IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

Facing the Void: Overcoming Missing Data in Multi-View Imagery

GABRIEL MACHADO[1], MATHEUS B. PEREIRA[1], KEILLER NOGUEIRA[2], AND JEFERSSON A. DOS SANTOS[1,2] (Senior Member, IEEE)

[1]Department of Computer Science, Universidade Federal de Minas Gerais, Brazil (e-mail: gabriel.lucas@dcc.ufmg.br; matheuspereira@dcc.ufmg.br; jefersson@dcc.ufmg.br)

[2]Computing Science and Mathematics, University of Stirling, Stirling, FK9 4LA, Scotland, UK (e-mail: keiller.nogueira@stir.ac.uk)

Corresponding author: Gabriel Machado (e-mail: gabriel.lucas@dcc.ufmg.br).

The authors acknowledge the Serrapilheira Institute [grant number Serra – R-2011-37776]. This work was supported in part by FAPEMIG [grant number APQ-00449-17], CNPq [grant number 306955/2021-0], and CAPES – Finance Code 001.

ABSTRACT In some scenarios, a single input image may not be enough to allow the object classification. In those cases, it is crucial to explore the complementary information extracted from images presenting the same object from multiple perspectives (or views) in order to enhance the general scene understanding and, consequently, increase the performance. However, this task, commonly called multi-view image classification, has a major challenge: missing data. In this paper, we propose a novel technique for multi-view image classification robust to this problem. The proposed method, based on state-of-the-art deep learning-based approaches and metric learning, can be easily adapted and exploited in other applications and domains. A systematic evaluation of the proposed algorithm was conducted using two multi-view aerial-ground datasets with very distinct properties. Results show that the proposed algorithm provides improvements in multi-view image classification accuracy when compared to state-of-the-art methods. The code of the proposed approach is available at https://github.com/Gabriellm2003/remote_sensing_missing_data.

INDEX TERMS Remote Sensing, Image Classification, Multi-Modal Machine Learning, Metric Learning, Cross-View Matching, Multi-view Missing Data Completion

I. INTRODUCTION

Standard image classification tasks are trained by using a single data point as input. However, in some cases, using only one input image is not enough to allow its categorization. One reason for this is the perspective of the object presented in the image, which may not have enough information to allow its identification. For example, aerial images allow us to observe scenes from above, providing information about the general shape and structure of the objects and facilitating the classification of some entities such as bridges and streets. On the other hand, ground images give us a closer and frontal view of the object, providing information about fine details and helping the recognition of, for instance, statues and specific facade buildings (such as hospitals).

In this context, researchers [1], [2] noticed that it would be essential to exploit the complementary information extracted from images depicting the same object from multiple perspectives (or views) in order to enhance the general scene understanding and, consequently, increase the performance. This important task, commonly called multi-view image classification, has been successfully explored for distinct applications, including geo-localization [3], mammography analysis [4], and land use mapping [5]. However, although essential and impactful, such task has a major challenge: missing data. When working with multi-view data, it is really common to have one or more views missing due to malfunction of a sensor, noise, or simply lack of data. This is even worse when considering that multi-view samples with missing data are often discarded entirely, resulting in a severe loss of available information, as presented in Figure 1. This issue is especially relevant for domains in which it is difficult to obtain annotated multi-view samples, such as the medical and remote sensing ones [1].

In this paper, we propose a novel framework for multi-view image classification capable of dealing with missing data using information extracted from real (non-synthetic) images. Technically, this method can be divided into two parts. In the first one, a retrieval network, trained using metric learning, receives an input instance and retrieves samples that can be used to fill its missing data gap. Then, in the second

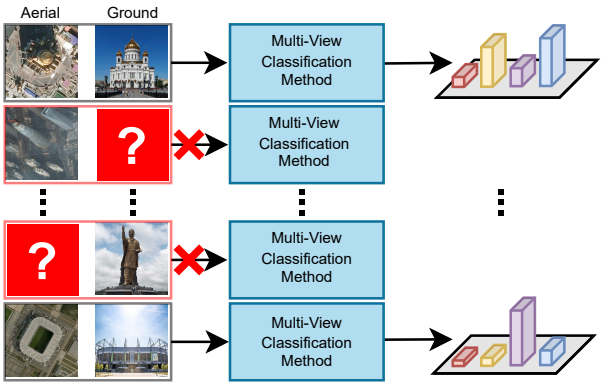


FIGURE 1. Example of the impact of missing views in the multi-view image classification. A missing view actually represents an instance that can not be used for training, *i.e.*, it represents a severe loss of available information that could be used for the learning process.

part, information extracted from both the input instance and the top-k retrieved images are further processed using state-of-the-art deep learning-based approaches and then late-fused using standard algorithms in order to perform the final classification. We evaluate the proposed framework on two 2-view (aerial and ground) datasets from the literature, achieving state-of-the-art results. Our methodology, however, can be easily expanded for more complex scenarios with more than two views.

In practice, the main contributions of this work are:

- A novel and versatile multi-view classification technique capable of efficiently handling missing data and that can be easily adapted to other applications with any number of views;
- A multi-view retrieval model that can, with a single training process, recover images (to fill the missing data gap) for all views, thus reducing the computational load.

The paper is structured as follows. Related works are presented in Section II while the proposed technique is introduced in Section III. Section IV presents the experimental protocol and Section V reports and discusses the obtained results. Finally, in Section VI we conclude the paper and point at promising directions for future work.

II. RELATED WORK

Although several methods have been proposed to tackle multi-view image classification [1], [6], only a few works have investigated and conceived approaches to handle missing data [5], [7], [8], [9], [10], [11].

Zhang *et al.* [7] proposed a feature-level completion method for missing view of multi-view data. Technically, their approach first linearly maps multi-view data to a feature-isomorphic subspace, unfolding the shared information from different views. Then, features of this isomorphic space are used to train a model, which is responsible for retrieving features to represent the missing view and, consequently, completing the multi-view data. In [9], the authors proposed a multi-modal classification framework, called EmbraceNet, that is robust to missing data. This framework uses a multinomial distribution to select the most relevant features of each view. In order to make the model robust to the partial absence of data, they readjust this multinomial distribution to select features only of the existing modalities. By doing this, they argue that the missing information due to data loss of a modality can be covered by the other modalities. Finally, Srivastava *et al.* [5] proposed a two-stream network that extracts discriminative features from aerial and ground images and then combines them for the final classification. In order to make their approach robust to missing data, they used these discriminative features to create an embedding space (using Canonical Correlation Analysis (CCA) [12]), which is exploited to retrieve samples that would be similar to the missing data and thus complete the data.

More recently, Generative Adversarial Networks (GANs) [13] have gained popularity in the missing data completion field, due to their ability to generate synthetic samples. Although there are several works [8], [10], [11] handling multi-view missing data completion using GANs, those are not thoroughly discussed here given that, as introduced, they are outside of the scope of this work, which focuses on information extracted from real (non-synthetic) images. The main reason for this is the fact that training models to generate good quality synthetic data is troublesome mainly when having a small amount of data, a common case in multi-view scenarios.

Considering this, in this work, we propose a multi-view image classification framework that uses a deep network, trained using metric learning/cross-view matching, to retrieve potential images that can be used to fill the missing data gap of input multi-view instances and, consequently, improve the general classification performance. Our framework is capable of recovering samples from a different domain just by using the data from the available view. Several differences may be pointed out between the proposed approach and the aforementioned works:

1) Instead of handling missing data at the feature level [7], [9], the proposed approach deals with such an important issue at the input level, allowing a better understanding and (a certain level of) explainability of the final results.
2) The proposed technique learns the embedding space using deep metric learning, which allows the model to capture specific features that optimize the class distribution in such space (i.e., end-to-end learning), thus being completely different from other works, such as Srivastava *et al.* [5], which use classification features extracted from an independent model to optimize the embedding space, usually using an external technique, such as CCA [12].

III. METHODOLOGY

The proposed multi-view data classification approach, presented in Figure 2, can be split into two parts. In the first one

(the retrieval part), a multi-view sample with missing data is processed using a retrieval network responsible for ranking the images (of an **auxiliary** database composed of scenes of several classes but from the same domain of the missing data) based on their potential to be paired with this input data. The original image and the generated ranking are then forwarded to the second part, i.e., the multi-view classification. In this step, retrieved images and original input are processed using their corresponding networks and then fused to produce the final classification. Observe that instead of pairing the input image with only the best retrieved image, we select the top-k images in order to alleviate any potential bias and improve generalization.

More details on each of those components, i.e., the retrieval and the classification parts, are presented in the next sections.

A. RETRIEVAL

As introduced, the retrieval network is responsible for recovering images that could potentially be used as pair for an input example with missing data.

Technically, during the training, such a model receives, as input, multi-view image pairs, and is optimized using the weighted soft-margin triplet loss [14], in which the main objective is to pull these input pairs (commonly called anchor and positive samples) close together whereas pushing dissimilar data (i.e., negative examples) apart. Based on previous works [15], [16], [17], the negative instances are mined from the batch using the exhaustive mini-batch strategy [18], which proposes to use all other samples (excluding the current anchor and positive pairs) as negative examples. Due to this, each batch must have only one image pair per class, in order to ensure that samples from the same class will never be used as negative examples [19].

Formally, suppose the model receives, as input, a batch $B = \{b^1, b^2, ..., b^C\}$ composed of C elements, one of each class, and in which each element is actually a multi-view image pair $b^i = \{I^i_{v_1}, I^i_{v_2}\}$ (i.e., 2 views). The network first extracts the features for each image pair $F^i = \{F^i_{v_1}, F^i_{v_2}\}$ and then uses these features to create a matrix of distances α between all images, as presented in Equation 1.

$$\alpha^{i,j} = 2 \times (1 - F^i_{v_1} F^{j\intercal}_{v_2}) \quad \forall\, i, j \in C \tag{1}$$

Given the matrix of distances α, it is possible to easily calculate the distance between anchors and positive samples ($d_{ap} = \alpha^{i,i}$), and between anchors and negative examples ($d_{an} = \alpha^{i,j} \;\; \forall\, i \neq j$), and then use those to optimize the model following the aforementioned weighted soft-margin triplet loss [14]:

$$\mathcal{L}_{weighted} = \ln 1 + \epsilon^{\gamma(d_{ap} - d_{neg})} \tag{2}$$

where γ is a hyper-parameter that controls the loss convergence [14].

It is important to highlight two main aspects of the training procedure: (i) instead of using the exact pair of images as input, random pairs within the same class are employed in order to increase the robustness of the model to missing data (given that, during the testing phase, there will be no corresponding pair to the query image due to missing data); (ii) for each input image pair, the model is optimized considering one as an anchor and the other as a positive sample and also vice-versa. This, which can be easily done using the matrix of distances α, allows the model to be prepared for missing data from all possible domains without increasing the computational load, as aforementioned.

During the inference, a multi-view sample with missing data (also called query) is paired with other images (from the previously established auxiliary database) that could potentially fill this missing data gap. All pairs are then processed by the retrieval network and ranked based on their similarity (i.e., on their euclidean distance). All images and the ranking are then forwarded to the classification network to be further processed, as explained in the next Section.

B. CLASSIFICATION

In the second step of the proposed methodology, the input image with missing data is finally classified. Towards this, first, the top-k most similar images are selected based on the aforementioned ranking. Such images are employed to fill the missing data gap of the original query image. The idea of using the top-k images is to alleviate any potential bias and improve generalization, given that using only one could produce under-representative multi-view pairs whereas using all images could generate instances composed of scenes from different classes.

Query and top-k images are then processed using networks trained specifically for their domains. Predictions σ for the top-k images are fused using the mean operation [20]:

$$\sigma_{f_2} = \frac{\sum_{i=1}^{k} \sigma_i}{k} \tag{3}$$

Finally, the predictions for the query image (σ_{f_1}) and the merged predictions of the top-k scenes (σ_{f_2}) are late-fused using the product operation [20], thus producing the final classification:

$$\hat{y} = \arg\max \prod_{i=1}^{2} \sigma_{f_i} \tag{4}$$

Observe that both mean and product fusion strategies were selected based on previous works [21], [20].

C. IMPLEMENTATION DETAILS

The architecture of the retrieval network consists of a two-stream encoder, in which each encoder is actually a pre-trained SwAV [22] model with ResNet-50 [23]. Precisely, this architecture has no classification layers and is trained using the aforesaid weighted soft-margin triplet loss [14].

For the inference, instead of using the retrieval network to extract features for all database images for every input query (a costly and time-consuming process), we extract, save, and

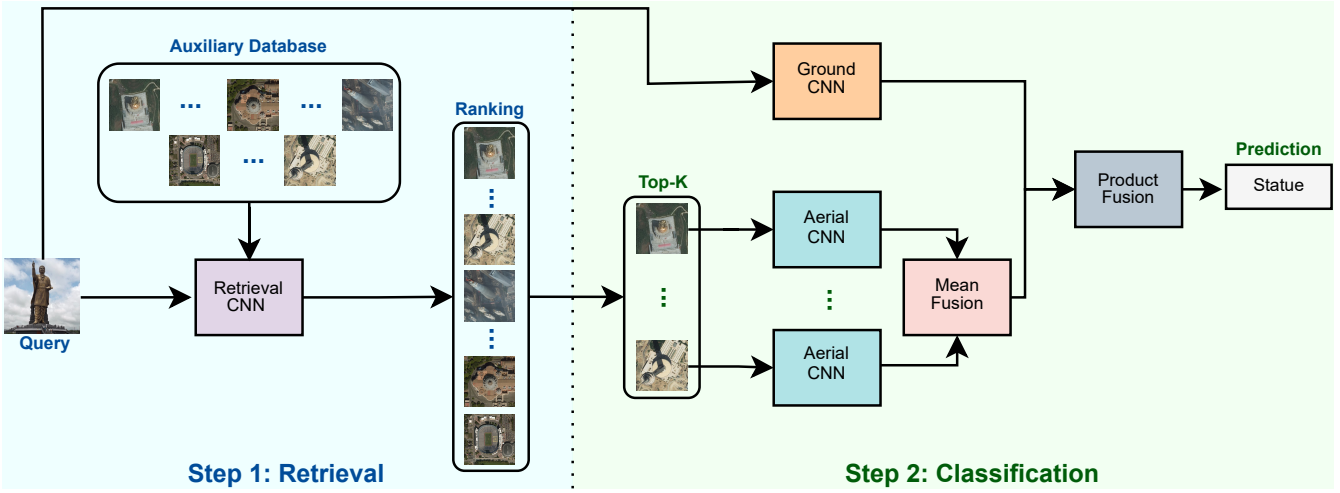


FIGURE 2. Pipeline of the proposed approach applied in a two-view scenario, where aerial images are missing. First, a retrieval network is employed to rank images (of an auxiliary database composed of scenes from the same domain of the missing data) based on their potential to be paired with the input sample with missing data (**Step 1**). Then, the top-k retrieved images and the input data are processed and fused, generating the final classification (**Step 2**). Following a similar logic, this methodology can be applied to different scenarios.

reuse the features of those database images in order to speed up the testing phase.

For the classification part, we evaluated different Convolutional Networks (VGG [24], DenseNet [25], and SKNet [26]). More details about this, as well as about the employed hyper-parameters, can be seen in Section IV-B.

D. GENERALIZATION

Although the methodology was described considering input samples composed of 2 views (v_1 and v_2), it can be easily adapted to scenarios with N views (with one of them missing).

Precisely, for the retrieval part, two changes would be required: (i) instead of using the features of the views to create the matrix of distances α (Equation 1), one would use the average of the features (from the non-missing views) to compute the distance against samples from the missing domain; and (ii) instead of using a two-stream network, an N-stream one would be required in order to support N-view inputs.

As for the classification part, two simple changes would be necessary: (i) a network for each view would need to be trained in order to extract features for that domain; and (ii) Equation 4 must be adjusted to late fuse the prediction scores of all N views (instead of 2).

With those small changes, it is possible to generalize the proposed approach to datasets composed of instances with N-views.

IV. EXPERIMENTAL SETUP

In this section, we describe the experimental setup used for the experiments. Section IV-A presents the datasets whereas Section IV-B describes the experimental protocol. Finally, baselines are described in Section IV-C.

A. DATASETS

Two multi-view datasets with very distinct properties were used for the experiments in order to better evaluate the effectiveness of the proposed approach. The first dataset, called AiRound [20], is composed of 11,753 images divided into 11 classes: airport, bridge, church, forest, lake, park, river, skyscraper, stadium, statue, and tower. Some examples of these classes are presented in Figure 3. The second one, named CV-BrCT [20], comprises approximately 24k pairs of images unevenly split into 7 urban classes: apartment, house, industrial, parking lot, religious, school, store. Samples of these classes are presented in Figure 4.

For both datasets, each multi-view sample is composed of a ground and an aerial perspective. Images for the former domain were collected from different sources (such as Google Images, Google Places, Google Street View) and have varying resolutions whereas scenes for the latter perspective were collected using Google Maps and have a fixed resolution of 500×500 pixels.

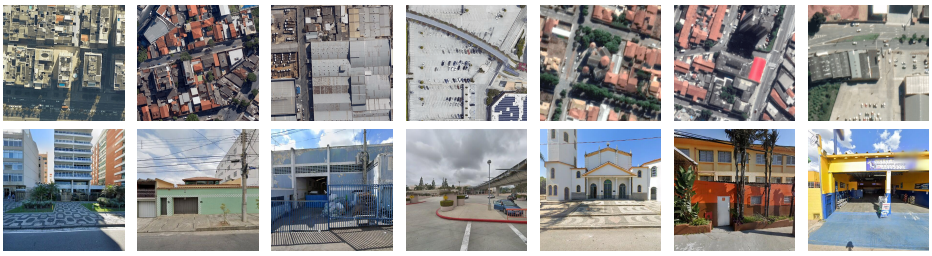
B. EXPERIMENTAL PROTOCOL

For both datasets, we employed a 5-fold cross-validation protocol, in which 80% of the images are used for training, 10% for validation, and the remaining 10% for testing. Following this protocol, we simulated and evaluated two different missing data scenarios using the test set: one for aerial and one for ground.

Considering this, the retrieval network (described in Section III-C) was trained using the following hyper-parameters: 200 epochs, batch size equal to the number of classes of the dataset (i.e., 11 for AiRound and 7 for CV-BrCT), γ (Equation 2) of 10, Adam [27] as optimizer, learning rate of 0.00001, and exponential decays of 0.9 and 0.999. Results related to this model are reported in terms of the average

(a) Airport (b) Bridge (c) Church (d) Forest (e) Lake (f) Park (g) River (h) Skyscr. (i) Stadium (j) Statue (k) Tower

FIGURE 3. Examples of the AiRound Dataset.



(a) Apartment (b) House (c) Industrial (d) Parking (e) Religious (f) School (g) Store

FIGURE 4. Examples of the CV-BrCT Dataset.

mean Average Precision at K (mAP@K) [28], taken from all 5-fold experiments with its corresponding standard deviation. Additionally, it is important to highlight that, during the inference, this network used the validation set as the retrieval database, i.e., images of the validation set are retrieved and then paired (based on their similarity) with the test samples for the final classification.

For the classification part, we evaluated three well-known architectures: VGG [24], DenseNet [25], and SKNet [26]. All networks were fine-tuned (from the ImageNet [29] dataset) for each of the domains (aerial and ground) using the following hyper-parameters: 200 epochs, early stop with 20 epochs, batch size of 32, stochastic gradient descent as optimizer, a learning rate of 0.001, and momentum of 0.9. In this case, all obtained results are reported in terms of the average F1-Score and standard deviation among all 5 folds.

C. BASELINES

Four techniques were considered as baselines for both datasets. The first baseline, referenced as "No Fusion", consists of using a single-view classification CNN evaluated in the data available in the test phase (without any completion). The idea of this baseline is to establish a lower bound for the other experiments. The second baseline, referenced hereafter as "Fully-Paired", performs final classification assuming that the test set has no missing data (i.e., that it is fully paired). To do so, this baseline uses two classification CNNs to extract features from both aerial and ground domains which are then fused (using the product fusion presented in Equation 4) to produce the final result. As for the second baseline, the idea is to set an upper bound for further experiments.

The remaining baselines come from the literature. One of those is the EmbraceNet [9], a multi-view framework that learns a multimodal distribution to select the most relevant features of each view. This distribution can be adjusted depending on the availability or absence of certain domains, thus being able to handle missing data. Based on such framework, we proposed a two-stream network in which the final layer is actually an EmbraceNet layer [9], capable of efficiently dealing with missing data. Such network was trained using the same set of hyper-parameters used for the classification models (Section IV-B).

The last baseline is the Canonical Correlation Analysis (CCA) [5]. Such an approach learns matrices that project the features from different views into a common latent space, bringing closer the varying perspectives of the same object [5]. Using such projection matrices (learned from the training data) we can project the available view of the test set into the common latent space and then retrieve the closest scene to finally fill the missing data gap. Following the guidelines introduced in the original work [5], we first project the features extracted from the last layer before the classification using a Principal Component Analysis (PCA), and then use them as input for the CCA [5].

V. RESULTS

In this section, we present and discuss the obtained results. Section V-A presents the results related to the retrieval part whereas Section V-B evaluates and compares different configurations for the classification part of the proposed frame-

work. Finally, Section V-C compares the proposed framework with state-of-the-art baselines.

A. RETRIEVAL ANALYSIS

In this Section, we analyze the retrieval network, a main component of the proposed framework. Precisely, Figure 5 reports the retrieval results, in terms of mAP@K [28], for both AiRound and CV-BrCT datasets.

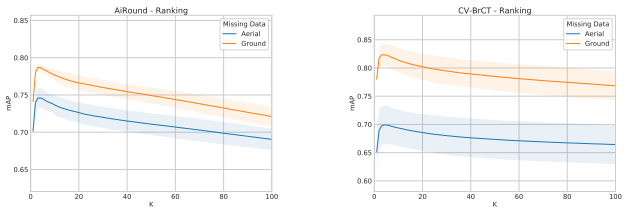


FIGURE 5. Results, in terms of mean Average Precision (mAP) and ranking size (K), of the retrieval network for the aerial and ground missing data scenarios in AiRound and CV-BrCT datasets. The shaded areas represent the standard deviation across the folds.

Analyzing the results, it is possible to observe that the retrieval network tends to produce better outcomes when using aerial data as input query (i.e., when the ground view is missing). This may be justified by the fact that most aerial images tend to provide more context information than ground scenes to the retrieval model, that in turn exploits such useful data to recover similar images. Another important aspect to discuss is the fact that, in general, the top-1 image does not produce the best results in terms of mAP. This corroborates with our initial analysis about the potential of using top-k (instead of top-1) images to fill the missing data gap and, consequently, produce better classification results. A better discussion about the use of top-k images to fill this missing data gap is presented in the next Section.

B. CLASSIFICATION ANALYSIS

In this Section, we analyze the impact of the different classification networks (VGG [24], DenseNet [25], and SKNet [26]) and investigate the influence of the ranking size (i.e., top-k) in the final performance of the framework considering two distinct missing data scenarios (described in Section IV-B), i.e., one for aerial and one for ground.

Obtained results for the AiRound and CV-BrCT datasets are presented in Tables 1 and 2, respectively. For both datasets and assessed scenarios, the strategy of combining information from multiple images to fill the missing data gap produced better outcomes than using just the top retrieved image (an approach commonly exploited in the literature [7], [5]). Again, this directly corroborates with our initial analysis about the potential of combining information from the top-k images to fill the missing data gap.

Aside from this, for the AiRound dataset, the best result for the aerial missing scenario was produced by the SKNet model [26] whereas the best outcome for the ground missing scenario was yielded by the DenseNet network [25]. In both cases, the ranking size of 100 (i.e., the top-100 images) yielded the best outcomes. For the CV-BrCT dataset, all assessed models achieved very similar results when using a ranking size greater than or equal to 3 for the aerial missing scenario, and greater than or equal to 2 for the ground missing scenario. Given this, the simplest model, i.e., VGG network [24], with ranking size 100 was selected and used for further experiments using this dataset.

TABLE 1. Obtained results achieved by the proposed method for the AiRound dataset.

Ranking Size	Available Data	VGG [24]	DenseNet [25]	SKNet [26]
Top 1	Ground (Aerial Missing)	0.74 ± 0.01	0.75 ± 0.02	0.75 ± 0.02
Top 2		0.76 ± 0.01	0.76 ± 0.02	0.77 ± 0.02
Top 3		0.76 ± 0.01	0.77 ± 0.03	0.77 ± 0.02
Top 4		0.77 ± 0.01	0.77 ± 0.02	0.78 ± 0.01
Top 5		0.77 ± 0.01	0.77 ± 0.02	0.78 ± 0.01
Top 10		0.77 ± 0.02	0.77 ± 0.02	0.78 ± 0.01
Top 50		0.77 ± 0.02	0.77 ± 0.02	0.78 ± 0.00
Top 100		0.77 ± 0.02	0.77 ± 0.02	$\mathbf{0.79 \pm 0.00}$
Top 1	Aerial (Ground Missing)	0.82 ± 0.01	0.83 ± 0.01	0.83 ± 0.01
Top 2		0.83 ± 0.01	0.84 ± 0.02	0.83 ± 0.01
Top 3		0.83 ± 0.01	0.84 ± 0.02	0.84 ± 0.01
Top 4		0.83 ± 0.01	0.84 ± 0.02	0.84 ± 0.01
Top 5		0.83 ± 0.01	0.84 ± 0.02	0.84 ± 0.01
Top 10		0.83 ± 0.01	0.85 ± 0.02	0.84 ± 0.01
Top 50		0.83 ± 0.00	0.84 ± 0.02	0.84 ± 0.01
Top 100		0.83 ± 0.01	$\mathbf{0.85 \pm 0.01}$	0.84 ± 0.01

TABLE 2. Obtained results achieved by the proposed method for the CV-BrCT dataset.

Ranking Size	Available Data	VGG [24]	DenseNet [25]	SKNet [26]
Top 1	Ground (Aerial Missing)	0.70 ± 0.02	0.70 ± 0.03	0.70 ± 0.02
Top 2		0.72 ± 0.02	0.72 ± 0.02	0.72 ± 0.02
Top 3		0.72 ± 0.02	$\mathbf{0.73 \pm 0.02}$	0.72 ± 0.02
Top 4		$\mathbf{0.73 \pm 0.02}$	$\mathbf{0.73 \pm 0.02}$	$\mathbf{0.73 \pm 0.03}$
Top 5		$\mathbf{0.73 \pm 0.02}$	$\mathbf{0.73 \pm 0.02}$	$\mathbf{0.73 \pm 0.02}$
Top 10		$\mathbf{0.73 \pm 0.02}$	$\mathbf{0.73 \pm 0.01}$	$\mathbf{0.73 \pm 0.02}$
Top 50		$\mathbf{0.73 \pm 0.02}$	$\mathbf{0.73 \pm 0.02}$	$\mathbf{0.73 \pm 0.03}$
Top 100		$\mathbf{0.73 \pm 0.02}$	$\mathbf{0.73 \pm 0.02}$	$\mathbf{0.73 \pm 0.02}$
Top 1	Aerial (Ground Missing)	0.87 ± 0.02	0.87 ± 0.02	0.87 ± 0.02
Top 2		0.87 ± 0.01	$\mathbf{0.88 \pm 0.02}$	0.87 ± 0.02
Top 3		$\mathbf{0.88 \pm 0.01}$	$\mathbf{0.88 \pm 0.01}$	0.87 ± 0.02
Top 4		$\mathbf{0.88 \pm 0.01}$	$\mathbf{0.88 \pm 0.01}$	0.87 ± 0.02
Top 5		$\mathbf{0.88 \pm 0.01}$	$\mathbf{0.88 \pm 0.01}$	0.87 ± 0.02
Top 10		$\mathbf{0.88 \pm 0.01}$	$\mathbf{0.88 \pm 0.02}$	$\mathbf{0.88 \pm 0.01}$
Top 50		$\mathbf{0.88 \pm 0.01}$	$\mathbf{0.88 \pm 0.02}$	$\mathbf{0.88 \pm 0.01}$
Top 100		$\mathbf{0.88 \pm 0.01}$	$\mathbf{0.88 \pm 0.02}$	$\mathbf{0.88 \pm 0.01}$

C. STATE-OF-THE-ART COMPARISON

This Section compares and discusses the results obtained by the proposed framework and the state-of-the-art baselines. Observe that: (i) based on the analyses carried out in the previous Sections, only the best results for each dataset and scenario are presented. (ii) the baselines were conceived using the same network selected for the classification part of the proposed framework. Precisely, for the AiRound dataset, baselines are based on the SKNet model [26] (for the aerial missing scenario) and on the DenseNet [25] (for the ground missing scenario). For the CV-BrCT, all baselines were conceived based on the VGG network [24]. (iii) all results presented here were verified using a fold-by-fold paired t-test with confidence level of 95%.

TABLE 3. Results of the proposed method and baselines for AiRound dataset.

Method	Available Data	F1 Score
No Fusion (Lower bound)	Ground (Aerial Missing)	0.76 ± 0.00
EmbraceNet [9]		0.69 ± 0.03
CCA [5]		0.75 ± 0.02
Ours		$\mathbf{0.79 \pm 0.00}$
No Fusion (Lower bound)	Aerial (Ground Missing)	0.83 ± 0.00
EmbraceNet [9]		0.83 ± 0.00
CCA [5]		0.83 ± 0.01
Ours		$\mathbf{0.85 \pm 0.01}$
Fully-Paired (Upper bound)	Aerial + Ground	0.91 ± 0.01

Tables 3 and 4 present the results for the AiRound and CV-BrCT dataset, respectively. Overall, the proposed approach outperformed all baselines except, as expected, the upper bound one. In fact, for the CV-BrCT dataset, the difference between the results achieved by the proposed framework and the fully-paired (upper bound) baseline for the ground missing scenario is almost irrelevant. However, for all other scenarios and datasets, this difference is considerable, which shows that there is still room for improvements.

TABLE 4. Results of the proposed method and baselines for CV-BrCT dataset.

Method	Available Data	F1 Score
No Fusion (Lower bound)	Ground (Aerial Missing)	0.72 ± 0.01
EmbraceNet [9]		0.53 ± 0.04
CCA [5]		0.72 ± 0.02
Ours		$\mathbf{0.73 \pm 0.02}$
No Fusion (Lower bound)	Aerial (Ground Missing)	0.86 ± 0.01
EmbraceNet [9]		0.77 ± 0.03
CCA [5]		0.86 ± 0.01
Ours		$\mathbf{0.88 \pm 0.01}$
Fully-Paired (Upper bound)	Aerial + Ground	0.89 ± 0.02

Aside from this, it is interesting to observe that, for both datasets, the obtained results for the aerial missing scenario are worse than the outcomes for the ground missing scenario. As previously explained, this may be justified by the fact that most aerial images tend to provide more context information (than ground scenes), thus assisting in the classification process and, consequently, yielding better results.

VI. CONCLUSIONS

In this paper, we propose a novel framework to handle multi-view image classification with missing data. The proposed approach is composed of two main parts: (i) a retrieval one, responsible for recovering similar samples that can be used to fill the missing data gap of the input query image; and (ii) a classification one, that fuses information extracted from both the input image and the top-k retrieved scenes in order to perform the final classification.

Experiments were conducted using two multi-view aerial-ground datasets (AiRound and CV-BrCT) and considering two distinct scenarios: one in which aerial images are considered absent and another where ground scenes are considered missing. Results have showed that the proposed technique is efficient and robust. Precisely, it achieved state-of-the-art results in both datasets outperforming all baselines, except for the fully paired (upper-bound) one. Additionally, experimental outcomes showed that the strategy of combining information from multiple (i.e., top-k) images to fill the missing data gap (instead of using just the top retrieved image, as commonly employed in the literature [7], [5]) is remarkably effective.

As future work, we intend to evaluate the proposed approach using other datasets with more (than two) views per instance. We also would like to assess other retrieval methods as well as other classification networks.

REFERENCES

[1] M. Seeland and P. Mäder, "Multi-view classification with convolutional neural networks," Plos one, vol. 16, no. 1, p. e0245230, 2021.
[2] Y. Li, M. Yang, and Z. Zhang, "A survey of multi-view representation learning," IEEE transactions on knowledge and data engineering, vol. 31, no. 10, pp. 1863–1883, 2018.
[3] 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, 2019.
[4] G. Carneiro, J. Nascimento, and A. P. Bradley, "Unregistered multiview mammogram analysis with pre-trained deep learning models," in International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 652–660, Springer, 2015.
[5] 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.
[6] H. Zhang, H. Xu, X. Tian, J. Jiang, and J. Ma, "Image fusion meets deep learning: A survey and perspective," Information Fusion, vol. 76, pp. 323–336, 2021.
[7] 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.
[8] 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, pp. 1158–1166, 2018.
[9] J.-H. Choi and J.-S. Lee, "Embracenet: A robust deep learning architecture for multimodal classification," Information Fusion, vol. 51, pp. 259–270, 2020.
[10] D. Lee, J. Kim, W.-J. Moon, and J. C. Ye, "Collagan: Collaborative gan for missing image data imputation," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2487–2496, 2019.
[11] G. Aversano, M. Jarraya, M. Marwani, I. Lahouli, and S. Skhiri, "Mic: Multi-view image classifier using generative adversarial networks for missing data imputation," in 2021 18th International Multi-Conference on Systems, Signals & Devices (SSD), pp. 283–288, IEEE, 2021.
[12] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," Neural computation, vol. 16, no. 12, pp. 2639–2664, 2004.
[13] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," Advances in neural information processing systems, vol. 27, 2014.
[14] B. Yu, T. Liu, M. Gong, C. Ding, and D. Tao, "Correcting the triplet selection bias for triplet loss," in Proceedings of the European Conference on Computer Vision (ECCV), pp. 71–87, 2018.
[15] S. Hu, M. Feng, R. M. Nguyen, and G. H. Lee, "Cvm-net: Cross-view matching network for image-based ground-to-aerial geo-localization," in IEEE/CVF Computer Vision and Pattern Recognition, pp. 7258–7267, 2018.
[16] Y. Shi, L. Liu, X. Yu, and H. Li, "Spatial-aware feature aggregation for image based cross-view geo-localization," Advances in Neural Information Processing Systems, vol. 32, 2019.

[17] Y. Shi, X. Yu, D. Campbell, and H. Li, "Where am i looking at? joint location and orientation estimation by cross-view matching," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4064–4072, 2020.
[18] N. N. Vo and J. Hays, "Localizing and orienting street views using overhead imagery," in European Conference on Computer Vision, pp. 494–509, Springer, 2016.
[19] 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, pp. 35–44, 2018.
[20] G. Machado, E. Ferreira, K. Nogueira, H. Oliveira, M. Brito, P. H. T. Gama, and J. A. dos 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, 2020.
[21] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 41, no. 2, 2018.
[22] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, "Unsupervised learning of visual features by contrasting cluster assignments," Advances in Neural Information Processing Systems, vol. 33, pp. 9912–9924, 2020.
[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in IEEE/CVF Computer Vision and Pattern Recognition, pp. 770–778, 2016.
[24] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.
[25] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in IEEE/CVF Computer Vision and Pattern Recognition, pp. 4700–4708, 2017.
[26] X. Li, W. Wang, X. Hu, and J. Yang, "Selective kernel networks," in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 510–519, 2019.
[27] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv preprint arXiv:1412.6980, 2014.
[28] D. Valcarce, A. Bellogín, J. Parapar, and P. Castells, "Assessing ranking metrics in top-n recommendation," Information Retrieval Journal, vol. 23, no. 4, pp. 411–448, 2020.
[29] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in IEEE/CVF Computer Vision and Pattern Recognition, pp. 248–255, Ieee, 2009.

GABRIEL MACHADO is a computer vision scientist. In 2021, he received the MSc degree in Computer Science from Universidade Federal de Minas Gerais, Brazil. In 2018, he granted by the same institution his B.Sc. degree in Computer Science. His research interests include Deep and Machine Learning, Remote Sensing, Pattern Recognition, and Computer Vision.

MATHEUS B. PEREIRA is currently pursuing the PhD degree at the Department of Computer Science from Universidade Federal de Minas Gerais, Brazil, where he also received the MSc. degree in 2019. He received the Bsc. degree in Computer Science from Universidade Federal de Lavras, Brazil. His research interests include Machine Learning, Computer Vision, Remote Sensing, Open Long-Tailed Recognition and Semantic Segmentation.

KEILLER NOGUEIRA is a lecturer in the Division of Computing Science and Mathematics at the University of Stirling, UK. He received the MSc and PhD degrees in Computer Science from the Universidade Federal de Minas Gerais, Brazil, in 2015 and 2019, respectively. He has got the BSc degree in Computer Science from Universidade Federal de Viçosa (Brazil) in 2012. Keiller has published several high-quality articles in leading journals and conferences. His research interests include Deep and Machine Learning, Pattern Recognition, Image Processing, Computer Vision, and Remote Sensing.

JEFERSSON ALEX DOS SANTOS has got a Ph.D in Computer Science from Université de Cergy-Pontoise (France) and from University of Campinas (Brazil) in 2013. Currently, he is an Associate Professor in the Department of Computer Science at the Universidade Federal de Minas Gerais, Brazil. He has a Research Productivity scholarship from the Brazilian Research Council (CNPq) since 2016. Jefersson has published several articles in journals with high impact factor and selective editorial policy. He has also published more than thirty articles in important conferences of remote sensing, image processing and computer vision areas. Jefersson has experience in coordinating research with Brazilian funding agencies and R&D projects with companies in those topics. Jefersson is founder and coordinator of the Laboratory of Pattern Recognition and Earth Observation (PATREO - www.patreo.dcc.ufmg.br), one of Brazil's pioneer groups focused on the development of computer vision and machine learning for remote sensing applications.