Scene Classification using a Combination of Aerial and Ground Images

Gabriel Machado[1], Keiller Nogueira[2], Jefersson Alex dos Santos[1]
[1]Department of Computer Science, Universidade Federal de Minas Gerais, Belo Horizonte, MG - Brazil
Email: {gabriel.lucas, jefersson}@dcc.ufmg.br
[2]Computing Science and Mathematics, University of Stirling, Stirling, FK9 4LA, Scotland, UK
E-mail: keiller.nogueira@stir.ac.uk

***Abstract*—It is undeniable that aerial images can provide useful information for a large variety of tasks, such as disaster relief, and urban planning. But, since these images only see the Earth from one point of view, some applications may benefit from complementary information provided by other perspective views of the scene, such as ground-level images. Despite a large number of public image repositories for both georeferenced photos and aerial images (such as Google Maps, and Street View), there is a lack of public datasets that allow studies that exploit the complementarity of aerial+ground imagery. Given this, we present two new publicly available datasets named AiRound and CV-BrCT. Using both, we tackled the scene classification task in 2 different scenarios. The first one has a fully-paired image set, while the second has missing samples. In both situations, we used deep learning and feature fusion algorithms. To handle missing samples, we proposed a content-based image retrieval framework.**

***Index Terms*—deep learning; machine learning; remote sensing; image classification; multi-modal machine learning; metric learning; cross-view matching;**

I. Introduction

Satellite images become more accessible to civilian applications each year. New technologies are enabling the wide usage of better and cheaper images in comparison with the past few decades. Nowadays, it is also possible to access many free remote sensing image repositories with a variety of spatial, spectral and temporal resolutions [1]. Images with aerial perspective give us a unique view of the world, allowing the capture of relevant information (not provided by any other type of image) that may assist in several applications, such as automatic geographic mapping and urban planning [2]–[4].

Despite the clear benefits of optical aerial imagery, the fact that they are always looking from the same perspective of the Earth may limit their use. More precisely, the presence of vegetation cover, clouds or simply the need of more detailed on-the-ground information can decrease the effectiveness of such images in some applications. Because of this, in some scenarios (known as multi-view), it is crucial to combine complementary information of aerial and ground images in order to efficiently tackle a problem. Such combination of multiple sources images can benefit many applications in different fields, like 3D human pose estimation [5], geo-localization of places [6], and urban land use [7]. Motivated by these benefits, several approaches [8]–[12] have been proposed to exploit multi-view datasets to face distinct tasks. Although important, it is not easy to find multi-view datasets for specific image tasks, given the difficulty in creating and labeling such data. In fact, as far as we know, there is no publicly available multi-view (aerial and ground) dataset for the scene classification task in the literature.

Due to the lack of publicly available multi-view (aerial and ground) datasets for image classification tasks in the literature, in this dissertation[1] we present two novel multi-view image datasets. The main purpose of creating these datasets is to make them publicly available so that the scientific community can carry out image classification experiments in multi-view scenarios. One of the datasets is composed of $11,753$ triplets of images, each of which consisting of a ground scene, a high-resolution aerial image, and a multi-spectral aerial image. The images are unevenly divided into 11 classes, including airport, bridge, church, forest, lake, park, river, skyscraper, stadium, statue, and tower. An interesting property of our dataset is that it was designed to contain a high intra-class variety, so it was composed with selected places from all around the world. The other dataset is composed of $24k$ pairs of images, each one containing a ground-level scene and a high-resolution aerial image. Those pairs are labeled into 9 different classes, which include apartment, hospital, house, industrial, parking lot, religious, school, store, and vacant lot.

In this dissertation, we designed a series of experiments to exploit the complementary information that pairs of aerial and ground-level images have. In those experiments, we used our datasets to evaluate several state-of-the-art convolutional neural networks (ConvNets) for the scene classification task. To assess the gains from the supplementary information provided by aerial/ground imagery, we evaluated several feature fusion techniques, including early and late fusion. Those algorithms were incorporated with 8 different ConvNets. We also compared the fusion algorithms and analyzed which classes benefit more from them. Since it is not always possible to acquire pairs of aerial + ground images, we also proposed a framework to handle this problem. The main purpose of it is to ensure a way of exploring the complementary information that both images have in a missing data scenario. To evaluate this framework, we simulate a scenario, where we have all the samples from a specific view but do not have their

[1]This work relates to a MSc dissertation.

Dataset	Image Type			Publicly Available	Paired Aerial & Ground Images	Total of Samples	Number of Classes	Task	Year
	Aerial RGB	Ground	Multispectral						
CV-USA [16]	✓	✓	✗	✓	✓	$\sim 44k$	-	Cross-View Matching	2015
Cities [17]	✓	✓	✗	✗	✓	$\sim 156k$	-	Cross-View Matching	2015
Pasadena Urban Trees [10]	✓	✓	✗	✓	✓	$\sim 100k$	18	Object Detection	2016
Vo and Hays [18]	✓	✓	✗	✓	✓	$\sim 1m$	-	Cross-View Matching	2016
Brooklyn and Queens [19]	✓	✓	✗	✓	✓	$\sim 213k$	-	Instance Segmentation	2017
Urban Environments [20]	✓	✓	✗	✗	✓	$\sim 18k$	-	Cross-View Matching	2017
CV-ACT [12]	✓	✓	✗	✓	✓	$\sim 128k$	-	Cross-View Matching	2019
Buildings [8]	✓	✓	✗	✗	✓	$\sim 261k$	4	Classification	2019
Ile-de-France land use [7]	✓	✓	✗	✗	✓	∼50k	16	Classification	2019
CV-London [21]	✓	✓	✗	✗	✓	∼2k	-	Image Synthesis	2020
AiRound [22]	✓	✓	✓	✓	✓	∼35.4k	11	Classification	2020
CV-BrCT [22]	✓	✓	✗	✓	✓	∼48k	9	Classification	2020

TABLE I: Properties of other datasets found in the literature that are similar to AiRound and CV-BrCT.

correspondent pairs from the other domain.

The remainder of this document is structured as follows. Section II introduces the datasets, the used feature fusion algorithms and our missing data framework. Section III briefly discusses some of the results obtained in this dissertation. And lastly, in Section IV we conclude our work.

II. Methodology

A. Datasets

Recent advances in satellite data acquisition and cloud computing, access to high-resolution satellite images and other types of data was made easier. Despite the great advantages that aerial images provide, some applications demand information that an aerial perspective may lack. In these cases, an alternative solution is to use complementary perspectives of the same location, *i.e.*, ground-level view, to better seek such information [13]–[15]. Due to the high demand for images to be used in multi-view tasks, naturally some multi-view datasets were proposed in the literature. In Table I, we summarize some datasets that are similar to the novel ones proposed for this dissertation.

Differentiating our datasets from the ones in Table I, some of the existing datasets were designed in a way that each image pair can be seen as a class. Such datasets do not contain groups of classes that share the same label, which ends up making its use for image classification impractical. Other datasets are quite different from the ones proposed here, given that the main task for which they were proposed is different. That difference resides in the fact that those problems require different types of label as input and also generates distinct outputs (bounding boxes and segmentation). Relating to multi-view image classification datasets, two datasets [7], [8] are quite similar to both datasets proposed here. However, neither of these existing datasets are publicly available, while ours are.

1) AiRound: The first dataset is named AiRound, and is composed of $11,753$ images distributed into 11 classes, including: airport, bridge, church, forest, lake, park, river, skyscraper, stadium, statue, and tower. Each sample is composed of a triplet, that contains images in 3 distinct points of view: (i) a ground perspective image; (ii) a high resolution RGB aerial image; and (iii) a multi-spectral image from the Sentinel-2 satellite. All the images collected for this dataset correspond to real places around the world.

This dataset was created using two methods. In the first, to download the samples, two types of metadata were required: (i) the name of the place; and (ii) its correspondent geographical coordinates. These metadata were collected using web crawlers in various lists of Wikipedia web pages.

Given the metadata, the RGB aerial images were collected using Bing Maps API[2]. The zoom level was empirically selected in order to adapt a proper vision for the samples of each class.

In order to collect the ground level samples, it was checked if the correspondent class exists in the Google Places' database. If the sample class exists, a query was built using this place's geographical coordinates as input. The outputs returned by this API were all manually checked, and if they did not correspond to the class, a second protocol was performed. The second protocol was used for cases that the class did not exist in Google Places database or the image retrieved did not correspond to the query requested. This protocol consists of crawling the top 5 images from Google Images using, as query, the place's name. Finally, an image was manually selected to represent each sample on AiRound, the best instance between the 5 images downloaded.

Relating to the Sentinel-2 images acquisition, we followed exactly the same protocol that was proposed by [23]. In this protocol, Google Earth Engine [1] was used to download the data using the place's geographical coordinates. After careful analysis, we decided to resize all images to 224x224 pixels, a resolution that could cover most of the classes' areas.

After working with this methodology for a while, we noticed that it was not scalable because of limited metadata (per class) available in the Wikipedia lists. Due to this, we decided to move to another, more scalable, method. So, a second methodology was applied to build this dataset, the metadata were obtained from the publicly available data of the OpenStreetMap[3], and collected using the Overpass API[4]. These lists consist of only geographic coordinates, for most

[2]https://docs.microsoft.com/en-us/bingmaps/

[3]www.openstreetmap.org/

[4]https://overpass-turbo.eu/

of the classes, with exception of the classes forest, lake, river, and park, which we collected only places that have a name assigned to it. The lists are then fed to scripts that utilize the Google StaticMap API[5], to collect the aerial images, and the Google StreetView API[6], to collect the frontal images. Except for the zoom parameter, which was set to a proper value per class empirically, the default values of the Google APIs were used for the aerial images. Since we could not retrieve street-level images for the classes forest, lake, river, and park, we used their names in a query to a Google Images crawler. We followed the same protocol used in the first methodology to download images from these classes. To download the Sentinel-2 images, we also applied the same protocol used in the first methodology. Finally, since we gathered a large collection of locations, we ignored points where we could not retrieve an image from each view.

2) CV-BrCT: The CV-BrCT dataset, which stands for **Cross-View Brazilian Construction Types**, comprises approximate $24k$ pairs of images split into 9 urban classes, *i.e.*, apartment, hospital, house, industrial, parking lot, religious, school, store, and vacant lot. The pairs are composed of images from two different views: an aerial view, and a frontal view of a location.

The lists of coordinates used to build this dataset were collected using the second methodology used to build AiRound dataset, which was applied to all classes, except Vacant Lot, which was manually annotated. To download the samples, we also use the same APIs of the second methodology previously described. Finally, it is important to mention that we decided to not collect Sentinel-2 samples for this dataset. This decision was made considering the nature of all the classes, which only include objects that would not benefit from Sentinel-2's resolution.

B. Feature Fusion

To enhance scene classification results, we evaluated several models for early and late fusion in order compare both approaches. In this work, those techniques were applied to fuse aerial/ground/satellite features, acquiring new information, and using them to enhance the final predictions.

1) Early Fusion Methods: The early fusion strategy performed in this work consists of using the first feature extraction layers of the target network as a backbone. This backbone is replicated to aerial and ground images. The fusion of the features is made by applying a concatenation layer on the low-level features, which results in a tensor that contains twice as many kernels as the original ones. The choice of where those concatenations were performed is based on the total of kernels that each convolution layer has. So, in order to be possible to fully explore pre-trained models, we decided to concatenate those feature vectors before the first convolution layer that doubles its amount of kernels in the target network. In this way, we ignore the convolution that duplicates this amount of kernels and substitute it for a fusion that also duplicates the amount of feature vectors. Figure 1 represents the early fusion methodology proposed for this work. The first L_e layers (blue and green blocks in the figure) represent the early feature extraction process, which is made individually for each view. After a few layers, the features are concatenated and transported to the remainder of the architecture, which was used as a base (yellow block in the figure). After that, when the high-level features are extracted, the classification process is performed.

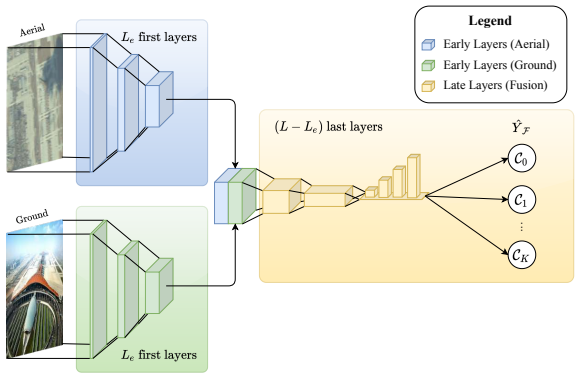

Fig. 1: Example of the proposed early fusion methodology.

2) Late Fusion Methods: The late fusion strategy performed in this work is illustrated in Figure 2. One should note that the softmax scores were used to combine the results. For the experiments, we used 5 different classic late fusion algorithms [24]. Those include: sum, minimum, majority voting, weighted sum, and product.

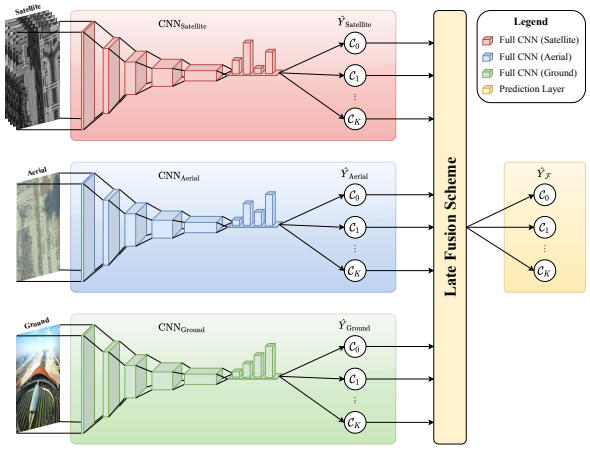

Fig. 2: A typical late fusion pipeline. As can be seen, each ConvNet is trained individually and their outputs are combined using a late fusion algorithm, resulting in the final prediction.

C. Classification with Missing Data

Due to the undeniable importance of studying techniques to handle missing data for multi-modal machine learning [7], [25]–[27], in this dissertation, we also decided to propose

[5]https://developers.google.com/maps/documentation/maps-static/intro
[6]https://developers.google.com/maps/documentation/streetview/intro

a framework to tackle this task. This framework operates using a retrieval approach, *viz.* cross-view matching. Since neither AiRound nor CV-BrCT datasets have missing pairs for any sample, we simulate a missing data scenario on them to test our framework. In this simulation, we have ground-level images but remove their correspondent pairs from the aerial domain. Our missing data framework is represented in Figure 4.

In our simulation, both datasets were split according to the protocol illustrated by Figure 3. One may note that the aerial samples from the test suites do not exist. To supply these data, we decided to re-use the validation set of aerial images as a database for the content-based image retrieval task. The reason that we decided to re-use such data is because we wanted to have as much data as possible to test our framework. And, since those validation sets were not used to update weights during the training phase, we argue that this strategy is fair.

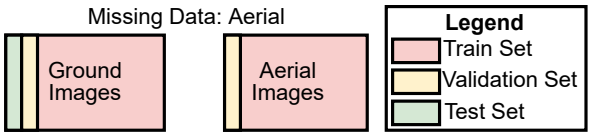

Fig. 3: This figure represents the exact scheme that was used to split AiRound and CV-BrCT datasets for missing data experiments.

In the training procedure, three different networks are trained, two of which are used for image classification, and one for image retrieval. Unlike the cross-view matching task, in our missing data classification approach, we do not have image pairs in the testing phase. Because of that, we performed two major changes in the original training protocol. The first modification in the training procedure consists of using as input random samples from the same class, instead of only image pairs. This is because in the testing phase of our missing data scenario, there is no original correspondent pair to the query (Figure 3). So, it is better to train the encoders to approximate samples of the same class, instead of just using image pairs.

The second change was performed in the batch construction. Since most of the recent cross-view matching networks [6], [12], [28], [29] uses the exhaustive mini-batch strategy[7], we also applied it. Because of that, we only used batches that contain exactly one sample of the same class from aerial and ground domains. So, it ensures that samples from same class of the query will never be used as negative anchors. It is important to mention that the decision of modifying the construction of batches was inspired by the semantic triplet loss, proposed by [34].

$$\alpha_{i\times j} = 2 \times (1 - F_g F_a^{\intercal}) \tag{1}$$

Figure 4 shows how the inference process is performed in our framework. In the first step, a forward through the retrieval network is executed in the entire test set. All the output feature vectors are stored in two different matrices, being F_a the aerial features matrix, and F_g the ground one. After that, a distance matrix $\alpha_{i\times j}$ is calculated using F_a and F_g, as can be checked in equation 1.

Through this framework, we can compute a top-k image ranking to a query by using Equation 2. In this formula, α represents the distance matrix (computed using Equation 1), q represents the correspondent line to the query image in matrix α, and k corresponds to the ranking size.

$$Rank(q,k) = \arg\min(\alpha_{q,c}, k)\ \forall c \in [0, j) \tag{2}$$

The second step of Figure 4 is where the classification is performed. First, through a forward process, we calculate the softmax scores for the query image q using the ground CNN. After that, we execute inference processes through aerial CNN for all images contained in the ranking, saving all the scores returned by the network.

Finally, to calculate the final prediction, we use both softmax scores returned by the inferences, and calculate a fusion of them using the product fusion [22].

III. Experiments

We carry out 4 different sets of experiments in order to evaluate if it is possible to exploit complementary information from aerial and ground images. The first one consists of networks trained using only one view, and it is used to check if the networks with Aerial+Ground fusions lead to improvements. For the second set, we carry out experiments using the proposed early fusion methodology. For the third set of experiments, we employed late fusion algorithms using different types of views. Lastly, for the fourth set, we employed experiments using our missing data framework.

A. Single-View Classification

1) AiRound: The results for AiRound for the experiments using a single type of image are presented in Table II.

2) CV-BrCT: The results for CV-BrCT for the experiments using a single type of image are presented in Table III.

B. Early Fusion

1) AiRound: In Table IV, comparing the same architecture models, it is notable that most of the results show improvements, as compared to the 1-view results reported in Table II.

[6]Cross-view matching is essentially a geo-localization task, which involves data from aerial and ground domains. The main objective of this task is to estimate the corresponding image pair from one domain, given a query from the other domain.

[7]The exhaustive mini-batch strategy was proposed by [18] to maximize the number of triplets within each batch. Since defining a negative sample for triplet training is essentially a hard task [30]–[33], this technique proposes to use all samples (excluding the positive anchor to the triplet) as negative anchors. To formally define this technique, consider β as the batch size. In each iteration of exhaustive mini-batch strategy, there is a total of $\beta \times 2 \times (\beta - 1)$ triplets. This is because in each batch we have β positive pairs, and this strategy uses all the batch not-paired samples as negative anchors, so for each ground image and/or aerial image, there is also $\beta - 1$ negative pairs, resulting in $2 \times (\beta - 1)$ negative triplets, totaling $\beta \times 2 \times (\beta - 1)$.

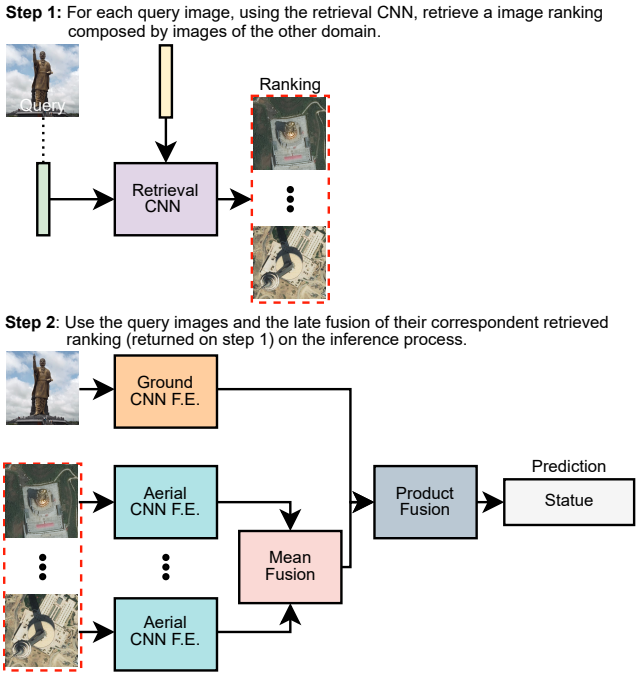

Fig. 4: Testing pipeline of the schema proposed to simulate a missing data scenario. Since there is not any paired image on the proposed test sets, in **Step 1**, an inference on a cross-view matching (retrieval) network is performed to acquire possible candidates. In this inference, all the ground images are used as queries, and for each query, the retrieval network returns a ranking of aerial images, taken from the test set of the dataset. In **Step 2**, the classification process is performed. One may note that the features extracted from the images of the ranking are fused before being used to classify the query image. As a last step, the result of the fusion of this ranking is combined with the ground image (query) features, resulting in the final prediction.

Network	Input Data			
	Aerial		Ground	
	B. Acc.	F1	B. Acc.	F1
AlexNet	0.76 ± 0.01	0.76 ± 0.01	0.71 ± 0.01	0.70 ± 0.01
VGG	0.82 ± 0.01	0.81 ± 0.01	0.76 ± 0.01	0.76 ± 0.01
Inception	$\mathbf{0.85 \pm 0.01}$	$\mathbf{0.84 \pm 0.01}$	$\mathbf{0.78 \pm 0.01}$	$\mathbf{0.78 \pm 0.01}$
ResNet	0.81 ± 0.01	0.80 ± 0.00	0.75 ± 0.01	0.75 ± 0.01
DenseNet	0.84 ± 0.01	$\mathbf{0.84 \pm 0.01}$	0.77 ± 0.01	0.77 ± 0.02
SqueezeNet	0.76 ± 0.00	0.76 ± 0.01	0.72 ± 0.02	0.72 ± 0.02
SENet	0.83 ± 0.01	0.83 ± 0.01	0.77 ± 0.01	0.76 ± 0.01
SKNet	0.84 ± 0.01	$\mathbf{0.84 \pm 0.01}$	0.77 ± 0.01	0.77 ± 0.01

TABLE II: Results in terms of F1 Score of the evaluated models for AiRound dataset. Bold values represent the best results achieved for each input data domain.

2) CV-BrCT: Comparing the Tables III and V, the results of networks using only ground images were worse than the aerial image models. Through the experiments, we noted that using a network that merges and combines features of both images from the start leads to no improvements, if we compare to networks using only aerial images.

Network	Input Data			
	Aerial		Ground	
	B. Acc.	F1	B. Acc.	F1
AlexNet [35]	0.75 ± 0.02	0.84 ± 0.01	0.54 ± 0.01	0.66 ± 0.01
VGG [36]	0.79 ± 0.03	0.87 ± 0.01	$\mathbf{0.60 \pm 0.02}$	$\mathbf{0.71 \pm 0.01}$
Inception [37]	$\mathbf{0.80 \pm 0.02}$	0.87 ± 0.00	$\mathbf{0.60 \pm 0.03}$	$\mathbf{0.71 \pm 0.01}$
ResNet [38]	0.78 ± 0.02	0.86 ± 0.01	0.58 ± 0.04	0.69 ± 0.02
DenseNet [39]	$\mathbf{0.80 \pm 0.02}$	0.87 ± 0.01	$\mathbf{0.60 \pm 0.01}$	$\mathbf{0.71 \pm 0.01}$
SqueezeNet [40]	0.70 ± 0.02	0.80 ± 0.01	0.56 ± 0.02	0.68 ± 0.01
SENet [41]	$\mathbf{0.80 \pm 0.02}$	0.87 ± 0.01	$\mathbf{0.60 \pm 0.01}$	$\mathbf{0.71 \pm 0.01}$
SKNet [42]	$\mathbf{0.80 \pm 0.03}$	$\mathbf{0.88 \pm 0.01}$	$\mathbf{0.60 \pm 0.02}$	$\mathbf{0.71 \pm 0.01}$

TABLE III: Results of the evaluated models for CV-BrCT dataset. Bold values represent the best results achieved for each training strategy, metric, and input data domain.

Early Fusion Network	B. Acc.	F1
AlexNet [35]	0.81 ± 0.02	0.80 ± 0.02
VGG [36]	0.84 ± 0.02	0.84 ± 0.02
Inception [37]	0.84 ± 0.01	0.84 ± 0.01
ResNet [38]	0.83 ± 0.02	0.83 ± 0.02
DenseNet [39]	0.83 ± 0.01	0.83 ± 0.01
SqueezeNet [40]	0.78 ± 0.02	0.77 ± 0.02
SENet [41]	0.84 ± 0.02	0.83 ± 0.01
SKNet [42]	$\mathbf{0.86 \pm 0.02}$	$\mathbf{0.86 \pm 0.02}$

TABLE IV: Results of the evaluated early fusion networks for AiRound dataset. Bold values represent the best results achieved for each training strategy and metric.

Early Fusion Network	B. Acc.	F1
AlexNet [35]	0.72 ± 0.02	0.82 ± 0.01
VGG [36]	0.76 ± 0.02	0.84 ± 0.02
Inception [37]	0.79 ± 0.03	$\mathbf{0.87 \pm 0.01}$
ResNet [38]	0.74 ± 0.02	0.83 ± 0.01
DenseNet [39]	0.72 ± 0.03	0.81 ± 0.01
SqueezeNet [40]	0.67 ± 0.04	0.79 ± 0.02
SENet [41]	0.78 ± 0.02	0.86 ± 0.01
SKNet [42]	$\mathbf{0.80 \pm 0.02}$	$\mathbf{0.87 \pm 0.01}$

TABLE V: Results of the evaluated early fusion networks for CV-BrCT dataset. Bold values represent the best results achieved for each training strategy and metric.

C. Late Fusion

1) AiRound: The late fusion results for the AiRound dataset are presented in Table VI. Comparing the results using only one type of data (Table II) with the fusion outcomes, it is possible to notice that the late fusion outperformed all approaches using only one view. This corroborates our initial insight that the combination of multi-source data could improve the results for the scene classification task. In Table VI, the best overall results were achieved by the DenseNet ([39]) architecture.

2) CV-BrCT: The same set of late fusion experiments were perfomed for CV-BrCT dataset. Overall, all fusion methods improved the results of the networks trained with a single type. The results across fusion methods are similar, although some techniques show a consistent improvement, e.g., Weighted

Sum, and others do not appear to have a noticeable effect, e.g., Minimum.

D. Aerial+ground fusion with missing data evaluation

1) AiRound: The results of our framework applied in the AiRound dataset can be checked in Table VII. In these experiments, a network trained using only ground images is used as a baseline. This is because without performing a retrieval using our framework, the only data available for inference are the ground images.

Relating to the outcomes, all the networks downgraded their results, compared to a network using only ground data, using fusion with only the retrieved aerial images from the top 1. However, as the ranking retrieved by CirVGG [29] is increased, the results from Table VII improve. By using the fusion of top 50 retrieved images, all the results from the benchmarked methods were improved, comparing to the baseline. Given this, as expected, the best outcome was produced when using the retrieved top 100 images. The best result and also the biggest gain by using the framework was achieved by SKNet [42]. This network achieved a score of 0.77 in both metrics, which consists of a gain of 0.03 comparing to the baseline.

Data Used	Classification Model					
	VGG [36]		DenseNet [39]		SKNet [42]	
	B. Acc.	F1	B. Acc.	F1	B. Acc.	F1
Only Ground	0.74 ± 0.02	0.74 ± 0.02	0.75 ± 0.01	0.74 ± 0.02	0.74 ± 0.02	0.74 ± 0.02
Top 1	0.70 ± 0.02	0.69 ± 0.01	0.71 ± 0.02	0.71 ± 0.01	0.70 ± 0.02	0.70 ± 0.02
Top 5	0.73 ± 0.01	0.73 ± 0.00	0.74 ± 0.02	0.74 ± 0.02	0.74 ± 0.01	0.74 ± 0.01
Top 10	0.74 ± 0.01	0.74 ± 0.00	0.75 ± 0.01	0.74 ± 0.01	0.75 ± 0.01	0.75 ± 0.01
Top 50	0.75 ± 0.01	0.75 ± 0.01	$\mathbf{0.76 \pm 0.01}$	0.75 ± 0.01	0.76 ± 0.01	0.76 ± 0.01
Top 100	$\mathbf{0.76 \pm 0.01}$	$\mathbf{0.76 \pm 0.01}$	$\mathbf{0.76 \pm 0.00}$	$\mathbf{0.76 \pm 0.00}$	$\mathbf{0.77 \pm 0.01}\dagger$	$\mathbf{0.77 \pm 0.02}\dagger$

TABLE VII: Classification results in the proposed missing data scenario using the AiRound dataset. The red coloring indicates a downgrade compared to the baseline, while the blue coloring represents an upgrade. The values in bold indicate the best result for each metric and network, while the † symbols mark the best overall results for each metric.

2) CV-BrCT: Table VIII contains the results achieved by our missing data framework. Just like happened for the AiRound dataset (Table VII), the framework downgraded the baseline results by using only the first image retrieved by CirVGG. By using the top 5 ranking fusion, we achieved some gains for VGG [36] and DenseNet [39]. Despite that, we did not achieve any gain for SKNet [42]. We also highlight that using the remaining rankings (top 10, 50, and 100) did not lead to better results. The best results and also improvements were achieved by DenseNet [39]. Those results were achieved by using the top 5, 10, 50, and 100 ranking fusions, resulting in a score of 0.72 in F1 score metric and 0.66 in balanced accuracy. Comparing to the baseline, this represents a gain of 0.03 and 0.02 in F1 score and balanced accuracy, respectively.

Data Used	Classification Model					
	VGG [36]		DenseNet [39]		SKNet [42]	
	B. Acc.	F1	B. Acc.	F1	B. Acc.	F1
Only Ground	0.64 ± 0.04	0.68 ± 0.06	0.64 ± 0.03	0.69 ± 0.02	0.65 ± 0.02	$\mathbf{0.71 \pm 0.01}$
Top 1	0.59 ± 0.01	0.65 ± 0.02	0.59 ± 0.00	0.66 ± 0.02	0.58 ± 0.02	0.65 ± 0.03
Top 5	$\mathbf{0.65 \pm 0.01}$	$\mathbf{0.70 \pm 0.02}$	$\mathbf{0.66 \pm 0.02}\dagger$	$\mathbf{0.72 \pm 0.00}\dagger$	0.65 ± 0.01	$\mathbf{0.71 \pm 0.02}$
Top 10	$\mathbf{0.65 \pm 0.01}$	$\mathbf{0.70 \pm 0.02}$	$\mathbf{0.66 \pm 0.02}\dagger$	$\mathbf{0.72 \pm 0.01}\dagger$	0.65 ± 0.01	$\mathbf{0.71 \pm 0.02}$
Top 50	$\mathbf{0.65 \pm 0.01}$	$\mathbf{0.70 \pm 0.02}$	$\mathbf{0.66 \pm 0.03}\dagger$	$\mathbf{0.72 \pm 0.01}\dagger$	$\mathbf{0.66 \pm 0.02}$	$\mathbf{0.71 \pm 0.02}$
Top 100	$\mathbf{0.65 \pm 0.01}$	$\mathbf{0.70 \pm 0.02}$	$\mathbf{0.66 \pm 0.03}\dagger$	$\mathbf{0.72 \pm 0.01}\dagger$	0.65 ± 0.02	$\mathbf{0.71 \pm 0.02}$

TABLE VIII: Classification results in the proposed missing data scenario using the CV-BrCT dataset. The red coloring indicates a downgrade compared to the baseline, while the blue coloring represents an upgrade. The values in bold indicate the best result for each metric and network, while the † symbols mark the best overall results for each metric.

IV. Conclusion

In this dissertation, we proposed two novel cross-view image datasets that can be used for multi-purpose tasks. Those datasets were named AiRound and CV-BrCT, and can be downloaded in our project's web page (http://www.patreo.dcc.ufmg.br/multi-view-datasets/). Using those datasets, we performed a series of experiments to address the important task of fusing features from images of different perspectives, which resulted in a publication in an international journal paper [22]. Lastly, we also address the task of multi-view missing data classification that resulted in a framework, which achieved promising results, *i.e.*, improvements compared to the use of networks with only one-view data. It is important to mention that the publication of this framework is still on progress.

For future work, we plan to explore different possibilities to perform missing data completion. One good alternative is to simulate the opposite scenario to the one used in this work, *i.e.*, a situation where images from an aerial perspective are available, while their correspondent pairs from the ground domain are not. Other aspects that could be exploited are the use of different metric learning losses, different architectures, or even clustering algorithms.

Network	Fusion Strategy									
	Sum		M. Voting		W. Sum		Minimum		Product†	
	B. Acc.	F1 Score	B. Acc.	F1 Score	B. Acc.	F1 Score	B. Acc.	F1 Score	B. Acc.	F1 Score
AlexNet [35]	0.84 ± 0.00	0.84 ± 0.00	0.83 ± 0.01	0.83 ± 0.01	0.83 ± 0.01	0.83 ± 0.01	0.85 ± 0.01	0.85 ± 0.01	0.86 ± 0.01	0.86 ± 0.01
VGG [36]	0.88 ± 0.01	0.88 ± 0.01	0.88 ± 0.01	0.87 ± 0.01	$\mathbf{0.88 \pm 0.01}$	0.87 ± 0.01	0.89 ± 0.01	0.89 ± 0.01	0.90 ± 0.00	0.90 ± 0.00
Inception [37]	0.88 ± 0.01	0.88 ± 0.01	0.88 ± 0.01	0.87 ± 0.01	$\mathbf{0.88 \pm 0.01}$	0.87 ± 0.01	0.89 ± 0.01	0.89 ± 0.01	0.90 ± 0.00	0.90 ± 0.00
ResNet [38]	0.88 ± 0.01	0.87 ± 0.01	0.87 ± 0.01	0.87 ± 0.01	0.86 ± 0.02	0.86 ± 0.02	0.88 ± 0.01	0.88 ± 0.01	0.89 ± 0.01	0.89 ± 0.01
DenseNet [39]†	$\mathbf{0.90 \pm 0.01}$	$\mathbf{0.89 \pm 0.01}$	$\mathbf{0.89 \pm 0.01}$	$\mathbf{0.89 \pm 0.01}$	$\mathbf{0.88 \pm 0.01}$	$\mathbf{0.88 \pm 0.01}$	$\mathbf{0.90 \pm 0.01}$	$\mathbf{0.90 \pm 0.01}$	$\mathbf{0.91 \pm 0.01}$	$\mathbf{0.91 \pm 0.01}$
SqueezeNet [40]	0.85 ± 0.01	0.85 ± 0.01	0.85 ± 0.01	0.85 ± 0.01	0.85 ± 0.01	0.84 ± 0.01	0.86 ± 0.01	0.86 ± 0.01	0.87 ± 0.01	0.86 ± 0.01
SENet [41]	0.89 ± 0.02	$\mathbf{0.89 \pm 0.02}$	0.88 ± 0.02	0.88 ± 0.02	0.87 ± 0.01	0.87 ± 0.01	$\mathbf{0.90 \pm 0.01}$	0.89 ± 0.01	0.90 ± 0.01	0.90 ± 0.01
SKNet [42]	$\mathbf{0.90 \pm 0.01}$	$\mathbf{0.89 \pm 0.01}$	$\mathbf{0.89 \pm 0.01}$	$\mathbf{0.89 \pm 0.01}$	$\mathbf{0.88 \pm 0.01}$	$\mathbf{0.88 \pm 0.01}$	$\mathbf{0.90 \pm 0.01}$	$\mathbf{0.90 \pm 0.00}$	0.90 ± 0.01	0.90 ± 0.01

TABLE VI: Results of the evaluated late fusion techniques for AiRound dataset using fine-tuned models. Bold values represent the best results achieved for each fusion strategy and metric. While the † symbols mark the best overall network and fusion strategy.

REFERENCES

[1] N. Gorelick, M. Hancher, M. Dixon, S. Ilyushchenko, D. Thau, and R. Moore, "Google earth engine: Planetary-scale geospatial analysis for everyone," *Remote Sensing of Environment*, 2017.

[2] S. S. Rwanga, J. M. Ndambuki *et al.*, "Accuracy assessment of land use/land cover classification using remote sensing and gis," *International Journal of Geosciences*, vol. 8, no. 04, p. 611, 2017.

[3] P. Zhang, Y. Ke, Z. Zhang, M. Wang, P. Li, and S. Zhang, "Urban land use and land cover classification using novel deep learning models based on high spatial resolution satellite imagery," *Sensors*, vol. 18, no. 11, p. 3717, 2018.

[4] C. Zhang, I. Sargent, X. Pan, H. Li, A. Gardiner, J. Hare, and P. M. Atkinson, "An object-based convolutional neural network (ocnn) for urban land use classification," *Remote Sensing of Environment*, vol. 216, pp. 57–70, 2018.

[5] H. Qiu, C. Wang, J. Wang, N. Wang, and W. Zeng, "Cross view fusion for 3d human pose estimation," in *IEEE International Conference on Computer Vision*, October 2019.

[6] S. Hu, M. Feng, R. M. Nguyen, and G. Hee Lee, "Cvm-net: Cross-view matching network for image-based ground-to-aerial geo-localization," in *IEEE/CVF Computer Vision and Pattern Recognition*, 2018.

[7] S. Srivastava, J. E. Vargas-Muñoz, and D. Tuia, "Understanding urban landuse from the above and ground perspectives: A deep learning, multimodal solution," *Remote Sensing of Environment*, 2019.

[8] E. J. Hoffmann, Y. Wang, M. Werner, J. Kang, and X. X. Zhu, "Model fusion for building type classification from aerial and street view images," *Remote Sensing*, vol. 11, no. 11, p. 1259, May 2019.

[9] N. Ghouaiel and S. LefÃ¨vre, "Coupling ground-level panoramas and aerial imagery for change detection," *Geo-spatial Information Science*, vol. 19, no. 3, pp. 222–232, 2016.

[10] J. D. Wegner, S. Branson, D. Hall, K. Schindler, and P. Perona, "Cataloging public objects using aerial and street-level images-urban trees," in *IEEE/CVF Computer Vision and Pattern Recognition*, 2016.

[11] R. Cao, J. Zhu, W. Tu, Q. Li, J. Cao, B. Liu, Q. Zhang, and G. Qiu, "Integrating aerial and street view images for urban land use classification," *Remote Sensing*, vol. 10, no. 10, p. 1553, Sep 2018.

[12] L. Liu and H. Li, "Lending orientation to neural networks for cross-view geo-localization," in *International Conference on Pattern Recognition*, 2019.

[13] A. L. Majdik, Y. Albers-Schoenberg, and D. Scaramuzza, "Mav urban localization from google street view data," in *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2013, pp. 3979–3986.

[14] M. Rumpler, A. Tscharf, C. Mostegel, S. Daftry, C. Hoppe, R. Prettenthaler, F. Fraundorfer, G. Mayer, and H. Bischof, "Evaluations on multi-scale camera networks for precise and geo-accurate reconstructions from aerial and terrestrial images with user guidance," *Computer Vision and Image Understanding*, vol. 157, pp. 255–273, 2017.

[15] M. Zhai, Z. Bessinger, S. Workman, and N. Jacobs, "Predicting ground-level scene layout from aerial imagery," in *IEEE/CVF Computer Vision and Pattern Recognition*, 2017.

[16] S. Workman, R. Souvenir, and N. Jacobs, "Wide-area image geolocalization with aerial reference imagery," in *IEEE International Conference on Computer Vision*, December 2015.

[17] T.-Y. Lin, Y. Cui, S. Belongie, and J. Hays, "Learning deep representations for ground-to-aerial geolocalization," in *IEEE/CVF Computer Vision and Pattern Recognition*, 2015, pp. 5007–5015.

[18] N. N. Vo and J. Hays, "Localizing and orienting street views using overhead imagery," in *European Conference on Computer Vision*. Springer, 2016.

[19] S. Workman, M. Zhai, D. J. Crandall, and N. Jacobs, "A unified model for near and remote sensing," in *IEEE International Conference on Computer Vision*, 2017, pp. 2688–2697.

[20] Y. Tian, C. Chen, and M. Shah, "Cross-view image matching for geo-localization in urban environments," in *IEEE/CVF Computer Vision and Pattern Recognition*, 2017, pp. 3608–3616.

[21] X. Lu, Z. Li, Z. Cui, M. R. Oswald, M. Pollefeys, and R. Qin, "Geometry-aware satellite-to-ground image synthesis for urban areas," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 859–867.

[22] G. Machado, E. Ferreira, K. Nogueira, H. Oliveira, M. Brito, P. H. T. Gama, and J. A. d. Santos, "Airound and cv-brct: Novel multiview datasets for scene classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 488–503, 2021.

[23] E. Ferreira, M. Brito, R. Balaniuk, M. S. Alvim, and J. A. dos Santos, "Brazildam: A benchmark dataset for tailings dam detection," in *2020 IEEE Latin American GRSS & ISPRS Remote Sensing Conference (LAGIRS)*. IEEE, 2020, pp. 339–344.

[24] E. Alpaydin, *Introduction to machine learning*. MIT press, 2014.

[25] M. K.-P. Ng, Q. Yuan, L. Yan, and J. Sun, "An adaptive weighted tensor completion method for the recovery of remote sensing images with missing data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 6, pp. 3367–3381, 2017.

[26] L. Cai, Z. Wang, H. Gao, D. Shen, and S. Ji, "Deep adversarial learning for multi-modality missing data completion," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery Data Mining*, ser. KDD '18. New York, NY, USA: Association for Computing Machinery, 2018, p. 1158â1166.

[27] L. Zhang, Y. Zhao, Z. Zhu, D. Shen, and S. Ji, "Multi-view missing data completion," *IEEE Transactions on Knowledge and Data Engineering*, vol. 30, no. 7, pp. 1296–1309, 2018.

[28] Y. Shi, L. Liu, X. Yu, and H. Li, "Spatial-aware feature aggregation for image based cross-view geo-localization," in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds. Curran Associates, Inc., 2019.

[29] Y. Shi, X. Yu, D. Campbell, and H. Li, "Where am i looking at? joint location and orientation estimation by cross-view matching," in *IEEE/CVF Computer Vision and Pattern Recognition*, 2020, pp. 4064–4072.

[30] H. Oh Song, Y. Xiang, S. Jegelka, and S. Savarese, "Deep metric learning via lifted structured feature embedding," in *IEEE/CVF Computer Vision and Pattern Recognition*, 2016, pp. 4004–4012.

[31] H. Liu, Y. Tian, Y. Yang, L. Pang, and T. Huang, "Deep relative distance learning: Tell the difference between similar vehicles," in *IEEE/CVF Computer Vision and Pattern Recognition*, 2016, pp. 2167–2175.

[32] K. Sohn, "Improved deep metric learning with multi-class n-pair loss objective," *Advances in Neural Information Processing Systems*, vol. 29, pp. 1857–1865, 2016.

[33] Y. Zhao, Z. Jin, G.-j. Qi, H. Lu, and X.-s. Hua, "An adversarial approach to hard triplet generation," in *European Conference on Computer Vision*, 2018, pp. 501–517.

[34] M. Carvalho, R. Cadène, D. Picard, L. Soulier, N. Thome, and M. Cord, "Cross-modal retrieval in the cooking context: Learning semantic text-image embeddings," in *The 41st International ACM SIGIR Conference on Research Development in Information Retrieval*, ser. SIGIR '18. New York, NY, USA: Association for Computing Machinery, 2018, p. 35â44.

[35] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, 2012, pp. 1097–1105.

[36] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[37] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *IEEE/CVF Computer Vision and Pattern Recognition*, 2016.

[38] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE/CVF Computer Vision and Pattern Recognition*, 2016.

[39] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *IEEE/CVF Computer Vision and Pattern Recognition*, 2017, pp. 4700–4708.

[40] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "Squeezenet: Alexnet-level accuracy with 50x fewer parameters and¡ 0.5 mb model size," *arXiv preprint arXiv:1602.07360*, 2016.

[41] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *IEEE/CVF Computer Vision and Pattern Recognition*, 2018, pp. 7132–7141.

[42] X. Li, W. Wang, X. Hu, and J. Yang, "Selective kernel networks," in *IEEE/CVF Computer Vision and Pattern Recognition*, 2019, pp. 510–519.