Urban Land Cover Classification with Missing Data Using Deep Convolutional Neural Networks

Michael Kampffmeyer, *Student Member, IEEE,* Arnt-Børre Salberg, *Member, IEEE,* and Robert Jenssen, *Member, IEEE*

***Abstract*—Automatic urban land cover classification is a classical problem in remote sensing and good urban land cover maps build the foundation for many tasks, such as e.g. environmental monitoring. It is a particularly challenging problem, as classes generally have high inter-class and low intra-class variance. A common technique to improve urban land cover classification performance in remote sensing is the fusing of data from different sensors with different data modalities. However, all modalities are rarely available for all test data, and this missing data problem poses severe challenges for multi-modal learning. Inspired by recent successes in deep learning, we propose as a remedy a convolutional neural network (CNN) architecture for urban remote sensing image segmentation trained on data modalities which are not all available at test time. We train our architecture with a cost function particularly suited for imbalanced classes, as this is a frequent problem in remote sensing, especially in urban areas. We demonstrate the method using two benchmark datasets, both consisting of optical and digital surface model (DSM) images. We simulate missing data, by assuming that the DSM images are missing during testing and show that our method outperforms both CNNs trained on optical images as well as an ensemble of two CNNs trained only on optical images. We further evaluate the potential of our method to handle situations where only some DSM images are missing during training and show that we can clearly exploit training time information of the missing modality during testing.**

***Index Terms*—Deep learning, convolutional neural networks, remote sensing, missing data, land cover classification**

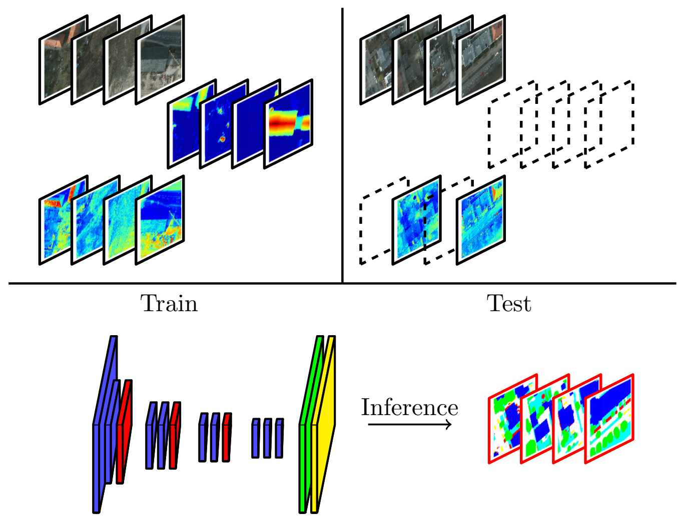


Fig. 1: We address the issue of producing urban land cover classification in a remote sensing context, when some data modalities are missing partly or all the time during the test phase as illustrated in the top right part of the Figure. Our method leverages all available training modalities (top left part of the figure) to increase overall performance when performing inference (bottom part of figure).

I. Introduction

More than half of the world population now lives in cities, and 2.5 billion more people are expected to move into cities by 2050 [1]. Although constituting only a small percentage of global land cover, urban areas significantly alter climate, biogeochemistry, and hydrology at local, regional, and global scales. Thus, in order to support sustainable urban development, accurate information on urban land cover is needed.

A frequently studied topic in land cover classification, since it often leads to improved accuracy, is data fusion. Data fusion aims to integrate the information acquired potentially with different spatial resolution, spectral bands and imaging modes from sensors mounted on satellites, aircraft and ground platforms to produce fused data that contains more detailed information than each of the individual sources [2], [3]. However, often one or more of the data sources are missing due to e.g. cloud coverage, sensor failure, or simply lack of image data. This is illustrated in Figure 1. To handle such missing data scenarios, several classification strategies have been investigated [4], [5], [6]. However, these classifiers do generally not make use of all available data sources to improve the accuracy when one or more data sources are missing.

Another challenge that is often encountered when designing classifiers for land cover mapping is class imbalance. Land covers within the area of interest are often highly imbalanced, where some land cover types are frequent, whereas others are rare. Moreover, many objects of interest in remote sensing are small compared to the overall image. A solution based on optimizing the overall classification accuracy is often not satisfactory, since "small" classes will often be suppressed [7]. In a recent paper [8], the influence of imbalanced classes was reduced by introducing weights to the cost function. Classes with few samples where given high weights, whereas classes with many samples were given small weights.

A concrete example, where all these aspects are highly likely to occur are in disaster monitoring and assessment systems. These systems are often time critical, especially in highly populated areas, and require analysis immediately after

M. Kampffmeyer and R. Jenssen are with the Machine Learning Group, UiT The Arctic University of Norway, 9019 Tromsø, Norway, http://site.uit.no/ml/
A.-B. Salberg and R. Jenssen are with the Norwegian Computing Center, 0373 Oslo, Norway.

disasters happen. Due to this requirement, it is likely that not all data modalities have been gathered, in which case it would be an advantage to still exploit the information that has been gathered during training. Another example is the monitoring of urban areas in the presence of dense cloud cover, where optical imagery can not be used and instead sensors with longer wavelength, such as synthetic aperture radar, have to be employed.

Due to the successes of deep learning architectures, convolutional neural networks (CNNs) have found increased use also in the field of remote sensing [8], [9], [10], [11], [12], outperforming more traditional approaches [13]. The strength of CNNs is the ability to learn features that exploit the spatial context, and thereby provide land cover maps with high accuracy.

In this paper, we propose a CNN for urban land cover mapping that makes use of data modalities during the test-phase that are only available during training. The proposed system will build upon the *hallucination network* strategy proposed by Hoffman et al. [14] for training CNNs for object detection from data modalities when one of the modalities is completely missing during testing. Because of the inherent difficulties in urban land cover classification caused by small objects and imbalanced classes, we specially tailor the network by incorporating a class-wise median frequency balancing approach when designing the cost function. Moreover, we extend the hallucination network idea itself to handle situations where more than one modality is missing and analyze its potential to handle both situations where data modalities are completely missing and situations where data modalities are only partly missing. Section IV describes the approach taken in detail.

A preliminary version of this paper appeared in [15]. Here, we extend our work by (i) providing a more thorough literature background discussion, (ii) extending and evaluating the methodology to more than one missing modality, (iii) expand the experiment section to include an additional dataset, (iv) evaluate the effect of medium frequency balancing on the small (imbalanced) classes, and (v) evaluate the potential of the trained model to also handle situations where both modalities are available during testing.

This paper is structured as follows. Section II provides an overview of related past works. Section III introduces the datasets used in our work and the evaluation procedure. In Section IV we present the methodology and training procedure. Experimental evaluation of the method is performed in Section V and finally in Section VI we draw conclusions and point to future directions.

II. Related Work

Our approach to segmentation builds on the recent successes that deep learning techniques have achieved for image segmentation using CNNs. Instead of requiring elaborate feature design, CNNs are able to learn relevant features from data. Even though segmentation can be viewed as a pixel-wise classification problem, currently, most state-of-the-art CNN models for image segmentation are inspired by the idea of fully convolutional pixel-to-pixel end-to-end learnable architectures [16]. These architectures generally consist of a encoder-decoder architecture, where the encoder creates a lower resolution representation of the image and the decoder produces the pixel-wise prediction from this representation. Upsampling in the decoder can be performed in several ways, with one common approach being the fractional-strided convolution (also referred to as Deconvolution) [16]. Alternative approaches avoid the explicit learning of an upsampling by using the max-pooling locations from the downsampling stage to place activations, thereby achieving a sparse upsampling map, which is then converted into a dense representation with the help of convolutional filters [17]. All layers in these networks are based on convolutions and do not make use of, as in previous approaches, fully connected layers, which allows their application on test images of varying size.

Lately, deep learning and CNNs have also increasingly been applied to remote sensing tasks such as for instance domain adaptation [18], but especially for object detection and image segmentation in remote sensing. For instance, Paisitkriangkrai et al. [19] proposed a scheme for high-resolution land cover classification using a combination of a patch-based CNN and a random forest classifier that is trained on hand-crafted features. To increase the classification accuracy further, a conditional random field (CRF) was used to smooth the final pixel labeling results. Further, Krylov et al. [10] proposes the use of a patch-based neural network to perform land cover classification in SPOT-5 imagery. More recent works, to a large extend make use of fully convolutional architectures [8], [9], [11], achieving better overall accuracy and being more computationally efficient [8], [11]. Building on the good performance of fully convolutional architectures for land cover classifications, two stage systems to object detection have been proposed, where the segmentation is followed by a classification step [20]. For scenarios, where only few labeled training images are available greedy layerwise unsupervised pretraining was proposed to extract useful features using CNNs [21] and ways of fusing heterogeneous data modalities in a land cover classification context have been proposed [12].

Missing data, is a common problem in real-world data that can occur due to for instance sensors being corrupt. In remote sensing missing data can also occur due to the properties of certain sensors. The occurrence of cloud cover in optical images or the acquisition of images during the night using optical sensors are two examples of missing data that can occur due to the inherent nature of the sensor. Common approaches to handling missing data for pattern recognition can be grouped into four main groups [22]. Incomplete data can either be deleted, leading to the removal of potentially large amounts of training data, or missing values can be imputed [23]. Alternatively, model based approaches can be used, such as e.g. the expectation-maximization algorithm by modeling the data distribution, or approaches can be designed that incorporate missing data into the machine learning approach, such as in the case of e.g. decision trees [22].

Previously, most works on handling missing data in remote sensing are based on partly missing data, situations where only parts of a modality are available. One common approach of

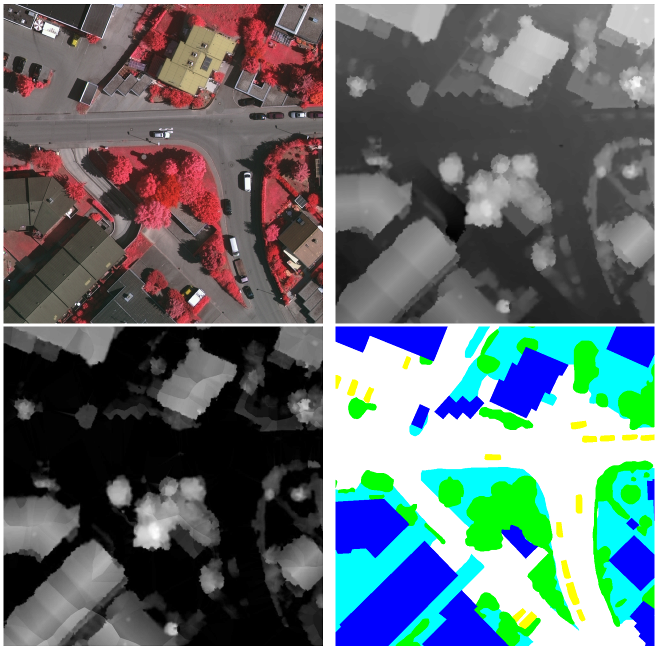


Fig. 2: A small example patch from the Vaihingen validation dataset. From left to right and top to bottom: RG+I image, DSM, normalized DSM, and ground truth image. It illustrates the difference in size between classes, such as the car class (yellow) and the building class (blue).

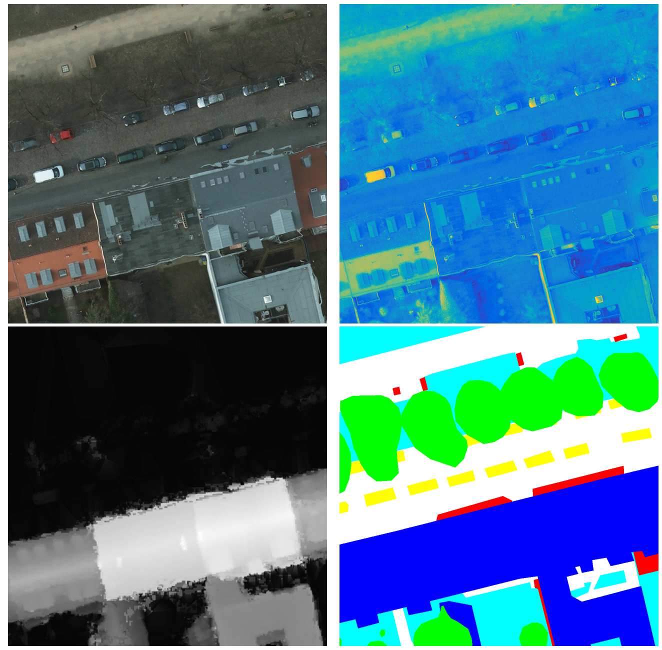


Fig. 3: A small example patch from the Potsdam validation dataset. From left to right and top to bottom: RGB image, IR, normalized DSM, and ground truth image. The DSM modality has been omitted in this figure, as we only use the normalized DSM in our experiments.

many algorithms is to simply discard these incomplete data samples [24]. However, more advanced methods have been proposed, such as for instance, by Latif and Mercier [6], who leverage Self-Organizing Maps to estimate and impute missing values. Further, various data imputation techniques have been used to fill the missing parts in the data [25]. However, imputation may often produce suboptimal results, since the assumption of missingness at random [23] may not hold when missingness appears due to the properties of the sensors. Other works that utilize the missingness directly as part of the machine learning approach include Salberg and Jenssen [5], who propose to train SVMs for all types of missingness patterns for landcover classification, and Aksoy et al. [24], who present a decision tree based model to data fusion when some modalities are partly missing. However, none of these approaches are able to exploit the knowledge provided by all data sources to improve the accuracy when one or more data sources are completely missing.

III. Datasets

The performance of our proposed method has been evaluated using two commonly used large-scale benchmark datasets for urban land cover classification that each consist of more than one data modality, namely the ISPRS Vaihingen and the ISPRS Potsdam 2D semantic labeling benchmark datasets [26]. The Vaihingen dataset consists of 33 high resolution true ortho photo images with the three bands corresponding to near infrared, red and green bands (RG+I). The images are acquired with a ground sampling distance of 9 cm and of varying size (ranging from approximately 3 million to 10 million pixels), with ground truth being available for 16 images. The Potsdam dataset consists of 38 four channel true ortho photo images with the channels corresponding to red, green, blue and infrared (RGB+I). The ground sampling distance is 5 cm with each images containing 36 million pixels and with ground truth available for 24 images. Additionally, both the digital surface model (DSM) and the normalized DSM (produced by Gerke [27]) is available for all images in both datasets. Figure 2 shows two example patches, one from the Vaihingen dataset and one from the Potsdam dataset. The datasets contain six classes, impervious surface (white in ground truth), buildings (blue), low vegetation (cyan), trees (green), cars (white) and background/clutter (red). To evaluate the approach and its use for missing data modalities in remote sensing, the normalized DSM was included only during the training phase.

We evaluate our results according to the ISPRS specification [26]. The F1-score

$$\mathrm{F1} = 2 \cdot \mathrm{precision} \cdot \mathrm{recall} / (\mathrm{precision} + \mathrm{recall}) \qquad (1)$$

is measured per class and the overall accuracy is the percentage of correctly labeled pixels. Following the specifications, the class boundaries were eroded with a disk of radius 3 and ignored in the evaluation to reduce boundary effects. For evaluation, the labeled part of the Vaihingen dataset was divided into a training set, validation set and test set containing 11, 2 and 3 images, respectively. The Potsdam dataset was similarly split into training, validation and test set (19, 2 and 3 images).

IV. Approach

In this section we describe the implementation and the training details of the proposed architecture. We start by explaining the main idea of hallucination networks in IV-A, since this framework is at the core of the proposed approach. The concept of median frequency balancing for handling small classes is explained in IV-B. In IV-C we provide details on the overall implementation and describe how we combine median frequency balancing and the idea of hallucination network to perform urban land cover classification in the presence of missing data. Finally, we propose in IV-D an extention of the halluciantion network methodology from dealing solely with one missing modality to handling the case of multiple missing modalities.

A. Hallucination Networks

There are multiple ways of combining different modalities in convolutional neural networks, such as for instance learning low-level (early layer) features over all modalities [8], [9]. However, in this work we will focus on another approach, where features are extracted independently from each modality using separate dedicated CNNs and where the final scores are combined. Such approaches have previously shown promising results for combining multiple data modalities, such as for example RGB and Depth information [28], [12].

Hallucination Networks [14] are recent attempts to use data modalities that are available solely during the training phase to improve object detection performance. They are inspired by recent works on transferring information between neural networks through network distillation, which initially was proposed as a way to compress the knowledge of large models into a smaller one by training small neural networks on the softmax output of the larger network [29].

Utilizing all available training data modalities in the case of missing data is done by adding an additional network, the hallucination network, in addition to the existing modalities. This network takes a data modality as input that we assume to be available during both training and testing and attempts to learn a mapping function from this modality to the modality that is missing. This is done by adding a loss for the mid-level features of the hallucination network to mimic the mid-level features of the potentially missing data modality. In the case that the modality is missing during testing, inference is performed by not only passing the available data to its dedicated network, but in addition also passing it to the hallucination network. The raw scores of the two networks are then averaged and the land-cover classification map is produced through a final softmax layer.

B. Medium frequency balancing

Medium frequency balancing is a weighted cross-entropy loss function that has been shown to yield good performance for imbalanced classes [30], [17], [8]. Instead of optimizing the commonly used cross-entropy loss, where small classes contribute less to the overall loss and are therefore often neglected compared to the larger classes, each class in the loss

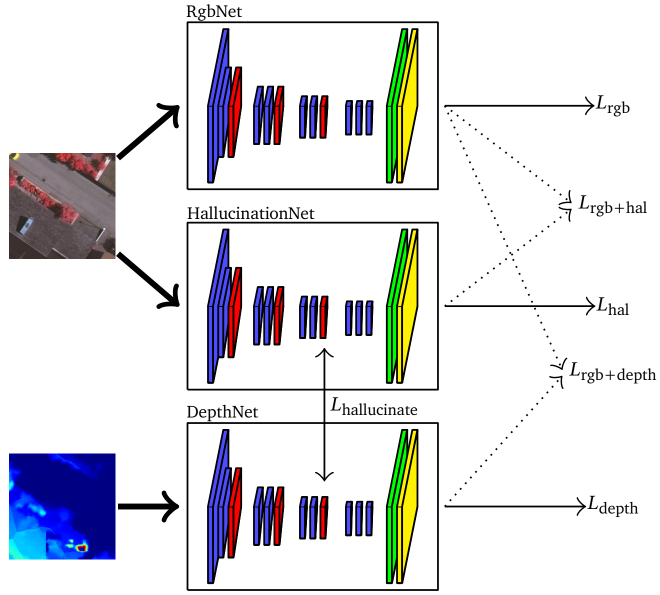


Fig. 4: Network architecture of the proposed method. Note that we choose the layer where the hallucination loss is applied (D) to be the third pooling layer based on the fact that Hoffman et al. [14] reported the best accuracy when hallucination is applied on the mid/late-level features. A more thorough analysis of this choice is left for future work.

function is weighted by the ratio of the median class frequency and the class frequency (computed over the training dataset). The updated cost formulation is

$$L = -\frac{1}{N} \sum_{n=1}^{N} \sum_{c \in \mathcal{C}} l_c^n \log\left(\widehat{p}_c^n\right) w_c \ , \tag{2}$$

where

$$w_c = \frac{\text{median}\left(\{f_c \mid c \in \mathcal{C}\}\right)}{f_c} \tag{3}$$

denotes the weight for class c, f_c the frequency of pixels in class c, $\widehat{p}_c^n$ is the softmax probability of sample n being in class c, l_c^n corresponds to the label of sample n for class c when the label is given in one-hot encoding, $\mathcal{C}$ is the set of all classes and N is the number of samples in the mini-batch.

C. Object segmentation for imbalanced classes with missing data

Our architecture for the individual networks follows the architecture of Kampffmeyer et al. [8] and is based on the fully convolutional network architecture [16]. It consists of four sets of two 3×3 convolutional layers, each set followed by a 2×2 max-pooling layer and each individual convolution layer followed by a ReLU nonlinearity and batch normalization layer [31]. The first convolutional layer has stride 2 due to memory restrictions during the test-phase when considering the large images, whereas all other convolutions are of stride 1. Three of these networks are trained jointly, one for the RG+I or RGB+I images, which will be refer to as RGB network for simplicity, one for the depth image and the hallucination network. Figure 4 illustrates the complete Architecture. Training is performed on image patches of size 256×256, which have been extracted from the original images with 50% overlap and have been flipped and rotated at 90 degree intervals as part of the data augmentation step. For the Potsdam dataset we only rotate the images due to the fact that much more data is available, making additional data augmentation unnecessary.

The total loss consists of six individual losses and is

$$L = \gamma L_{\text{hallucinate}} + L_{\text{depth}} + L_{\text{rgb}} + L_{\text{hal}} + L_{\text{rgb+depth}} + L_{\text{rgb+hal}} \ , \tag{4}$$

which is optimized using backpropagation. $L_{\text{hallucinate}}$, L_{rgb}, L_{depth} and L_{hal} are the hallucination loss, the loss of the RGB network, the loss of the depth network and the loss of the hallucination network, respectively, and $L_{\text{rgb+depth}}$ and $L_{\text{rgb+hal}}$ are the joint losses. The parameter γ is the weight parameter for the hallucination loss. The individual losses for the three networks, L_{rgb}, L_{depth} and L_{hal}, ensure that the features learned by individual networks will be useful also independently from each other. The joint losses, $L_{\text{rgb+depth}}$ and $L_{\text{rgb+hal}}$, instead ensure that the combination of the networks that can be observed during testing perform well. Here we assume that the data for the RGB network is always available, but that the depth data might be missing some or all of the time during testing. In the case that the depth data is missing during testing the combination of the RGB and the hallucination network can be used, whereas in the case where the depth data is available, the combination of the RGB and the depth network can be used. The potential of using the same architecture for both scenarios is analyzed in Section V.

Following Hoffman et al. [14] the hallucination loss is

$$L_{\text{hallucinate}} = ||\sigma(A_{\text{depth}}) - \sigma(A_{\text{hal}})||_2^2 \ , \tag{5}$$

where $\sigma(\cdot)$ is the sigmoid function and $A(\cdot)$ refers to the activation of the networks at a certain network depth D. To avoid that the depth features are adapting during the end-to-end training procedure, the learning rate for the layers before network depth D are set to zero. The hallucination loss in our experiments is based on the activations after the third pooling layer (D) and medium frequency balancing was used for all terms in the loss function except the hallucination loss.

Training details: Training is performed by first training the RGB and depth network separately and then finetuning the whole architecture end-to-end. The depth network was used to initialize the hallucination network and the weight of the hallucination network loss, γ, was set such that the hallucination loss is roughly 10 times the loss of the largest loss of the remaining terms in Eq. 4, following the example of Hoffman et al. [14]. A more thorough analysis of how to optimally set the γ parameter is left for future work.

To avoid large variations in the magnitude of the gradients, gradient clipping [32] is performed to clip outlier gradients to a an acceptable range, which is determined by monitoring the gradients during training. Training is performed using Adam [33].

D. Multiple missing modalities

In this section we propose an extention of the hallucination network to handle more than one missing data modality.

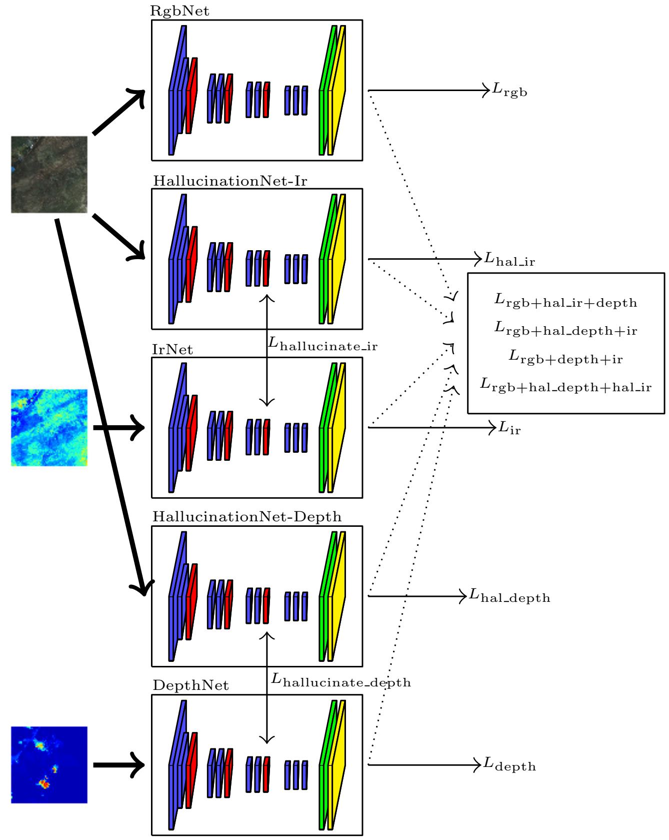


Fig. 5: Network architecture of the proposed method when handeling multiple missing modalities.

We focus especially on the Potsdam dataset as it comes with 3 unique modalities, RGB, infrared and depth. Figure 5 illustrates the updated architecture, where an additional hallucination network is introduced to handle missing infrared information. The updated cost function consists of 11 terms, which are optimized jointly following the same procedure as in the one-missing modality case, and is

$$\begin{aligned} L = {} & \gamma(L_{\text{hallucinate\_ir}} + L_{\text{hallucinate\_depth}}) + L_{\text{ir}} + L_{\text{depth}} \\ & + L_{\text{rgb}} + L_{\text{hal\_depth}} + L_{\text{hal\_ir}} + L_{\text{rgb+hal\_ir+depth}} \\ & + L_{\text{rgb+ir+hal\_depth}} + L_{\text{rgb+ir+depth}} + L_{\text{rgb+hal\_ir+hal\_depth}} \ , \end{aligned} \tag{6}$$

where in addition to the two hallucination losses, $L_{\text{hallucinate\_ir}}$ and $L_{\text{hallucinate\_depth}}$, which are used to transfer information from the training phase about the relationship between RGB and Depth images to the test stage, we have 5 losses, one for each network. As in the single missing modality scenario, these losses ensure that the models learn good individual feature representation. In addition, losses are added for the various combinations of available and missing modalities, namely, $L_{\text{rgb+hal\_ir+depth}}$, $L_{\text{rgb+ir+hal\_depth}}$, $L_{\text{rgb+ir+depth}}$ and $L_{\text{rgb+hal\_ir+hal\_depth}}$, where the subscript indicates the network losses that are combined. These losses ensure, that the network is able to utilize the availability of data modalities during testing for the cases that modalities are missing only sometimes. For instance, given that both infrared and depth information is missing, testing is performed by using the RGB and the two hallucination networks, which means that the loss $L_{\text{rgb+hal\_ir+hal\_depth}}$ should be low. On the other hand, if for example the depth information is missing, but infrared is available, we would in the test phase use the RGB network, the depth hallucination network and the infrared network, which means that the loss $L_{\text{rgb+ir+hal\_depth}}$ should be low.

V. Experiments and results

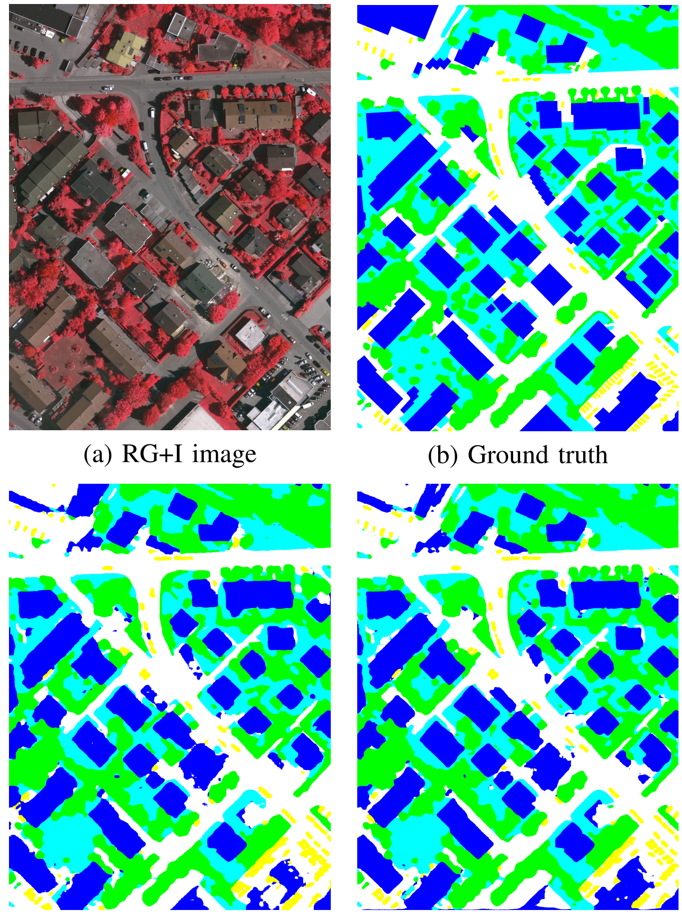
In this section we perform experiments on the Vaihingen (Section V-A) and Potsdam (Section V-B) datasets and report both qualitative and quantitative results. We further analyze for the trained Potsdam hallucination network its ability to perform inference in situations where both modalities are available. Section V-C focuses on the case of multiple missing modalities for the Potsdam scenario, where both infrared and depth measurements are missing. For all datasets, we compare our proposed approach to a CNN trained only on the RGB+I (or RG+I) image. To ensure fairness with respect to the number of network parameters, we also compare the approach to an ensemble of two CNNs trained on the RGB+I (or RG+I) images. For the ensemble the softmax output of the two CNNs was averaged during the test-phase.

A. Vaihingen

Table I shows the results for the models when trained on the Vaihingen dataset with and without medium frequency balancing. It illustrates that the hallucination network outperforms both the single RG+I model as well as the ensemble when considering overall accuracy and increases in accuracy can be observed in most classes with large increases being observed for the impervious surface class and the building class. This indicates that some of the additional information contained in the depth data is benefiting the test-phase even though the depth data is missing during testing.

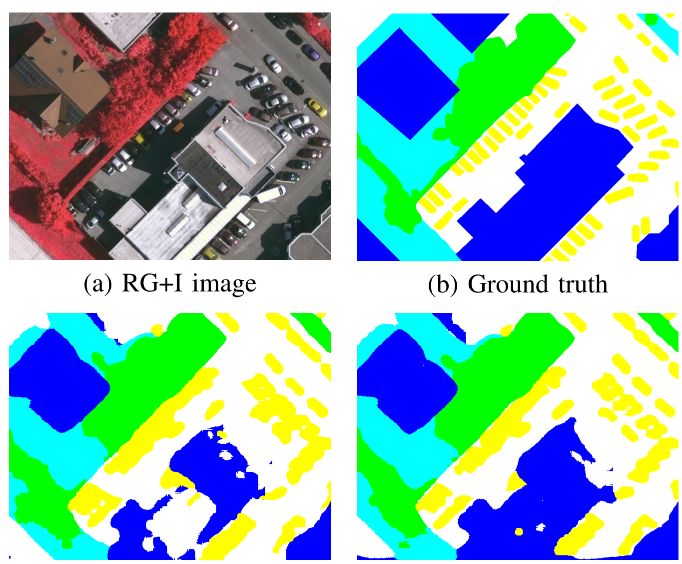
To illustrate some of the differences between the results achieved by the ensemble and the hallucination method, Figure 6 shows one of the RG+I images from the test set, the ground truth, and the achieved segmentation using the proposed method with the median frequency balancing cost function as well as the RG+I ensemble. It can be seen that both models perform well, however, when comparing the results of the two models it can be observed that the ensemble assigns more impervious surface pixels to the building class. This is as expected, as the colour and shape of some roof areas can be considered similar to impervious surfaces. Including, however, the additional depth information, as done in our proposed method, allows a better separation between these classes as the normalized depth measurements indicate differences between buildings and impervious surfaces. Looking at a close-up of the bottom right corner of the image in Figure 6, Figure 7 illustrates this difference clearly, as large parts of the building gets classified as impervious surface by the RG+I ensemble. By making use of the depth information during the training phase, the proposed model is able to capture the building class more accurately.

For comparison, we also investigate the performance of a model when the RG+I and depth data are available both during



(a) RG+I image (b) Ground truth (c) Segmentation Ensemble (d) Segmentation Hallucination

Fig. 6: Segmentation results for an image in the test dataset.



(a) RG+I image (b) Ground truth (c) Ensemble (d) Hallucination

Fig. 7: Closeup of the segmentation results for the bottom left corner of the image.

		Imp Surf		Building		Low veg		Tree		Car		Overall	
Method	MFB	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	Avg F1	Acc
RG+I	✓	90.38	87.26	91.23	86.43	81.29	72.27	89.86	86.95	90.93	83.82	88.74	83.50
RG+I-ensemble	✓	90.72	87.83	91.17	86.14	82.07	73.52	90.01	87.01	**91.77**	**85.19**	89.15	83.86
Hallucination	✓	**92.02**	**89.58**	**92.65**	**88.30**	**82.74**	**74.33**	**90.33**	**87.47**	91.31	84.30	**89.81**	**85.20**
RG+I&Depth	✓	92.54	89.33	94.57	91.42	83.40	75.68	90.35	87.38	90.82	83.42	90.34	86.33
RG+I	✗	91.73	89.13	93.50	89.81	82.70	75.18	89.96	86.36	61.45	44.38	83.87	84.96
RG+I-ensemble	✗	91.98	89.47	93.55	89.79	**83.05**	**75.74**	89.99	86.39	62.13	45.09	84.14	85.17
Hallucination	✗	**92.94**	**91.14**	**94.14**	**90.58**	82.69	74.31	**90.96**	**88.34**	**77.20**	**62.90**	**87.59**	**86.22**
RG+I&Depth	✗	93.00	91.35	94.62	91.36	83.42	75.69	90.78	87.54	66.66	50.01	85.70	86.39

TABLE I: Performance of the different models for the **Vaihingen** dataset. The F1 scores and accuracies are shown as percentages.

		Imp Surf		Building		Low veg		Tree		Car		Overall	
Method	MFB	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	Avg F1	Acc
RGB+I	✓	90.30	90.57	94.33	91.84	88.83	82.73	90.53	84.78	98.52	97.61	92.50	85.74
RGB+I-ensemble	✓	90.85	91.28	**95.08**	**92.99**	89.06	82.98	**90.77**	**85.08**	**98.67**	97.86	92.89	86.43
Hallucination	✓	**91.38**	**93.01**	94.73	91.41	**89.38**	**83.72**	90.42	84.25	98.65	**97.88**	**92.91**	**86.53**
RGB+I&Depth	✓	91.76	91.00	97.65	97.79	87.14	79.36	91.62	87.22	98.62	98.02	93.36	87.58
RGB+I	✗	90.99	92.82	0.9503	93.07	88.99	83.20	86.94	78.34	89.93	81.78	90.38	85.79
RGB+I-ensemble	✗	91.03	91.98	**95.84**	**95.14**	**89.83**	**84.62**	86.86	77.96	91.60	84.58	91.03	86.39
Hallucination	✗	**91.85**	**94.51**	95.68	93.47	89.24	83.13	**88.34**	**80.42**	**94.98**	**90.62**	**92.02**	**86.96**
RGB+I&Depth	✗	92.58	94.50	97.73	97.26	89.47	83.90	87.92	79.85	83.28	71.37	90.20	87.76

TABLE II: Performance of the different models for the **Potsdam** dataset. The F1 scores and accuracies are shown as percentages.

		Imp Surf		Building		Low veg		Tree		Car		Overall	
Method	MFB	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	Avg F1	Acc
RGB+I+Hallucination	✓	91.38	**93.01**	94.73	91.41	**89.38**	**83.72**	**90.42**	**84.25**	**98.65**	**97.88**	92.91	86.53
RGB+I&Depth	✓	**91.75**	93.00	**96.15**	**94.14**	88.93	82.78	90.20	83.91	98.55	97.64	**93.11**	**87.20**
RGB+I+Hallucination	✗	91.85	94.51	95.68	93.47	89.24	83.13	**88.34**	**80.42**	**94.98**	**90.62**	**92.02**	86.96
RGB+I&Depth	✗	**92.12**	**94.92**	**96.59**	**95.06**	**89.57**	**83.86**	86.61	77.45	91.83	85.01	91.34	**87.20**

TABLE III: Performance of the hallucination **Potsdam** model. The F1 scores and accuracies are shown as percentages.

training and testing and we provide the results in Table I. The overall accuracy for this model is, as expected, higher as more information is available for the network. However, we are able to observe the largest increases for the building and the low vegetation class. This corresponds to our intuition that the depth data is most useful for the building class and vegetation class. It also illustrates that the hallucination model, with regards to overall accuracy, is able to capture a significant part of the information contained in the depth data.

Table I also illustrate the effect of using medium frequency balancing, as it can be observed that the smaller car class in our experiments generally performs much worse when not using medium frequency balancing. Overall it can be seen that the overall accuracy drops slightly when using the balanced cost function, however, due to the large improvements on the smaller classes this appears to be a negligible difference.

B. Potsdam

Table II illustrates the results when performing the same set of experiments on the Potsdam dataset. Overall we can observe similar results, where the overall accuracy for the hallucination network outperforms both the RGB+I and the ensemble method. However, the overall increase in accuracy is lower for the Potsdam dataset, which the authors hypothesize is due to the higher resolution, of the images. Due to the higher resolution there are more factors in the RGB+I bands, which make it possible to distinguish between for example buildings and impervious surfaces. Table II also shows that the medium frequency balancing has an effect for the Potsdam dataset, however, also due to the higher resolution, resulting in cars being larger in the small training image patches, the effect is lower compared to the Vaihingen dataset.

Further, we explore the effect of using the trained hallucination model to evaluate images during the test phase where both RGB+I and Depth images are available for the Potsdam dataset. To illustrate this, we take the trained hallucination model and evaluate its performance when both image modalities are available during testing. The RGB image is fed only into the RGB part of the hallucination network and the depth image is fed into the hallucination networks depth network. The final classification is again produced by applying a softmax to the averaged raw scores of the two networks. Note, no additional training is performed and the hallucination model is applied as is. These results are presented in Table III, which also, to allow the reader to perform an easier comparison, includes a duplicate of the hallucination networks performance when the data is missing (from Table II). It can be seen that the hallucination model, when presented with the depth modality also during testing, performs better than when the modality is missing. This means that the trained hallucination

		Imp Surf		Building		Low veg		Tree		Car		Overall	
Method	MFB	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	Avg F1	Acc
RGB	✓	89.79	90.01	93.95	91.41	87.60	81.14	88.68	81.57	98.60	97.81	91.72	84.77
RGB-ensemble	✓	89.89	89.64	94.71	92.80	**89.15**	**83.71**	89.15	82.16	98.75	**98.03**	92.33	85.55
Hallucination	✓	**90.91**	**92.16**	**95.30**	**93.10**	89.12	83.03	**89.99**	**83.41**	**98.76**	97.95	**92.82**	**86.56**
RGB&I&Depth	✓	92.03	92.81	97.35	96.32	88.70	82.91	88.64	81.15	98.46	97.46	93.04	87.49

TABLE IV: Performance of the different models for the **Potsdam** dataset when we assume multiple missing modalities, namely infrared and depth. The F1 scores and accuracies are shown as percentages.

model can use the depth information when it is not missing to achieve better performance, and can therefore be used both in situations where both modalities are available, as well as in situations where the depth information is missing. This flexibility, however, comes at a price, as the overall accuracy for the model when both modalities are available is slightly lower than RGB+I&Depth in Table II, where the network is only optimized for situations where both modalities are available.

C. Multiple missing modalities

In this experiment, we evaluate the proposed method on the multiple missing modality scenario, where we assume that the RGB data modality of the Potsdam dataset is available, while the infrared and the depth modalities are missing. Table IV shows the results for the different methods. It can, like in the case of a single missing modality, be seen that the Hallucination network outperforms both the RGB network and the RGB-ensemble network, while performing worse than the network trained on all available modalities.

VI. Conclusions and Future Work

In this paper we have proposed a method for image segmentation in urban remote sensing that makes use of data modalities that are only available during the training phase. Our experiments show that the method performs better than both a single model using only the available data as well as an ensemble of two models. Additionally, by making use of the medium frequency balancing cost function, we achieve good performances on small classes. We therefore consider it an attractive choice for handling missing data in urban remote sensing.

The drawback of the proposed approach for multiple missing data modalities is the fact that it does not scale well with the number of (potentially) missing modalities, as the number of losses grows exponentially. This is in many cases not problematic, as it is common to only consider a limited amount of data modalities. However, in future work we intend to extend this work to a larger amount of missing modalities, trying to devise end-to-end learnable networks that in a scalable way can 'hallucinate' several modalities and that can provide a single model for all combinations of modalities. Also, currently we assume that one of the data modalities is always available, for example in the Vaihingen example we assume that the true ortho photo is available during training and testing. However, in future work we would like to explore ways of avoiding this restriction. Further, we will investigate transfer learning methodology and the usage of pretrained models for the sub-networks where pretrained models are available, such as for example in the case where we have optical image data. Another potential research direction is to utilize larger networks that achieve state-of-the-art performances on segmentation tasks, however, as the focus of this work was to investigate the potential for handeling missing data and not achieving overall state-of-the-art performance, this is left for future work.

Acknowledgment

We gratefully acknowledge the support of NVIDIA Corporation with the donation of the GPU used for this research and the German Society for Photogrammetry, Remote Sensing and Geoinformation (DGPF) for providing the Vaihingen and the Potsdam data set [34]. This work was partially funded by the Norwegian Research Council FRIPRO grant no. 239844 on developing the *Next Generation Learning Machines* and IKTPLUSS grant no. 270738 on *Deep Learning for Health*.

References

[1] D. Un, "World urbanization prospects: The 2014 revision," 2015.

[2] J. Zhang, "Multi-source remote sensing data fusion: status and trends," *International Journal of Image and Data Fusion*, vol. 1, no. 1, pp. 5–24, 2010.

[3] F. Bovolo and L. Bruzzone, "The time variable in data fusion: a change detection perspective," *IEEE Geoscience and Remote Sensing Magazine*, vol. 3, no. 3, pp. 8–26, 2015.

[4] A.-B. Salberg, "Land cover classification of cloud contaminated multitemporal high-resolution images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 49, no. 1, pp. 377–387, 2011.

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

[6] B. A. Latif and G. Mercier, *Self-Organizing Maps for processing of data with missing values and outliers: application to remote sensing images*. INTECH, 2010.

[7] A. Estabrooks, T. Jo, and N. Japkowicz, "A multiple resampling method for learning from imbalanced data sets," *Computational Intelligence*, vol. 20, no. 1, pp. 18–36, 2004.

[8] M. Kampffmeyer, A.-B. Salberg, and R. Jenssen, "Semantic segmentation of small objects and modeling of uncertainty in urban remote sensing images using deep convolutional neural networks," in *Proceedings IEEE Conferance Computer Vision Pattern Recognition Workshops*, 2016, pp. 1–9.

[9] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez, "Fully convolutional neural networks for remote sensing image classification," in *2016 International Geoscience and Remote Sensing Symposium*, 2016, pp. 5071–5074.

[10] V. A. Krylov, M. De Martino, G. Moser, and S. B. Serpico, "Large urban zone classification on spot-5 imagery with convolutional neural networks," in *2016 International Geoscience and Remote Sensing Symposium*. IEEE, 2016, pp. 1796–1799.

[11] M. Volpi and D. Tuia, "Dense semantic labeling of subdecimeter resolution images with convolutional neural networks," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 2, pp. 881–893, 2017.

[12] N. Audebert, B. Le Saux, and S. Lefèvre, "Fusion of heterogeneous data in convolutional networks for urban semantic labeling," in *Urban Remote Sensing Event (JURSE), 2017 Joint*. IEEE, 2017, pp. 1–4.
[13] A. Lagrange, B. Le Saux, A. Beaupère, A. Boulch, A. Chan-Hon-Tong, S. Herbin, H. Randrianarivo, and M. Ferecatu, "Benchmarking classification of earth-observation data: from learning explicit features to convolutional networks," in *2015 International Geoscience and Remote Sensing Symposium*, 2015, pp. 4173–4176.
[14] J. Hoffman, S. Gupta, and T. Darrell, "Learning with side information through modality hallucination," in *Proceedings IEEE Conferance Computer Vision Pattern Recognition (CVPR)*, June 2016.
[15] M. Kampffmeyer, A.-B. Salberg, and R. Jenssen, "Urban land cover classification with missing data using deep convolutional neural networks," in *2017 International Geoscience and Remote Sensing Symposium*, 2017.
[16] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings IEEE Conferance Computer Vision Pattern Recognition*, 2015, pp. 3431–3440.
[17] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *arXiv preprint arXiv:1511.00561*, 2015.
[18] E. Riz, B. Demir, and L. Bruzzone, "Domain adaptation based on deep denoising auto-encoders for classification of remote sensing images," in *SPIE Remote Sensing*. International Society for Optics and Photonics, 2016, pp. 100 040K–100 040K.
[19] S. Paisitkriangkrai, J. Sherrah, P. Janney, V.-D. Hengel *et al.*, "Effective semantic pixel labelling with convolutional networks and conditional random fields," in *Proceedings IEEE Conferance Computer Vision Pattern Recognition Workshops*, 2015, pp. 36–43.
[20] N. Audebert, B. L. Saux, and S. Lefèvre, "Segment-before-detect: Vehicle detection and classification through semantic segmentation of aerial images," *Remote Sensing*, vol. 9, no. 4, p. 368, 2017.
[21] A. Romero, C. Gatta, and G. Camps-Valls, "Unsupervised deep feature extraction for remote sensing image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 3, pp. 1349–1362, 2016.
[22] P. J. García-Laencina, J.-L. Sancho-Gómez, and A. R. Figueiras-Vidal, "Pattern classification with missing data: a review," *Neural Computing and Applications*, vol. 19, no. 2, pp. 263–282, 2010.
[23] J. L. Schafer, *Analysis of incomplete multivariate data*. CRC press, 1997.
[24] S. Aksoy, K. Koperski, C. Tusk, and G. Marchisio, "Land cover classification with multi-sensor fusion of partly missing data," *Photogrammetric Engineering & Remote Sensing*, vol. 75, no. 5, pp. 577–593, 2009.
[25] H. Shen, X. Li, Q. Cheng, C. Zeng, G. Yang, H. Li, and L. Zhang, "Missing information reconstruction of remote sensing data: A technical review," *IEEE Geoscience and Remote Sensing Magazine*, vol. 3, no. 3, pp. 61–85, 2015.
[26] "ISPRS 2d semantic labeling contest," http://www2.isprs.org/commissions/comm3/wg4/semantic-labeling.html.
[27] M. Gerke, "Use of the stair vision library within the isprs 2d semantic labeling benchmark (vaihingen)," Technical report, University of Twente, 2015., Tech. Rep., 2015.
[28] S. Gupta, R. Girshick, P. Arbeláez, and J. Malik, "Learning rich features from rgb-d images for object detection and segmentation," in *European Conference on Computer Vision*. Springer, 2014, pp. 345–360.
[29] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.
[30] D. Eigen and R. Fergus, "Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 2650–2658.
[31] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.
[32] R. Pascanu, T. Mikolov, and Y. Bengio, "Understanding the exploding gradient problem," *Computing Research Repository (CoRR) abs/1211.5063*, 2012.
[33] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
[34] M. Cramer, "The DGPF-test on digital airborne camera evaluation–overview and test design," *Photogrammetrie-Fernerkundung-Geoinformation*, no. 2, pp. 73–82, 2010.

Michael Kampffmeyer received Masters degrees at the Physics and Technology (2014) and Computer Science Department (2015) at UiT The Arctic University of Norway, Tromsø, Norway, where he currently pursues a PhD in Machine Learning. His research interests include the development of unsupervised deep learning methods for representation learning and clustering by utilizing ideas from kernel machines and information theoretic learning. Further, he is interested in computer vision, especially related to remote sensing and health applications. His paper "Deep Kernelized Autoencoders" with S. Løkse, F. M. Bianchi, R. Jenssen and L. Livi won the Best Student Paper Award at the Scandinavian Conference on Image Analysis, 2017. Since September 2017, he has been a Guest Researcher with Carnegie Mellon University, Pittsburgh, PA, USA, where he will stay until June 2018.

Arnt-Børre Salberg (M03) received the Diploma in applied physics and the Ph.D. degree in physics from the University of Tromsø, Tromsø, Norway, in 1998 and 2003, respectively. From August 2001 to June 2002, he was a Visiting Researcher with the U.S. Army Research Laboratory, Adelphi, MD, USA. From February 2003 to December 2005, he had a postdoctoral and research position with the Institute of Marine Research, Tromsø, Norway. From December 2005 to October 2008, he was the Head of R&D with Dolphiscan AS, Moelv, Norway. Since October 2008, he has been with the Norwegian Computing Center, Oslo, Norway, where he is currently a Senior Research Scientist in earth observation. His research interests are in the areas of earth observation, signal and image processing, machine learning, and computer vision.

Robert Jenssen (M02) received the PhD (Dr. Scient.) in Electrical Engineering from the University of Tromsø, in 2005. Currently, he is an Associate Professor with the Department of Physics and Technology, UiT The Arctic University of Norway, Tromsø, Norway. He directs the UiT Machine Learning Group: http://site.uit.no/ml. The group's focus is on innovating information theoretic learning, kernel machines, and deep learning research, an activity that is funded, e.g., over the Norwegian Research Council's prestigious FRIPRO program, grant 239844 (Next-Generation Learning Machines). He is also a Research Professor with the Norwegian Computing Center, Oslo, Norway. He was a Guest Researcher with the Technical University of Denmark, Kongens Lyngby, Denmark, from 2012 to 2013, with the Technical University of Berlin, Berlin, Germany, from 2008 to 2009, and with the University of Florida, Gainesville, FL, USA, from 2002 to 2003. Jenssen was a senior researcher at the Norwegian Center for E-Health Research, 2012-2015. Jenssen received the Honorable Mention for the 2003 Pattern Recognition Journal Best Paper Award, the 2005 IEEE ICASSP Outstanding Student Paper Award, and the 2007 UiT Young Investigator Award. His paper Kernel Entropy Component Analysis was the featured paper of the May 2010 issue of the IEEE Transactions on Pattern Analysis and Machine Intelligence. In addition, his paper Kernel Entropy Component Analysis for Remote Sensing Image Clustering with L. Gomez-Chova and G. Camps-Valls received the IEEE Geoscience and Remote Sensing Society Letters Prize Paper Award in 2013. His student M. Kampffmeyer won the Best Student Paper Award for the paper "Deep Kernelized Autoencoders" at the Scandinavian Conference on Image Analysis, 2017. Jenssen is on the IEEE Technical Committe on Machine Learning for Signal Processing (MLSP), he is the president of the Norwegian section of IAPR (NOBIM - nobim.no) and serves on the IAPR Governing Board. He is an associate editor with the journal Pattern Recognition since 2010.