ETH - Swiss Federal Institute of Technology Zurich

Master thesis

A Shared Deep Feature Embedding of Sentinel-1 and Sentinel-2 for Building Detection

Arno Rüegg
arrueegg@student.ethz.ch
18-914-788

Professorship
Prof. Dr. Konrad Schindler

supervised by
Nando Metzger
Olivier Dietrich
Dr. Anton Obukhov

Zurich, 3th July 2023

Abstract

This thesis tackles the challenge of accurate building segmentation in single-domain remote sensing data, where existing methods relying on data fusion techniques with multiple sources may be impractical. We propose a combined training procedure that leverages both domains during training and enhances single domain inference at prediction time. To improve the performance of single-domain segmentation, we evaluate various loss functions for consistency between the sources. Based on the evaluation, we introduce two innovative approaches: the Discriminator and Input Augmentation. The Discriminator approach incorporates a Discriminator network with a gradient reversal layer, replacing the traditional consistency loss function. This integration enhances the similarity between the source domains, leading to improved segmentation results. On the other hand, the Input Augmentation approach augments the input data with additional features extracted from the variational autoencoder of the Stable Diffusion framework. This augmentation enriches the representation of the input data, enabling the network to capture more comprehensive spatial information. In our experimental evaluation, we demonstrate that both the Discriminator and Input Augmentation approaches yield slight metric improvements in terms of accuracy and precision. However, the visual analysis of the prediction maps reveals significant enhancements in the case of similarity between the source domains. The predicted segmentations exhibit a higher level of detail and better alignment with ground truth annotations. Our research enables more frequent and accurate building segmentations, particularly in scenarios with limited data availability.

Acknowledgement

I would like to thank the following people:

- Prof. Dr. Konrad Schindler for making the work possible.
- Nando Metzger for his support during the implementation and for the numerous inputs and tips I received from him.
- Olivier Dietrich for his support during the implementation and organisation of the thesis.
- Dr. Anton Obukhov for the expertise on domain adaptation and the many ideas for implementations and simplifications.

Contents

List of Figures

List of Tables

1 Introduction

The reliable and accurate segmentation of buildings plays a crucial role in various applications, such as urban planning, disaster management, and population estimation. The availability of high-resolution satellite imagery from sources like Sentinel-1 and Sentinel-2 has revolutionized the field of remote sensing, providing a wealth of data for building segmentation. However, existing approaches often rely on data fusion techniques, combining information from Sentinel 1 and 2 during the training phase. This raises a significant challenge when it comes to inferring segmentations, when only one source modality (also called domain) is available at inference time, limiting their practical applicability. (Hafner, 2022, p.3-4)

In this master's thesis, we aim to address this challenge by investigating the possibility of improving single domain inference for building segmentation, thereby advancing upon the existing research presented in the paper "Unsupervised domain adaptation for global urban extraction using Sentinel-1 SAR and Sentinel-2 MSI data" by Hafner, Ban, and Nascetti. While they primarily focus on data fusion approaches that leverage information from both Sentinel-1 and Sentinel-2 domains separately, our research aims to explore methods that allow for accurate segmentations even when only one domain is available at inference time.

The motivation behind our research stems from the need for more frequent segmentations, especially in scenarios where access to data from both domains may be limited or impractical. This need is particularly evident in regions affected by conflicts or wars, where monitoring building damage in a timely manner is of utmost importance. By developing a combined training procedure that benefits from both domains during the training phase and enhances the metrics for single domain inference at the time of prediction, we aim to overcome this limitation.

Our methodology builds upon the foundation laid by Hafner, Ban, and Nascetti and introduces modifications to the existing pipeline. Specifically, we adjust the pipeline to accommodate the scenario of single domain inference. To better model the complexity of the task and encourage similarity in the extracted features, we replace the consistency loss function with a Discriminator network and a gradient reversal layer.

The goal of this research are improved building segmentations, even in situations where only one domain is available for prediction. This advancement allows for more frequent and reliable monitoring of building damages in war-torn regions. Moreover, the availability of accurate and timely segmentations can have broader implications, including applications such as population estimations in remote areas and urban planning.

The thesis is structured as follows: section 2 provides a comprehensive review of related work in the field of building segmentation and domain adaptation techniques. In section 3, we present the dataset used in our project and outline the data processing steps involved. In section 4 the theoretical basics used in the project are explained before section 5 details the methodology employed, describing the adjustments made to the pipeline and the introduction of the Discriminator network with a gradient reversal layer. In section 6, we present the results of our experiments and analyze their implications. Finally, section 7 offers a conclusion of the findings, highlighting the significance of our research and its potential applications.

In conclusion, this master's thesis aims to enhance building segmentations by improving single domain inference, specifically targeting scenarios where only one domain is available at inference time. By leveraging the benefits of both domains during the training phase, we anticipate achieving more frequent and accurate segmentations, thereby enabling various applications such as damage monitoring in war regions and population estimations in remote areas.

2 Related Work

In this chapter, we provide a description of the publications that form the foundation for this work. While the segmentation of built-up areas is not a novel subject, significant advancements have been made by other researchers in this field. The primary focus of this thesis is the utilization of combined Sentinel-1 and Sentinel-2 data during the training process, allowing for predictions when both or even when only one of these satellite sources are available. Notably, Hafner, Ban, and Nascetti have conducted notable research in this specific direction through their paper titled "Unsupervised domain adaptation for global urban extraction using Sentinel-1 SAR and Sentinel-2 MSI data." Their objective was to develop an innovative and enhanced domain adaptation solution for built-up area segmentation, leveraging the potential of Sentinel-1 and Sentinel-2 data. Due to the close alignment of their work with the topic of this thesis, their research serves as a baseline for our project.

2.1 Unsupervised domain adaptation for global urban extraction using Sentinel-1 SAR and Sentinel-2 MSI data

The baseline paper, titled "Unsupervised Domain Adaptation for Global Urban Extraction using Sentinel-1 SAR and Sentinel-2 MSI data", addresses the challenges of mapping built-up areas on a global scale using Earth Observation (EO) data from the Sentinel-1 SAR and Sentinel-2 MSI missions. The paper acknowledges that previous urban mapping efforts relied on fully supervised training or assumed satisfactory generalization across different regions, both of which have limitations. To overcome these shortcomings, Hafner, Ban, and Nascetti propose a novel approach that leverages Semi-Supervised Learning (SSL) algorithms and Domain Adaptation (DA) techniques. This approach is refereed as "Fusion-DA" in the following.

The proposed methodology involves training Convolutional Neural Networks (CNNs) separately on the SAR and optical images using SSL. Two identical sub-networks are incorporated into the model for built-up area segmentation, with the assumption that consistent segmentation should be achieved across data modalities. To encourage consistency, an unsupervised loss penalizes inconsistent segmentation results from the two sub-networks. This approach effectively utilizes the complementary information from SAR and optical data for improved generalization.

Experimental evaluations were conducted on a test set consisting of sixty representative sites worldwide as described in section 3. The results demonstrate the effectiveness of the proposed DA approach, achieving strong improvements in F1 score compared to fully supervised learning from individual data modalities and their fusion at the input level using the same LU-Net as for the Fusion-DA approach (see tab. 1). Furthermore, a comparison with state-of-the-art products, GHS-BUILT-S2 and WSF 2019, revealed that the proposed model can produce built-up area maps of comparable or even better quality. (Hafner, Ban, & Nascetti, 2022b)

	F1 score		**Precision**		**Recall**		**IoU**	
	Val.	**Test**	**Val.**	**Test**	**Val.**	**Test**	**Val.**	**Test**
SAR	0.664	0.574	0.664	0.570	0.664	0.579	0.497	0.403
Optical	0.739	0.580	0.727	0.699	0.752	0.496	0.586	0.409
Fusion	**0.774**	0.651	**0.745**	**0.712**	0.805	0.599	**0.631**	0.482
Fusion-DA	0.764	**0.694**	0.691	0.661	**0.855**	**0.730**	0.618	**0.531**

Table 1: All metrics of SSL domain adaptation approach by Hafner, Ban, and Nascetti compared to fully supervised single data modalities and input level fusion.

The findings highlight the potential of multi-modal DA to generate easily updateable global human settlement maps, providing accurate and up-to-date information to support sustainable urban development efforts.

In the baseline paper by Hafner et al., a comprehensive comparison between the predictions obtained from the SAR and optical data only at inference using their domain adaptation approach was not conducted. This omission is noteworthy because such a comparison would have been highly informative, particularly for situations where only one satellite sensor is available for inference. It would have been interesting to evaluate the performance of SAR and optical predictions individually and assess their suitability for standalone use, allowing for more frequent predictions. By examining the differences between SAR and optical outputs, it becomes evident that there are considerable disparities, especially when visually inspecting the predictions. SAR predictions tend to exhibit more blurriness in their structural details compared to the predictions derived from optical data.

Given the significant differences observed between the SAR and optical predictions in terms of visual quality and structural clarity, it becomes imperative to address this research gap. While the baseline paper focuses on domain adaptation and consistency regularization to align the predictions from both modalities, the achieved similarity falls short of expectations. Consequently, there is a clear motivation to explore alternative loss functions or methodologies to enhance the consistency and similarity between SAR and optical predictions. One potential approach that could be considered is incorporating a pixelwise domain Discriminator, similar to the one employed later in our approach. By integrating such a Discriminator or exploring different loss functions, it may be possible to improve the visual quality and structural coherence of SAR predictions, aligning them more closely with the optical predictions.

In summary, the baseline paper by Hafner et al. does not provide a direct comparison between the predictions obtained from SAR and optical data using their domain adaptation approach. This comparison would have been valuable for scenarios where only one satellite sensor is available, enabling more frequent predictions. Moreover, the differences observed between SAR and optical outputs, particularly the blurry structural details in SAR predictions, necessitate further investigation. Therefore, this research seeks to address this gap by exploring alternative loss functions or incorporating a pixelwise Discriminator to improve the consistency and similarity between SAR and optical predictions. The subsequent sections will delve into the methodology and findings of our approach, which aims to enhance the visual quality and coherence of SAR predictions, ultimately bridging the gap between SAR and optical-based urban mapping techniques. (Hafner, Ban, & Nascetti, 2022b)

2.2 Building Mapping with Sentinel-1 and Sentinel-2

Earth Observation data, such as Sentinel-1 Synthetic Aperture Radar (SAR) and Sentinel-2 MultiSpectral Instrument (MSI) missions, have provided valuable resources for large-scale urban mapping. SAR data, in particular, has been used in fully automatic building segmentation, leveraging the high backscattering from double-bounce effects of buildings (Chini, Pelich, Hostache, Matgen, & Lopez-Martinez, 2018). For example, the Global Urban Footprint (GUF) was derived from TerraSAR-X and TanDEM-X SAR images, accurately mapping global human settlement at 12 m resolution (Esch et al., 2017). More recent studies have demonstrated the potential of Sentinel-1 SAR data for global urban mapping, achieving high agreement with GUF, although Sentinel-1 only has a resolution of 20m (Chini et al., 2018). On the other hand, optical sensors like Sentinel-2 MSI provide information on surface reflectance characteristics, allowing for the detection of urban areas. (Hafner, Ban, & Nascetti, 2022b)

To harness the complementary information from SAR and optical data, sensor fusion techniques have been explored for improved urban mapping. These techniques aim to combine features extracted from different data modalities to enhance mapping capabilities. This is known as input-level fusion, where features from SAR and optical data are combined before inputting them into traditional ML classifiers. This improves the results of building mapping already significantly as it can learn from more input data and benefit from the sensor-specific advantages of both sensors (shown in tab. 1). In our method and in the Fusion-DA approach the SAR and optical data does not get fused on the input-level, but both sources have their own feature extractor and afterwards a consistency loss is introduced to generate similar prediction maps for both source domains. (Hafner, Ban, & Nascetti, 2022b)

3 Datasets

The dataset for this project is the same as the dataset by Hafner, Ban, and Nascetti for their paper on "Unsupervised domain adaptation for global urban extraction using Sentinel-1 SAR and Sentinel-2 MSI data". This dataset with exactly the same processing steps as described in the following has been used for all methods and tests provided in this thesis.
The entire dataset can be accessed at `https://doi.org/10.5281/zenodo.6914898`. It encompasses satellite images acquired from Sentinel-1 SAR (VV + VH band) and Sentinel-2 MSI (10 spectral bands). These images represent 96 training and validation sites, as well as an additional 60 test sites located in various cities worldwide. Building labels obtained from Microsoft's open-access building footprints are available for 30 training and validation sites, limited to the United States, Canada, and Australia. The remaining 66 training sites outside of these countries do not have labels but can still be utilized for semi-supervised learning. All 60 test sites in the dataset have manually labeled building footprints sourced from the SpaceNet7 dataset. (Hafner, Ban, & Nascetti, 2022a) The locations of training, validation and test sites are displayed in figures 1 and 2.

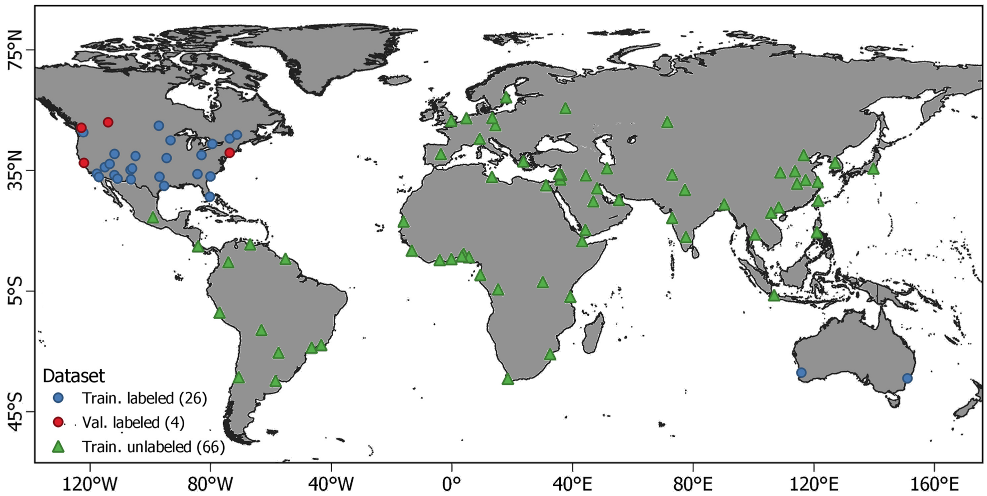


Figure 1: Overview of locations for training and validation sites. (modified from (Hafner, Ban, & Nascetti, 2022b))

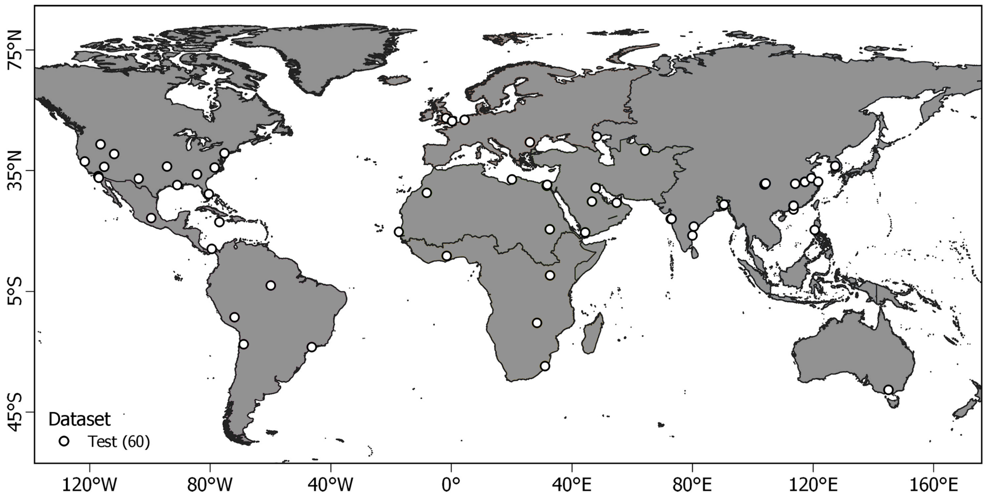


Figure 2: Overview of locations for test sites. (modified from (Hafner, Ban, & Nascetti, 2022b))

The preprocessing steps for all data sources are shown in figure 3, where all operations are visualized.

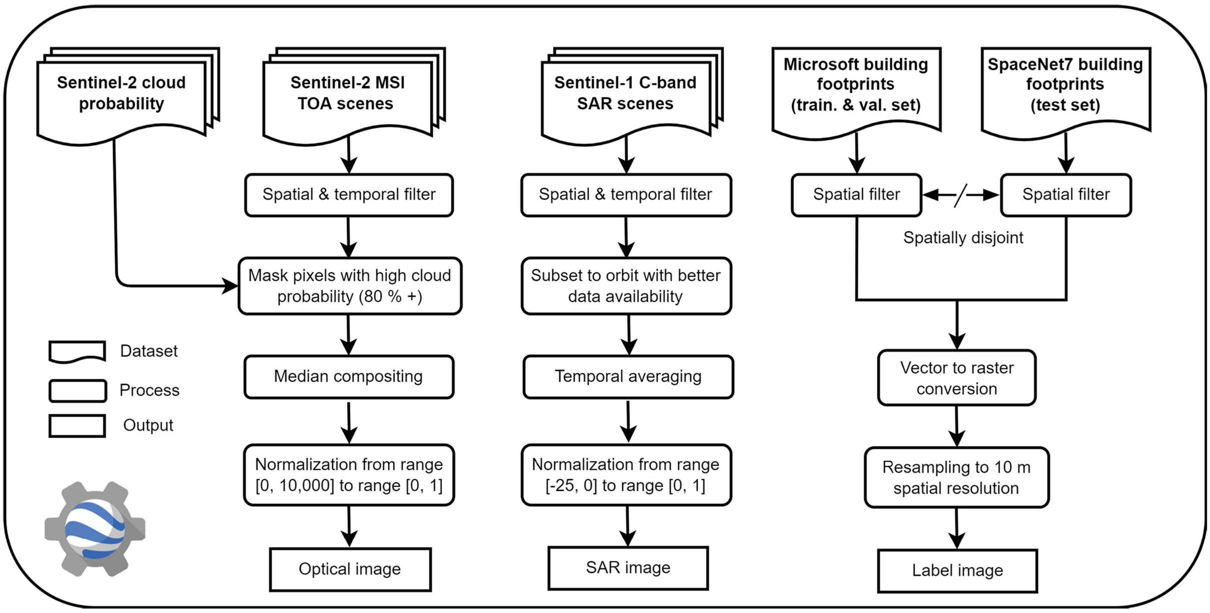

Sentinel-2 cloud probability
Sentinel-2 MSI TOA scenes
Sentinel-1 C-band SAR scenes
Microsoft building footprints (train. & val. set)
SpaceNet7 building footprints (test set)
Spatial & temporal filter
Spatial & temporal filter
Spatial filter
Spatial filter
Spatially disjoint
Mask pixels with high cloud probability (80 % +)
Subset to orbit with better data availability
Dataset
Process
Output
Median compositing
Temporal averaging
Vector to raster conversion
Normalization from range [0, 10,000] to range [0, 1]
Normalization from range [-25, 0] to range [0, 1]
Resampling to 10 m spatial resolution
Optical image
SAR image
Label image

Figure 3: Overview of the preprocessing operations for Sentinel 1 and 2, Microsoft building footprints and SpaceNet7 building footprints. (Hafner, Ban, & Nascetti, 2022b, p.3)

3.1 Sentinel 1

The Sentinel 1 mission, which is part of Copernicus, utilizes the C band to gather information with a ground resolution of 20m. The mission's capabilities allow for the distinction of two channels: HH + HV and VV + VH, based on different polarizations. To process the acquired images, Ground Range Detection was employed followed by further preprocessing using Google Earth Engine (GEE) (Gorelick et al., 2017). This preprocessing entailed various steps, including thermal noise removal, radiometric calibration, and terrain correction.

It was subsequently applied time filtering to the images for each region of interest, and only the orbits with better data availability were selected. This selection was crucial due to the significant impact of the signal's incidence angle on backscattering for buildings. Since multiple observations were available for most regions as a result of temporal filtering, a proven method of reducing speckle noise was adopted by taking the mean value per 20m ground resolution pixel (Chini et al., 2018, p.5). This approach not only minimized noise but also ensured the inclusion of useful information.

To facilitate analysis, the signal was converted to decibels using a logarithmic scale. Then, the noise was eliminated by applying a mask that excluded all values below -25dB. The remaining values ranged between [-25,0] before being normalized to the interval [0,1]. These preprocessed images were subsequently utilized for training, validation, and test sites (Hafner, Ban, & Nascetti, 2022b).

3.2 Sentinel 2

Sentinel 2, like Sentinel 1, is part of ESA's Copernicus program. It collects data in 13 spectral bands provided at different processing levels (ESA, 2023). For this thesis, processing level 1C was chosen due to its global availability shortly after acquisition and because it is already geometrically corrected on the cartographic map. The 13 spectral bands include RGB channels (Band 2, 3, 4), Visible and Near Infrared channels (Band 5, 6, 7, 8, 8a), and Short Wave Infrared channels (Band 11, 12). Only the channels with ground resolutions of 10m and 20m are utilized for further processing, as the 60m resolution channels contain atmospheric information not considered relevant for building segmentation.

Band	Resolution	Central Wavelength	Description
B1	60 m	443 nm	Ultra Blue (Coastal and Aerosol)
B2	10 m	490 nm	Blue
B3	10 m	560 nm	Green
B4	10 m	665 nm	Red
B5	20 m	705 nm	Visible and Near Infrared (VNIR)
B6	20 m	740 nm	Visible and Near Infrared (VNIR)
B7	20 m	783 nm	Visible and Near Infrared (VNIR)
B8	10 m	842 nm	Visible and Near Infrared (VNIR)
B8a	20 m	865 nm	Visible and Near Infrared (VNIR)
B9	60 m	940 nm	Short Wave Infrared (SWIR)
B10	60 m	1375 nm	Short Wave Infrared (SWIR)
B11	20 m	1610 nm	Short Wave Infrared (SWIR)
B12	20 m	2190 nm	Short Wave Infrared (SWIR)

Table 2: Sentinel 2 bands with resolution, Central Wavelength and Description (GIS-Geography, 2022)

The Sentinel 2 data undergoes additional preprocessing, including temporal and spatial filtering similar to the Sentinel 1 images. Since Sentinel 2 observations are susceptible to weather conditions, a cloud masking technique based on the Sentinel Hub's cloud detector is applied (Sentinel-Hub, 2023). Pixels with a cloud probability higher than

80% are masked, and for the remaining pixels, the median value from all recordings is used to generate the final image. Lastly, the data is normalized to a range of [0, 1]. Like the Sentinel 1 images, also the Sentinel 2 images were subsequently utilized for training, validation, and test sites. (Hafner, Ban, & Nascetti, 2022b)

3.3 Microsoft building footprints

For training and validation the Microsoft building footprints were used because they were available before 2021 for the United States, Canada and Australia only. Microsoft used a two step process to derive the building footprints from aerial images. First, they performed a semantic segmentation to recognize building pixels with a deep neural network. Second, they converted building pixel blobs to polygons, which were then published as GeoJSON. In table 3 are all the metrics for the three regions shown. (Microsoft, 2023)

Region	**Precision**	**Recall**	**IoU**
USA	98.5%	92.4%	0.86
Canada	98.7%	72.3%	0.76
Australia	98.6%	65.0%	0.79

Table 3: Evaluation metrics for all regions of Microsoft building footprints (Microsoft, 2023)

For the training and validation set, these building footprint polygons were transfered to GEE. The polygons which are located inside one of the labeled training or validation sites are rasterized with a resolution of 10m that they fit to the sentinel 1 & 2 raster cells. Due to the temporal difference between the Sentinel images and the Microsoft building footprints it is possible that they not reach the metric values stated in table 3 in the final dataset.

3.4 SpaceNet 7 building footprints

For the test dataset the labels were taken from SpaceNet 7, as it is a very large dataset with manually annotated building footprints. It consists of 101 fast growing sites for which monthly imagery is available for a median period of 24 months between 2017 and 2020. (Van Etten et al., 2021)

After a spatial disjoint from training and validation sites to prevent any overlap, the remaining 60 sites were taken for the test dataset. The selected timestamp of the test images depend mainly on the fact that the observations had to be cloud free, allowing an accurate mapping of all buildings. The data preprocessing follows the same work flow as for the Microsoft building footprints and is also converted to a raster map with 10m resolution. (Hafner, Ban, & Nascetti, 2022b)

3.5 Dataset overview

The final dataset consists of a training, validation and test dataset, where the labels for the training and validation set are derived from Microsoft building footprints and the test set labels from manually annotated SpaceNet 7 building footprints. This dataset equals the dataset used by Hafner, Ban, and Nascetti. All datasets include Sentinel 1 & 2 images with a timestamp as close as possible to the corresponding labels. In figure 4 the sizes of the training and validation set are displayed. The labeled training set includes roughly 1.32 billion pixels, covering an area of over 132,000 square kilometers. Around 11% of this area represents Built-Up Area. Comparatively, the unlabeled training set is larger than the labeled one, encompassing about 1.79 billion pixels and covering over 179,000 square kilometers. As for the validation set, it is considerably smaller, it consists of approximately 222 million pixels, covering more than 22,000 square kilometers, with approximately 10% of that area representing Built-Up Area. The test set includes 11.1 million pixels and covers a total area of 1,112 square kilometers, whereas 15% of the pixels are denoted as buildings.

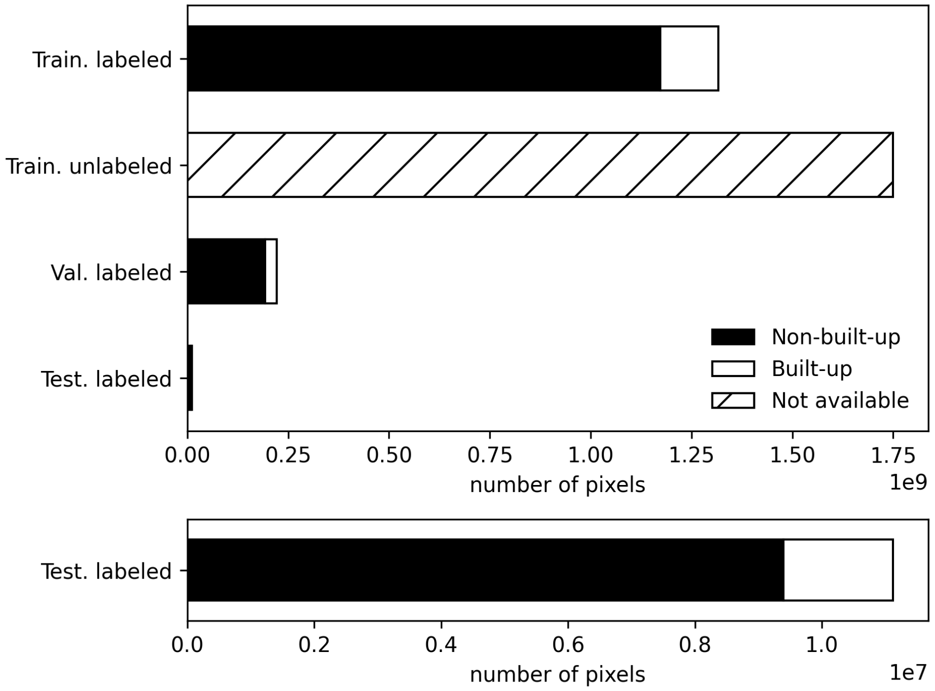


Figure 4: Pixel distribution for training, validation and test set on top and separate test set on bottom with different scaling. (modified from Hafner, Ban, and Nascetti)

4 Theoretical Background

The theoretical background chapter of this master thesis provides a foundation of key concepts and techniques crucial to the research of the thesis. This chapter focuses on semi-supervised learning, U-Net architecture, Discriminator models, Gradient Reversal layer, and evaluation metrics. By exploring these topics, we establish a strong understanding of the theoretical underpinnings that form the basis of our work.

4.1 U-Net

U-Net is a widely used deep-learning architecture for semantic segmentation, initially developed for medical images. It has shown remarkable success in various fields, including satellite images and handwritten characters. Semantic segmentation involves assigning a class to each pixel in an image, and U-Net was specifically designed for this task. The architecture consists of an encoder, decoder, and skip connections, which enable accurate segmentations with limited data. The network architecture from the original paper is shown in figure 5 and shows the encoder on the left and the decoder on the right. (O'Sullivan, 2023) (Ronneberger et al., 2015)

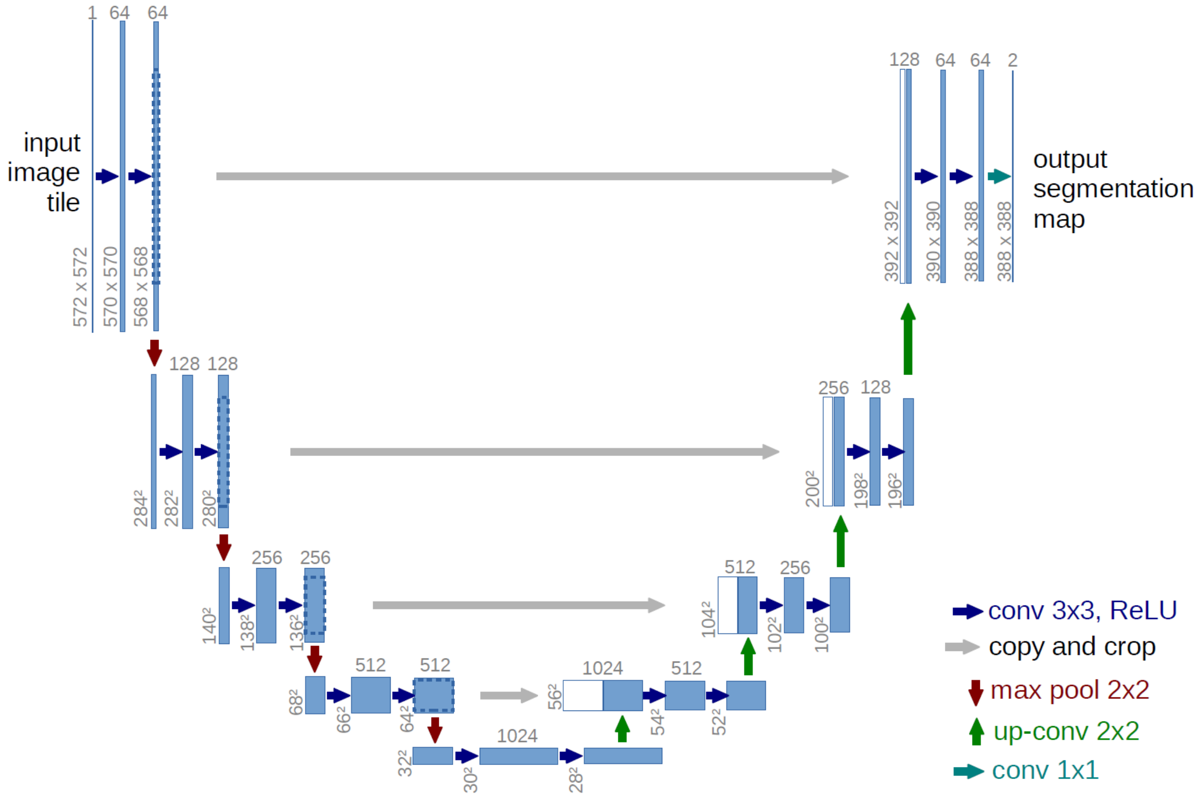


Figure 5: U-Net architecture with encoder and decoder. (Ronneberger et al., 2015)

The encoder plays a crucial role in creating a compact representation of the input image, extracting important features through convolutional and pooling layers. However, this representation lacks information about the location of the features, which is essential

for segmentation. To address this, the decoder reconstructs an image from the compact representation using deconvolution layers. It increases the dimensionality of the image, allowing for accurate reconstruction. (O'Sullivan, 2023)
However, the decoder alone cannot retain the location information without a significant amount of training data. An important innovation of U-Net lies in its use of skip connections, which connect the encoder and decoder. These connections pass feature information from earlier convolutional layers to the deconvolutional layers while preserving the location information. This is achieved by concatenating the corresponding layers and performing convolution on the concatenated tensor. (Ronneberger et al., 2015)
By combining feature extraction and feature localization, U-Net enhances the performance of semantic models and reduces the data required for training. Skip connections enable the network to effectively learn from limited training images and produce accurate segmentations. (O'Sullivan, 2023)

4.2 Discriminator

The discriminator network has first been described in 2014 by Goodfellow et al. in their paper on "Generative Adversarial Nets". In that, they introduced the concept of generative adversarial networks (GANs), which consist of a generator network and a discriminator network engaged in a game-theoretic framework as shown in figure 6.

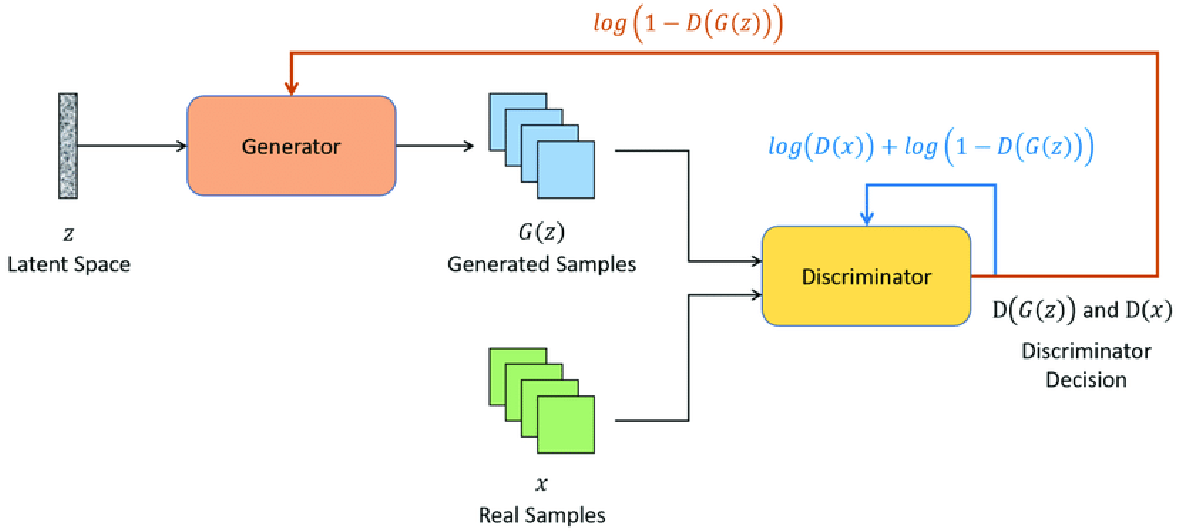


Figure 6: Typical GAN architecture including a discriminator. (Vint et al., 2021)

The goal of the generator in a GAN is to generate realistic samples, such as images or texts, that resemble the training data. The generator takes random noise or latent vectors as input and tries to generate output samples that are indistinguishable from real data. On the other hand, the goal of the discriminator is to accurately classify and distinguish between real samples from the training data and the generated samples produced by the generator. The discriminator is trained to improve its ability to correctly identify whether a given sample is real or generated. The optimum is reached when the generator approximates the training data distribution well and the discriminator accuracy is 1/2 and therefore cannot distinguish anymore between real and fake. (Goodfellow et al., 2014)

The discriminator is used to distinguish between to classes (real and fake in the example above), in a classification manner for entire images. It can also be changed to a U-Net like architecture to do this classification on pixel level. It is expected to be more robust, since not the entire segmentation of an image has the same quality. This allows to locate areas which cause a high loss for the discriminator and apply the corrections on specific features. (Cai et al., 2022; Souly, Spampinato, & Shah, 2017)

4.3 Gradient reversal layer

In the paper "Unsupervised Domain Adaptation by Backpropagation" by Ganin and Lempitsky the authors propose a method for unsupervised domain adaptation, aiming to adapt a model trained on a source domain to perform well on a target domain without relying on labeled target data. The proposed framework utilizes a domain classifier and a feature extractor to learn domain-invariant representations (see fig. 7). (Ganin & Lempitsky, 2015)
The feature extractor, responsible for extracting relevant features from input data, commonly employs a siamese architecture where two identical feature encoders share weights. This architectural choice encourages the feature extractor to learn representations that capture the underlying structure and semantics of the data, rather than being influenced by domain-specific characteristics.
To achieve domain-agnostic feature representations, the authors introduce a domain confusion objective. A domain classifier, another neural network, predicts the domain label (source or target) given the extracted features. The goal is to generate feature representations that do not contain information distinguishing between the domains.
To enforce domain invariance, a Gradient Reversal Layer (GRL) is introduced. Placed between the feature extractor and the domain classifier as shown in figure 7. The GRL reverses gradients during backpropagation, negating the gradient signal typically provided by the domain classifier to the feature extractor. By confusing the feature extractor, the GRL encourages the learning of domain-invariant representations. This mechanism aligns the feature distributions across domains, reducing the influence of domain-specific information. (Bolte et al., 2019; Ganin & Lempitsky, 2015)

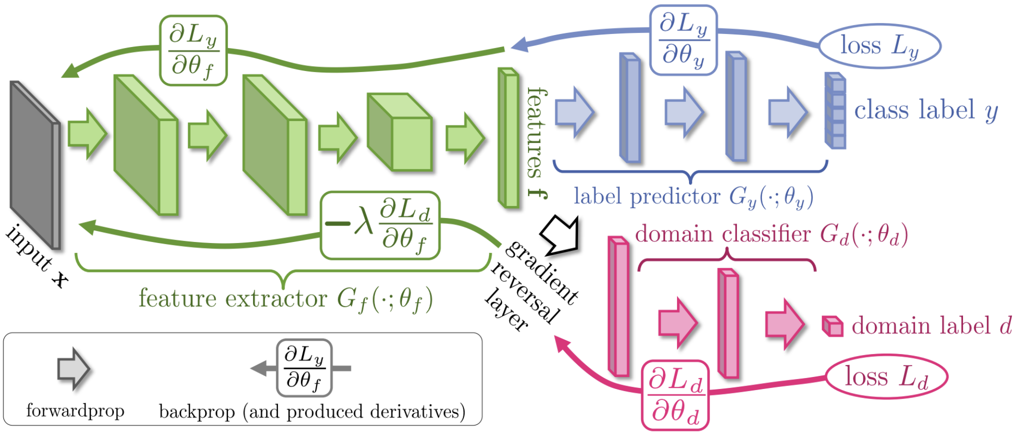

$\frac{\partial L_y}{\partial \theta_f}$
$\frac{\partial L_y}{\partial \theta_y}$
loss L_y
class label y
features $\mathbf{f}$
label predictor $G_y(\cdot;\theta_y)$
$-\lambda \frac{\partial L_d}{\partial \theta_f}$
input $\mathbf{x}$
feature extractor $G_f(\cdot;\theta_f)$
gradient reversal layer
domain classifier $G_d(\cdot;\theta_d)$
domain label d
loss L_d
$\frac{\partial L_d}{\partial \theta_d}$
forwardprop
backprop (and produced derivatives)

Figure 7: Network architecture with feature extractor (green), domain classifier (pink), label predictor (blue) and gradient reversal layer. (Ganin & Lempitsky, 2015)

During training, the feature extractor aims to generate representations that confuse the domain classifier, while the domain classifier attempts to correctly classify the domains. By jointly optimizing the feature extractor and the domain classifier, the model learns to extract features that are invariant across domains, enabling effective adaptation to the target domain even with limited or no labeled target data. (Ganin & Lempitsky, 2015)

4.4 Semi-supervised learning

In remote sensing, where obtaining labeled data can be difficult and expensive, semi-supervised learning has become an increasingly popular approach to improve the accuracy of classification and segmentation tasks. Semi-supervised learning is a type of machine learning that uses both labeled and unlabeled data to train models. Obtaining labeled data in remote sensing can be challenging due to the high cost and time required for manual annotation. Semi-supervised learning can leverage the large amounts of unlabeled data available to improve the performance of models trained on limited labeled data. (Z. Chen et al., 2020; Li, Zhang, Li, & Ye, 2023)

One possible approach is called consistency regularization, that is a technique that leverages unlabeled data by applying strong augmentations to create multiple versions of the data. The objective is to ensure that these augmented samples produce consistent predictions, regardless of their specific augmentation. In a similar manner, multi-modal data, like Sentinel-1 and Sentinel-2 data from the same location, can be utilized to simulate these augmentations, aiming to achieve consistent prediction outputs. The different modalities serve as real-world perturbations, as they are called in the Paper by Hafner, Ban, and Nascetti. (Hafner, Ban, & Nascetti, 2022b; Oliver, Odena, Raffel, Cubuk, & Goodfellow, 2018)

4.5 Metrics

In the field of semantic segmentation, several metrics are commonly used to evaluate the performance and accuracy of segmentation algorithms. This chapter discusses some of the key metrics employed in the assessment of semantic segmentation models.

4.5.1 Precision

Precision is a metric that measures the proportion of correctly identified positive predictions out of all positive predictions made by the model. In the context of semantic segmentation, precision indicates the accuracy of the model in correctly classifying pixels belonging to the target class. A high precision value indicates a low rate of false positive predictions.

The precision is calculated as follows:

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$$

4.5.2 Recall

Recall, also known as sensitivity or true positive rate, measures the proportion of correctly identified positive predictions out of all ground truth positive instances. In semantic segmentation, recall evaluates the model's ability to accurately detect and capture all pixels belonging to the target class. A high recall value indicates a low rate of false negatives.

The recall is calculated as follows:

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$$

4.5.3 F1 score

The F1 score combines precision and recall into a single metric, providing a balanced assessment of the model's performance. It is calculated as the harmonic mean of precision and recall, giving equal weight to both metrics. The F1 score is particularly useful when there is an imbalance between positive and negative instances in the dataset.

The F1 score is calculated as follows:

$$\text{F1 score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

4.5.4 Intersection over Union

Intersection over Union (IoU), also referred to as the Jaccard index, quantifies the overlap between the predicted segmentation and the ground truth. It is calculated as the ratio of the intersection of the predicted and ground truth regions to their union. IoU measures the accuracy of the model in localizing and delineating the target class.

The IoU is calculated as follows:

$$\text{IoU} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive} + \text{False Negative}}$$

4.5.5 Structural Similarity Index Measure

The Structural Similarity Index Measure (SSIM) assesses the similarity between two images, considering various aspects such as luminance, contrast, and structure. In semantic segmentation, SSIM can be used to evaluate the perceptual quality and visual similarity between the predicted segmentation and the ground truth. It provides a comprehensive measure of the model's ability to preserve structural information during the segmentation process.

The SSIM is calculated as follows:

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}$$

where:

- x and y are the input images
- μ_x and μ_y are the means of x and y
- σ_x and σ_y are the standard deviations of x and y
- σ_{xy} is the covariance between x and y
- C_1 and C_2 are constants to stabilize the division.

5 Methodology

In this chapter, we present the methodology employed for accurate building segmentation using Sentinel-1 SAR and Sentinel-2 MSI data. The workflow is based upon the paper "Unsupervised domain adaptation for global urban extraction using Sentinel-1 SAR and Sentinel-2 MSI data", as they already researched a similar direction. After modifying the Fusion-DA code including different loss functions, we replaced it with a Discriminator and decided to develop a new pipeline. By integrating the Discriminator into the segmentation process, we aimed to improve the model's performance, F1-score, and segmentation structure. Throughout this chapter, we provide an overview of the revised methodology, including data preprocessing, model architectures and training setup.

5.1 Baseline "Fusion-DA"

The goal of this project is to make good predictions for built-up areas based on Sentinel-1 and Sentinel-2 data, even when only one modality is available. In the paper ”Unsupervised domain adaptation for global urban extraction using Sentinel-1 SAR and Sentinel-2 MSI data" by Hafner, Ban, and Nascetti a similar goal has been explored, which provides a useful baseline architecture for our goal as well refereed as "Fusion-DA" in the following. This architecture will be explained in the following and the results of this baseline are provided along the results of our approach in section 6.

The overall architecture consists of a Dual Stream U-Net as proposed in Hafner, Nascetti, et al. (2022) and shown in figure 8. This includes two feature encoder streams, one for Sentinel-1 and one for Sentinel-2 data ($\mathrm{x}^{\mathrm{sar}}$ and $\mathrm{x}^{\mathrm{opt}}$), which are referred as lightweight-U-Nets (LU-Net). After extracting the features the samples with labels get concatenated and passed trough a last convolutional layer followed by a sigmoid activation (referred as Out), which reduces the features to a single channel that indicates the pixel-wise probability for built-up area. This gives predictions for the fusion, sar and optical stream ($\mathrm{p}^{\mathrm{fus}}$, $\mathrm{p}^{\mathrm{sar}}$ and $\mathrm{p}^{\mathrm{opt}}$), that are then used to calculate the supervised loss. The unlabeled data get directly passed trough the Out block and is then used to calculate the consistency loss, which should lead to similar encodings by the SAR and optical stream.

The LU-Net retains the fundamental structure of the original U-Net, consisting of a contracting path (left side) with downsampling blocks and an expansive path (right side) with upsampling blocks. The number of downsampling and upsampling blocks, also known as network depth, plays a crucial role in the complexity and performance of the model. While the original U-Net with a depth of 4 has around 30 million trainable parameters, the LU-Net only has depth 2 and therefore approximately 1 million trainable parameters. This reduction in depth not only helps in reducing model complexity but also reduces the receptive filed, limits overfitting and enhances computational efficiency. (Hafner, Ban, & Nascetti, 2022b)

The architecture of LU-Net is illustrated in figure 9. Similar to the original U-Net, the LU-Net comprises a contracting path and an expansive path. Each downsampling step in the contracting path involves a sequence of operations, including a 3 × 3 convolution (Conv) with padding, batch normalization (BatchNorm), and rectified linear unit activation function (ReLU). The downsampling is performed by a 2 × 2 max pooling

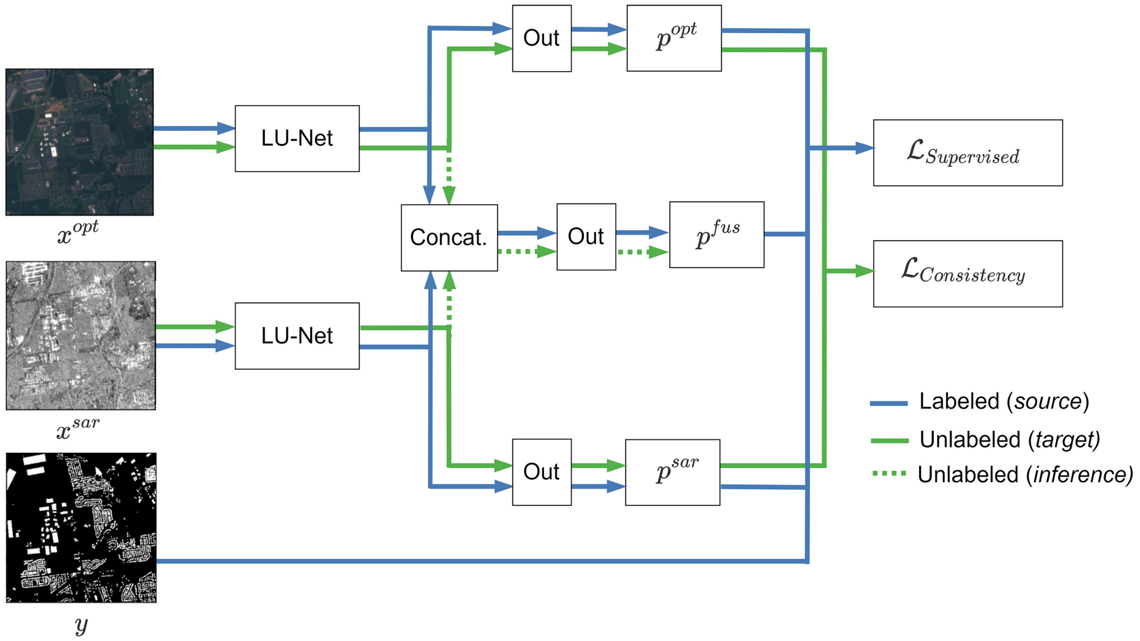


Figure 8: Network architecture with Dual Stream U-Net (Hafner, Ban, & Nascetti, 2022b; Hafner, Nascetti, et al., 2022)

(MaxPool) operation. With each downsampling step, the number of feature channels is doubled, starting from 64 channels, and the size of the feature maps is halved.
The LU-Net consists of two downsampling steps in the contracting path and two upsampling steps in the expansive path. The upsampling steps reverse the operations of downsampling by doubling the size of the feature maps and halving the number of channels. This is achieved using a 2 × 2 transpose convolution operation followed by the same operation triplet utilized in downsampling steps. Additionally, skip connections are incorporated to directly transmit feature maps from the contracting path to corresponding steps in the expansive path. Ultimately, the feature map reaches the original size of the input image and contains 64 channels. A 1 × 1 convolution followed by the sigmoid activation function is applied to convert this feature map into a single-channel output representing the probability of built-up areas (reffered as Out in figure 8). (Hafner, Ban, & Nascetti, 2022b)

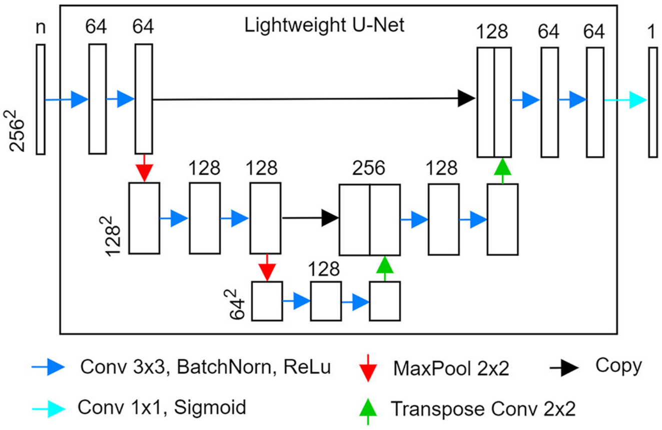


Figure 9: LU-Net architecture with width/height (left) and number of channels (top) for each step (Hafner, Ban, & Nascetti, 2022b)

As one can see already in figure 8 the Fusion-DA approach involves a data path in the pipeline for the labeled and unlabeled data separately. This means the training procedure is based on semi-supervised learning (SSL), namely on a consistency regularization as it is explained in section 4.4. The core idea is to minimize the discrepancy between the model outputs for different augmented versions of the same unlabeled sample. In the Fusion-DA, the hypothesis is that the distinct yet complementary urban information present in SAR and optical data can be effectively utilized for domain adaptation through the principle of consistency regularization. Hence, Sentinel 1 and Sentinel 2 observations of the same location are referred as the strong augmentations of the same sample.

From $\mathrm{p}^{\mathrm{fus}}$, $\mathrm{p}^{\mathrm{sar}}$, $\mathrm{p}^{\mathrm{opt}}$ and the ground truth (y) the supervised loss and the consistency loss is calculated with the Power Jaccard loss (PJL) (Duque-Arias et al., 2021), which is defined as follows:

$$\mathfrak{L}_{pJacc}(p, y, c) = 1 - \frac{(p \cdot y) + \varepsilon}{(p^c + y^c - p \cdot y) + \varepsilon}$$

where:

- y represents the ground truth binary mask
- p represents the predicted binary mask
- c is a constant parameter
- ε is a small value added for numerical stability (to avoid division by zero)

The total loss is then calculated as:

$$\begin{aligned}
\mathfrak{L}_{\text{total}} &= \mathfrak{L}_{\text{supervised}} + \varphi \cdot \mathfrak{L}_{\text{consistency}} \\
\mathfrak{L}_{\text{supervised}} &= \mathfrak{L}_{\text{pJacc}}(p^{\text{fus}}, y, 2) + \mathfrak{L}_{\text{pJacc}}(p^{\text{sar}}, y, 2) + \mathfrak{L}_{\text{pJacc}}(p^{\text{opt}}, y, 2) \\
\mathfrak{L}_{\text{consistency}} &= \mathfrak{L}_{\text{pJacc}}(p^{\text{sar}}, p^{\text{opt}}, 2)
\end{aligned}$$

The proposed unsupervised domain adaptation approach, with the combination of consistency regularization and multi-modal data, aims to bridge the gap between the source and target domains, enabling the model to generalize well to the target domain without requiring labeled data. By exploiting the complementary information in SAR and optical data, the model can effectively adapt to the characteristics of the target domain, resulting in improved performance for the urban area segmentation task. (Hafner, Ban, & Nascetti, 2022b)

5.2 Loss functions

In the original setup proposed by Hafner, Ban, and Nascetti, the predictions based on a single modality exhibit noticeable differences in terms of F1-score and visual appearance when comparing samples. These dissimilarities arise due to the inherent characteristics of the SAR and optical streams within the network, which ideally should produce highly similar features. To encourage such similarity, a Consistency Loss is employed, calculated using the Power Jaccard loss. However, the obtained results are not satisfactory, necessitating the exploration of alternative loss functions or weighting schemes that promote the generation of more similar features by the LU-Nets.

Initially, we tried to enhance the importance of the consistency loss by increasing it's weighting, aiming to encourage the network to generate more similar features. However, this approach was found to be unsuccessful. Increasing the weight of the consistency loss did not have a notable impact on performance. In fact, beyond a certain point, when the weight became too high, the model's efficiency in learning decreased. Surprisingly, instead of becoming more similar, the feature outputs actually became more dissimilar compared to using a weight of 0.5. As an alternative, various similarity loss functions were tested to replace the Power Jaccard loss. These alternative functions included the Binary Cross Entropy loss, Contrastive loss, Maximum Mean Discrepancy (MMD) loss, L1 loss, and L2 loss, which are described in detail in the following.

5.2.1 Binary Cross Entropy Loss

Binary Cross Entropy (BCE) loss, also known as the log loss, is a commonly used loss function for binary classification tasks. It measures the dissimilarity between predicted and target binary values. BCE loss is particularly suitable for similarity-based tasks when the goal is to determine the similarity or dissimilarity between two instances. By minimizing the BCE loss, the network learns to assign higher probabilities to similar instances and lower probabilities to dissimilar instances.

The formula for BCE loss is given by:

$$\mathfrak{L}_{\mathrm{BCE}} = -\frac{1}{N}\sum_{i=1}^{N}\left(y_i \cdot \log(\hat{y}_i) + (1-y_i)\cdot \log(1-\hat{y}_i)\right)$$

where y_i represents the binary target value (0 or 1), $\hat{y}_i$ represents the predicted probability, and N is the total number of instances. (Godoy, 2022)

5.2.2 Intersection over Union Loss

Intersection over Union (IoU) loss is a commonly used loss function in object detection and segmentation tasks. It measures the overlap between predicted and target bounding boxes or masks. IoU loss encourages accurate localization by penalizing low overlap. Minimizing IoU loss improves the alignment of predicted regions with the ground truth, leading to more precise object localization. It is often used alongside other loss functions and enables quantitative evaluation of region matching.

The formula for IoU loss is:

$$\mathfrak{L}_{\mathrm{IoU}} = 1 - \frac{\mathrm{I}}{\mathrm{U}} = 1 - \frac{|GT * Pred|}{|GT + Pred - (GT * Pred)|}$$

where the Area of Intersection (I) represents the overlapping area between the predicted and target regions, and the Area of Union (U) represents the combined area of the predicted and target regions. GT and Pred stand for the true segmentation and for the predicted segmentation respectively. By incorporating IoU loss into the training process, models can be trained to produce more accurate and visually appealing object detection or segmentation results. (van Beers, Lindström, Okafor, & Wiering, 2019)

5.2.3 Contrastive Loss

Contrastive loss is a loss function commonly used in unsupervised learning tasks. It aims to map similar input samples to similar representations in a latent space while pushing dissimilar samples apart. To calculate the contrastive loss, the cosine distances between the positive example and negative examples are computed. The cosine distance measures the angle between vectors, where similar vectors have larger values closer to 1 and dissimilar vectors have values closer to 0 or negative values. The distances are treated as prediction probabilities and passed through a softmax function, which normalizes the values to a range of 0 to 1. The softmax output represents the probabilities of the positive and negative examples being similar.

The contrastive loss function is similar to the softmax function with the addition of the cosine similarity and a temperature parameter. The numerator of the loss function consists of the cosine distance of the positive example, while the denominator includes the cosine distances from the negative examples. The loss is computed by taking the negative logarithm of the positive example's softmax probability divided by the sum of all softmax probabilities. The aim is to minimize the loss, making the similar examples have probabilities close to 1 and dissimilar examples have probabilities close to 0.

The formula for Contrastive Loss is given by:

$$\mathfrak{L}_{\text{contrastive}} = -\log \frac{\exp(\text{sim}(z_i, z_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(\text{sim}(z_i, z_j)/\tau)}$$

where $\text{sim}(u, v)$ equals the cosine similarity which is calculated as follows in a batch of N Sentinel-1 and Sentinel-2 pairs.

$$\text{sim}(u, v) = \frac{u^T v}{\|u\| \, \|v\|}$$

Contrastive loss is a powerful tool for various unsupervised learning tasks. It enables the creation of meaningful representations in a latent space by encouraging similarity for positive examples and dissimilarity for negative examples. By optimizing the siamese network with contrastive loss, impressive results have been achieved in papers like SimCLR. (T. Chen, Kornblith, Norouzi, & Hinton, 2020; Williams, 2023)

In this project positive samples are represented by Sentinel-1 and Sentinel-2 images of the same area, whereas negative samples are images from different locations. This means in one batch with 8 samples from Sentinel-1 and Sentinel-2 each, there were 8 positive pairs and 56 negative pairs, if each Sentinel-1 image can be combined with all Sentinel-2 images of the batch.

5.2.4 Maximum Mean Discrepancy Loss

Maximum Mean Discrepancy (MMD) Loss is a kernel-based statistical measure that quantifies the discrepancy between two probability distributions. In the context of deep learning, MMD loss can be used as a regularization term or a similarity metric. By minimizing the MMD loss, the network is encouraged to generate feature representations that are statistically similar across different modalities or domains. MMD loss is advantageous as it does not rely on explicit class labels and can capture higher-order statistics of the data distribution.

The MMD loss can be computed using the RBF kernel, which measures the similarity between two feature vectors based on their Euclidean distance in the feature space. By incorporating the RBF kernel within the MMD framework, we can effectively quantify the dissimilarity between the embeddings.

In the case of Sentinel-1 (X) and Sentinel-2 (Y) feature embeddings with one-to-one correspondence, the MMD loss with the RBF kernel is expressed as:

$$\mathfrak{L}_{\text{MMD}}(X,Y) = \left\| \frac{1}{n_x^2} \sum_{i=1}^{n_x} \sum_{j=1}^{n_x} k(x_i, x_j) - \frac{2}{n_x n_y} \sum_{i=1}^{n_x} \sum_{j=1}^{n_y} k(x_i, y_j) + \frac{1}{n_y^2} \sum_{i=1}^{n_y} \sum_{j=1}^{n_y} k(y_i, y_j) \right\|^2$$

Here, n_x and n_y represent the number of feature embeddings in sets X and Y, respectively. The RBF kernel function k between x_i from set X and y_j from set Y is given by:

$$k(x_i, y_j) = \exp\left(-\gamma \|x_i - y_j\|^2\right)$$

where γ is the bandwidth parameter of the RBF kernel.

Minimizing the MMD loss encourages similarity between Sentinel-1 and Sentinel-2 feature embeddings from the same geographic location, while preserving dissimilarity between embeddings from different locations. (Gretton, Borgwardt, Rasch, Schölkopf, & Smola, 2012, p.728-729)

5.2.5 Mean Absolute Error Loss

L1 loss, also known as the mean absolute error (MAE) loss, calculates the absolute differences between predicted and target values. It promotes sparsity in the learned features by penalizing large deviations from the ground truth. MAE loss is less sensitive to outliers compared to other loss functions like mean squared error (MSE) loss, making it suitable for scenarios where robustness against outliers is important.

The formula for MAE loss is given by:

$$\mathfrak{L}_{\text{MAE}} = \frac{1}{N} \sum_{i=1}^{N} |y_i - \hat{y}_i|$$

where y_i represents the target value, $\hat{y}_i$ represents the predicted value, and N is the total number of instances. (Seif, 2022)

5.2.6 Mean Squared Error Loss

L2 loss, also known as the mean squared error (MSE) loss, calculates the squared differences between predicted and target values. It penalizes larger errors more severely than MAE loss and encourages the network to converge towards the mean of the target values. MSE loss is widely used in various regression tasks and often leads to smoother and more distributed feature representations.

The formula for MSE loss is given by:

$$\mathfrak{L}_{\mathrm{L2}} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

where y_i represents the target value, $\hat{y}_i$ represents the predicted value, and N is the total number of instances. (Seif, 2022)

5.3 Proposed network with the Discriminator

As in section 6 described the approaches described above did not improve the performance metrics significantly. Therefore the consistency loss is entirely replaced by a Discriminator network, as it is expected to outperform the loss functions, as it can learn the important metrics to produce similar features itself. The proposed architecture is displayed in figure 10 and is refereed as "Discriminator approach" troughout this thesis.

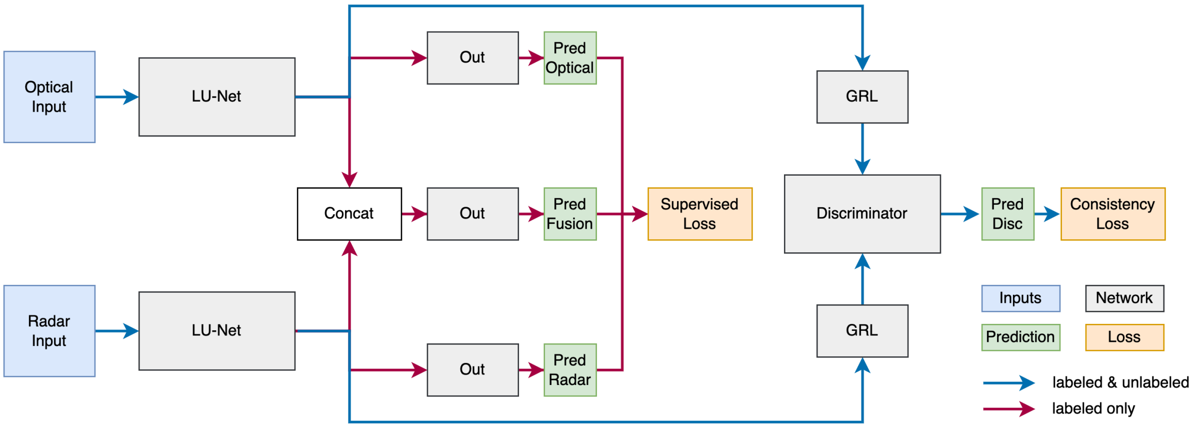


Figure 10: Proposed network architecture for multi model building segmentation with discriminator for similar feature embeddings of Sentinel-1 and Sentinel-2 data. (own illustration)

The LU-Nets employed in this study are equivalent to those utilized by Hafner, Ban, and Nascetti in their network, ensuring improved comparability (see fig. 9).

However, there are several additional modifications compared to the Fusion-DA that contribute to the stability and effectiveness of the training process. Notably, a learning rate scheduler is incorporated into the network architecture. This scheduler dynamically adjusts the learning rate during training, gradually reducing it towards the end of the training process. This technique helps us to stabilize the learning process and prevent abrupt changes that hinder convergence. Furthermore, another significant distinction is the utilization of a different data sampler in the training process. Unlike the Fusion-DA approach, where batches are randomly sampled from the entire dataset, this study employs a data sampler that includes an equal proportion of labeled and unlabeled data in each batch. This approach aims to enhance training stability by providing a balanced representation of both labeled and unlabeled samples during the learning process.

Concerning the general pipeline of our approach, the supervised loss is calculated in the same manner as in the Fusion-DA. Regarding the unsupervised loss, our approach differs from the Fusion-DA. While the Fusion-DA only utilizes unlabeled data, our method passes all samples through the pipeline, effectively doubling the amount of data used to calculate the consistency loss at each step.

As previously mentioned, the consistency loss function is replaced in our approach with a discriminator network, which is expected to yield superior results compared to a simple metric loss function. Additionally, rather than using prediction maps, we employ the entire feature embedding from the LU-Nets as input for the discriminator. The primary objective of the discriminator is to distinguish the data source modality based on the feature embedding. By training a robust discriminator, the LU-Nets are forced to generate similar feature embeddings regardless of the data source, thereby confusing the discriminator. To enable this, a gradient reversal layer (GRL) is employed during backpropagation, which multiplies all gradients by -1. This encourages the feature embeddings to become more similar, making it challenging for the discriminator to distinguish the domain of origin, while simultaneously improving the discriminator's performance through optimization of the consistency loss. The optimum discriminator accuracy is attained at 0.5, indicating a scenario where the feature embeddings are indistinguishably similar, such that the discriminator is unable to accurately classify the region of origin (Goodfellow et al., 2014). At this point, the discriminator's predictive capability is equivalent to random guessing, with no meaningful differentiation between domains based on the given feature embeddings.

The discriminator network was designed to make pixelwise decisions for the origin domain, rather than relying on global decisions for the entire image. This approach aimed to prevent misclassifications based on small, local features and therefore improve the accuracy of domain differentiation. By focusing on individual pixels, the network aimed to provide more precise and reliable results, considering the discriminative power. This design choice aimed to enhance the overall performance of the discriminator network in domain classification tasks. It enables the network to apply gradients precisely where dissimilarities in the feature embeddings allowed the discriminator to distinguish between domains.

In figure 11 the discriminator architecture used is depicted. Many different architectures have been tested with various depths and as it is also shown in the implementation of the GRL that a small discriminator network is sufficient. (Fungtion, n.d.; Ganin & Lempitsky, 2015)

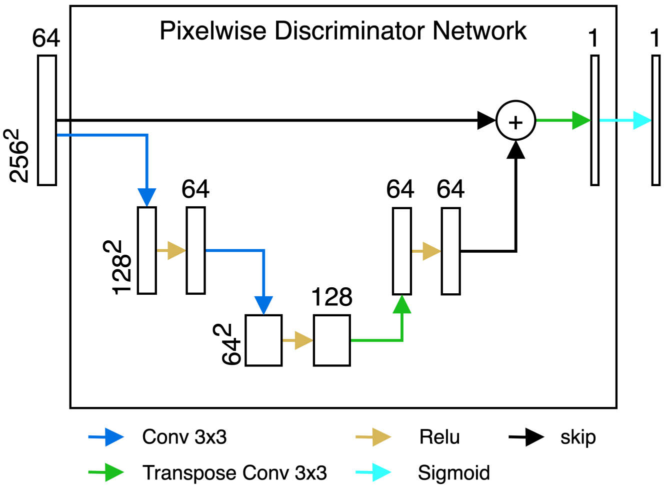


Figure 11: Discriminator network with width/height (left) and number of channels (top) for each step. (own illustration)

The discriminator network architecture employed in this thesis consists of convolutional layers with a kernel size of 3 and stride 2, facilitating feature downsampling on the encoder side. These layers are accompanied by the rectified linear unit (ReLU) activation function to enhance non-linearity in the network. On the decoder side, transpose convolutions with kernel size 3 and stride 2 as well, similar to those utilized in the LU-Net architecture, are employed to upsample the features back to their original input size.

Additionally, skip connections are incorporated into the network design. These connections serve the purpose of including information from the original input on the decoder side, enabling the network to leverage both high-level and low-level features for more accurate domain classification. By combining information from multiple levels of abstraction, the network gains a more comprehensive understanding of the input data, leading to improved performance.

Furthermore, the final step in the network architecture involves applying a sigmoid activation function to obtain probability-like values for each pixel. This allows the network to produce output indicating the origin domain for each pixel, providing a measure of confidence in the domain classification.

It is worth noting that the receptive field of this network is limited to 18 pixels. This deliberate choice ensures that the network focuses on a localized area, disregarding artifacts or influences from distant regions. By confining the receptive field to a small area relative to the image size, the network avoids potential biases caused by irrelevant information and maintains the ability to make accurate domain distinctions.

Numerous network architectures were systematically evaluated to explore their potential for domain classification. These architectures varied in depth, ranging from 1 to 4 layers, thereby influencing the receptive field of the network (ranging from 7x7 up to 270x270 pixels). Additionally, the LU-Net architecture, depicted in figure 9, was also tested at multiple depths. However, none of these alternative architectures demonstrated noteworthy improvements in performance compared to the selected architecture illustrated in figure 11, which delivered the best results considering the evaluation metrics, losses and discriminator accuracies.

The ground truth labels for the discriminator were generated as tensors of the same size as the input images. These tensors were filled with ones or zeros, depending on the corresponding domain of the image, establishing a binary classification task. To calculate the discriminator loss, the Binary Cross Entropy (BCE) loss with logits was employed. This loss function measures the dissimilarity between the predicted probabilities and the ground truth labels.

5.4 Input Augmentation

As an alternative approach to enhance the built-up area segmentation network proposed by Hafner, Ban, and Nascetti, an attempt was made to augment the training process by incorporating Stable Diffusion feature embeddings into the network's input (Rombach, Blattmann, Lorenz, Esser, & Ommer, 2022).

The stable diffusion model is based on the concept of diffusion, which involves two fundamental steps. The first step is the forward diffusion process, where noise is incrementally added to an image. The second step is the reverse diffusion process, represented by a denoising U-Net, which aims to predict and subtract the noise from the original input.

Once this U-Net is trained, it can generate images by taking pure noise as input and attempting to denoise it, similar to the training phase. This step can be iterated multiple times using the output from the previous iteration until a final noise-free image is obtained. However, this process becomes computationally intensive when numerous iterations are required or when dealing with large images. To address this, the same procedure can be applied in the latent space of the image, known as the "Latent Diffusion Model" (LDM), or in this case, referred to as stable diffusion. To enable this, a variational autoencoder (VAE) is trained to compress the images into a meaningful latent space (Steins, 2023). In this study, the employed VAE compresses RGB images of size 512x512 into latent spaces of size [4x64x64], preserving the information with a compression factor of 48 (Wong, 2023).

Additionally, the stable diffusion models can generate images from textual inputs. To facilitate this capability, the internal diffusion model of stable diffusion can incorporate conditions in the form of segmentations, other images, or text. While this feature holds great importance and fascination for various applications and individuals, it is not the focus of this thesis (Rombach et al., 2022; Steins, 2023).

To improve the built-up area segmentation, only the feature embeddings extracted by the VAE from the stable diffusion model are utilized, disregarding the entire stable diffusion framework. Since the VAE must effectively enable the diffusion process in the latent space, the extracted features are expected to represent the most significant structures in the image. By incorporating these features, it becomes easier for the LU-Nets used in the Fusion-DA approach to extract the relevant features for the segmentation task, as they do not need to learn all the structures from scratch.

Given that the VAE is trained to extract features from RGB images with three channels, while the Sentinel-1 and Sentinel-2 images have two and ten channels, respectively, each channel is separately passed through the VAE by replicating it three times along the channel axis. The VAE used, was taken from the stable diffusion model v2.1 base (Rombach et al., 2022). During the training process, when the images are loaded, the [4x64x64] latent spaces are upsampled to match the size of the original inputs and concatenated with them. The impact of these additional inputs is shown in section 6 and this setup is refereed as "Input Augmentation approach".

6 Results and Discussion

In this chapter, we will share the results of all the experiments conducted in this thesis. This includes testing different loss functions, measuring the effectiveness of the Discriminator and Input Augmentation approach using metrics, and providing visual representations of the predictions. Despite using the exact same code provided on Github by Hafner, Ban, and Nascetti, we were not able to reproduce the exact same metric values as shown in their paper. Therefore in the following all results for the Fusion-DA are based on the reproduced runs to enable a better comparability of the test cases. Overall the metrics of the test set have a very high variance and therefore the following results are all based on the metrics of the last epoch. This does not mean the models did not converge, because they did according to the training and validation metrics. There are multiple reasons, which can cause the difference between validation and test metrics. First, the size of the test set is much smaller than the training and validation set, also the test set was not generated in the same manner. Further, the test set covers locations, which are not included in the training set due to the original goal of domain adaptation by Hafner, Ban, and Nascetti.

6.1 Loss functions

In sections 5.1 and 5.2 different loss functions for the Fusion-DA approach have been described and tested. The results for all loss functions are shown in tables 4 to 6 for the validation and test set.

In table 4 the metrics for the output segmentations based on Sentinel-1 and Sentinel-2 are shown. One can see that the IoU loss outperforms the other loss functions for most of the metrics. For example the F1-score which combines recall and precision in one metric and therefore is concerned as the most important indicator followed by IoU, outperforms PJL and contrastive loss by 3% on the test set. On the other hand for the validation set the F1-score and the IoU are almost equal, and since the validation set give a better indicator for the methodology used, as the variance is much smaller, the different loss functions are not concerned to work significantly better than PJL. Also if one looks at the second to last epoch on the test set, PJL would have performed best for F1-score and IoU instead of the IoU loss.

Fusion from	**F1 score**		**Precision**		**Recall**		**IoU**		**SSIM**	
Fusion-DA	**Val.**	**Test**	**Val.**	**Test**	**Val.**	**Test**	**Val.**	**Test**	**Val.**	**Test**
PJL	0.768	0.634	0.801	0.747	**0.738**	0.551	0.623	0.464	0.726	0.544
Contrastive	0.768	0.638	0.818	0.717	0.724	0.574	0.623	0.468	0.723	0.558
MMD	**0.771**	0.606	0.812	0.741	0.735	0.513	**0.628**	0.435	0.723	0.548
BCE	0.768	0.515	0.812	**0.750**	0.729	0.392	0.624	0.347	0.718	0.534
IoU	0.767	**0.667**	**0.836**	0.709	0.708	**0.630**	0.621	**0.500**	**0.733**	**0.584**
L1	0.766	0.622	0.826	0.726	0.714	0.545	0.621	0.452	0.710	0.544
L2	0.766	0.609	0.809	0.735	0.727	0.519	0.621	0.437	0.696	0.497

Table 4: Comparison for multiple consistency loss functions evaluated on the fusion data of validation and test set.

In the scenario where only Sentinel-2 data is available during inference, the resulting metrics are shown in table 5. It can be observed that for the crucial metrics of F1-score and IoU, both PJL and contrastive loss yield the best performance. While the improvements of contrastive loss for the validation set are relatively minor (below 1%), PJL achieves the highest scores for the test set. Compared to table 4, when Sentinel-1 and Sentinel-2 data is available, the IoU loss does not achieve competing results to PJL. Overall, PJL demonstrates superior performance compared to all the other tested loss functions.

Optical from	**F1 score**		**Precision**		**Recall**		**IoU**		**SSIM**	
Fusion-DA	**Val.**	**Test**	**Val.**	**Test**	**Val.**	**Test**	**Val.**	**Test**	**Val.**	**Test**
PJL	0.738	**0.654**	0.772	0.603	0.706	0.714	0.584	**0.486**	0.537	0.336
Contrastive	**0.744**	0.589	0.777	0.702	0.714	0.508	**0.592**	0.418	0.704	**0.563**
MMD	0.633	0.499	0.815	0.415	0.515	0.627	0.463	0.333	0.144	0.081
BCE	0.723	0.312	0.712	**0.704**	**0.734**	0.201	0.566	0.185	0.623	0.389
IoU	0.728	0.629	0.815	0.515	0.659	0.808	0.573	0.459	**0.723**	0.556
L1	0.416	0.358	0.911	0.221	0.270	**0.945**	0.263	0.218	0.101	0.039
L2	0.489	0.403	**0.911**	0.257	0.334	0.934	0.323	0.252	0.092	0.038

Table 5: Comparison for multiple consistency loss functions evaluated on the Sentinel-2 data of validation and test set.

In the third case, where only Sentinel-1 data is available during inference, the metric scores are presented in table 6. Notably, the IoU loss achieves favorable results for several metrics, particularly when compared to PJL on the test set. However, for the validation set, the IoU loss exhibits lower scores for F1-score and IoU compared to PJL. Consequently, the IoU loss does not consistently enhance the model in a robust manner. Moreover, while certain tested loss functions show improvement in specific metrics, they fail to deliver across all metrics. As a result, no single loss function clearly outperforms the baseline configuration.

SAR from	**F1 score**		**Precision**		**Recall**		**IoU**		**SSIM**	
Fusion-DA	**Val.**	**Test**	**Val.**	**Test**	**Val.**	**Test**	**Val.**	**Test**	**Val.**	**Test**
PJL	0.671	0.571	0.761	0.622	0.600	0.528	0.504	0.400	0.463	0.333
Contrastive	0.665	0.585	0.743	0.567	0.602	0.603	0.498	0.413	0.605	0.422
MMD	**0.679**	0.560	0.733	0.624	0.632	0.508	**0.514**	0.389	0.581	0.426
BCE	0.550	0.314	0.474	0.621	**0.654**	0.210	0.379	0.186	**0.680**	**0.585**
IoU	0.647	**0.597**	**0.832**	0.571	0.530	**0.626**	0.479	**0.426**	0.651	0.558
L1	0.601	0.406	0.556	0.616	0.653	0.302	0.429	0.254	0.604	0.464
L2	0.672	0.522	0.708	**0.631**	0.640	0.446	0.506	0.353	0.594	0.451

Table 6: Comparison for multiple consistency loss functions evaluated on the Sentinel-1 data of validation and test set.

In summary, upon examining the metrics presented in tables 4 to 6, it is evident that no loss function exhibits a substantial improvement over PJL, as employed in the baseline configuration. While certain loss functions demonstrate better performance on specific metrics, it is challenging to draw a straightforward conclusion. However, it is clear that PJL remains a valid approach that performs well across a majority of the metrics. Despite the potential advantages offered by alternative loss functions in certain scenarios, the overall effectiveness and versatility of PJL make it a reliable choice for the task at hand. Therefore, in the following the Fusion-DA approach is always referring to the configuration with the PJL as its loss function.

6.2 Discriminator and Input Augmentation

In the following tables, the metrics for the fusion-DA baseline approach, the single/fusion sensor approach, as well as the Discriminator- and Input Augmentation approach are shown.

In table 7 the metrics for all four approaches when both modalities are available at inference time are displayed. The first approach, referred to as Fusion only, concatenates Sentinel-1 and Sentinel-2 data at input level. Detailed explanations of the other approaches can be found in section 5.

The Fusion-DA approach shows only partial improvements compared to Fusion only. Unfortunately, the results obtained by Hafner, Ban, and Nascetti, who achieved significant metric improvements with their Fusion-DA approach, could not be replicated. The Discriminator approach performs similarly to the Fusion-DA approach, with only slight differences in the metric scores.

The most substantial improvement is observed with the incorporation of the input features from the VAE from the stable diffusion network as additional inputs. On the validation set, the F1-score and IoU improve significantly by 2.3% and 3.1% respectively. These metrics also exhibit improvement on the test set, although the specific values tend to vary strongly across epoch.

Fusion	**F1 score**		**Precision**		**Recall**		**IoU**		**SSIM**	
	Val.	**Test**	**Val.**	**Test**	**Val.**	**Test**	**Val.**	**Test**	**Val.**	**Test**
Fusion only	0.758	0.664	0.803	0.642	0.718	**0.688**	0.610	0.497	0.658	0.397
Fusion-DA	0.768	0.634	0.801	**0.747**	0.738	0.551	0.623	0.464	0.726	**0.544**
Discriminator	0.765	0.634	0.809	0.720	0.726	0.566	0.620	0.464	0.713	0.503
Input Aug.	**0.781**	**0.671**	**0.820**	0.695	**0.746**	0.648	**0.641**	**0.505**	**0.749**	0.503

Table 7: Comparison of described approaches with Sentinel-1 and Sentinel-2 data of validation and test set.

The table below shows different measurements when only Sentinel-2 data is used for analysis. Different methods have their own advantages. For example, the fusion-DA approach performs the best in terms of F1-score and IoU, but it has the lowest results in SSIM. On the other hand, the Input Augmentation approach has nearly the same performance as Fusion-DA and also has competitive SSIM scores. The Discriminator approach strongly improves the SSIM value, which measures the structural similarity between the ground truth and the prediction and therefore gives an indicator of the visual perception. On the validation set the Discriminator scores are almost as good as in the other methods, while on the test set all scores are significantly worse. Overall, the Input Augmentation approach is the preferred choice based on all the measurements, although there are no significant improvements except for the SSIM.

Optical	**F1 score**		**Precision**		**Recall**		**IoU**		**SSIM**	
	Val.	**Test**	**Val.**	**Test**	**Val.**	**Test**	**Val.**	**Test**	**Val.**	**Test**
Sent.-2 only	0.721	0.634	0.785	0.603	0.667	0.668	0.564	0.464	0.642	0.438
Fusion-DA	**0.738**	**0.654**	0.772	0.603	0.706	0.714	**0.584**	**0.486**	0.537	0.336
Discriminator	0.733	0.572	0.761	**0.695**	**0.707**	0.486	0.579	0.400	**0.665**	**0.512**
Input Aug.	0.737	0.651	**0.834**	0.569	0.659	**0.759**	0.583	0.482	0.654	0.392

Table 8: Comparison of described approaches with Sentinel-2 data of validation and test set.

When only Sentinel-1 data is used, the overall performance across all metrics is moderate. The F1-score and IoU values are relatively low compared to the other modalities. The Fusion-DA approach does not improve the metrics compared to using only Sentinel-1 data. However, there is an increase in precision while the recall is lower. On the other hand, the Discriminator approach achieves the highest F1 score and IoU values on the validation set. The Input Augmentation approach performs well on most of the metrics, even when it not scores the top results in some. It achieves the highest recall out of all the methods for the test set and the highest precision for the validation set. However, it does not significantly improve the F1 scores due to having the lowest precision values on the test set and the lowest recall values on the validation set by a large margin. These discrepancies between the test and validation sets indicate that the metrics should be interpreted carefully. Overall, based on the results, the Input Augmentation approach appears to be the most favorable choice among the discussed methods when considering the F1 score and IoU. However, when taking a closer look at precision and recall to understand how the F1 score was achieved, one should consider the Discriminator approach as the best method, as it provides more stability due to the lower discrepancies between the validation and test metrics.

SAR	**F1 score**		**Precision**		**Recall**		**IoU**		**SSIM**	
	Val.	**Test**	**Val.**	**Test**	**Val.**	**Test**	**Val.**	**Test**	**Val.**	**Test**
Sent.-1 only	0.675	0.571	0.720	0.599	**0.635**	0.547	0.509	0.400	0.549	0.310
Fusion-DA	0.671	0.571	0.761	**0.622**	0.600	0.528	0.504	0.400	0.463	**0.333**
Discriminator	**0.679**	0.584	0.753	0.594	0.617	0.574	**0.513**	0.412	**0.557**	0.291
Input Aug.	0.675	**0.611**	**0.821**	0.528	0.573	**0.726**	0.509	**0.440**	0.541	0.304

Table 9: Comparison of described approaches with Sentinel-1 data of validation and test set.

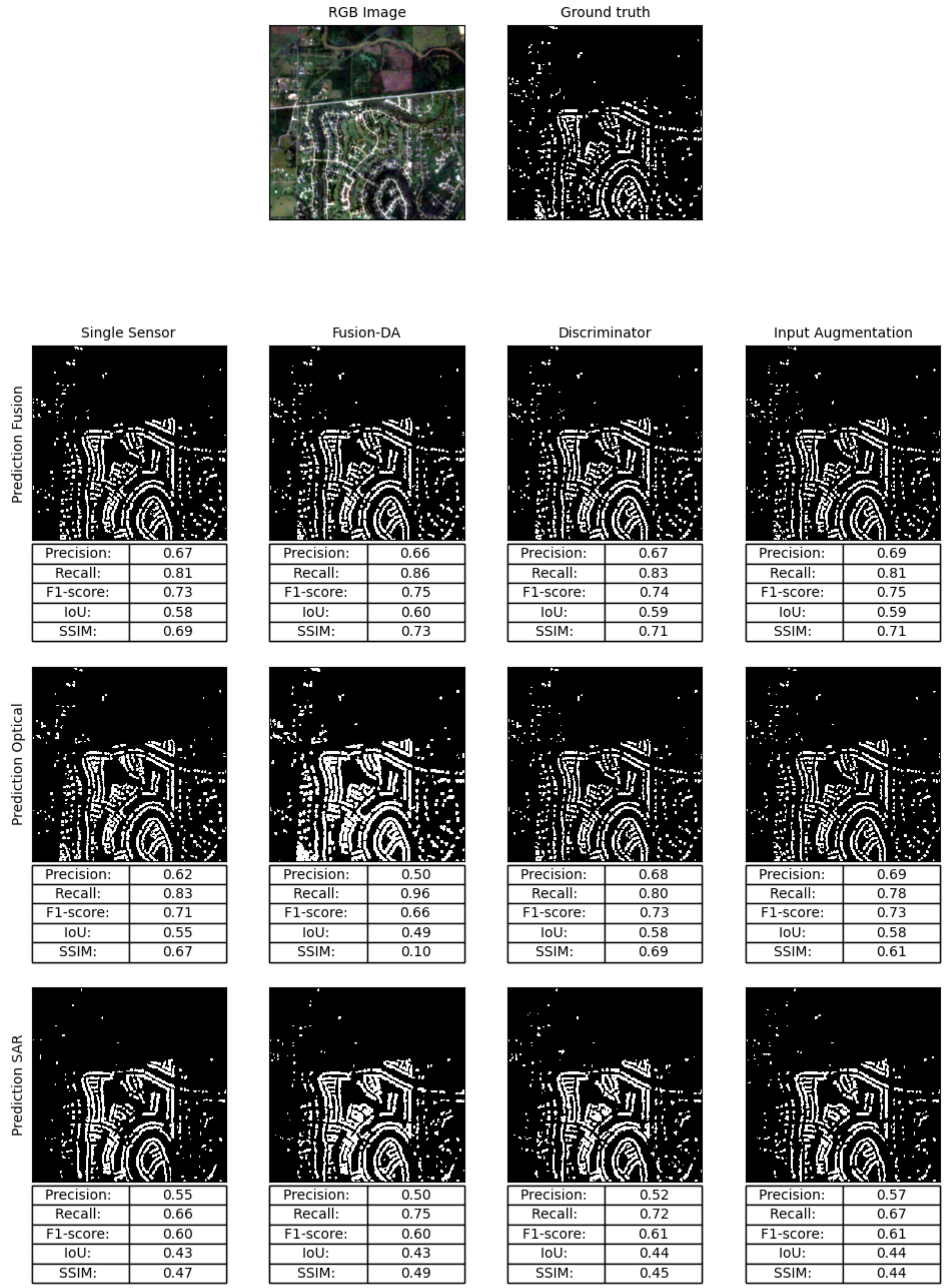
In conclusion the Input Augmentation approach performs best concerning all possible metrics for all cases of data availability. The metrics are very similar for all methods and therefore the improvements are not significant for all metrics. Therefore the results get inspected visually with some examples from the training and test set in the following section.

6.3 Examples

In the following section, we present visual examples of the training and test sets, illustrating the input with RGB images and their corresponding ground truth. Additionally, we provide four columns displaying prediction maps for built-up areas. Each row represents one of the four methods introduced earlier: Single Sensor, Fusion-DA, Discriminator, and Input Augmentation.
These examples demonstrate different scenarios of data availability. The first row corresponds to the case where both Sentinel-1 and Sentinel-2 data are available, while the second and third rows represent cases where only one of them is present. Also each segmentation includes a small table below, which indicate the metric values for the corresponding prediction.
The figures 12 and 13 showcase visualizations based on the training set for Houston, Los Angeles, and Miami in the USA. Similarly, figures 14 and 15 depict results from the test set, specifically locations in Egypt, Saudi Arabia, and the USA, respectively. Notably, the scores obtained for the test set generally exhibit lower values compared to those of the training set, as illustrated in tables 7 to 9.

The predictions for satellite observations in Houston, USA are depicted in figure 12. Initially, all methods exhibit similar performance, accurately recognizing the primary structures with only minor local variations. The provided metrics below the predictions also indicate minimal variation among the four methods, particularly for the Fusion and SAR segmentations. However, noticeable discrepancies emerge in the optical predictions, which solely rely on Sentinel-2 data. Upon examining all four methods, it becomes apparent that most metrics display consistent and comparable results, except for the Fusion-DA approach. Although the Fusion-DA approach achieves a comparable F1-score, its SSIM and IoU values are significantly lower. Additionally, the F1-score is influenced by distinct Precision and Recall scores, where the Recall approaches almost one while Precision is considerably lower in comparison. Consequently, the Fusion-DA approach detects nearly all buildings within the patch, resulting in an excessive number of detections. This leads to oversegmentation, and the elevated F1-scores are primarily attained by classifying regions resembling buildings, resulting in a blurred segmentation that lacks the level of detail observed in the other methods.

In figure 13 a more densely populated region situated in Los Angeles, USA is shown. Similar to the observations made in figure 12, the metrics demonstrate minimal variation across all methods. However, when analyzing the predictions derived from optical data, the recurring issue of blurring becomes evident, accompanied by a corresponding impact on the metric values. This effect is also noticeable to a lesser extent in the case of the Input Augmentation approach. Through visual examination, the Discriminator approach outperforms the others, because it seems to be the closest to the ground truth in all three scenarios of data availability, as also indicated by the SSIM score.



Prediction Fusion

	Single Sensor	Fusion-DA	Discriminator	Input Augmentation
Precision:	0.67	0.66	0.67	0.69
Recall:	0.81	0.86	0.83	0.81
F1-score:	0.73	0.75	0.74	0.75
IoU:	0.58	0.60	0.59	0.59
SSIM:	0.69	0.73	0.71	0.71

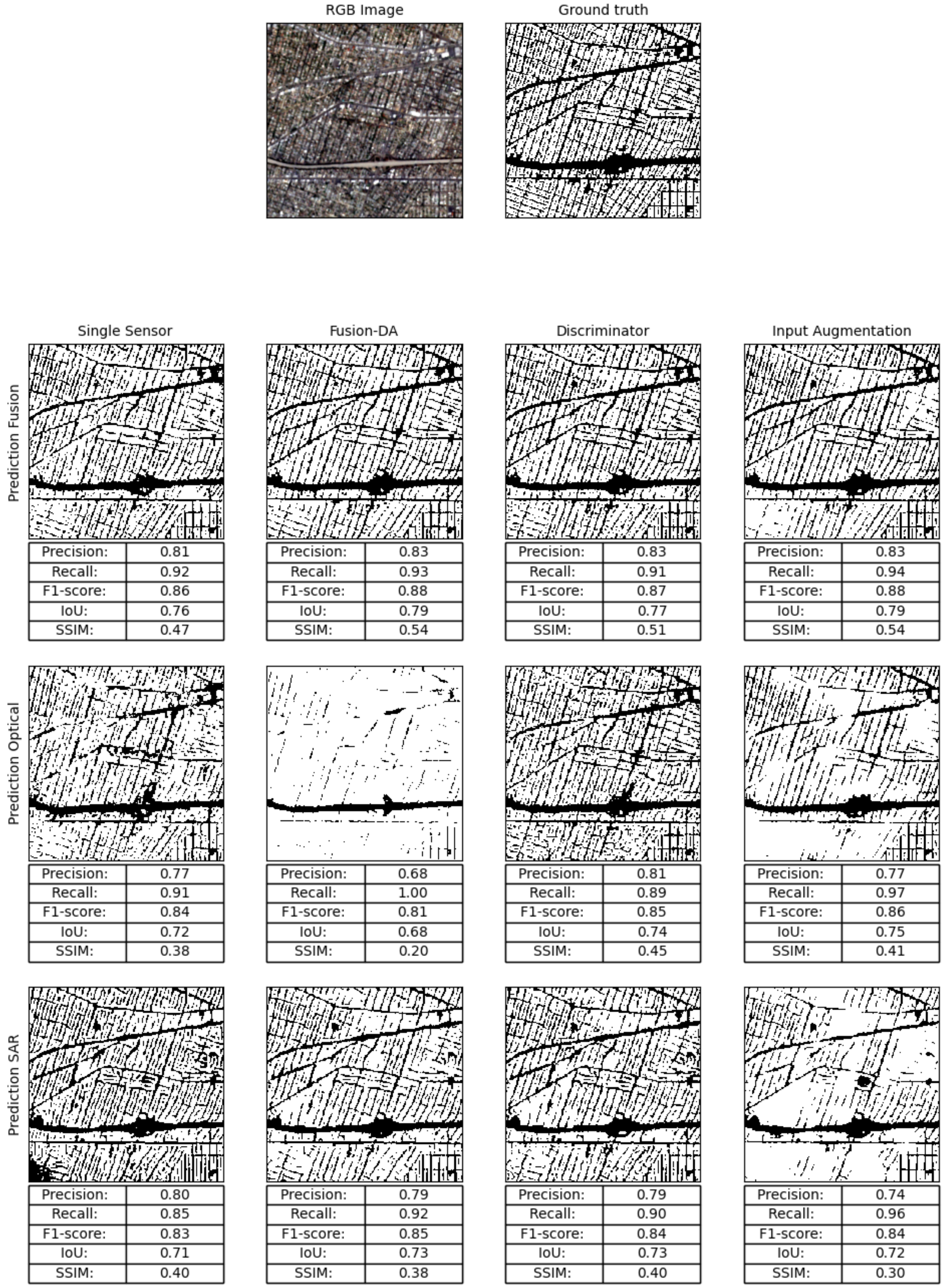
Prediction Optical

	Single Sensor	Fusion-DA	Discriminator	Input Augmentation
Precision:	0.62	0.50	0.68	0.69
Recall:	0.83	0.96	0.80	0.78
F1-score:	0.71	0.66	0.73	0.73
IoU:	0.55	0.49	0.58	0.58
SSIM:	0.67	0.10	0.69	0.61

Prediction SAR

	Single Sensor	Fusion-DA	Discriminator	Input Augmentation
Precision:	0.55	0.50	0.52	0.57
Recall:	0.66	0.75	0.72	0.67
F1-score:	0.60	0.60	0.61	0.61
IoU:	0.43	0.43	0.44	0.44
SSIM:	0.47	0.49	0.45	0.44

Figure 12: Example output segmentations for all proposed methods and data availability cases for a training set sample located in Houston, USA



Prediction Fusion

	Single Sensor	Fusion-DA	Discriminator	Input Augmentation
Precision:	0.81	0.83	0.83	0.83
Recall:	0.92	0.93	0.91	0.94
F1-score:	0.86	0.88	0.87	0.88
IoU:	0.76	0.79	0.77	0.79
SSIM:	0.47	0.54	0.51	0.54

Prediction Optical

	Single Sensor	Fusion-DA	Discriminator	Input Augmentation
Precision:	0.77	0.68	0.81	0.77
Recall:	0.91	1.00	0.89	0.97
F1-score:	0.84	0.81	0.85	0.86
IoU:	0.72	0.68	0.74	0.75
SSIM:	0.38	0.20	0.45	0.41

Prediction SAR

	Single Sensor	Fusion-DA	Discriminator	Input Augmentation
Precision:	0.80	0.79	0.79	0.74
Recall:	0.85	0.92	0.90	0.96
F1-score:	0.83	0.85	0.84	0.84
IoU:	0.71	0.73	0.73	0.72
SSIM:	0.40	0.38	0.40	0.30

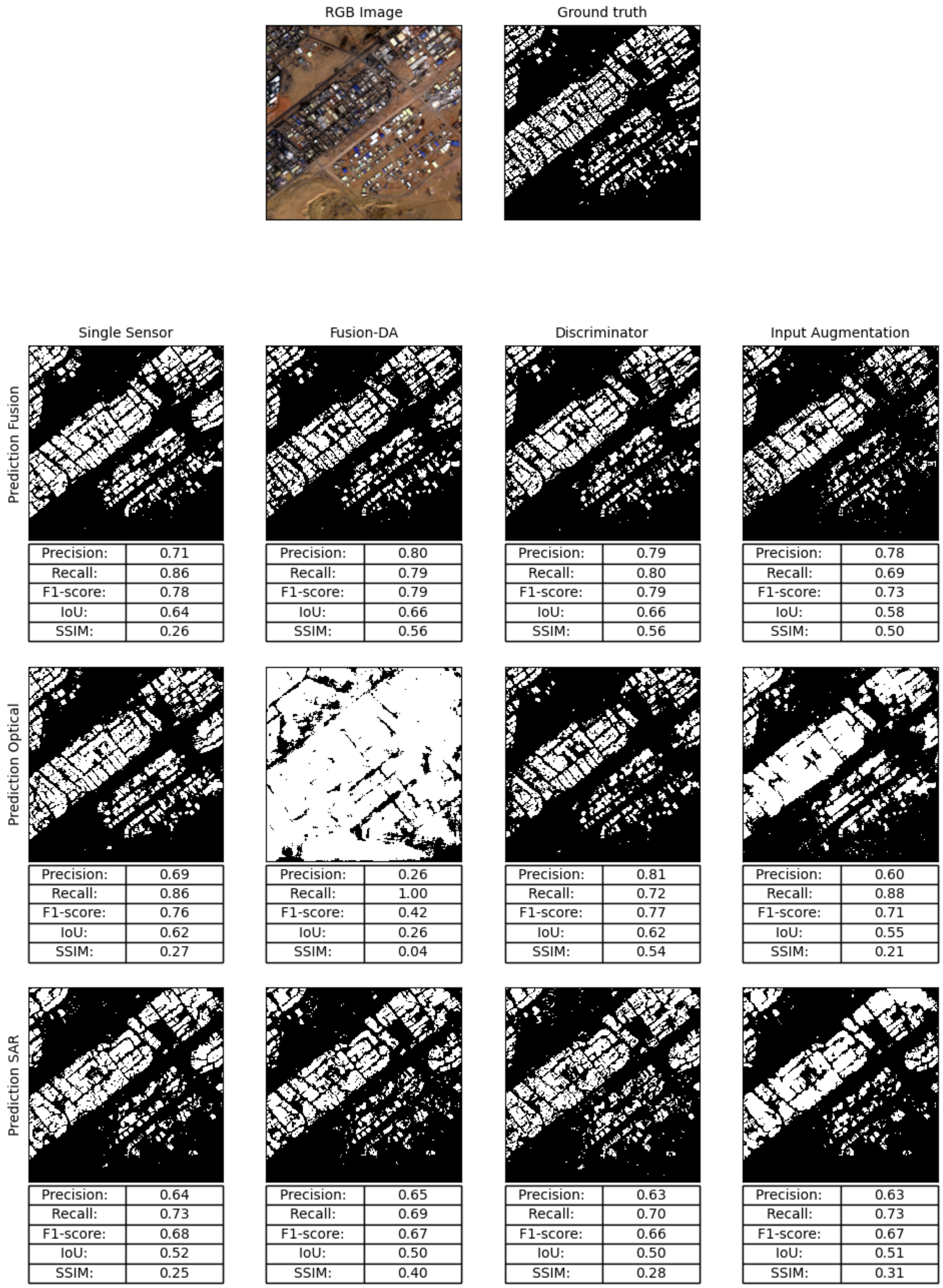
Figure 13: Example output segmentations for all proposed methods and data availability cases for a training set sample located in Los Angeles, USA

The first example of the test set is shown in figure 14, where a region in Egypt is depicted together with the corresponding prediction maps. The observations made on the training set for figures 12 and 13 can be applied as well on the test set, where most of the metrics are very similar, except for the blurry outputs of the optical predictions of the Fusion-DA and Input Augmentation approach. Also very interesting to mention is, that the inclusion of the additional input channels augmented by the VAE of the stable diffusion framework, strongly improves the Fusion-DA segmentation of the optical data. The influence on the SAR data is not significant, but one can see an increased blurriness for the Input Augmentation approach as well. Overall the Discriminator performs joint best with the Single Sensor approach, whereas the Discriminator achieves higher precision, to the cost of lower recall, but this means that out of the detected buildings the probability that it is indeed a building is higher. Also the Precision and Recall values are in general more balanced for the Discriminator than for the other methods. This example also shows that the models are able to make comparable accurate predictions even outside the source area of the training set. Except the Fusion-DA approach, which seems to struggle with change of surface structure in the optical data stream, which is probably not included in the training set as it is shown in the rgb image of figure 14. On the other the SAR data stream seems to adjust better to the varying surface, as it still achieves good metric scores.

The last example in figure 15 shows a region in the USA, and therefore is not affected as strongly as the example in Egypt by the shift of location, as most cities of the training set are located in the USA as well. The main difference in this example is the generation of the ground truth and therefore it is possible to focus more on this effect. One can see that the metric values show a similar pattern as for the examples from the training set in figures 12 and 13. Therefore the different ground truths seem not to have a major impact.

Overall the metrics for all methods are very similar, as also shown in tables 7 to 9. In the examples one can see some predictions, which show that close F1-scores not automatically indicate similar segmentation maps. For example in figure 15 the F1-score of the Fusion-DA and Discriminator approach are very close for the optical data stream, but the predictions look very different, what is also represented in precision and recall.

The Discriminator approach improves the results by visual inspection, especially for the case when only Sentinel-2 data is available at inference. This effect has been measures by the SSIM and is also shown in table 8, where the SSIM is improved strongly compared to the other methods also when looking not only at a single example. Also, the Discriminator network strongly improves the similarity of the prediction maps between the case when only Sentinel-1 data is available to the case when only Sentinel-2 data is available, compared to the PJL used in the Fusion-DA approach as a consistency loss. Therefore, the use of a Discriminator network instead of a simple loss function definitely improved consistency between the modalities.



Prediction Fusion	Single Sensor	Fusion-DA	Discriminator	Input Augmentation
Precision:	0.71	0.80	0.79	0.78
Recall:	0.86	0.79	0.80	0.69
F1-score:	0.78	0.79	0.79	0.73
IoU:	0.64	0.66	0.66	0.58
SSIM:	0.26	0.56	0.56	0.50

Prediction Optical	Single Sensor	Fusion-DA	Discriminator	Input Augmentation
Precision:	0.69	0.26	0.81	0.60
Recall:	0.86	1.00	0.72	0.88
F1-score:	0.76	0.42	0.77	0.71
IoU:	0.62	0.26	0.62	0.55
SSIM:	0.27	0.04	0.54	0.21

Prediction SAR	Single Sensor	Fusion-DA	Discriminator	Input Augmentation
Precision:	0.64	0.65	0.63	0.63
Recall:	0.73	0.69	0.70	0.73
F1-score:	0.68	0.67	0.66	0.67
IoU:	0.52	0.50	0.50	0.51
SSIM:	0.25	0.40	0.28	0.31

Figure 14: Example output segmentations for all proposed methods and data availability cases for a test set sample located in Egypt

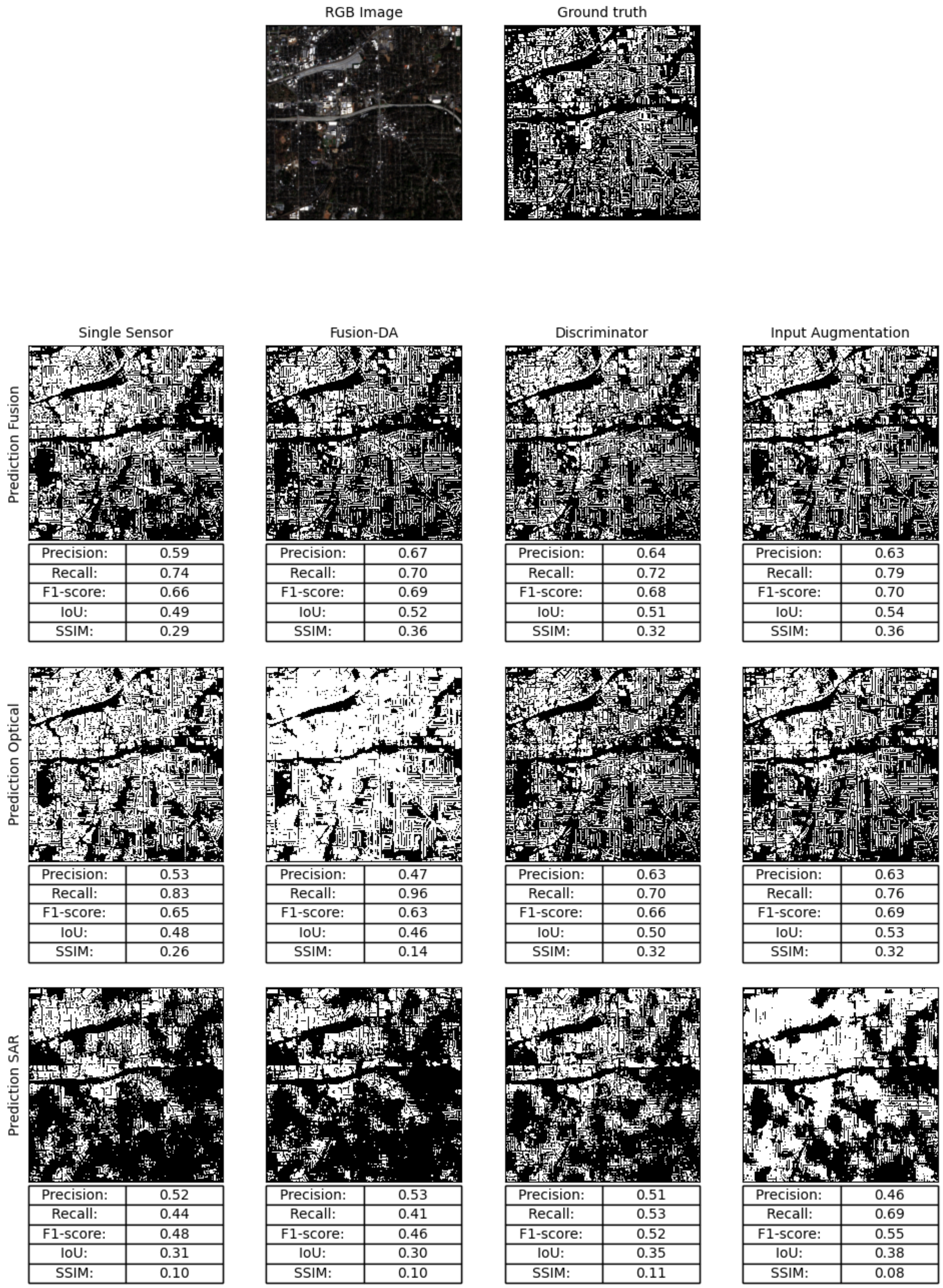


Figure 15: Example output segmentations for all proposed methods and data availability cases for a test set sample located in USA

7 Conclusion

In this thesis, we explored the domain adaptation techniques for semantic segmentation of remote sensing images using a fusion of Sentinel-1 and Sentinel-2 data, as well as both modalities separately. We started by discussing the importance of domain adaptation in remote sensing applications and the challenges associated with transferring knowledge from one domain to another. Our goal was to develop a model that can effectively leverage the complementary information provided by both Sentinel-1 and Sentinel-2 data sources, even when only one data source is available at inference time.

In the literature review, we examined existing methods for domain adaptation in semantic segmentation and identified the Fusion-DA approach proposed by Hafner, Ban, and Nascetti as a suitable framework for our research. We analyzed the key components of Fusion-DA and tested different loss functions in this framework, while we also implemented two new strategies based on the Fusion-DA approach, namely the Discriminator and the Input Augmentation approach.

We conducted extensive experiments to evaluate the effectiveness of different loss functions in improving the performance of the Fusion-DA model. We compared several consistency loss functions, including Power Jaccard loss, contrastive loss, Maximum Mean Discrepancy loss, Binary Cross-Entropy loss, Intersection over Union loss, L1 loss and L2 loss. We measured the performance of these loss functions on validation and test sets using various metrics such as F1-score, precision, recall, IoU, and Structural Similarity Index (SSIM).

We observed that the choice of the loss function did not significantly influence the performance of the Fusion-DA model. In the case of fusion data (Sentinel-1 and Sentinel-2), the IoU loss outperformed other loss functions, achieving the highest F1-score and IoU on the test set. However, PJL demonstrated competitive performance on the validation set. When only Sentinel-2 data was available during inference, PJL and contrastive loss showed the most promising results. For the scenario where only Sentinel-1 data was available, the IoU loss yielded favorable results in some metrics but did not consistently improve performance across all metrics. Despite achieving improvements with certain loss functions, no single loss function emerged as a clear winner in all cases.

In order to further improve the results, two new models were introduced, incorporating the Discriminator and Input Augmentation approaches. The performance of these models was evaluated based on different data availability scenarios.

For the case when both Sentinel-1 and Sentinel-2 data were available, the Input Augmentation approach exhibited superior performance in terms of F1-score, IoU, and SSIM for both the validation and test sets. This highlights the effectiveness of the Input Augmentation technique in enhancing the fusion of the two data sources and achieving better segmentation results.

When only Sentinel-2 data was present during inference, the Discriminator approach significantly improved the SSIM metric, while the other evaluation metrics remained at a similar level compared to the other methods on the validation set. However, the performance on the test set was noticeably worse. Despite this, overall, the Input Augmentation approach still outperformed other methods in this particular data availability scenario.

In the case where only Sentinel-1 data was available, all methods yielded comparable results for most metrics, with minor differences observed in precision and recall. Notably, the Input Augmentation approach exhibited larger disparities between precision and recall compared to the Discriminator approach.

Additionally, visual inspections of selected examples from the training and test sets were conducted to assess the quality of the segmentation maps. These visualizations revealed that pixel-based metrics may not always be sensitive enough to capture the true quality of the segmentation. However, in terms of visual performance, the Discriminator approach consistently achieved the best results across all three data availability scenarios, with the most significant deviations observed when only Sentinel-2 data was used. The Fusion-DA approach, on the other hand, tended to oversegment buildings, resulting in higher recall but lower precision and blurry prediction maps.

Overall, the findings highlight the effectiveness of the Discriminator approach in predicting fine-grained segmentations, what is also represented in high SSIM values especially for the optical data stream. Considering the comprehensive evaluation of metrics, as well as the visual assessments, the Input Augmentation approach and the Discriminator approach are the preferred choice for improving segmentation results depending on what metrics are most important for different usecases.

In conclusion, the Input Augmentation approach improves the metrics the most and would be a simple and useful tool for a better feature extraction process and especially improves segmentations for the case when only Sentinel-2 data is present at inference. Nevertheless it is important to also have a look at the final prediction maps and when doing that, one can see that also the Discriminator approach can be a beneficial approach, especially when finegrained segmentations are required.

As a further step it would be interesting to evaluate this methods on other datasets. For example the training set could be extended with ground truth data from South America, which would be available by now. Also, averaging multiple runs and calculate mean and variance would be very interesting to better quantify the results on the test set, which seems to be affected by a high variance at the moment. Finally a combination of the proposed approaches would be interesting, by using the augmented inputs from the VAE also for the Discriminator approach, as they provide advantages in different areas and a combination of both may lead to improved metric scores, while preserving the ability to produce finegrained prediction maps.

References

Bolte, J.-A., Kamp, M., Breuer, A., Homoceanu, S., Schlicht, P., Hüger, F., ... Fingscheidt, T. (2019). Unsupervised domain adaptation to improve image segmentation quality both in the source and target domain. In *2019 ieee/cvf conference on computer vision and pattern recognition workshops (cvprw)* (p. 1404-1413). doi: 10.1109/CVPRW.2019.00181

Cai, Y., Hu, X., Wang, H., Zhang, Y., Pfister, H., & Wei, D. (2022). *Learning to generate realistic noisy images via pixel-level noise-aware adversarial training.*

Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). *A simple framework for contrastive learning of visual representations.*

Chen, Z., Chen, G., Zhou, F., Yang, B., Wang, L., Liu, Q., & Chen, Y. (2020). A novel general semisupervised deep learning framework for classification and regression with remote sensing images. *IGARSS 2020 - 2020 IEEE International Geoscience and Remote Sensing Symposium*, 1323-1326.

Chini, M., Pelich, R., Hostache, R., Matgen, P., & Lopez-Martinez, C. (2018). Towards a 20 m global building map from sentinel-1 sar data. *Remote Sensing*, *10*(11). Retrieved from https://www.mdpi.com/2072-4292/10/11/1833 doi: 10.3390/rs10111833

Duque-Arias, D., Velasco-Forero, S., Deschaud, J.-E., Goulette, F., Serna, A., Decencière, E., & Marcotegui, B. (2021, February). On power Jaccard losses for semantic segmentation. In *VISAPP 2021 : 16th International Conference on Computer Vision Theory and Applications.* Vienne (on line), Austria. Retrieved from https://hal.science/hal-03139997

ESA. (2023). *Sentinel 2 overview.* Retrieved from https://sentinel.esa.int/web/sentinel/missions/sentinel-2/overview

Esch, T., Heldens, W., Hirner, A., Keil, M., Marconcini, M., Roth, A., ... Strano, E. (2017). Breaking new ground in mapping human settlements from space – the global urban footprint. *ISPRS Journal of Photogrammetry and Remote Sensing*, *134*, 30-42. Retrieved from https://www.sciencedirect.com/science/article/pii/S0924271617301880 doi: https://doi.org/10.1016/j.isprsjprs.2017.10.012

Fungtion. (n.d.). *Fungtion/dann: Pytorch implementation of domain-adversarial training of neural networks.* Retrieved from https://github.com/fungtion/DANN

Ganin, Y., & Lempitsky, V. (2015). *Unsupervised domain adaptation by backpropagation.*

GISGeography. (2022). *Sentinel 2 bands and combinations.* Retrieved from https://gisgeography.com/sentinel-2-bands-combinations/

Godoy, D. (2022, Jul). *Understanding binary cross-entropy / log loss: A visual explanation.* Towards Data Science. Retrieved from https://towardsdatascience.com/understanding-binary-cross-entropy-log-loss-a-visual

-explanation-a3ac6025181a

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... Bengio, Y. (2014). *Generative adversarial networks.*

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., & Moore, R. (2017). Google earth engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment*, *202*, 18-27. Retrieved from https://www.sciencedirect.com/science/article/pii/S0034425717302900 (Big Remotely Sensed Data: tools, applications and experiences) doi: https://doi.org/10.1016/j.rse.2017.06.031

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., & Smola, A. (2012). A kernel two-sample test. *Journal of Machine Learning Research*, *13*(25), 728-729. Retrieved from http://jmlr.org/papers/v13/gretton12a.html

Hafner, S. (2022, May). *Multi-modal deep learning with sentinel-1 and sentinel-2 data for urban mapping and change detection.* KTH Royal Institute of Technology. Retrieved from https://www.diva-portal.org/smash/get/diva2:1661349/FULLTEXT02.pdf

Hafner, S., Ban, Y., & Nascetti, A. (2022a). *Sen12 global urban mapping dataset.* Zenodo. Retrieved from https://doi.org/10.5281/zenodo.6914898 doi: 10.5281/zenodo.6914898

Hafner, S., Ban, Y., & Nascetti, A. (2022b). Unsupervised domain adaptation for global urban extraction using sentinel-1 sar and sentinel-2 msi data. *Remote Sensing of Environment*, *280*, 113192. Retrieved from https://www.sciencedirect.com/science/article/pii/S0034425722003029 doi: https://doi.org/10.1016/j.rse.2022.113192

Hafner, S., Nascetti, A., Azizpour, H., & Ban, Y. (2022). Sentinel-1 and sentinel-2 data fusion for urban change detection using a dual stream u-net. *IEEE Geoscience and Remote Sensing Letters*, *19*, 1-5. doi: 10.1109/LGRS.2021.3119856

Li, Y., Zhang, S., Li, X., & Ye, F. (2023). Remote sensing image classification with few labeled data using semisupervised learning. *Wireless Communications and Mobile Computing.*

Microsoft. (2023). *Global microsoft ml building footprints.* GitHub. Retrieved from https://github.com/microsoft/GlobalMLBuildingFootprints

Oliver, A., Odena, A., Raffel, C. A., Cubuk, E. D., & Goodfellow, I. (2018). Realistic evaluation of deep semi-supervised learning algorithms. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 31). Curran Associates, Inc. Retrieved from https://proceedings.neurips.cc/paper_files/paper/2018/file/c1fea270c48e8079d8ddf7d06d26ab52-Paper.pdf

O'Sullivan, C. (2023). *U-net explained: Understanding its image segmentation architecture.* Towards Data Science. Retrieved from https://towardsdatascience

.com/u-net-explained-understanding-its-image-segmentation-architecture-56e4842e313a

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022, June). High-resolution image synthesis with latent diffusion models. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition (cvpr)* (p. 10684-10695).

Ronneberger, O., Fischer, P., & Brox, T. (2015). *U-net: Convolutional networks for biomedical image segmentation.*

Seif, G. (2022, Feb). *Understanding the 3 most common loss functions for machine learning regression.* Towards Data Science. Retrieved from https://towardsdatascience.com/understanding-the-3-most-common-loss-functions-for-machine-learning-regression-23e0ef3e14d3

Sentinel-Hub. (2023). *Sentinel hub's cloud detector for sentinel-2 imagery.* GitHub. Retrieved from https://github.com/sentinel-hub/sentinel2-cloud-detector

Souly, N., Spampinato, C., & Shah, M. (2017). *Semi and weakly supervised semantic segmentation using generative adversarial network.*

Steins. (2023, Jun). *Stable diffusion clearly explained!* Medium. Retrieved from https://medium.com/@steinsfu/stable-diffusion-clearly-explained-ed008044e07e#6f06

van Beers, F., Lindström, A., Okafor, E., & Wiering, M. (2019, 02). Deep neural networks with intersection over union loss for binary image segmentation.. doi: 10.5220/0007347504380445

Van Etten, A., Hogan, D., Manso, J. M., Shermeyer, J., Weir, N., & Lewis, R. (2021, June). The multi-temporal urban development spacenet dataset. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition (cvpr)* (p. 6398-6407).

Vint, D., Anderson, M., Yang, Y., Ilioudis, C., Di Caterina, G., & Clemente, C. (2021, 02). Automatic target recognition for low resolution foliage penetrating sar images using cnns and gans. *Remote Sensing*, *13*, 596. doi: 10.3390/rs13040596

Williams, B. (2023, Mar). *Contrastive loss explained.* Towards Data Science. Retrieved from https://towardsdatascience.com/contrastive-loss-explaned-159f2d4a87ec

Wong, A. (2023, Jun). *How does stable diffusion work?* Retrieved from https://stable-diffusion-art.com/how-stable-diffusion-work/

ETH

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Declaration of originality

The signed declaration of originality is a component of every semester paper, Bachelor's thesis, Master's thesis and any other degree paper undertaken during the course of studies, including the respective electronic versions.

Lecturers may also require a declaration of originality for other written papers compiled for their courses.

I hereby confirm that I am the sole author of the written work here enclosed and that I have compiled it in my own words. Parts excepted are corrections of form and content by the supervisor.

Title of work (in block letters):

A Shared Deep Feature Embedding of Sentinel-1 and Sentinel-2 for Building Detection

Authored by (in block letters):

For papers written by groups the names of all authors are required.

Name(s):	**First name(s):**
Rüegg	Arno

With my signature I confirm that

- I have committed none of the forms of plagiarism described in the 'Citation etiquette' information sheet.
- I have documented all methods, data and processes truthfully.
- I have not manipulated any data.
- I have mentioned all persons who were significant facilitators of the work.

I am aware that the work may be screened electronically for plagiarism.

Place, date	**Signature(s)**
03.07.2023	

For papers written by groups the names of all authors are required. Their signatures collectively guarantee the entire content of the written paper.