# EVALUATION OF SELF-SUPERVISED LEARNING APPROACHES FOR SEMANTIC SEGMENTATION OF INDUSTRIAL BURNER FLAMES

S. Landgraf[1,*], L. Kühnlein[2], M. Hillemann[1], M. Hoyer[2], S. Keller[1, 2], M. Ulrich[1]

[1] Institute of Photogrammetry and Remote Sensing (IPF), Karlsruhe Institute of Technology (KIT), Germany -
(steven.landgraf, markus.hillemann, markus.ulrich)@kit.edu
[2] ci-tec GmbH, Germany -
(l.kuehnlein, s.keller, m.hoyer)@ci-tec.de

**Commission II, WG II/6**

**KEY WORDS:** Semantic Segmentation, Self-Supervised Learning, Deep Learning, Industrial Burner, Flame Segmentation, Industrial Automation.

## ABSTRACT:

In recent years, self-supervised learning has made tremendous progress in closing the gap to supervised learning due to the rapid development of more sophisticated approaches like SimCLR, MoCo, and SwAV. However, these achievements are primarily evaluated on common benchmark datasets. In this paper, we focus on evaluating self-supervised learning for semantic segmentation of industrial burner flames. Our goal is to build an intuition on how self-supervision performs in a scenario relevant for industrial application where training labels and the opportunities for hyperparameter tuning are limited. We demonstrate that self-supervised pre-training can constitute an alternative to the state-of-the-art approach of pre-training on ImageNet. Across all scenarios, the self-supervised approaches are less susceptible to sub-optimal learning rates and achieve higher mean accuracies than ImageNet pre-training, especially when training labels are scarce.

## 1. INTRODUCTION

Image segmentation aims to assign a class label to each pixel in an image and hence can be thought of as a pixel-wise classification task. This key computer vision task serves many applications, like scene understanding, medical image analysis, autonomous driving (Minaee et al., 2021), or industrial automation (Steger et al., 2018). Semantic segmentation, a subcategory of image segmentation, provides each image pixel with a semantic label of a set of object categories. Deep convolutional neural networks (CNNs) are the state-of-the-art technique for semantic image segmentation (Long et al., 2015; Minaee et al., 2021). However, for these methods to achieve high performance, thousands of pixel-precise labels must be provided during training, which are laborious to create and often depend on the personal assessment of the human operator, as Figure 2(b) shows.

A relatively new approach that strives to significantly reduce the need for label annotations is self-supervised learning (SSL) for computer vision (Doersch et al., 2015; Zhang et al., 2016). This approach is particularly useful for applications which require strenuous or ambiguous labeling work. An SSL pipeline can generally be divided into the *pre-training stage* and the *fine-tuning stage*. The latter is equivalent to fine-tuning a model with pre-trained weights and biases, whereas the former does not require labeled data. Instead, a proxy task is designed that enables the network to learn feature representations based on self-derived supervisory signals from unlabeled images. For instance, a possible proxy task could train a CNN that determines the two-dimensional (2D) rotation between an artificially rotated image and the original image (Gidaris et al., 2018). More sophisticated approaches like ”Simple framework for Contrastive Learning of visual Representations” (SimCLR) (Chen et al., 2020), ”Momentum Contrast” (MoCo) (He et al., 2020), and ”Swapping Assignments between multiple Views” (SwAV) (Caron et al., 2020) approach the pre-training stage by using various augmentation techniques paired with a contrastive loss function.

Recent research shows that SSL approaches can outperform a supervised baseline on ImageNet (Tomasev et al., 2022). In contrast to many conventional pre-training approaches, SSL does not suffer from a domain gap since the pre-training and fine-tuning stages use the same dataset. As a consequence, self-supervision is getting more popular in very domain-specific applications, such as chest X-rays or dermatological images (Azizi et al., 2021), for example.

This work aims to investigate the effect of self-supervision on a dataset vastly different from ImageNet. To achieve this, we evaluate three promising SSL approaches on the task of semantic segmentation of industrial burner flames. This segmentation task is motivated by our long-term goal: to enable the automatic extraction of process-relevant parameters and thereby contribute to optimize industrial combustion processes regarding energy efficiency and $CO_2$ emission reduction.

## 2. METHODOLOGY

In this section, we summarize information about the dataset, network architecture, self-supervised pre-training, fine-tuning, and the evaluation methodology. Our goal is to assess the impact of SSL for semantic segmentation of industrial burner flames in comparison to the well-established practice of ImageNet pre-training.

[*] Corresponding author

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLIII-B2-2022-601-2022 | © Author(s) 2022. CC BY 4.0 License.

### 2.1 Dataset

Our evaluation is based on the freely available industrial burner flames dataset provided by Großkopf et al. (2021). It contains 3000 labeled grayscale images of two industrial burner flames in an augmented and a non-augmented configuration. We use the non-augmented configuration for our evaluation and refer to this subset as the dataset in the remainder of this paper. The dataset is already split into training ($80\,\%$), validation ($10\,\%$), and test ($10\,\%$) subsets. We do not perform any data augmentation but normalize each image based on the dataset's mean and standard deviation of the gray values.

### 2.2 Network Architecture

For every approach, we use a feature pyramid network (FPN) (Lin et al., 2017), consisting of a pyramidal hierarchy of CNNs. It contains a bottom-up feature encoder pathway, for which we use a 50-layer residual network (ResNet-50). Additionally, a top-down pathway with lateral connections builds high-level semantic feature maps at all scales (Lin et al., 2017). Together with the FPN decoder, the architecture consists of roughly 26M trainable parameters. In comparison to the 1.3B parameter state-of-the-art "SElf-supERvised" (SEER) model from Goyal et al. (2021a), our architecture is rather compact, which reduces training time and memory usage significantly.

In this paper, the following pre-training scenarios are evaluated and compared:

1. Initialization of a ResNet-50 backbone with pre-trained weights and biases from ImageNet,
2. Initialization of a ResNet-50 backbone with pre-trained weights and biases from three SSL approaches:
   (a) SimCLR as proposed in Chen et al. (2020),
   (b) MoCo as proposed in He et al. (2020), and
   (c) SwAV as proposed in Caron et al. (2020).

### 2.3 Self-Supervised Pre-training

We use the computer vision library for state-of-the-art self-supervised learning research (VISSL) (Goyal et al., 2021b). The three SSL approaches, which we compare for pre-training in this paper, are implemented according to their original publications. All pre-training methods use a contrastive loss to bring similar representations closer together.

SimCLR (Chen et al., 2020) uses a pool of image augmentations consisting of random crops, flips, color distortions, and Gaussian blur. These are composed randomly to produce positive pairs of augmented images. The network is then trained to maximize the agreement between the differently augmented views of the same image via a contrastive loss.

MoCo (He et al., 2020) views contrastive learning as dictionary look-up problems similar to natural language processing. It builds a dynamic dictionary with a queue and a moving-averaged encoder. In essence, there is a query $q$ that matches one of the many keys $k$ in the dictionary. The input images or patches to the queries and keys are denoted by $x_q$ and $x_k$. $k$ is computed from $x_k$ by a momentum encoder that is slowly updated. $q$ is computed from $x_q$ by a second encoder and matched to the dictionary keys using a contrastive loss that becomes small for high similarity. As a result, it builds a large and consistent dictionary. MoCo provides competitive results on ImageNet classification and transfers well to the fine-tuning stage.

SwAV (Caron et al., 2020), in essence, predicts a cluster assignment of a view from the representation of another view. Multiple crops of the same image are produced and augmented to achieve this, similar to SimCLR. These crops are passed through a backbone model. A shallow non-linear network then predicts a projection vector, mapped to trainable prototype vectors by a single linear layer (prototype layer). The output of this layer is used for cluster assignment with the Sinkhorn Knopp algorithm (Cuturi, 2013). These assignments are then swapped and used to predict the swapped targets.

### 2.4 Fine-Tuning

In the fine-tuning stage, the learned representations from pre-training are used to initialize the weights and biases of the network. Thereafter, the network is fine-tuned with the labeled data.

The following scenarios are considered:

1. Training on $1\,\%$ of the dataset with a frozen backbone,
2. Training on $1\,\%$ of the dataset,
3. Training on $10\,\%$ of the dataset with a frozen backbone,
4. Training on $10\,\%$ of the dataset,
5. Training on $100\,\%$ of the dataset with a frozen backbone, and
6. Training on $100\,\%$ of the dataset.

Hereby, we choose random subsets that are identical for all scenarios. Each fine-tuning process was given 30 epochs to fit the given training data with the widely used region-based Dice Loss (Jadon, 2020) and an Adam optimizer (Kingma and Ba, 2014). We decide to fine-tune frozen backbone scenarios to gain further intuition on how well SSL works for our domain-specific application and to find out whether a frozen backbone scenario is a worthwhile, resource-saving alternative. For these scenarios, the weights and biases of the ResNet-50 backbone are not affected by the backward propagation during fine-tuning.

### 2.5 Evaluation Methodology

To evaluate the effect of SSL for semantic segmentation of industrial burner flames, we choose 30 learning rates from a log-uniform distribution in the range of $1 \times 10^{-6}$ to 1. Preliminary experiments have shown that other hyperparameters have a small impact on the training. For this reason, and to reduce the dimensionality of the evaluation space, all other hyperparameters are set to fixed values. We use a mini-batch size of 4, a momentum of 0.9, and no learning rate schedule. This method mimics a random search for the optimal learning rate in hyperparameter-tuning based on the recommendations by Bengio (2012).

To compare the different scenarios, we use the accuracy as our main metric. It reports the proportion of pixels in an image that are classified correctly. Note that accuracy can be misleading when the class representation is small within the image, as the metric will basically report how well negative cases were identified. However, across all images in the dataset used in this paper, class representation of flames is relatively high, and thus accuracy is not heavily biased by true negatives.

Over the 30 learning rates, we compute the mean and maximum accuracy, as well as count the number a certain accuracy threshold is exceeded for each approach and scenario.

| Index | Scenario | Pre-training | A. > 90 % [%] and absolute counts | | A. > 95 % [%] and absolute counts | | Max. A. [%] | Mean A. [%] |
|---|---|---|---|---|---|---|---|---|
| 1 | frozen backbone, 1 % | ImageNet | 0.0 | 0 | 0.0 | 0 | 76.3 | 75.9 |
| | | MoCo | 16.7 | 5 | 0.0 | 0 | 94.7 | 79.0 |
| | | SimCLR | 23.3 | 7 | 0.0 | 0 | 94.7 | 80.0 |
| | | SwaV | 13.3 | 4 | 0.0 | 0 | 94.5 | 78.7 |
| 2 | unfrozen backbone, 1 % | ImageNet | 26.7 | 8 | 10.0 | 3 | 95.8 | 81.0 |
| | | MoCo | 33.3 | 10 | 6.6 | 2 | 95.1 | 82.1 |
| | | SimCLR | 33.3 | 10 | 3.3 | 1 | 95.3 | 82.0 |
| | | SwaV | 30.0 | 9 | 10.0 | 3 | 95.4 | 81.9 |
| 3 | frozen backbone, 10 % | ImageNet | 23.3 | 7 | 10.0 | 3 | 95.4 | 80.4 |
| | | MoCo | 26.7 | 8 | 20.0 | 6 | 95.7 | 81.1 |
| | | SimCLR | 30.0 | 9 | 23.3 | 7 | 95.7 | 81.7 |
| | | SwaV | 23.3 | 7 | 20.0 | 6 | 95.6 | 80.6 |
| 4 | unfrozen backbone, 10 % | ImageNet | 33.3 | 10 | 26.7 | 8 | **96.7** | 82.6 |
| | | MoCo | 33.3 | 10 | 33.3 | **10** | 96.5 | 82.7 |
| | | SimCLR | 33.3 | 10 | 30.0 | 9 | 96.5 | 82.6 |
| | | SwaV | 40.0 | 12 | **33.3** | **10** | 96.5 | 83.9 |
| 5 | frozen backbone, 100 % | ImageNet | 30.0 | 9 | 0.0 | 0 | 94.3 | 81.3 |
| | | MoCo | 33.3 | 10 | 0.0 | 0 | 93.6 | 81.6 |
| | | SimCLR | 33.3 | 10 | 0.0 | 0 | 93.5 | 81.7 |
| | | SwaV | 30.0 | 9 | 0.0 | 0 | 94.2 | 81.2 |
| 6 | unfrozen backbone, 100 % | ImageNet | 33.3 | 10 | 23.3 | 7 | 96.3 | 82.5 |
| | | MoCo | 33.3 | 10 | 3.3 | 1 | 95.2 | 82.1 |
| | | SimCLR | **43.3** | **13** | 6.6 | 2 | 95.1 | **84.0** |
| | | SwaV | 40.0 | 12 | 10.0 | 3 | 95.3 | 83.4 |

Table 1. Evaluation results on the test dataset after fine-tuning the FPN with four different pre-training methods: ImageNet, MoCo, SimCLR, and SwAV. The third to sixth columns illustrate the proportion of models that reach a certain accuracy. Furthermore, the two last columns show the maximum and mean accuracy that is achieved. Results are based on 30 random learning rates uniformly distributed on a logarithmic scale from $1 \times 10^{-6}$ to 1. The highlighted fields mark the best results within the respective scenario.

## 3. RESULTS

In the following, we provide a quantitative evaluation in addition to a qualitative evaluation in the form of a visual comparison of inference results. Our goal is to build an intuition for how well self-supervision performs in scenarios where training labels and the opportunities for hyperparameter tuning are limited.

### 3.1 Quantitative Evaluation

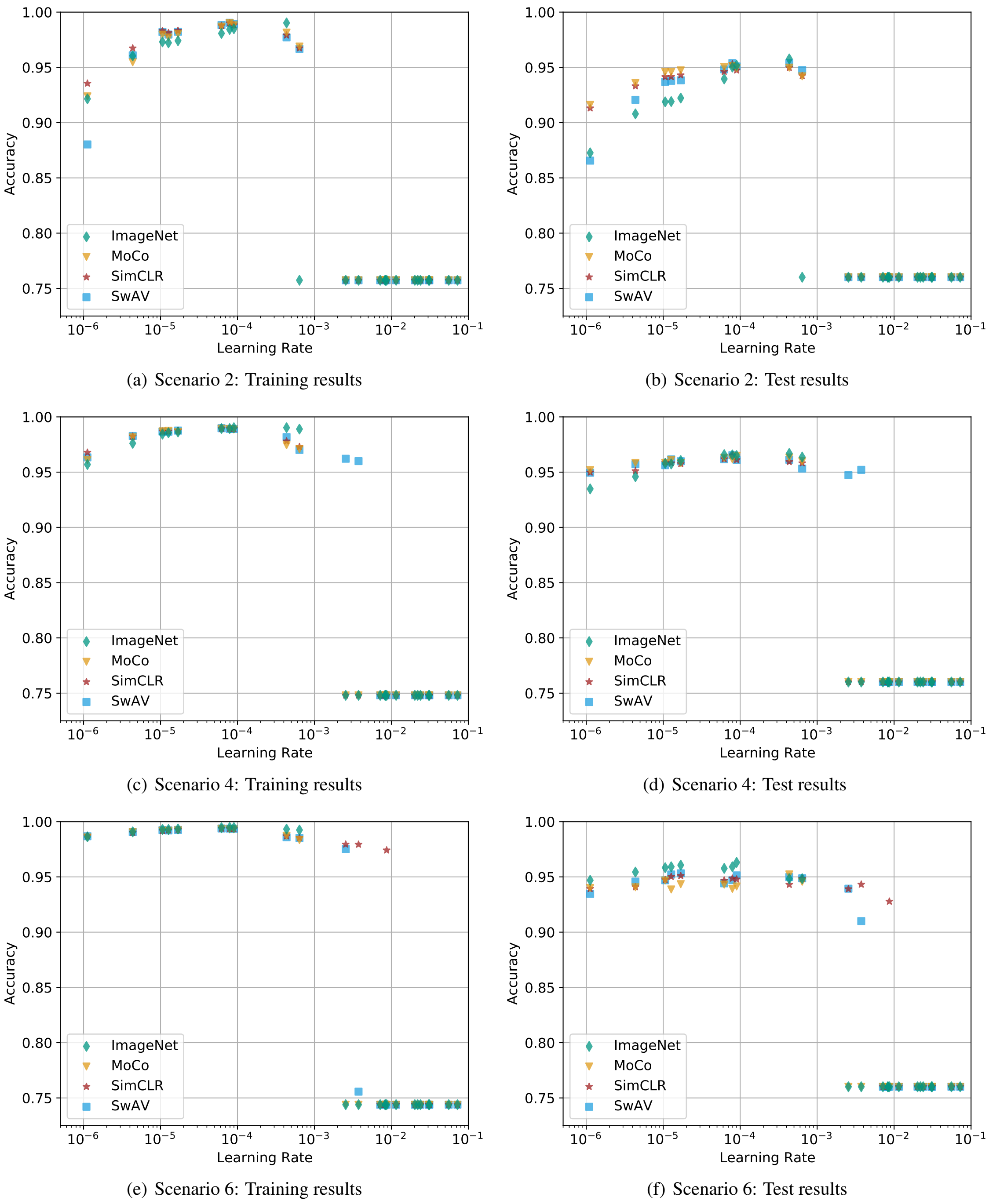
Table 1 shows the numerical results of our evaluation. As described in detail in Section 2.5, we train every scenario with 30 different learning rates selected from a log-uniform distribution. All metrics are based on the evaluations of these 30 trials. While the ImageNet-pre-trained approaches score the highest maximum accuracy in four of six scenarios, the highest mean accuracy is scored by one of the SSL approaches in every case. Although not beating the maximum accuracy of the ImageNet-pre-trained models in any unfrozen scenario, the SSL approaches still achieve comparable results and accomplish to beat the maximum accuracy in two of the three frozen backbone scenarios. Generally, the unfrozen scenarios converge more often properly than their respective frozen counterpart and achieve higher maximum and mean accuracies. In general, we observe better results by training with just 1 % of the training labels with an unfrozen backbone than training with 100 % of the training dataset but a frozen backbone.

Another aspect can be observed in columns three and four. These columns illustrate the number of cases in which a certain accuracy threshold is reached as a percentage of the total amount (30) of sampled learning rates and as an absolute number. Here, the SSL approaches reach the best results across all scenarios. This observation is especially prevalent in the scenarios where only 1 % of the training dataset is provided for fine-tuning and the ResNet-50 backbone's weights and biases are frozen. Here, the ImageNet-pre-trained approach did not converge properly a single time over all 30 learning rates resulting in a maximum accuracy of 76.3 % on the test dataset. In contrast, all three SSL approaches converge properly over the 90 % threshold for multiple learning rates.

As will be discussed in more detail in Section 3.3, all approaches score somewhat similarly. However, the SSL approaches always exceed the specified accuracy threshold more often than the models pre-trained on ImageNet with the exception of the second and sixth scenarios for the 95 % class. Moreover, the ImageNet-pre-trained approaches achieve the highest maximum accuracy in all of the scenarios except for the first and third ones.

In the 90 % threshold class, models pre-trained with SimCLR converge properly most frequently except for the fourth scenario. However, this is not the case in the 95 % threshold class. Here, the approaches that use SimCLR only score the best in one out of the six scenarios.

Below the 95 % threshold, the approaches that use MoCo and SwAV show similar behavior. In the second scenario, they score similarly well to the SimCLR approach. Still, they surpass the other approaches in one additional case each. Above the 95 % threshold, MoCo and SwAV score best in the same scenario and SwAV scores best in a second scenario.



(a) Scenario 2: Training results

(b) Scenario 2: Test results

(c) Scenario 4: Training results

(d) Scenario 4: Test results

(e) Scenario 6: Training results

(f) Scenario 6: Test results

Figure 1. Fine-tuning results of 30 learning rates for each pre-training scenario of the FPN with an unfrozen backbone.

### 3.2 Fine-Tuning Results

Figure 1 visualizes the results on the test dataset (right column) and, for comparison, on the training dataset (left column).

The training results on 100 % of the training dataset in Figure 1(e) show that there is little to no difference between all approaches for learning rates in the range of $10^{-6}$ to $10^{-4}$. Between $10^{-4}$ and $10^{-3}$ the ImageNet-pre-trained models fit the training data slightly better. The main differences between the four approaches can be observed for learning rates between $10^{-3}$ and $10^{-2}$. In this range, the ImageNet- and MoCo-approaches do not converge properly but one SwAV-model does and even three SimCLR-models do. All of the remaining learning rates from $10^{-2}$ to 1 lead to improper convergence in all approaches. Figure 1(f) reveals that the FPN that was pre-trained on ImageNet generalizes best after fine-tuning on 100 % of the training dataset.

By reducing the available training labels to 10 %, similar training behavior can be noticed, as Figure 1(c) visualizes. However, the SSL approaches manage to fit the training data slightly better with the smallest learning rates in the range of $10^{-6}$ to $10^{-5}$. Figure 1(d) manifests this observation as all of the SSL approaches achieve higher accuracies in this range of learning rates. With learning rates larger than $10^{-3}$, only SwAV manages to converge properly.

As shown in Figure 1(a), our evaluations find the most significant difference when only 1 % of the training labels are available

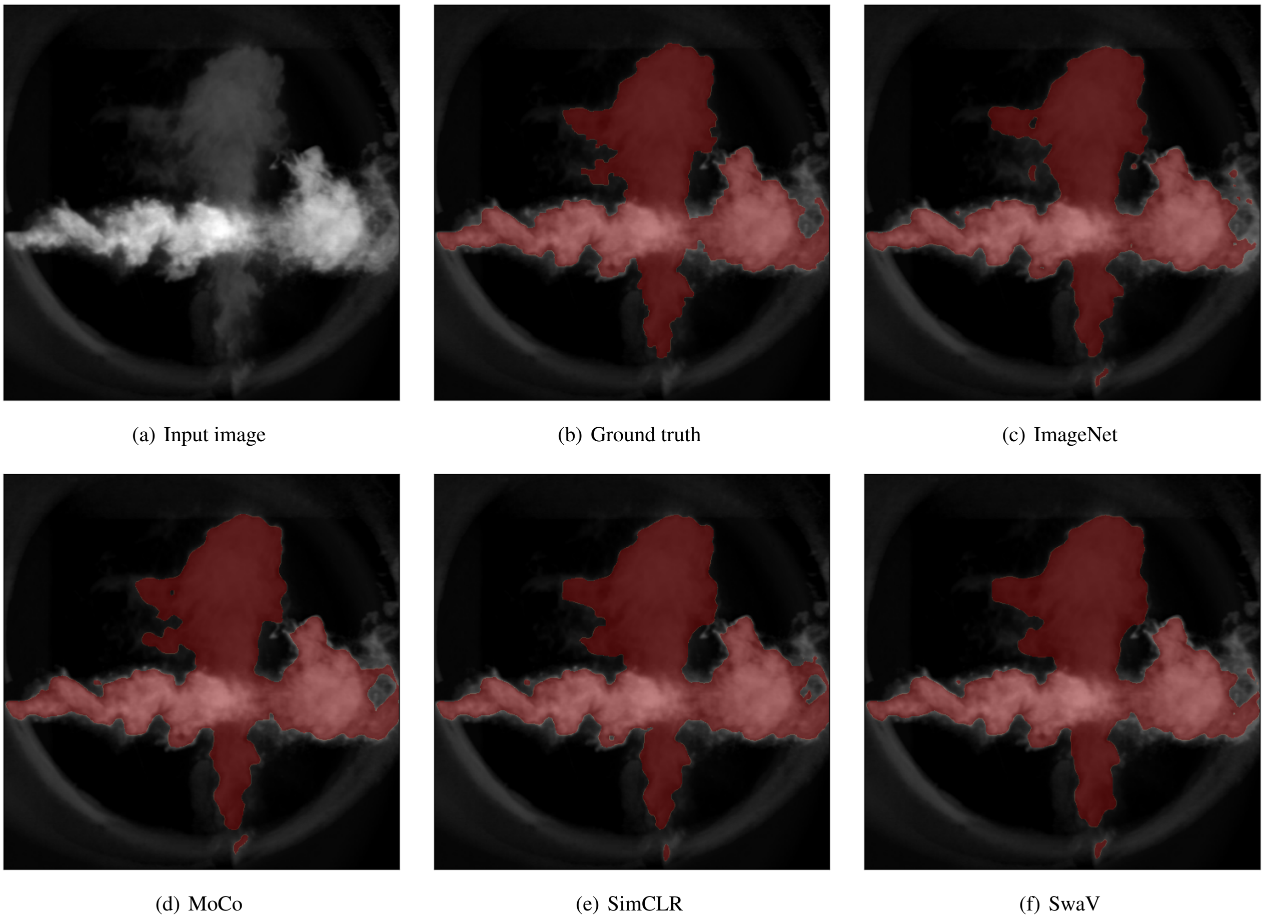


Figure 2. Visual comparison between an input image from the test dataset (a), the ground truth (b), and inference results from ImageNet (c), MoCo (d), SimCLR (e), and SwaV (f). We used the models that achieved the highest accuracy on the test dataset for inference.

for fine-tuning. Except for very few exceptions, all of the SSL approaches achieve higher accuracies on the training dataset across all learning rates. However, none of the four approaches converge with learning rates larger than $10^{-3}$. Figure 1(b) reveals that the SSL approaches also generalize better onto the test dataset than the FPN pre-trained with ImageNet.

The frozen backbone scenarios show largely the same behavior, although magnified for the $1\,\%$ scenario. Here, the ImageNet-pre-trained approach does not fit the training data once, whereas the SSL approaches manage to converge properly multiple times. They can be found in Figure 3 in the Appendix.

### 3.3 Qualitative Evaluation

Figure 2 shows a visual comparison between a sample input image in Figure 2(a), the corresponding ground truth in Figure 2(b) and inference results of the models that achieved the highest accuracies on the test dataset for the approaches pre-trained using ImageNet, MoCo, SimCLR, and SwAV in Figure 2(c) through Figure 2(f). At first glance, the results of all four models are not easily distinguishable from the ground truth. In Figure 2, the SSL approaches seem to segment more coherent flames than the model pre-trained on ImageNet. Across $100\,\%$ of the dataset, though, we found no significant visual differences between the inference results of the four models. However, it has to be mentioned that the provided ground truth masks by Großkopf et al. (2021) are questionable in minor details. As shown in Figure 2(b), the ground truth label is missing a darker part of the flame in the middle right part of the input image and is not pixel-perfect in other parts. These findings are not an exception and can be observed across $100\,\%$ of the dataset.

By looking at the inference results of models that did not properly train, meaning they were stuck in a local minimum in training and thus did not converge properly, we find that the FPN predicts no flames for the entire image. In these cases, the accuracy of $76.3\,\%$ on the test dataset is achieved. This observation shows that imbalanced data is a significant problem for practitioners.

## 4. DISCUSSION

Overall, the experiments confirm that a suitable choice of the learning rate is important for the success of all tested approaches. If the learning rate is not chosen appropriately, the approaches do not segment a flame, but instead classify the entire image as background. In these cases, an accuracy of about $76\,\%$ is achieved.

We train our models with a frozen backbone in the scenarios one, three, and five. Our tests show that even though training with an entirely frozen backbone saves up to 50 % of the necessary training time on 100 % of the dataset, the resulting models cannot converge properly as often as the models of the scenarios with unfrozen backbones. This observation is especially true for the ImageNet-pre-trained approaches but less applicable for the SSL approaches. In this case, a tradeoff has to be accepted between training time and memory usage on one hand and accuracy on the other hand.

Goyal et al. (2021a) state that successful SSL approaches need to have two key ingredients: They have to include massive models and massive datasets. Both of these conditions are not met by our application. Nonetheless, we observe that the SSL approaches in our case converge more often than the models in the entirely supervised approaches pre-trained on ImageNet. Our results indicate that, although much fewer labels need to be provided compared to ImageNet pre-training, recent SSL approaches can learn meaningful representations even when a small dataset and a medium-sized model architecture are used.

## 5. CONCLUSIONS

In this paper, we focus on the evaluation of three recent SSL approaches for semantic segmentation of industrial burner flames. We show how self-supervision performs in a practical scenario where training labels and the opportunities for hyperparameter tuning are limited.

We demonstrate that the SSL approaches are more robust to sub-optimal learning rates, which leads to proper convergence more often. Therefore, average accuracies achieved by pre-training in a self-supervised manner are higher across all scenarios. This advantage is apparent when the amount of training labels are reduced.

In future work, we plan to use and improve on current self-supervision techniques to enable the automatic extraction of process-relevant parameters in industrial secondary combustion chambers and thereby contribute to optimizing industrial combustion processes regarding energy efficiency and $CO_2$ emission reduction.

## ACKNOWLEDGMENT

This research is funded by Invest BW on behalf of the Ministry of Economics, Labor and Tourism of Baden-Württemberg.

## References

Azizi, S., Mustafa, B., Ryan, F., Beaver, Z., Freyberg, J., Deaton, J., Loh, A., Karthikesalingam, A., Kornblith, S., Chen, T., Natarajan, V., Norouzi, M., 2021. Big Self-Supervised Models Advance Medical Image Classification.

Bengio, Y., 2012. Practical Recommendations for Gradient-Based Training of Deep Architectures. 437–478.

Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A., 2020. Unsupervised Learning of Visual Features by Contrasting Cluster Assignments. *arXiv Preprint arXiv:2006.09882*.

Chen, T., Kornblith, S., Norouzi, M., Hinton, G., 2020. A Simple Framework for Contrastive Learning of Visual Representations.

Cuturi, M., 2013. Sinkhorn Distances: Lightspeed Computation of Optimal Transport. 26.

Doersch, C., Gupta, A., Efros, A. A., 2015. Unsupervised Visual Representation Learning by Context Prediction. *Proceedings of the IEEE International Conference on Computer Vision*, 1422–1430.

Gidaris, S., Singh, P., Komodakis, N., 2018. Unsupervised Representation Learning by Predicting Image Rotations. *arXiv Preprint arXiv:1803.07728*.

Goyal, P., Caron, M., Lefaudeux, B., Xu, M., Wang, P., Pai, V., Singh, M., Liptchinsky, V., Misra, I., Joulin, A., Bojanowski, P., 2021a. Self-supervised Pretraining of Visual Features in the Wild.

Goyal, P., Duval, Q., Reizenstein, J., Leavitt, M., Xu, M., Lefaudeux, B., Singh, M., Reis, V., Caron, M., Bojanowski, P., Joulin, A., Misra, I., 2021b. VISSL.

Großkopf, J., Matthes, J., Vogelbacher, M., Waibel, P., 2021. Evaluation of Deep Learning-Based Segmentation Methods for Industrial Burner Flames. *Energies*, 14(6), 1716.

He, K., Fan, H., Wu, Y., Xie, S., Girshick, R., 2020. Momentum Contrast for Unsupervised Visual Representation Learning.

Jadon, S., 2020. A Survey of Loss Functions for Semantic Segmentation. *2020 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, 1–7.

Kingma, D. P., Ba, J., 2014. Adam: A Method for Stochastic Optimization. *arXiv Preprint arXiv:1412.6980*.

Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S., 2017. Feature Pyramid Networks for Object Detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2117–2125.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully Convolutional Networks for Semantic Segmentation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3431–3440.

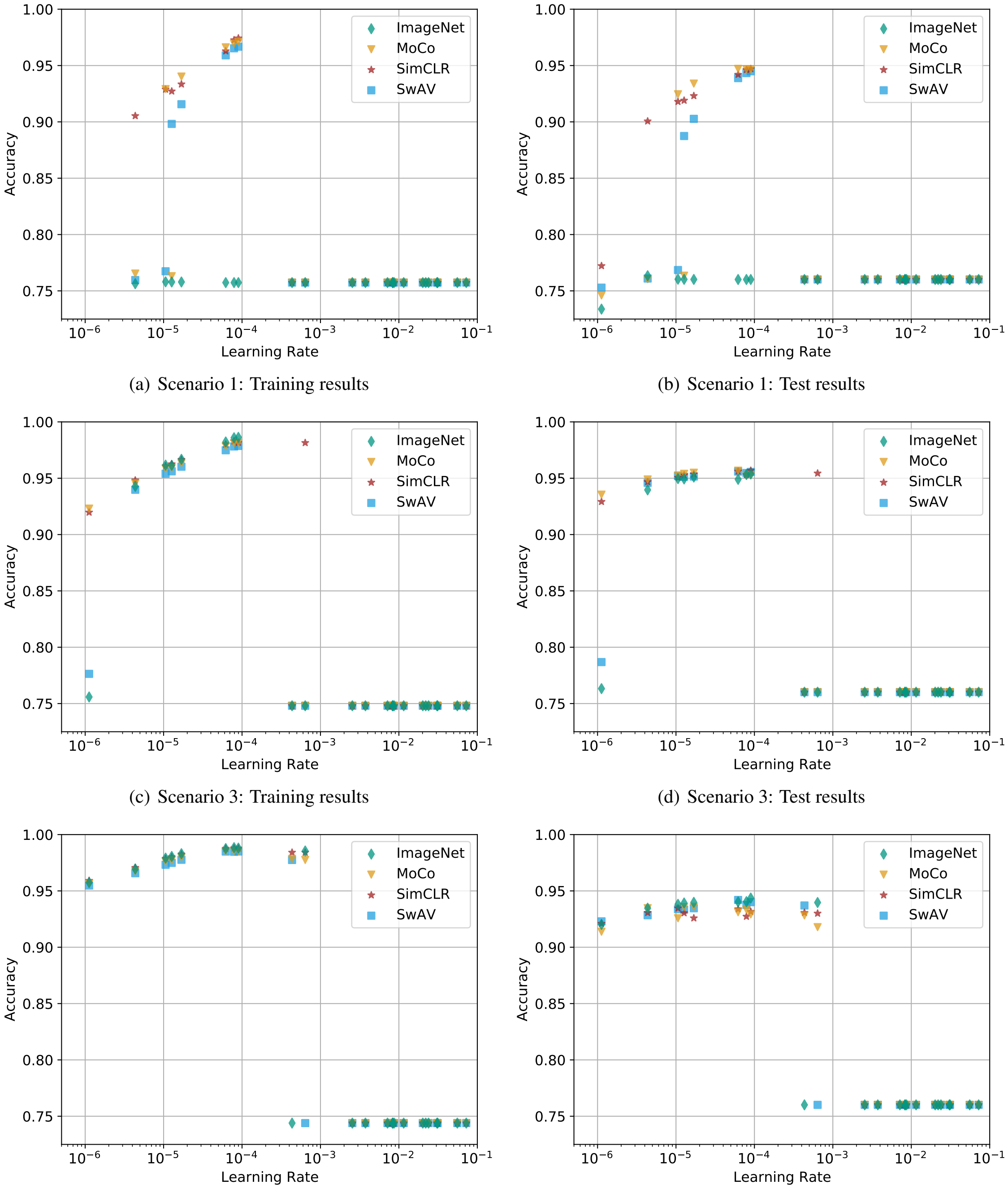
Minaee, S., Boykov, Y. Y., Porikli, F., Plaza, A. J., Kehtarnavaz, N., Terzopoulos, D., 2021. Image Segmentation Using Deep Learning: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Steger, C., Ulrich, M., Wiedemann, C., 2018. *Machine Vision Algorithms and Applications*. John Wiley & Sons.

Tomasev, N., Bica, I., McWilliams, B., Buesing, L., Pascanu, R., Blundell, C., Mitrovic, J., 2022. Pushing the Limits of Self-supervised ResNets: Can we Outperform Supervised Learning without Labels on ImageNet?

Zhang, R., Isola, P., Efros, A. A., 2016. Colorful Image Colorization. *European Conference on Computer Vision*, 649–666.

## APPENDIX



(a) Scenario 1: Training results

(b) Scenario 1: Test results

(c) Scenario 3: Training results

(d) Scenario 3: Test results

(e) Scenario 5: Training results

(f) Scenario 5: Test results

Figure 3. Fine-tuning results of 30 learning rates for each pre-training scenario of the FPN with a frozen backbone.