Common Practices and Taxonomy in Deep Multiview Fusion for Remote Sensing Applications

Francisco Mena, *Graduate Student Member, IEEE*, Diego Arenas, Marlon Nuske, and Andreas Dengel

***Abstract*—The advances in remote sensing technologies have boosted applications for Earth observation. These technologies provide multiple observations or views with different levels of information. They might contain static or temporary views with different levels of resolution, in addition to having different types and amounts of noise due to sensor calibration or deterioration. A great variety of deep learning models have been applied to fuse the information from these multiple views, known as deep multiview (MV) or multimodal fusion learning. However, the approaches in the literature vary greatly since different terminology is used to refer to similar concepts or different illustrations are given to similar techniques. This article gathers works on MV fusion for Earth observation by focusing on the common practices and approaches used in the literature. We summarize and structure insights from several different publications concentrating on unifying points and ideas. In this manuscript, we provide a harmonized terminology while at the same time mentioning the various alternative terms that are used in literature. The topics covered by the works reviewed focus on supervised learning with the use of neural network models. We hope this review, with a long list of recent references, can support future research and lead to a unified advance in the area.**

***Index Terms*—Data fusion, deep learning, multimodal learning, multiview (MV) learning, remote sensing (RS), supervised learning.**

Nomenclature

CNN	Convolutional neural network.
DSM	Digital surface model.
EO	Earth observation.
GRU	Gated recurrent unit.
HS	Hyperspectral.
LiDAR	Light detection and ranging.
LSTM	Long-short term memory.
LULC	Land-use land-cover.
L8	Landsat-8.
MS	Multispectral.
MLP	Multilayer Perceptron.
MV	Multiview.
NIR	Near infra-red.
NDVI	Normalized difference vegetation index.
NN	Neural network.
RS	Remote sensing.
RNN	Recurrent neural network.
SAR	Synthetic aperture radar.
S1	Sentinel-1.
S2	Sentinel-2.
UAV	Unmanned aerial vehicles.

I. Introduction

Earth observation (EO) allows the study and analysis of different aspects of human life and natural resources, where RS technologies are a crucial factor in providing a global perspective on the Earth. The final purpose is to make better data-informed decisions based on the current and future state of the planet. Many applications in this context express phenomena or objects that could be represented by multiple observations. For instance, the classification of a crop-type using observations from multiple satellites [1], the estimation of agricultural yield using ground-based weather and RS-based optical information [2], or the estimation of evapotranspiration (water evaporation into the atmosphere) based on meteorological factors [3]. Deep neural networks have been successfully applied to different areas of EO [4] for their capacity to learn complex nonlinear functions and heterogeneous patterns. In a comprehensive learning scenario, the object of interest can be represented by multiple views, making it necessary to suggest the appropriate approaches to combine diverse types of information.

Combining multiple views presents numerous challenges. For example, sensors have four types of resolution: spectral, spatial, temporal, and radiometric, which need to be considered when combining the views. In addition, different data sources use different sensors and data collection, introducing different levels of noise to the data [5]. Besides, machine learning models are prone to errors caused by inductive bias. Therefore, if we try to include too many views, the model may collapse due to overparameterization [5] or the curse of dimensionality [6], [7]. Then, the goal of the MV learning models is to extract the most valuable information for a predictive task from the available views. This manuscript focuses on deep learning models, addressing the challenges of model design. The types of questions that we are aiming to answer in this manuscript are: What are the modeling options? Which types of architecture and strategies of data fusion to use? What are the common approaches from the

Manuscript received 1 September 2023; revised 7 January 2024; accepted 28 January 2024. Date of publication 2 February 2024; date of current version 20 February 2024. The work of Francisco Mena was supported by the chair of Prof. A. Dengel with University of Kaiserslautern-Landau. *(Corresponding author: Francisco Mena.)*

Francisco Mena and Andreas Dengel are with the Computer Science Department, University of Kaiserslautern-Landau (RPTU), 67663 Kaiserslautern, Germany, and also with Smart Data and Knowledge Services, German Research Center for Artificial Intelligence (DFKI), 67663 Kaiserslautern, Germany (e-mail: f.menat@rptu.de; andreas.dengel@dfki.de).

Diego Arenas and Marlon Nuske are with Smart Data and Knowledge Services, German Research Center for Artificial Intelligence (DFKI), 67663 Kaiserslautern, Germany (e-mail: diego.arenas@dfki.de; marlon.nuske@dfki.de).

Digital Object Identifier 10.1109/JSTARS.2024.3361556

literature? We notice that employing advanced fusion strategies involving additional components and/or multiple fusion layers results in superior predictive performance. The relevance of this work is the examination of recent studies in a way to support ongoing research efforts and encourages researchers to embrace appropriate fusion approaches based on the available research within the field.

The rest of this article is organized as follows. In Section II, the conceptual framework is introduced. Section III, presents the challenges of the MV learning. In Section IV, we attempt to answer some main questions about fusion from the reviewed literature. In Section V, common approaches (characterized by the components used) are highlighted. The current resources and promising results for fusion in MV learning are described in Section VI. Finally, Section VII, open questions and Section VIII concludes this article.

A. Background Concepts

EO is the gathering and study of information about the biological, chemical, and physical systems of the planet Earth. RS involves observing objects from platforms that are distant from the object being observed, e.g., satellites, aerial images, or UAV. Sensors in RS have different types of resolution: spectral, as electromagnetic bands or channels obtained by different filters; spatial, as the area covered by each pixel based on the distance of the object to the observed area or the resolution of the instrument; and temporal, as the frequency with which an area is swept, and radiometric, affected by the sensor sensibility, calibration, and deterioration.

Regression, classification, and segmentation are common tasks explored in the literature. Regression tasks predict a continuous value from the input data, e.g., estimate the crop yield produced in a particular field during a growing season (agricultural yield prediction), estimate the amount of precipitation for the next days (precipitation forecasting), or estimate the snow depth. Classification tasks predict a label from a set of categories, e.g., identify a target object on the Earth's surface (automatic target recognition), identify the crop type growing in a field (vegetation recognition), or identify whether a field is irrigated or not (irrigation recognition). Segmentation tasks to assign a class in a mesh of a given region (with pixel-wise information), e.g., identify the type of use given to a piece of land (LULC mapping), identify which section of a region is flooded (flood mapping), or identify which pixels of an RS image are covered by clouds (cloud segmentation).

A single *view* is a data point, observation, information channel, or feature set associated with an object of interest that contains information about it (direct or indirect) [8], [9]. In the context of EO, optical images are the most common type of view, coming from passive RS, which measures the solar radiation reflection on the Earth's surface. Red–green–blue (RGB) bands, MS (with more bands on the spectrum than RGB), HS (with more than hundreds of bands), or panchromatic (PAN, a single band with a broad spectrum) are the most common options for optical views. The missions Sentinel-2 (S2), Landsat (L7 and L8), MODIS, and custom UAV are current common sources of optical views. Some studies have explored the use of spatial indexes derived from the optical views as input data [10], [11], [12], [13]. The NDVI and the enhanced vegetation index are the most widely used for agricultural purposes, while the normalized difference water index is most used in water-related applications. Other types of views come from active RS sources, sending electromagnetic pulses to the Earth's surface and recording the reflected energy. SAR that uses microwaves and LiDAR that uses infrared waves emitting pulses, are two commonly used active RS. Sentinel-1 (S1), Radarsat, or Envisat are example sources of the SAR view. Private UAVs can be used to obtain LiDAR data. With active RS, it is possible to generate a digital elevation map/model representing the topographic surface of the Earth. There are two types of these models: DSMs and digital terrain models (DTMs). DSMs contain only the bare ground, while DTMs additionally contain objects such as vegetation and buildings. DSM is the most common model used in EO applications. Finally, meteorological variables can be represented in one or more views collected from RS or ground-based instruments. For example, temperature, precipitation, solar radiation, wind speed, humidity, and vapor pressure.

The MV learning scenario assumes that multiple views are (always) available for each object. We use the term "MV" as a general concept that includes the concepts of multimodal, multisource, or multisensor used in literature. MV does not constrain that views must be complementary, represent different physical quantities [14], or be from different data (e.g., images, signals, and metadata). For instance, MV data may contain multiband images [15], different groups of spectral bands, such as RGB and NIR [16], or additional views extracted from the optical view, such as NDVI [17], [18] or NN features [19].

In this article, we focus on *data fusion* within the context of MV learning. The term data fusion is commonly used in the context of database management, referring to integrating heterogeneous data sources into a single, consistent, and clean representation [20]. Image fusion can be understood as combining the geometric detail of multiband images to produce a final image [21], [22], e.g., spatio-temporal fusion [23], [24], [25] or spatio-temporal-spectral fusion [26]. However, data fusion may have different interpretations in MV learning, depending on the application. Therefore, we suggest the following interpretation: to integrate and merge the information in MV data with machine learning models to maximize the predictive performance on a given downstream task.

Deep learning refers to the use of NNs as machine learning models, usually fed with raw-format data, such as images or time series. These models could be trained in a supervised (e.g., with labels) or unsupervised (e.g., without labels) way. The unsupervised training is usually focused on pattern recognition or representation learning tasks such as clustering and dimensionality reduction, e.g., by using an autoencoder or self-supervision [27]. However, when training in a supervised fashion, it is usually for a *downstream task* (aka predictive task or learning task), which is some task that the model needs to learn to predict through minimizing a *predictive loss*. Thus, a *prediction head* (aka classifier) is used to give the final decision or prediction for the task based on the input data. However, when the models learn from raw data, usually an *encoder* model (aka backbone or extraction network) is used. These encoder models obtain a representation that compresses the most valuable information from the input, which feeds the prediction head. The standard NN architecture used as a prediction head is the MLP, which incorporates fully connected layers. While the standard NN architecture used as encoder of images or spatial information is CNNs, the one for sequential data is RNN with LSTM or GRU layers. We refer to the predictive quality of the supervised trained model as the predictive performance.

II. Literature Review

The primary purpose of data fusion in MV learning is to combine the information from different perspectives (views) to provide a broader understanding of the phenomena and improve the predictive performance of machine learning models [28]. However, sometimes the goal could be just to get an embedding to search for similar views, as is the case of MV alignment or representation learning [8], [29], [30]. This alignment is the base of contrastive learning [27], where a model is used to project the data into a shared subspace for each view. Nevertheless, this article only covers data fusion within the MV learning topic.

Given the MV nature of the EO data, primarily attributed to RS technologies, several fusion approaches have been proposed in the literature. While some approaches may share fundamental similarities, it is common to observe different use of terms, creating variations in their presentation. For instance, the S1 and S2 missions are mentioned as *multisensor* [31], [32], *multisource* [15], or *multimodal* [33], [34] in LULC applications. Another case is the term used when fusing intermediate representations extracted by NNs, such as *middle-level fusion* [35], *layer-level fusion* [1], or *late-level fusion* [36]. This manuscript provides a unified taxonomy and common practices from the literature, discussing the advantages, and limitations of different approaches.

Some studies have already offered valuable reviews of MV learning. Such as Sun et al. [37], focusing on unsupervised and semisupervised MV learning with a theoretical perspective. Later, Zhao et al. [38] provided a review on the same line with updated references, including open problems in the area. Lahat et al. [14] focused on multisensor, medical, and environmental applications. Recent reviews compared conventional with NN models for MV learning [39], [40]. Several public MV datasets were shared from these works, and most of them based on human and action recognition. In addition, some surveys focused on the EO domain. For instance, Gomez-Chova et al. [41] presented a review of RS image classification, while the authors in [6] and [21] focused on RS image fusion. Image fusion in the sense of a spatio-temporal fusion occurs before any learning occurs for a downstream task [23]. Salcedo-Sanz et al. [42] reviewed several data fusion methods in different applications of the EO domain. Recently, Li et al. [43] gathered open-source code and RS datasets for specific EO applications, focusing on data fusion and multimodal learning. They focused on specific models categorized into two main types of data fusion: homogeneous and heterogeneous. For our study, Fig. 1 summarizes the query-based search process. We initiated our article collection from the year 2014 onwards, aligning with the notable advances in CNN for image classification [44], [45]. This temporal choice was further substantiated by the observable surge in the number of research articles in 2014. Nonetheless, we have included a select few seminal articles predating 2014 as a reference to conventional approaches that predated the deep learning era. Our exhaustive search yielded a total of 160 articles related to MV learning in the context of RS image-based applications. These underwent a thorough review to ensure alignment with the specific MV learning scenario we delineate in this manuscript. It can be seen that the main sources are EO-related journals, such as *Remote Sensing*, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, and IEEE Transactions on Geoscience and Remote Sensing.

A. History of MV Learning

MV learning beginnings can be traced to 1936, to the mathematical statistician Hotelling and his correlation canonical analysis (CCA) proposal [8]. In this work, linear mapping is learned to maximize the correlation between two views. Kettering et al. [46] extended CCA to multiple feature sets in 1971. More recently, Andrew et al. [47] used NN models to learn nonlinear mappings and correlations in 2013. However, to the best of our knowledge, the first work mentioning the concept "view" providing the first theoretical foundations of MV learning in classification tasks is Blum et al. [9] in 1998, which was extended by Muslea et al. [48]. Furthermore, according to our research, the first mention of the "MV" concept in the EO domain is associated with an LULC application in 2015 [49], where different scales (zooms) of an image were used as MV data. The "MV" concept was also used in the more recent research of contrastive learning in applications with RS images [30], [50].

The use of deep MV learning in the EO domain has received considerable attention due to the recent advances in NN models and open science culture [22]. The EO community has been very active in generating open access code, benchmark datasets, or pretrained models. As in other areas of machine learning, the research started using machine learning models that learn from tabular data (e.g., metadata). To give an idea, linear models such as the perceptron was used in 1989 [51] followed by classical nonlinear models such as SVM in 2006–2007 [31], [52] for LULC, decision trees in 2006 [53], and MLP in 2008 [54]. Later, the community explored highly nonlinear functions through deep NNs. For example, using multiple CNN models for each view in LULC during 2016–2017 [17], [32], [55]. Camps-Valls et al. [22] recognized three phases of research in the evolution of the application of NN models for EO [aligning with the research trend we observe in Figs. 1 and 4(b)]. The early stage, starting in 2014, involved the exploration of different EO tasks using NN models; followed by the release of standard datasets for benchmarking in 2016, which enabled researchers to validate and compare the performance of different approaches; and in more recent years, there has been a shift toward EO-driven methodological research, with a focus not only on the applicability of NN models but also on other aspects, such as uncertainty

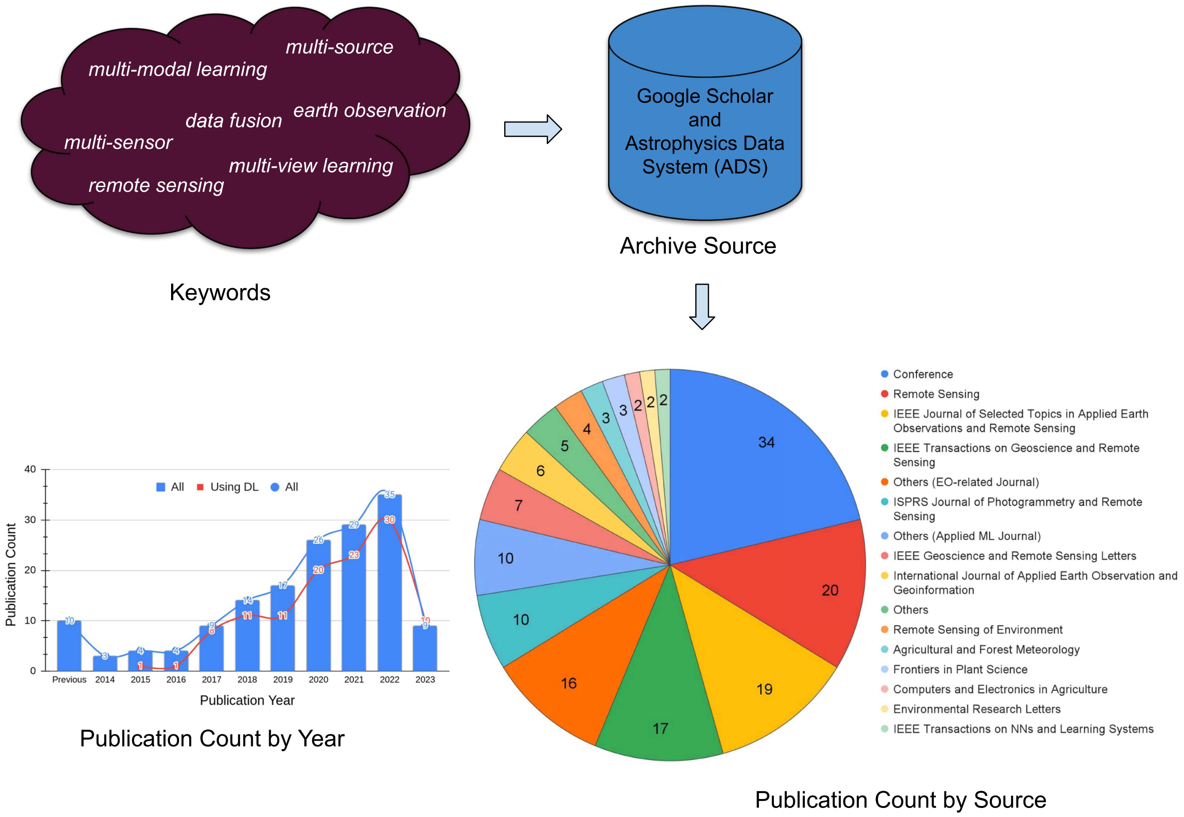


Fig. 1. Illustration of the articles search. Multiple keywords were used, yielding varying publications across EO journals, conferences, and machine learning journals. “Others” count the sources that have less than two articles reviewed. The articles were obtained until April 2023.

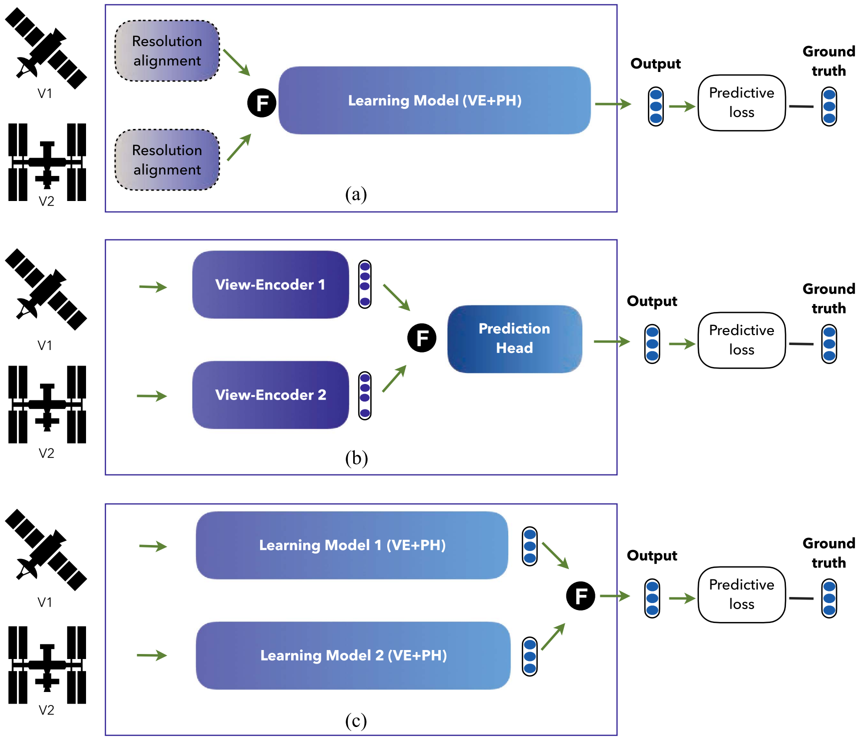


Fig. 2. *Where to fuse*. Illustration of three alternative fusion strategies: (a) input-level fusion at the top, (b) feature-level fusion at the middle, and (c) decision-level fusion at the bottom. The model forward pass is from left to right (green arrows), the VE stands for view-encoder and PH for prediction head.

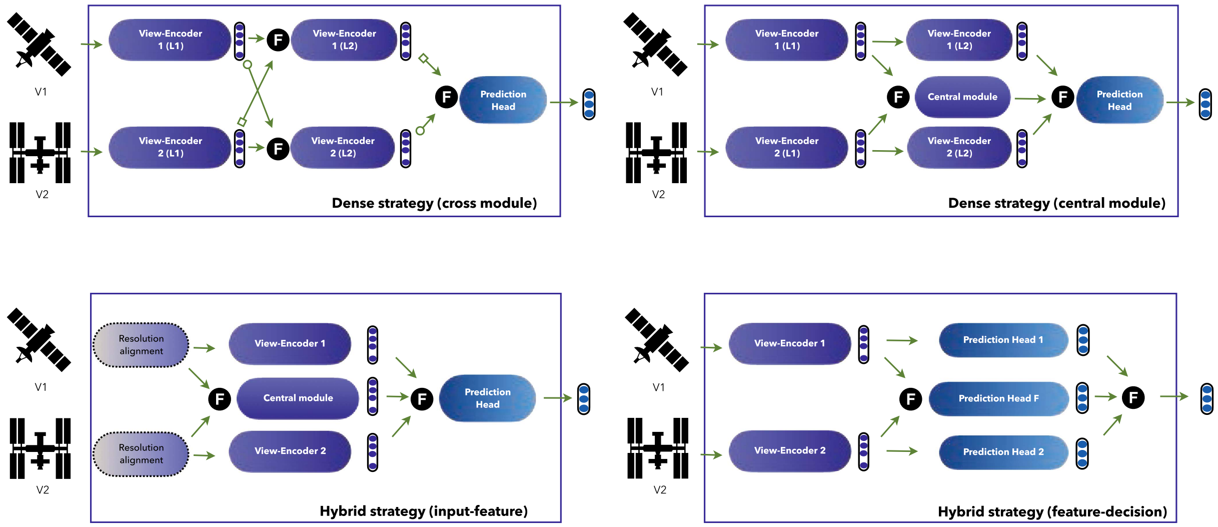
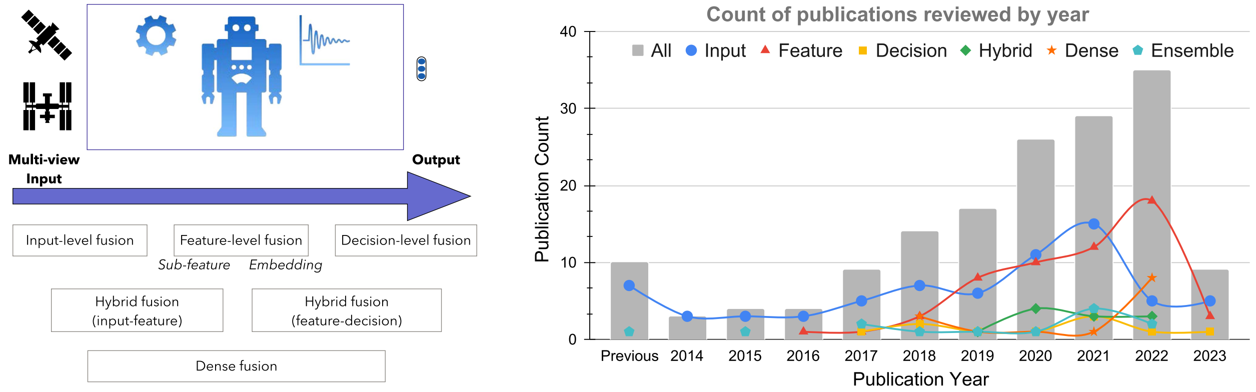


Fig. 3. *Where to fuse*. Illustration of additional fusion strategies from literature. Two versions of dense fusion methods are at the top of the figure, and two hybrid fusions are at the bottom. The forward pass of the model is from left to right (green arrows). Since dense strategies could use different connections in the crossing layers, the paths are distinguished with a circle and square at the arrows beginning.



(a)

(b)

Fig. 4. Illustration of (a) fusion terms and (b) publication count for all the fusion strategies described in Section IV-A.

and reasoning. These developments represent a significant step forward in the application of NN models in the field of EO.

III. Challenges of Deep MV Learning

In the following section, we discuss the intrinsic challenges associated with deep MV learning using RS data. For a comprehensive review of the deep MV learning topic outside the EO domain, refer to [39].

Wang et al. [5] presented interesting insights into the difficulties of training MV learning models in the vision domain. We comment on the results that also apply to RS image data.

1) *Heterogeneity in modeling:* Views can have different resolutions (e.g., spatial and temporal), requiring different model or NN architectures to process them. For instance, a CNN for spatial data, an RNN for temporal data, and an MLP for tabular data.
2) *Different information levels:* Views may concentrate the information at different levels (e.g., high-level versus low-level feature or high versus low noise), and each one will require different network complexity to process them. For example, an optical image might require more layers in the NN than a cloud mask.
3) *More patterns:* Each view has its own distribution, magnitude, and behavior patterns. Therefore, by feeding more input views (patterns) to the network, the model must determine the most efficient way to relate these views to the output. For instance, the curse of dimensionality[1] [6] could occur when RS views are concatenated [7].
4) *Overfitting:* The number of parameters increases with each additional network, compared to using a single network,

[1] In machine learning, the curse of dimensionality can manifest itself through the decrease in predictive performance when increasing the number of input-features (dimensionality).

i.e., single-view learning. This scenario could cause overfitting due to overparameterization if there is not enough labeled data to learn from.

The amount and quality of labeled data are crucial to reduce the error of supervised trained models. Besides, MV learning models are more difficult to learn than single-view learning models [5] because each view overfits and generalizes at a different rate. However, MV models usually optimize in a single framework that learns all the parameters together at the same learning rate, i.e., does not consider the difference in views overfitting.

In addition, shared (or common) and complementary (or specific) information could be presented in the views [9], [56] that is usually ignored in the model design and included in the model bias. Christoudias et al. [57] explored the problem when complementary information among views is high enough to have view disagreement. View disagreement is when views express contradictory information about the ground truth [57], causing problems in training. Ideally, one looks for views that balance complementary and similar/correlated information. Moreover, determining the optimal views required to describe an object accurately can be challenging. Whether additional views will improve or worsen the model's performance is often unclear. This raises the question of selecting the appropriate number of views to ensure that the model performs optimally without overloading with redundant information. With this discussion, we try to give an impression to researchers outside the field on why this topic is currently being studied. To date, many works have been proposed in the literature to properly handle and merge the information contained in MV data, of which we discussed in this manuscript.

IV. Main Questions

Learning from multiple views is a complex task for learning models (see Section III), which might require important decisions by the *practitioner*. We refer to the "practitioner" as the person in charge of the model design and experimentation in the machine learning context. Numerous studies [18], [32], [34], [36], [55], [58] have demonstrated that the choices made in the fusing method can significantly impact the predictive performance of machine learning models. These choices may generate some questions that we try to answer in this section, such as in which layer to fuse the data (*where to fuse*), which functions to use for fusing the information (*how to fuse*), in which part to assign more resources (*what to focus*), or which approaches are more common for specific EO scenarios.

A. Where to Fuse?

A common question is at which stage of the deep learning model it is recommended to fuse data, in the early, middle, or late stages. However, the concepts of early, middle, and late can be ambiguous in the context of NNs. For instance, *which layers define the boundaries between partitions?* Some works [16], [18] considered late as the fusion on the decision layer of multiple models, while others [34], [59] considered late as the fusion one layer before (hidden representation) the decision layer. If the number of features in the hidden representation is large, it could make a big difference between merging in this representation or in the decision layer (which usually has a few features). In Fig. 2, we categorize a more explicit terminology for early, middle, and late fusion strategies inspired by the literature [1], [34], [36], [60]. We categorized (not mutually exclusive[2]) 160 articles into one or more fusion strategies (see all the articles categorized in Table IV in the Appendix).

Input-level fusion (early or data level fusion) is a naive approach involving concatenation of the input data and feeding it to a single model. The learning setting is similar to a single-view model. A resolution alignment step is often required to match all the dimensions of the views (see Fig. 2 for an illustration) before the concatenation, e.g., spatio-temporal alignment using resampling or interpolation operations or more sophisticated operations such as feature extraction. This fusion in the data layer was the most common strategy in our review, used in 72 articles with RS image-based applications.

Feature-level fusion (middle, intermediate, or layer level fusion) yields a joint representation (ideally compact) that is useful for a predictive task. This method uses NN view encoders to generate an intermediate representation for each view, followed by a fusion module and a prediction head (see Fig. 2). This fusion strategy at the hidden representation layer was found in 56 articles. Hidden features at the first layers of the NN are usually called low-level features, while hidden features at the last layer of the NN are usually called high-level features. Motivated by this, we divided feature-level fusion into two subcategories. Fusion of subfeatures when the fused features have temporal, spatial, and/or spectral dimension(s), such as image or time-series features. For example, Audebert et al. [18] fused feature maps (spatial features) inside convolutional blocks of CNNs for an LULC application. On the other side, a fusion of embeddings implies fusing vector features. For example, Chen et al. [32] fused vector features after extracting an embedding with CNN for LULC.

Decision-level fusion (late level or classifier fusion) combines view-based predictions (e.g., probabilities, logits, or numerical values) from parallel single-view models processing each view and yielding a decision (see Fig. 2). This strategy bears resemblance to the mixture of expert model [61] but where each expert receives a different input data. The decision-level fusion was the least used strategy in the literature reviewed, with ten articles. As Fig. 2 shows, feature and decision-level fusion use multiple models to process each view, while input-level fusion uses a single model.

Table I outlines key considerations for practitioners when selecting a fusion strategy. The most appropriate fusion strategy may depend on the specific RS image-based application.

An alternative late-level strategy, distinct from the previous decision-level mentioned, has been employed in some studies [31], [49], [62], [63], [64], [65], [66], [67], [68], [69], [70], [71], [72], [73]. We refer to this approach as *ensemble-based* aggregation (with 13 articles in total), since it is based on a two-step process where fusion is not learned but added. The first

[2] More than one fusion strategy could be presented in each article.

TABLE I
TECHNICAL DIFFERENCES OF THE MAIN FUSION STRATEGIES PRESENTED IN THIS MANUSCRIPT, WITH V THE NUMBER OF VIEWS IN THE MODELING

Considerations	Input level	Feature-level	Decision-level
General comment	Straightforward modeling. However, the alignment step might require domain knowledge and can add bias. Besides, it relies on the conditional independence of the input views.	Avoids the alignment step and moves the fusion inside NN layers, allowing nonlinear feature exchange.	Avoids the alignment step by moving the fusion to the last layer of parallel NN models. However, due to the lack of feature exchange modules, it is difficult to learn cross-modal features.
Suitable for	Views with the same data type (e.g. only images or only time series).	Views with different data types (e.g. images, time series, and metadata).	Views with different data types, where individual views are sufficiently discriminatory for the predictive task.
Architectural decisions	One encoder and one prediction head.	V encoders and one prediction head.	V encoders and V prediction heads.
Additional decisions	Alignment step	Merge function	Merge function
Designed to learn	-	Cross-features between the views.	Fusion conditional independent of the input data.
Nonlinear mapping of views	Unknown	Yes (explicit)	Yes (explicit)
Critic part in modeling	Input dimensionality	Model parameters	Model parameters
Number of learnable parameters	Relatively low	Middle	Relatively high
Model complexity per view	Same	Any (same or different)	Any (same or different)

step trains a model for each view independently. The second step happens at test time, where a prediction aggregation function merges the view-based predictions (e.g., through the average or majority vote), similar to an ensemble framework. This case is a model-agnostic fusion, where the information from multiple views does not interact with each other, nor the relationship between these is exploited, i.e., the fusion is detached from the learning.

More recent works have explored different fusion strategies based on the flexibility of NN models and their representation capability (at feature level). One strategy is *hybrid* fusion (with 11 articles found), which combines two-level fusions in the same model. To illustrate, input and feature fusions [see Fig. 3(bottom-left)] [72], [74], [75], [76], [77], [78], [79], [80] or feature and decision fusions [see Fig. 3(bottom-right)] [33], [58], [81] are used together, improving the prediction performance compared to using a single layer fusion. These hybrid fusion strategies can be applied to different views, e.g., input-level fusion for temporal features and feature-level fusion to integrate tabular auxiliary data [74], [76], [77]. A natural extension of the hybrid strategy is the case when the feature fusion is integrated into multiple layers of the model. We refer to this as *dense* fusion [82] (with 14 articles found). One option involves the use of cross modules between the intermediate view representations [see Fig. 3(top-left)] [18], [65], [83], [84], [85], [86], [87], [88], which could be directed to a specific view (illustrated with a circles and squares in the Fig. 3), or to use an additional central model that stores the previously combined features [see Fig. 3 (top-right)] [18], [58], [89]. In summary, Fig. 4 provides an overview of fusion strategies. It can be seen that input-level fusion is the standard method used by several works across the years, while feature-level fusion shows an increasing trend. In addition, it shows that hybrid and dense fusion are more current strategies gaining popularity among researchers.

B. How to Fuse?

The practitioner must define how the fusion will be performed. Many data fusion architectures with different ways to merge data have been explored thanks to the flexibility of using deep MV learning. For instance, using different merge functions such as uniform-sum, weighted-sum, product, concatenation, or with gated modules (as in adaptive fusion) [90], [91]. For input-level fusion, concatenation is the most common merge function. In the following paragraphs, we will discuss some common merge functions used in the other fusion strategies. For simplicity, consider the representation (or prediction) obtained by V view encoders (or -prediction heads) on each view $\{z_v\}_{v=1}^{V}$ and the merge function $F(\cdot)$, the final fused (or joint) representation could be expressed by $z^* = F(\{z_1, z_2, \ldots, z_V\})$. Table II summarizes some function options used in the literature, which are simplified for mathematical comparison. The merge function F could even be a submodule that includes complex components or layers. Some options include cross layers with direction from one view to another (asymmetric) [18], [65], [83], [84], [86], [87], [88], [92], [93], cross layers with multiple directions among all views (symmetric) [18], [34], [85], [94], [95], [96], [97], [98], a central additional submodel [18], [58], [89], submodel with average correction in decision-level fusion [17], or even an RNN submodel for sequential views [99], [100]. Nevertheless, these could be seen as a special case of dense fusion, since the fusion is already integrated across the model. Therefore, a dense fusion model could eventually learn this one-step (or two-step) fusion within the model layers.

To the best nowledge, we have not encountered any prior research that specifically examines and compares merge functions, particularly within the EO domain. While most of the merge functions produce a pooling aggregation (fused dimension is the same as individual-view dimensions), the most commonly used in literature, concatenation, gets a dimensionality equal to the sum of individual dimensions. However, a few works in LULC have shown that the pooling functions perform better than concatenation when fusing an optical-MS with a LiDAR view [81], or an optical-RGB with a DSM view [58]. When comparing different pooling aggregations, other publications show that the convex combination of a uniform sum was better than the maximum for land-use characterization with RS and ground-based views [101], or than the product for crop-type mapping with optical-MS and SAR views [1], or than the majority voting for building extraction with optical and DSM views [58]. This could be because a convex combination does not possess the same level of pooling strength as maximum or product operators. In addition, Hong et al. [102] showed that view-specific features are more useful for prediction than

TABLE II
COMMON MERGE FUNCTIONS IN THE LITERATURE

Name	Merge function choice (F)	Additional	Commonly used in	Mode	Articles
Concatenation	$z^* = [z_1, z_2, \ldots, z_V]$	-	Early and Feature	stack	57
Attention	$z^* = \alpha \odot [z_1, z_2, \ldots, z_V]$	with $\alpha = G(z^{'}) \in [0, 1]$	Feature and Dense	stack	7
Directed Attention	$z^* = \alpha \odot z_v$	with $\alpha = G(z^{'}) \in [0, 1]$, and v a target view	Feature and Dense	stack	6
Uniform-sum	$z^* = \sum_{v=1}^{V} z_v$	(optional normalization) $z^* = z^*/V$	Feature and Decision	pool	23
Weighted-sum	$z^* = \sum_{v=1}^{V} g_v \odot z_v$	with $\sum_{v=1}^{V} g_v = 1$ learnable or fixed parameters	Ensemble and Decision	pool	9
Gated-sum	$z^* = \sum_{v=1}^{V} g_v \odot z_v$	with $\sum_{v=1}^{V} g_v = 1, g_v = G(z^{'})$	Dense and Feature	pool	7
Product	$z^* = z_1 \odot z_2 \odot \ldots \odot z_V$	(optional normalization) $z^* = z^*/C$	No preference	pool	4
Maximum	$z^* = \max\{z_1, z_2, \ldots, z_V\}$	-	Feature and Hybrid	pool	4
Majority Vote	$z^* = \text{mode}\{z_1, z_2, \ldots, z_V\}$	-	Ensemble and Decision	pool	4

With $z* = F(\{z_1, z_2, \ldots, z_V\})$, $\odot$ the Hadamard (element-wise) product, z' an auxiliary joint representation, $G(\cdot)$ an auxiliary function, and C a normalization factor. The entire list of articles is in Table V.

view-shared features, suggesting a step toward exploiting the individual information within each view.

C. What to Focus?

As an additional concern, the practitioner could be curious about which aspects of the modeling warrant particular attention. For example, in which part to design a more complex architecture to obtain better predictive performance. Since RS-based views have different resolutions, e.g., an image, a sequence of images, vectors, or metadata, it is necessary to define an appropriate view-encoder model to process each view. The most common choice of view encoders for RS image data is MLP or well-known models and architectures for the specific data [34], [103], [104], [105], [106]. For instance, when using optical images (usually for LULC), the common choice is to use known CNN architectures [107], such as AlexNet [108], [109], [110], [111], [112], [113], VGG [55], [114], [115], [116], [117], [118], InceptionNet/GoogleNet [110], [118], [119], ResNet [69], [76], [89], [99], [120], [121], [122], [123], [124], DenseNet [76], SENet [69], and EfficienNet [76], [125], [126]. While SegNet [127] and Faster R-CNN [16] have been used for optical-MS image segmentation. On the other hand, RNN models (with LSTM or GRU) are usually selected when using temporal data [105], [106].

As observed in other fields, increasing the complexity of MV models result in an improvement in the predictive performance for RS image-based tasks. Some examples are increasing the network layers and branches on LULC [18] or increasing the network parameters on automatic target recognition [100]. However, some cases present the opposite evidence [34], [106]. Using a less complex model (few layers) achieves a good predictive performance in MV learning.

Some works [128], [129] recommend pretraining the view-encoder or having prediction heads for each view independently. The application of a pretrained model (aka transfer learning) to the EO domain varies from fine-tuned models pretrained on large image datasets outside the EO domain [19], [55], [117], to models pretrained with the same task in a different geographical region [130], including pretrained models on a different task in the same geographical region [76], [131]. For example, Khaki et al. [131] proposed a crop-type classification transfer on the same region, i.e., fine tuning in a different crop type.

Sahu and Vechtomova [132] mentioned that simple view encoders could be combined with a complex fusion mechanism, which can make it compete against complex single-view models (such as transformers or deep networks). This highlights the importance of focusing on the fusion modules rather than only on the complexity of the view encoders. However, the empirical evidence of Gadiraju et al. [106] for crop classification showed that having a linear model (SVM) after the fusion step worked better than a nonlinear model (MLP). Their approach implemented complex view encoders to process the optical-RGB image and NDVI time series views. Ienco et al. [15] obtained similar results for LULC when comparing conventional learning models such as random forest versus MLP with the fusion of S1 and S2 data. Research on RS image-based applications has been mostly restricted to limited comparisons of the complexity of NN models before and after the fusion process. Such a comparison would be of great interest as it could shed light on which stage of the modeling process, prefusion or postfusion, requires more resources to achieve optimal performance. In addition, it is essential to pay attention to certain aspects such as regularization techniques, including dropout, batch normalization, pretraining, data augmentation, and early stopping, to ensure the model does not overfits toward a specific type of pattern.

D. What to Consider in Specific EO Applications?

As mentioned in Section I-A, there are different types of RS image-based applications requiring a downstream task in EO. For applications involving tasks like LULC classification or segmentation, cloud detection, flood mapping, and vegetation recognition, it is common to encounter the use of static RS images, e.g., a single time frame. For these application scenarios, the most commonly used views are optical and radar images, DSM, and LiDAR maps. Another application scenario that uses temporal data, like a sequence of RS images or a signal of different measurements, involves applications like agricultural yield prediction, vegetation recognition (including crop-type classification), environmental parameter retrieval, and change detection. In these applications, it is common to use weather, optical, and radar views.

For the scenario of static RS views, LULC is the most explored application, with feature-level fusion as the most common fusion strategy. It is quite common to use residual connections and/or layers that operate at multiscales in this scenario [92], [93], [94], [95], [121], [133], [134], [135], [136], [137]. For LULC segmentation, we found three standard approaches: image-to-image mapping, pixel-wise classification, and neighborhood pixel-wise classification. Furthermore, when we filter approaches that employ feature-level fusion, the neighborhood pixel-wise classification approach is the most common one regardless of the application [32], [34], [35], [59], [93], [94], [96], [122], [133], [136], [138], [139]. To address the limited amount of labeled data in static MV image data for different applications, some works use pretrained models on ImageNet as initialization of one or multiple layers inside the MV learning model [19], [55], [69], [89], [117], [121], [122], [125], or use different data augmentation techniques [18], [117], [134], [140], such as random crops, flips, and rotations.

For the scenario of temporary RS views, we found input-level fusion as the most explored fusion strategy. Often, by extracting (and aggregating) features across the temporal dimension, as described in [2], [10], [12], [54], [141], [142], [143], [144], [145], [146], [147], [148], and [149], or by aligning all the views to the same temporal resolution [7], [130], [149], [150], [151], [152], [153], [154], [155], [156], [157], [158], [159]. A common approach to aggregate the temporary information is to stack a temporary NN and return the last state. For instance, 2-D CNN with an LSTM [153], 3-D CNN with a convolutional-LSTM [158], or 2-D CNN with a GRU. Another approach would be to stack a pooling layer. For instance, a convolutional-LSTM network with an average layer [160], LSTM with a temporal attention layer [155], [157], or MLP with a temporal self-attention layer [1], [36]. In addition, some works consider the temporal information as an additional input channel beyond the standard positional encoding [161] used in transformer-like models for EO [1], [162], [163]. For instance, the day of the year [160], [164], or temporal differences between consecutive times [165].

V. Modeling Considerations

In addition to the previous questions and choices. Some works have proposed different components or modules that can help in the stability and predictive performance of MV models. In particular, the dropout[3] [166] has been mentioned as a crucial technique to include throughout the MV model for better learning [32], [60], [70], [104], [157]. Other works [17], [32], [34], [35], [70], [81], [87], [88], [104], [125] mentioned the batch normalization[4] [167] with the same purpose. Recent research has proposed innovative techniques such as sharing parameters between the prediction heads for each view while maintaining view-specific batch normalization, as demonstrated by Wang et al. [168], and applied to RS image-based applications by fusing HS and SAR views [135].

Some techniques used to improve model generalization and to reduce overfitting and inductive bias are as follows:

1) *Feature reduction or selection [7], [144], [157], [160], [169], [170], [171], [172], [173]:* Reduce the number of features or bands in the case of images, in each view. The idea is to remove the redundant or nonrelevant information. This technique helps to prevent the curse of dimensionality [7]. For instance, Ghamisi et al. [171] reduced the number of bands in optical-HS images with extinction profiles and kernel principal component analysis. Others used NNs to learn new features, e.g., for bands in optical-MS images [19], or for image and text data [174].
2) *Share model parameters [33], [81], [100], [103], [117], [135], [168], [175], [176], [177]:* To avoid overparameterization in the MV learning model, this technique employs shared parameters in NN layers across views. Similarly to SiameseNets [178] (see [179] for a survey) or TwinNets [180], as has been called in recent works. The parameter sharing is usually applied in the whole NN for each view, but it has also been applied in a few layers, e.g., with 2-D CNN that processes optical-MS and SAR views [33], [176]. Sharing parameters is possible when the same NN architectures are designed for each view.
3) *Group views [17], [18], [75], [76], [77], [78], [79], [92], [181]:* To reduce overparametrization when multiple models are used. Views are grouped (in practice concatenation) based on their semantic or structural similarity. For instance, Wang et al. [75] grouped temporal features (optical-MS and weather data) as one view and static features (soil information) as another view for agricultural yield prediction. Grouping is possible when the views are from the same data type (e.g., images and vectors).
4) *Use pretrained NNs:* Pretrained networks on large-scale datasets can be used to transfer the learned knowledge. It requires the RS data to have a similar resolution to the one used in the pretrained model, e.g., Imagenet-based CNNs need RGB or three bands images. However, some works have used multiband images by initializing randomly the first layer or by some heuristic followed by the pretrained model [76], [89], [99], [113], [120], [121], [123], [124], [125], [126].
5) *Add prediction loss on the views [15], [33], [36], [81], [86], [182], [183]:* Incorporating an auxiliary predictive loss for each view might force all views to be used for the downstream task. This is done by including one prediction head for each view instead of having a single model after the fusion; see Fig. 5. There are some differences in the proposal regarding how to apply weights to the additional MV losses.
   a) In some LULC works [81], [182], [183], the authors set different weights for the loss of each view, which

[3] Regularization technique used to drop or sample some neurons on a NN layer. It uses a dropout ratio (practitioner defined) as the dropping probability.

[4] Technique used for reducing internal covariance-shift of the layers (and also used as regularization). It normalizes the layer features and subsequently applies an affine transformation with learnable parameters.

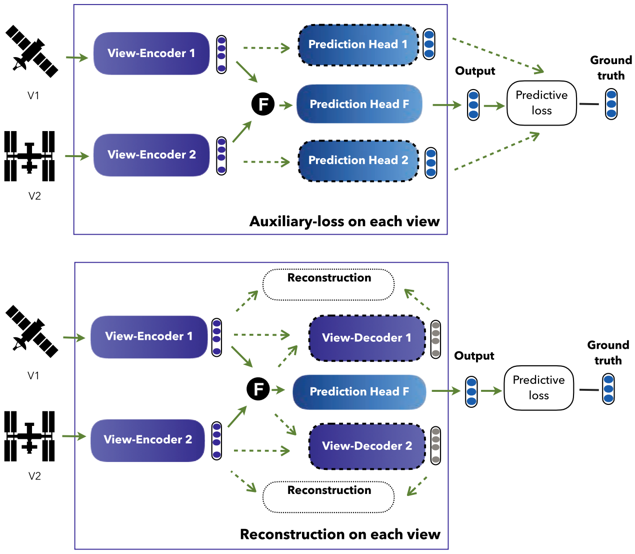


Fig. 5. Illustration of additional components in feature-level fusion: Auxiliary loss on each view at the top and reconstruction on each view at the bottom. The dashed lines are auxiliary steps, only used during training. The "prediction head F" represents the prediction head that is fed with the fused representation.

is consistent with the argument that views converge at different rates [5].

b) Ienco et al. [15] assigned one weight to the total sum of the additional MV losses.

c) In [33] and [86], the authors used the sum of all the MV losses.

6) *Reconstruct views [35], [59], [137]:* Incorporate auxiliary losses and layers that focus on reconstructing each view entirely [59], [137] or the view's representation before fusion [35] (see Fig. 5). The idea is that the learned representation contains enough information for the single-view reconstruction.

Other techniques used to help model convergence and learning are mentioned in the following paragraphs.

7) *Perform pretraining and fine tuning:* Pretrain each view encoder individually on the same RS data used for the downstream task. For example, learn to predict the downstream task based on each view individually [71], [133], [136], [184] or learn to reconstruct the views as a pretraining [137], then fine tune with the downstream task. With this technique, the parameters of the view encoders have already learned information about the RS patterns in the corresponding data.

8) *Include attention modules:* Inspired by the success of attention modules [161], [185], [186]. The main focus of application in RS data has been to perform temporal attention [1], [15], [36], [74], [155], [157], [165], [182], [187], [188]. Nevertheless, in RS image-based applications, the use cases have been extended as a way to enhance the information. Then, attention has been applied across spatial [100], [177], spectral [123], [134], [181], spatio-spectral [78], [80], [136], [139], or vector [138] dimensions. The motivation is that attention mechanisms lead to adaptively enhance the most relevant features of the data.

9) *Fuse with attention:* One application of attention modules involves to fuse the MV data adaptively (see Section IV-B). For instance, gated fusion for LULC with the attention across vector [122], spatial [85], [86], [88], [92], [124], or spectral [58] dimensions. Another case corresponds to crossing layers that apply attention from one view to another (and vice versa) enhancing spatial [89], [95], spectral [85], [88], [92], [97], [113], or spatio-spectral [87], [93], [94], [96] dimensions.

10) *Include residual learning:* Inspired by residual NN (or ResNet [189]) and its ability to build very deep models, some studies use skip connections in the MV learning model. For instance, skip connections could be included between each view (before fusion) [78], [86], [97], [123], [134], [135], [136], [183], [190], from each view (before fusion) to after fusion [17], [18], [65], [85], [94], [95], [98], [124], [135], [137], [191], [192], or between after fusion layers [84], [88], [193], [194].

11) *Normalize view representation:* To handle different feature scales in the MV data, a normalization of the learned view representations could be included before applying the merge function. For example, Marmanis et al. [55] used it on optical (RG+NIR) and DSM view representations, and Zhang et al. [137] used it on fusing optical-HS and DSM view representations. Li et al. [135] remarked in their proposal that having specific normalization layers for each view is crucial.

12) *Use a different learning on the views:* This technique assigns different learning rates for each view in the model. To give an idea, Zhang et al. [58] used a higher learning rate on a DSM view and a lower rate on an optical-RGB view. Other approaches with RS data are in the same line [81], [182].

Furthermore, there is the option to include more fusion channels, as is the case of hybrid or dense fusion. The main purpose is to allow explicit fusion at multiple levels of abstraction of the NN layers. However, it has been shown in the vision domain that fusion in just some layers is better than in all of them [82], [129]. Other approaches have used postprocessing after training, e.g., updating the predictions based on the application and domain-knowledge [52], [68], [69], [70], [72], [169], [195]. Furthermore, some works have included context information in the modeling, e.g., neighboring pixels as input data on pixel-wise predictions [19], [32], [34], [35], [59], [87], [93], [96], [106], [122], [133], [136], [138], [169], [171], [182], [196], [197] or graph-based constraints on similar pixels/patches [102], [169], [171], [175], [198].

VI. Current Resources and Results

The following analysis provides a general perspective of the current RS sources and views, as well as fusion approaches, focusing on the predictive performance for EO applications.

A. Which Views are Most Used in Earth Observation?

Optical (surface reflectance) data are the views that usually provide further information, and therefore, are the most used for various downstream RS image-based tasks. Meanwhile, active-based views are the quintessential views chosen to complement (and improve by fusion) the optical in classification tasks with MV models, e.g., by using SAR view [1], [15], [33], [34], [36], [99], [113], [120], [124], [126], [143], [163], [173], [174], [188], [190], [193], [199], [200], [201], [202] or LiDAR view [94], [139]. Furthermore, the DSM view has been widely used together with the optical view [17], [18], [19], [55], [58], [64], [65], [76], [80], [84], [85], [88], [89], [92], [104], [123], [184], where on some occasions is a LiDAR-derived DSM [32], [34], [35], [59], [62], [81], [86], [87], [93], [96], [133], [134], [136], [137], [139], [169], [170], [171], [183], [184], [198], [203]. The visible light of the spectrum (RGB bands concretely) have been shown to be more relevant than other spectral bands in the optical view when considering the predictive performance of NN models [16], [99], [204]. Besides, views with coarse resolutions are usually worst in predictive performance than finer resolution views [106], [120], [201], and multitemporal views have better predictive performance than static views [99], [106], [188], [191], [205]. Although RS-based views provide valuable and more accessible information for downstream tasks than ground-based views [30], [117], some works have used interesting data sources to complement RS-based views. For instance, Heidler et al. [30] used ground-based audios in addition to the optical view to classify an observed place, Mantsis et al. [174] included images and text from tweets to estimate the snow depth of an observed place, and He et al. [206] included people density as a temporal view from geospatial Big Data in China. These works suggest that social views could be a powerful source to estimate disasters [207], such as earthquakes [208] and floods [209]. There are other cases of domain-specific views depending on the application. One is the case of the agricultural yield prediction, where the weather and soil views are chosen to complement (and improve) the optical view [2], [75], [77], [103], [146], [181], or even used without the optical view [73], [79], [131]. In addition, different types of metadata can be used, e.g., statistics of the planted crop [71], [74], [210], the region where it was planted [211] or irrigation factors [60], [142].

Different time periods of the same sensors could be considered as views in MV learning models, as is the case of change detection [52], [72], [98], [177], [192], [194], [212], while some works generate multiple views from single-view data, e.g., by different image operations (zoom ratios [49], color alteration [213], and rotations [175]). Others, partition the single-view data to generate MV data, e.g., splitting the spectral bands of S2 optical image for vegetation recognition [19] or water body detection [121]. Furthermore, feature extraction of spectral indexes [15], [17], [18], [71], [97], [103], [125], [172], model-based (such as principal component analysis [138]) or domain-specific feature generation [78], [102], [160], [165] have been used as an additional view to the raw bands of optical images. Nevertheless, the use of redundant optical views (multiple optical sensors) has shown compelling evidence in the literature, such as learning from S2 and L8 satellites [69], [126], [199], [201], from satellite data and UAV [72], [120], or from low and high spatial resolution images [72], [106], [120], [160], [182], [214]. This shows that different types of views can be considered for different predictive tasks, and that the concept of “view” grants flexibility in exploring methods within the same framework.

There exist several publicly accessible datasets that can be used for downstream tasks involving multiple RS views. In Table VII (in the appendix), we provide an overview of some datasets that are also used as benchmarks for specific applications and validation. Another valuable resource for researchers is the annual data fusion contest hosted by the IEEE Geoscience and Remote Sensing Society, where challenging datasets are made publicly available.[5]

B. Does the Use of Additional Views Improve Predictive Performance?

Outside the EO domain, substantial evidence suggests that additional views or modalities improve the predictive performance on downstream tasks regarding using single-view data [91], [129], [215], [216]. Does this result apply to RS-based applications? In the following, we describe some results in this direction.

There is plenty of evidence from works using two-view data on various RS image-based tasks. These works show that the predictive performance improves with respect to training on any of the single views, e.g., with optical and active-based views (SAR/LiDAR/DSM) [1], [13], [13], [15], [18], [32], [33], [34], [35], [58], [59], [65], [81], [86], [92], [93], [94], [96], [99], [102], [113], [133], [136], [143], [163], [169], [190], [191], [193], [199], [202]. This indicates that the views complement each other in the MV learning for EO tasks, in addition to the fact that there is evidence when other diverse views are chosen to supplement or replace the optical view [16], [70], [71], [106], [117], [159], [160], [181], [183], [210], [211]. Several publications have shown that improvements regarding the optical view appear even when using more than two views. Nguyen et al. [103] demonstrated that optical-MS images, NVDI, and soil properties perform better than individual views for agricultural yield prediction. Pageot et al. [217] showed that optical-MS image, SAR image, and weather improve over individual views for irrigation recognition. Song et al. [201] fused optical-MS from S2, optical from MODIS, optical-MS from L7 and L8, and SAR from S1 views and found better predictive performance as compared to individual views for crop-type mapping. Irvin et al. [76] showed that optical and auxiliary data extracted from DSM, weather, and soil properties improve over optical for detecting deforestation. In [49], multiple zooms of an optical image improved predictive performance over a single perspective for LULC. Wang et al. [100] found that multiple angles of an object improve over a single angle for automatic target

[5] www.classic.grss-ieee.org/community/technical-committees/data-fusion/ (Accessed 19 December 2023)

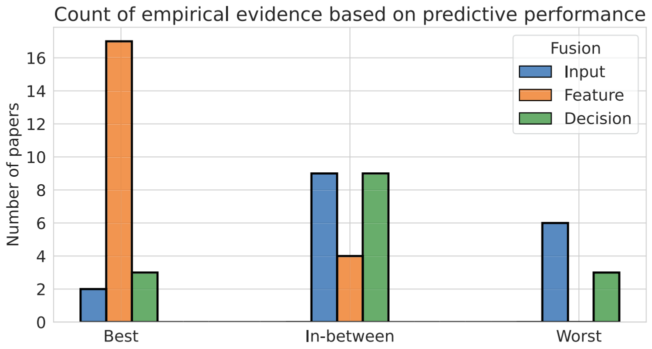


Fig. 6. Number of articles that show empirical evidence of being best, in-between, or worst predictive performance within three fusion strategies (input, feature, and decision fusion). Individual articles are in Table VI.

recognition. Finally, some works [19], [60], [75], [83], [104], [120], [146], [205], [217] presented a monotonically increasing trend in predictive performance by including additional views in the MV model. This evidence suggests that learning from MV data outperforms in terms of predictive performance as compared to learning from single-view data in RS image-based applications. Heidler et al. [30] showed an interesting result that MV learning improves over single-view learning when inferring on single-view data. They showed that an MV model trained with optical and audio views outperforms a single-view trained model when predicting using only the optical view. This suggests that additional views have the potential to enhance the learning process, even if they are not used during the inference stage.

Even though all the previously commented works report that using additional data improves predictive performance, there are some works that report the opposite regarding the number of features in single-view learning with conventional learning models. In these works [7], [144], [196], a subset of the features (selected with feature engineering techniques) improved the predictive performance of the models. This might suggest that conventional models with single-view learning fail to extract the patterns adequately from the additional views needed for the downstream task. However, it is essential to note that the results usually depend on the empirical data and the downstream task [201], the difficulty of that task (e.g., number of classes or task granularity) [131], the number of training examples [50], [130], [218], and the model used [172].

C. Which Strategy Leads to Better Predictive Performance?

Regarding current results in the field, some works have compared and studied the effectiveness of fusion strategies. Hong et al. [34] explored the question of where to fuse RS images in NNs. Their conclusions indicate that middle (cross) fusion has the best performance, followed by late fusion. Nevertheless, the conclusions were based on the results in two EO datasets, while in this article, a collection of different experimental evidence is gathered and summarized for comparison. In Fig. 6, we gather empirical evidence from the literature on different RS datasets by comparing the three main fusion strategies discussed in previous sections (see Fig. 2). In this case, the comparison is categorical and corresponds to whether the fusion strategy shows the best, worst, or in-between predictive performance. The main outcome is that feature-level fusion has better predictive performance compared with input and decision-level fusion in most of the cases [18], [32], [34], [35], [36], [58], [86], [101], [104], [105], [133], [200]. Furthermore, based on the analyzed evidence, the feature-level fusion performance stands above the worst performance alternatives. Input-level and decision-level fusion strategies are similar in results to each other. In some cases, input fusion is the best [1], [36] (or worst [34], [35], [190], [200]), while in others, decision fusion is the best option [16], [33] (or worst [1], [36]), showing that the results strongly depend on the data. The available evidence suggests that input and decision-level fusion techniques are more unpredictable regarding their effectiveness in RS image-based applications, making it challenging to determine a priori whether they will yield superior or inferior results. Srivastava et al. [101] compared empirically that when no fusion is performed, feature-based aggregation (sum of NN features before training) is better than ensemble-based aggregation (majority voting of predictions after training) when using RS-based and ground-based views for LULC. In addition, many works reported that additional fusion layers improve the predictive performance compared to using just one fusion layer. For instance, in hybrid fusion [18], [33] or dense fusion [58], [83], [97], [177], [219]. These works give credit to the idea that the model can exchange more information from the views and correct the fusion of the earlier stages. Furthermore, it is worth noting that the differences in predictive performance observed across various fusion strategies in the literature are generally modest.

Considering that the selection of the fusion strategy might depend on the EO application, we offer some suggestions for practitioners in Table III. Since different solutions can be proposed for the same application using different RS sources, we group the applications in the three task types defined in Section I-A as follows: segmentation, classification, and regression.

An evaluation of the methods is not provided as the results reported in the literature exhibit significant variability depending on various factors, including the downstream tasks, geographic region, RS satellites, and views used. As such, we refrain from offering incomplete or biased insights on this topic. Nevertheless, some authors have made the code of the MV learning model available alongside their manuscripts, allowing reproducibility and progressive research. These valuable resources are summarized in Table VIII (in the Appendix).

VII. Open Questions

Although many fusion approaches have been explored in the MV learning topic, there are still some open challenges that could motivate new research and proposals in the EO domain.

1) *Missing views:* The usual assumption of MV learning is that all views are available for each sample during and after training. However, RS-based scenarios are dynamic environments that do not necessarily follow this assumption, e.g., remote sensors may fail or be unavailable, causing a MV learning with missing views [220]. Only a few works

TABLE III
ALTERNATIVES FOR FUSION STRATEGIES DEPENDING ON THE TASK TYPES

Task type	EO application	Most used strategy	Most successful strategy	Most used merge function
Segmentation	LULC mapping, vegetation mapping, flood mapping, cloud detection, change detection, deforestation detection, wildfire detection.	Feature, Input, then Dense.	Feature, then Decision.	Concatenation, then Uniform-sum.
Classification	LULC classification, scene understanding, vegetation recognition, automatic target recognition, crop irrigation detection.	Feature, Input, then Decision.	Feature, then Input.	Concatenation, then Uniform-sum.
Regression	Cloud removal, crop yield prediction, weather forecast, moisture estimation	Input, Feature, then Hybrid.	Feature, then Input.	Concatenation, then Weighted-sum.

The “most successful” column was generated from Table VI.

have explored the effect of fusion when this occurs. For instance, fusions further away from the input data (e.g., decision and feature fusion) are more robust to missing data[6] in HS and LiDAR images [34]. Other works present the same results when missing[7] optical-MS images at some time steps of an optical and SAR MV model [1], [36], [163] or in cloudy conditions [70], [121], [143], [190]. Hong et al. [34] also showed that the robustness to missing views could be increased by including additional components to the standard fusion approaches (see Fig. 2). These solutions are usually data-specific techniques and require knowing in advance what and when the missing views will be. However, NN models have shown the ability to reconstruct complex RS image patterns [205]. Adapting the MV learning model to missing views still has open questions that could motivate further research. For example, when more than two views are used, the missing views can be dynamic, and a robust model could be suitable.

2) *MV uncertainty analysis:* Not many studies have analyzed the prediction uncertainty when using MV data. It is reasonable to ask whether additional views reduce or increase the uncertainty. It might happen that if views are too different from each other, the uncertainty increases, or if views are more similar, the uncertainty is reduced. Ofori-Ampofo et al. [1] showed that, when running multiple times, the MV models obtain a lower variance compared to the single-view ones. Ebel et al. [188] showed that additional views reduce the aleatoric and epistemic uncertainty on model predictions for cloud removal. Nevertheless, there is an open opportunity for research in alternative applications and RS sources.

3) *Complex MV models:* Several studies in RS image-based applications have proposed diverse models and architectures to address MV learning [15], [55], [85], [86], [88], [89], [137], [187], e.g., by making the model more complex, and thereby, able to extract better information. However, the neural architecture search field might assist in searching an optimal MV learning [221]. This might reduce the human effort of a manual model design if proper use is done with RS data.

4) *Explainability on MV learning:* The explainability of a single-view model is a research line exploring questions such as: What is the impact on the explainability as the complexity of the model increases? since current approaches use multiple models for each view to deal with MV data, increasing the complexity for each view. It is challenging to understand what the model is doing under abstract fusion operations. Therefore, significant attention needs to be put on this subject.

5) *Theories on MV learning:* Some theoretical aspects of MV learning have been explored in the ML domain, such as view consistency ($f_1(x_1) = f_1(x_2)$, [9]), sufficiency for correct classification ($f_1(x_1) = f_1(x_2) = y$, [9]), complementary ($\epsilon > 0$), and redundancy ($\epsilon = 0$) in view information ($I(y, x_2|x_1) \leq \epsilon$, [222]). With x_v the input view and $f_v(\cdot)$ the prediction model for view v, y the prediction target, and $I(\cdot)$ the mutual information. However, just a few works have used a theoretical framework for view-alignment in RS-based applications, e.g., an MV linear discriminant analysis like formulation in an LULC application [29], a canonical correlation analysis (CCA) for missing view retrieval [117], or CCA for learning kernels in CNN [213]. Nevertheless, given the current scarcity of research on theories for RS data fusion within MV learning, there is an open field to explore.

VIII. Conclusion

This manuscript analyzes different data fusion aspects in MV learning in the context of RS data. Multiple approaches from the literature were grouped based on their similarity, providing a unified structure to compare methods in the field. Our observations indicate that the utilization of advanced fusion strategies, which incorporate supplementary components and/or multiple fusion layers, leads to a notable enhancement in predictive performance. To the best of our knowledge, the works included in this review reflect the current trends of MV fusion learning in RS image-based applications.

Appendix

Table IV contains the references of articles of the categorization made in Section IV. While Table V contains the references used to create Table II in the main content of this manuscript, Table VI contains the data used to generate Fig. 6. On the side of the current resources in the literature, Table VII provides an overview of public available datasets, while Table VIII shows public available code of different research proposals.

[6] In practice, the missing views are filled with zero when feeding the model.
[7] In practice, the missing is ignored by the model or masked out.

TABLE IV
NONEXCLUSIVE CATEGORIZATION OF ARTICLES REVIEWED

Strategy	References
Input	[1]–[3], [7], [10], [12], [13], [19], [34], [36], [51]–[54], [65], [68], [71], [99], [104], [125], [130], [141]–[159], [163], [169]–[172], [174], [188], [190], [193]–[201], [203], [205], [212], [214], [217], [223]–[232]
Feature	sub-feature based: [1], [34], [36], [55], [92]–[96], [100], [120], [123], [135], [137], [173], [176], [190], [191], [203], [233], [234], embedding based: [1], [15], [32], [34], [35], [59], [60], [71], [101]–[106], [113], [117], [121], [122], [126], [131], [133], [134], [136], [138]–[140], [165], [181]–[183], [187] [192], [200], [206], [210], [211], [235], [236]
Decision	[1], [17], [18], [35], [36], [160], [175], [184], [190], [200]
Ensemble	[31], [49], [62]–[64], [66]–[73]
Hybrid	[33], [58], [72], [74]–[81]
Dense	[18], [58], [65], [83]–[89], [97], [98], [124], [177]

The discussion about each type of categorization could be found in Section IV.

TABLE V
NONEXCLUSIVE CATEGORIZATION OF MERGE FUNCTIONS USED IN SOME ARTICLES

Merge function	Reference
Concatenation	[1], [15], [17], [18], [32], [34]–[36], [55], [59], [60], [71], [72], [74]–[79], [81], [83], [87], [94], [96], [98], [102]–[106], [117], [120], [121], [123], [126], [131], [133], [134], [136], [137], [139], [140], [160], [165], [176], [177], [181], [182], [187], 190]–[192], [200], [203], [206], [210], [211]
Attention	[89], [94], [95], [97], [113], [124], [138]
Directed attention	[87], [92], [94]–[96], [234]
Uniform-sum	[1], [17], [18], [33], [35], [36], [49], [58], [63], [65], [80], [81], [97], [101], [106], [117], [135], [173], [175], [184], [200], [233], [235]
Weighted-sum	[33], [68], [70], [71], [73], [81], [84], [87], [93], [200]
Gated	[58], [85], [86], [88], [92], [97], [122]
Product	[1], [66], [113], [124]
Maximum	[81], [101], [117], [183]
Majority	[31], [58], [64], [67]

This table is a supplement of Table II.

TABLE VI
DATA USED TO GENERATE FIG. 6

Source	Views used	Inp	Fea	Dec
[16]	optical RGB, NIR	2	-	1
[32] (San Francisco)	optical-MS, DSM	2	1	-
[32] (Houston)	optical-HS, LiDAR	2	1	-
[18]	optical-MS, DSM	-	1	2
[133] (Houston-Trento)	optical-HS, LiDAR	2	1	-
[133] (Pavia-Salinas)	optical HS, RGB	2	1	-
[101]	multiple optical-RGB	-	1	2
[105]	soil, cultivation, yield stats	2	1	-
[58]	optical-RGB, DSM	-	1	2
[104]	optical-MS, thermal, DSM	2	1	-
[35]	optical-HS, LiDAR	3	1	2
[1]	optical-MS, SAR	1	2	3
[33]	optical, SAR	-	2	1
[34] (Houston)	optical-HS, LiDAR	3	1	2
[34] (LCZ)	optical-MS, SAR	3	1	2
[200]	optical-MS, SAR	3	1	2
[86]	optical-RGB, DSM	-	1	2
[36] (Classification)	optical-MS, SAR	2	1	3
[36] (Segmentation)	optical-MS, SAR	2	1	2
[36] (Panoptic)	optical-MS, SAR	1	2	-
[97]	optical RGB, infrared	1	2	-
[190] (clean)	optical-MS, SAR	3	2	1
[190] (cloudy)	optical-MS, SAR	3	1	2

The numbers correspond to: 1: Best, 2: In-between, 3: Worst, of the predictive performance when comparing the input (Inp), feature (Fea) and decision (Dec) fusion strategies. The different datasets or use-cases from the same paper are shown in parentheses.

TABLE VII
RS IMAGE-BASED MV DATASETS

Dataset	RS-based views	Temp.	Task	Region	Link
University of Houston	UAV-based: HS and RGB optical, DSM (LiDAR)	×	Image Segmentation	USA	hyperspectral.ee.uh.edu/?page_id=1075
MSLCC	MS optical (S2), SAR (S1)	×	Image Segmentation	Germany	www.dlr.de/eoc/en/desktopdefault.aspx/ tabid-12760/22294_read-51180
ISPRS 2D Challenge	UAV-based: MS optical, DSM	×	Image Segmentation	Germany	www.isprs.org/education/benchmarks/ UrbanSemLab
So2Sat LCZ42 [237]	MS Optical (S2), SAR (S1)	×	Image Classification	Global	mediatum.ub.tum.de/1483140
Sen1Floods11 [238]	MS optical (S2), SAR (S1)	×	Image Segmentation	Global	github.com/cloudtostreet/Sen1Floods11
AI4Food [239]	2 MS optical (S2, PlanetFusion), SAR (S1)	✓	Image Classification	Germany	doi.org/10.34911/rdnt.z9y7vu
ForestNet [76]	MS optical (L8), DSM (SRTM), weather (NCEP)	✓	Image Segmentation	Indonesia	stanfordmlgroup.github.io/projects/forestnet/
Dams [240]	MS optical (S2) including VI, SAR (S1), DSM	×	Image Segmentation	Global	www.kaggle.com/datasets/gdonchyts/ global-dams-from-space
LandCoverNet [241]	2 MS optical (S2, L8), SAR (S1)	✓	Image Segmentation	Global	doi.org/10.34911/rdnt.7s12zu
EarthNet [242]	MS optical (S2), weather (E-OBS), DSM (EUDEM)	✓	Image Regression	Europe	www.earthnet.tech
CropHarvest [218]	MS optical+NDVI (S2), SAR (S1), weather (ERA5), DSM (SRTM)	✓	Pixel-wise Classification	Global	github.com/nasaharvest/cropharvest
DynamicEarthNet [243]	2 MS optical (S2, PlanetFusion), SAR (S1)	✓	Image Segmentation	Global	doi.org/10.14459/2018mp1483140
Ombria [191]	MS optical (S2), SAR (S1)	✓	Image Segmentation	Global	github.com/geodrak/OMBRIA
PASTIS-R [36]	MS optical (S2), SAR (S1)	✓	Image Segmentation	France	github.com/VSainteuf/pastis-benchmark
SEN12MS-CR-TS [193]	MS optical (S2), SAR (S1)	✓	Image Segmentation	Global	patricktum.github.io/cloud_removal
WHU-OPT-SAR [124]	MS Optical (G1), SAR (G3)	×	Image Segmentation	China	github.com/AmberHen/ WHU-OPT-SAR-dataset
CloudSEN12 [244]	MS optical (S2), SAR (S1), DSM (MERIT)	×	Image Segmentation	Global	cloudsen12.github.io/
Satlas [245]	MS optical (S2 and NAIP)	✓	Multiple Tasks	Global	github.com/allenai/satlas/
MultiEarth [246]	MS Optical (S2 and L8), SAR (S1)	✓	Multiple Tasks	Global	sites.google.com/view/rainforest-challenge
MDAS [247]	MS (S2) and HS (EnMAP, HySpex) optical, SAR (S1), DSM (DLR 3K)	×	Pixel-wise Classification	Germany	doi.org/10.14459/2022mp1657312
TreeSatAI [248]	2 MS optical (S2, UAV), SAR (S1)	×	Pixel-wise Classification	Germany	doi.org/10.5281/zenodo.6598390
BEN-GE [249], [250]	MS optical (S2), SAR (S1), DSM (GLO-30), weather (ERA5)	×	Image Segmentation	Europe	github.com/HSG-AIML/ben-ge

The RS Views column contains the views (and source where it was obtained), Temp. column indicates if the labels have a temporal scope. G1 and G3 refer to Gaofen-1 and Gaofen-3, respectively. Last Access to Links January 7, 2024. A Copy of this information can be found at github.com/fmenat/multiviewRS-Datasets.

TABLE VIII
PUBLIC CODE FOR MV LEARNING MODELS PROPOSED IN RS IMAGE-BASED APPLICATIONS

Name	Reference	Description	Code link
V-FuseNet	Audebert et al. [18]	Dense fusion with 2D CNN and central model.	github.com/nshaud/DeepNetsForEO
Multi3Net	Rudner et al. [120]	Feature-level fusion with 2D CNN.	github.com/FrontierDevelopmentLab/multi3net
UNet-CLSTM	Rustowicz et al. [160]	Decision-level fusion with 2D CNN and convolutional-LSTM	github.com/roserustowicz/crop-type-mapping
HRWN	Zhao et al. [198]	Input-level fusion with 2D CNN and pixel graph constraints.	github.com/xudongzhao461/HRWN
FusAtNet	Mohla et al. [94]	Feature-level fusion with 2D CNN and cross attention.	github.com/ShivamP1993/FusAtNet
CCR-Net	Wu et al. [35]	Feature-level fusion with 2D CNN and cross view-reconstruction.	github.com/danfenghong/IEEE_TGRS_CCR-Net
MV PSE-TAE	Ofori-Ampofo et al. [1]	Multiple fusion strategies with PSE-TAE.	github.com/ellaampy/CropTypeMapping
MDL-RS	Hong et al. [34]	Multiple fusion strategies with NN.	github.com/danfenghong/IEEE_TGRS_MDL-RS
CMGFNet	Hosseinpour et al. [86]	Dense fusion with 2D CNN and gated attention.	github.com/hamidreza2015/CMGFNet-Building_Extraction
S2FL	Hong et al. [102]	Feature-level fusion with feature contrains.	github.com/danfenghong/ISPRS_S2FL
CFCNN	He et al. [206]	Feature-level fusion with 2D and 1D CNN.	github.com/SysuHe/MultiSourceData_CFCNN
MV NN	Danilevicz et al. [71]	Feature-level fusion with tabular NN and 2D CNN.	github.com/mdanilevicz/maize_early_yield_prediction
IP-CNN	Zhang et al. [137]	Feature-level fusion with 2D CNN and view-reconstruction.	github.com/HelloPiPi/IP-CNN-code
MV CNN	Lu et al. [122]	Feature-level fusion with 2D CNN and adaptive attention.	github.com/GeoX-Lab/UnifiedDL-UFZ-extraction
SE^{2}Net	Fang et al. [96]	Feature-level fusion with 2D CNN.	github.com/likyoo/Multimodal-Remote-Sensing-Toolkit
EndNet	Hong et al. [59]	Feature-level fusion with 2D CNN and view-reconstruction.	github.com/danfenghong/IEEE_GRSL_EndNet
MAHiDFNet	Wang et al. [87]	Dense feature fusion with 2D CNN.	github.com/SYFYN0317/-MAHiDFNet
AM^{3}Net	Wang et al. [93]	Feature-level fusion with 2D CNN and cross attention.	github.com/Cimy-wang/AM3Net_Multimodal_Data_Fusion
AMM-FuseNet	Ma et al. [123]	Feature-level fusion with 2D CNN and attention.	github.com/oktaykarakus/ReSIF/tree/main/AMM-FuseNet
MCANet	Li et al. [124]	Dense fusion with 2D CNN and cross attention.	github.com/yisun98/SOLC
ChangeFormer	Bandara et al. [177]	Dense fusion with transformer and attention.	github.com/wgcban/ChangeFormer
CMAFF	Qingyun et al. [97]	Dense fusion with 2D CNN and cross attention.	github.com/DocF/CMAFF
OmbriaNet	Drakonakis et al. [191]	Feature fusion with 2D CNN and skip-connections	github.com/geodrak/OMBRIA
DCSA-Net	Wang et al. [80]	Hybrid fusion with 2D CNN and attention.	github.com/Julia90/DCSA-Net
Siamese U-Net	Cummings et al. [98]	Dense fusion with 2D CNN and skip-connections	github.com/solcummings/earthvision2021-weakly-supervised
ELECTS	Russwurm et al. [214]	Input-level fusion with LSTM.	github.com/marccoru/elects
MV CNN	Ferrari et al. [190]	Multiple fusion strategies with 2D CNN (encoder-decoder)	github.com/felferrari/deforestation-from-data-fusion
AFCF3D-Net	Ye et al. [194]	Input-level fusion with 3D CNN.	github.com/wm-Githuber/AFCF3D-Net
UnCRtainTS	Ebel et al. [188]	Input fusion with 2D CNN and attention.	github.com/PatrickTUM/UnCRtainTS

Last access to links February 6, 2024. A copy of this information can be found at github.com/fmenat/multiviewRS-Models.

Authors' contributions: F. Mena: Conceptualization, methodology, investigation, writing original draft, and visualization; D. Arenas: Conceptualization, methodology, review and editing, and supervision; M. Nuske: Conceptualization, review and editing, and project administration; and A. Dengel: Supervision, funding acquisition, and resources.

References

[1] S. Ofori-Ampofo, C. Pelletier, and S. Lang, "Crop type mapping from optical and radar time series using attention-based deep learning," *Remote Sens.*, vol. 13, no. 22, 2021, Art. no. 4668.

[2] L. Meng, H. Liu, S. L. Ustin, and X. Zhang, "Predicting maize yield at the plot scale of different fertilizer systems by multi-source data and machine learning methods," *Remote Sens.*, vol. 13, no. 18, 2021, Art. no. 3760.

[3] M. Kumar, A. Bandyopadhyay, N. S. Raghuwanshi, and R. Singh, "Comparative study of conventional and artificial neural network-based ETo estimation models," *Irrigation Sci.*, vol. 26, no. 6, pp. 531–545, 2008.

[4] Q. Yuan et al., "Deep learning in environmental remote sensing: Achievements and challenges," *Remote Sens. Environ.*, vol. 241, 2020, Art. no. 111716.

[5] W. Wang, D. Tran, and M. Feiszli, "What makes training multi-modal classification networks hard?," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 12695–12705.

[6] P. Ghamisi et al., "Multisource and multitemporal data fusion in remote sensing: A comprehensive review of the state of the art," *IEEE Geosci. Remote Sens. Mag.*, vol. 7, no. 1, pp. 6–39, Mar. 2019.

[7] Y. Kang, M. Ozdogan, X. Zhu, Z. Ye, C. Hain, and M. Anderson, "Comparative assessment of environmental variables and machine learning algorithms for maize yield prediction in the US midwest," *Environ. Res. Lett.*, vol. 15, no. 6, 2020, Art. no. 064005.

[8] H. Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, no. 3/4, pp. 321–377, 1936.

[9] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," in *Proc. 11th Annu. Conf. Comput. Learn. Theory*, 1998, pp. 92–100.

[10] N. Kim and Y.-W. Lee, "Machine learning approaches to corn yield estimation using satellite images and climate data: A case of iowa state," *J. Korean Soc. Surveying, Geodesy, Photogrammetry Cartography*, vol. 34, no. 4, pp. 383–390, 2016.

[11] M. D. Johnson, W. W. Hsieh, A. J. Cannon, A. Davidson, and F. Bédard, "Crop yield forecasting on the canadian prairies by remotely sensed vegetation indices and machine learning methods," *Agricultural Forest Meteorol.*, vol. 218–219, pp. 74–84, 2016.

[12] D. Gómez, P. Salvador, J. Sanz, and J. L. Casanova, "Potato yield prediction using machine learning techniques and sentinel 2 data," *Remote Sens.*, vol. 11, no. 15, 2019, Art. no. 1745.

[13] G. Konapala, S. V. Kumar, and S. Khalique Ahmad, "Exploring Sentinel-1 and Sentinel-2 diversity for flood inundation mapping using deep learning," *ISPRS J. Photogrammetry Remote Sens.*, vol. 180, pp. 163–173, 2021.

[14] D. Lahat, T. Adali, and C. Jutten, "Multimodal data fusion: An overview of methods, challenges, and prospects," *Proc. IEEE*, vol. 103, no. 9, pp. 1449–1477, Sep. 2015.

[15] D. Ienco, R. Interdonato, R. Gaetano, and D. Ho Tong Minh, "Combining Sentinel-1 and Sentinel-2 satellite image time series for land cover mapping via a multi-source deep learning architecture," *ISPRS J. Photogrammetry Remote Sens.*, vol. 158, pp. 11–22, 2019.

[16] I. Sa, Z. Ge, F. Dayoub, B. Upcroft, T. Perez, and C. McCool, "DeepFruits: A fruit detection system using deep neural networks," *Sensors*, vol. 16, no. 8, 2016, Art. no. 1222.

[17] N. Audebert, B. Le Saux, and S. Lefèvre, "Semantic segmentation of earth observation data using multimodal and multi-scale deep networks," in *Proc. Asian Conf. Comput. Vis.*, 2017, pp. 180–196.

[18] N. Audebert, B. Le Saux, and S. Lefèvre, "Beyond RGB: Very high resolution urban remote sensing with multimodal deep networks," *ISPRS J. Photogrammetry Remote Sens.*, vol. 140, pp. 20–32, 2018.

[19] R. Nijhawan, H. Sharma, H. Sahni, and A. Batra, "A deep learning hybrid CNN framework approach for vegetation cover mapping using deep features," in *Proc. Int. Conf. Signal-Image Technol. Internet-Based Syst.*, 2017, pp. 192–196.

[20] J. Bleiholder and F. Naumann, "Data fusion," *ACM Comput. Surv.*, vol. 41, no. 1, pp. 1–41, 2009.

[21] X. Zhu, F. Cai, J. Tian, and T. K. -A. Williams, "Spatiotemporal fusion of multisource remote sensing data: Literature survey, taxonomy, principles, applications, and future directions," *Remote Sens.*, vol. 10, no. 4, 2018, Art. no. 527.

[22] G. Camps-Valls, D. Tuia, X. X. Zhu, and M. Reichstein, *Deep Learning for the Earth Sciences: A Comprehensive Approach to Remote Sensing, Climate Science, and Geosciences*. Hoboken, NJ, USA: Wiley, 2021.

[23] F. Gao, J. Masek, M. Schwaller, and F. Hall, "On the blending of the Landsat and MODIS surface reflectance: Predicting daily landsat surface reflectance," *IEEE Trans. Geosci. Remote Sens.*, vol. 44, no. 8, pp. 2207–2218, Aug. 2006.

[24] G. Scarpa, M. Gargiulo, A. Mazza, and R. Gaetano, "A CNN-based fusion method for feature extraction from sentinel data," *Remote Sens.*, vol. 10, no. 2, 2018, Art. no. 236.

[25] D. Lei, M. Bai, L. Zhang, and W. Li, "Convolution neural network with edge structure loss for spatiotemporal remote sensing image fusion," *Int. J. Remote Sens.*, vol. 43, no. 3, pp. 1015–1036, 2022.

[26] H. Shen, X. Meng, and L. Zhang, "An integrated framework for the spatio–temporal–spectral fusion of remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 12, pp. 7135–7148, Dec. 2016.

[27] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 1597–1607.

[28] Y. Li, M. Yang, and Z. Zhang, "A survey of multi-view representation learning," *IEEE Trans. Knowl. Data Eng.*, vol. 31, no. 10, pp. 1863–1883, Oct. 2019.

[29] D. Hong, J. Chanussot, N. Yokoya, J. Kang, and X. X. Zhu, "Learning-shared cross-modality representation using multispectral-LiDAR and hyperspectral data," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 8, pp. 1470–1474, Aug. 2020.

[30] K. Heidler et al., "Self-supervised audiovisual representation learning for remote sensing data," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 116, 2023, Art. no. 103130.

[31] B. Waske and J. A. Benediktsson, "Fusion of support vector machines for classification of multisensor data," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 12, pp. 3858–3866, Dec. 2007.

[32] Y. Chen, C. Li, P. Ghamisi, X. Jia, and Y. Gu, "Deep fusion of remote sensing data for accurate classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 8, pp. 1253–1257, Aug. 2017.

[33] L. E. Cué La Rosa, D. A. B. Oliveira, and R. Q. Feitosa, "Investigating fusion strategies on encoder-decoder networks for crop segmentation using SAR and optical image sequences," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2021, pp. 2405–2408.

[34] D. Hong et al., "More diverse means better: Multimodal deep learning meets remote-sensing imagery classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 4340–4354, May 2021.

[35] X. Wu, D. Hong, and J. Chanussot, "Convolutional neural networks for multimodal remote sensing data classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Nov. 2022, Art. no. 5517010.

[36] V. Sainte Fare, L. Garnot Landrieu, and N. Chehata, "Multi-modal temporal attention models for crop mapping from satellite time series," *ISPRS J. Photogrammetry Remote Sens.*, vol. 187, pp. 294–305, 2022.

[37] S. Sun, "A survey of multi-view machine learning," *Neural Comput. Appl.*, vol. 23, no. 7, pp. 2031–2038, 2013.

[38] J. Zhao, X. Xie, X. Xu, and S. Sun, "Multi-view learning overview: Recent progress and new challenges," *Inf. Fusion*, vol. 38, pp. 43–54, 2017.

[39] D. Ramachandram and G. W. Taylor, "Deep multimodal learning: A survey on recent advances and trends," *IEEE Signal Process. Mag.*, vol. 34, no. 6, pp. 96–108, Nov. 2017.

[40] X. Yan, S. Hu, Y. Mao, Y. Ye, and H. Yu, "Deep multi-view learning methods: A review," *Neurocomputing*, vol. 448, pp. 106–129, 2021.

[41] L. Gómez-Chova, D. Tuia, G. Moser, and G. Camps-Valls, "Multimodal classification of remote sensing images: A review and future directions," *Proc. IEEE*, vol. 103, no. 9, pp. 1560–1584, Sep. 2015.

[42] S. Salcedo-Sanz et al., "Machine learning information fusion in earth observation: A comprehensive review of methods, applications and data sources," *Inf. Fusion*, vol. 63, pp. 256–272, 2020.

[43] J. Li et al., "Deep learning in multimodal remote sensing data fusion: A comprehensive review," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 112, 2022, Art. no. 102926.

[44] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Adv. Neural Inf. Process. Syst.*, vol. 25, pp. 1097–1105, 2012.

[45] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Representation*, 2015.

[46] J. R. Kettenring, "Canonical analysis of several sets of variables," *Biometrika*, vol. 58, no. 3, pp. 433–451, 1971.

[47] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *Proc. Int. Conf. Mach. Learn.*, 2013, pp. 1247–1255.

[48] I. Muslea, S. Minton, and C. A. Knoblock, "Active semi-supervised learning = Robust multi-view learning," in *Proc. Int. Conf. Mach. Learn.*, 2002, pp. 435–442.

[49] F. P. S. Luus, B. P. Salmon, F. van den Bergh, and B. T. J. Maharaj, "Multiview deep learning for land-use classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 12, pp. 2448–2452, Dec. 2015.

[50] V. Stojnic and V. Risojevic, "Self-supervised learning of remote sensing scene representations using contrastive multiview coding," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 1182–1191.

[51] J. Benediktsson, P. Swain, and O. Ersoy, "Neural network approaches versus statistical methods in classification of multisource remote sensing data," in *Proc. Can. Symp. Remote Sens. Geosci. Remote Sens. Symp.*, 1989, pp. 489–492.

[52] H. Nemmour and Y. Chibani, "Multiple support vector machines for land cover change detection: An application for mapping urban extensions," *ISPRS J. Photogrammetry Remote Sens.*, vol. 61, no. 2, pp. 125–133, 2006.

[53] L. Gomez-Chova, D. Fernández-Prieto, J. Calpe, E. Soria, J. Vila, and G. Camps-Valls, "Urban monitoring using multi-temporal SAR and multi-spectral data," *Pattern Recognit. Lett.*, vol. 27, no. 4, pp. 234–243, 2006.

[54] G. Ruß, R. Kruse, M. Schneider, and P. Wagner, "Data mining with neural networks for wheat yield prediction," in *Advances in Data Mining. Medical Applications, E-Commerce, Marketing, and Theoretical Aspects*(Lecture Notes in Computer Science). Berlin, Germany: Springer, 2008, pp. 47–56.

[55] D. Marmanis, J. D. Wegner, S. Galliani, K. Schindler, M. Datcu, and U. Stilla, "Semantic segmentation of aerial images with an ensemble of CNSS," in *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. III-3. Göttingen, Germany: Copernicus Publications, 2016, pp. 473–480.

[56] F. Castanedo, "A review of data fusion techniques," *Sci. World J.*, vol. 2013, pp. 704504–704504, 2013.

[57] C. M. Christoudias, R. Urtasun, and T. Darrell, "Multi-view learning in the presence of view disagreement," in *Proc. Conf. Uncertainty Artif. Intell.*, 2008, pp. 88–96.

[58] P. Zhang et al., "A hybrid attention-aware fusion network (HAFNet) for building extraction from high-resolution imagery and LiDAR data," *Remote Sens.*, vol. 12, no. 22, 2020, Art. no. 3764.

[59] D. Hong, L. Gao, R. Hang, B. Zhang, and J. Chanussot, "Deep encoder–decoder networks for classification of hyperspectral and LiDAR data," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, Aug. 2022, Art. no. 5500205.

[60] L. Zhang, Z. Zhang, Y. Luo, J. Cao, and F. Tao, "Combining optical, fluorescence, thermal satellite, and environmental data to predict county-level maize yield in China using machine learning approaches," *Remote Sens.*, vol. 12, no. 1, 2020, Art. no. 21.

[61] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton, "Adaptive mixtures of local experts," *Neural Comput.*, vol. 3, no. 1, pp. 79–87, 1991.

[62] Y. Liu, S. Piramanayagam, S. T. Monteiro, and E. Saber, "Dense semantic labeling of very-high-resolution aerial imagery and LiDAR with fully-convolutional neural networks and higher-order CRFs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 76–85.

[63] K. Ahmad, K. Pogorelov, M. Riegler, N. Conci, and P. Halvorsen, "CNN and GAN based satellite and social media data fusion for disaster detection," in *Proc. Conf. MediaEval Benchmarking Initiative Multimedia Eval.*, 2017.

[64] T. Liu and A. Abd-Elrahman, "Deep convolutional neural network training enrichment using multi-view object-based analysis of unmanned aerial systems imagery for wetlands classification," *ISPRS J. Photogrammetry Remote Sens.*, vol. 139, pp. 154–170, 2018.

[65] W. Zhang, H. Huang, M. Schmitz, X. Sun, H. Wang, and H. Mayer, "Effective fusion of multi-modal remote sensing data in a fully convolutional network for semantic labeling," *Remote Sens.*, vol. 10, no. 1, 2018, Art. no. 52.

[66] S. Valero, L. Arnaud, M. Planells, E. Ceschia, and G. Dedieu, "Sentinel's classifier fusion system for seasonal crop mapping," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2019, pp. 6243–6246.

[67] S. M. Jameel, M. A. Hashmani, M. Rehman, and A. Budiman, "Adaptive CNN ensemble for complex multispectral image analysis," *Complexity*, vol. 2020, pp. 1–21, 2020.

[68] C. Robinson et al., "Global land-cover mapping with weak supervision: Outcome of the 2020 IEEE GRSS data fusion contest," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 3185–3199, 2021.

[69] Y. Ma et al., "The outcome of the 2021 IEEE GRSS data fusion contest - track DSE: Detection of settlements without electricity," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 12375–12385, Nov. 2021.

[70] D. Rashkovetsky, F. Mauracher, M. Langer, and M. Schmitt, "Wildfire detection from multisensor satellite imagery using deep semantic segmentation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 7001–7016, Jun. 2021.

[71] M. F. Danilevicz, P. E. Bayer, F. Boussaid, M. Bennamoun, and D. Edwards, "Maize yield prediction at an early developmental stage using multispectral images and genotype data for preliminary hybrid selection," *Remote Sens.*, vol. 13, no. 19, 2021, Art. no. 3976.

[72] Z. Li et al., "The outcome of the 2021 IEEE GRSS data fusion contest-track MSD: Multitemporal semantic change detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 1643–1655, Jan. 2022.

[73] P. J. Mitchell, F. Waldner, H. Horan, J. N. Brown, and Z. Hochman, "Data fusion using climatology and seasonal climate forecasts improves estimates of australian national wheat yields," *Agricultural Forest Meteorol.*, vol. 320, 2022, Art. no. 108932.

[74] T. Gangopadhyay, J. Shook, A. K. Singh, and S. Sarkar, "Deep time series attention models for crop yield prediction and insights," in *Proc. Neural Inf. Process. Syst. Workshop Mach. Learn. Phys. Sci.*, 2019.

[75] X. Wang, J. Huang, Q. Feng, and D. Yin, "Winter wheat yield prediction at county level and uncertainty analysis in main wheat-producing regions of China with deep learning approaches," *Remote Sens.*, vol. 12, no. 11, 2020, Art. no. 1744.

[76] J. Irvin et al., "ForestNet: Classifying drivers of deforestation in Indonesia using deep learning on satellite imagery," *Adv. Neural Inf. Process. Syst.*, vol. 34, 2020.

[77] J. Cao et al., "Integrating multi-source data for rice yield prediction across China using machine learning and deep learning approaches," *Agricultural Forest Meteorol.*, vol. 297, 2021, Art. no. 108275.

[78] H. Kamangir et al., "FogNet: A multiscale 3D CNN with double-branch dense block and attention mechanism for fog prediction," *Mach. Learn. Appl.*, vol. 5, 2021, Art. no. 100038.

[79] A.K. Srivastava et al., "Winter wheat yield prediction using convolutional neural networks from environmental and phenological data," *Sci. Rep.*, vol. 12, no. 1, 2022, Art. no. 3215.

[80] X. Wang, Y. Zhang, T. Lei, Y. Wang, Y. Zhai, and A. K. Nandi, "Dynamic convolution self-attention network for land-cover classification in VHR remote-sensing images," *Remote Sens.*, vol. 14, no. 19, 2022, Art. no. 4941.

[81] R. Hang, Z. Li, P. Ghamisi, D. Hong, G. Xia, and Q. Liu, "Classification of hyperspectral and LiDAR data using coupled CNNs," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 7, pp. 4939–4950, Jul. 2020.

[82] V. Vielzeuf, A. Lechervy, S. Pateux, and F. Jurie, "CentralNet: A multilayer approach for multimodal fusion," in *Proc. Eur. Conf. Comput. Vis. Workshops*, 2018, pp. 575–589.

[83] J. Pei, Y. Huang, W. Huo, Y. Zhang, J. Yang, and T.-S. Yeo, "SAR automatic target recognition based on multiview deep learning framework," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 4, pp. 2196–2210, Apr. 2018.

[84] C. Peng, Y. Li, L. Jiao, Y. Chen, and R. Shang, "Densely based multi-scale and multi-modal fully convolutional networks for high-resolution remote-sensing image semantic segmentation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 8, pp. 2612–2626, Aug. 2019.

[85] Z. Cao, W. Diao, X. Sun, X. Lyu, M. Yan, and K. Fu, "C3Net: Cross-modal feature recalibrated, cross-scale semantic aggregated and compact network for semantic segmentation of multi-modal high-resolution aerial images," *Remote Sens.*, vol. 13, no. 3, 2021, Art. no. 528.

[86] H. Hosseinpour, F. Samadzadegan, and F. D. Javan, "CMGFNet: A deep cross-modal gated fusion network for building extraction from very high-resolution remote sensing images," *ISPRS J. Photogrammetry Remote Sens.*, vol. 184, pp. 96–115, 2022.

[87] X. Wang, Y. Feng, R. Song, Z. Mu, and C. Song, "Multi-attentive hierarchical dense fusion net for fusion classification of hyperspectral and LiDAR data," *Inf. Fusion*, vol. 82, pp. 1–18, 2022.

[88] J. Zhao et al., “Multi-source collaborative enhanced for remote sensing images semantic segmentation,” *Neurocomputing*, vol. 493, pp. 76–90, 2022.

[89] W. Zhou, J. Jin, J. Lei, and J.-N. Hwang, “CEGFNet: Common extraction and gate fusion network for scene parsing of remote sensing images,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–10, Sep. 2022.

[90] J. Arevalo, T. Solorio, M. Montes-y Gómez, and F. A. González, “Gated multimodal units for information fusion,” in *Proc. Int. Conf. Learn. Representation Workshop*, 2017.

[91] J. Arevalo, T. Solorio, M. Montes-y Gómez, and F. A. González, “Gated multimodal networks,” *Neural Comput. Appl.*, vol. 32, no. 14, pp. 10209–10228, 2020.

[92] X. Zheng, X. Wu, L. Huan, W. He, and H. Zhang, “A gather-to-guide network for remote sensing semantic segmentation of RGB and auxiliary image,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–15, Aug. 2022.

[93] J. Wang, J. Li, Y. Shi, J. Lai, and X. Tan, “AM ^{3} Net: Adaptive mutual-learning-based multimodal data fusion network,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 8, pp. 5411–5426, Aug. 2022.

[94] S. Mohla, S. Pande, B. Banerjee, and S. Chaudhuri, “FusAtNet: Dual attention based spectrospatial multimodal fusion network for hyperspectral and LiDAR classification,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshop*, 2020, pp. 416–425.

[95] J. Wang, W. Li, Y. Gao, M. Zhang, R. Tao, and Q. Du, “Hyperspectral and SAR image classification via multiscale interactive fusion network,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 12, pp. 10823–10837, Dec. 2023.

[96] S. Fang, K. Li, and Z. Li, “S ^{2} ENet: Spatial–spectral cross-modal enhancement network for classification of hyperspectral and LiDAR data,” *IEEE Geosci. Remote Sens. Lett.*, vol. 19, Oct. 2022, Art. no. 6504205.

[97] F. Qingyun and W. Zhaokui, “Cross-modality attentive feature fusion for object detection in multispectral remote sensing imagery,” *Pattern Recognit.*, vol. 130, 2022, Art. no. 108786.

[98] S. Cummings, L. Kondmann, and X. X. Zhu, “Siamese attention U-net for multi-class change detection,” in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2022, pp. 211–214.

[99] C. Rambour, N. Audebert, E. Koeniguer, B. Le Saux, M. Crucianu, and M. Datcu, “Flood detection in time series of optical and SAR images,” in *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. XLIII-B2-2020. Göttingen, Germany: Copernicus GmbH, 2020, pp. 1343–1346.

[100] C. Wang, X. Liu, J. Pei, Y. Huang, Y. Zhang, and J. Yang, “Multiview attention CBB-LSTM network for SAR automatic target recognition,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 12504–12513, Nov. 2021.

[101] S. Srivastava, J. E. Vargas Muñoz, S. Lobry, and D. Tuia, “Fine-grained landuse characterization using ground-based pictures: A deep learning solution based on globally available data,” *Int. J. Geographical Inf. Sci.*, vol. 34, no. 6, pp. 1117–1136, 2020.

[102] D. Hong, J. Hu, J. Yao, J. Chanussot, and X. X. Zhu, “Multimodal remote sensing benchmark datasets for land cover classification with a shared and specific feature learning model,” *ISPRS J. Photogrammetry Remote Sens.*, vol. 178, pp. 68–80, 2021.

[103] L. H. Nguyen et al., “Spatial-temporal multi-task learning for within-field cotton yield prediction,” in *Advances in Knowledge Discovery and Data Mining* (Lecture Notes in Computer Science). Berlin, Germany: Springer, 2019, pp. 343–354.

[104] M. Maimaitijiang, V. Sagan, P. Sidike, S. Hartling, F. Esposito, and F. B. Fritschi, “Soybean yield prediction from UAV using multimodal data fusion and deep learning,” *Remote Sens. Environ.*, vol. 237, 2020, Art. no. 111599.

[105] I. E. Livieris, S. D. Dafnis, G. K. Papadopoulos, and D. P. Kalivas, “A multiple-input neural network model for predicting cotton production quantity: A case study,” *Algorithms*, vol. 13, no. 11, 2020, Art. no. 273.

[106] K. K. Gadiraju, B. Ramachandra, Z. Chen, and R. R. Vatsavai, “Multimodal deep learning based crop classification using multispectral and multitemporal satellite imagery,” in *Proc. ACM Int. Conf. Knowl. Discov. Data Mining*, 2020, pp. 3234–3242.

[107] G. J. Scott, K. C. Hagan, R. A. Marcum, J. A. Hurt, D. T. Anderson, and C. H. Davis, “Enhanced fusion of deep neural networks for classification of benchmark high-resolution image data sets,” *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 9, pp. 1451–1455, Sep. 2018.

[108] A. K. Reyes, J. C. Caicedo, and J. E. Camargo, “Fine-tuning deep convolutional networks for plant recognition,” in *Proc. Conf. Labs Eval. Forum*, 2015, vol. 1391, pp. 467–475.

[109] S. H. Lee, C. S. Chan, P. Wilkin, and P. Remagnino, “Deep-plant: Plant identification with convolutional neural networks,” in *Proc. IEEE Int. Conf. Image Process.*, 2015, pp. 452–456.

[110] S. P. Mohanty, D. P. Hughes, and M. Salathé, “Using deep learning for image-based plant disease detection,” *Front. Plant Sci.*, vol. 7, 2016, Art. no. 1419.

[111] H. Yalcin, “Phenology recognition using deep learning,” in *Proc. Electric Electron.Comput. Sci. Biomed. Eng. Meeting*, 2018, pp. 1–5.

[112] J. Valente, M. Doldersum, C. Roers, and L. Kooistra, “Deteting rumex obtusifolius weed plants in grasslands from UAV RGB imagery using deep learning,” in *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. IV-2-W5. Göttingen, Germany: Copernicus GmbH, 2019, pp. 179–185.

[113] X. Li, L. Lei, Y. Sun, M. Li, and G. Kuang, “Multimodal bilinear fusion network with second-order attention-based channel selection for land cover classification,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 1011–1026, Mar. 2020.

[114] R. Chew et al., “Deep neural networks and transfer learning for food crop identification in UAV images,” *Drones*, vol. 4, no. 1, 2020, Art. no. 7.

[115] S. Bargoti and J. Underwood, “Deep fruit detection in orchards,” in *Proc. IEEE Int. Conf. Robot. Automat.*, 2017, pp. 3626–3633.

[116] Z. Jiang, “A novel crop weed recognition method based on transfer learning from VGG16 implemented by keras,” *IOP Conf. Series, Mater. Sci. Eng.*, vol. 677, no. 3, 2019, Art. no. 032073.

[117] S. Srivastava, J. E. Vargas-Muñoz, and D. Tuia, “Understanding urban landuse from the above and ground perspectives: A deep learning, multimodal solution,” *Remote Sens. Environ.*, vol. 228, pp. 129–143, 2019.

[118] A. Nowakowski et al., “Crop type mapping by using transfer learning,” *Int. J. Appl. Earth Observ. Geoinf.*, vol. 98, 2021, Art. no. 102313.

[119] A. Ramcharan, K. Baranowski, P. McCloskey, B. Ahmed, J. Legg, and D. P. Hughes, “Deep learning for image-based cassava disease detection,” *Front. Plant Sci.*, vol. 8, 2017, Art. no. 1852.

[120] T. G. J. Rudner et al., “Multi3Net: Segmenting flooded buildings via fusion of multiresolution, multisensor, and multitemporal satellite imagery,” in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, no. 01, pp. 702–709, 2019.

[121] K. Yuan, X. Zhuang, G. Schaefer, J. Feng, L. Guan, and H. Fang, “Deep-learning-based multispectral satellite image segmentation for water body detection,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 7422–7434, Jul. 2021.

[122] W. Lu, C. Tao, H. Li, J. Qi, and Y. Li, “A unified deep learning framework for urban functional zone extraction based on multi-source heterogeneous data,” *Remote Sens. Environ.*, vol. 270, 2022, Art. no. 112830.

[123] W. Ma, O. Karakuş, and P. L. Rosin, “AMM-FuseNet: Attention-based multi-modal image fusion network for land cover mapping,” *Remote Sens.*, vol. 14, no. 18, 2022, Art. no. 4458.

[124] X. Li et al., “MCANet: A joint semantic segmentation framework of optical and SAR images for land use classification,” *Int. J. Appl. Earth Observ. Geoinf.*, vol. 106, 2022, Art. no. 102638.

[125] H. Sheng, X. Chen, J. Su, R. Rajagopal, and A. Ng, “Effective data fusion with generalized vegetation index: Evidence from land cover segmentation in agriculture,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 60–61.

[126] T. Di Martino, M. Lenormand, and E. C. Koeniguer, “Multi-branch deep learning model for detection of settlements without electricity,” in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2021, pp. 1847–1850.

[127] P. Bosilj, E. Aptoula, T. Duckett, and G. Cielniak, “Transfer learning between crop types for semantic segmentation of crops versus weeds in precision agriculture,” *J. Field Robot.*, vol. 37, no. 1, pp. 7–19, 2020.

[128] C. Feichtenhofer, A. Pinz, and A. Zisserman, “Convolutional two-stream network fusion for video action recognition,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 1933–1941.

[129] J.-M. Perez-Rua, V. Vielzeuf, S. Pateux, M. Baccouche, and F. Jurie, “MFAS: Multimodal fusion architecture search,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 6966–6975.

[130] A. X. Wang, C. Tran, N. Desai, D. Lobell, and S. Ermon, “Deep transfer learning for crop yield prediction with remote sensing data,” in *Proc. 1st ACM SIGCAS Conf. Comput. Sustain. Soc.*, 2018, pp. 1–5.

[131] S. Khaki, L. Wang, and S. V. Archontoulis, “A CNN-RNN framework for crop yield prediction,” *Front. Plant Sci.*, vol. 10, 2020, Art. no. 1750.

[132] G. Sahu and O. Vechtomova, “Adaptive fusion techniques for multimodal data,” in *Proc. Conf. Eur. Chapter Assoc. Comput. Linguistics*, 2021, pp. 3156–3166.

[133] X. Xu, W. Li, Q. Ran, Q. Du, L. Gao, and B. Zhang, "Multisource remote sensing data classification based on convolutional neural network," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 2, pp. 937–949, Feb. 2018.

[134] Q. Feng, D. Zhu, J. Yang, and B. Li, "Multisource hyperspectral and LiDAR data fusion for urban land-use mapping based on a modified two-branch convolutional neural network," *ISPRS Int. J. Geo-Inf.*, vol. 8, no. 1, 2019, Art. no. 28.

[135] W. Li, Y. Gao, M. Zhang, R. Tao, and Q. Du, "Asymmetric feature fusion network for hyperspectral and SAR image classification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 10, pp. 8057–8070, Oct. 2023.

[136] H. Zhang, J. Yao, L. Ni, L. Gao, and M. Huang, "Multimodal attention-aware convolutional neural networks for classification of hyperspectral and LiDAR data," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 3635–3644, Jul. 2023.

[137] M. Zhang, W. Li, R. Tao, H. Li, and Q. Du, "Information fusion for classification of hyperspectral and LiDAR data using IP-CNN," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Jul. 2022, Art. no. 5506812.

[138] H. Gao, Z. Chen, and F. Xu, "Adaptive spectral-spatial feature fusion network for hyperspectral image classification using limited training samples," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 107, 2022, Art. no. 102687.

[139] S. Falahatnejad and A. Karami, "Deep fusion of hyperspectral and LiDAR images using attention-based CNN," *SN Comput. Sci.*, vol. 4, no. 1, 2022, Art. no. 1.

[140] Q. Yang, L. Shi, J. Han, Y. Zha, and P. Zhu, "Deep convolutional neural networks for rice grain yield estimation at the ripening stage using UAV-based remotely sensed images," *Field Crops Res.*, vol. 235, pp. 142–153, 2019.

[141] D. M. Johnson, "An assessment of pre- and within-season remotely sensed variables for forecasting corn and soybean yields in the United States," *Remote Sens. Environ.*, vol. 141, pp. 116–128, 2014.

[142] A. T. M. S. Ahamed et al., "Applying data mining techniques to predict annual yield of major crops and recommend planting different crops in different districts in Bangladesh," in *Proc. Int. Conf. Softw. Eng. Artif. Intell. Netw. Parallel/Distrib. Comput.*, 2015, pp. 1–6.

[143] J. Inglada, A. Vincent, M. Arias, and C. Marais-Sicre, "Improved early crop type identification by joint use of high temporal resolution SAR and optical image time series," *Remote Sens.*, vol. 8, no. 5, 2016, Art. no. 362.

[144] F. F. Bocca and L. H. A. Rodrigues, "The effect of tuning, feature engineering, and feature selection in data mining applied to rainfed sugarcane yield modelling," *Comput. Electron. Agriculture*, vol. 128, pp. 67–76, 2016.

[145] G. Niedbała, "Simple model based on artificial neural network for early prediction and simulation winter rapeseed yield," *J. Integrative Agriculture*, vol. 18, no. 1, pp. 54–61, 2019.

[146] Y. Cai et al., "Integrating satellite and climate data to predict wheat yield in Australia using machine learning approaches," *Agricultural Forest Meteorol.*, vol. 274, pp. 144–159, 2019.

[147] B. Peng et al., "Assessing the benefit of satellite-based solar-induced chlorophyll fluorescence in crop yield prediction," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 90, 2020, Art. no. 102126.

[148] S. H. Bhojani and N. Bhatt, "Wheat crop yield prediction using new activation functions in neural network," *Neural Comput. Appl.*, vol. 32, no. 17, pp. 13941–13951, 2020.

[149] V. Sagan et al., "Field-scale crop yield prediction using multi-temporal WorldView-3 and PlanetScope satellite data and deep learning," *ISPRS J. Photogrammetry Remote Sens.*, vol. 174, pp. 265–281, 2021.

[150] J. You, X. Li, M. Low, D. Lobell, and S. Ermon, "Deep Gaussian process for crop yield prediction based on remote sensing data," in *Proc. AAAI Conf. Artif. Intell.*, 2017, pp. 4559–4565.

[151] H. Russello and W. Shang, "Convolutional neural networks for crop yield prediction using satellite," 2018. [Online]. Available: https://scripties.uba.uva.nl/document/658789?setlang=en

[152] Z. Jiang, C. Liu, N. P. Hendricks, B. Ganapathysubramanian, D. J. Hayes, and S. Sarkar, "Predicting county level corn yields using deep long short term memory models," 2018, *arXiv:1805.12044*.

[153] J. Sun, L. Di, Z. Sun, Y. Shen, and Z. Lai, "County-level soybean yield prediction using deep CNN-LSTM model," *Sensors*, vol. 19, no. 20, 2019, Art. no. 4363.

[154] H. Jiang et al., "A deep learning approach to conflating heterogeneous geospatial data for corn yield estimation: A case study of the US corn belt at the county level," *Glob. Change Biol.*, vol. 26, no. 3, pp. 1754–1766, 2020.

[155] T. Lin et al., "DeepCropNet: A deep spatial-temporal learning framework for county-level corn yield estimation," *Environ. Res. Lett.*, vol. 15, no. 3, 2020, Art. no. 034016.

[156] S. Khaki, H. Pham, and L. Wang, "Simultaneous corn and soybean yield prediction from remote sensing data using deep transfer learning," *Sci. Rep.*, vol. 11, no. 1, 2021, Art. no. 11132.

[157] H. Tian et al., "A deep learning framework under attention mechanism for wheat yield estimation using remotely sensed indices in the guanzhong plain, PR China," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 102, 2021, Art. no. 102375.

[158] K. Gavahi, P. Abbaszadeh, and H. Moradkhani, "DeepYield: A combined convolutional neural network with long short-term memory for crop yield forecasting," *Expert Syst. Appl.*, vol. 184, 2021, Art. no. 115511.

[159] D. Dobrinić, M. Gašparović, and D. Medak, "Sentinel-1 and 2 time-series for vegetation mapping using random forest classification: A case study of northern Croatia," *Remote Sens.*, vol. 13, no. 12, 2021, Art. no. 2321.

[160] R. M. Rustowicz, R. Cheong, L. Wang, S. Ermon, M. Burke, and D. Lobell, "Semantic segmentation of crop type in Africa: A novel dataset and analysis of deep learning methods," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2019, pp. 75–82.

[161] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Representations*, 2021.

[162] V. S. F. Garnot, L. Landrieu, S. Giordano, and N. Chehata, "Satellite image time series classification with pixel-set encoders and temporal self-attention," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 12325–12334.

[163] F. Weilandt et al., "Early crop classification via multi-modal satellite data fusion and temporal attention," *Remote Sens.*, vol. 15, no. 3, 2023, Art. no. 799.

[164] M. Rußwurm and M. Körner, "Multi-temporal land cover classification with sequential recurrent encoders," *ISPRS Int. J. Geo-Inf.*, vol. 7, no. 4, 2018, Art. no. 129.

[165] A. Stergioulas, K. Dimitropoulos, and N. Grammalidis, "Crop classification from satellite image sequences using a two-stream network with temporal self-attention," in *Proc. IEEE Int. Conf. Imag. Syst. Techn.*, 2022, pp. 1–6.

[166] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.

[167] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 448–456.

[168] Y. Wang, W. Huang, F. Sun, T. Xu, Y. Rong, and J. Huang, "Deep multimodal fusion by channel exchanging," *Adv. Neural Inf. Process. Syst.*, vol. 33, pp. 4835–4845, 2020.

[169] C. Debes et al., "Hyperspectral and LiDAR data fusion: Outcome of the 2013 GRSS data fusion contest," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 6, pp. 2405–2418, Jun. 2014.

[170] M. Khodadadzadeh, J. Li, S. Prasad, and A. Plaza, "Fusion of hyperspectral and LiDAR remote sensing data using multiple feature learning," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 6, pp. 2971–2983, Jun. 2015.

[171] P. Ghamisi, B. Höfle, and X. X. Zhu, "Hyperspectral and LiDAR data fusion using extinction profiles and deep convolutional neural network," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 6, pp. 3011–3024, Jun. 2017.

[172] J. Denize, L. Hubert-Moy, J. Betbeder, S. Corgne, J. Baudry, and E. Pottier, "Evaluation of using Sentinel-1 and -2 time-series to identify winter land use in agricultural landscapes," *Remote Sens.*, vol. 11, no. 1, 2019, Art. no. 37.

[173] J. Li et al., "Fusion of optical and SAR images based on deep learning to reconstruct vegetation NDVI time series in cloud-prone regions," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 112, 2022, Art. no. 102818.

[174] D. F. Mantsis et al., "Multimodal fusion of Sentinel 1 images and social media data for snow depth estimation," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, Oct. 2022, Art. no. 4004105.

[175] Q. Liu, M. C. Kampffmeyer, R. Jenssen, and A.-B. Salberg, "Multi-view self-constructing graph convolutional networks with adaptive class weighting loss for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2020, pp. 44–45.

[176] M. Tom, Y. Jiang, E. Baltsavias, and K. Schindler, "Learning a joint embedding of multiple satellite sensors: A case study for lake ice monitoring," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Sep. 2022, Art. no. 4306315.

[177] W. G. C. Bandara and V. M. Patel, "A transformer-based siamese network for change detection," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2022, pp. 207–210.

[178] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah, "Signature verification using a "siamese" time delay neural network," in *Proc. Adv. Neural Inf. Process. Syst.*, 1993, pp. 737–744.

[179] D. Chicco, "Siamese neural networks: An overview," in *Artificial Neural Networks* (Methods in Molecular Biology). Berlin, Germany: Springer, 2021, pp. 73–94.

[180] J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny, "Barlow twins: Self-supervised learning via redundancy reduction," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 12 310–12320.

[181] H. Huang et al., "Developing a dual-stream deep-learning neural network model for improving county-level winter wheat yield estimates in China," *Remote Sens.*, vol. 14, no. 20, 2022, Art. no. 5280.

[182] P. Benedetti, D. Ienco, R. Gaetano, K. Ose, R. G. Pensa, and S. Dupuy, "M ^{3} Fusion: A deep learning architecture for multiscale multimodal multitemporal satellite data fusion," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 12, pp. 4939–4949, Dec. 2018.

[183] H. Li et al., "A multi-sensor fusion framework based on coupled residual convolutional neural networks," *Remote Sens.*, vol. 12, no. 12, 2020, Art. no. 2067.

[184] M. Kampffmeyer, A.-B. Salberg, and R. Jenssen, "Urban land cover classification with missing data modalities using deep convolutional neural networks," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 6, pp. 1758–1768, Jun. 2018.

[185] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, Springer, 2018, pp. 3–19.

[186] C. Shi, D. Liao, T. Zhang, and L. Wang, "Hyperspectral image classification based on 3D coordination attention mechanism network," *Remote Sens.*, vol. 14, no. 3, 2022, Art. no. 608.

[187] S. De Alwis, Y. Zhang, M. Na, and G. Li, "Duo attention with deep learning on tomato yield prediction and factor interpretation," in *Proc. Pacific Rim Int. Conf. Artif. Intell.*, 2019, pp. 704–715.

[188] P. Ebel, V. S. F. Garnot, M. Schmitt, J. D. Wegner, and X. X. Zhu, "UnCRtainTS: Uncertainty quantification for cloud removal in optical satellite time series," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 2085–2095.

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

[190] F. Ferrari, M. P. Ferreira, C. A. Almeida, and R. Q. Feitosa, "Fusing Sentinel-1 and Sentinel-2 images for deforestation detection in the Brazilian amazon under diverse cloud conditions," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, Feb. 2023, Art. no. 2501005.

[191] G. I. Drakonakis, G. Tsagkatakis, K. Fotiadou, and P. Tsakalides, "Ombrianet-supervised flood mapping via convolutional neural networks using multitemporal Sentinel-1 and Sentinel-2 data fusion," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 2341–2356, 2022.

[192] X. Zhang, S. Cheng, L. Wang, and H. Li, "Asymmetric cross-attention hierarchical network based on CNN and transformer for bitemporal remote sensing images change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Feb. 2023, Art. no. 2000415.

[193] P. Ebel, Y. Xu, M. Schmitt, and X. X. Zhu, "SEN12MS-CR-TS: A remote-sensing data set for multimodal multitemporal cloud removal," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Jan. 2022, Art. no. 5222414.

[194] Y. Ye, M. Wang, L. Zhou, G. Lei, J. Fan, and Y. Qin, "Adjacent-level feature cross-fusion with 3D CNN for remote sensing image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Aug. 2023, Art. no. 5618214.

[195] N. Yokoya et al., "Open data for global multimodal land use classification: Outcome of the 2017 IEEE GRSS data fusion contest," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 5, pp. 1363–1377, May 2018.

[196] Y. Chen, Z. Lin, X. Zhao, G. Wang, and Y. Gu, "Deep learning-based classification of hyperspectral data," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 6, pp. 2094–2107, Jun. 2014.

[197] N. Kussul, M. Lavreniuk, S. Skakun, and A. Shelestov, "Deep learning classification of land cover and crop types using remote sensing data," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 5, pp. 778–782, May 2017.

[198] X. Zhao et al., "Joint classification of hyperspectral and LiDAR data using hierarchical random walk and deep CNN architecture," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 10, pp. 7355–7370, Oct. 2020.

[199] N. Torbick, X. Huang, B. Ziniti, D. Johnson, J. Masek, and M. Reba, "Fusion of moderate resolution earth observations for operational crop type mapping," *Remote Sens.*, vol. 10, no. 7, 2018, Art. no. 1058.

[200] A. Sebastianelli, M. P. Del Rosso, P. P. Mathieu, and S. L. Ullo, "Paradigm selection for data fusion of SAR and multispectral sentinel data applied to land-cover classification," 2021, *arXiv:2106.11056*.

[201] X.-P. Song, W. Huang, M. C. Hansen, and P. Potapov, "An evaluation of landsat, Sentinel-2, Sentinel-1 and MODIS data for crop type mapping," *Sci. Remote Sens.*, vol. 3, 2021, Art. no. 100018.

[202] L. Zhao and S. Ji, "CNN, RNN, or ViT? An evaluation of different deep learning architectures for spatio-temporal representation of sentinel time series," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 44–56, Nov. 2022.

[203] Y. Xu et al., "Advanced multi-sensor optical remote sensing for urban land use and land cover classification: Outcome of the 2018 IEEE GRSS data fusion contest," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 6, pp. 1709–1724, Jun. 2019.

[204] P. Nevavuori, N. Narra, and T. Lipping, "Crop yield prediction with deep convolutional neural networks," *Comput. Electron. Agriculture*, vol. 163, 2019, Art. no. 104859.

[205] C.-A. Diaconu, S. Saha, S. Günnemann, and X. X. Zhu, "Understanding the role of weather data for earth surface forecasting using a ConvLSTM-based model," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 1362–1371.

[206] J. He et al., "Accurate estimation of the proportion of mixed land use at the street-block level by integrating high spatial resolution images and geospatial Big Data," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 8, pp. 6357–6370, Aug. 2021.

[207] N. Said et al., "Natural disasters detection in social media and satellite imagery: A survey," *Multimedia Tools Appl.*, vol. 78, no. 22, pp. 31267–31302, 2019.

[208] S. Cresci, M. Avvenuti, M. La Polla, C. Meletti, and M. Tesconi, "Nowcasting of earthquake consequences using big social data," *IEEE Internet Comput.*, to be published, doi: 10.1109/MIC.2017.265102211.

[209] B. Bischke, P. Helber, Z. Zhao, J. de Bruijn, and D. Borth, "The multimedia satellite task at MediaEval 2018 emergency response for flooding events," in *Proc. MediaEval Workshop*, 2018.

[210] M. Shahhosseini, G. Hu, S. Khaki, and S. V. Archontoulis, "Corn yield prediction with ensemble CNN-DNN," *Front. Plant Sci.*, vol. 12, 2021, Art. no. 709008.

[211] Z. Chu and J. Yu, "An end-to-end model for rice yield prediction using deep learning fusion," *Comput. Electron. Agriculture*, vol. 174, 2020, Art. no. 105471.

[212] G. Camps-Valls, L. Gomez-Chova, J. Munoz-Mari, J. L. Rojo-Alvarez, and M. Martinez-Ramon, "Kernel-based framework for multitemporal and multisource remote sensing data classification and change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 6, pp. 1822–1835, Jun. 2008.

[213] X. Yang, W. Liu, D. Tao, J. Cheng, and S. Li, "Multiview canonical correlation analysis networks for remote sensing image recognition," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 10, pp. 1855–1859, Oct. 2017.

[214] M. Rußwurm, N. Courty, R. Emonet, S. Lefèvre, D. Tuia, and R. Tavenard, "End-to-end learned early classification of time series for in-season crop type mapping," *ISPRS J. Photogrammetry Remote Sens.*, vol. 196, pp. 445–456, 2023.

[215] E. Georgiou, C. Papaioannou, and A. Potamianos, "Deep hierarchical fusion with application in sentiment analysis," in *Proc. Interspeech*, 2019, pp. 1646–1650.

[216] Q. Wang, C. Boudreau, Q. Luo, P.-N. Tan, and J. Zhou, "Deep multi-view information bottleneck," in *Proc. SIAM Int. Conf. Data Mining*, 2019, pp. 37–45.

[217] Y. Pageot, F. Baup, J. Inglada, N. Baghdadi, and V. Demarez, "Detection of irrigated and rainfed crops in temperate areas using Sentinel-1 and Sentinel-2 time series," *Remote Sens.*, vol. 12, no. 18, 2020, Art. no. 3044.

[218] G. Tseng, I. Zvonkov, C. L. Nakalembe, and H. Kerner, "CropHarvest: A global dataset for crop-type classification," in *Proc. Conf. Neural Inf. Process. Syst. Datasets Benchmarks Track*, 2021.

[219] C. Hazirbas, L. Ma, C. Domokos, and D. Cremers, "FuseNet: Incorporating depth into semantic segmentation via fusion-based CNN architecture," in *Proc. Asian Conf. Comput. Vis.*, 2017, pp. 213–228.

[220] H. Shen et al., "Missing information reconstruction of remote sensing data: A technical review," *IEEE Geosci. Remote Sens. Mag.*, vol. 3, no. 3, pp. 61–85, Sep. 2015.

[221] J.-M. Pérez-Rúa, V. Vielzeuf, S. Pateux, M. Baccouche, and F. Jurie, "MFAS: Multimodal fusion architecture search," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 6966–6975.

[222] K. Sridharan and S. M. Kakade, "An information theoretic framework for multi-view learning," in *Proc. 21st Annu. Conf. Learn. Theory*, 2008, pp. 403–404.

[223] M. Kumar, N. S. Raghuwanshi, R. Singh, W. W. Wallender, and W. O. Pruitt, "Estimating evapotranspiration using artificial neural network," *J. Irrigation Drainage Eng.*, vol. 128, no. 4, pp. 224–233, 2002.

[224] S. Aksoy, K. Koperski, C. Tusk, and G. Marchisio, "Land cover classification with multi-sensor fusion of partly missing data," *Photogrammetric Eng. Remote Sens.*, vol. 75, no. 5, pp. 577–593, 2009.

[225] A.-B. Salberg and R. Jenssen, "Land-cover classification of partly missing data using support vector machines," *Int. J. Remote Sens.*, vol. 33, no. 14, pp. 4471–4481, 2012.

[226] C. Doña et al., "Integrated satellite data fusion and mining for monitoring lake water quality status of the Albufera de Valencia in Spain," *J. Environ. Manage.*, vol. 151, pp. 416–426, 2015.

[227] M. Albughdadi, D. Kouamé, G. Rieu, and J.-Y. Tourneret, "Missing data reconstruction and anomaly detection in crop development using agronomic indicators derived from multispectral satellite images," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2017, pp. 5081–5084.

[228] E. Santi, S. Paloscia, S. Pettinato, L. Brocca, L. Ciabatta, and D. Entekhabi, "On the synergy of SMAP, AMSR2 and Sentinel-1 for retrieving soil moisture," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 65, pp. 114–123, 2018.

[229] L. Landuyt, N. E. C. Verhoest, and F. M. B. Van Coillie, "Flood mapping in vegetated areas using an unsupervised clustering approach on Sentinel-1 and -2 imagery," *Remote Sens.*, vol. 12, no. 21, 2020, Art. no. 3611.

[230] G. Sumbul et al., "BigEarthNet-MM: A large-scale, multimodal, multilabel benchmark archive for remote sensing image classification and retrieval [software and data sets]," *IEEE Geosci. Remote Sens. Mag.*, vol. 9, no. 3, pp. 174–180, Sep. 2021.

[231] C. Requena-Mesa, V. Benson, M. Reichstein, J. Runge, and J. Denzler, "EarthNet2021: A large-scale dataset and challenge for earth surface forecasting as a guided video prediction task.," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 1132–1142.

[232] E. Cheng et al., "Wheat yield estimation using remote sensing data based on machine learning approaches," *Front. Plant Sci.*, vol. 13, 2022, Art. no. 1090970.

[233] W. G. Chaminda Bandara and V. M. Patel, "Revisiting consistency regularization for semi-supervised change detection in remote sensing images," 2022, *arXiv:2204.08454*.

[234] M. Bernhard, N. Strauß, and M. Schubert, "MapFormer: Boosting change detection by using pre-change information," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 16837–16846.

[235] G. Mao, Y. Yuan, and L. Xiaoqiang, "Deep cross-modal retrieval for remote sensing image and audio," in *Proc. IAPR Workshop Pattern Recognit. Remote Sens.*, 2018, pp. 1–7.

[236] L. Xie, R. Han, S. Xie, D. Chen, and Y. Chen, "Multi-view fusion network for crop disease recognition," in *Proc. ACM Int. Conf. Algorithms Comput. Syst.*, 2021, pp. 121–126.

[237] X. X. Zhu et al., "So2Sat LCZ42: A benchmark data set for the classification of global local climate zones [software and data sets]," *IEEE Geosci. Remote Sens. Mag.*, vol. 8, no. 3, pp. 76–89, Sep. 2020.

[238] D. Bonafilia, B. Tellman, T. Anderson, and E. Issenberg, "Sen1Floods11: A georeferenced dataset to train and test deep learning flood algorithms for Sentinel-1," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2020, pp. 210–211.

[239] L. Kondmann et al., "DENETHOR: The DynamicEarthNET dataset for harmonized, inter-operable, analysis-ready, daily crop monitoring from space," in *Proc. 35th Conf. Neural Inf. Process. Syst. Datasets Benchmarks Track*, 2021, pp. 1–13.

[240] G. Donchyts, A. Moreno-Rodenas, D. Valero, P. N. Tovar, N. Gorelick, and M. Franca, "Segmentation of dams in medium-resolution satellite imagery, U-net type CNN architecture, and cloud computing," in *Proc. AGU Fall Meeting Abstr.*, 2021, Art. no. GC43D-05.

[241] H. Alemohammad and K. Booth, "LandCoverNet: A global benchmark land cover classification training dataset," 2020, *arXiv:2012.03111*.

[242] C. Requena-Mesa, V. Benson, M. Reichstein, J. Runge, and J. Denzler, "EarthNet2021: A large-scale dataset and challenge for earth surface forecasting as a guided video prediction task," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 1132–1142.

[243] A. Toker et al., "DynamicEarthNet: Daily multi-spectral satellite dataset for semantic change segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 21126–21135.

[244] C. Aybar et al., "CloudSEN12, a global dataset for semantic understanding of cloud and cloud shadow in Sentinel-2," *Sci. Data*, vol. 9, no. 1, 2022, Art. no. 782.

[245] F. Bastani, P. Wolters, R. Gupta, J. Ferdinando, and A. Kembhavi, "Satlas: A large-scale, multi-task dataset for remote sensing image understanding," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 16772–16782.

[246] M. Cha et al., "Multiearth 2023–Multimodal learning for earth and environment workshop and challenge," 2023, *arXiv:2306.04738*.

[247] J. Hu et al., "MDAS: A new multimodal benchmark dataset for remote sensing," *Earth Syst. Sci. Data*, vol. 15, no. 1, pp. 113–131, 2023.

[248] S. Ahlswede et al., "TreeSatAI benchmark archive: A multi-sensor, multi-label dataset for tree species classification in remote sensing," *Earth Syst. Sci. Data*, vol. 15, no. 2, pp. 681–695, 2023.

[249] M. Mommert, N. Kesseli, J. Hanna, L. Scheibenreif, D. Borth, and B. Demir, "Ben-ge: Extending BigEarthNet with geographical and environmental data," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2023, pp. 1016–1019.

Francisco Mena (Graduate Student Member, IEEE) received the master's and bachelor's degree in computer engineering from the Federico Santa Maria Technical University (UTFSM), Valparaíso, Chile, in 2020. He is currently working toward the Ph.D. degree in computer science with the University of Kaiserslautern-Landau, Kaiserslautern, Germany.

During 2020 and 2021, he gave some lectures on computational statistics and artificial neural networks in the computer engineering program with the UTFSM. He is currently researching with the German Research Center for Artificial Intelligence (DFKI GmbH), Kaiserslautern, under the supervision of Prof. A. Dengel. His research interests include deep neural networks, dimensionality reduction, multiview or multimodal learning, data fusion, and Earth observation applications.

Mr. Mena is involved as a Member in the Geoscience and Remote Sensing Society (GRSS) and a Reviewer for IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING.

Diego Arenas received the bachelor's degree in computer science from the Universidad de Talca, Talca, Chile, in 2007, the M.Sc. degree in data science from the University of Edinburgh, Edinburgh, U.K., in 2016, and the Eng.D. degree in computer science from the University of St. Andrews, St. Andrews, U.K., in 2021.

He worked in information systems as a Consultant and Project Manager for ten years in sectors such as finance, banking, retail, education, human resources, transport, services, and telecommunications in Chile. He is volunteering and collaborating in applied data related projects with academia and the third sector for the last 15 years. Since 2022, he has been a Senior Researcher with the German Research Center for Artificial Intelligence (DFKI GmbH), Kaiserslautern, Germany, working in the intersection of artificial intelligence and Earth observation. His research interests include anticorruption, biodiversity, AI for Good, food security, and remote sensing among others.

Marlon Nuske received the master's and Ph.D. degrees in physics from the University of Hamburg, Hamburg, Germany in 2015 and 2020, respectively.

He is currently working with the German Research Center for Artificial Intelligence (DFKI GmbH), Kaiserslautern Germany, as a Senior Researcher leading the Earth and Space Applications team since 2021. His research interests include machine learning applications in Earth Observation, data fusion, hybrid modeling techniques, and physics-aware machine learning.

Andreas Dengel received the diploma degree in computer science from the University of Kaiserslautern, Kaiserslautern, Germany, in 1986, and the Ph.D. degree in computer science from the University of Stuttgart, Stuttgart, Germany, in 1989.

In 1993, he became a Professor with the Computer Science Department, University of Kaiserslautern, where he holds the chair Knowledge-Based Systems. Since 2009, he has been a Professor (Kyakuin) with the Department of Computer Science and Information Systems, Osaka Prefecture University, Sakai, Japan. He was also with IBM, Siemens, and Xerox Parc. He is currently a Scientific Director with the German Research Center for Artificial Intelligence (DFKI GmbH), Kaiserslautern. Moreover, he has coedited international computer science journals and has written or edited 12 books. He has authored more than 300 peer-reviewed scientific publications and has supervised more than 170 Ph.D. and master's theses. His research interests include the areas of pattern recognition, document understanding, information retrieval, multimedia mining, semantic technologies, and social media.

Dr. Dengel is a member of several international advisory boards, has chaired major international conferences, and founded several successful start-up companies. He is an IAPR Fellow and has received prominent international awards.