# EVALUATING CONVNET AND TRANSFORMER BASED SELF-SUPERVISED ALGORITHMS FOR BUILDING ROOF FORM CLASSIFICATION

G. Mutreja[1],*, K. Bittner[1]

[1]Remote Sensing Technology Institute, German Aerospace Center (DLR), Weßling, Germany –
(guneet.mutreja, ksenia.bittner)@dlr.de

**KEY WORDS:** Roof-form classification, Self-supervised learning, SimCLR, MoCo, ConvNets, Vision transformers, BYOL, BEiT

**ABSTRACT:**

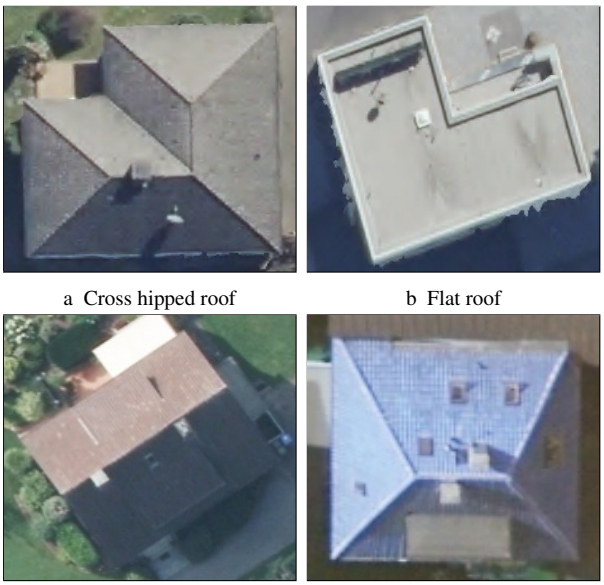
This research paper presents a comprehensive evaluation of various self-supervised learning models for building roof type classification. We conduct linear evaluation experiments for the models pretrained on both the ImageNet1K dataset and a custom building roof type dataset to assess the models' performance for the roof type classification task. The results demonstrate the effectiveness of the ViT-based BEiTV2 model, which outperforms other models on both datasets, achieving an accuracy of 96.8% from the model pretrained on ImageNet1K dataset and 92.67% on the model pretrained on building roof type dataset. The class activation maps further validate the strong performance of MoCoV3, BarlowTwins, and DenseCL models. These findings emphasize the potential of self-supervised learning for accurate building roof type classification, with the ViT-based BEiTV2 model showcasing state-of-the-art results.

## 1. INTRODUCTION

In the era of smart cities, the efficient planning of infrastructure, accurate forecasts, and timely disaster response have become crucial for sustainable urban development. The emergence of digital twins, enabled by advancements in 3D modeling, offers immense potential to support governments in addressing these challenges. One key element in the reconstruction and maintenance of digital twins is the identification of building roof types. Unfortunately, in many cases, the existing records maintained by local governments lack this critical information. This limitation has raised the demand for automated methods utilizing artificial intelligence (AI) and aerial/satellite imagery to generate building roof type information.

While supervised models in computer vision have achieved remarkable success in various tasks, building roof type classification presents unique challenges due to the scarcity of high-quality training data. Acquiring labeled data for diverse roof types at scale is a laborious and costly process. Hence, there is a need to explore alternative approaches that can leverage existing data without relying heavily on annotated labels. In recent years, unsupervised and self-supervised learning techniques ( Larsson et al. (2017), Gidaris et al. (2018), He et al. (2019), Caron et al. (2021), Zbontar et al. (2021), Saad et al. (2021)) have gained significant attention in the computer vision community. These methods eliminate the dependence on labeled training data and instead focus on leveraging the inherent structure and information present in the data itself. With recent innovations in this field, self-supervised models have demonstrated comparable accuracies, and in some cases, even surpassed their supervised counterparts on well-known classification benchmarks such as ImageNet and CIFAR-10.

Motivated by the success of self-supervised models and recognizing the importance of building roof type information, we aim to investigate the application of different convolutional and vision transformer-based self-supervised algorithms for the downstream task of building roof type classification. By exploiting the intrinsic patterns and relationships within the unlabeled aerial/satellite imagery data, we seek to develop an effective and efficient approach for automatically identifying and classifying building roof types.



a Cross hipped roof b Flat roof

c Gable roof d Hip roof

Figure 1. An example of roof forms from each of the four classes.

In this paper, we present our comprehensive evaluation of various self-supervised learning techniques and their effectiveness in addressing the challenge of building roof type classification. We conduct extensive experiments, compare the performance of different models. Through our research, we aim to advance the state-of-the-art in automated building roof type classification, facilitating the development of smarter cities with improved infrastructure planning, accurate forecasts, and enhanced disaster response capabilities.

* Corresponding author

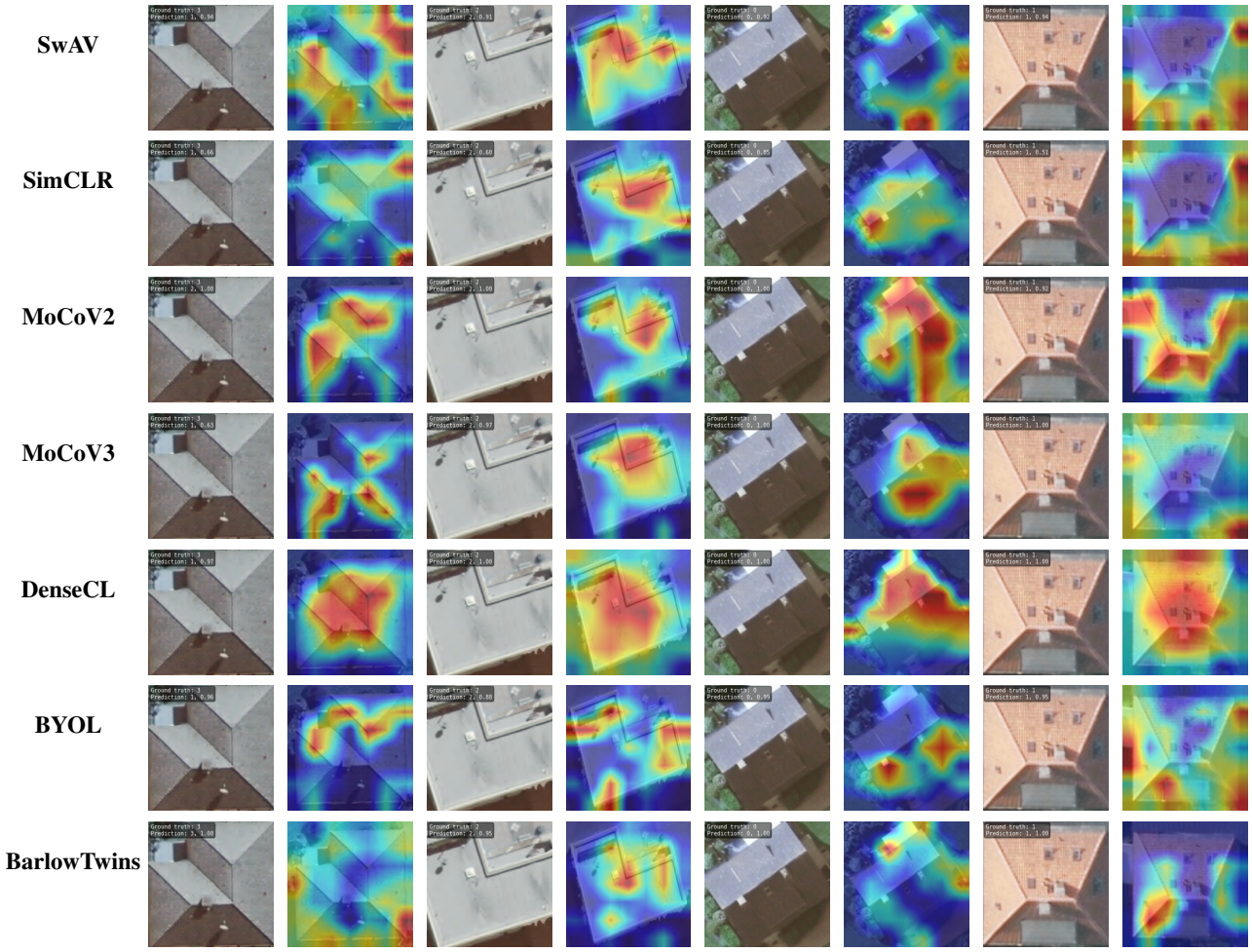



Figure 2. Predictions and class activation maps for different ResNet50 based architectures pretrained on ImageNet1k dataset, encompassing cross-hipped, flat, gable and hip roof types (left to right).

## 2. RELATED WORK

### 2.1 Self-Supervised Learning

Self-supervised learning has emerged as a promising approach to leverage unlabeled data for representation learning in computer vision tasks. In recent years, contrastive learning-based models have gained significant popularity, outperforming non-contrastive methods. Following the pivotal work of van den Oord et al. (2018) on contrastive predictive coding (CPC), several notable self-supervised algorithms have been proposed. Early approaches focused on tasks such as image inpainting and colorization to learn meaningful representations. Pathak et al. (2016) introduced Context Encoders, which leveraged the reconstruction of missing image regions for representation learning. Similarly, Zhang et al. (2016) proposed colorization as a pretext task for learning image representations.

Another line of research explored instance discrimination, where models are trained to differentiate between instances of the same object class. This idea was popularized by He et al. (2020) with the introduction of the MoCo framework. The MoCo approach utilized a contrastive loss to learn discriminative representations by contrasting positive and negative pairs of instances. Further advancements in self-supervised learning led to the development of Chen et al. (2020a). SimCLR introduced a contrastive learning framework that maximized agreement between differently augmented views of the same image. It achieved state-of-the-art performance on several benchmark datasets and demonstrated the efficacy of contrastive learning on large-scale datasets. Building upon SimCLR, the BYOL (Bring Your Own Latent) framework proposed by Grill et al. (2020) eliminated the need for negative samples during training. BYOL achieved competitive results and showcased the potential of self-supervised learning without contrastive negative pairs. Recent advancements have extended self-supervised learning beyond convolutional architectures to transformer-based models. Notably, the BeiT (BERT-like Encoder with Transformer) framework proposed by Bao et al. (2021) demonstrated the effectiveness of self-supervised learning for vision transformers. BeiT achieved remarkable performance on ImageNet, surpassing previous state-of-the-art models.

While the above-mentioned algorithms represent a subset of the extensive research in self-supervised learning, they highlight the evolution and success of different approaches. In our study, we draw inspiration from these models to explore their applicability in the specific task of building roof type classification.

### 2.2 Building Roof Type Classification

Building roof type classification plays a vital role in the reconstruction and maintenance of digital twins for smart cities. Previous works like in Alidoost Fatemeh (2018), Buyukdemircio-

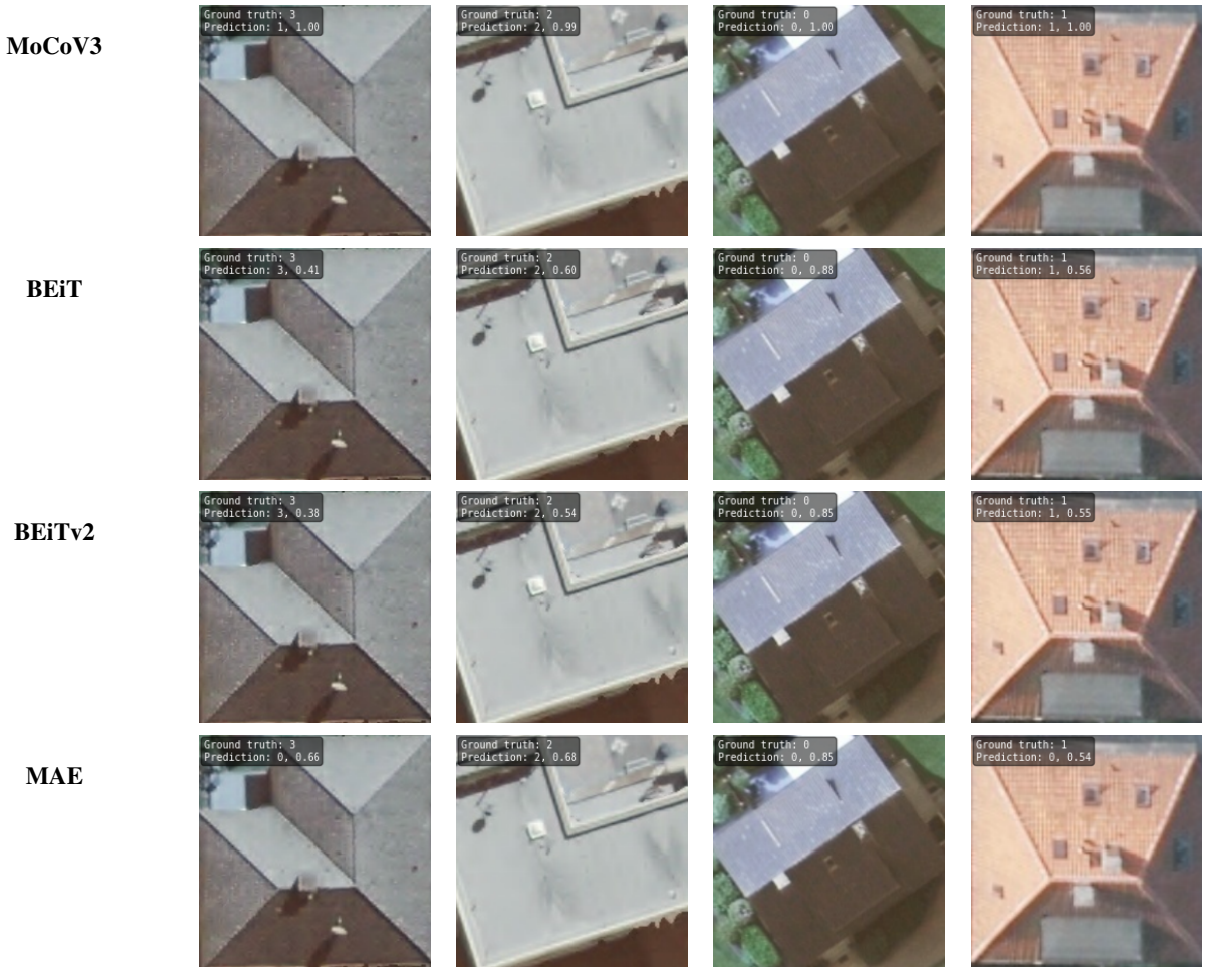



Figure 3. Predictions of different ViT based architectures pretrained on ImageNet1K dataset, encompassing cross-hipped, flat, gable and hip roof types (left to right).

glu et al. (2021), and Wang et al. (2022) focused on utilizing Convolutional Neural Networks (CNNs) for detection and recognition of roof shapes using labeled training data. These methods demonstrated effectiveness with high-quality rates in detection and recognition. However, the reliance on limited and high-quality training data, as well as challenges in accurate segmentation, remained as notable limitations. While Saad et al. (2021)'s work investigated the application of contrastive learning methods like SimCLR and BYOL to mitigate the challenge of limited labeled training data, it is important to consider the potential of more accurate self-supervised convolutional and vision transformer-based algorithms for this task. The advancements in self-supervised techniques offer an opportunity to improve the performance of building roof type classification, and thus, it becomes essential to evaluate their effectiveness in this context.

## 3. METHODOLOGY

### 3.1 Pretraining

Pretraining serves as a crucial step in self-supervised learning to learn effective representations from unlabeled data. In our study, we adopted pretrained backbone weights on the ImageNet1K dataset for multiple ResNet (He et al., 2015) and ViT-based (Dosovitskiy et al., 2020) self-supervised models. These pretrained models were chosen due to their demonstrated effectiveness in capturing high-level visual features.

By utilizing pretrained weights, we aim to leverage the generalization capabilities of these models and evaluate their effectiveness on aerial imagery and building roof type classification. The pretrained models provide a strong initial feature extractor, enabling us to transfer knowledge from the large-scale ImageNet dataset to our specific task.

### 3.2 Linear Evaluation

The linear evaluation phase focuses on training a linear classifier on top of the pretrained backbone weights, allowing us to assess the models' performance specifically for building roof type classification. During the linear evaluation, we froze the weights of the pretrained backbone and trained only the linear classifier head. This approach allowed us to fine-tune the classifier specifically for the task of building roof type classification, while keeping the learned features from the pretrained models intact. The linear evaluation enabled us to assess the effectiveness of the pretrained models in differentiating between the main roof types: Gable, Hip, Flat, and Cross-Hipped.

In the following sections, we present the experimental setup, including details on the dataset, model architectures, training procedures, and evaluation metrics. By combining the power

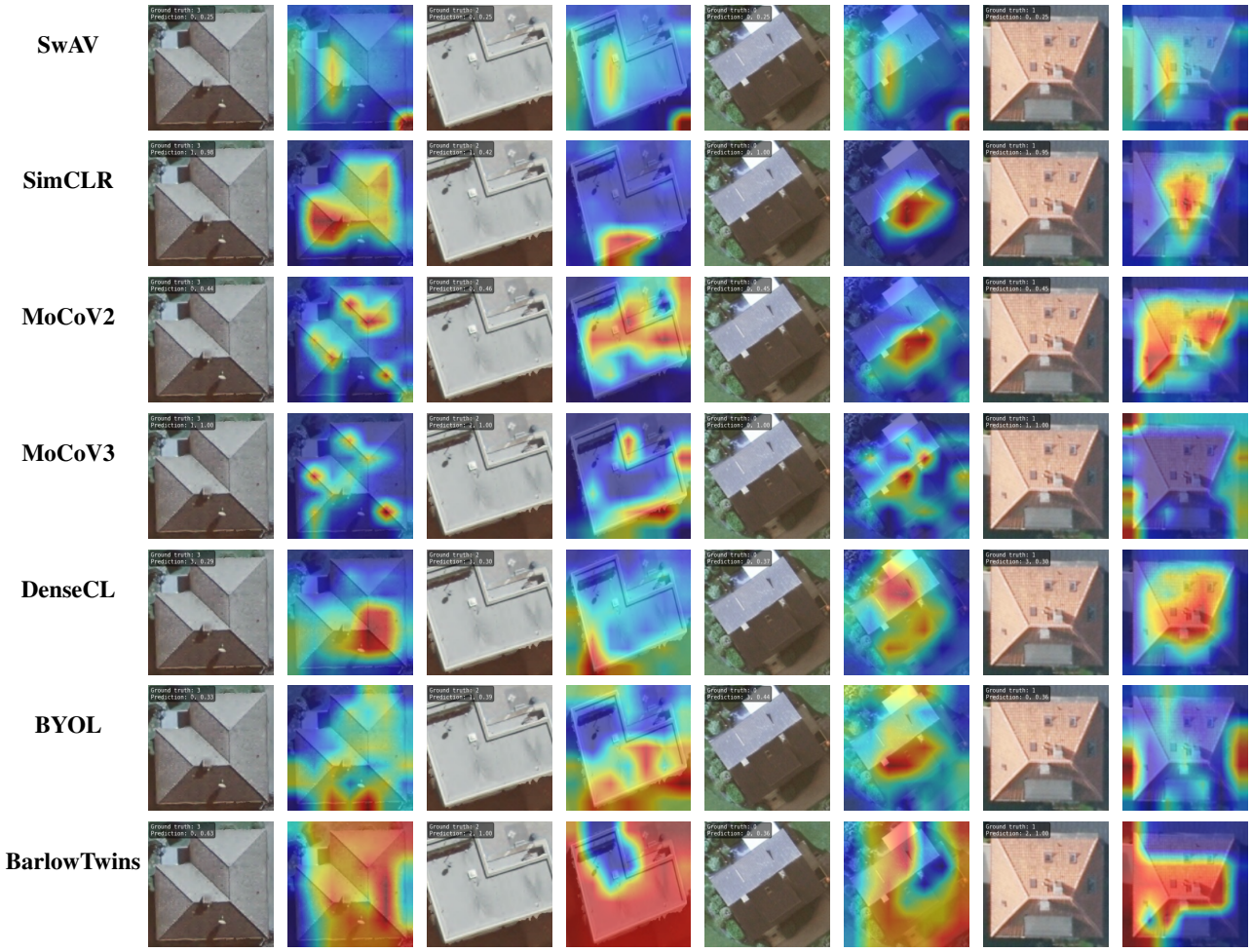



Figure 4. Predictions and class activation maps for different ResNet50 based architectures pretrained on Building roof type dataset, encompassing cross-hipped, flat, gable and hip roof types (left to right).

of pretraining and linear evaluation, our methodology seeks to provide insights into the effectiveness of self-supervised learning models for building roof type classification.

## 4. EXPERIMENTS

### 4.1 Dataset

For our experiments, we conducted linear evaluation using building roof vectorization dataset published by Hensel et al. (2021). However, the dataset lacked roof type information. Therefore, we manually assigned labels to images, categorizing them into four distinct roof type classes: Gable, Hip, Flat, and Cross-Hipped. To provide a visual representation, Figure 1 showcases an example image for each of these four classes. To maintain consistency, we preserved the original train-test split provided in the dataset. This resulted in a training set comprising approximately 7500 images and a validation set containing 765 images.

### 4.2 Implementation Details

In our research, we conducted linear evaluation experiments on various self-supervised algorithms. Specifically, we evaluated SimCLR (Chen et al., 2020a), SwAV (Caron et al., 2020), MoCoV2 ( Chen et al. (2020b), MoCoV3 (Chen et al., 2021), DenseCL (Wang et al., 2020), BYOL (Grill et al., 2020), BarlowTwins (Zbontar et al., 2021) with ResNet50, and BEiT (Bao et al., 2021), BEiTV2 (Peng et al., 2022), and MAE (He et al., 2021) with Vision Transformers as backbones.

We performed two sets of experiments. Firstly, we trained a linear classifier directly on the frozen weights of the pretrained models, which were pretrained on the ImageNet1k dataset. In the second experiment, we pretrained the backbone weights from scratch on the building dataset without labels for 300 epochs. Subsequently, we trained a linear classifier in a supervised fashion using the frozen weights for 90 epochs.

During the linear evaluation, we applied random resize crops and horizontal flips as data augmentation techniques. We evaluated the accuracy using a central crop for all models. For SimCLR, SwAV, BYOL, MoCoV3 (ResNet-based), and BarlowTwins, we utilized the LARS optimizer (You et al., 2017) with a learning rate of 0.6 and a momentum of 0.9. We gradually decreased the learning rate using a cosine schedule that started from the first epoch and continued for all 90 epochs. For MoCoV2 and DenseCL, we employed Stochastic Gradient Descent (SGD) as the optimizer with a learning rate of 0.1, momentum of 0.9, and weight decay of $1e-4$. We decayed the learning rate by a factor of 0.1 with milestones at 60 and 80 epochs using a multi-step strategy. For BEiT, BEiTV2, and MAE, we used the AdamW optimizer with a weight decay of

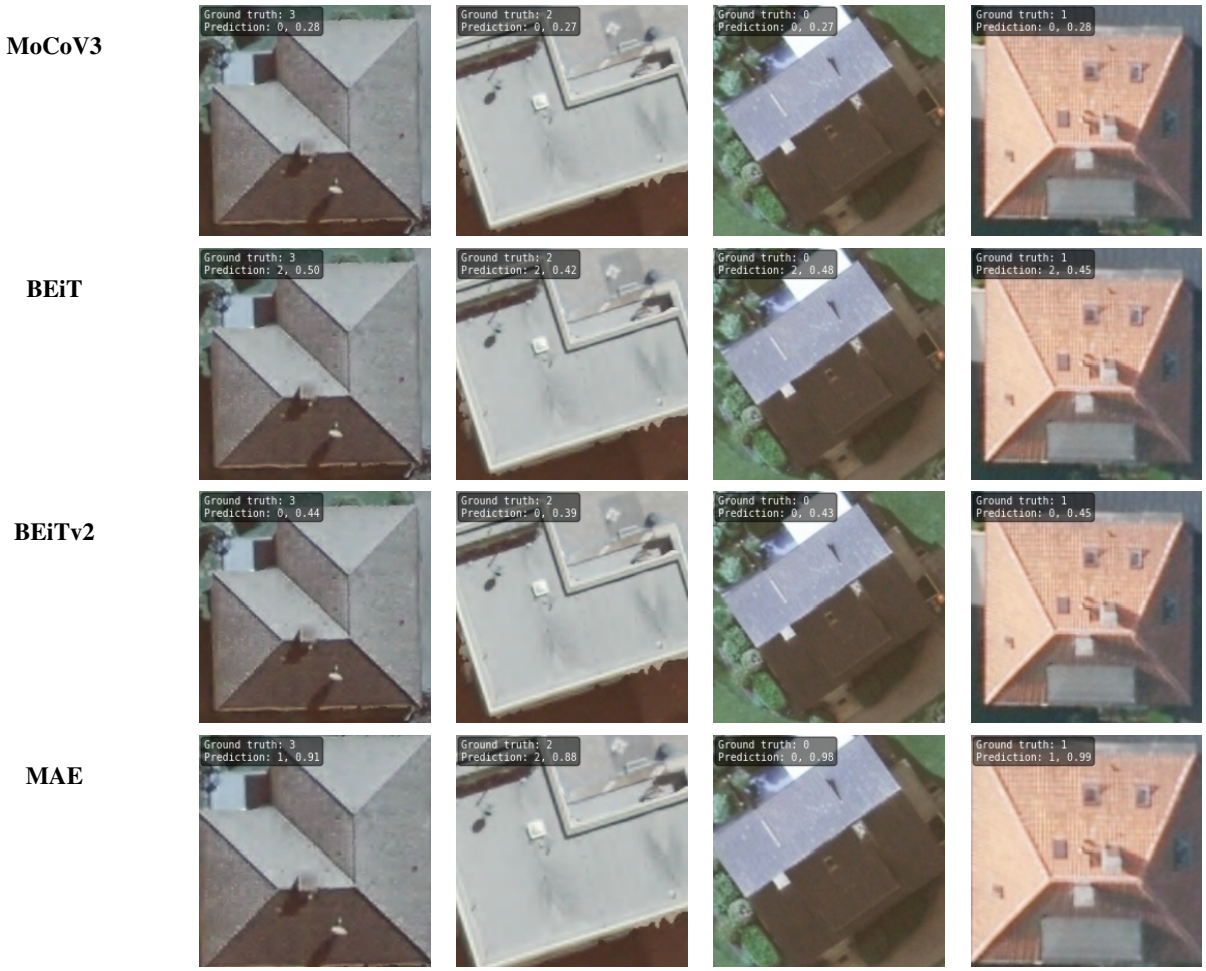



Figure 5. Predictions of different ViT based architectures pretrained on Buildings roof type dataset, encompassing cross-hipped, flat, gable and hip roof types (left to right).

0.05. The initial learning rate was determined using a linear scaling rule, where lr was $1e-3$, gradually increasing it during the first 20 epochs. We decayed the learning rate using a cosine schedule.

During the pretraining phase, we applied data augmentation techniques as prescribed in the original papers. For SimCLR, SwAV, BYOL, MoCoV3 (ResNet-based), and BarlowTwins, we utilized the LARS optimizer with a linear scaling rule to determine the initial learning rate, gradually increasing it during the first 10 epochs. We decayed the learning rate using a cosine schedule and applied a weight decay of $1e-6$. For MoCoV2 and DenseCL, we employed Stochastic Gradient Descent (SGD) as the optimizer with a learning rate of 0.03, momentum of 0.9, and weight decay of $1e-4$. We gradually decreased the learning rate using a cosine schedule. For BEiT, BEiTV2, and MAE, we used the AdamW optimizer with a weight decay of 0.1. The initial learning rate was set to $1e-4$, with a warmup of 40 epochs. We decayed the learning rate using a cosine schedule. All models were pretrained for 300 epochs to capture rich and meaningful representations from the unlabeled building dataset.

## 5. RESULTS AND DISCUSSION

In this section, we present the results and discussion from our experiments, comparing the performance of the mentioned self-supervised learning models for the task of building roof type classification.

**Linear evaluation results for models pretrained on ImageNet1k dataset.** We performed a linear evaluation for all the mentioned self-supervised learning models pretrained on the ImageNet1K dataset. First column in Table 1 summarizes the accuracy results obtained from this evaluation. Among these models, the ViT-based model BEiTV2 demonstrated superior performance, achieving an accuracy of 96.8%. This indicates the effectiveness of Vision Transformers for building roof type classification. Other notable performers include ResNet50 based MoCoV3, DenseCL, and BarlowTwins, which achieved accuracy scores exceeding 95%. The class activation maps in Figure 2 provide visual evidence supporting the performance of MoCoV3, BarlowTwins, and DenseCL. These maps highlight the model's ability to focus on relevant image regions associated with specific roof types, reinforcing their accuracy and robustness.

**Linear evaluation results for models pretrained on buildings roof type dataset.** Next, we conducted a linear evaluation on the models pretrained without labels using the building roof types dataset. Second column in Table 1 summarizes the accuracy results obtained from this evaluation. In this evaluation scenario, the BEiT model achieved the highest accuracy of 96%, closely followed by MoCoV3 ViT with an accuracy of

| Model | Results ImageNet1k pretrained | Results Building dataset pretrained |
|---|---|---|
| SimCLR | 80.78 | 87.98 |
| SwAV | 87.9 | 86.5 |
| MoCoV2 | 71.63 | 85.62 |
| MoCoV3 (ResNet50) | 95.8 | 86.2 |
| MoCoV3 (ViT) | 92.9 | 92.4 |
| DenseCL | 94.1 | 83.2 |
| BYOL | 92.1 | 84.05 |
| BEit | 96.4 | 95.94 |
| BEitV2 | 96.8 | 92.67 |
| BarlowTwins | 95.03 | 59.3 |
| MAE | 77.1 | 88.4 |

Table 1. Linear evaluation results from different ResNet50 and ViT based self-supervised models pretrained on ImageNet1k and Building roof type datasets.

92.41%. These results demonstrate the robustness and transferability of these models to the specific task of building roof type classification. However, it is worth noting that some models, such as BarlowTwins, exhibited lower accuracy in this evaluation.

To provide further insights into the performance of these models, we showcase predictions from these models for one image each from the four different roof classes in the validation dataset and also the class activation maps in Figure 2 to Figure 5, which offer visual explanations of the model's attention to specific regions. These visualizations will provide additional evidence of the models' ability to capture meaningful features related to building roof types.

In summary, our experiments demonstrate the effectiveness of self-supervised learning models for building roof type classification. The ViT-based model BEiTV2, along with other high-performing models such as MoCoV3, DenseCL, and BarlowTwins, showcase the potential of self-supervised learning for capturing meaningful representations. The class activation maps and the accuracy results further support the efficacy of these models in accurately classifying building roof types.

## 6. CONCLUSION

In this study, we investigated the effectiveness of various self-supervised learning models for building roof type classification. Through extensive experiments and evaluations, we obtained insightful results that contribute to the understanding and advancement of self-supervised learning in computer vision tasks.

Our findings highlight the potential of self-supervised learning models for building roof type classification and demonstrate the efficacy of Vision Transformers in this task. The results also emphasize the importance of leveraging large-scale unlabeled datasets and transferring learned representations to achieve high accuracy in real-world applications. The performance of various models, coupled with visual evidence from class activation maps, reinforces the importance of leveraging self-supervised learning algorithms and their ability to capture meaningful features. These results contribute to advancing computer vision research in building reconstruction and provide valuable insights for intelligent infrastructure planning in smart cities.

## References


Alidoost Fatemeh, A. H., 2018. A CNN-Based Approach for Automatic Building Detection and Recognition of Roof Types Using a Single Aerial Image. *Journal of Photogrammetry, Remote Sensing and Geoinformation Science*.

Bao, H., Dong, L., Wei, F., 2021. BEiT: BERT Pre-Training of Image Transformers. *CoRR*, abs/2106.08254.

Buyukdemircioglu, M., Can, R., Kocaman, S., 2021. Deep Learning Based Roof Type Classification Using Very High Resolution Aerial Imagery. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLIII-B3-2021, 55-60.

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

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A., 2021. Emerging Properties in Self-Supervised Vision Transformers. *CoRR*, abs/2104.14294.

Chen, T., Kornblith, S., Norouzi, M., Hinton, G. E., 2020a. A Simple Framework for Contrastive Learning of Visual Representations. *CoRR*, abs/2002.05709.

Chen, X., Fan, H., Girshick, R. B., He, K., 2020b. Improved Baselines with Momentum Contrastive Learning. *CoRR*, abs/2003.04297.

Chen, X., Xie, S., He, K., 2021. An Empirical Study of Training Self-Supervised Vision Transformers. *CoRR*, abs/2104.02057.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N., 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *CoRR*, abs/2010.11929.

Gidaris, S., Singh, P., Komodakis, N., 2018. Unsupervised Representation Learning by Predicting Image Rotations. *CoRR*, abs/1803.07728.

Grill, J., Strub, F., Altché, F., Tallec, C., Richemond, P. H., Buchatskaya, E., Doersch, C., Pires, B. Á., Guo, Z. D., Azar, M. G., Piot, B., Kavukcuoglu, K., Munos, R., Valko, M., 2020. Bootstrap Your Own Latent: A New Approach to Self-Supervised Learning. *CoRR*, abs/2006.07733.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R. B., 2021. Masked Autoencoders Are Scalable Vision Learners. *CoRR*, abs/2111.06377.

He, K., Fan, H., Wu, Y., Xie, S., Girshick, R., 2020. Momentum contrast for unsupervised visual representation learning. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9726–9735.

He, K., Fan, H., Wu, Y., Xie, S., Girshick, R. B., 2019. Momentum Contrast for Unsupervised Visual Representation Learning. *CoRR*, abs/1911.05722.

He, K., Zhang, X., Ren, S., Sun, J., 2015. Deep Residual Learning for Image Recognition. *CoRR*, abs/1512.03385.

Hensel, S., Goebbels, S., Kada, M., 2021. Building Roof Vectorization with Ppgnet. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 46W4, 85-90.

Larsson, G., Maire, M., Shakhnarovich, G., 2017. Colorization as a Proxy Task for Visual Understanding. *CoRR*, abs/1703.04044.

Pathak, D., Krähenbühl, P., Donahue, J., Darrell, T., Efros, A. A., 2016. Context Encoders: Feature Learning by Inpainting. *CoRR*, abs/1604.07379.

Peng, Z., Dong, L., Bao, H., Ye, Q., Wei, F., 2022. Beit v2: Masked image modeling with vector-quantized visual tokenizers.

Saad, A. B., Drouyer, S., Hell, B., Gavoille, S., Gaiffas, S., Facciolo, G., 2021. A review on contrastive learning methods and applications to roof-type classification on aerial images. *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, 4960–4963.

van den Oord, A., Li, Y., Vinyals, O., 2018. Representation Learning with Contrastive Predictive Coding. *CoRR*, abs/1807.03748.

Wang, X., Zhang, R., Shen, C., Kong, T., Li, L., 2020. Dense Contrastive Learning for Self-Supervised Visual Pre-Training. *CoRR*, abs/2011.09157.

Wang, Y., Li, S., Teng, F., Lin, Y., Wang, M., Cai, H., 2022. Improved Mask R-CNN for Rural Building Roof Type Recognition from UAV High-Resolution Images: A Case Study in Hunan Province, China. *Remote Sensing*, 14(2).

You, Y., Gitman, I., Ginsburg, B., 2017. Scaling SGD Batch Size to 32K for ImageNet Training. *CoRR*, abs/1708.03888.

Zbontar, J., Jing, L., Misra, I., LeCun, Y., Deny, S., 2021. Barlow Twins: Self-Supervised Learning via Redundancy Reduction. *CoRR*, abs/2103.03230.

Zhang, R., Isola, P., Efros, A. A., 2016. Colorful Image Colorization. *CoRR*, abs/1603.08511.