Cross-Modal Learning for Classification of Optical and SAR Imagery

Ekatarina Senchugova[a] and Ronny Hänsch[b]
[a]Technische Universität Berlin, Germany
[b]German Aerospace Center (DLR), Germany

Abstract

Most automatic systems to analyze Earth observation data are designed for a particular combination of sensor and task. This limits their applicability in real world scenarios where for a certain point in time and space an acquisition of a certain sensor might not be available. We propose a cross-modal learning system that trained on multiple modalities can be applied to any of these modalities during inference time. We show that the proposed model not only maintains performance compared to the baseline approach of having independent modality specific models but also provides predictions with increased homogeneity regarding the modalities.

1 Introduction

Remote sensing is progressing at an unmatched speed leading to a continued increase of the amount, quality, and diversity of Earth observation imagery [14]. More and more satellites produce images of higher and higher spatial and spectral resolutions offering new opportunities to observe natural and artificial processes on the surface of the Earth. Despite this progress, many approaches either focus on a single modality or aim to leverage multiple modalities in parallel. The former has the disadvantage, that it creates hard constraints on the availability of the used sensor modality. This can be critical in applications that require either a timely response, i.e. an acquisition at a specific time, or a dense time series of acquisitions. If acquisitions are not available, e.g. due to cloud cover or other occlusions for optical images or simply because the sensor is not yet in the right position, one loses important information. Multimodal learning only increases those shortcomings as every sensor brings its own constraints on the acquisitions.

A typical example is the response to a natural disaster such as a flood event or earthquake [12]. While there might be machine learning based models available to aid the first responders, e.g. by detecting damaged buildings or blocked roads in satellite imagery, these are very likely to be trained on images from a specific sensor. In this case, one would need to wait until this satellite is in place to make an acquisition. If at this time point the scene is covered by clouds or smoke, one has to postpone the image acquisition further loosing precious time and making a quick response leveraging such information impossible.

The alternative is to have a system that can be applied to various sensor modalities yet provides predictions with similar characteristics for each of them.

The analysis of Earth observation data often leverages multimodal learning [4] but mostly as data fusion approaches requiring all modalities being present during inference [3]. In contrast to such multimodal methods, cross-modal models are trained on multiple modalities, e.g. imagery of different satellites, but use only a single of these modalities during prediction.

The straight-forward solution to this task is to create independent modality-specific models. This has the disadvantage that the models cannot share any information despite aiming to solve the same task and that their predictions can hardly be combined (e.g. into consistent time series) since their errors and uncertainties will be very different and modality-specific, too.

Modality translation aims to estimate how the data would look like in one modality given input from another modality, e.g. creating an optical image based on SAR data. This has been leveraged in multiple applications including estimating hyperspectral images from multi-spectral data [11], inpainting regions occluded by clouds in optical images based on SAR [1], change detection between images of different modalities [9], and densification of time series [8]. However, for use cases where each modality is equally likely to be useful during inference, selecting a target modality as basis for any further processing is arbitrary and might represent a suboptimal choice. Related to modality transfer are domain adaptation and transfer learning [18] where circumstances of the data acquisition (such as season, geographic location, etc.) might change but not the sensor itself. This is usually realized by enforcing the same distribution in source and target domain [20, 15, 16]. Closest to our method are approaches for manifold alignment aiming at finding a common embedding for all modalities [19, 10, 17], which have been used for classification [17] and visualisation of multi-/hyper-spectral imagery [7]. Most approaches are based on linear projections or their kernelized versions while deep learning approaches are sparse [5, 6].

Our approach is based on [21] where a very similar framework is proposed in the context of cross-modal retrieval between images and text. We apply the general idea in the context of multi-label classification of land use/cover from remote sensing imagery provided by different sensors, i.e. SAR (Sentinel-1) and multi-spectral (Sentinel-2).

While in this work we implement the proposed approach with only two modalities, it is in general by no means limited to that. We do not specify a target modality and do not aim to transform one modality into the another, neither on a data nor feature level. Instead, we rely on modality-specific networks to extract meaningful features that are subsequently projected into a common latent space. Different losses during the learning phase encourage that features extracted from images of different modalities fall into similar regions in this latent space, if their semantic content is similar. We do not require aligned images of the different modalities but only training datasets with unified reference data, i.e. the target variable (e.g. class labels) need to follow the same definition.

2 Methodology

In this work, we focus on Sentinel-1 and Sentinel-2 imagery. Dealing with only two modalities simplifies the problem while working jointly with SAR and multi-spectral images keeps it sufficiently challenging to remain relevant. SAR and multi-spectral images show extreme differences regarding geometry (distance vs. angle), look-angle (side-looking vs. nadir), and sensitivity (electric permittivity vs. colors) - to name a few. This means we assume a training dataset D^m for each modality m, i.e. $D^m \subset \mathcal{X}^m \times \mathcal{Y}$ where $\mathcal{X}^m$ are the input images and $\mathcal{Y}$ the reference data, i.e. in our case labels for land use/cover classes. It should be noted that while obviously the image space depends on the modality (e.g. the dimensionality and value range will differ for different sensors), the label space does not, i.e. all modality specific datasets are assumed to be consistently labeled. We also do not assume that images of different modalities are aligned.

As shown in Figure 1, the general framework [21] of the proposed approach can be roughly divided into three modules: Feature extraction, feature merging, and optimization. The first step leverages modality-specific networks to extract features $f^m(x_i^m)$ from the i-th input image x_i^m of modality m, i.e. $m \in \{S1, S2\}$ depending on whether a Sentinel-1 or Sentinel-2 image is used. These can be trained from scratch (like in the proposed work) or pre-trained on similar tasks for each modality. Their only purpose is to extract meaningful features from the imagery that are descriptive for the downstream task and any network that fulfills this quality can be used. We use a ResNet50 network for both branches. As both networks are applied to very different modalities, weights are not shared.

In general, both networks are able to extract features that are meaningful for the downstream task. However, even if the same output dimensionality would be enforced so that feature vectors of both modalities lie in the same general space, feature vectors of images of different modalities showing similar content are not likely to be close to each other. Thus, the second step in the proposed method is to use two modality-specific networks to project the modality-specific features in a shared latent space, i.e. $g^m(f_i^m)$ (where f_i^m denotes $f^m(x_i^m)$ for a less cluttered notation) with the goal that $g^{S1}(f^{S1}) \approx g^{S2}(f^{S2})$ if x^{S1} and x^{S2} show the same scene at the same time. This is achieved by fully-connected networks that take the modality-specific feature vectors to create a new vector for which modality-specific information is decreased (in the best case even erased) while information required to solve the downstream task is maintained. The two ResNet50 networks used for feature extraction produce 2048-dimensional feature vectors which are reduced by three fully-connected layers in the merging networks to 1024 dimensional vectors.

The last step of the framework are modules enabling the end-to-end optimization of the network. The first module is a network dedicated to the given downstream task, i.e. in our case multi-label classification. It is a fully-connected network that uses the feature vector computed by the merging networks to predict the semantic classes present in the image. We use the Kullback-Leibler-Divergence loss

$$\mathcal{L}_C = \frac{1}{N} \sum_{i=1}^{N} \sum_{c=1}^{C} y_i(c) \cdot \log \left(\frac{y_i(c)}{\hat{y}_i^m(c)} \right), \tag{1}$$

where y is the label distribution over C classes and $\hat{y}^m$ its prediction by the network based on an image from modality m. This loss ensures that the merged latent space retains the information necessary to solve the downstream task, i.e. that classes in the latent space are well distinguishable. However, it does not (explicitly) encourage feature vectors extracted from images of the two modalities showing the same scene to be similar to each other.

To encourage that features in the merged latent space follow similar overall statistics, we add a modality classifier, i.e. a network that given a vector determines from which modality it originated. Similar to the discriminator in a GAN, the inverse loss of the modality classifier is used to adjust the parameters of the upstream networks. The classifier itself is trained via a binary cross entropy loss

$$\mathcal{L}_M = \frac{1}{N} \sum_{i=1}^{N} m_i \cdot \log(\hat{m}_i), \tag{2}$$

where m is the actual input modality and $\hat{m}$ its prediction.

Neither the loss of the modality classifier nor the loss of the downstream classifier encourage that features in the merged latent space are similar for images showing the same scene but coming from different modalities. While the latter loss can be minimized despite having completely different distributions for both modalities, the former only ensures that the general distributions are similar. It would be possible to add a loss that ensures a minimal distance between the features extracted from images that are different acquisitions of the same scene in close temporal proximity. However, this would require spatially and temporally well aligned image pairs from all modalities which would add a significant constraint on data acquisition. Instead, the last module is a triplet constraint that minimizes the distance between images having similar semantic content while maximizing the distance between images showing different classes. For each sample x_i^a within a batch that comes from modality a, we select two images of the other modality b: One with similar x_+^b and one with different

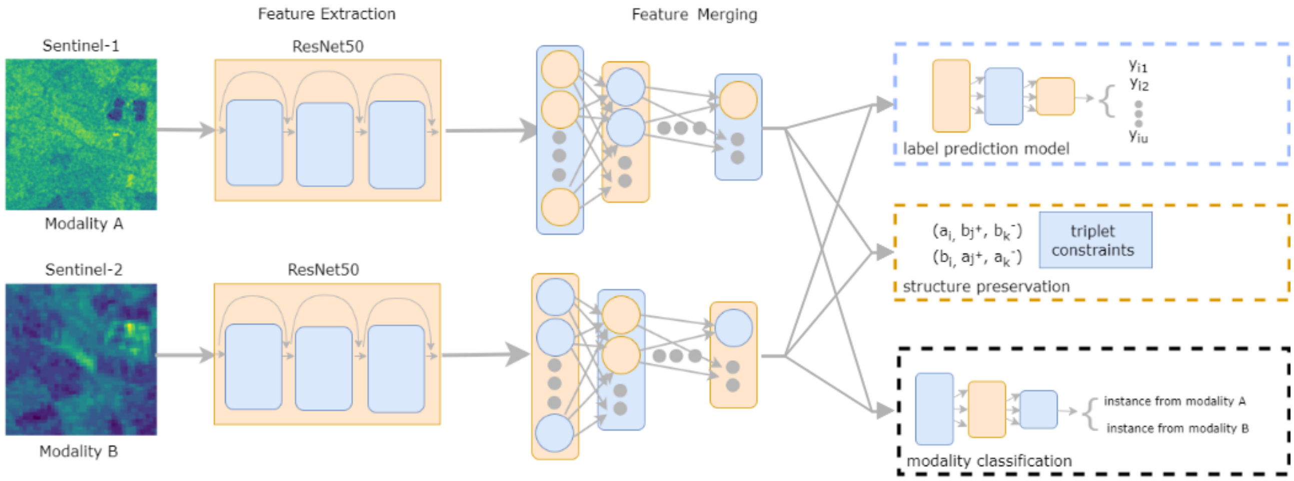

Figure 1 General framework of the proposed approach: Following [21], we use modality-specific networks to extract features from images of two different sensors (Sentinel-1 and Sentinel-2) and project them into a merged latent space. These projected features are then used for the downstream task, i.e. multi-label classification, which is solved by a network that is not modality-specific anymore. A modality classification loss and a triplet constraint aim to encourage that feature vectors of similar images are at similar positions in the latent space.

class content x^b_-. The latter is selected from a group of semantically different images based on the maximal similarity in appearance (based on cosine similarity). These image triplets are then used in a loss

$$\mathcal{L}_T = \sum_{i=1}^{N} ||x^a_i - x^b_+||_2 + \lambda \cdot \max(0, \mu - ||x^a_i - x^b_-||_2) \quad (3)$$

where λ, μ are hyperparameters that control the influence of the negative samples.

The final loss is a composition of the three individual loss functions, i.e.

$$\mathcal{L} = \alpha\mathcal{L}_T + \beta\mathcal{L}_C - \mathcal{L}_M, \quad (4)$$

where α, β are tuneable hyperparameters. The different nature of the modality loss compared to classification and triplet loss (reflected by a negative sign) is implemented as Gradient Reversal Layer [2] which ensures that the forward-propagation stays unchanged, while during back-propagation the sign of modality classifier gradient is switched.

3 Experiments

3.1 Data and Performance Metrics

We use the Sen12MS dataset [13] which provides aligned images from Sentinel-1 and Sentinel-2 together with different MODIS-derived land cover maps. Following [22], we use the simplified version of the IGBP label scheme which provides nine different classes. The low-resolution semantic map of each image is converted into a relative histogram of class occurrence which is used as a soft-label for this image. To set our work into the context of small training datasets, we limit the data to the astronomical winter (but since the images are taken from both hemispheres, winter images still contain two meteorological seasons). This gives us 31,825 images of which we use 27,825 for training and 4,000 for testing.

To evaluate the performance of the multi-label classification task, we use the F1 score which is the harmonic mean of precision $P = TP/(TP + FP)$ and recall $R = TP/(TP + FN)$ (with TP, FP, FN being the true positives, false positives, and false negatives, respectively) where we count a sample as positive for a class if the estimated probability for this class exceeds 30%. To account for the class imbalance of the Sen12MS dataset, we focus on the macro and weighted F1 scores, i.e. the weighted average of the class-wise F1 scores where the weight is uniform for the former and proportional to the number of samples of a class for the latter.

3.2 Results and Discussion

Figure 2 shows the results obtained by the proposed method in various configurations as macro and weighted F1 scores for both modalities, i.e. for Sentinel-1 (S1) and Sentinel-2 (S2). In general, the results for S2 are superior to S1 which is consistent with previous findings. The baseline consists of training the whole network on only one modality with the classification loss only, i.e. without the triplet and modality losses. This allows a fair comparison since the number of parameters and the general architecture remain the same compared to the proposed approach. This baseline achieves 0.58 (0.63) and 0.68 (0.77) in macro (weighted) F1 score for S1 and S2, respectively.

Applying the full proposed approach, i.e. using all three losses (triplet loss T, modality loss M, and classification loss C) decreases the macro F1 score to 0.55 and 0.64 (i.e. by roughly 0.03-0.04) for S1 and S2, respectively, but increases the weighted F1 score to 0.7 and 0.79 for S1 and S2 (i.e. by roughly 0.02-0.07) showing that in particular majority classes benefit. Please also note that the discrepancy of the classification performance between both modalities decreased from 0.1 (0.14) for the macro (weighted) F1

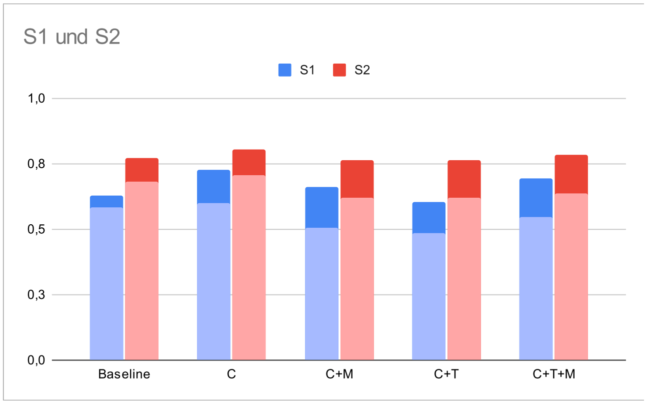

Figure 2 Obtained results in terms of macro (opaque) and weighted (half-transparent) F1 score for images of Sentinel-1 (blue) and Sentinel-2 (red). From left to right: Baseline, using only the classifier loss (C), using classifier and modality loss (C+M), using classifier and triplet loss (C+T), and using all three losses (C+T+M). The proposed approach mostly maintains the accuracy of the baseline, while using only the classification loss leads to the overall highest accuracy.

score of the baseline to 0.09 (0.09) indicating a more homogeneous classification result for the two modalities.

Leaving either triplet loss or modality loss out decreases performance in all cases. Interestingly, however, using only the classification loss leads to the overall best results. The macro (weighted) F1 score improves for all cases, i.e. to 0.6 (0.73) for S1 and 0.71 (0.8) for S2. The reason is that for the baseline all network submodules including the classification network are only trained with samples of one modality. The proposed approach, however, leverages data from both modalities practically doubling the number of samples. This is in particular true for the classification network which is truly performing twice as many forward and backprop passes. In contrast to the other settings, however, the proposed method without modality and triplet loss puts no constraints on the latent space which is then fully optimized for accuracy without regard to homogeneity of the two modalities.

Figure 3 shows the influence of the number of samples on the weighted F1 score of the baseline, the full proposed approach with all three losses, and only using the classification loss and confirms the above findings. The full approach shows results that are either on-par or superior to the baseline indicating that the additional constraints to enforce homogeneity of the two modalities in the merged latent space does not decrease the information content. On the other hand, removing these constraints leads to consistently better results in terms of accuracy.

To analyze whether the proposed approach really leads to more homogenous outputs regarding the two modalities, we exploit the fact that the Sen12MS dataset actually does provide aligned images of Sentinel-1 and Sentinel-2. After training, we compute the average histogram intersection over the test data for each image pair $(x^{S1}, x^{S2})_{i=1,\dots,N}$,

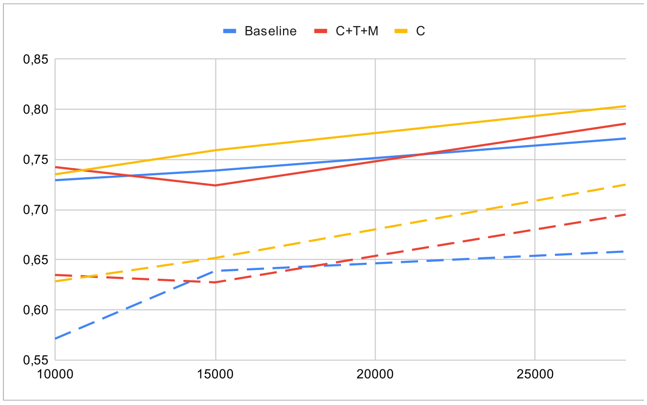

Figure 3 Influence of the number of samples on the accuracy (weighted F1 score) of the baseline (blue), the full approach with all three losses (C+T+M), and using only the classification loss (C) applied to images of either Sentinel-1 (dashed line) and Sentinel-2 (solid line).

i.e.

$$s = \frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C}\min[\hat{y}_i^{S1}(c), \hat{y}_i^{S2}(c)], \tag{5}$$

where $\hat{y}_i^m$ is the model prediction for sample x^m of modality m. The baseline achieves a score of $s = 0.63$ which is expected as both modality classifiers make mostly correct decisions. The full model achieves a score of $s = 0.67$ showing that while the baseline performance is maintained (or in some aspects even surpassed), the classification decisions are much more homogeneous regarding the two modalities. If only the classification loss is used, the score is with $s = 0.66$ only a little bit smaller showing that with improved performance the homogeneity of the classification decisions also increases.

4 Conclusion and Future Work

This work proposes a deep neural network model that aims to mitigate the dependence of current expert systems on the availability of specific modalities during inference. This is achieved by leveraging a merging network that takes image features extracted by a modality-specific network and projects them into a common latent space. Specific loss functions during the learning phase encourage features of images with similar semantic content to be at similar locations within this latent space. We show that the proposed approach is successful in maintaining (or in some aspects surpassing) the performance of the baseline (i.e. the naive approach of having independent modality-specific networks) while increasing the homogeneity of the predictions.

Future work will extend the proposed approach to more modalities. Since we do not make any assumptions about relations between the training data other then a consistent annotation, i.e. we do not require aligned images of the different modalities, extending the approach to more than two modalities is easily possible. Furthermore, more work has to be done to better understand the properties of the shared latent space.

5 Literature

[1] Jose D. Bermudez, Patrick Nigri Happ, Dario Augusto Borges Oliveira, and Raul Queiroz Feitosa. Sar to optical image synthesis for cloud removal with generative adversarial networks. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 2018.

[2] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation, 2015.

[3] Pedram Ghamisi, Behnood Rasti, Naoto Yokoya, Qunming Wang, Bernhard Hofle, Lorenzo Bruzzone, Francesca Bovolo, Mingmin Chi, Katharina Anders, Richard Gloaguen, Peter M. Atkinson, and Jon Atli Benediktsson. Multisource and multitemporal data fusion in remote sensing: A comprehensive review of the state of the art. *IEEE Geoscience and Remote Sensing Magazine*, 7(1):6–39, 2019.

[4] Luis Gomez-Chova, Devis Tuia, Gabriele Moser, and Gustau Camps-Valls. Multimodal classification of remote sensing images: A review and future directions. *Proceedings of the IEEE*, 103(9):1560–1584, 2015.

[5] Danfeng Hong, Jing Yao, Deyu Meng, Zongben Xu, and Jocelyn Chanussot. Multimodal gans: Toward crossmodal hyperspectral–multispectral image segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 59(6):5103–5113, 2021.

[6] Danfeng Hong, Naoto Yokoya, Gui-Song Xia, Jocelyn Chanussot, and Xiao Xiang Zhu. X-modalnet: A semi-supervised deep cross-modal network for classification of remote sensing data. *ISPRS Journal of Photogrammetry and Remote Sensing*, 167:12–23, 2020.

[7] Danping Liao, Yuntao Qian, Jun Zhou, and Yuan Yan Tang. A manifold alignment approach for hyperspectral image visualization with natural color. *IEEE Transactions on Geoscience and Remote Sensing*, 54(6):3151–3162, 2016.

[8] Xun Liu, Chenwei Deng, Baojun Zhao, and Jocelyn Chanussot. Multimodal-temporal fusion: Blending multimodal remote sensing images to generate image series with high temporal resolution. In *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*, pages 10083–10086, 2019.

[9] Luigi Tommaso Luppino, Michael Kampffmeyer, Filippo Maria Bianchi, Gabriele Moser, Sebastiano Bruno Serpico, Robert Jenssen, and Stian Normann Anfinsen. Deep image translation with an affinity-based change prior for unsupervised multimodal change detection. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–22, 2022.

[10] Giona Matasci, Michele Volpi, Mikhail Kanevski, Lorenzo Bruzzone, and Devis Tuia. Semisupervised transfer component analysis for domain adaptation in remote sensing image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 53(7):3550–3564, 2015.

[11] Hao Peng, Xiaomei Chen, and Jie Zhao. Residual pixel attention network for spectral reconstruction from rgb images. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2012–2020, 2020.

[12] Claudio Persello, Jan Dirk Wegner, Ronny Hänsch, Devis Tuia, Pedram Ghamisi, Mila Koeva, and Gustau Camps-Valls. Deep learning and earth observation to support the sustainable development goals: Current approaches, open challenges, and future opportunities. *IEEE Geoscience and Remote Sensing Magazine*, 10(2):172–200, 2022.

[13] M. Schmitt, L. H. Hughes, C. Qiu, and X. X. Zhu. Sen12ms – a curated dataset of georeferenced multispectral sentinel-1/2 imagery for deep learning and data fusion. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, IV-2/W7:153–160, 2019.

[14] Michael Schmitt, Seyed Ali Ahmadi, Yonghao Xu, Gülşen Taşkin, Ujjwal Verma, Francescopaolo Sica, and Ronny Hänsch. There are no data like more data: Datasets for deep learning in earth observation. *IEEE Geoscience and Remote Sensing Magazine*, 11(3):63–97, 2023.

[15] Onur Tasar, S. L. Happy, Yuliya Tarabalka, and Pierre Alliez. Semi2i: Semantically consistent image-to-image translation for domain adaptation of remote sensing data. *CoRR*, abs/2002.05925, 2020.

[16] Onur Tasar, Yuliya Tarabalka, Alain Giros, Pierre Alliez, and Sebastien Clerc. Standardgan: Multi-source domain adaptation for semantic segmentation of very high resolution satellite images by data standardization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020.

[17] Devis Tuia, Diego Marcos, and Gustau Camps-Valls. Multi-temporal and multi-source remote sensing image classification by nonlinear relative normalization. *ISPRS Journal of Photogrammetry and Remote Sensing*, 120:1–12, 2016.

[18] Devis Tuia, Claudio Persello, and Lorenzo Bruzzone. Domain adaptation for the classification of remote sensing data: An overview of recent advances. *IEEE Geoscience and Remote Sensing Magazine*, 4(2):41–57, 2016.

[19] Devis Tuia, Michele Volpi, Maxime Trolliet, and Gustau Camps-Valls. Semisupervised manifold alignment of multimodal remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 52(12):7708–7720, 2014.

[20] Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. Deep domain confusion: Maximizing for domain invariance. *CoRR*, abs/1412.3474, 2014.

[21] Bokun Wang, Yang Yang, Xing Xu, Alan Hanjalic, and Heng Tao Shen. Adversarial cross-modal retrieval. In *Proceedings of the 25th ACM International Conference on Multimedia*, MM '17, page 154–162, New York, NY, USA, 2017. Association for Computing Machinery.

[22] Naoto Yokoya, Pedram Ghamisi, Ronny Haensch,

and Michael Schmitt. 2020 ieee grss data fusion contest: Global land cover mapping with weak supervision [technical committees]. *IEEE Geoscience and Remote Sensing Magazine*, 8(1):154–157, 2020.