Extending Investigations on the Frobenius Norm Objective in SSL

Marcus Warnerfjord
warner@kth.se

Hanns Hahn
hmhahn@kth.se

Narese Michele
narese@kth.se

Abstract

This work aims to extend the Frobenius Norm Objective for self-supervised learning (SSL), presented in [8]. First, we reproduce the main results of the paper and extend them by testing different settings and approaches that were not investigated in the original work, despite being mentioned as valid alternatives. Subsequently, we evaluate the quality of the embeddings produced by performing tasks that assess their effectiveness and compare them to competitive methods. We proceed to evaluate the FroSSL objective on a different task, namely scene classification on satellite images. This is motivated by two reasons: first, as the original authors suggested, their work focuses only on object-centric datasets; second, this task is fundamentally different from simple classification as the samples are not standard images but contain spectral information. Finally, we investigate Semi-Supervised Learning (Semi-SL) for domain adaptation. We train a Semi-SL model on labeled and unlabeled data originating from two different domains, then evaluate our model on a new, unseen domain. This approach has been explored previously in [5]. Our model, which utilizes the FroSSL objective as a loss for unlabeled samples, demonstrates superior results. Our repository can be found at `https://github.com/mitchchessnoob/FroSSL`.

1 Introduction

The problem of producing meaningful representations without the need for human supervision has become increasingly relevant in machine learning, as we see an increasing amount of unlabeled data that would be wasted with standard supervised methods. The objective of self-supervised learning (SSL) is to train networks that are able to extract semantic and augmentation-invariant features.
This can mainly be done in three ways: sample-contrastive methods, asymmetric networks, and dimension-contrastive methods. The last one employs a combination of variance and invariance terms and uses only positive samples.
The objective function in [8] has been explicitly tailored in order to satisfy four desirable criteria: invariance to projection rotations, manipulating eigenvalues explicitly, scaling quadratically in batch size and dimension, and scaling linearly in views. This results in a function that is both dimension-contrastive and sample-contrastive.
The paper presents some promising results. However, its focus lies on object-centric classification. To show that the method is not limited to that kind of image, we extend it by performing scene classification. More precisely, we classify 13-channel satellite images from the EuroSAT dataset. By fine-tuning the FroSSL embedding, we created a competitive model that clearly beats the baseline. The idea of working with EuroSAT was suggested to us when we contacted one of the original authors, Oscar Skean, asking what he thought a valuable contribution would be.

Project report for DD2412 or FDD3412

Furthermore, we investigate the ability to generalize over different domains. This is a desirable quality since most models fail drastically on input that is from a different distribution than the training dataset. Using the Office31 dataset, we train the model on two domains and evaluate it on a third. We obtain excellent results that clearly beat not only the benchmark but also recent attempts by others.

2 Related Works

2.1 Scene classification on satelite data

Many people have tested various methods on the scene classification task with the EuroSAT dataset [3] and with a Benchmark accuracy of over 95% it is known as a simple classification task. However it is an interesting dataset, since its images go beyond human visible wavelengths, which especially influences the choice of augmentation methods.
To evaluate the embeddings created by the FroSSL objective, we compare them with an embedding created by the I-VNE objective [4]. This objective plays a central role in the development of FroSSL objective and has shown to be very competitive in our experiments.

2.2 Domain adaptation with Semi-SL

Traditional semi-supervised learning (Semi-SL) assumes that the feature distributions of labeled and unlabeled data are consistent which rarely holds in realistic scenarios. More recent applications, as in [5], assume a distribution shift between labeled and unlabeled data. The work proposed by the authors of [5] is a generic framework where a self-supervised task is incorporated into a Semi-Supervised Learning framework in order to extract features that better fit the distribution of unlabeled data. This enables them to improve performance when evaluating on labeled, unlabeled, and even unseen distributions.

3 Methods

3.1 FroSSL objective

A dimension contrastive loss usually consists of a term that forces the covariance to be close to the identity and a term that minimizes the distance between augmentations. This can be seen in the I-VNE objective

$$\max \mathcal{L}_{I-VNE} = \sum_{v=1}^{V} track(\Sigma_v) \ln(\Sigma_v) + \sum_{v=1}^{V-1} \sum_{r=v+1}^{V} \frac{Z_v^T Z_r}{||Z_v||_2 ||Z_r||_2}, \tag{1}$$

proposed in [4] where V is the number of views, N the batch size, D the dimension of the projection, $Z_v \in \mathbb{R}^{N \times D}$ a batch of projections and $\Sigma_v = \frac{1}{N} \hat{Z}^T \hat{Z}$ the corresponding covariance of the centered projections $\hat{Z}$.
From this objective, the authors of [8] derived the FroSSL objective by utilizing properties of the frobenius norm, that relieves computational complexity:

$$\min \mathcal{L}_{FroSSL} = \sum_{v=1}^{V} \ln(||\Sigma_v||_F^2) + \gamma ||Z_v - \overline{Z}||_F^2, \tag{2}$$

where $\overline{Z} = \frac{1}{V} \sum_{v=1}^{V} Z_v$ and γ is the invariance weight, which is usually choosen between 1 and 1.4.

3.2 FroSSL in Semi-SL for domain adaptation

In general, a Semi-Supervised loss includes a component that is specifically designed for the labeled part (usually a standard supervised loss) and a component that is able to handle unlabeled data. The loss we propose is the following

$$\mathcal{L}_{SemiSL} = \mathcal{L}_s(\mathcal{D}_l) + \lambda_u \mathcal{L}_u(\mathcal{D}_u), \tag{3}$$

where $\mathcal{L}_s$ is the Cross Entropy loss, λ_u is a hyper parameter denoting the relative weight and $\mathcal{L}_u$ is FroSSL loss (2). This Semi-supervised loss will aim towards a correct classification but, with the

help of the SSL loss acting as regularizer, the model will be pushed towards learning meaningful and most importantly domain invariant features. This will help our model to generalize well even for unseen test distributions.

4 Data

We use four datasets in our experiments: CIFAR10, STL10, EuroSAT, and OFFICE-31. CIFAR10 and STL10 are standard datasets extensively used in SSL research, and their characteristics are well-documented in prior works, including the FroSSL paper [8]. A detailed summary of dataset characteristics is provided in Table 7 (Appendix A). Below, we elaborate on EuroSAT and OFFICE-31, as they are newly introduced in our study.

EuroSAT [3]

EuroSAT is a land-cover classification dataset consisting of 27,000 images across 10 classes, with each image comprising 13 spectral bands ($64 \times 64 \times 13$). Unlike object-centric datasets like CIFAR10 and STL10, EuroSAT focuses on scene classification, where information is distributed across the entire image rather than centered on specific objects. This introduces a fundamentally different learning challenge and feature extraction process. Following standard practice, we split the dataset into 20% test data, with the remaining 80% split into 80/20% train/validation subsets. Figure 3 in Appendix E visualizes the 10 classes and 13 channels.

OFFICE-31

OFFICE-31 contains 4,110 images spanning 31 object categories in three domains: Amazon (A), DSLR (D), and Webcam (W). The dataset is widely used for domain adaptation tasks, as the domain differences arise from distinct backgrounds, noise, and resolution levels. FroSSL was not originally designed for domain adaptation, so this dataset enables us to evaluate its generalizability to such tasks. Appendix C provides domain-specific visualizations.

5 Experiments and Findings

5.1 Paper Results Reproduction and Multicrop Extension

We partially reproduced the official paper's results, as no checkpoint was provided. This established strong baseline models for further analyses in Section 5.2 and enabled testing different data augmentation strategies, including an asymmetric approach absent in the original paper. As shown in Table 1, this approach improved performance and was thus used for subsequent tasks.

We focused on smaller datasets (STL10 and CIFAR10) to reproduce the official results for 2, 4, and 8 views. Although 2-view performance on STL10 matched the original results, extending to 4 views using the authors' preconfigured GitHub settings yielded only 18.4% top-1 accuracy. This discrepancy led us to identify multiple inconsistencies between the report, the repository [1], and the results. For instance, the invariance hyperparameter for CIFAR10 did not match the provided YAML files, contradicting Appendix D.2, and color jitter settings were nearly halved relative to the paper's claims. Interestingly, removing Random Resize Crop (RRC) improved 4-view accuracy on STL10 to 50.1% and on CIFAR10 to 50.2%. In contrast, using their default CIFAR10 training script for FroSSL achieved only 51.6% accuracy, far below the reported 92.8%. At this point, we decided to not proceed training on 8 views.

We also explored a multicrop extension, testing multiple views and additional smaller crops. Since multiple-view training on STL10 was not viable, we focused on CIFAR10. Due to the small image size (32x32), we evaluated 8x8 and 16x16 crops. We trained with two full views plus 2, 4, or 8 smaller crops, running each model for 100 epochs with consistent hyperparameters, augmentation, and optimization settings (detailed in Appendix B). As Table 2 shows, multicrop did not markedly improve accuracy over two-view baselines, although 16x16 crops performed slightly better than 8x8 crops.

5.2 Embeddings visualization and evaluation

The authors of [6] employ pretrained backbones as encoders to create a database of feature embeddings rather than storing the entire model. They then perform classification using the k-NN algorithm

Method	CIFAR-10	STL-10
2 views	51.6	87.3
4 views [no rrc]	50.2	50.1
4 views [github]	-	18.4
4 views [report]	-	16.8

Table 1: Top-1 accuracies of FroSSL on STL10 and CIFAR10, *no rrc* is the result of training without Random Resize Crop. *github* was trained using the preconfigured YAML files and *report* was configured with the explicit configurations detailed in the report.

Crop Size	2 Views	4 Views	6 Views
8x8	50.3	48.4	48.6
16x16	**50.7**	49.8	50.4

Table 2: Top-1 accuracies for the multicrop approach on CIFAR10. Each model uses two full views plus 2, 4, or 8 smaller crops.

applied to these stored embeddings. This approach depends heavily on the quality of the encoded features, therefore we can assess how effectively our trained backbones have learned by adopting the same methodology and comparing the resulting performance.
We produce the embeddings corresponding to our training images, normalize them and apply PCA when it's needed, we then produce the embeddings of the samples we want to classify and normalize them with the parameters obtained on the training set. Finally we evaluate k-NN accuracy using as metric the cosine similarity. Given that most of the proposed backbones are significantly larger and have been trained for a longer period of time on bigger datasets, we also evaluate accuracy using a ResNet18 pretrained on ImageNet1K to ensure a somewhat fair comparison.
The results provided in Table 3 show how our encoders trained with FroSSL produce different results based on the dataset they have been trained on: it seems that this SSL method suffer from poor data quality as in cifar10, producing results that are worse than MAE and far from our ResNet 18 baseline. When testing on STL-10 instead, the results are more promising as we gain a substantial improvement compared to MAE while reducing the gap with our baseline model. It is worth mentioning that a

Method	CIFAR-10	STL-10	EuroSAT
Proposed in [6]			
CLIP [7]	**94.4**	**98.9**	-
MAE [2]	51.8	66.6	-
Baseline ResNet 50	75.6	94.3	-
Our evaluations			
Pretrained ResNet 18	59.48	87.75	-
FroSSL ResNet	(18) 49.34	(18) 85.49	(50) 75.68

Table 3: Performance of various methods for k-NN embedding classification (in FroSSL we indicated the type of ResNet used as backbone)

deeper analysis shows that the nature of most frequent mistakes is somehow humanly comprehensible: for example if we look at a given class, the majority of misclassified embeddings have been classified as a class *semantically* related to it (e.g. cat with dogs, trucks with ships). Related visualizations and analysis can be found in Appendix D.

5.3 Scene classification

In Order to test the proposed method in [8] on a different task, we choose to do Scene classification with the EuroSAT Dataset [3].

5.3.1 Augmentation

The given augmentation pipeline in [1] are rgb specific. However these are not suitable for the 13 channel satelite images. Four simple and reasonable augmentations have shown the most promising results, that are rotation of max 30 degree, vertical flip with probability 0.5, horizontal flip with probability 0.5 and channel wise gaussian noise with 1% of the standard deviation. From that, one might already consider, that there is not enough variety for more than 2 views. Also, since the images have a fixed length and width of 64 pixel and are all taken in a very similar distance, there is no reason to use mutlicropping. We further justify our choice of augmentation with the results in Table 4.

5.3.2 Pretraining Backbone

Using the Augmentation pipeline as described above, we trained a ResNet50 backbone model, where we adapted the first convolution to the 13 channels. Further, with accordance to [1], the classification head is removed and replaced by a Projection head (Linear, BatchNorm, Relu, Linear, BatchNorm, Relu, Linear) with a hidden projection dimension of 2048 and an output projection dimension of 1024. For our main experiments, shown in Table 5, we trained the backbone for 400 epochs using the FroSSL-loss with an invariance weight of 1.4.

5.3.3 Classification Head

Utilizing the pretrained backbone, we fitted a classification head (linear layer with cross-entropy-loss) for 100 epochs. In Table 5 we show the test accuracy and compare it with other similar methods, as well as the baseline.

Parameter	Test Accuracy [%]
final model	85.85
4 views	84.81
multicropping	52.83
max rotation of 45	83.52

Table 4: Test accuracy of the FroSSL method on the EuroSAT dataset. Each backbone has been trained for 100 epochs with the FroSSL-loss and an invariance weight of 1. Using the frozen backbone we trained a classification head for 50 epochs. The final model uses 2 views, a max rotation of 30 degree and no multicropping.

Method	Test Accuracy [%]
Benchmark	
CNN (two layers)	87.96
GoogleNet	96.02
ResNet-50	96.43
Our evaluations with frozen backbone	
mmcr	83.43
simclr	89.63
i-vne	90.85
frossl	91.35
Our evaluations with finetuned backbone	
i-vne	97.96
frossl	**98.11**

Table 5: Performance of the FroSSL method on the EuroSAT dataset compared with the Benchmark [3] and three other competing SSL-methods, mentioned in [8](same hyper-parameters as described above, but different loss).

5.3.4 Discussion

We discuss our results in Table 5. We can see in the evaluation with frozen backbone, that the embedding, created by the FroSSL-loss beats the compared embeddings. However the achieved accuracy is far from the Benchmark accuracy, which is reasonable, since the embedding is learned unsupervised. We managed to close the gap and beat the benchmark by fine-tuning the backbone to the classification task.

5.4 Semi-SL for unseen domain adaptation

5.4.1 Problem setting

Following [5], we train a model on two different domains of Office31 dataset and evaluate its performance on the third. We do the training in a semi supervised way by having one domain as labeled and the other one as unlabeled.

5.4.2 Experiments

Experimental setting
While in [5] they are using a pre-trained ResNet50 backbone, we are already satisfied with a pre-trained ResNet18 backbone. Further we utilize the strong and asymmetric augmentation pipeline as in [8] for ImageNet dataset (more details in F) for our training datasets, since we believe that it is important to avoid overfitting, given the relative small amount of samples we have. We use $\lambda_u = 0.3$ and $\lambda_u = 0.5$ as they proved to be the best values based on our experiments.
For the following experiments, we compare our model with three groups of methods. (1) Supervised models: ResNet50 (from [5]) and ResNet18 (trained by us for a fair comparison, with the same settings used for our method and trained for 400 epochs). Both were pre-trained on ImageNet. (2) The other methods proposed in [5]. (3) Our method as described above but using a ResNet18 from

scratch, trained for 200 epochs. The standard hyperparameters are the same as those used in [8]: LARS as optimizer and a warm up cosine learning rate scheduler. We used a batch size of 64 for both labeled and unlabeled dataset and an initial learning rate of 5e-3.

Results
We comment our results summarized in Table 6. Our method outperforms previous works and supervised models, highlighting the effectiveness of the FroSSL objective in learning augmentation-invariant representation that are effective in domains that have similar semantic content while presenting a distribution shift.
It is even more surprising, that our method reaches state of the art accuracies in a very small amount of epochs (less than 30), which makes us suggest that the FroSSL loss has indeed a fast convergence as claimed by the authors in [8].

Method	**D/A/W**	**W/A/D**
Proposed in [5]		
supervised	84.9	92.4
DANN	16.7	4.2
CDAN	2.6	18.5
FixMatch	50.6	62.6
FM-Rot	50.6	66.9
FM-SSFA	85.9	92.8
FroSSL semi-SL	52.3	68.5
supervised (resnet18)	88.5	91
FroSSL semi-SL + pretrained ResNet18	**95.8**	**98.9**

Table 6: Comparison of accuracy for unseen domain adaptation on Office-31 (first letter is labeled domain, second is unlabeled domain, third is test domain)

6 Challenges

We think the paper is very well written and understandable and also the method is very good nested into the code base they used. However a few challenges came across when using their implementation [1] as basis for our investigations.
First of all, it was not straight forward doable to run the code on a CPU, which made the development very annoying. Also they did not provide any checkpoint and since the model is pretrained up to 1000 epochs, it is very costly to train it yourself. Last but not least, as described in Section 5.1, we found some misalignment between the parameters in the paper and the configurations in the repository. This cost us a lot of time since we had a hard time understanding if we were missing something when training our model and therefore, reiterated many times.

7 Conclusion

In our our experiments we have seen the strength of self-supervised learning in two different tasks. First we obtained that fine-tuning an unsupervised pretrained backbone can be better than learning directly in a supervised way. The pretraining is quite costly, especially when trying out more complex datasets like MIT67. However, once the backbone is trained, the resulting model/embedding can be used for various tasks.
With our experiments in domain adaption, the FroSSL-objective has proven to generate very valuable embeddings. This amazing result encourages to investigate further into similar settings. One can start by extending Table 6 to the other cases and try out more datasets. Further it would be interesting to analyze the embeddings and also see how the model performs when the set of classes in the two training sets are different from each other.
When using the Frossl model, you should definetly have access to a gpu. Further we recommend to think well about hyperparameters and maybe optimize them with only a subset of classes, since the pretraininig is very costly. Augmentation has shown to be very crucial for the performance resulting embedding and our experience is that the asymmetric augmentation works better than the symmetric.

8 Limitations

This project aims at highlighting the main limitations of the original work proposed. One of them, that has been confirmed by the authors themselves, is that the original experiments focused only on object centric dataset for classification tasks. We tried to overcome this issues by performing scene

classification on a non-object centric dataset. Another limitation might be that the authors didn't evaluate the quality of the embeddings created, that is why we tried to insert the features obtained in a framework that was created to evaluate them directly. Finally, FroSSL objective was used only for self-supervised learning tasks, while we tried to use it in a Semi-SL framework. The results seems promising but it's important to notice that the experiments done are very limited (in terms of dataset as one may extend it to Office-Home and in terms of instances tested) due to computational constraints. It might be relevant to continue investigating the goodness of FroSSL objective in such framework.

9 Self Assessment

From the given grading criteria and feedback received from our anointed TA, we deem that we have met the criteria for an excellent project. We have reached beyond the original paper in ways that are non-trivial, as we have followed many suggestions by the original authors and followed the requirements set by our TA.

Fair, conclusive but new comparisons with few other papers. Our experiments have been compared with state of the art papers on the same subject, resulting on a meaningful and valuable comparison.
Novel, interesting, and original application to a different task. Both Scene-Classification and Domain Adaptation for unseen classification are tasks that weren't tested in the original work, moreover our results proved that employing such method in this new tasks might indeed be valuable.
Novel, interesting, and/or informative modification or incremental improvement of the proposed method / justifiable combination of the methods proposed in two or more papers and corresponding experiments. We used the method proposed in [8] in the framework proposed in [5] with some adjustments. This combination outperformed previous results in a smaller amount of time, on a smaller net and by a wide margin.

Further we think that we fulfill the following bonus criteria.

Complexity of the new application/dataset. While the FroSSL paper and most of the course content is based on object centric RGB images, we applied the method to 13 channel satellite images. Therefore we had to rethink all the augmentations since they are crucial for the method.
Noticeably interesting and informative observations from the experiments. Especially with our experiments in Section 5.4 we observed very good results that beat all the models proposed by others by far.

References

[1] https://github.com/ofskean/frossl.

[2] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross B. Girshick. Masked autoencoders are scalable vision learners. *CoRR*, abs/2111.06377, 2021.

[3] et al. Helber. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *ArXiv preprint arXiv:1709.00029*, 2017.

[4] Jaeill Kim, Suhyun Kang, Duhun Hwang, Jungwook Shin, and Wonjong Rhee. Vne: An effective method for improving deep representation by manipulating eigenvalue distribution, 2023.

[5] Jiachen Liang, Ruibing Hou, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen. Generalized semi-supervised learning via self-supervised feature adaptation, 2024.

[6] Kengo Nakata, Youyang Ng, Daisuke Miyashita, Asuka Maki, Yu-Chieh Lin, and Jun Deguchi. Revisiting a knn-based image classification system with high-capacity storage, 2022.

[7] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *CoRR*, abs/2103.00020, 2021.

[8] Oscar Skean, Aayush Dhakal, Nathan Jacobs, and Luis Gonzalo Sanchez Giraldo. Frossl: Frobenius norm minimization for efficient multiview self-supervised learning, 2024.

10 Appendix

A Datasets

Dataset	Classes	Samples	Image Dimensions
CIFAR10	10	60,000	$32 \times 32 \times 3$
STL10	10	105,000	$96 \times 96 \times 3$
EuroSAT	10	27,000	$64 \times 64 \times 13$
OFFICE-31	31	4,110	Varies (Domain-Specific)

Table 7: Summary of datasets used in this study. Detailed dataset descriptions and visualizations are in Appendix A.

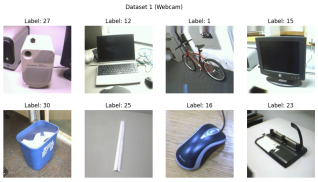
B Multicrop configuration

Settings used for all multicrop experiments on CIFAR10.

- Random Resized Crop - Crop scale ranges from 0.08 to 1.0
- Color Jitter with probability 0.8 and (brightness, contrast, saturation, hue) values of (0.8, 0.8, 0.8, 0.2)
- Grayscale with probability 0.2
- Gaussian blur with probability 0.5
- Horizontal flip with probability 0.5

C Office-31 domain visualization

(a) Amazon domain

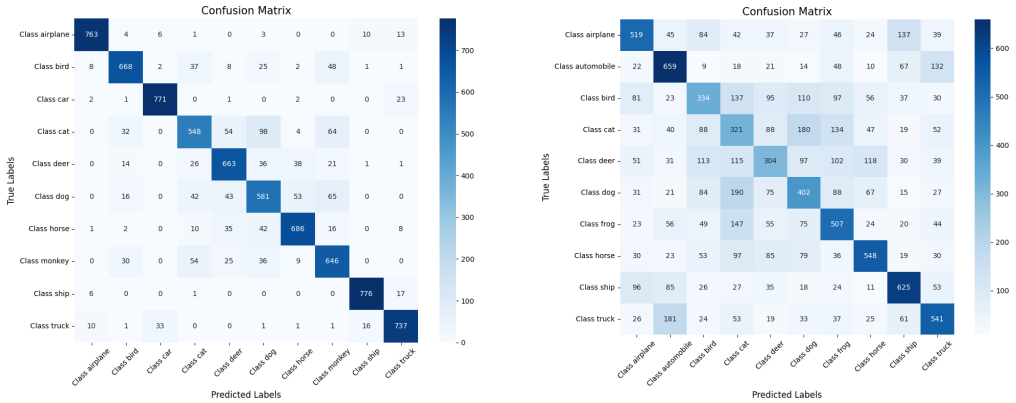


(b) Webcam domain

(c) DSLR domain

Figure 1: Office 31 different domains: Amazon images were captured from a website of online merchants, they are captured against clean background and at a unified scale. DSLR domain contains 498 low-noise high resolution images (4288×2848). There are 5 objects per category. Each object was captured from different viewpoints on average 3 times. In Webcam domain, the 795 images are of low resolution (640×480) and exhibit significant noise and color as well as white balance artifacts.

D Further analysis of embeddings capacity of FroSSL backbones



(a) STL-10 (b) CIFAR-10

Figure 2: Results on k-NN classification of test embeddings: evidence shows that many misclassified embeddings are assigned to semantically related classes. This can be noticed since most of animals related embeddings are classified as classes that are strongly related (e.g. in STL-10: deer and horse for the similar shape, bird and monkey for their natural habitat). One of the few exceptions in CIFAR-10 is the tuple (airplane, bird): the logical explanation it that, given the low quality of the images, our model "focused" on the common background and general shape, resulting in a high number of classification errors related to this.

E EuroSAT MSI dataset

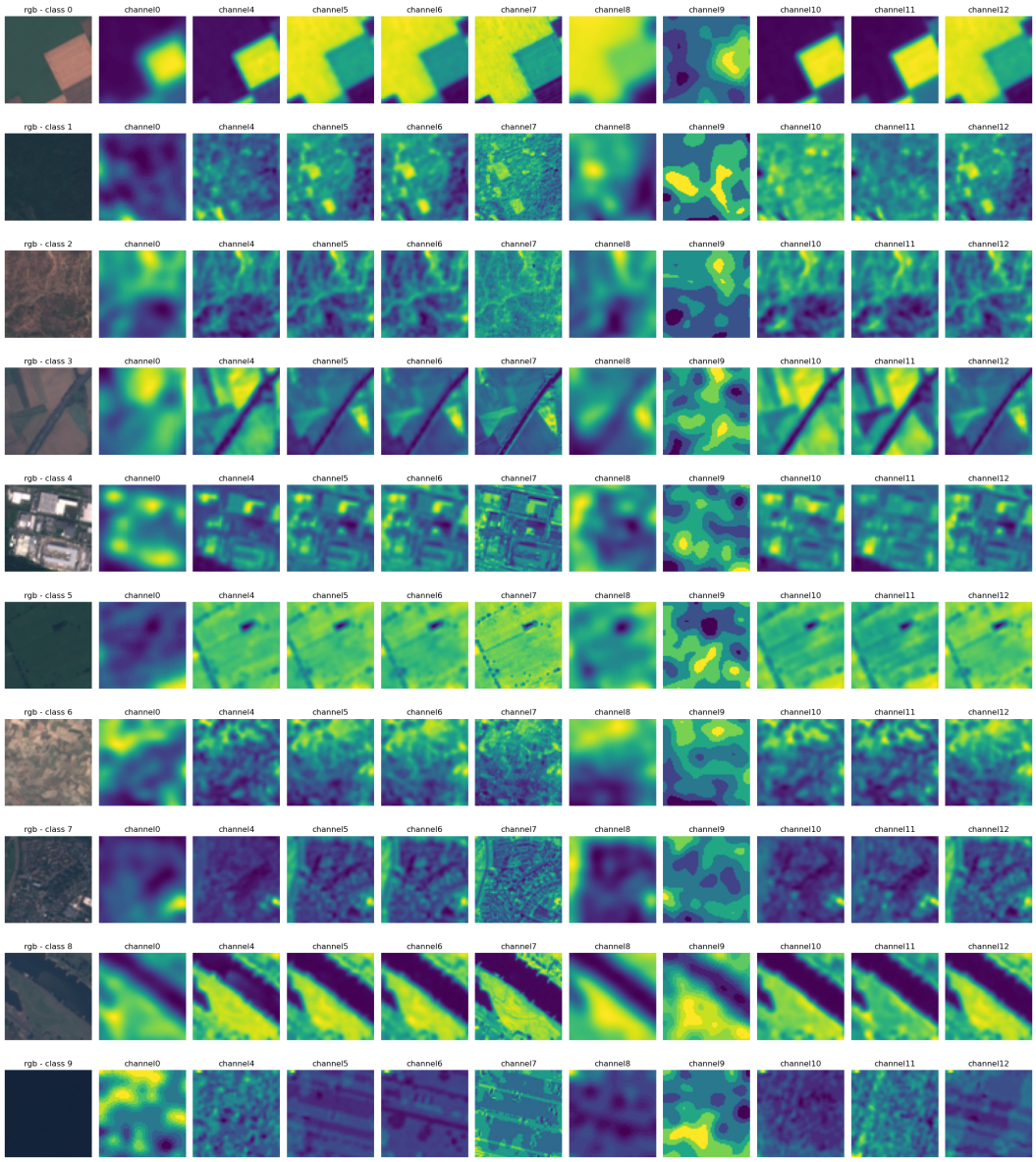


Figure 3: Images from the EuroSAT MSI dataset. Each row corresponds to one of the 10 classes. The first column are RGB channels, where the pixel values are multiplied by 3 for better visibility. The other columns are corresponding to the remaining channels.

F Semi-SL augmentation

Settings used for the augmentations in the Semi-SL method. Since we use an asymmetric augmentation with 2 views we generate 2 samples from each image transformed with different parameters.

- Random Resized Crop - Crop scale ranges from 0.08 to 1.0 in both views.
- Color Jitter with probability 0.8 and (brightness, contrast, saturation, hue) values of (0.8, 0.8, 0.8, 0.2) in both views.
- Grayscale with probability 0.2 for both views
- Gaussian blur with probability 1.0 for first view and 0.1 for the second view

- Solarization with probability 0.0 for first view and 0.2 for the second view
- Horizontal flip with probability 0.5 for both views