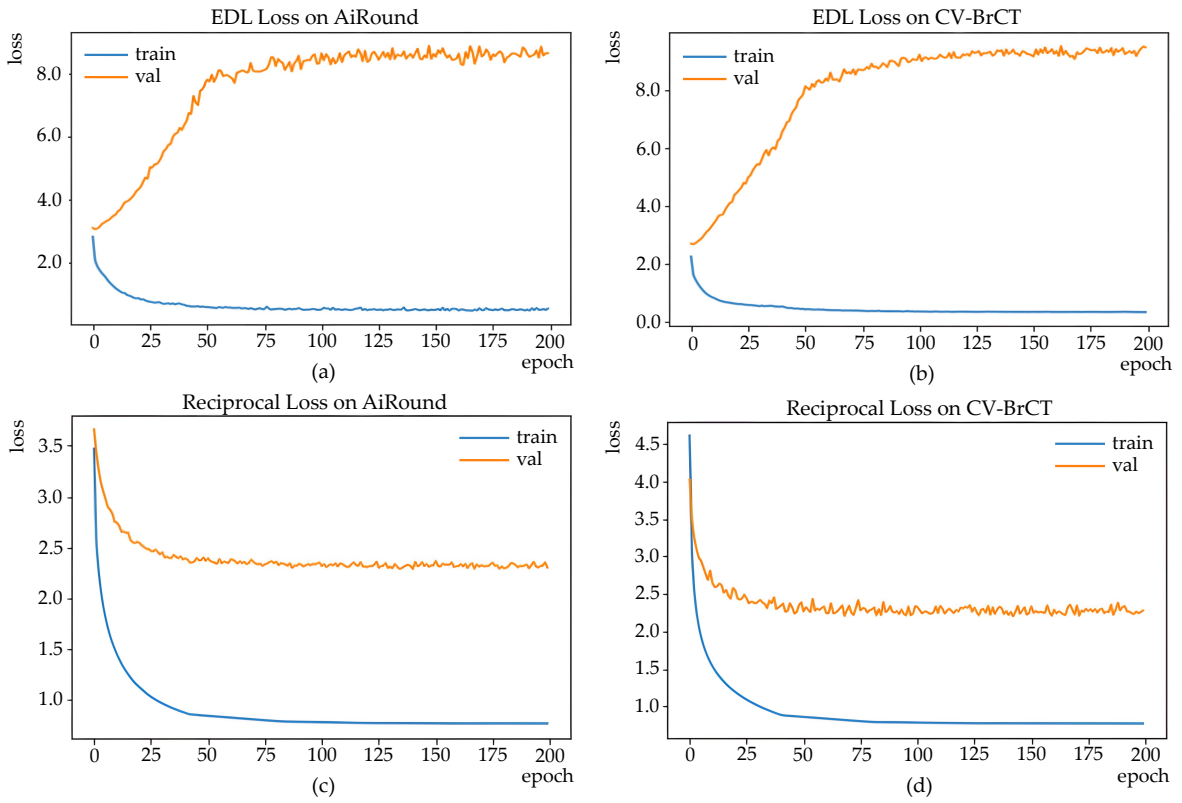


Figure 13. The training and validation loss using (**a**) EDL Loss on AiRound and (**b**) on CV-BrCT, and using (**c**) Reciprocal Loss on AiRound and (**d**) on CV-BrCT.

5. Conclusions

Deep learning models are easily influenced by data quality, especially when dealing with massive amounts of data [47,48]. Li's team proposed the concept of "trustworthy AI" [49], with a focus on data quality. As multi-view data is increasingly used in various remote sensing tasks, the issue of data quality in the original single view becomes more apparent. In this paper, the theory of evidence was introduced to quantify the credibility of samples. On this basis, a decision-level multi-view fusion strategy was proposed to assign higher weights to views with lower decision-making risk. The proposed evidential fusion network achieves the best performance on the two classical datasets in the task of remote sensing scene classification on aerial-ground dual-view images, outperforms the best decision-level fusion method by 1.26%, the best feature-level fusion method by 4.96% and the data-level fusion method by 17.04%.

Focusing on data quality in multi-view tasks is a new area of research, and much work remains to be done. First, there are few publicly available multi-view datasets for remote sensing tasks. Large-scale, instance-level aligned remote sensing multi-view data sets are urgently needed for public release for related research. Furthermore, effective objective evaluation of sample uncertainty estimation is lacking. Datasets with sample quality annotation have yet to appear in the field of remote sensing. Finally, more explicit representation methods of sample uncertainty need to be further explored.

Author Contributions: All the authors made significant contributions to this work. Project administration, L.Z.; innovations and original draft writing, K.Z.; coding, Q.G.; review and editing, S.H. and J.S. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the National Natural Science Foundation of China under Grant 62171247.

Data Availability Statement: The code and datasets of this article are available at the following address: https://github.com/gaopiaoliang/Evidential (accessed on 3 January 2023).

Acknowledgments: The authors would like to thank Ruxuan Bi, Zhiwei He and Mengshuo Fan, the experts in urban planning from the BIM Research Center, Qingdao Research Institute of Urban and Rural Construction for their professional guidance on the subjective evaluation of sample credibility. Thanks to those who participated in the subjective evaluation: Chunting Zhao, Xiayue Wang, Qi Liu, Mingdong Ding, Jinhong Guo and Zhengnan Fang.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Cheng, G.; Han, J.; Lu, X. Remote Sensing Image Scene Classification: Benchmark and State of the Art. *Proc. IEEE* **2017**, *105*, 1865–1883. [CrossRef]
2. Cheng, G.; Xie, X.; Han, J.; Guo, L.; Xia, G.S. Remote Sensing Image Scene Classification Meets Deep Learning: Challenges, Methods, Benchmarks, and Opportunities. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 3735–3756. [CrossRef]
3. Zhou, Y.; Chen, P.; Liu, N.; Yin, Q.; Zhang, F. Graph-Embedding Balanced Transfer Subspace Learning for Hyperspectral Cross-Scene Classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 2944–2955. [CrossRef]
4. Chen, L.; Cui, X.; Li, Z.; Yuan, Z.; Xing, J.; Xing, X.; Jia, Z. A New Deep Learning Algorithm for SAR Scene Classification Based on Spatial Statistical Modeling and Features Re-Calibration. *Sensors* **2019**, *19*, 2479. [CrossRef]
5. Wang, Q.; Liu, S.; Chanussot, J.; Li, X. Scene Classification With Recurrent Attention of VHR Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 1155–1167. [CrossRef]
6. Li, B.; Guo, Y.; Yang, J.; Wang, L.; Wang, Y.; An, W. Gated Recurrent Multiattention Network for VHR Remote Sensing Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–13. [CrossRef]
7. Qiao, Z.; Yuan, X. Urban land-use analysis using proximate sensing imagery: A survey. *Int. J. Geogr. Inf. Sci.* **2021**, *35*, 2129–2148. [CrossRef]
8. Baltrušaitis, T.; Ahuja, C.; Morency, L.P. Multimodal Machine Learning: A Survey and Taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *41*, 423–443. [CrossRef]
9. Nunez, J.; Otazu, X.; Fors, O.; Prades, A.; Pala, V.; Arbiol, R. Multiresolution-based image fusion with additive wavelet decomposition. *IEEE Trans. Geosci. Remote Sens.* **1999**, *37*, 1204–1211. [CrossRef]
10. Amolins, K.; Zhang, Y.; Dare, P. Wavelet based image fusion techniques—An introduction, review and comparison. *ISPRS J. Photogramm. Remote Sens.* **2007**, *62*, 249–263. [CrossRef]

11. Lin, T.Y.; Belongie, S.; Hays, J. Cross-View Image Geolocalization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Portland, OR, USA, 23–28 June 2013; pp. 891–898.
12. Lin, T.Y.; Cui, Y.; Belongie, S.; Hays, J. Learning Deep Representations for Ground-to-Aerial Geolocalization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 5007–5015.
13. Zhang, X.; Du, S.; Wang, Q. Hierarchical semantic cognition for urban functional zones with VHR satellite images and POI data. *ISPRS J. Photogramm. Remote Sens.* **2017**, *132*, 170–184. [CrossRef]
14. Workman, S.; Zhai, M.; Crandall, D.J.; Jacobs, N. A Unified Model for Near and Remote Sensing. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2688–2697.
15. Zhang, W.; Li, W.; Zhang, C.; Hanink, D.M.; Li, X.; Wang, W. Parcel-based urban land use classification in megacity using airborne LiDAR, high resolution orthoimagery, and Google Street View. *Comput. Environ. Urban Syst.* **2017**, *64*, 215–228. [CrossRef]
16. Deng, Z.; Sun, H.; Zhou, S. Semi-Supervised Ground-to-Aerial Adaptation with Heterogeneous Features Learning for Scene Classification. *ISPRS Int. J. Geo-Inf.* **2018**, *7*, 182. [CrossRef]
17. Cao, R.; Zhu, J.; Tu, W.; Li, Q.; Cao, J.; Liu, B.; Zhang, Q.; Qiu, G. Integrating Aerial and Street View Images for Urban Land Use Classification. *Remote Sens.* **2018**, *10*, 1553. [CrossRef]
18. Hoffmann, E.J.; Wang, Y.; Werner, M.; Kang, J.; Zhu, X.X. Model Fusion for Building Type Classification from Aerial and Street View Images. *Remote Sens.* **2019**, *11*, 1259. [CrossRef]
19. Srivastava, S.; Vargas-Muñoz, J.E.; Tuia, D. Understanding urban landuse from the above and ground perspectives: A deep learning, multimodal solution. *Remote Sens. Environ.* **2019**, *228*, 129–143. [CrossRef]
20. Hong, D.; Yokoya, N.; Chanussot, J.; Zhu, X.X. CoSpace: Common Subspace Learning From Hyperspectral-Multispectral Correspondences. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 4349–4359. [CrossRef]
21. Wang, X.; Feng, Y.; Song, R.; Mu, Z.; Song, C. Multi-attentive hierarchical dense fusion net for fusion classification of hyperspectral and LiDAR data. *Inf. Fusion* **2022**, *82*, 1–18. [CrossRef]
22. Fan, R.; Li, J.; Song, W.; Han, W.; Yan, J.; Wang, L. Urban informal settlements classification via a transformer-based spatial-temporal fusion network using multimodal remote sensing and time-series human activity data. *Int. J. Appl. Earth Obs. Geoinf.* **2022**, *111*, 102831. [CrossRef]
23. Hu, T.; Yang, J.; Li, X.; Gong, P. Mapping Urban Land Use by Using Landsat Images and Open Social Data. *Remote Sens.* **2016**, *8*, 151. [CrossRef]
24. Liu, X.; He, J.; Yao, Y.; Zhang, J.; Liang, H.; Wang, H.; Hong, Y. Classifying urban land use by integrating remote sensing and social media data. *Int. J. Geogr. Inf. Sci.* **2017**, *31*, 1675–1696. [CrossRef]
25. Jia, Y.; Ge, Y.; Ling, F.; Guo, X.; Wang, J.; Wang, L.; Chen, Y.; Li, X. Urban Land Use Mapping by Combining Remote Sensing Imagery and Mobile Phone Positioning Data. *Remote Sens.* **2018**, *10*, 446. [CrossRef]
26. Tu, W.; Hu, Z.; Li, L.; Cao, J.; Jiang, J.; Li, Q.; Li, Q. Portraying Urban Functional Zones by Coupling Remote Sensing Imagery and Human Sensing Data. *Remote Sens.* **2018**, *10*, 141. [CrossRef]
27. Zhang, F.; Du, B.; Zhang, L. Scene Classification via a Gradient Boosting Random Convolutional Network Framework. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 1793–1802. [CrossRef]
28. Yu, Y.; Liu, F. Aerial Scene Classification via Multilevel Fusion Based on Deep Convolutional Neural Networks. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 287–291. [CrossRef]
29. Yang, N.; Tang, H.; Sun, H.; Yang, X. DropBand: A Simple and Effective Method for Promoting the Scene Classification Accuracy of Convolutional Neural Networks for VHR Remote Sensing Imagery. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 257–261. [CrossRef]
30. Vargas-Munoz, J.E.; Srivastava, S.; Tuia, D.; Falcão, A.X. OpenStreetMap: Challenges and Opportunities in Machine Learning and Remote Sensing. *IEEE Geosci. Remote Sens. Mag.* **2021**, *9*, 184–199. [CrossRef]
31. Machado, G.; Ferreira, E.; Nogueira, K.; Oliveira, H.; Brito, M.; Gama, P.H.T.; Santos, J.A.d. AiRound and CV-BrCT: Novel Multiview Datasets for Scene Classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 488–503. [CrossRef]
32. Han, Z.; Zhang, C.; Fu, H.; Zhou, J.T. Trusted Multi-View Classification. In Proceedings of the International Conference on Learning Representations (ICLR), Online, 3–7 May 2021.
33. Sensoy, M.; Kaplan, L.; Kandemir, M. Evidential Deep Learning to Quantify Classification Uncertainty. In Proceedings of the International Conference on Neural Information Processing Systems (NeurIPS), Montréal, QC, Canada, 3–8 December 2018; pp. 3183–3193.
34. Moon, J.; Kim, J.; Shin, Y.; Hwang, S. Confidence-Aware Learning for Deep Neural Networks. In Proceedings of the PMLR International Conference on Machine Learning (ICML), Online, 13–18 July 2020; Volume 119, pp. 7034–7044.
35. Van Amersfoort, J.; Smith, L.; Teh, Y.W.; Gal, Y. Uncertainty Estimation Using a Single Deep Deterministic Neural Network. In Proceedings of the PMLR International Conference on Machine Learning (ICML), Online, 13–18 July 2020; Volume 119, pp. 9690–9700.
36. Yager, R.R.; Liu, L. *Classic Works of the Dempster-Shafer Theory of Belief Functions*; Springer: Berlin/Heidelberg, Germany, 2010.
37. Jøsang, A. *Subjective Logic: A Formalism for Reasoning under Uncertainty*; Springer: Berlin/Heidelberg, Germany, 2016.
38. Lin, J. On The Dirichlet Distribution. Master's Thesis, Queen's University, Kingston, ON, Canada, 2016; pp. 10–11.
39. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. In Proceedings of the International Conference on Learning Representations (ICLR), San Diego, CA, USA, 7–9 May 2015.

40. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks. *Commun. ACM* **2017**, *60*, 84–90. [CrossRef]
41. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
42. Vo, N.N.; Hays, J. Localizing and Orienting Street Views Using Overhead Imagery. In Proceedings of the European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 11–14 October 2016; pp. 494–509.
43. Geng, W.; Zhou, W.; Jin, S. Multi-View Urban Scene Classification with a Complementary-Information Learning Model. *Photogramm. Eng. Remote Sens.* **2022**, *88*, 65–72. [CrossRef]
44. Zhou, M.; Xu, X.; Zhang, Y. An Attention-based Multi-Scale Feature Learning Network for Multimodal Medical Image Fusion. *arXiv* **2022**, arXiv:2212.04661.
45. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the Inception Architecture for Computer Vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 2818–2826.
46. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely Connected Convolutional Networks. In Proceedings of the EEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.
47. Ju, Y.; Jian, M.; Guo, S.; Wang, Y.; Zhou, H.; Dong, J. Incorporating lambertian priors into surface normals measurement. *IEEE Trans. Instrum. Meas.* **2021**, *70*, 1–13. [CrossRef]
48. Ju, Y.; Shi, B.; Jian, M.; Qi, L.; Dong, J.; Lam, K.M. NormAttention-PSN: A High-frequency Region Enhanced Photometric Stereo Network with Normalized Attention. *Int. J. Comput. Vis.* **2022**, *130*, 3014–3034. [CrossRef]
49. Liang, W.; Tadesse, G.A.; Ho, D.; Fei-Fei, L.; Zaharia, M.; Zhang, C.; Zou, J. Advances, challenges and opportunities in creating data for trustworthy AI. *Nat. Mach. Intell.* **2022**, *4*, 669–677. [CrossRef]