# Let me show you how it's done - Cross-modal knowledge distillation as pretext task for semantic segmentation

Rudhishna Narayanan Nair
Technische Universität Berlin

Ronny Hänsch
German Aerospace Center (DLR)
ronny.haensch@dlr.der

## Abstract

*While Synthetic Aperture Radar (SAR) images have several advantages including robustness to weather conditions and independence from sunlight, they are much harder to interpret by human annotators leading to less and smaller training datasets than for optical imagery. This is in particular true for tasks such as building footprint extraction, where the side-looking nature of SAR complicates the perception of the object of interest. This work aims to leverage the availability of the large amount of labeled optical remote sensing images along with unlabeled paired PolSAR data for semantic segmentation of SAR images through cross-modal knowledge distillation. A network trained on optical images acts as a teacher model to train a student model by providing pseudo-labels for aligned images of both modalities. We test the proposed framework with multiple architectures and observe significantly increased performance after fine-tuning the student, i.e. an increase of 5-20% IoU score compared to training a network based on SAR imagery from scratch.*

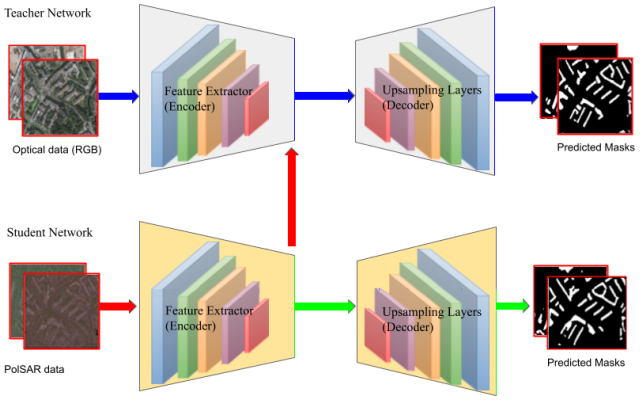



Figure 1. Workflow of the proposed framework: A teacher model is trained fully-supervised on an optical training dataset (blue arrows). The learned representations, i.e. layer activations, of the trained teacher model are subsequently used to provide pseudo-labels for the student model (red arrows) based on aligned optical and SAR imagery. The SAR network can finally be further fine-tuned on SAR training datasets yet achieve better performance compared to training it on these data alone.

## 1. Introduction

Semantic segmentation of remote sensing images plays a crucial role in a wide range of applications such as land-cover classification, crop yield forecasting, urban planning, disaster response, mapping, and monitoring progress towards the sustainable development goals [22].

Building footprint extraction is a particularly challenging field within semantic segmentation of remote sensing images. Being a binary problem (i.e. there are only two classes of interest: building and non-building) might make it appear simpler as general land cover classification which can easily result in dozens of categories. However, buildings have a very high variation of appearance regarding shape, geometry, and radiometric properties while sharing strong similarities with other objects (roads, parking lots, etc.).

An accurate extraction of building footprints is of relevance for mapping and monitoring urban growth, detection of informal settlements, population estimation, and damage assessment and disaster response during natural hazards such as floods, storms, and earthquakes. There is a lot of prior work addressing building footprint detection via deep learning mostly leveraging high-resolution optical imagery [2, 17] relying on multi-scale image features [20, 30, 34].

Optical images are rich in information for the task of building detection. Furthermore, they are readily available for example via open data programs of data providers. Given data availability and the fact that most humans are easily able to recognize buildings in overhead imagery led to the construction of large training datasets for building footprints (e.g. the SpaceNet[1] benchmark datasets).

Optical images, however, have the disadvantage that they are dependent on daylight and cloud cover. This becomes

[1] see spacenet.ai

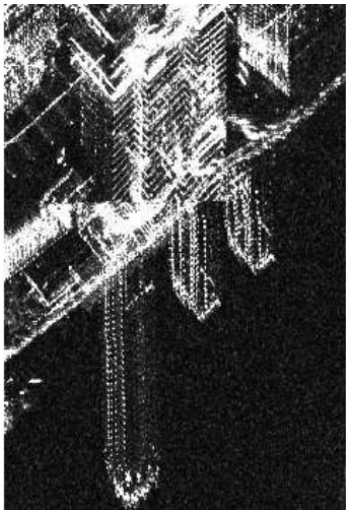


Figure 2. An example from the SpaceNet 6 data [26] showing the effect of layover, i.e. elevated objects such as high-rise buildings are projected towards the sensor (i.e. towards the bottom of the image in this example) rendering building footprint detection extremely challenging.

critical in time-sensitive applications such as disaster response where one might not have the luxury to wait for a cloudless day. Synthetic Aperture Radar (SAR), on the other hand, does not have these limitations. As an active sensor it can acquire images at any time during day and night and the used microwaves penetrate clouds, dust, and smoke, making it an extremely useful image source during natural disasters such as wildfires and floods where optical images are of limited use.

SAR data has been used for general semantic segmentation tasks (e.g. [7, 23, 31]) as well as for building detection (e.g. [4, 14]) e.g. for the use-case of damage assessment after earthquakes [18].

However, SAR images in particular of urban areas are much harder to interpret for human annotators as well as automatic procedures [33]. While access to SAR imagery improved in recent years, e.g. due to the European Copernicus program and open data programs of corresponding companies [21], it is still not at the level of optical imagery. Furthermore, building footprint detection from SAR is more challenging than from (nadir-looking) optical imagery, due to imaging effects such as layover and sensor shadow. Most of the visible building area in SAR imagery is actually from the building facade (see Figure 2 for an illustrative example) which is strongly overlaid with the roof structure, i.e. a point on the rood and a point at the facade having the same distance to the sensor are projected to the same image pixel.

Due to all of these aspects, the number of training datasets for building footprint detection from SAR imagery is much smaller and available datasets are smaller than for their optical counterparts [25]. This raises the question of whether annotated optical data can be used to aid the semantic segmentation of SAR data, in particular if (potentially unlabelled) aligned data of both modalities is available.

We frame this question in the context of knowledge distillation [11] in particular as cross-modal knowledge distillation for pretraining [9]. A model trained on optical data acts as a teacher to create pseudo-labels for a student model that is using SAR imagery as input. This allows the teacher model to be trained on large training datasets that are available for optical imagery and only requires aligned images of both modalities to train the student model.

In contrast to previous work that leverage optical and SAR data in the context of building detection based on data fusion (e.g. [19, 29]), we use optical data only during the training phase but not during inference time. Gupta *et al.* [9] proposed the general framework for cross-modal knowledge distillation and applied it to close-range optical and depth images while Gao *et al.* [8] and applied it to Earth observation data, i.e. optical and SAR images. In both cases, the main motivation is that only a limited amount of training data is available for the target domain. We show that cross-modal pretraining is beneficial even if based on the same amount of data for pre-training and training from scratch. Furthermore, while they limit the reconstruction to a single layer, we match the activations of multiple layers (similar to Huang *et al.* [13] but with a simpler regression loss instead of GANs). A similar approach is proposed by Kang *et al.* [15] who leverages the early layers of an optical-trained encoder and a dual-branch decoder to extract SAR- and optical-based features. Another example of leveraging aligned optical and SAR data to construct a pretext task is transcoding [16], i.e. pretraining the network on transforming one modality into the other. The learned features of the used encoder are then used in a classification network which for the use case of land cover classification is shown to perform far superior compared to training from scratch, in particular in the case of very limited training data. However, if there is no strong relationship between features that are efficient for transcoding and those for classification, the method will fail. In the proposed framework, the pretext task is directly related to the target task as both teacher and student models are solving the same problem only based on different input data.

In summary, the key points of our contribution are

- We propose using cross-modal feature reconstruction in a knowledge distillation framework as a pretext task for building footprint extraction from SAR imagery.
- We evaluate the potential of reconstructing activations of different layers within the network.
- We show that the observed effects of the proposed pretraining are mostly consistent across different network architectures.

## 2. Methodology

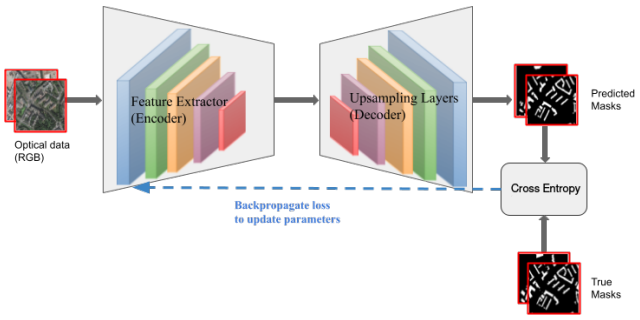
Knowledge distillation [11] is a framework originally intended to compress large models into simpler networks to achieve similar results with greatly reduced computational load. It uses a teacher-student architecture in which the student learns from both the reference labels and the pseudo-labels predicted by the teacher. In this work we lift the basic idea of knowledge distillation in the context of modality alignment [9] to create a pretext task for semantic segmentation, i.e. building footprint extraction from SAR data.

Compared to SAR images, optical images often offer a higher resolution, a lower noise level (due to speckle in SAR imagery), clearer building outlines, as well as larger spectral differences between a building and its surrounding. Effects such as layover (see Figure 2) and sensor-shadow that complicate building extraction from SAR data, are not present in optical imagery. Furthermore, optical data are often easier to access and to annotate than SAR imagery which leads to larger training sets for the former. Another aspect is that footprint extraction is easier for nadir-looking imagery since roof outlines are usually well correlated with the footprint than for side-looking imagery (such as SAR data) where the building footprint (in particular for high-rise buildings) is a rather small part of the visible building structure that mostly consists of the facade. While this results in optical-based models being usually superior to SAR-based models regarding accuracy (see also Section 3), the acquisition of optical imagery requires daylight and low cloud cover while SAR imagery depends on neither.

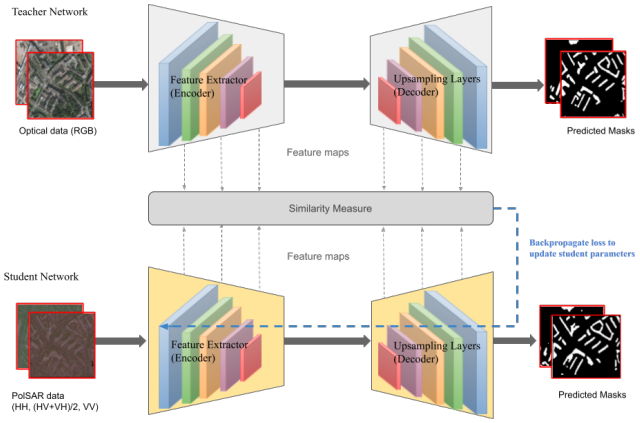
While optical and SAR data have clear and obvious differences, in particular in the context of building detection, the solutions to this task for both modalities can be expected to share similarities, e.g. using some form of edge detection even if building boundaries have very different appearance in both modalities. Thus, we aim to leverage a network trained on optical data (the teacher) to support the training of a SAR network (the student).

The overall workflow is shown in Figure 3 and consists of three phases: Training of the teacher model on optical data, training the student model via feature reconstruction on aligned optical and SAR data, and fine-tuning the student model on SAR imagery. Formally, there are three datasets containing samples of two modalities $\mathcal{M}^S$ (SAR) and $\mathcal{M}^O$ (optical):

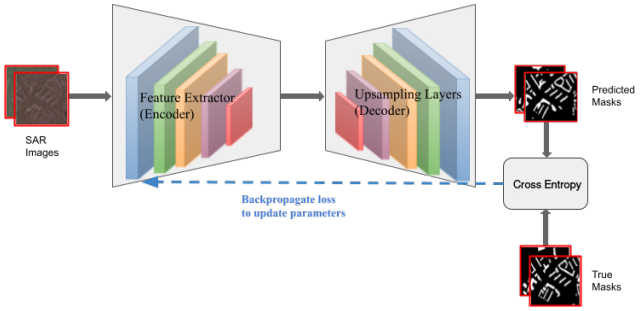
- The dataset to train the teacher in Phase 1, $\mathcal{D}^T = \{x_i^O, y_i\}_{i\in[1,N^T]}$, consisting of $N^T$ samples of modality $M^O$ with corresponding reference data, i.e. optical images and matching building footprints.
- The dataset to pretrain the student via feature reconstruction in Phase 2, $\mathcal{D}^R = \{x_i^O, x_i^S\}_{i\in[1,N^R]}$, consisting of $N^R$ tuples of aligned images from both modalities.
- The dataset to finetune the student in Phase 3, $\mathcal{D}^F = \{x_i^S, y_i\}_{i\in[1,N^F]}$, consisting of $N^F$ samples of modality $M^S$ with corresponding reference data, i.e. SAR images and matching building footprints.




(a) Phase 1: The teacher network $f^T$ is trained fully-supervised on optical training data $D^T$.




(b) Phase 2: Feature maps chosen from intermediate layers of the teacher serve as pseudo-labels for training the student network on a dataset $D_R$ of aligned optical and SAR images without the need of semantic reference data.




(c) Phase 3: The pretrained student model $f^S$ is finetuned on SAR training data $D^F$.

Figure 3. An overview of the proposed framework for cross modal knowledge distillation to leverage a teacher model trained on optical data to facilitate building footprint extraction from SAR imagery.

In general, cross-modal knowledge distillation assumes $N^R > N^T > N^F$, i.e. there is much more unlabeled but aligned data available than labeled data and that the available training dataset for the teacher model is larger than that

of the student. This, however, is not a strict requirement and it is possible to use the same dataset during all phases, i.e. $\mathcal{D} = \{x_i^O, x_i^S, y_i\}_{i\in[1,N]}$ with $N$ samples.

**Phase 1:** The teacher model $f^T$ is trained in a fully-supervised manner on the available optical training data $D^T$ (illustrated in Figure 3a). As loss we use the binary cross-entropy between the softmax predictions $f^T(x_i^O)$ and the reference maps $y_i$. Once converged, the network can be leveraged to create pseudo-labels for the student network.

**Phase 2:** In general, the proposed approach can be applied to different network architectures (indeed, we evaluate various networks with different backbones in Section 3) as long as they have some layers with the same dimensions in common. Otherwise, inter-network layers can project features of one network into the space of the other network [9] and attention layers help to identify network layers that refer to similar semantics [3]. We implement the reconstruction model as a Siamese network architecture (shown in Figure 3b) which automatically fulfills this requirement since the same model architecture is used for both student and teacher networks.

The reconstruction phase uses aligned SAR $x^S$ and optical $x^O$ images from $D^R$ but does not require reference maps (i.e. building footprints). Since we use the same architectures (apart from the number of input layers) for both models, we initialize the weights of the student with the trained weights of the teacher which was found to lead to faster convergence.

Given an image tuple $(x_i^O, x_i^S) \in D^R$, let $\Phi^T(x_i^O) = \{\phi_l^T\}_{l\in[1...L]}$ and $\Phi^S(x_i^S) = \{\phi_l^S\}_{l\in[1...L]}$ be the derived internal representations, i.e. layer outputs, of the teacher $f^T$ and student $f^S$ model, respectively. The reconstruction aims to achieve $\Phi^S(x_i^S) = \Phi^T(x_i^O)$ at least for a (small) subset of layers $\tilde{\mathcal{L}}$, i.e.

$$\forall \tilde{l} \in \tilde{\mathcal{L}} \subset [1, ..., L] : \phi_{\tilde{l}}^S \approx \phi_{\tilde{l}}^T. \tag{1}$$

This is achieved by fixing the weights of the teacher model and optimize the weights of the student by minimizing the cosine loss

$$L_{cos}\left(x^O, x^S\right) = 1 - \frac{1}{|\tilde{\mathcal{L}}|} \sum_{l\in\tilde{\mathcal{L}}} \cos\left(\phi_{\tilde{l}}^S, \phi_{\tilde{l}}^T\right). \tag{2}$$

**Phase 3:** After training the lower layers of the student model via feature reconstruction, they can be combined with the remaining layers of the teacher model. In principle, this provides a full network that is ready for inference without ever being trained on the task of building footprint extraction from SAR data. This renders this last phase optional.

However, given the significant differences between optical and SAR imagery, it can be expected that the feature reconstruction is only partially successful. Furthermore, the use of skip connections will introduce features into the higher levels that have not been optimized to match the characteristics of SAR data. Both aspects cause a distribution shift between the features the higher layers of the teacher model was trained on in Phase 1 and features these layers receive as input when used as part of the final network during prediction leading to suboptimal performance. Thus, it is advisable to finalize training by a fine-tuning phase that uses SAR images as training data, i.e. $D^F$ (Figure 3c. Similar to Phase 1, we use the binary cross-entropy between the softmax predictions $f^S(x_i^S)$ of the network and the reference maps $y_i$.

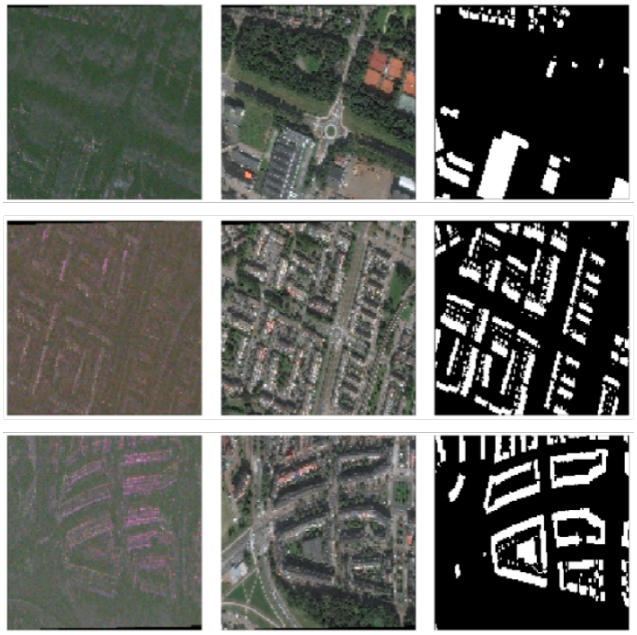


Figure 4. Three samples of the SpaceNet 6 dataset (rows). From left to right: Polarimetric SAR intensity, RGB, reference data.

## 3. Experiments

### 3.1. Data

For the following experiments we use the SpaceNet 6 [26] dataset which consists of over 120 km$^2$ of both high resolution SAR data and optical imagery of Rotterdam, The Netherlands, with about 48,000 building footprint annotations. The SAR data comes from Capella Space's X-band quad-pol sensor mounted on an aircraft and provides image strips featuring four polarizations. These are preprocessed to obtain the geocoded backscatter intensity in decibel at half-meter spatial resolution. The optical data is collected by Maxar's WorldView 2 satellite and provides images spanning about 92 km$^2$ at 0.5 m spatial resolution. Reference data was created by careful manual annotation.

A total of 3,401 triplets of pan-sharpened RGB images,

SAR intensity images, and reference maps each with a size of $900 \times 900$ pixels are provided. Training masks are created from geojson-formatted labels using solaris python API. Figure 4 shows several samples of pre-processed images, i.e. optical and SAR intensity images along with their reference mask of building footprints.

Since the reference data of the SN6 testset is not public, we randomly split the available SN6 training set into three subsets for training (90%), validation (5%), and testing (5%). As proposed by the SpaceNet 6 Challenge, we use the intersection over union computed over the test set as the performance measure.

The corresponding parts of this dataset are used for all three phases described in Section 2, i.e. optical images and footprint maps of the training part as $D^T$ for Phase 1, optical and SAR images of the training part as $D^R$ for Phase 2, SAR images and footprint maps of the training part as $D^F$ for Phase 3, and optical and SAR images as well as footprint maps of the test part for final evaluation of the teacher and student models, respectively.

### 3.2. Models

In order to allow drawing general conclusions about the proposed pretraining strategy, we evaluate several standard ConvNet architectures that are summarized in Table 1, i.e. UNet [24], PSPNet [32], SegNet [1], and DeepLabV3+ [5] with different backbone encoders including MobileNetV2 [12], XceptionNet [6], and ResNet50 [10][3].

### 3.3. Results and Discussion

In a first step we train two models fully-supervised using the available reference data, one using only the optical imagery and one using only the SAR data. On the one hand, we aim to establish a baseline to observe improvements in performance if knowledge-distillation based pretraining is used. On the other hand, the model trained on optical data serves as the teacher network in the following experiments.

The weights of the optical model are initialized via standard pre-training on ImageNet which showed to be beneficial, i.e. leads to an performance increase of about 2% compared to training models completely from scratch. Since ImageNet weights are not applicable to SAR imagery, the SAR network is initialized with random weights. Models are then trained for 300 epochs with a batch size of 32 images using a SGD optimizer with a learning rate of 0.1 and momentum of 0.9.

[3]A UNet with VGG16 [27] let to very similar results as the VGG19 backbone. A UNet with an EfficientNet [28] backbone, as well as an XceptionNet backbone in both PSPNet and SegNet led to very weak performance even for the baselines of training the network from scratch based on SAR or optical images. Thus, we excluded them from further experiments.

All optical models clearly outperform the corresponding SAR models by a large margin. They achieve IoU scores from 42.07 (DeepLabV3 with XceptionNet backbone) to 76.39 (PSPNet) while the performance of the SAR model lies between 25.09 (UNet with MobileNetV2) and 48.61 (UNet with XceptionNet). Noteworthy, this model ranked second for the optical data indicating that this particular architecture is beneficial for the given task for both modalities.

After the optical network is trained and achieves a sufficient performance, it acts as teacher to train a SAR model as student. As different layers across a ConvNet capture different features of the input data, we conduct multiple experiments to reconstruct features from lower as well as higher levels of a model, i.e. only the middle layer (Mid), only the final layer (Fin), middle and final layers (M&F), and multiple intermediate layers (Int).

Again we use an SGD optimizer with a learning rate of 0.1 and momentum of 0.9. All models are trained for 150 epochs with a batch size of 32.

For the mid-level reconstruction, we select the output layer of the encoder. Table 2 shows that this leads to very poor performance with the highest IoU score of 14.79 (DeepLabV3 with ResNet50) reaching not even half of the accuracy of the baseline SAR model. This is to be expected since tasking the student to only reconstruct the teacher's encoder output puts too weak constraints on the early encoder layers which will remain highly sensor-specific, i.e. differ significantly from the early layers of the teacher. However, the skip-connections are directly connecting them to the decoder which is thus confronted with very different data characteristics for most of the inputs (i.e. only the first layer of the decoder that processes the encoder output is not affected).

Finetuning the decoder (Mid+Ft-D) for 30 epochs remedies this effect to a large degree leading to a significant performance increase for all models and to IoU scores that are roughly on-par with the baseline SAR model (i.e. mostly slightly lower with 2-6%, one time identical, and one time 3.5% better). Finetuning the whole model (Mid+Ft-A) leads to further performance gains leading to superior IoU scores for UNet+MobileNetV2 (+3%), DeepLabV3+XceptionNet (+3%), and DeepLabV3+ResNet50 (+9%), while remaining inferior for UNet+XceptionNet (-5%) and PSPNet (-4%). The performance of UNet and SegNet with the VGG19 backbone are too weak to begin with which is why we forgo further experiments.

Reconstructing the final layer (Fin) is the original approach of knowledge distillation. It puts weak constraints on all layers (since all layers contribute to the final layer) but requires training the same amount of parameters as for the baseline model. However, potentially more data can be

| Model | Backbone | Parameters | Total Layers | Encoder Output Layer |
|---|---|---|---|---|
| UNet | MobileNetV2 | 2.25M | 191 | 149 |
| | VGG19 | 31.1M | 54 | 21 |
| | XceptionNet | 38.4M | 121 | 263 |
| | EfficientNetB4 | 19M | 153 | 293 |
| PSPNet | Vanilla Encoder[2] | 800K | 63 | 46 |
| SegNet | VGG19 | 29.5M | 44 | 88 |
| DeepLabV3+ | XceptionNet | 21.6M | 120 | 154 |
| | ResNet50 | 11.8M | 99 | 173 |

Table 1. Summary of the models used in the following experiments. The last column states the position of the last layer of the encoder which is used in some of the experiments to reconstruct optical features from a SAR input image.

| | UNet MobileNetV2 | UNet VGG19 | UNet XceptionNet | PSPNet Vanilla Encoder | SegNet VGG19 | DeepLabV3 XceptionNet | DeepLabV3 ResNet50 |
|---|---|---|---|---|---|---|---|
| Optical | 54.2 | 66.51 | 73.59 | 76.39 | 62.03 | 42.07 | 59.69 |
| SAR | 32.83 | 25.09 | 48.61 | 35.43 | 31.37 | 28.92 | 34.33 |
| Mid | 12.17 | 8.67 | 11.79 | 5.41 | 8.88 | 12.78 | 14.79 |
| Mid+Ft-D | 36.3 | - | 42.39 | 30.26 | - | 26.66 | 34.32 |
| Mid+Ft-A | 36.07 | - | 43.85 | 31.86 | - | 32.2 | **43.04** |
| Fin | 21.06 | 27.5 | 39.67 | 25.88 | 17.54 | 19.86 | 30.38 |
| Fin+Ft-D | 31.98 | 37.69 | 48.16 | 35.62 | - | 22.96 | 38.77 |
| Fin-Ft-A | 33.96 | 40.72 | 50.23 | **37.76** | - | 28.72 | 40.34 |
| M&F | 29.14 | 22.72 | *42.54* | *34.46* | 24.97 | 18.45 | 32.39 |
| M&F+Ft-D | 38.44 | 32.99 | 48.72 | 36.44 | - | 26.3 | 32.39 |
| M&F+Ft-A | 37.72 | 42.66 | 53.27 | 37.56 | - | 31.95 | 42.33 |
| Int | *30.41* | *32.14* | 41.14 | 26.63 | *29.87* | *21.05* | *35.06* |
| Int+Ft-D | 34.58 | 33.33 | **53.45** | 32.51 | 35.85 | 24.74 | 40.69 |
| Int+Ft-A | **37.99** | **46.09** | **53.45** | 29.88 | **42.73** | **32.96** | 42.73 |

Table 2. Accuracy as Intersection of Union scores for different network architectures (columns). The first two rows show the accuracy of the baseline models, i.e. training the network from scratch based on optical and SAR data, respectively. The following 3-row blocks show results for reconstructing activations of the output layer of the encoder ("Mid"), the last layer of the network ("Fin"), both subsequently ("M&F"), and multiple intermediate layers ("Int"). Each block reports accuracy directly after the pretraining and after finetuning either only the decoder ("-D") or the whole network ("-A"). If performance was too weak (less than 25%) for the pretrained model, finetuning experiments have not been performed. Best results per model are shown in boldface, second best in italics.

leveraged since the teacher's logits act as pseudo labels for the student. Furthermore, for some tasks the logits seem to be a more informative signal than class labels [11]. Table 2 shows that results are considerably better than reconstructing the mid-level layer (increasing accuracy by a factor of 2-3), but remain largely inferior to the baseline model. Only for the UNet+VGG19 performance increased by roughly 2%. However, this is the weakest model with an accuracy far below the performance of the other networks. Finetuning the decoder brings model performance close to the baseline, while finetuning the whole network leads to small performance gains (about +2%) for most models (+4% for DeepLabV3+ResNet50).

A combination of these two approaches is to reconstruct mid- and final level features sequentially (M&F). This is similar to M+Ft-D, i.e. pretraining by mid-level reconstruction and finetuning the decoder, with the only difference that instead of actual reference data (as in the case of finetuning), pseudo-labels in the form of teacher logits are used. It should be stressed that at this point the model has not been trained with reference data belonging to the SAR images but only with pseudo-labels generated by the teacher network on aligned optical data. Nevertheless, the achieved performance is - while still worse - close to the baseline

model (-1 to -10%) that is trained exclusively on SAR images and the corresponding reference maps. Finetuning the decoder increases performance further for all models (up to +10%) apart from DeepLabV3+ResNet50 which remains unchanged. If the whole network is finetuned, performance surpasses the baseline by +2 to +18%. The gain in performance is largest for UNet+VGG19 which increased from being with 25.1% the weakest baseline model to being the second strongest model with 42.7%.

The final experiment selects six target layers from en- and decoder for a simultaneous reconstruction. This distributes constraints on the extracted features well over the whole network instead only at specific single layers. The UNet+VGG19 (+7%) and DeepLabV3+ResNet50 (+1%) outperformed the baseline without any finetuning, while the other models performed almost on-par with the baseline. Finetuning improves performance of all models. As before, finetuning the whole model leads to larger gains than only finetuning the decoder (with the notable exception of PSPNet which shows a slight decrease).

In summary, all the UNet based models start to outperform the baseline starting from the final layer reconstruction after finetuning all layers. In particular the UNet model with an XceptionNet backbone shows a steady improvement in its performance as the number of reconstruction layers increases and achieves the highest performance when compared to other models and baseline.

Nearly all models outperform the baseline after multi-level feature reconstruction and finetuning, partially with large gains, e.g. 21% for UNet+VGG19. An exception is PSPNet that reaches top performance already for final level reconstruction that only decreases if more levels are used and DeepLabV3 which does not increase much in accuracy after finetuning the mid-level reconstruction network.

Figure 5 shows a few qualitative results. The optical and SAR input and reference data are shown in the last row. The other rows show prediction results of the different model architecture for the optical and SAR baselines as well as reconstructing the multiple intermediate layer without and with finetuning the whole network.

The optical baseline achieves very good and consistent segmentation results across all models. Inline with the quantitative results in Table 2, segmentation results vary among the different architectures for the SAR baseline. The UNet+VGG19 has strong omission errors and generally underestimates the building footprint area. PSPNet and DeepLabV3 overestimate the building footprint area and tend to fuse buildings in close proximity. Pretraining via feature reconstruction of multiple intermediate layers reduces the number of false negatives for all networks but SegNet where many previously detected buildings are now missing. However, building footprint areas are generally overestimated and precise footprint outlines are mostly lost.

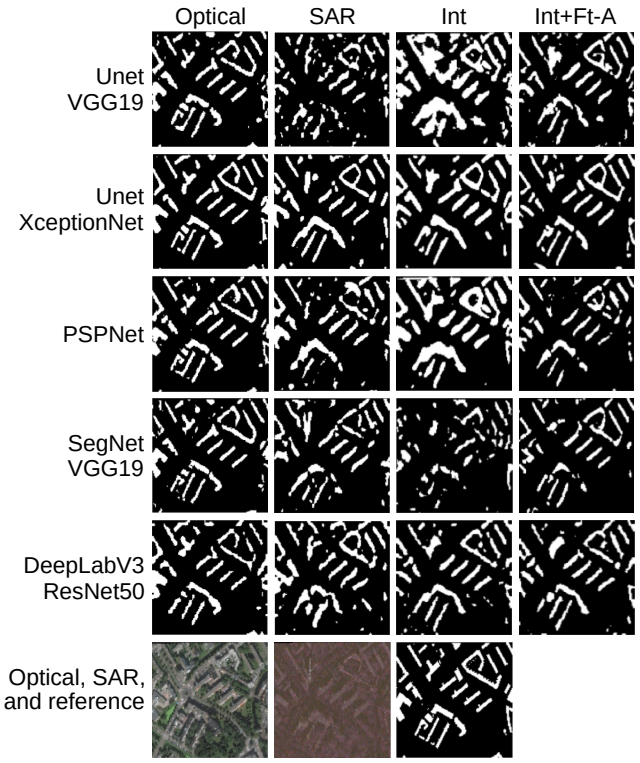



Figure 5. Qualitative results of the proposed approach of cross-modal knowledge distillation. The bottom row shows optical and SAR input images as well as the reference data. The other rows show results for different network architectures. The columns show from left to right the results of the optical and SAR baselines and the results of the models pretrained with reconstructing multiple intermediate layers without and with finetuning of the complete model, respectively.

This is corrected by finetuning the models leading to rather precise segmentation maps that are visually consistent with the results of the optical baseline as well as the reference data.

## 4. Conclusion

This paper proposes a framework leveraging cross-modal feature reconstruction as a pretext task for the semantic segmentation of SAR images. It is a three-step approach in which a model trained fully-supervised on optical images acts as a teacher to supervise in a second step the training of a student network on SAR data. The layer activations of the teacher model serve as pseudo-labels for the student network which only requires aligned input images but no reference data. The last step involves finetuning the pretrained student model on SAR data with available reference maps.

The proposed method is evaluated on several model architectures including UNet, SegNet, PSPNet and

DeepLabV3+ with different feature extractor networks as backbone.

Experimental results show a strong dependence on which layers are used during the feature reconstruction. Worst results are achieved by only reconstructing a single mid-level layer. Best results are achieved by reconstructing multiple intermediate layers. In that case, results are close to the baseline of training a model on SAR images only. This is an important outcome for scenarios were labels are only available for optical images but not for SAR data. Finetuning the decoder or the whole model improves results considerably and consistently for virtually all models. The final results surpass the SAR baseline by a large margin but (as expected) do not reach the performance of the optical baseline.

Future work will focus on leveraging a larger dataset of aligned images of the two modalities for the pretraining step instead of using the same dataset for all three phases and investigate the dependence of the performance on the number of samples available for finetuning.

## References


[1] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(12):2481–2495, 2017. 5

[2] Ksenia Bittner, Fathalrahman Adam, Shiyong Cui, Marco Körner, and Peter Reinartz. Building footprint extraction from vhr remote sensing images combined with normalized dsms using fused fully convolutional networks. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 11(8):2615–2629, 2018. 1

[3] Defang Chen, Jian-Ping Mei, Yuan Zhang, Can Wang, Zhe Wang, Yan Feng, and Chun Chen. Cross-layer distillation with semantic calibration. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(8):7028–7036, 2021. 4

[4] Jiankun Chen, Xiaolan Qiu, Chibiao Ding, and Yirong Wu. Cvcmff net: Complex-valued convolutional and multifeature fusion network for building semantic segmentation of insar images. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–14, 2022. 2

[5] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, 2018. 5

[6] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1800–1807, 2017. 5

[7] Emanuele Dalsasso, Clément Rambour, Nicolas Trouvé, and Nicolas Thome. Merlin-seg: Self-supervised despeckling for label-efficient semantic segmentation. *Computer Vision and Image Understanding*, 241:103940, 2024. 2

[8] Mengyu Gao, Jiping Xu, Jiabin Yu, and Qiulei Dong. Distilled heterogeneous feature alignment network for sar image semantic segmentation. *IEEE Geoscience and Remote Sensing Letters*, 20:1–5, 2023. 2

[9] Saurabh Gupta, Judy Hoffman, and Jitendra Malik. Cross modal distillation for supervision transfer, 2015. 2, 3, 4

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. 5

[11] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015. 2, 3, 6

[12] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications, 2017. 5

[13] Haoshuo Huang, Qixing Huang, and Philipp Krahenbuhl. Domain transfer through deep activation matching. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 2

[14] Hao Jing, Xian Sun, Zhirui Wang, Kaiqiang Chen, Wenhui Diao, and Kun Fu. Fine building segmentation in high-resolution sar images via selective pyramid dilated network. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:6608–6623, 2021. 2

[15] Jian Kang, Zhirui Wang, Ruoxin Zhu, Junshi Xia, Xian Sun, Ruben Fernandez-Beltran, and Antonio Plaza. Disoptnet: Distilling semantic knowledge from optical images for weather-independent building segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–15, 2022. 2

[16] Andreas Ley, Olivier Dhondt, Sebastien Valade, Ronny Haensch, and Olaf Hellwich. Exploiting gan-based sar to optical image transcoding for improved classification via deep learning. In *EUSAR 2018; 12th European Conference on Synthetic Aperture Radar*, pages 1–6, 2018. 2

[17] Qingyu Li, Lichao Mou, Yuansheng Hua, Yilei Shi, and Xiao Xiang Zhu. Building footprint generation through convolutional neural networks with attraction field representation. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–17, 2022. 1

[18] Tianyang Li, Chao Wang, Hong Zhang, Fan Wu, and Xiaohan Zheng. Ddformer: A dual-domain transformer for building damage detection using high-resolution sar imagery. *IEEE Geoscience and Remote Sensing Letters*, 20:1–5, 2023. 2

[19] Xue Li, Guo Zhang, Hao Cui, Shasha Hou, Yujia Chen, Zhijiang Li, Haifeng Li, and Huabin Wang. Progressive fusion learning: A multimodal joint segmentation framework for building extraction from optical and sar images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 195:178–191, 2023. 2

[20] Penghua Liu, Xiaoping Liu, Mengxi Liu, Qian Shi, Yang Jinxin, Xiaocong Xu, and Yuanying Zhang. Building footprint extraction from high-resolution images via spatial residual inception convolutional neural network. *Remote Sensing*, 11, 2019. 1

[21] Nirav Patel. Open source data programs from low-earth orbit synthetic aperture radar companies: Questions and answers [industry profiles and activities]. *IEEE Geoscience and Remote Sensing Magazine*, 11(4):171–C3, 2023. 2

[22] 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. 1

[23] Minh-Tan Pham and Sebastien Lefevre. Very high resolution airborne polsar image classification using convolutional neural networks. In *EUSAR 2021; 13th European Conference on Synthetic Aperture Radar*, pages 1–4, 2021. 2

[24] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *CoRR*, abs/1505.04597, 2015. 5

[25] 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. 2

[26] Jacob Shermeyer, Daniel Hogan, Jason Brown, Adam Van Etten, Nicholas Weir, Fabio Pacifici, Ronny Haensch, Alexei Bastidas, Scott Soenen, Todd Bacastow, and Ryan Lewis. Spacenet 6: Multi-sensor all weather mapping dataset, 2020. 2, 4

[27] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv 1409.1556*, 2014. 5

[28] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks, 2020. 5

[29] Wenfu Wu, Songjing Guo, Zhenfeng Shao, and Deren Li. Crofusenet: A semantic segmentation network for urban impervious surface extraction based on cross fusion of optical and sar images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 16:2573–2588, 2023. 2

[30] Lele Xu, Ye Li, Jinzhong Xu, Yue Zhang, and Lili Guo. Bct-net: Bi-branch cross-fusion transformer for building footprint extraction. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–14, 2023. 1

[31] Lingjuan Yu, Zhaoxin Zeng, Ao Liu, Xiaochun Xie, Haipeng Wang, Feng Xu, and Wen Hong. A lightweight complex-valued deeplabv3 for semantic segmentation of polsar image. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15:930–943, 2022. 2

[32] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6230–6239, 2017. 5

[33] Zhuo Zheng, Ailong Ma, Liangpei Zhang, and Yanfei Zhong. Deep multisensor learning for missing-modality all-weather mapping. *ISPRS Journal of Photogrammetry and Remote Sensing*, 174:254–264, 2021. 2

[34] Qing Zhu, Liao Cheng, Han Hu, Mei Xiaoming, and Haifeng Li. Map-net: Multiple attending path neural network for building footprint extraction from remote sensed imagery. *IEEE Transactions on Geoscience and Remote Sensing*, PP: 1–13, 2020. 1