MULTI-MODAL DEEP LEARNING WITH SENTINEL-3 OBSERVATIONS FOR THE DETECTION OF OCEANIC INTERNAL WAVES

L. Drees\textsuperscript{1,*}, J. Kusche\textsuperscript{1}, R. Roscher\textsuperscript{1,2}

\textsuperscript{1} IGG, University of Bonn, Germany - ldrees@uni-bonn.de, kusche@geod.uni-bonn.de, ribana.roscher@uni-bonn.de
\textsuperscript{2} Institute of Computer Science, University of Osnabrueck, Germany

KEY WORDS: multi-modal, deep learning, internal waves, sentinel-3, multi-stream, late fusion, neural network

ABSTRACT:

The observation of waves that propagate along density interfaces inside the ocean poses a significant challenge, as their visible surface signatures are much lower compared to their internal amplitudes. However, monitoring internal waves is important as they redistribute large amounts of energy, play a role in mixing and vertical heat transfer, and modify water and nutrient transports. Although satellite observations would allow global monitoring of internal waves at constant time intervals, their automatic detection is challenging: In optical images, internal waves are hardly visible and can be obscured by clouds, whereas radar data have limitations in coastal regions and their spatial coverage is not perfect. Furthermore, the occurrence of internal waves can be confused with other ocean phenomena. In this work, we present an automated detection framework for internal waves based on multiple data sources in order to compensate for the shortcoming of single data sources. In our application, we use Ocean and Land Color Imager and Synthetic Aperture Radar Altimeter data. Our contributions are (1) we develop a multi-modal deep neural network \textit{SONet} with multi-streams and late fusion, which performs a classification on the basis of training with both modalities, and (2) we establish a method to deal with missing modalities. Experiments in the Amazon Shelf region show \textit{SONet} achieves adequate results when both modalities are available, but also when only a single modality is available. By exploiting correlations between the modalities, \textit{SONet} classifies OLCI images off the SRAL ground track better than uni-modal network \textit{OLNet}, which describes a great advantage of our multi-modal network.

1. MOTIVATION

We witness a growth in the number of satellites with integrated sensors characterized by various spatial, spectral and temporal resolutions. It is therefore common in remote sensing that the same scene is observed simultaneously with different sensors and therefore multi-modal data is available for a joint analysis. Compared to data from individual sensors, the different modalities usually have certain properties and characteristics that can be exploited for a better understanding of the scene. Especially for satellite missions, where many research questions from different scientific fields are addressed simultaneously, multi-modal data is common. For the Sentinel-3 mission, for example, Ocean and Land Color Imager (OLCI) and Synthetic Aperture Radar Altimeter (SRAL) are mounted on the same satellite such that radar signals and optical images are acquired simultaneously in intersecting observation areas. We present a framework which combines SRAL and OLCI observations for an automatic detection of oceanic internal waves (IW), as illustrated in Figure 1. Oceanic IWs are gravity waves at internal density layers of the water. Compared to surface waves, they are significantly larger, both in amplitude (up to 200 meters) and in wavelength (up to multiple kilometers). For example, IWs play a key role in understanding the interaction of large-scale tides and smaller scale turbulences (Jackson et al., 2012). However, the detection of IWs is a challenge because they are hardly visible optically on the sea surface and active sensors like SRAL with a better detection rate do not observe the whole area. With this work we show that it is worthwhile to combine both modalities SRAL and OLCI in one model for the detection of IW. Although the data set we have created is currently still specific to our study site, our experiments can already show that the use of both modalities leads to an increase in accuracy compared to uni-modal OLCI-based methods.

The paper is structured as follows: In Sec. 2 we present current research on multi-modal learning in earth sciences and explain the connections to this work. Besides a brief insight into the investigation of IWs is given. Sec. 3 introduces the study site and our used multi-modal data set. The multi-modal deep learn-
ing framework and how it process the data set is presented in Sec. 4. In Sec. 5 we show a concrete implementation of the multi-stream procedure with late fusion and discuss the results afterwards.

2. STATE OF THE ART

In this section we first discuss multi-modal approaches using earth observation data. Afterwards we present the current state of research on IWs.

2.1 Multi-modal deep learning in earth sciences

In general, the potential of machine learning methods and especially of approaches of deep learning in remote sensing is large and includes classification and regression tasks as well as state prediction tasks such as now-casting of precipitation, seasonal forecasts and modelling of global mass transport (Reichstein et al., 2019, Ma et al., 2019). Related to this is the promising research area of multi-modal learning. Apart from conventional data fusion methods (Lahat et al., 2015, Gupta, Cheng, 2006) in the field of Earth sciences, so far, only a few multi-modal deep learning approaches exist. One multi-modal multi-source approach is used for underwater mapping of the seafloor. In order to analyze habitats for marine ecology, few visual images of autonomous underwater vehicles and a multitude of bathymetric data from ships are used to perform classification tasks (Rao et al., 2014). Another related approach is the multi-source classification of cloud, shadow and land cover scenes with data from different satellite missions (Shendy et al., 2019). But in contrast to our method, all input data (PlanetScope and Sentinel-2 imagery) are of optical nature. There are also temporal multi-modal networks with the aim to learn a common representation of the data, which have both different modalities and are variable in time (Yang et al., 2017). Similar to our architecture is the multi-stream approach for temporal Sentinel-2 data to generate land cover classes from VHSR images and time series with high spatial resolution (Benedetti et al., 2018). Apart from Earth Sciences, multi-modal deep learning is already more widespread, including audio-visual speech recognition (Mroueh et al., 2015), scene alignment (Aytar et al., 2017b, Aytar et al., 2017a), image captioning (Srivastava, Salakhutdinov, 2012) and video hyperlinking (Vukotić et al., 2016).

2.2 Investigation of internal waves

For many years, it has been an effort to detect oceanic IWs using remote sensing methods in order to determine their parameters and energy, to get information about the stratification of the water and the mixed layer depth, or to investigate their influence on tidal (or baroclinic) currents (Klemas, 2012). IWs are caused by external forces acting on stratified water levels, e.g. wind stress or tides over a region of bottom topography (Alpers et al., 2008, Robinson, 2010, Magalhães et al., 2016, Zhao et al., 2004). Although there are maps showing IWs hotspots in all areas of the oceans (Jackson et al., 2012), a global automatic detection procedure does not exist yet. This would be beneficial also for other satellite missions, like SWOT, which are interested in mesoscale processes like geostrophic velocities. These have an even greater amplitude than IWs, but since the phenomena overlap, it will be difficult to determine geostrophic velocities at scales smaller than about 70 km without knowledge of IWs locations (Qiu et al., 2017). In many cases, the sea surface elevations of IWs are too small to be visible, but rather

2.3 Multi-modal dataset with lack of modalities

We focus on an area in the Atlantic Ocean off the Amazon Shelf, which is known for large amplitude ISW (Magalhães et al., 2016, Santos-Ferreira et al., 2019). Spatially, we concentrate on certain relative orbits (RO) of the Sentinel-3 mission, namely RO 38, 95, 152 and 209 in the western part and RO 380, 52, 109 and 166 in the eastern part (Fig. 2). In the period from April 2017 to August 2019 we collected and annotated a total amount of 2373 data samples on these orbits. Also Sentinel-3B, which is equipped with the same sensors, has been contributing data since January 2019. However, Sentinel-3B flies 140 degree out of phase compared to Sentinel-3A, so the ground tracks are offset in such a way that the spatial coverage is approximately doubled.

3. DATA

3.1 Study site

We use OLCI Level-1b-EFR top-of-atmosphere (TOA) radiometric full resolution image data with 21 bands, which

![Figure 2. Map of the study region off the Amazon Shelf taken from (Santos-Ferreira et al., 2019) with ground tracks of 15 RO of Sentinel-3A (green dashed lines). Black bold curves indicate location of IWs crest observations by (Magalhães et al., 2016). The presumed IWs-origins are marked by labels A and B.](https://doi.org/10.5194/isprs-annals-V-2-2020-813-2020)
Since the originally provided data are too large to be processed by our multi-modal network, we extract image patches \( O = [351 \times 351 \times 21] \) as well as SLR tracks \( S = [313 \times 4] \) (due to 313 observations per parameter SWH, SLA, \( \sigma_0^N \), \( \delta \sigma_0^N \)) which are still georeferenced (Fig. 4). Thus, our dataset consists of three subsets \( \mathcal{P}_O \), \( \mathcal{P}_S \), and \( \mathcal{P}_P \). Subset \( \mathcal{P}_O \) consists only of optical data, subset \( \mathcal{P}_S \) only of radar signals, and \( \mathcal{P}_P \) contains both, that means multi-modal data (Fig. 3). In case that not both modalities are available for a sample we call this "lack of modality". Therefore the subsets \( \mathcal{P}_O \) and \( \mathcal{P}_P \) suffer from a lack of modality because the OLCI resp. SRAL modality is missing there. Just subset \( \mathcal{P}_S \) is a multi-modal data set without lack of modalities. All samples of the subsets are referenced in classes \( \text{IW} \) and \( \text{NoIW} \).

### 3.2.1 Subset \( \mathcal{P}_O \)
It consists of \( \mathcal{N} \) OLCI samples \( \mathcal{O} = \{O_1, \ldots, O_\mathcal{N}\} \) with labels \( \mathcal{Y}_O = [y_1, \ldots, y_\mathcal{N}]^T \). Images in this subset are mostly taken off the SRAL ground track. Multi-modal recorded data, where the SRAL signal is disturbed by land influences and therefore unusable, are also in this subset. The same applies to images where waves appear at the lateral edge of the image and are therefore not in the field of view of the SRAL sensor. The images show considerable variation in wave forms and sizes, brightness and cloud cover, which makes classification challenging.

### 3.2.2 Subset \( \mathcal{P}_S \)
Radar data lies in this subset if the underlying OLCI image is covered by clouds and therefore not applicable. A total number of \( \mathcal{N} \) SRAL samples \( \mathcal{S} = \{S_1, \ldots, S_\mathcal{N}\} \) with reference vector \( \mathcal{Y}_S = [y_1, \ldots, y_\mathcal{N}]^T \) is available. We focus on the parameters SWH, SLA and \( \sigma_0^N \) from Ku-band in 20 Hz resolution and use them in their original state. Additionally we compute the \( \delta \sigma_0^N \) from the \( \sigma_0^N \) as presented in (Santos-Ferreira et al., 2019).

### 3.2.3 Subset \( \mathcal{P}_P \)
This subset contains both modalities SRAL \( \mathcal{S} = \{S_1, \ldots, S_\mathcal{N}\} \) and OLCI \( \mathcal{O} = \{O_1, \ldots, O_\mathcal{N}\} \) with common label vector \( \mathcal{Y}_P = [y_1, \ldots, y_\mathcal{N}]^T \), where \( \mathcal{N} \) is the total number of samples in this subset. It concerns image patches where the SRAL signal is not corrupted by coastal topography and the OLCI image is covered by less than 25% of clouds. In any case, the point where SRAL indicates an IW must be visible in the OLCI image. Spatial and temporal synchronization of the modalities is achieved by ensuring SRAL ground track is in all samples exactly centered over the OLCI patch and ends at the boundaries of the image. This ensures a strong alignment between the data.

### 4. MULTI-MODAL DEEP LEARNING NETWORK
In our work, we design a multi-modal neural network called \( \text{S0Net} \), which is jointly trained on both modalities OLCI and SRAL. This is useful when a sensor fails, the image is obscured by clouds, or the radar signal is corrupted by coastal topography, as illustrated in Fig. 5. For instance, radar signals may indicate an IW, but in the OLCI image it is clearly visible that it is actually a rainstorm. Furthermore, correlations between the modalities can be utilized to control, discard or support results obtained from the other sensor. Besides, the spatial coverage of the OLCI image (swath width of 1270 km) is much larger than that of the SRAL signal (no swath, across width footprint diameter about 2km) on the ground.

Generally, processing of multi-modal data in machine learning is not trivial due to different characteristics, dimensions, units, scales and resolutions of input modalities. (Baltrušaitis et al., 2018) summarizes the core challenges of multi-modal learning as representation learning, alignment, fusion, translation and co-learning. In this work, the first three play an important role, while translation and co-learning are not required. Representation learning is about generalizing the data to exploit complementary and redundancy, for which we use multi-streams in \( \text{S0Net} \). The multi-stream technology is widely used.
in multi-modal deep learning because the modalities have different properties, and therefore require different operations for feature extraction (Wu et al., 2016, Huang, Kingsbury, 2013). Alignment describes the task to extract the connection between the modalities. In our dataset it is helpful that a spatial and temporal correlation already exist. Via fusion, both modalities are merged in the network to obtain a joint feature representation (Ngiam et al., 2011). At SONet, we opt for late fusion in terms of the depth of the network in which we are fusing. Although the interconnections in a late fusion are significantly weaker than in an early one, it is more suitable for modalities that have very different semantics. Furthermore it is easier to compensate a lack of modality (Liu et al., 2018). Other possibilities are early fusion, which requires similar semantics, much preprocessing and a high level of knowledge about the modality alignment, or to fuse multiple times and compute a weighted sum each (Vielzeuf et al., 2018).

4.1 Our multi-modal architecture

We have developed the neural network SONet, which handles the challenge of two-modality samples (Fig. 6). In the following we describe the complete structure starting with the required input form, the streams for each modality, the fusion of these streams and ending with the classification head.

For the input all OLCI data of the subsets form a matrix $\mathbf{O}X$ and all SRAL data form a matrix $\mathbf{S}X$. $\mathbf{O}X$ has the dimension $[N \times M \times M \times B]$, where $N$ is the number of samples, $M$ the side length of the OLCI patch and $B$ the number of used bands. $\mathbf{S}X$ has the dimension $[N \times L \times T]$, where $L$ is the number of SRAL observations that fall into a patch and $T$ is the number of actually used parameter. So $\mathbf{O}X$ represents the input for O-Stream and $\mathbf{S}X$ for S-Stream. Note that both input matrices have the same number of samples $N$ and correspond to the same reference label vector $y$.

The modalite-specific streams consist of recurring sequences of convolutional blocks. Each convolution block contain (in the order given) CONV$_{2D}$ or CONV$_{1D}$, ReLU (activation), BN (batch normalization), Pool$_{max}$ and DO (dropout), whereas regularization techniques BN and DO are optional. The output of the streams are flat layers, which can be considered as compact representation of the input modalities.

One key element of multi-modal learning is the fusion of the modalities. We decide for late fusion and merge both compact modality-specific representations at the end of the streams. There are several methods to fuse layers, like the operators addition, multiply, average, maximum, minimum and concatenation. Please note, that depending on the concrete fuse option the layers must have an identical size in at least one dimension. As result a joint representation of both modalities is obtained.

In deeper layers, the joint representation is subsequently compressed with fully-connected (FC) layers, until an output layer returns the classification result (classification head). The classification head consists of FC layers with ReLU activation. While in all hidden layers including the streams, ReLU is used as activation function, the output layer consisting of two neurons uses Softmax activation. Therefore, the value of the last two neurons can be interpreted as probability for the class assignment of the samples. One neuron stores the probability that the sample maps to $\mathbb{IW}$, the other neuron that it maps to $\mathbb{O}$.

4.2 Loss functions

In order to train the network, two loss functions are introduced. The first $CE(y, \hat{y})$ includes the measure of binary cross entropy

$$CE(y, \hat{y}) = -[y \cdot \log(\hat{y}) + (1-y) \cdot \log(1-\hat{y})]$$  \hspace{1cm} (1)
5.1 Data preprocessing

5.1.1 Brightness correction and normalization Since the brightness of OLCI images is very different from each other, it has also proved to be useful to correct each image individually. The 75% quantile values are calculated for all pixels of a band. The bands are then divided by their respective quantiles. After this operation all pixels that are larger than 1 are set to 1. Subsequently, a normalization is performed over each subset by calculating a z-transform over all samples of the subset. We also reduce the input image size to \([128 \times 128]\) for reasons of runtime. For SRAL only z-transform is performed.

where \(y\) represent the reference label and \(\hat{y}\) describes the probability of being class \(lw\). Further the focal loss \(FL(p_t)\) is used which is suitable to compensate for an imbalance in the data set (Lin et al., 2017).

\[
FL(p_t) = -\alpha_t \cdot (1 - p_t)^\gamma \cdot \log(p_t)
\]

(2)

Here \(\gamma\) is a previously defined integer scalar, \(\alpha_t\) the weighting factor

\[
\alpha_t = \begin{cases} 
\alpha & \text{if } y = 1 \\
1 - \alpha & \text{otherwise}
\end{cases}
\]

(3)

where \(\alpha\) is a predefined value between 0 and 1 and

\[
p_t = \begin{cases} 
\hat{y} & \text{if } y = 1 \\
1 - \hat{y} & \text{otherwise}
\end{cases}
\]

(4)

5. EXPERIMENTAL SETUP

First, information on data preprocessing, including brightness enhancement, normalization and data augmentation is given. This is followed by details on network architecture from input, streams and fusion to modifications for uni-modal baseline models. Finally the training procedure is explained with cross validation, hyperparameter settings and evaluation.

5.2 Detailed network structure

5.2.1 Stream-Input As already mentioned in Sec. 4.1, the input to 0-Stream is \(\mathcal{X} = \{\mathcal{O}, \mathcal{O}, \mathcal{O}\}\) and the input to S-Stream is \(\mathcal{X} = \{\mathcal{S}, \mathcal{O}, S\}\). Since the matrices \(\mathcal{O}\) and \(\mathcal{S}\) do not exist in their respective subsets, they are zeroed placeholder in the appropriate size so that the matrices can be concatenated together. So \(\mathcal{O}\) is a zero matrix with the size \([N \times M \times M \times B]\) and \(\mathcal{S}\) with dimension \([N^2 \times L \times T]\). Both streams have a common reference vector \(y = [y, y, y, y]\) with length \(N = [N + \alpha N + \alpha N]\), which is the total number of all samples in all subsets.

It has been evaluated in advance that accuracy cannot be increased by combining OLCI bands. In preliminary experiments, we observed that with band 16 the IWs are most visible. Therefore this band is used as the single input. Compared to the use of all bands, this offers a considerable reduction of runtime. For SRAL, we use information from several parameters, namely radar backscatter \(\sigma_0\), differenced-mean square slope \(\delta\), SWH, and SLA. After preprocessing, we first perform data augmentations for all subsets, resulting in \(N \rightarrow N\), \(\mathcal{O} \rightarrow \mathcal{O}\), and \(\mathcal{S} \rightarrow \mathcal{S}\), with \(N\) indicating the size after augmentation. After band selection, the 0-Stream input \(\mathcal{X}\) has a size of \([N^2 \times 128 \times 128 \times 1]\), and the S-Stream input \(\mathcal{X}\) has a size of \([N^2 \times 313 \times 4]\).

5.2.2 O-Stream Each of four convolutional block consists of three different layers in the order of CONV2D, ReLU and PoolMax. The kernel size in all CONV2D of all convolutional blocks is \((3 \times 3)\) at a stride of \((1 \times 1)\) in each direction. Besides, in
CONV_2D layers of the O stream, L2 kernel regularization with the parameter 0.01 is also applied to reduce overfitting. The number of filters increases continuously from 16 in the first convolutional block over 32 and 64 to 128 in the last one. PoolingMax uses a kernel of size $\left(2 \times 2\right)$ and a stride of $(2)$ in each direction. So the layer size is exactly halved by each pooling operation.

5.2.3 S-Stream 3 convolutional blocks consisting of CONV_ID, ReLU and PoolingMax are used in this stream. However, the main difference in the convolutional blocks, is the layer CONV_ID, so the parameters are modified in the direction of $l$ only. The number of filters is 16 in the first convolution block and 32 in the second and 64 in the last one. The kernel size of CONV_ID is $(3)$ with stride $(1)$. Accordingly, the POOL_MAX is one-dimensional with a pooling kernel of size $(2)$ and stride of $(2)$.

5.2.4 Fusion After flattening both streams, we use a fusion of width FC$_{128}$ and fusion type addition to merge the streams outputs. Therefore, we need both flat layers to have the same size. To get both flattened stream outputs to the size of 128 neurons, a dense layer is used $(\text{Flat} \rightarrow \text{FC}_{128})$. The joint representation also has the size of 128 neurons. Subsequently, the classification head is connected to the joint representation. The number of neurons in the classification head is slowly reduced by connecting two fully connected layers with ReLU activation of width 32 and 8. The softmax output layer with 2 neurons is the last layer.

5.2.5 Uni-modal baseline networks For comparison, uni-modal networks $\text{SNet}$ and $\text{ONet}$ are created, in which the classification head is directly attached to the respective stream. So the fusion is omitted, but stream and classification head are identical to $\text{SNet}$. The input matrices are slightly modified for the training of uni-modal networks, since it is not efficient to use the zero placeholder matrices for training. Therefore $\mathbf{O}$ is reduced for $\text{ONet}$ to $\mathbf{ONeX} = \mathbf{[O, O]}$ with shortened reference label $\mathbf{ORev} = \mathbf{[O, O]}$ and $\mathbf{X}$ is reduced for $\text{SNet}$ to $\mathbf{SNeX} = \mathbf{[S, S]}$ with $\mathbf{SRev} = \mathbf{[S, S]}$.

5.3 Training procedure

5.3.1 Cross Validation For the purpose of cross-validation, the data set (see Fig. 1) is quartered by combining the data from two RO in each case $(1: 38+172, 2: 95+209, 3: 380+109, 4: 52+166)$. The split of the total 2973 sample is done in this way to find a similar amount of data in each quarter $(1: 663, 2: 653, 3: 456, 4: 601)$. Moreover this ensures the samples in the different quarters are geographically separated from each other. Consciously we decided against a temporal split, because IWs at the same location at different times can look very similar, which would falsify the cross validation. In the experiments, training is performed sequentially on three quarters and testing is performed on the fourth one. This results in a total of 4 cross validation runs.

5.3.2 Pre-training and fine-tuning As a baseline, we first train models from the uni-modal networks. For $\text{SNet}$ we use 50 epochs, with a learning rate of $1e-4$, linear decay and a batchsize of 64. $\mathcal{CE}$ with Adam Optimizer has proven to be the most suitable loss-function. ONet is trained for a longer period of 200 epochs, but with a lower learning rate of $1e-5$. Decay and batchsize are identical to $\text{SNet}$. However, the loss function is $\mathcal{FL} \left(\alpha = 0.5, \gamma = 3\right)$ with Adam Optimizer.

Since uni-modal networks and $\text{SNet}$ have the same streams, it has been proven useful for training $\text{SNet}$ to use as initial stream-weights those from uni-modal networks. Thus the essential feature extraction is already given, so that only the connection between the modalities has to be learned. For this, $\text{SNet}$ is trained with 100 epochs with a learning rate of $1e-5$, a batch size of 64 and a linear decay. As with $\text{ONet}$, $\mathcal{FL} \left(\alpha = 0.5, \gamma = 3\right)$ is used with Adam optimizer.

5.3.3 Evaluation metrics To evaluate the classifiers, a confusion matrix is calculated from the reference class labels $\mathbf{y}$ and the predicted class labels $\hat{\mathbf{y}}$ (rounded to 0 and 1). The evaluation metrics overall accuracy (OA), average accuracy (AA) and $F_1$-Score are derived from this. As a further metric the MSE is introduced. This is calculated from the probabilities $\hat{y}$ the network returns in the output layer and the referenced labels $y$.

$$\text{MSE} = \frac{1}{N} \sum_{n=1}^{N} (y_n - \hat{y}_n)^2$$

It can be understood as a measure of how reliably the classifier can decide in the prediction for one of the classes. The smaller MSE, the higher the reliability of the prediction.

6. RESULTS

We intend to investigate which classification result is achievable on the specific subsets. For this, we compare the performance of the uni-modal networks $\text{SNet}$ and $\text{ONet}$ with the multi-modal network $\text{SNet}$. The results are summarized in Tab. 2.

For subset $\mathcal{S}$, it is noticeable that $\text{SNet}$ and $\text{SNet}$ perform about equally well. Both networks achieve an OA of approxim-
We have conducted additional experiments with a Random Forest (RF) (Breiman, 2001). Unlike SOnet, a RF has the weakness that it is poorly suited for multi-modal input that has different dimensions, which requires prior manual embedding or the application of a dimensionality reduction algorithm. We use PCA-obtained feature vectors of the same size for each modality independently, to avoid a potential weighting between both modalities. With uni-modal input, RF achieves maximum accuracies of SOnet and SNet. Nevertheless, the accuracy with multi-modal input does not increase over uni-modal input.

7. CONCLUSION

In this work, we demonstrated that our multi-modal deep learning framework is able to detect oceanic internal waves. We thus feel confident to suggest that such networks are a promising research direction in the earth sciences. Overall, the multi-stream technique with late fusion is well suited to exploit correlations and alignments between modalities. If both modalities are available, strong results (overall accuracy: 92%) are already achieved based just on Sentinel-3 SRAL data. However, the ground coverage of Sentinel-3 OLCI is much larger, which is essential for a continuous global observation of internal waves. Hence, areas which are not covered by SRAL, tracks, a multi-modal network significantly increases the overall accuracy (78% instead of 70%) and the average accuracy (71% instead of 64%), when compared to an uni-modal SOnet network. Meanwhile, SOnet also performs as well as SNet in areas where SRAL is present. Due to the higher reliability through different input data types, we recommended to use SOnet for classification for all subsets. We have also shown that zeroing the missing modality does not negatively affect the training of a multi-modal network. This allows a multi-modal data set to be extended very easily by uni-modal data. Future work will concern applications from satellites, remote sensing and the integration of further modalities, but also the joint use of close-range data from multi-sensor systems, as is the case in the field of precision agriculture.

ACKNOWLEDGEMENTS

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany’s Excellence Strategy – EXC 2070 – 390733234.

REFERENCES

Alpers, W., Brandt, P., Rubino, A., 2008. Internal waves generation: Strategy – EXC 2070 – 390732324. Remote sensing of the European seas, Springer, 319–330.

Table 2. Results achieved with SOnet compared with SNet and SNet after cross validation. The mean values of the cross validation runs and the standard deviations (indicated by ±) are given. The results are divided into tests with multi-modal data (subset 5p) and tests with uni-modal data (5p: OLCI, 1p: SRAL).

|       | OA [%]     | AA [%]     | F1 | MSE      | OA [%]     | AA [%]     | F1 | MSE      | OA [%]     | AA [%]     | F1 | MSE      |
|-------|------------|------------|----|----------|------------|------------|----|----------|------------|------------|----|----------|
| SOnet | 91.86 0.75 | 84.60 0.75 | 79.76 0.54 | 18 0.18 | 88.14 ±0.02 | 86.28 ±0.02 | 70.31 ±0.07 | 0.01 | 0.79 ±0.02 | 0.10 ±0.01 | 0.01 | 0.01 ±0.01 |
Aytn, Y., Castrejon, L., Vondrick, C., Pirsiavash, H., Torralba, A., 2017a. Cross-modal scene networks. *IEEE transactions on pattern analysis and machine intelligence*, 40(10), 2303–2314.

Aytn, Y., Vondrick, C., Torralba, A., 2017b. See, hear, and read: Deep aligned representations. *arXiv preprint arXiv:1706.00932.*

Baltas, T., Abuja, C., Morency, L.-P., 2018. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2), 423–443.

Benedetti, P., Ienco, D., Gaetano, R., Ose, K., Pensa, R. G., Dupuy, S., 2018. *M^3 Fusion*: A Deep Learning Architecture for Multiscale Multimodal Multitemporal Satellite Data Fusion. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 11(12), 4939–4949.

Breiman, L., 2001. Random forests. *Machine learning*, 45(1), 5–32.

Gupta, A. K., Cheng, Y. C., 2006. Method, system and module for multi-modal data fusion. US Paten 7,152,033.

Huang, J., Kingsbury, B., 2013. Audio-visual deep learning for noise robust speech recognition. *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, IEEE, 7596–7599.

Ikeda, M., 1995. *Oceanographic applications of remote sensing*. CRC press.

Jackson, C., 2007. Internal wave detection using the moderate resolution imaging spectroradiometer (MODIS). *Journal of Geophysical Research: Oceans*, 112(C11).

Jackson, C. R., Da Silva, J. C., Jeans, G., 2012. The generation of nonlinear internal waves. *Oceanography*, 25(2), 108–123.

Klemas, V., 2012. Remote sensing of ocean internal waves: An overview. *Journal of Coastal Research*, 28(3), 540–546.

Lahat, D., Adali, T., Jutten, C., 2015. Multimodal data fusion: an overview of methods, challenges, and prospects. *Proceedings of the IEEE*, 103(9), 1449–1477.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollár, P., 2017. Focal loss for dense object detection. *Proceedings of the IEEE international conference on computer vision*, 2980–2988.

Liu, K., Li, Y., Xu, N., Natarajan, P., 2018. Learn to combine modalities in multimodal deep learning. *arXiv preprint arXiv:1805.11730."

Ma, L., Liu, Y., Zhang, X., Ye, Y., Yin, G., Johnson, B. A., 2019. Deep learning in remote sensing applications: A meta-analysis and review. *ISPRS journal of photogrammetry and remote sensing*, 152, 166–177.

Magalhães, J. M., Da Silva, J., Bujsman, M. C., Garcia, C., 2016. Effect of the North Equatorial Counter Current on the generation and propagation of internal solitary waves off the Amazon shelf (SAR observations). *Ocean Science*, 12(1), 243–255.

Mroueh, Y., Marcheret, E., Goel, V., 2015. Deep multimodal learning for audio-visual speech recognition. *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2130–2134.

Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., Ng, A. Y., 2011. Multimodal deep learning. *Proceedings of the 28th international conference on machine learning (ICML-11)*, 689–696.

Qiu, B., Nakano, T., Chen, S., Klein, P., 2017. Submesoscale transition from geostrophic flows to internal waves in the northwestern Pacific upper ocean. *Nature communications*, 8(1), 1–10.

Rao, D., De Deuge, M., Nourani-Vatani, N., Douillard, B., Williams, S. B., Pizarro, O., 2014. Multimodal learning for autonomous underwater vehicles from visual and bathymetric data. *2014 IEEE International Conference on Robotics and Automation (ICRA)*, IEEE, 3819–3825.

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N. et al., 2019. Deep learning and process understanding for data-driven Earth system science. *Nature*, 566(7743), 195.

Robinson, I. S., 2010. *Discovering the Ocean from Space: The unique applications of satellite oceanography*. Springer Science & Business Media.

Santos-Ferreira, A., da Silva, J., Magalhaes, J., 2018. SAR Mode Altimetry Observations of Internal Solitary Waves in the Tropical Ocean Part 1: Case Studies. *Remote Sensing*, 10(4), 644.

Santos-Ferreira, A., da Silva, J., Srokosz, M., 2019. SAR Mode Altimetry Observations of Internal Solitary Waves in the Tropical Ocean Part 2: A Method of Detection. *Remote Sensing*.

Shendryk, Y., Rist, Y., Ticehurst, C., Thorburn, P., 2019. Deep learning for multi-modal classification of cloud, shadow and land cover scenes in PlanetScope and Sentinel-2 imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 157, 124–136.

Srivastava, N., Salakhutdinov, R. R., 2012. Multimodal learning with deep boltzmann machines. *Advances in neural information processing systems*, 2222–2230.

Vielzeuf, V., Lechervy, A., Pateux, S., Jurie, F., 2018. CentralNet: a multilayer approach for multimodal fusion. *Proceedings of the European Conference on Computer Vision (ECCV)*.

Vukotic, V., Raymond, C., Gravier, G., 2016. Multimodal and crossmodal representation learning from textual and visual features with bidirectional deep neural networks for video hyperlinking. *Proceedings of the 2016 ACM workshop on Vision and Language Integration Meets Multimedia Fusion*, ACM, 37–44.

Wu, Z., Jiang, Y.-G., Wang, X., Ye, H., Xue, X., 2016. Multistream multi-class fusion of deep networks for video classification. *Proceedings of the 24th ACM international conference on Multimedia*, ACM, 791–800.

Yang, X., Ramesh, P., Chitta, R., Madhavanath, S., Bernal, E. A., Luo, J., 2017. Deep multimodal representation learning from temporal data. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5447–5455.

Zhao, Z., Klemas, V., Zheng, Q., Yan, X.-H., 2004. Remote sensing evidence for baroclinic tide origin of internal solitary waves in the northeastern South China Sea. *Geophysical Research Letters*, 31(6).