Shared-space Autoencoders with Randomized Skip Connections for Building Footprint Detection with Missing Views

Giannis Ashiotis[1], James Oldfield[1,3], Charalambos Chrysostomou[1], Theodoros Christoudias[1,2], and Mihalis A. Nicolaou[1]

[1] Computation-based Science and Technology Research Center, The Cyprus Institute, Cyprus
[2] Climate and Atmosphere Research Center, The Cyprus Institute, Cyprus
[3] School of Electronic Engineering and Computer Science, Queen Mary University of London, UK

Abstract. Recently, a vast amount of satellite data has become available, going beyond standard optical (EO) data to other forms such as synthetic aperture radars (SAR). While more robust, SAR data are often more difficult to interpret, can be of lower resolution, and require intense pre-processing compared to EO data. On the other hand, while more interpretable, EO data often fail under unfavourable lighting, weather, or cloud-cover conditions. To leverage the advantages of both domains, we present a novel autoencoder-based architecture that is able to both (i) fuse multi-spectral optical and radar data in a common shared-space, and (ii) perform image segmentation for building footprint detection under the assumption that one of the data modalities is missing–resembling a situation often encountered under real-world settings. To do so, a novel randomized skip-connection architecture that utilizes autoencoder weight-sharing is designed. We compare the proposed method to baseline approaches relying on network fine-tuning, and established architectures such as `UNet`. Qualitative and quantitative results show the merits of the proposed method, that outperforms all compared techniques for the task-at-hand.

Keywords: Shared-space · Footprint Detection · Missing Views.

1 Introduction

Deep learning is becoming a necessity for tackling problems that emerge in the analysis of geospatial data [24], as researchers are faced with an ever-increasing volume of data, with sizes exceeding 100 petabytes, and the need to interpret and make predictions in diverse contexts. In this paper, we focus on the problem of detecting building footprints from satellite images and radars. This task carries significant impact, and is a necessary step for a wide range of applications, such as estimating economic factors [17], disaster and crisis response [35, 7],

human activity monitoring and urban dynamics [9], as well as population density estimation[25][4].

Inspired by the recent `Spacenet`[5] challenge, in this paper we tackle the task of multi-view learning given diverse data from multiple sensors. Concretely, we focus on fusing Synthetic-Aperture Radar (SAR) images, along with their electro-optical (EO) counterparts. Using SAR and EO data in-tandem is a common approach, since these multiple data views are considered to hold complimentary information. Specifically, SAR data are becoming more relevant, as the specific wavelengths used can penetrate clouds (Fig. 1), and can carry significant information even when captured in unfavourable weather conditions - while can also be collected independently of the day-night cycle. However, optical data - although easier to interpret and usually of higher resolution - fail to capture meaningful information when occlusions are present in the optical range, for example when insufficient lighting is present, or when the area is covered with clouds. We adopt the challenging setting where EO data is only available during training, and considered as a missing modality (view) during test-time. This is a realistic assumption inspired from real-world settings, where as aforementioned, EO data are likely to be missing due to conditions at capture time.

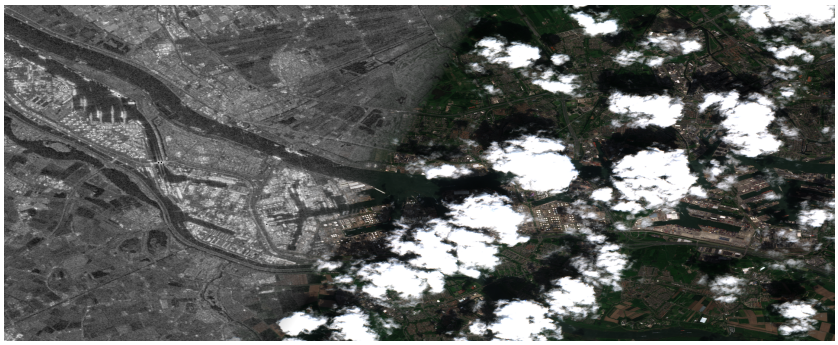

Fig. 1: SAR (left) and EO (right) composite from the Rotterdam area, with data collected on the same day by the Sentinel-1 and Sentinel-2 satellites.

The straightforward approach towards such a challenging setting is to pre-train a network with the view that is missing at test time, and subsequently *fine-tune* the network using data available both during training and testing - an approach commonly employed in satellite imagery analysis [28, 10, 29, 1]. At the same time, several approaches have been proposed on multi-view learning for

[4] Human activity monitoring and population density estimation are critical steps for building robust epidemiological models to tackle global events such as pandemics and natural disasters

[5] https://spacenet.ai/sn6-challenge/

satellite imagery for a variety of tasks [22, 18, 19]. However, these methods are *not* tailored for handling missing modalities at test-time, and are thus unsuitable for the specific setting under consideration. In this light, we propose a novel, multi-view method for building footprint detection that introduces a weight-sharing mechanism for learning a common shared-space where the input modalities are fused. To retain the advantages of skip connections in terms of retaining high-frequency information and facilitate inference with missing modalities, we further propose a novel randomized skip connection mechanism. In summary, the contributions of this work are as follows:

- We propose a segmentation method based on weight-sharing, that enforces a shared latent-space that leverages information from multiple satellite views at train-time. As shown, this facilitates learning more informative representations at test-time given entirely missing views.
- We introduce the concept of randomized skip-connections to *enhance* the shared-space representations, whilst maintaining the benefits of propagating high-frequency information directly to later layers in the network - without circumventing the shared space.
- Through a set of rigorous qualitative and quantitative experiments, we demonstrate how the proposed method outperforms typically used architectures such as `UNet`[26], as well as other commonly used fine-tuning approaches, in the presence of missing modalities.

The rest of the paper is organized as follows. In Section 2, we provide a brief summary of related work in the areas of building footprint detection using SAR, image segmentation, transfer-learning, as well as multi-view fusion. In Section 3 we present the proposed methodology, while in Section 4 we describe the employed dataset along with the relevant pre-processing steps. Finally, in Section 5 we present both qualitative and quantitative results demonstrating the merits of the proposed method.

2 Related Work

Building detection using SAR data. Several works have been proposed for building footprint detection using SAR data. In [33], an approach where SAR data is used to extract lines defined by building faces, which are then used in conjunction with the corresponding optical data. In [32], high resolution images are used to generate Digital Surface Models (DSM) of urban areas using Markovian fusion. In [39], Conditional Random Fields (CRF) were used to detect building footprints on pairs of high-resolution interferometric SAR (InSAR) and orthorectified images. More recently, deep learning approaches such as [27] have been proposed. For a detailed review on deep learning applications on SAR data, the reader is referred to [42, 41].

Image segmentation. The goal of semantic image segmentation is to map an input image to a matrix of class predictions for each pixel. A whole range of modern methods and architectures exploiting the power of deep neural networks have

been proposed in the literature, like the Fully Convolutional Networks (FCN) [28], and occasionally even Recurrent Neural Networks (RNN) [37, 36], and adversarial losses [16]. But the most prominent architectures used for the purposes of semantic image segmentation are based on the encoder-decoder paradigm. The seminal work of Ronneberger et al. combined in UNet a general-purpose autoencoder-based architecture, with the use of symmetric skip connections, in order to allow for fine-grained details to be recovered in the prediction. SegNet [3] uses a similar encoder-decoder architecture, where the maxpooling indices from the encoding sequence are used in the upsampling procedure. ResNet-DUC [38] is another encoder-decoder based architecture where the encoder part is made up of ResNet [12] blocks while the upsampling is handled by a series of Dense Upsampling Convolutions (DUC). DeepLab [4] makes use of atrous convolutions and atrous spatial pyramid pooling to better handle variations in scale, while it also adds a fully connected Conditional Random Field (CRF) to improve localization performance. FC-DenseNet [30] takes the symmetric skip connections of the UNet and combines them with DenseNet [13] blocks, in which each layer takes input from all the preceding layers in the block. This allows the creation of very deep networks of this kind.

Transfer Learning. By pretraining on comprehensive image datasets such as ImageNet [6], multiple works have achieved state of the art results and/or faster convergence for many segmentation tasks [28, 11, 10, 15], including the SpaceNet baseline model [29]. Recent work [29, 15] has also shown that improvements can be gained by pretraining on datasets more closely aligned to the task at hand, such as the Carvana dataset [1]. In attempts to leverage multiple satellite views, the SpaceNet baseline also pretrains the networks on the more informative EO satellite views. We build off such ideas, and propose an approach that models both domains simultaneously at train-time.

Multi-view fusion. A huge range of different types of image satellite data are readily available [2, 5, 40]–each providing their own unique benefits and drawbacks. For this reason, many detection and segmentation networks employ techniques to fuse these views to combine the useful features of each. Multi-view satellite fusion methods have been successfully utilized for tasks including crop segmentation [22], target detection [18, 19], and a whole host of other application areas [8, 23, 20].

In contrast to such approaches however, in our case the EO modality is missing at test-time. We thus design our network to leverage both views at train-time to learn a more informative shared representation: inspired from works on coupled and Siamese networks [21, 14], we design a weight-sharing scheme to map the two views to a shared representation.

3 Methodology

In this section we describe the proposed multi-view shared-space method that facilitates building footprint extraction under missing modalities. We first de-

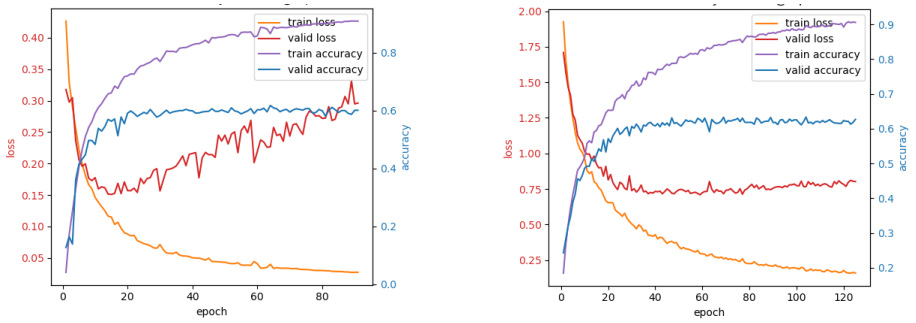

Fig. 2: Learning curves of the models trained with (a) BCE only vs with (b) BCE+DL. As can be seen, including the DL loss leads to better stability, and higher validation accuracy.

scribe the task at hand in Section 3.1. We then detail how we enrich the SAR representations using the EO data available at train time, using a weight-sharing mechanism (Section 3.2) and randomized skip-connections (Section 3.3). Finally, we detail our choice of loss functions used to train the network in Section 3.4. An overview of the proposed method is visualized in Fig. 3.

3.1 Problem setting

Our goal is to learn a mapping f from satellite image data $\mathcal{X} \in \mathbb{R}^{V \times C \times H \times W}$ (comprised of V views) to its corresponding ground-truth binary mask $\mathbf{Y} \in \mathbb{R}^{H \times W}$, from which we extract its building footprints. In the setting of the SpaceNet challenge we have multiple informative views at train-time, including SAR and EO imagery. However, the EO view is *missing* at test-time, and therefore can only be leveraged during training.

To address this problem, we adopt a UNet [26] as a base segmentation network (as is commonly employed for segmentation of satellite imagery [34]), comprised of an encoder and decoder with symmetric skip connections between these two sub-networks. However, one cannot straight-forwardly pass the entire data tensor $\mathcal{X}$ through such a network at test-time, due to these missing views. We therefore propose accordingly a novel method, building on this UNet architecture, to facilitate learning a shared latent representation of both views that can lead to more accurate mask predictions at test-time, whilst requiring only the SAR data.

Throughout the rest of the paper, we denote view $v \in \{E, S\}$ (denoting the EO and SAR views respectively; both comprised of 4 channels) of our satellite data as $\mathcal{X}^{(v)} \in \mathbb{R}^{C \times H \times W}$, scalars as lower-case Latin letters p, and random variables as lower-case Greek letters ψ.

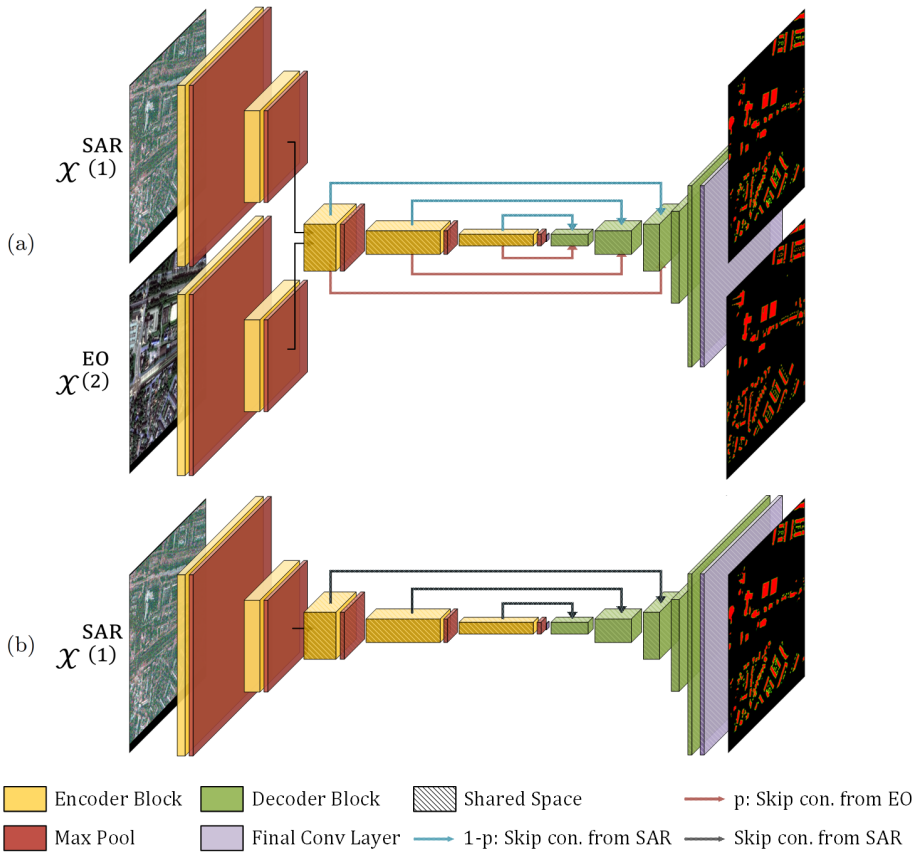

Fig. 3: An overview of the proposed method for enforcing a shared space in our segmentation network. During training (a), we first process the two images $\mathcal{X}^{(i)}$ separately with a series of view-specific layers to transform each to a common representation. We then pass these representations from both views through a shared encoder and decoder to generate the corresponding masks predictions, as a means of encouraging the most informative features of both views to be represented in the common encodings. We further enforce the shared-space constraint by introducing stochastic skip connections (shown with red and blue lines) to mix the representations. During testing (b), the available modality (SAR) can be straightforwardly utilized with the respective encoder, while still leveraging information from the missing modality, infused in the shared-space representation obtained during training.

3.2 Enforcing a shared space

In order to jointly utilize all views of our satellite data at train-time and learn a shared representation, we make the assumption that they live in the same low-dimensional subspace. To enforce such a shared space, we introduce two siamese *specific* encoders, followed by a final *shared* encoder and decoder. The two specific encoders first map the two separate views to this shared representation, and the single shared encoder and decoder take these common representations and transform them back to image space. By processing both views this way, the common representation extracted from the SAR data can be influenced and enriched by the specific information present in the EO data, without requiring any access to it whatsoever at test-time.

3.3 Randomized Skip Connections

Symmetric skip connections are extremely useful for image transformation tasks that need to retain some semblance of the input image, due to the provided ability for the network to pass low-level image information directly to later layers of the network. In this section, we propose a modification to the skip connection paradigm that not only retains such a desirable property, but jointly encourages shared representations of the multiple satellite views.

Concretely, the representation $\mathcal{D}_i$ at the i^{th} layer in the decoder is concatenated with the $N-i^{\text{th}}$ layer of the shared encoder $\mathcal{E}_{N-i}$'s representations of the multiple views in a stochastic manner. We thus modify the i^{th} activation in the decoder $\mathcal{D}_i$ with a skip connection as

$$\mathcal{D}_i := \left[\psi_i \mathcal{E}^{(S)}_{N-i} + (1-\psi_i)\mathcal{E}^{(O)}_{N-i}, \mathcal{D}_i\right], \tag{1}$$

where $\psi_i \overset{\text{i.i.d.}}{\sim} \text{Bern}(p)$, and $[.,.]$ denotes channel-wise concatenation. At train-time $p \in [0,1]$, and at test-time we fix the parameter $p := 1$ to deterministically pass the only available SAR representations.

We posit that by sending a mixture of representations of both views via the skip connections, we further encourage the shared representations to be generic and retain the informative features from both views in order to best predict the corresponding mask. We show experimentally in Section 5.2 the benefit of using such non-deterministic skip connections at train-time. We note that we recover the vanilla skip connection setup when each random variables x_i's 'success' parameter is set to $p = 0$ or $p = 1$.

3.4 Loss Functions

In order to train the network to map input images $\mathcal{X}^{(i)}$ to its binary mask counterpart $\mathbf{Y}$, we impose a typical pixel-wise binary cross-entropy loss, defined

as

$$\mathcal{L}_B = \mathbb{E}_{\mathcal{X}^{(i)},\mathbf{Y}}\Bigg[-\frac{1}{CHW}\sum_{c,h,w} y_{hw}\log f(\mathcal{X}^{(i)})_{chw} \\ +(1-y_{hw})\log(1-f(\mathcal{X}^{(i)})_{chw})\Bigg], \tag{2}$$

where f is the segmentation network, and where we compute the average BCE loss over all spatial and channel dimensions in each image.

We also impose, in addition to this regular BCE loss, the so-called Dice Loss (DL) [31]. This objective more closely encodes our goal of having high Intersection over Union (IoU) of the predicted and ground-truth masks, and is defined as

$$\mathcal{L}_D = \mathbb{E}_{\mathcal{X}^{(i)},\mathbf{Y}}\left[1 - \frac{2\sum_{c,h,w} y_{hw} f(\mathcal{X}^{(i)})_{chw}}{\sum_{c,h,w} y_{hw} + \sum_{c,h,w} f(\mathcal{X}^{(i)})_{chw}}\right]. \tag{3}$$

We find the additional loss term defined in Eq. (3) to increase the stability of the model's training process, along with reducing overfitting. We show this impact of the two loss terms by plotting the level curves for the model trained with the BCE loss and BCE+DL in Fig. 2a and Fig. 2b respectively. We compute the total loss as an average over both views of the satellite images, so as to facilitate the weight sharing described in section Section 3.2. This leads to the final combined objective for our segmentation network

$$\mathcal{L} = \lambda_D \mathcal{L}_D + \lambda_B \mathcal{L}_B, \tag{4}$$

where λ_D, λ_B are the weights for the two loss separate loss terms.

4 Data and Data Preprocessing

4.1 Data

For the training and testing of our model we used the publicly available training dataset from the SpaceNet 6 challenge. It consists of 3401 half-meter resolution SAR images (provided by Capella Space) together with their half-meter resolution RGB-NIR counterparts (provided by Maxar's WorldView 2 satellite) of the city of Rotterdam. The RGB-NIR images were obtained through reconstruction using several other EO images that were also provided in the dataset. In addition, annotations for over 48,000 building footprints were provided, together with look-angle information (north or south facing) for each of the SAR image tiles. It should be noted that although future applications of this technology will be most likely using remote sensing data obtained by satellites, this proof-of-concept dataset was obtained through aerial means.

4.2 Data Preprocessing

Each of the 900×900 pixel images was zero-padded to a size of 1024×1024 and then normalised/standardised based on values gathered from the whole of the training set (no-data pixels were not taken into account). Also, since SAR images are affected by the direction from which the data was collected (North vs South), they had to be flipped according to their orientation.

The ground truth masks are comprised of two separate channels, one marking the interiors of each of the building footprints and one marking the borders. Using two separate channels allowed to have the border and the interior overlap over a few pixels to facilitate extraction of the polygons from the predicted masks.

5 Experiments and Results

5.1 Training

We randomly selected 20% (680 pairs of images) of the test dataset to use for testing purposes. A further 20% of the remaining test dataset was used for validation during training. The images were preprocessed similarly for all tests. We used a UNet16 as our 'baseline' model, which was firstly trained using the EO data for 150 epochs, and was subsequently trained over a further 50 epochs using the SAR data. All versions of the SS model were trained for 100 epochs using both SAR and EO data, using hyperparameters $\lambda_D = \lambda_B = 1$ for all experiments. Furthermore, when training our models, the loss was taken to be the average of the separate losses produced by the SAR and EO data. In all cases the Adam optimizer was used with a learning rate of 0.001 and with all the other parameters set to their default value. The batch size was limited to 8 by the capacity of the GPUs used. As mentioned in Section 3.3, in the versions of the model using the randomized skip connections, the value of p is randomly chosen as $p \in [0, 1]$ at each occurrence of a skip connection. All models were trained on single nodes of the Cyclone supercomputer of the Cyprus Institute, each featuring dual Intel Xeon Gold 6248 CPUs and quad NVidia V100 GPUs. The training time did not exceed 12 hours in any of the experiments.

5.2 Performance Evaluation

The evaluation of the performance of the different models was done using the average pixel-wise intersection-over-union score (IoU) of the prediction versus the ground truth masks and the SpaceNet metric which is the F1 score (harmonic average of precision and recall), where a polygon-wise IoU of a predicted and a ground truth polygon greater than 0.5 is considered to be a true positive.

A summary of the test results is presented in Table 1, where as can be seen, the proposed method outperforms compared approaches. The best results are achieved when utilizing both weight-sharing as well as skip mixing, which is a consistent conclusion with respect to both evaluation metrics employed. Furthermore, in Fig. 4 we present indicative results of our implementation and

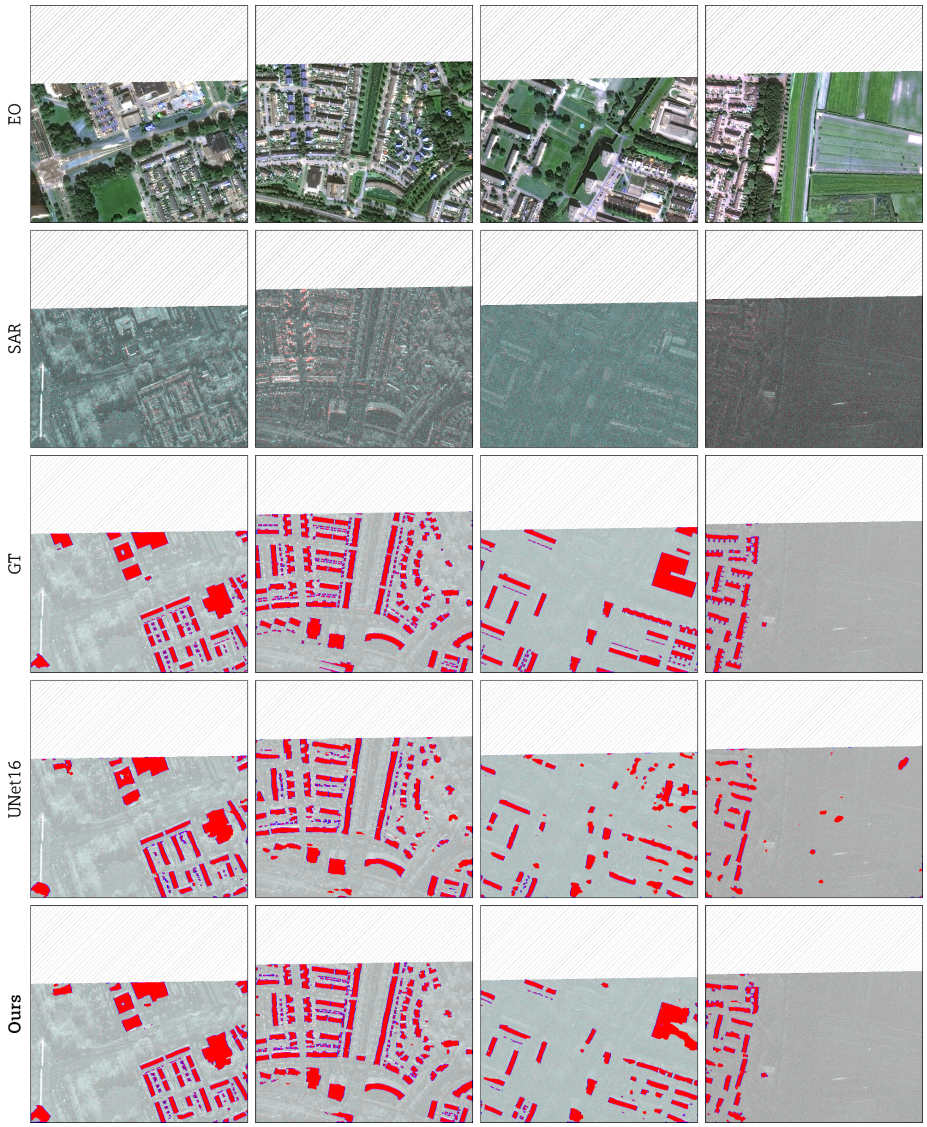

Fig. 4: Different views for 4 tiles from the test datatset, showcasing a wide variety of urban environments. From top to bottom: EO Image, SAR Image, Ground truth, UNet16 mask prediction, Shared Space (ours) mask prediction.

Table 1: Testing Scores for UNet16 (using fine-tuning), and different variations of the proposed shared-space (SS) model, showing the merits of employing randomized skip connections (RSC) and weight sharing. For the last entry on the table, the skip connections originating from non-shared encoder layers were removed. The number of parameters of each model is also presented.

Implementation	No of params	Pixel-wise IoU	SpaceNet metric
UNet16 with fine-tuning	44M	0.596	0.522
SS with UNet skip-conns removed	36M	0.577	0.498
SS with Unet skip-conns	44M	**0.639**	0.592
SS with randomized skip-conns	44M	0.635	0.604
SS with RSC only within the SS	44M	**0.639**	**0.616**

compare with baseline results using UNet16, along with the ground truth and EO and SAR inputs for each case. In the first two columns, one can see how that our model outperforms UNet16 in detecting more fine-grained details (i.e., small buildings). The fourth column of the figure shows how our model generates considerably less false positives than UNet16. This becomes more apparent in Fig. 5 and Fig. 6, were predictions from all the variants of our model are presented together with predictions from UNet16 and the ground truth. It can be clearly seen that utilizing both weight-sharing *in-tandem* with randomized skip-connections facilitates the detection of smaller buildings, that are often missed in the baseline models. We have also observed that the proposed model is able to detect footprints of complex, large buildings with much better fidelity (e.g., Fig. 6b). To further verify the positive effect of the proposed randomized skip-connection architecture, we evaluate a variant of the proposed shared-space architecture where skip connections are entirely removed. As can be seen from results presented in Fig. 5b and Fig. 6b, removing the skip-connections results in segmentation maps that fail to capture fine-grained details. This verifies the successful propagation of high-frequency information through the network layers by employing the randomized skip-connections.

6 Conclusions

In this work, we presented a novel approach for building footprint detection utilising multiple satellite imagery views, in the challenging setting where only one modality is available at test time. To this end, we presented a novel shared-space autoencoder method that utilizes randomized skip connections to facilitate propagating high-frequency information to the later layers without circumventing the shared-space property. We highlight that the proposed method can be considered as a generic approach to fusion under missing modalities at test-time, and can be readily incorporated into more complex architectures. With a set of rigorous experiments, we presented qualitative and quantitative results that demonstrate the merits of the proposed approach, in comparison to typically employed vanilla UNet architectures as well as other fine-tuning approaches.

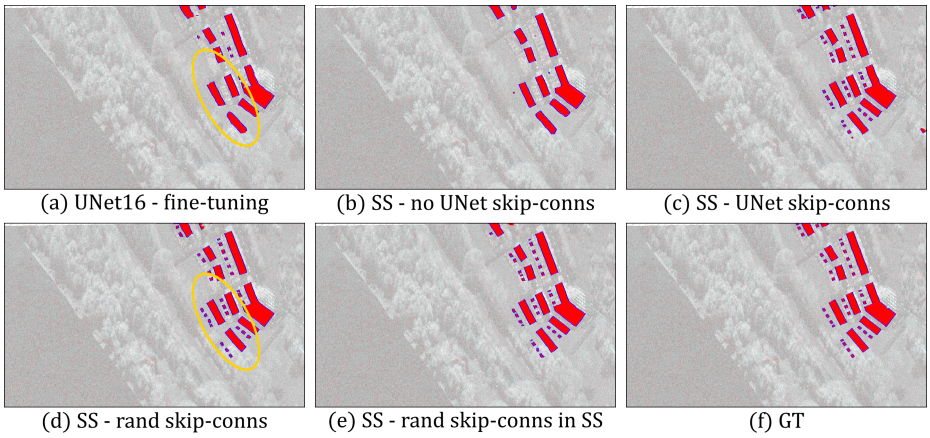

Fig. 5: Predictions from all models together with the ground truth. Yellow circles mark where UNet16 was unable to detect fine details in the image

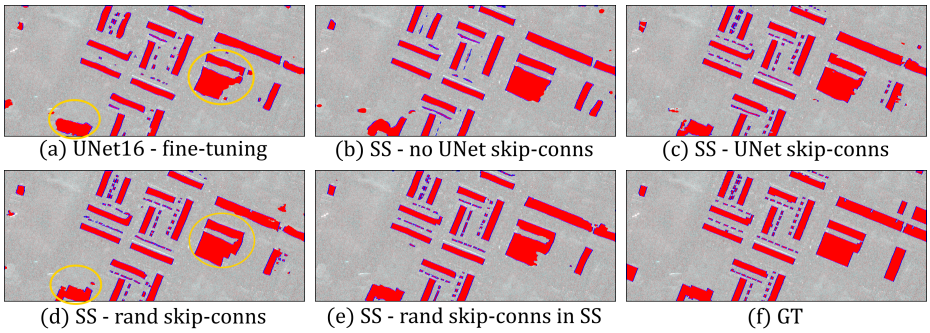

Fig. 6: Predictions from all models together with the ground truth. Our model maintains higher fidelity when predicting complex features like the ones marked by yellow circles

References

1. The carvana masking challenge: https://www.kaggle.com/c/carvana-image-masking-challenge
2. Spacenet on amazon web services (aws). "datasets." the spacenet catalog. last modified april 30, 2018. accessed on 15th july 2020. https://spacenetchallenge.github.io/datasets/datasethomepage.html. (2020)
3. Badrinarayanan, V., Kendall, A., Cipolla, R.: Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence **39**(12), 2481–2495 (Dec 2017). https://doi.org/10.1109/tpami.2016.2644615, http://dx.doi.org/10.1109/TPAMI.2016.2644615
4. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Transactions on Pattern Analysis and Machine Intelligence **40**(4), 834–848 (Apr 2018). https://doi.org/10.1109/tpami.2017.2699184, http://dx.doi.org/10.1109/TPAMI.2017.2699184
5. Chiu, M.T., Xu, X., Wei, Y., Huang, Z., Schwing, A., Brunner, R., Khachatrian, H., Karapetyan, H., Dozier, I., Rose, G., Wilson, D., Tudor, A., Hovakimyan, N., Huang, T.S., Shi, H.: Agriculture-vision: A large aerial image database for agricultural pattern analysis (2020)
6. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A Large-Scale Hierarchical Image Database. In: CVPR09 (2009)
7. Dubois, D., Lepage, R.: Object-versus pixel-based building detection for disaster response. In: 2012 11th International Conference on Information Science, Signal Processing and their Applications (ISSPA). pp. 5–10. IEEE (2012)
8. d'Angelo, P., Cerra, D., Azimi, S.M., Merkle, N., Tian, J., Auer, S., Pato, M., de los Reyes, R., Zhuo, X., Bittner, K., Krauss, T., Reinartz, P.: 3d semantic segmentation from multi-view optical satellite images. In: IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium. pp. 5053–5056 (2019)
9. Gavankar, N.L., Ghosh, S.K.: Object based building footprint detection from high resolution multispectral satellite image using k-means clustering algorithm and shape parameters. Geocarto International **34**(6), 626–643 (2019)
10. He, K., Girshick, R., Dollar, P.: Rethinking imagenet pre-training. 2019 IEEE/CVF International Conference on Computer Vision (ICCV) (Oct 2019). https://doi.org/10.1109/iccv.2019.00502, http://dx.doi.org/10.1109/ICCV.2019.00502
11. He, K., Gkioxari, G., Dollar, P., Girshick, R.: Mask r-cnn. 2017 IEEE International Conference on Computer Vision (ICCV) (Oct 2017). https://doi.org/10.1109/iccv.2017.322, http://dx.doi.org/10.1109/ICCV.2017.322
12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
13. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4700–4708 (2017)
14. Hughes, L.H., Schmitt, M., Mou, L., Wang, Y., Zhu, X.X.: Identifying corresponding patches in sar and optical images with a pseudo-siamese cnn. IEEE Geoscience and Remote Sensing Letters

15(5), 784–788 (May 2018). https://doi.org/10.1109/lgrs.2018.2799232, http://dx.doi.org/10.1109/LGRS.2018.2799232
15. Iglovikov, V., Shvets, A.: Ternausnet: U-net with vgg11 encoder pre-trained on imagenet for image segmentation (2018)
16. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (Jul 2017). https://doi.org/10.1109/cvpr.2017.632, http://dx.doi.org/10.1109/CVPR.2017.632
17. Jean, N., Burke, M., Xie, M., Davis, W.M., Lobell, D.B., Ermon, S.: Combining satellite imagery and machine learning to predict poverty. Science **353**(6301), 790–794 (2016)
18. Kim, J., Kwag, Y.K.: Multi-sensor fusion based target detection using eo/ sar (2014)
19. Kim, S., Song, W.J., Kim, S.H.: Robust ground target detection by sar and ir sensor fusion using adaboost-based feature selection. Sensors (Basel, Switzerland) **16** (2016)
20. Leotta, M.J., Long, C., Jacquet, B., Zins, M., Lipsa, D.R., Shan, J., Xu, B., Li, Z., Zhang, X., Chang, S.F., Purri, M., Xue, J., Dana, K.J.: Urban semantic 3d reconstruction from multiview satellite imagery. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) pp. 1451–1460 (2019)
21. Liu, M.Y., Tuzel, O.: Coupled generative adversarial networks (2016)
22. Pena, J., Boonpook, W., Tan, Y.: Semantic segmentation based remote sensing data fusion on crops detection (07 2019)
23. Purri, M., Xue, J., Dana, K., Leotta, M., Lipsa, D., Li, Z., Xu, B., Shan, J.: Material segmentation of multi-view satellite imagery (04 2019)
24. Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., et al.: Deep learning and process understanding for data-driven earth system science. Nature **566**(7743), 195–204 (2019)
25. Robinson, C., Hohman, F., Dilkina, B.: A deep learning approach for population estimation from satellite imagery. In: Proceedings of the 1st ACM SIGSPATIAL Workshop on Geospatial Humanities. pp. 47–54 (2017)
26. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015 p. 234–241 (2015)
27. Shahzad, M., Maurer, M., Fraundorfer, F., Wang, Y., Zhu, X.X.: Buildings detection in vhr sar images using fully convolution neural networks. IEEE transactions on geoscience and remote sensing **57**(2), 1100–1116 (2018)
28. Shelhamer, E., Long, J., Darrell, T.: Fully convolutional networks for semantic segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence **39**(4), 640–651 (Apr 2017). https://doi.org/10.1109/tpami.2016.2572683, http://dx.doi.org/10.1109/TPAMI.2016.2572683
29. Shermeyer, J., Hogan, D., Brown, J., Etten, A.V., Weir, N., Pacifici, F., Haensch, R., Bastidas, A., Soenen, S., Bacastow, T., Lewis, R.: Spacenet 6: Multi-sensor all weather mapping dataset (2020)
30. Simon Jegou, Michal Drozdzal, D.V.A.R.Y.B.: The one hundred layers tiramisu:fully convolutional densenets for semantic segmentation (2017)
31. Sørensen, T.: A Method of Establishing Groups of Equal Amplitude in Plant Sociology Based on Similarity of Species Content and Its Application to Analyses of the Vegetation on Danish Commons. Biologiske skrifter, I kommission hos E. Munksgaard (1948), https://books.google.co.uk/books?id=rpS8GAAACAAJ

32. Tison, C., Tupin, F., Maitre, H.: A fusion scheme for joint retrieval of urban height map and classification from high-resolution interferometric sar images. Geoscience and Remote Sensing, IEEE Transactions on **45**, 496 – 505 (03 2007). https://doi.org/10.1109/TGRS.2006.887006
33. Tupin, F., Roux, M.: Detection of building outlines based on the fusion of sar and optical features. ISPRS Journal of Photogrammetry and Remote Sensing **58**(1), 71 – 82 (2003). https://doi.org/https://doi.org/10.1016/S0924-2716(03)00018-2, http://www.sciencedirect.com/science/article/pii/S0924271603000182
34. Ulmas, P., Liiv, I.: Segmentation of satellite imagery using u-net models for land cover classification (2020)
35. Van Westen, C.: Remote sensing for natural disaster management. International archives of photogrammetry and remote sensing **33**(B7/4; PART 7), 1609–1617 (2000)
36. Visin, F., Kastner, K., Cho, K., Matteucci, M., Courville, A., Bengio, Y.: Renet: A recurrent neural network based alternative to convolutional networks (2015)
37. Visin, F., Romero, A., Cho, K., Matteucci, M., Ciccone, M., Kastner, K., Bengio, Y., Courville, A.: Reseg: A recurrent neural network-based model for semantic segmentation. 2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (Jun 2016). https://doi.org/10.1109/cvprw.2016.60, http://dx.doi.org/10.1109/CVPRW.2016.60
38. Wang, P., Chen, P., Yuan, Y., Liu, D., Huang, Z., Hou, X., Cottrell, G.: Understanding convolution for semantic segmentation. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 1451–1460. IEEE (2018)
39. Wegner, J.D., Hänsch, R., Thiele, A., Soergel, U.: Building detection from one orthophoto and high-resolution insar data using conditional random fields. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing **4**(1), 83–91 (2011). https://doi.org/10.1109/JSTARS.2010.2053521
40. Weir, N., Lindenbaum, D., Bastidas, A., Etten, A., Kumar, V., Mcpherson, S., Shermeyer, J., Tang, H.: Spacenet mvoi: A multi-view overhead imagery dataset. 2019 IEEE/CVF International Conference on Computer Vision (ICCV) (Oct 2019). https://doi.org/10.1109/iccv.2019.00108, http://dx.doi.org/10.1109/ICCV.2019.00108
41. Zhu, X.X., Montazeri, S., Ali, M., Hua, Y., Wang, Y., Mou, L., Shi, Y., Xu, F., Bamler, R.: Deep learning meets sar. arXiv preprint arXiv:2006.10027 (2020)
42. Zhu, X.X., Tuia, D., Mou, L., Xia, G.S., Zhang, L., Xu, F., Fraundorfer, F.: Deep learning in remote sensing: A comprehensive review and list of resources. IEEE Geoscience and Remote Sensing Magazine **5**(4), 8–36 (2017)