Evaluation of Semi-supervised Semantic Segmentation for Remote Sensing, Medical Imaging, and Machine Vision Settings

Steven Landgraf *, Johannes Huber, Markus Hillemann, Markus Ulrich

Institute of Photogrammetry and Remote Sensing (IPF), Karlsruhe Institute of Technology (KIT), Germany -
(steven.landgraf, markus.hillemann, markus.ulrich)@kit.edu
johannes.huber@student.kit.edu

Keywords: Semi-supervised Learning, Semantic Segmentation, Remote Sensing, Medical Imaging, Machine Vision

Abstract

Semi-supervised semantic segmentation (S^4) has garnered significant attention in recent years due to the time-consuming and costly process of creating pixel-level annotations. Instead of only relying on labeled data, semi-supervised approaches leverage both labeled and unlabeled data to mitigate the issue of the labor-intense annotation process. Although current state-of-the-art methods in S^4 achieve impressive results, they are often only evaluated in specific domains, which are not fully representative of many real-world applications. For this reason, we evaluate the foundational Mean Teacher approach together with UniMatch, one of the current state-of-the-art methods, on multiple datasets spanning remote sensing, medical imaging, and machine vision settings. Our results demonstrate that semi-supervised approaches are able to achieve significant performance gains in label-scarce environments and even surpass the fully supervised baseline with 100% of the labels in the machine vision setting.

1. Introduction

Deep learning revolutionized our ability to deal with high-dimensional data, dramatically improving the state-of-the-art in natural language processing, genomics, and computer vision tasks like image classification, object detection, or semantic segmentation (LeCun et al., 2015). However, obtaining the large amount of labeled data required to train deep neural networks is particularly time-consuming and costly for semantic segmentation tasks, as they demand labor-intensive pixel-level annotations. To mitigate this issue, an ever-growing amount of researchers developed semi-supervised semantic segmentation (S^4) methods, exploiting both labeled and unlabeled images to improve performance without the need for exhaustive manual annotations (Peláez-Vegas et al., 2023).

Despite the rapid progress, current state-of-the-art methods (Zhao et al., 2023; Sun et al., 2024; Mai et al., 2024; Wang et al., 2024a,b; Hu et al., 2024) in S^4 are often only evaluated in specific domains, primarily PASCAL VOC (Everingham et al., 2010) and Cityscapes (Cordts et al., 2016). In particular, PASCAL VOC consists mainly of object-centered images, whereas Cityscapes only contains street scenes. Although they are widely used as benchmark datasets, they are not fully representative for many more general real-world applications and conditions. Additionally, many evaluations still rely on outdated or small architectures, limiting the generalizability and comparability of their findings to more modern, large-scale networks.

To address the aforementioned limitations, we conduct a comprehensive comparison between two key methods: the foundational Mean Teacher approach (Tarvainen and Valpola, 2017) and UniMatch (Yang et al., 2023). We choose Mean Teacher for its widespread use in semi-supervised learning, and because many state-of-the-art approaches still use it as their methodological foundation. UniMatch, on the other hand, is a leading state-of-the-art method recognized for its strong performance and usability across diverse domains. We evaluate four datasets in three distinct application domains: Two remote sensing, one medical imaging, and one machine vision dataset. Notably, we are the first to apply S^4 to the T-LESS dataset (Hodan et al., 2017), representing the machine vision setting with industry-relevant, texture-less objects.

*Corresponding author

2. Semi-supervised Semantic Segmentation

Based on the taxonomy proposed by Peláez-Vegas et al. (2023), most existing S^4 methods can be classified into five categories: adversarial methods, consistency regularization, pseudo-labeling, contrastive learning, and hybrid methods.

Adversarial methods. Adversarial methods for S^4 can be divided into two groups based on the use of generative models during training. The first group includes approaches that incorporate a generative model to create synthetic images, which are then used as additional input for the segmentation task (Souly et al., 2017; Li et al., 2021). The second group, in contrast, does not use a separate generative model. Instead, these methods employ a GAN-like structure where the segmentation network itself acts as the generator, and the discriminator distinguishes between the predicted segmentation maps and the real ground truth maps (Hung et al., 2018; Mittal et al., 2019; Mendel et al., 2020; Ke et al., 2020; Zhang et al., 2021b; Jin et al., 2021).

Consistency regularization. Consistency regularization-based semi-supervised learning methods leverage unlabeled data by applying perturbations and training models to remain invariant to these changes. This is typically accomplished by introducing an additional regularization term in the loss function, which measures the distance between the original and perturbed predictions. Mean Teacher (Tarvainen and Valpola, 2017) is a foundational approach, which enforces consistency between a student and a teacher model. The primary distinction among consistency regularization methods for S^4 is how perturbations

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-G-2025-527-2025 | © Author(s) 2025. CC BY 4.0 License.

are applied. Following the taxonomy of Peláez-Vegas et al. (2023), these methods can be further categorized into four subgroups. The first group applies perturbations directly to the input images, requiring the model to predict the same label for both the original and the augmented input (French et al., 2019; Olsson et al., 2021; Chen et al., 2021b; Zhao et al., 2023). While the second group focuses on feature perturbations (Ouali et al., 2020), the third category generates perturbed predictions by using auxiliary models (Chen et al., 2021a; Peng et al., 2020). Lastly, there are hybrid approaches (Liu et al., 2022; Wu et al., 2023; Yang et al., 2023) that combine multiple types of perturbations from the previous categories.

Pseudo-labeling. Pseudo-labeling methods, sometimes also referred to as bootstrapping, are fundamental to S^4. These methods generate pseudo-labels of unlabeled images by using predictions from a model that was pre-trained on labeled data. Conventionally, these pseudo-labels are then added to the labeled dataset, and a new model is trained on the expanded dataset. Peláez-Vegas et al. (2023) further categorize these pseudo-labeling methods into self-training and mutual-training methods. Self-training methods (Yang et al., 2022; Teh et al., 2022; Zhu et al., 2021; Chen et al., 2020c; Yuan et al., 2021; He et al., 2021; Sun et al., 2024; Wang et al., 2024a) generate pseudo-labels mostly based on high-confidence predictions of a single supervised model, whereas mutual-training methods (Feng et al., 2022; Zhou et al., 2022; Li et al., 2023a,b; Na et al., 2024) involve multiple models.

Contrastive learning. Contrastive learning methods aim to organize the feature space by grouping similar samples together and pushing dissimilar samples away. Inspired by the groundbreaking success in self-supervised learning by methods like SimCLR (Chen et al., 2020a,b), a series of S^4 approaches have been proposed (Liu et al., 2021; Chen and He, 2021; Alonso et al., 2021).

Hybrid methods. Naturally, there is also a large variety of hybrid methods (Qiao et al., 2023; Wang et al., 2023b,c; Ma et al., 2023; Liang et al., 2023; Wang et al., 2023a; Hu et al., 2024; Mai et al., 2024; Wang et al., 2024b; Yang et al., 2023) that fuse characteristics from the other categories, particularly pseudo-labeling and consistency regularization.

3. Methodological Background

In this section, we outline the two methods employed in this study: Mean Teacher (Tarvainen and Valpola, 2017) and UniMatch (Yang et al., 2023). Mean Teacher serves as the semi-supervised baseline due to its widespread use and foundational role for many modern methods. UniMatch represents a leading state-of-the-art approach recognized for its performance and generalizability across diverse settings.

3.1 Mean Teacher

The Mean Teacher approach (Tarvainen and Valpola, 2017) is a foundational method for semi-supervised learning that enforces consistency between a student and a teacher model. Conventionally, both the student and teacher networks share the same architecture.

The central idea is to train the student network using standard backpropagation and gradient descent, while the teacher network's weights θ'_t are updated as an exponential moving average (EMA) of the student's weights:

$$\theta'_t = \alpha\theta'_{t-1} + (1-\alpha)\theta_t \ . \tag{1}$$

Here, t is the current training step, θ_t represents the student's weights at step t, and α is a smoothing coefficient that controls the rate of the EMA update. By applying this EMA-based weight update, the teacher model maintains a more stable representation over time, providing more consistent pseudo-labels, which in turn regularize the student model during training.

The student network is trained with

$$\mathcal{L}_{\text{MeanTeacher}} = \mathcal{L}_s + \lambda\mathcal{L}_c \ , \tag{2}$$

where $\mathcal{L}_s$ represents the regular cross-entropy loss for labeled data with

$$\mathcal{L}_s = -\frac{1}{N}\sum_{n=1}^{N}\sum_{c=1}^{C} y_{n,c}\cdot\log(p(z)_{n,c}) \ , \tag{3}$$

where N is the number of pixels in an image, C is the number of classes, $y_{n,c}$ is the corresponding ground truth label, and $p(z)_{n,c}$ is the predicted softmax probability for class c.

$\mathcal{L}_c$ is a consistency loss between the student and the teacher's predictions using the unlabeled data, weighted by λ. It encourages the student to produce predictions similar to the teacher's pseudo-labels. We follow (Chen et al., 2021a) and employ the common mean squared error for $\mathcal{L}_c$.

3.2 UniMatch

UniMatch (Yang et al., 2023) is a hybrid method based on FixMatch (Sohn et al., 2020), which combines consistency regularization with pseudo-labeling. This approach still remains as one of the state-of-the-art methods for S^4.

The central idea of UniMatch is to enforce consistency between dual perturbations and feature perturbations, while pseudo-labels are generated from the model's own predictions on unlabeled data. Weak perturbations include transformations like resizing, cropping, and flipping, while strong perturbations apply more aggressive transformations such as color jitter and CutMix (Yun et al., 2019). The pseudo-labels are treated as ground truth if their softmax confidence exceeds a certain threshold, ensuring that only high-quality pseudo-labels contribute to training.

The total loss function of UniMatch is defined as the average of the supervised and unsupervised losses:

$$\mathcal{L}_{\text{UniMatch}} = \frac{1}{2}\left(\mathcal{L}_s + \mathcal{L}_u\right) \ . \tag{4}$$

For unlabeled images, the unsupervised loss $\mathcal{L}_u$ is applied only if the softmax confidence of the weakly augmented prediction exceeds a pre-defined threshold τ. This ensures that the model only learns from pseudo-labels that are reliable enough:

$$\mathcal{L}_u = \mathbb{1}\left(max(p^{\text{w}}) \geq \tau\right)\cdot \left(\lambda\mathcal{L}_s(p^{\text{w}}, p^{\text{fp}}) + \frac{\mu}{2}\left(\mathcal{L}_s(p^{\text{w}}, p^{\text{s}_1}) + \mathcal{L}_s(p^{\text{w}}, p^{\text{s}_2})\right)\right) \ , \tag{5}$$

where p^{w} represents the weakly augmented softmax probability, p^{fp} is the feature-perturbed softmax prediction using a dropout layer (Srivastava et al., 2014), and p^{s_1}, p^{s_2} are the softmax probabilities from two strongly augmented input images. The unsupervised loss $\mathcal{L}u$ is composed of two parts: a consistency term between weak and feature-perturbed predictions, and a consistency term between weak and strongly augmented predictions, ensuring that the model's predictions remain consistent under different perturbations. $\mathcal{L}_s$ denotes the regular cross-entropy loss with p^{w} as the pseudo-label.

4. Experimental Setup

Hereinafter, we describe our experimental setup. We provide details on all four datasets and on implementation details to ensure reproducibility.

4.1 Datasets

Table 1 provides an overview of the four datasets, each presenting unique challenges and characteristics.

Remote Sensing. The WHU-CD (Ji et al., 2018) and LEVIR-CD (Chen and Shi, 2020) datasets consist of aerial or satellite images taken years apart to monitor urban changes. WHU-CD captures building changes in post-earthquake Christchurch, New Zealand, while LEVIR-CD monitors urban development in Texas, USA. Both datasets are commonly used for urban change detection.

Medical Imaging. The ACDC dataset (Bernard et al., 2018) comprises 200 MRI frames that consist of a series of sequential slices, which together form a 3D representation of the heart. The task is to segment the left ventricle, myocardium, and right ventricle. Following Luo et al. (2022), the dataset is divided into 140 frames from 70 patients for training and 60 frames from 30 patients for testing.

Machine Vision. The T-LESS dataset (Hodan et al., 2017) is an RGB-D dataset for 6D pose estimation of textureless objects and is part of the BOP Challenge (Hodan et al., 2024). It consists of 30 objects without significant texture, color, or reflection features, often similar in shape and size, characteristics typical of industrial settings.

4.2 Implementation Details

Architecture. For all experiments, we adopt DeepLabv3+ (Chen et al., 2018) with a ResNet-101 backbone (He et al., 2016), ensuring competitive performance across diverse domains. This well-established architecture not only guarantees robust results but also facilitates future comparisons, as its widespread use allows for consistent benchmarking across studies.

Training. In terms of training, we largely follow the suggestions of Mean Teacher (Tarvainen and Valpola, 2017) and UniMatch (Yang et al., 2023) in terms of data augmentations and initial hyperparameters. We use a Stochastic Gradient Descent optimizer with a momentum of 0.9, and weight decay of 0.0001 as optimizer-specific hyperparameters. Dataset-specific hyperparameters can be found in Table 2. Additionally, we employ a polynomial learning rate scheduler:

$$lr = lr_{\mathrm{base}} \cdot (1 - \frac{\mathrm{iteration}}{\mathrm{total\ iterations}})^{0.9} \quad , \tag{6}$$

where lr is the current learning rate and lr_{base} is the initial base learning rate.

Metrics. For the remote sensing datasets, WHU-CD and LEVIR-CD, we report the changed-class Intersection over Union. In contrast, for the medical imaging dataset, ACDC, we follow previous work (Tarvainen and Valpola, 2017; Yang et al., 2023) and report the Mean Dice. Finally, we report the mean Intersection over Union, also known as the Jaccard Index, for the machine vision setting with the T-LESS dataset.

5. Results

In this section, we present the evaluation results. First, we provide quantitative comparisons of Mean Teacher and UniMatch on WHU-CD, LEVIR-CD, ACDC, and T-LESS, benchmarking them against a fully supervised baseline. The supervised baseline uses the same architecture and hyperparameters but is trained solely with the supervised loss, as defined in Equation 3. Next, we present qualitative results of UniMatch for all four datasets.

For the training splits, we follow the methodology of Bandara and Patel (2022) for WHU-CD and LEVIR-CD. For ACDC, we adopt the approach used by Yang et al. (2023), utilizing images from one, three, or seven patients, corresponding to approximately 2.5%, 5%, and 10% of the dataset, respectively. As this is the first evaluation of S^4 on T-LESS, we use the training splits from MS COCO (Lin et al., 2014), given its comparable dataset size.

5.1 Quantitative Evaluation

WHU-CD. As shown by Table 3, utilizing the ResNet-101 (RN-101) backbone consistently improves performance over the smaller ResNet-50 (RN-50). Reducing the number of available labels from 40% to 5% significantly deteriorates performance for the supervised baseline (80.6% to 60.8%). Mean Teacher shows notable improvements but also struggles with only 5% of labels (64.0%). In contrast, UniMatch consistently performs well across all splits, achieving 80.9% to 86.6%.

LEVIR-CD. Table 3 reveals that the larger RN-101 backbone offers minimal improvement over RN-50. The performance gap between Mean Teacher and UniMatch is also smaller. Mean Teacher achieves 78.7% to 81.9%, while UniMatch slightly outperforms it, ranging from 80.7% to 82.7%. Notably, with the 40% label split, the difference between the supervised baseline and UniMatch is only 2.3%.

ACDC. Based on the results highlighted in Table 4, swapping the U-Net architecture with a DeepLabv3+ using the RN-101 backbone results in significant improvements for the supervised baseline (from 28.5% to 62.5% using U-Net, to 50.1% to 86.5% with DeepLabv3+). Similar, albeit less striking, improvements are observed for the Mean Teacher approach. UniMatch's gains are more modest, improving from 85.4% to 89.9% (compared to 86.7% to 90.7% with the updated architecture). However, UniMatch's performance is particularly strong in the 1-case scenario, where only labeled images of a single patient were used, showing only a 4.0% drop in performance compared to a 36.4% decrease for the supervised baseline.

T-LESS. As Table 5 shows, for T-LESS, the supervised baseline yields poor results with smaller fractions of labeled

	Domain	Resolution (W × H)	Training Images	Test Images
WHU-CD (Ji et al., 2018)	Remote Sensing	256 × 256	5947	743
LEVIR-CD (Chen and Shi, 2020)	Remote Sensing	256 × 256	7120	1024
ACDC (Bernard et al., 2018)	Medical Imaging	256 × 256	1312	610
T-LESS (Hodan et al., 2017)	Machine Vision	720 × 540	87584 (synth. + real)	10080 (real)

Table 1. Overview of the datasets that we used for evaluation.

	Batch Size	Learning Rate	Epochs	Crop Size
WHU-CD (Ji et al., 2018)	8	0.02	80	256 × 256
LEVIR-CD (Chen and Shi, 2020)	8	0.02	80	256 × 256
ACDC (Bernard et al., 2018)	8	0.25	400	256 × 256
T-LESS (Hodan et al., 2017)	8	0.005	30	512 × 512

Table 2. Overview of the hyperparameters used for training.

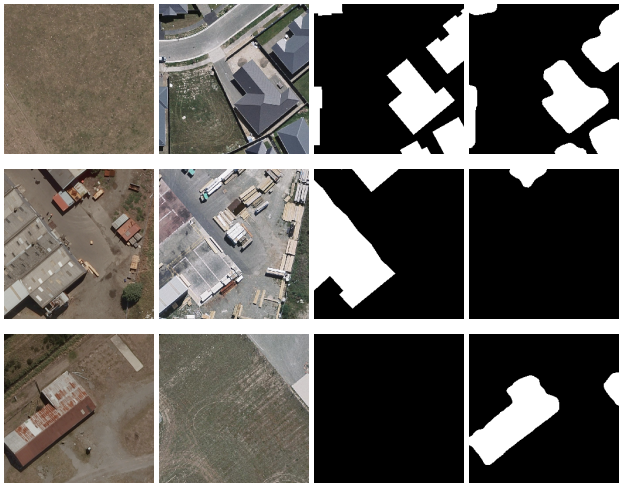


Figure 1. Qualitative examples of UniMatch (20% split) on WHU-CD. From left to right: image A, image B, ground truth, and prediction.

data, achieving only 19.5% with the 1/512 split. In comparison, when training with all available data, 65.6% are attained. Mean Teacher demonstrates considerable improvements across all splits, ranging from 38.5% to 67.5%. However, UniMatch outperforms both the baseline and Mean Teacher, with results spanning from 55.6% to 75.9%. Notably, UniMatch surpasses the fully supervised baseline with only 685 labels (1/128 split) already, achieving an impressive 71.9%. These striking findings will be discussed further in Section 6.

5.2 Qualitative Evaluation

Figures 1, 2, and 4 provide qualitative examples from the WHU-CD, LEVIR-CD, and T-LESS datasets, respectively, showcasing the performance of the UniMatch models. In each figure, the first row represents a positive example where the predictions are nearly indistinguishable from the ground truth, demonstrating the model's ability to accurately solve the task. In contrast, the second row illustrates cases where the model struggles, either missing an entire building for WHU-CD and LEVIR-CD or producing noisy and incomplete segmentation masks for T-LESS. Notably, the third row highlights instances where the ground truth appears to be inaccurate — missing changes in WHU-CD and LEVIR-CD or overlapping object boundaries in T-LESS — yet the model's predictions seem more plausible and closer to the expected outcomes.

Similarly, Figure 3 presents qualitative examples from the

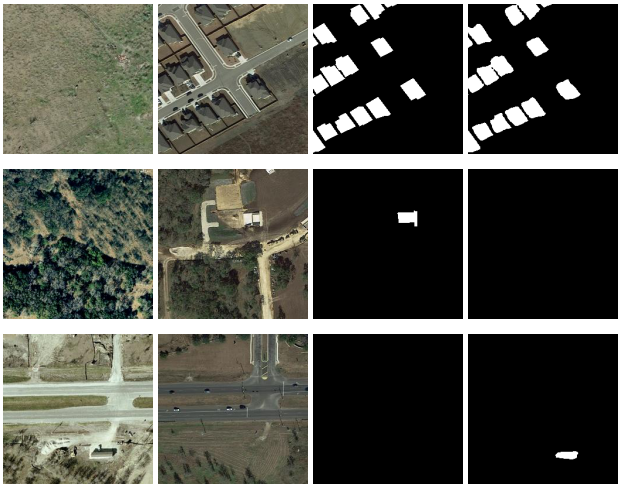


Figure 2. Qualitative examples of UniMatch (20% split) on LEVIR-CD. From left to right: image A, image B, ground truth, and prediction.

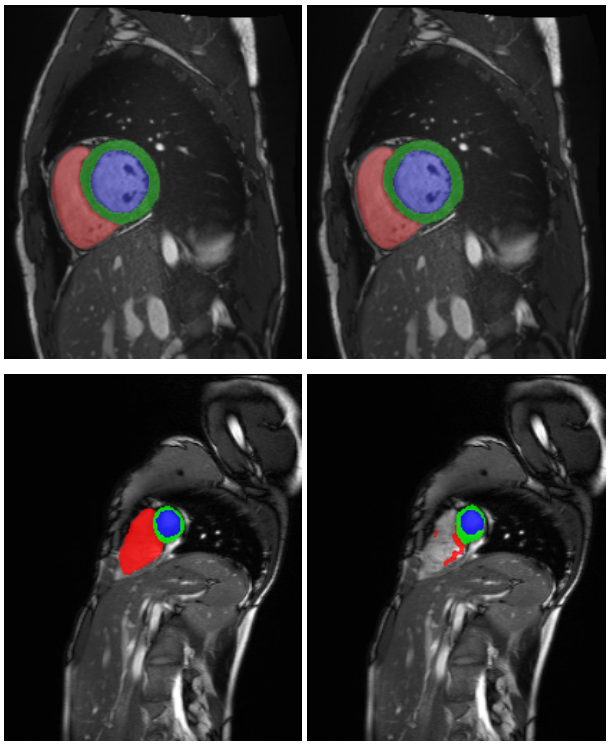


Figure 3. Qualitative examples of UniMatch (7 cases) on ACDC. The left column shows the input image with the ground truth overlaid, while the right column shows the corresponding predictions.

ACDC dataset, where the first row shows near-perfect segmentation across all three classes, while the second row illustrates

	Remote Sensing	WHU-CD				LEVIR-CD			
		5% (297)	10% (594)	20% (1189)	40% (2378)	5% (356)	10% (712)	20% (1424)	40% (2848)
RN-50	SupBaseline (Yang et al., 2023)	54.1	60.9	68.4	76.2	69.3	76.0	77.6	80.5
	SemiCD (Bandara and Patel, 2022)	65.8	68.1	74.8	77.2	72.5	75.5	76.2	77.2
	UniMatch (Yang et al., 2023)	80.2	81.7	81.7	85.1	**80.7**	82.0	81.7	82.1
RN-101	SupBaseline	60.8	59.6	66.4	80.6	69.7	75.1	78.0	80.4
	Mean Teacher	64.0	73.9	80.6	84.4	78.7	80.6	81.6	81.9
	UniMatch	**80.9**	**85.5**	**85.6**	**86.6**	**80.7**	**82.2**	**82.4**	**82.7**

Table 3. Quantitative comparisons of the changed-class Intersection over Union on the WHU-CD and LEVIR-CD datasets for different training splits and architectures. Best respective results are marked in **bold**.

	Medical Imaging	1 case	3 cases	7 cases
RN-50	SupBaseline (Yang et al., 2023)	28.5	41.5	62.5
	MeanTeacher (Tarvainen and Valpola, 2017)	-	56.6	81.0
	UniMatch (Yang et al., 2023)	85.4	88.9	89.9
RN-101	SupBaseline	50.1	72.1	86.5
	Mean Teacher	59.6	80.9	88.8
	UniMatch	**86.7**	**90.1**	**90.7**

Table 4. Quantitative comparisons of Mean Dice on the ACDC dataset for different training splits and architectures. Best respective results are marked in **bold**.

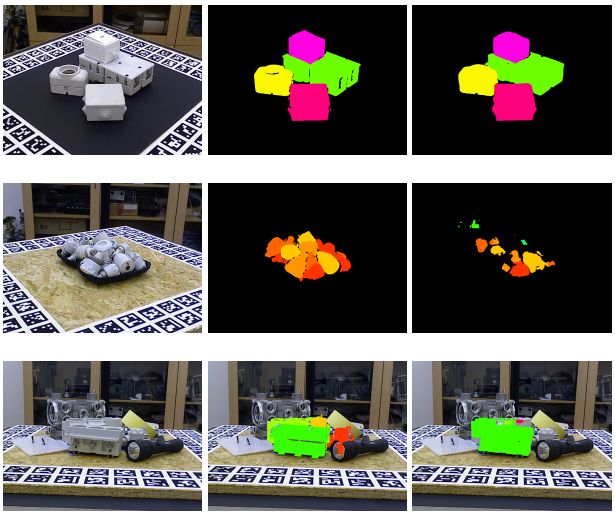


Figure 4. Qualitative examples of UniMatch (1/4 split) on T-LESS. From left to right: input image, ground truth, and prediction.

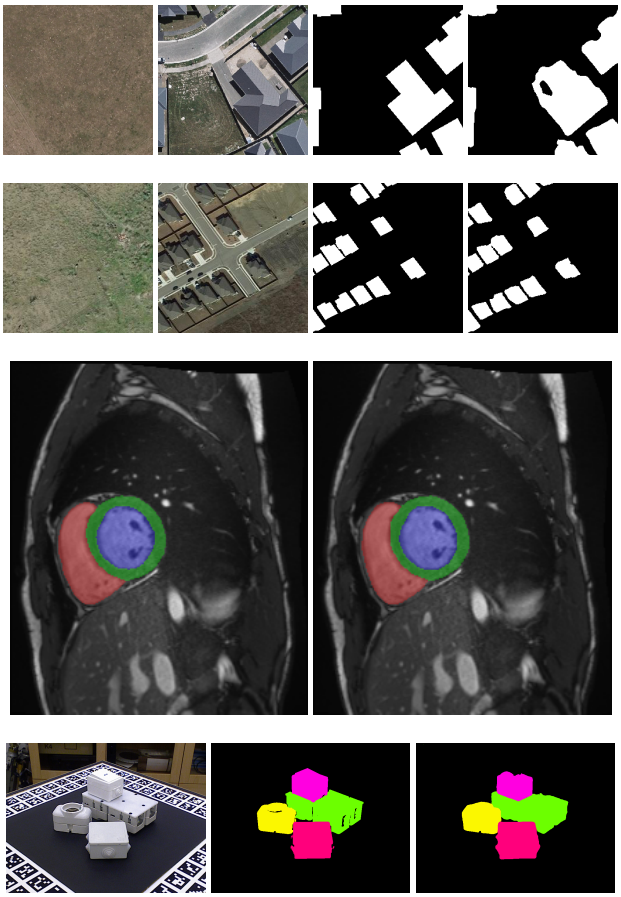


Figure 5. Qualitative examples of Mean Teacher across all domains. From top to bottom: WHU-CD (20% split), LEVIR-CD (20% split), ACDC (7 cases), and T-LESS (1/4 split).

the model's difficulty in detecting one class, indicating its limitations in more challenging scenarios.

Additionally, Figure 5 provides qualitative examples of Mean Teacher for the positive samples from Figures 1-4. While the results from Mean Teacher are generally satisfactory, they fall short of matching the performance of UniMatch, supporting the quantitative comparisons.

6. Discussion

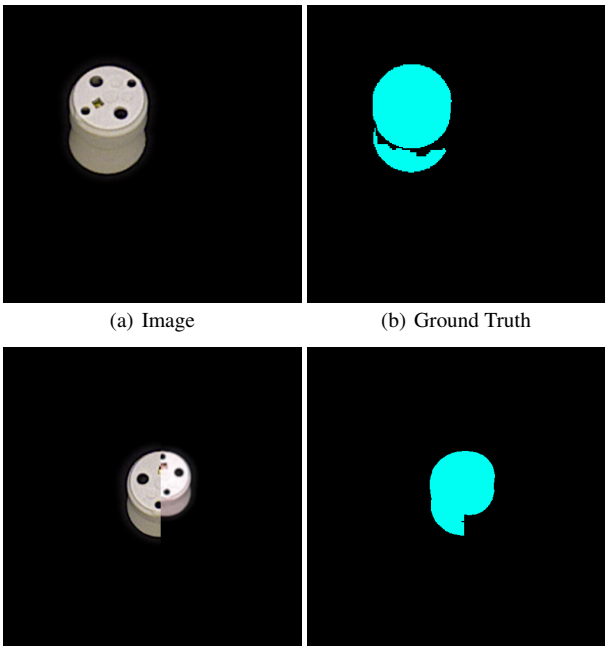
The results demonstrate that semi-supervised approaches consistently outperform fully supervised models, particularly in label-scarce scenarios. UniMatch not only shows greater resilience to reductions in labeled data than Mean Teacher but also delivers the best results across all experiments.

A key observation of our study is that UniMatch even surpasses the fully supervised baseline with 100% of the labels in several T-LESS splits. This may seem counterintuitive, but as shown in Figure 6, the training labels often fail to accurately capture object boundaries, while the pseudo-labels generated by UniMatch provide better results. Fully supervised models likely overfit on these noisy or inaccurate labels, which weakens their generalization, especially in the presence of the domain gap between synthetic training and real test images. These findings are supported by the work of Zhang et al. (2021a), which highlights how large models can overfit noisy training data. Similarly, French et al. (2019) stress the importance of strong augmentations in semi-supervised segmentation, a plausible factor in UniMatch's superiority over Mean Teacher, which relies mainly on consistency regularization. Finally, recent work on foundation models in monocular depth estimation (Yang et al., 2024) emphasizes the benefits of strong perturbations on unlabeled data, further supporting UniMatch's robust performance.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-G-2025-527-2025 | © Author(s) 2025. CC BY 4.0 License.

Machine Vision	1/512 (172)	1/256 (342)	1/128 (685)	1/64 (1370)	1/32 (2738)	1/16 (5474)	1/8 (10948)	1/4 (21896)	1/2 (43792)	1 (87584)
SupBaseline	19.4	31.0	39.1	47.3	55.7	52.3	55.1	54.4	59.9	**65.6**
MeanTeacher	38.5	47.3	56.4	58.5	61.4	62.9	64.2	65.0	67.5	-
UniMatch	**55.6**	**61.8**	**71.9**	**74.0**	**75.3**	**74.3**	**75.3**	**75.9**	**73.0**	-

Table 5. Quantitative comparisons of the mean Intersection over Union on the T-LESS dataset for different training splits, using a DeepLabv3+ architecture with a ResNet-101 backbone. Best respective results are marked in **bold**.



(a) Image (b) Ground Truth

(c) Strongly Perturbed Image (d) Pseudo-label

Figure 6. Representative example of UniMatch pseudo-label of strongly perturbed image after just 5 training epochs in comparison to the original GT.

7. Conclusion

In this work, we demonstrated that semi-supervised approaches are able to provide a significant performance gain over fully supervised models for semantic segmentation in label-scarce environments across remote sensing, medical imaging, and machine vision settings. UniMatch, in particular, not only proved remarkably resilient to reductions in labeled data but also outperformed the fully supervised baseline on T-LESS with just 1/128 of the labels. It shows that by leveraging more accurate pseudo-labels and strong perturbations, the model is less likely to overfit to noisy labels, which is a common limitation of traditional supervised methods. These results confirm the value of applying semi-supervised techniques across diverse domains, emphasizing their potential for real-world applications.

For future work, leveraging pseudo-labels from large-scale teacher models, as suggested by Yang et al. (2024), could help overcome the limitations of synthetic images, such as domain gaps and inaccurate labels, while enhancing their advantages. Additionally, incorporating uncertainty estimation, which has proven useful in semantic segmentation (Landgraf et al., 2024a,b,c), could be explored within semi-supervised pipelines for further improvements in label-scarce environments where creating pixel-level annotations is time-consuming and costly.

References

Alonso, I., Sabater, A., Ferstl, D., Montesano, L., Murillo, A. C., 2021. Semi-supervised semantic segmentation with pixel-level contrastive learning from a class-wise memory bank. *Proceedings of the IEEE/CVF international conference on computer vision*, 8219–8228.

Bandara, W. G. C., Patel, V. M., 2022. Revisiting consistency regularization for semi-supervised change detection in remote sensing images. *arXiv preprint arXiv:2204.08454*.

Bernard, O., Lalande, A., Zotti, C., Cervenansky, F., Yang, X., Heng, P.-A., Cetin, I., Lekadir, K., Camara, O., Ballester, M. A. G. et al., 2018. Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved? *IEEE transactions on medical imaging*, 37(11), 2514–2525.

Chen, H., Shi, Z., 2020. A spatial-temporal attention-based method and a new dataset for remote sensing image change detection. *Remote Sensing*, 12(10), 1662.

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H., 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. *Proceedings of the European conference on computer vision (ECCV)*, 801–818.

Chen, T., Kornblith, S., Norouzi, M., Hinton, G., 2020a. A simple framework for contrastive learning of visual representations. *International conference on machine learning*, PMLR, 1597–1607.

Chen, T., Kornblith, S., Swersky, K., Norouzi, M., Hinton, G. E., 2020b. Big self-supervised models are strong semi-supervised learners. *Advances in neural information processing systems*, 33, 22243–22255.

Chen, X., He, K., 2021. Exploring simple siamese representation learning. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 15750–15758.

Chen, X., Yuan, Y., Zeng, G., Wang, J., 2021a. Semi-supervised semantic segmentation with cross pseudo supervision. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2613–2622.

Chen, Y., Ouyang, X., Zhu, K., Agam, G., 2021b. Complexmix: Semi-supervised semantic segmentation via mask-based data augmentation. *2021 IEEE International Conference on Image Processing (ICIP)*, IEEE, 2264–2268.

Chen, Z., Zhang, R., Zhang, G., Ma, Z., Lei, T., 2020c. Digging into pseudo label: a low-budget approach for semi-supervised semantic segmentation. *IEEE Access*, 8, 41830–41837.

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B., 2016. The cityscapes dataset for semantic urban scene understanding. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3213–3223.

Everingham, M., Van Gool, L., Williams, C. K., Winn, J., Zisserman, A., 2010. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88, 303–338.

Feng, Z., Zhou, Q., Gu, Q., Tan, X., Cheng, G., Lu, X., Shi, J., Ma, L., 2022. Dmt: Dynamic mutual training for semi-supervised learning. *Pattern Recognition*, 130, 108777.

French, G., Laine, S., Aila, T., Mackiewicz, M., Finlayson, G., 2019. Semi-supervised semantic segmentation needs strong, varied perturbations. *arXiv preprint arXiv:1906.01916*.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

He, R., Yang, J., Qi, X., 2021. Re-distributing biased pseudo labels for semi-supervised semantic segmentation: A baseline investigation. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6930–6940.

Hodan, T., Haluza, P., Obdržálek, Š., Matas, J., Lourakis, M., Zabulis, X., 2017. T-less: An rgb-d dataset for 6d pose estimation of texture-less objects. *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, IEEE, 880–888.

Hodan, T., Sundermeyer, M., Labbe, Y., Nguyen, V. N., Wang, G., Brachmann, E., Drost, B., Lepetit, V., Rother, C., Matas, J., 2024. Bop challenge 2023 on detection segmentation and pose estimation of seen and unseen rigid objects. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5610–5619.

Hu, X., Jiang, L., Schiele, B., 2024. Training vision transformers for semi-supervised semantic segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4007–4017.

Hung, W.-C., Tsai, Y.-H., Liou, Y.-T., Lin, Y.-Y., Yang, M.-H., 2018. Adversarial learning for semi-supervised semantic segmentation. *arXiv preprint arXiv:1802.07934*.

Ji, S., Wei, S., Lu, M., 2018. Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set. *IEEE Transactions on geoscience and remote sensing*, 57(1), 574–586.

Jin, G., Liu, C., Chen, X., 2021. Adversarial network integrating dual attention and sparse representation for semi-supervised semantic segmentation. *Information Processing & Management*, 58(5), 102680.

Ke, Z., Qiu, D., Li, K., Yan, Q., Lau, R. W., 2020. Guided collaborative training for pixel-wise semi-supervised learning. *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16*, Springer, 429–445.

Landgraf, S., Hillemann, M., Kapler, T., Ulrich, M., 2024a. Efficient Multi-task Uncertainties for Joint Semantic Segmentation and Monocular Depth Estimation. *arXiv preprint arXiv:2402.10580*.

Landgraf, S., Hillemann, M., Wursthorn, K., Ulrich, M., 2024b. Uncertainty-aware Cross-Entropy for Semantic Segmentation. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 10, 129–136.

Landgraf, S., Wursthorn, K., Hillemann, M., Ulrich, M., 2024c. Dudes: Deep uncertainty distillation using ensembles for semantic segmentation. *PFG–Journal of Photogrammetry, Remote Sensing and Geoinformation Science*, 92(2), 101–114.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. *nature*, 521(7553), 436–444.

Li, D., Yang, J., Kreis, K., Torralba, A., Fidler, S., 2021. Semantic segmentation with generative models: Semi-supervised learning and strong out-of-domain generalization. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8300–8311.

Li, P., Purkait, P., Ajanthan, T., Abdolshah, M., Garg, R., Husain, H., Xu, C., Gould, S., Ouyang, W., Van Den Hengel, A., 2023a. Semi-supervised semantic segmentation under label noise via diverse learning groups. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1229–1238.

Li, S., He, Y., Zhang, W., Zhang, W., Tan, X., Han, J., Ding, E., Wang, J., 2023b. Cfcg: Semi-supervised semantic segmentation via cross-fusion and contour guidance supervision. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 16348–16358.

Liang, C., Wang, W., Miao, J., Yang, Y., 2023. Logic-induced diagnostic reasoning for semi-supervised semantic segmentation. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 16197–16208.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C. L., 2014. Microsoft coco: Common objects in context. *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, Springer, 740–755.

Liu, S., Zhi, S., Johns, E., Davison, A. J., 2021. Bootstrapping semantic segmentation with regional contrast. *arXiv preprint arXiv:2104.04465*.

Liu, Y., Tian, Y., Chen, Y., Liu, F., Belagiannis, V., Carneiro, G., 2022. Perturbed and strict mean teachers for semi-supervised semantic segmentation. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4258–4267.

Luo, X., Hu, M., Song, T., Wang, G., Zhang, S., 2022. Semi-supervised medical image segmentation via cross teaching between cnn and transformer. *International conference on medical imaging with deep learning*, PMLR, 820–833.

Ma, J., Wang, C., Liu, Y., Lin, L., Li, G., 2023. Enhanced soft label for semi-supervised semantic segmentation. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1185–1195.

Mai, H., Sun, R., Zhang, T., Wu, F., 2024. Rankmatch: Exploring the better consistency regularization for semi-supervised semantic segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3391–3401.

Mendel, R., De Souza, L. A., Rauber, D., Papa, J. P., Palm, C., 2020. Semi-supervised segmentation based on error-correcting supervision. *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIX 16*, Springer, 141–157.

Mittal, S., Tatarchenko, M., Brox, T., 2019. Semi-supervised semantic segmentation with high-and low-level consistency. *IEEE transactions on pattern analysis and machine intelligence*, 43(4), 1369–1379.

Na, J., Ha, J.-W., Chang, H. J., Han, D., Hwang, W., 2024. Switching temporary teachers for semi-supervised semantic segmentation. *Advances in Neural Information Processing Systems*, 36.

Olsson, V., Tranheden, W., Pinto, J., Svensson, L., 2021. Classmix: Segmentation-based data augmentation for semi-supervised learning. *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 1369–1378.

Ouali, Y., Hudelot, C., Tami, M., 2020. Semi-supervised semantic segmentation with cross-consistency training. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12674–12684.

Peláez-Vegas, A., Mesejo, P., Luengo, J., 2023. A survey on semi-supervised semantic segmentation. *arXiv preprint arXiv:2302.09899*.

Peng, J., Estrada, G., Pedersoli, M., Desrosiers, C., 2020. Deep co-training for semi-supervised image segmentation. *Pattern Recognition*, 107, 107269.

Qiao, P., Wei, Z., Wang, Y., Wang, Z., Song, G., Xu, F., Ji, X., Liu, C., Chen, J., 2023. Fuzzy positive learning for semi-supervised semantic segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15465–15474.

Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C. A., Cubuk, E. D., Kurakin, A., Li, C.-L., 2020. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33, 596–608.

Souly, N., Spampinato, C., Shah, M., 2017. Semi supervised semantic segmentation using generative adversarial network. *Proceedings of the IEEE international conference on computer vision*, 5688–5696.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1), 1929–1958.

Sun, B., Yang, Y., Zhang, L., Cheng, M.-M., Hou, Q., 2024. Corrmatch: Label propagation via correlation matching for semi-supervised semantic segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3097–3107.

Tarvainen, A., Valpola, H., 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30.

Teh, E. W., DeVries, T., Duke, B., Jiang, R., Aarabi, P., Taylor, G. W., 2022. The gist and rist of iterative self-training for semi-supervised segmentation. *2022 19th Conference on Robots and Vision (CRV)*, IEEE, 58–66.

Wang, C., Xie, H., Yuan, Y., Fu, C., Yue, X., 2023a. Space engage: Collaborative space supervision for contrastive-based semi-supervised semantic segmentation. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 931–942.

Wang, H., Zhang, Q., Li, Y., Li, X., 2024a. Allspark: Reborn labeled features from unlabeled in transformer for semi-supervised semantic segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3627–3636.

Wang, X., Bai, H., Yu, L., Zhao, Y., Xiao, J., 2024b. Towards the uncharted: Density-descending feature perturbation for semi-supervised semantic segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3303–3312.

Wang, X., Zhang, B., Yu, L., Xiao, J., 2023b. Hunting sparsity: Density-guided contrastive learning for semi-supervised semantic segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3114–3123.

Wang, Z., Zhao, Z., Xing, X., Xu, D., Kong, X., Zhou, L., 2023c. Conflict-based cross-view consistency for semi-supervised semantic segmentation. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 19585–19595.

Wu, Y., Liu, C., Chen, L., Zhao, D., Zheng, Q., Zhou, H., 2023. Perturbation consistency and mutual information regularization for semi-supervised semantic segmentation. *Multimedia Systems*, 29(2), 511–523.

Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., Zhao, H., 2024. Depth Anything V2. *arXiv preprint arXiv:2406.09414*.

Yang, L., Qi, L., Feng, L., Zhang, W., Shi, Y., 2023. Revisiting weak-to-strong consistency in semi-supervised semantic segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7236–7246.

Yang, L., Zhuo, W., Qi, L., Shi, Y., Gao, Y., 2022. St++: Make self-training work better for semi-supervised semantic segmentation. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4268–4277.

Yuan, J., Liu, Y., Shen, C., Wang, Z., Li, H., 2021. A simple baseline for semi-supervised semantic segmentation with strong data augmentation. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8229–8238.

Yun, S., Han, D., Oh, S. J., Chun, S., Choe, J., Yoo, Y., 2019. Cutmix: Regularization strategy to train strong classifiers with localizable features. *Proceedings of the IEEE/CVF international conference on computer vision*, 6023–6032.

Zhang, C., Bengio, S., Hardt, M., Recht, B., Vinyals, O., 2021a. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3), 107–115.

Zhang, J., Li, Z., Zhang, C., Ma, H., 2021b. Stable self-attention adversarial learning for semi-supervised semantic image segmentation. *Journal of Visual Communication and Image Representation*, 78, 103170.

Zhao, Z., Yang, L., Long, S., Pi, J., Zhou, L., Wang, J., 2023. Augmentation matters: A simple-yet-effective approach to semi-supervised semantic segmentation. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11350–11359.

Zhou, Y., Jiao, R., Wang, D., Mu, J., Li, J., 2022. Catastrophic forgetting problem in semi-supervised semantic segmentation. *IEEE Access*, 10, 48855–48864.

Zhu, Y., Zhang, Z., Wu, C., Zhang, Z., He, T., Zhang, H., Manmatha, R., Li, M., Smola, A., 2021. Improving semantic segmentation via efficient self-training. *IEEE transactions on pattern analysis and machine intelligence*, 46(3), 1589–1602.