# Few-Shot Segmentation of Historical Maps via Linear Probing of Vision Foundation Models

Rafael Sterzinger, Marco Peer, and Robert Sablatnig

Computer Vision Lab, TU Wien, Vienna, AUT
{firstname.lastname}@tuwien.ac.at

**Abstract.** As rich sources of history, maps provide crucial insights into historical changes, yet their diverse visual representations and limited annotated data pose significant challenges for automated processing. We propose a simple yet effective approach for few-shot segmentation of historical maps, leveraging the rich semantic embeddings of large vision foundation models combined with parameter-efficient fine-tuning. Our method outperforms the state-of-the-art on the Siegfried benchmark dataset in vineyard and railway segmentation, achieving +5% and +13% relative improvements in mIoU in 10-shot scenarios and around +20% in the more challenging 5-shot setting. Additionally, it demonstrates strong performance on the ICDAR 2021 competition dataset, attaining a mean PQ of 67.3% for building block segmentation, despite not being optimized for this shape-sensitive metric, underscoring its generalizability. Notably, our approach maintains high performance even in extremely low-data regimes (10- & 5-shot), while requiring only 689k trainable parameters – just 0.21% of the total model size. Our approach enables precise segmentation of diverse historical maps while drastically reducing the need for manual annotations, advancing automated processing and analysis in the field. Our implementation is publicly available at: https://github.com/RafaelSterzinger/few-shot-map-segmentation.

**Keywords:** Few-Shot Learning· Low-Rank Adaptation· Semantic Segmentation· Foundation Models· Historical Maps · Historical Documents

## 1 Introduction

As invaluable records of the past, historical maps provide critical insights into geographic, infrastructural, and sociopolitical changes. However, their automated analysis remains a significant challenge due to the wide stylistic variability, inconsistent annotations, and frequent physical degradation inherent in these documents [22,33]. In contrast to modern maps, which often conform to standardized cartographic norms, historical maps are profoundly heterogeneous, both in visual appearance and semantic content, as illustrated in Figure 1.

As a result, models struggle to generalize effectively due to the domain gap between maps, making it difficult to transfer knowledge from one corpus to

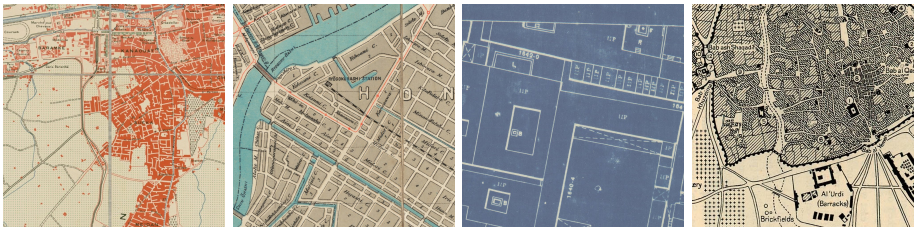


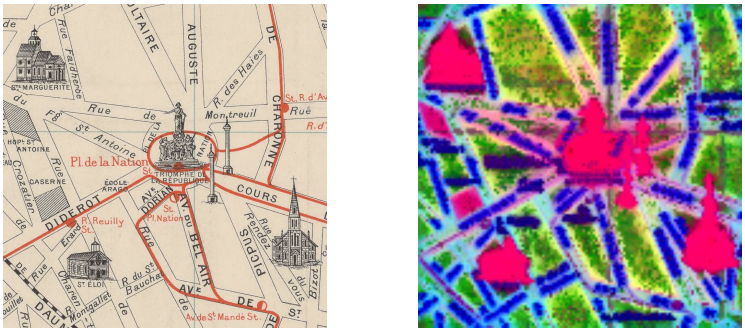
Fig. 1: An excerpt of historical city maps from around the world, published between 1720 and 1950. ©Petitpierre et al.

another [22]. Another key obstacle in developing robust models for automated historical map segmentation is the scarcity of labeled data, a substantial problem that extends beyond historical maps: as annotating historical artifacts demands expert knowledge and considerable manual effort, which renders the task cost-ineffective, it is impractical to build the large training corpora typically required for supervised learning [27,28].

As a consequence of these challenges, historical map segmentation presents a promising, yet largely untapped, application for few-shot learning, a paradigm which enables models to learn from limited labeled data by leveraging pretrained representations of large vision foundation models [26,7,11,30]. Over the years, numerous vision foundation models have emerged that are trained on vast corpora of natural images to serve as powerful backbones for various downstream tasks across diverse domains [23]. When processing images, these models can generate rich features, which are particularly beneficial for dense prediction tasks such as segmentation, sometimes even for domains such as historical maps. As illustrated in Figure 2, our visualization shows the first three principal components of embeddings obtained from RADIO [23], which already capture meaningful semantic structures, offering a strong basis for downstream tasks.

Ideally, these expressive representations can bridge the gap across stylistic differences and deliver satisfactory results even in low-data regimes. However, while often pretrained on natural images, historical maps present a unique challenge: their structural complexity means features such as roads, railways, and river patterns can appear visually similar, although semantically different. Based on this observation, we necessitate that the models need to learn the abstract distinctions between these elements, highlighting the crucial need for model adaptation, for example, through parameter-efficient fine-tuning [33].

Given these challenges – the heterogeneity of historical maps, the scarcity of annotated data, and the domain gap between natural images and historical documents – Xia et al. [33] explored leveraging SAM [15] and its representations for improved few-shot map segmentation. We extend their work and investigate other types of foundation models such as DINOv2 [20] or RADIO [23], which have fundamentally different training goals that yield richer and more general spatial representations.




(a) Original (b) PCA Radio

Fig. 2: An illustration of the first three principal components of RADIO-H [23] feature embeddings of a map of Paris: despite no prior training in this specialized domain, meaningful classes have already emerged: landmarks, building blocks, streets, and street names are clearly distinguishable.

Although our approach builds on MapSAM by Xia et al. [33], such as leveraging vision foundation models and parameter-efficient fine-tuning, its strength lies in its reduced complexity. Unlike MapSAM, which retrains the full SAM decoder, we use a lightweight linear classifier – yet achieve significantly better results. Our approach exemplifies *Occam's Razor*: when a simpler method outperforms more complex ones, added complexity becomes not only unnecessary but even detrimental, as more parameters increase the risk of overfitting and reduce generalization.

In summary, our contributions include:

- We provide the first comprehensive evaluation of three SOTA vision foundation models for historical map segmentation, delivering both qualitative visualizations and quantitative analysis of their feature embeddings' effectiveness in this specialized domain.
- We demonstrate that combining low-rank adaptation with linear probing of foundation models yields exceptional segmentation performance while maintaining parameter efficiency. Our extensive ablation studies validate this approach as both computationally lightweight and highly effective for historical document analysis.
- We perform an extensive evaluation on two datasets, the Siegfried and the ICDAR 2021 competition datasets. On the Siegfried dataset, we substantially outperform in all few-shot settings, especially in the challenging 5-shot setting with relative improvements of around 20% in IoU, while requiring training of only 689k parameters (a mere 0.21% of the overall model size).
- To facilitate future research and practical applications in historical map digitization, we release our code at: https://github.com/RafaelSterzinger/few-shot-map-segmentation.

## 2 Related Work

Motivated by the limited generalization of models trained on homogeneous or single-map datasets, Petitpierre et al. [22] explore training on two stylistically diverse corpora: Paris city maps and global city maps. By evaluating CNN-based segmentation performance, they show that neural networks trained on large, diverse corpora can integrate abstract reasoning (e.g., morphology, topology, semantic hierarchy) and remain robust despite stylistic variation. Although a large and diverse corpus improves generalization, annotated data remains limited, and geographic or stylistic biases can hinder transfer to truly unseen or underrepresented map styles. To mitigate this, Xia et al. [32] introduce a contrastive pretraining strategy for Transformers, leveraging image pairs of the same location from different historical map series. To further reduce annotation requirements, in a subsequent work, they introduce MapSAM [33], to leverage the strong zero-shot segmentation capabilities of SAM [15] for historical map segmentation. By integrating domain-specific knowledge via DoRA [19] into the image encoder, automating prompt generation, and enhancing both positional prompts and the attention mechanisms, their approach improves effective automatic segmentation of historical maps.

MapSAM is built upon two key pillars of modern computer vision: large-scale, pre-trained vision foundation models, which provide powerful, general-purpose feature representations and the ability to adapt these massive models to new domains efficiently. As mentioned, one prominent example is SAM [15], which has demonstrated remarkable generalization capabilities across various segmentation tasks. SAM is optimized for instance-level distinctions to segment clearly defined, prompt-specific objects based on supervised mask annotations [15]. An alternative to leveraging SAM's embeddings is the features from DINOv2 [20], which learns rich semantic representations by enforcing consistency across different views of the same scene using a self-supervised objective, without relying on labeled data. A model that combines SAM, DINOv2, and other foundation models is RADIO [23] by Ranzinger et al., who use agglomeration learning to distill multiple foundation models into a single one, yielding even more expressive feature representations.

Concurrently with the development of large vision models, there has also been growing interest in adapting specifically SAM for few-shot segmentation: He et al. [11] proposed APSeg with domain-agnostic feature transformation for cross-domain applications. Sun et al. [30] developed VRP-SAM to utilize annotated reference images as segmentation prompts with various annotation formats. Moreover, Liu et al. [18] as well as Zhang et al. [37], propose similar training-free prototype-based methodologies to extend SAM for few-shot segmentation.

As underscored by Xia et al. [33] for effectively applying large foundation models, with their general-purpose features, to highly specialized tasks such as historical map segmentation, adaptation is crucial. However, due to the sheer size of these models, often containing billions of parameters, full fine-tuning becomes increasingly impractical. In order to be able to adapt models of this size efficiently, Hu et al. [12] introduced LoRA, which initiated a new research direc-

tion focused entirely on efficiently fine-tuning large foundation models through low-rank decomposition. Further extensions like DoRA [19] improve representational capacity without significantly increasing the number of parameters by decomposing pre-trained weights into magnitude and direction, applying LoRA specifically to the directional component to reduce the number of trainable parameters during fine-tuning.

Apart from these advancements, few-shot segmentation for historical documents has fostered an active research community – though with a stronger focus on pixel-precise layout analysis of handwritten manuscripts than on historical maps: De Nardin et al. [5] propose an efficient learning-based method combined with classical binarization techniques, requiring only two labeled pages per manuscript. Subsequently, they demonstrate that even a single annotated page can suffice when paired with lightweight augmentations and balanced loss functions [6]. Architecture-wise, all three rely on smaller and simpler models such as DeepLabv3+ [2] or its predecessor, rather than adapting large foundation models.

## 3 Methodology

Our methodology follows a three-stage approach to adapt a vision foundation model with only a handful of labels to the specialized domain of historical maps: first, we extract semantically rich image embeddings from a vision foundation model such as SAM [15], DINOv2 [20], or RADIO [23]; second, we upscale the embeddings back to the original size and, third, use a linear classifier at the pixel level to obtain logit masks. Given that these foundation models have been trained predominantly on natural images [15], we incorporate low-rank adaptation techniques to enhance predictive performance further.

### 3.1 Extracting Image Embeddings

In this work, we primarily focus on three different foundation models: DINOv2 and RADIO, which have been specifically trained for image feature extraction, as well as SAM, a model that has explicitly been trained for open-vocabulary instance segmentation. For the latter, we extract solely the image encoder from the segmentation framework and discard the mask decoder head, as opposed to previous work. All three of these foundation models use the Vision Transformer (ViT) as their base architecture, which has been introduced by Dosovitskiy et al. [8], and comes in three different variants: ViT-Base, ViT-Large, and ViT-Huge, with parameter counts ranging from 86 million to 632 million. While DINOv2 and RADIO use standard ViTs, SAM employs ViTDet [16] as a means to reduce computational and memory complexity, which is specific to SAM as it enforces a fixed input size of $1024 \times 1024$ pixels to enable pixel-precision segmentation at the cost of requiring more compute. DINOv2 and RADIO support arbitrary resolutions and aspect ratios [23].

Given an input image $\mathbf{x} \in \mathbb{R}^{H \times W \times C}$, the image is first preprocessed with the vision foundation model $\mathcal{F}_\theta$ specific preprocessor and subsequently converted into a sequence of token embeddings, which consists of dividing the image into non-overlapping patches of size $P \times P$, resulting in a sequence of flattened patches:

$$\mathbf{x}_p \in \mathbb{R}^{N \times (P^2 \times C)}, \tag{1}$$

where $N = HW/P^2$ is the number of patches [8]. Next, patches are linearly projected into a $D$-dimensional latent space using a trainable embedding layer:

$$\mathbf{z}_0 = [\mathbf{x}_{\text{class}}; \mathbf{x}_p^1\mathbf{E}; \mathbf{x}_p^2\mathbf{E}; \ldots; \mathbf{x}_p^N\mathbf{E}] + \mathbf{E}_{\text{pos}}, \tag{2}$$

where $\mathbf{E} \in \mathbb{R}^{(P^2 \times C) \times D}$ is the projection matrix, and $\mathbf{E}_{\text{pos}} \in \mathbb{R}^{(N+1) \times D}$ represents a learnable positional encoding. Additionally, a special token, $\mathbf{x}_{\text{class}}$, is prepended to the sequence, which serves as a global image embedding.

Next, the token sequence is processed by a Transformer encoder (cf. [31]) consisting of $L$ layers (24 for ViT-L) of Multi-Head Self-Attention (MSA) and feed-forward networks, with residual connections and layer normalization applied at each step, yielding a final representation: $\mathbf{z}_L$. At position $\mathbf{z}_L^0$ is the special token, now with global context information. Although important for downstream tasks such as classification, this token is discarded in our setting, where we require solely spatial features. After reshaping the output, we end up with strong spatial feature representations in the form of:

$$\mathbf{z} = \mathcal{F}_\theta(\mathbf{x}_p), \quad \mathbf{z} \in \mathbb{R}^{H' \times W' \times D}, \tag{3}$$

where $H' = H/P$ and $W' = W/P$ correspond to the number of patches along each spatial dimension, which can be used to perform segmentation or other downstream tasks such as dense feature matching [20]. When visualizing these embeddings, e.g., by applying principal component analysis and plotting the first three components, clear semantic classes emerge (cf. Fig. 2), even for specialized data such as historical maps.

### 3.2 Linear-Probing

Next, keeping the image encoder $\mathcal{F}_\theta$ frozen, we map the embeddings to soft mask prediction. To map from $\mathbf{z}$ to a soft mask output which can be compared to a given binary annotation mask $\mathbf{m} \in \mathbb{B}^{H \times W}$, we first perform bilinear up-sampling to resize the embeddings back to their original size of $H \times W$ and then perform linear pixel-wise classification via $f$ on each feature vector $\mathbf{z}_{i,j}$:

$$f(\mathbf{z}_{i,j}) = \sigma\Big(\mathbf{z}_{i,j}^T\mathbf{w} + b\Big) = \hat{\mathbf{m}}_{i,j}, \quad \mathbf{w} \in \mathbb{R}^D, \quad b \in \mathbb{R}, \tag{4}$$

where $\mathbf{w}, b$ are learned parameters, and $\sigma(\cdot)$ is the sigmoid function, mapping the logit to a probability. In our experiments, we found that performing the pixel classification step after the up-sampling is crucial, not to interpolate logits but feature embeddings to obtain more accurate masks.

We explored more complex classifiers, such as the decoder head of the MaskFormer [4] or learned transpose convolutions as well. However, in our few-shot setting with limited training examples, high-capacity models tended to overfit, resulting in poor generalization. Thus, linear probing emerges as a simple yet effective and robust solution.

### 3.3 Fine-Tuning via Low-Rank Adaptation

Although a prediction can be obtained with the previous two steps, the data domain gap between these vision foundation models might still be significant. For instance, SAM [15] has been trained on modern object-centric datasets, so its feature space might not align well with the semantic meaning of cartographic symbols, resulting in inferior performance than what could be achieved if adapted to the domain of historical maps. However, full-finetuning, which retrains all model parameters, in times of foundation models, becomes impractical due to their enormous parameter count, lying in the order of hundreds of millions for the ViT-L version [12]. According to Kim et al. [14], full-finetuning can be disadvantageous as fully fine-tuned feature extractors can distort the rich and strong pre-trained representations.

Hence, we opt for so-called low-rank adaptation (LoRA), a technique developed by Hu et al. [12]. LoRA is a parameter-efficient fine-tuning method that injects trainable low-rank decomposition matrices into pre-trained weight matrices, reducing the number of trainable parameters while maintaining expressiveness. LoRA is typically applied to the query, key, and value matrices $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V \in \mathbb{R}^{D \times D_h}$, with $D_h$ being $D$ divided by the number of MSA heads, and the output projection layer $\mathbf{W}_O \in \mathbb{R}^{D \times D}$ of MSA. Hence, instead of updating, for instance, $\mathbf{W}_Q$ directly, LoRA reparametrizes it as:

$$\mathbf{W}'_Q = \mathbf{W}_Q + \mathbf{BA}, \tag{5}$$

with $\mathbf{A} \in \mathbb{R}^{D \times r}$ and $\mathbf{B} \in \mathbb{R}^{r \times D_h}$, where $r$ is the rank of the decomposition and is typically chosen such that $r \ll \min(D_h, D)$ to reduce the number of trainable parameters significantly. By setting the rank $r$ to be the same rank as the pre-trained weights, one essential recovers full-finetuning. We, therefore, select $r = 4$ to avoid overfitting with our limited data.

In addition to LoRA, many other parameter-efficient fine-tuning techniques emerged over the years, including LoKr and LoHa by Hyeon et al. [13], as well as DoRA [19], which we evaluate in this work.

### 3.4 Objective and Optimization

In order to tune the weights $\mathbf{w}$ and bias $b$ of our probing head $f$, as well as adapting $\mathcal{F}_\theta$ to our task-specific domain of historical maps, we use a mixture of Focal [17] and Dice [29] loss as proposed by Cheng et al. [4]. In detail, our training objective is as follows:

$$\mathcal{L} = \alpha\mathcal{L}_{\text{focal}} + \beta\mathcal{L}_{\text{dice}}, \tag{6}$$

where $\alpha$ and $\beta$ hyperparameters to weigh the two terms. Compared to Perera et al. [21], we found that $\alpha = 10$ and $\beta = 1$, a down-weighing of the first term benefited predictive performance.

We optimize the objective by utilizing an AdamW optimizer. In addition to that, we employ a scheduler to stabilize and expedite training. Specifically, we opted for the one-cycle policy described in the works by Smith et al. [25] which anneals the learning rate from an initial rate ($10^{-4}$) to some maximum ($10^{-3}$) and then from that to some minimum much lower than the initial learning rate.

## 4 Experiments & Results

In the following, we describe the evaluated datasets, a general overview of training and evaluation specifics, as well as the ablations we conducted, which include:

- **Foundation Models:** We compare the performance of DINOv2, RADIO, and SAM as feature extractors, i.e., keeping the encoder frozen and solely training a lightweight classifier.
- **Low-Rank Adaptation:** We assess four different parameter-efficient fine-tuning methods, namely LoRA, DoRA, LoHa, and LoKr, and compare them to pure linear probing of the encoder.
- **Input Resolution:** We examine the model's performance under different input resolutions, ranging from 224 pixels to the computationally much more intensive resolution of 1120 pixels.

We conclude this section with our final results, where we evaluate **Few-Shot Performance**. Specifically, we evaluate our best model in a multitude of few-shot settings, spanning $k$ from a single sample to the whole training set.

### 4.1 Datasets

We evaluate our proposed approach on two datasets of historical maps: the *Siegfried* [33] dataset and the *ICDAR 2021* [1] dataset, specifically on the task of segmenting building blocks from non-building blocks.

**Siegfried** consists of two sub-datasets: a railway dataset representing linear features and a vineyard dataset representing areal features. Both datasets are divided into $224 \times 224$ pixel patches from the Swiss Siegfried Maps*. In total, the full railway dataset consists of 5,872 training tiles, 839 validation tiles, and 1,679 testing tiles, maintaining an approximate 7:1:2 split. When examining predictive performance in few-shot settings, we systematically reduce the number of labeled training tiles from 100% to 10% (587 tiles), 1% (58 tiles), 10 tiles, 5 tiles, and

*see https://www.swisstopo.admin.ch/en/siegfried-map, last access: 15.03.2025.

1 tile. In the same fashion, the much smaller vineyard dataset is split into 613 training tiles, 87 validation tiles, and 177 testing tiles. Here, based on the number of available samples, we solely conduct training with 100%, 10-shot, 5-shot, and 1-shot experiments. Finally, overall performance is evaluated on the test set in each scenario.

**ICDAR 2021** comprises historical maps of Paris (1860s–1940s) with the task of detecting closed shapes representing building blocks, which are separated by streets, rivers, or fortifications. Concerning technical details, the dataset includes large-scale scans (up to $8000 \times 8000$ pixels) with a frame mask distinguishing relevant from non-relevant areas. For evaluation, we crop, for simplicity, non-overlapping $448 \times 448$ patches, resize them to $224 \times 224$, and rescale predictions accordingly. In total, it consists of one training image, one validation image, and three test images.

### 4.2 Experimental Setup

We report the following standard evaluation metrics for semantic segmentation: the mean Intersection over Union (mIoU) as well as the F1-Score (F1). To ensure the validity of our results, we trained a minimum of three models and averaged their results. In general, our ablations are performed on the Siegfried dataset, using the same ten samples as by Xia et al. [33] and training for 300 epochs with a batch size of four. With respect to our three image encoders, we use the ViT-L variants. While training, we track the mIoU of the validation set, which we use to select the model for the final evaluation on the test set.

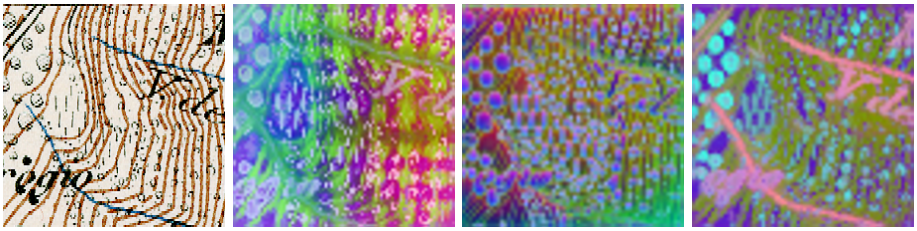
To increase data variety in our few-shot setting, we employ $D_4$-dihedral group symmetry transformations. These transformations include *rotations* (90°, 180°, 270°), *flips* (vertical, horizontal, diagonal), and *transpositions*. A great benefit of these augmentations is that input data remains within its original data distribution. We also considered more drastic augmentations in the form of MixUp [36] and CutMix [35]. However, we found this to be not beneficial in our case, which is in accordance with the findings by Kim et al. [14], who found out that augmentations with strong intensity degrade performance in few-shot learning.

We pursue a similar yet less drastic approach to input resolution than SAM. To obtain spatially larger feature embeddings, we rescale the input image by default to three times its size ($3 \cdot 224$ pixels), to improve per-pixel accuracy.

Concerning hardware, our experiments were conducted on NVIDIA RTX A5000/A6000 GPUs, depending on the required memory. More specifically, evaluations using the image encoder from SAM or evaluating input resolutions greater than $3 \cdot 224$) pixels required us to use GPUs with bigger memory and reduce the batch size down to only two samples.

### 4.3 Ablation Study

In the following sections, we report the results of our conducted ablations. As a teaser, we find that among the three foundation models, when probing them for



(a) Input (b) DINOv2 (c) SAM (d) RADIO

Fig. 3: A visualization of the first three principal components of feature embeddings from vision foundation models: subjectively speaking, RADIO exhibits the strongest spatial features, followed by DINOv2 and, lastly, SAM.

segmentation, RADIO has the richest feature embeddings. We find DoRA and LoRA to perform very similarly for parameter-efficient fine-tuning, but opted for DoRA due to its generally strong reported performance [33]. Finally, we find up-scaling the input beneficial, with three times the original input resolution performing best.

**Foundation Models.** In the following, we quantitatively and qualitatively assess DINOv2, RADIO, and SAM when keeping their weights fixed and training a linear classifier to predict segmentation masks per-pixel.

*Qualitative Results.* Similar to the works by Oquab et al. [20], we use PCA to reduce the spatial feature dimensionality of patch features from the three foundation models and visualize their first three principal components. Fig. 3 illustrates the ability of the selected foundation models to separate historical maps semantically into different classes. In all three cases, classes are separated; however, subjectively speaking, RADIO exhibits the strongest spatial features, followed by DINOv2 and, lastly, SAM. RADIO clearly distinguishes between text, contour lines, empty land, and vegetation symbols. Opposed to this is SAM, which puts text, vegetation symbols, and empty land semantically closer.

As SAM is optimized for instance-level segmentation to distinct clearly defined, prompt-specific objects based on supervised mask annotations, this is to be expected [15]. In contrast, DINOv2 is trained with a self-supervised objective that encourages consistency across different views of the same scene, without relying on labeled data [20]. Based on these fundamentally different training goals, it is clear why DINOv2 – and by extension, RADIO through its agglomerative learning – offers richer and more general representations.

*Quantitative Results.* Next, we examine the three foundation models quantitatively with the results denoted in Table 1. Continuing with the assumption that RADIO offers the most expressive features, the results are not as definitive: for the railway dataset, RADIO performs best by a large margin (mIoU of 52),

Table 1: Comparison of segmentation performance for DINOv2, SAM, and RADIO under linear probing with frozen backbones.

| Method | | Railways | | Vineyards | |
|---|---|---|---|---|---|
| | | F1 | IoU | F1 | IoU |
| DINOv2 | [20] | 35.5 | 19.1 | 41.3 | 23.9 |
| SAM | [15] | 45.5 | 16.4 | 44.2 | **47.4** |
| RADIO | [23] | **73.4** | **52.4** | **62.1** | 36.1 |

followed by DINOv2 (mIoU of 19); in contrast, for the vineyard dataset, SAM performs best (mIoU of 47), followed by RADIO (mIoU of 36). Although ambiguous, we show in subsequent sections that RADIO is equal to or better than SAM (cf. Fig. 4), with the additional benefit of requiring less memory.

Table 2: Comparison of parameter-efficient fine-tuning methods by segmentation performance and number of trainable parameters.

| Method | | Railways | | Vineyards | | Parameters | |
|---|---|---|---|---|---|---|---|
| | | F1 | IoU | F1 | IoU | Total | in % |
| None | | 73.4 | 52.4 | 62.1 | 36.1 | 1k | 0.00% |
| LoRA | [12] | 89.6 | 82.5 | **78.6** | **67.9** | 590k | 0.18% |
| LoKr | [13] | 87.1 | 78.8 | 76.1 | 63.7 | 77k | 0.02% |
| LoHa | [13] | 88.1 | 80.2 | 77.9 | 66.8 | 1,180k | 0.37% |
| DoRA | [19] | **89.9** | **83.7** | 77.8 | 67.2 | 689k | 0.21% |

**Low-Rank Adaptation.** In the following, we evaluated four different low-rank adaptation methods: LoRA, LoKr, LoHa, and DoRA. We define parameter-efficient fine-tuning configurations tailored to different data regimes. For few-shot learning with limited samples ($k \leq 10$), we employed a lower rank ($r = 4$), moderate scaling factor ($\alpha = 8$), and higher dropout rate (20%) to mitigate overfitting. Conversely, for datasets with more samples ($> 10$), we increased the rank ($r = 8$) and scaling factor ($\alpha = 16$) while reducing the dropout rate to 10%, improving learning stability and reducing the risk of overfitting.

Considering the results shown in Table 2, we find that simply using the embeddings as is and probing the network for semantic segmentation yields the worst performance. In contrast, adapting the model via low-rank adaptation shows significant improvements, with either LoRA or DoRA performing best depending on the dataset. Our results match previous findings, as according to Kim et al. [14] fine-tuning is a better strategy than linear probing.

Although almost indifferent, we selected DoRA for all subsequent experiments to allow for a better comparison to MapSAM [33]. Notably, implementing

parameter-efficient fine-tuning for the RADIO-L variant required only an additional 689k parameters, representing just 0.21% of the total trainable parameters.

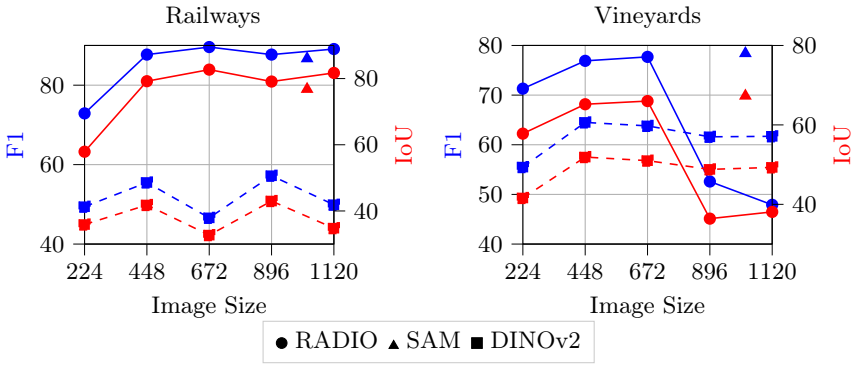
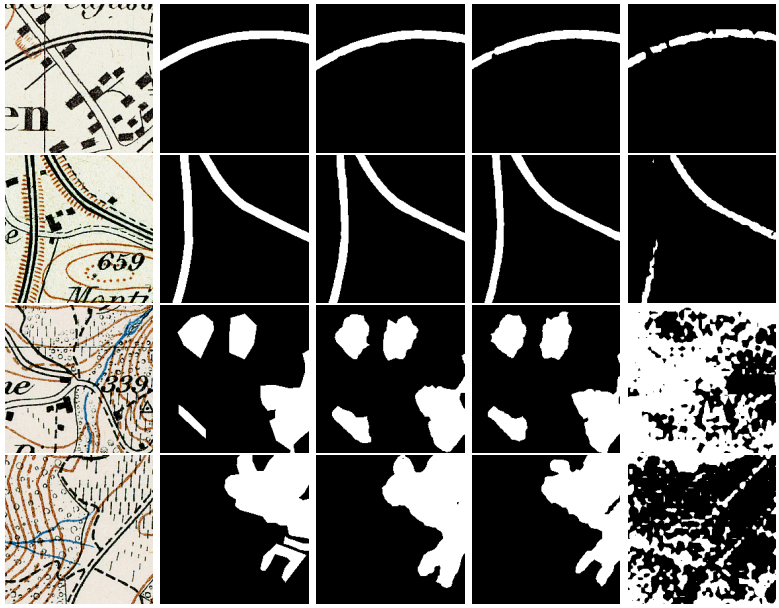



Fig. 4: Analyzing the impact of resolution: RADIO is on par with / better than SAM, while being computationally more efficient [23].

**Input Scaling.** In the next ablation, we consider the impact of input resolution and report the median over three runs due to strong variations in performance exhibited. Ablating input size plays a crucial role, as larger inputs result in higher-dimensional feature embeddings from the encoder. For instance, in SAM, resizing the input to a resolution of $1024 \times 1024$ pixels is the default to benefit precision in segmentation [15]. However, improved per-pixel performance comes at the cost of requiring much more compute, with the MSA being the bottleneck as its complexity is in $O(N^2)$, i.e., compute grows quadratically w.r.t. the number of patches. In our experiments, SAM took about five times longer than RADIO with an input resolution of $3 \cdot (224 \times 224)$.

We summarize our findings in Fig. 4, which depicts the performance of the three foundation models, adapted using DoRA [19] at different input resolutions. As SAM operates at a fixed input resolution, we report its performance at 1024 pixels. Analyzing the two graphs shows that RADIO peaks at a resolution of 672 pixels and performs on par with / better than SAM while being computationally cheaper and requiring less memory. Although this performance remains constant for the railway dataset, we experience a drastic drop in performance for vineyards with higher pixel values. We assume that this is due to the mode switch reported by Ranzinger et al. [23], where performance drops at a resolution greater than 720px; however, the performance on railways seems not to be impacted by this. Generally speaking, we experienced fluctuations in vineyards for both RADIO and SAM (not for DINOv2), and we hypothesize that sample quality is more important than for railways. Based on these insights, we keep the input resolution at 672 pixels and proceed with RADIO for the final evaluation.



(a) Input (b) Label (c) 10-Shot (d) 5-Shot (e) 1-Shot

Fig. 5: Examples of the Siegfried dataset [33], showing (a) the input map, (b) the ground truth, and the predictions at (c) 10-shots, (d) 5-shots, and (e) 1-shot.

### 4.4 Results

In the final evaluation, we compare our method to existing approaches on the Siegfried and ICDAR 2021 datasets. Our method, a RADIO-L model, low-rank adapted with DoRA and probed with a linear classifier for per-pixel segmentation, achieves competitive results despite its simplicity: it outperforms the SOTA in all few-shot settings on the Siegfried dataset for both railways and vineyards.

**Siegfried** In this section, we evaluate our model on the Siegfried dataset against other baselines and MapSAM [33]. For this, we train a baseline U-Net [24] with a ResNet-50 [10] encoder for 100 epochs. In addition, we evaluate two other model variants: DeepLabV3+ [2] & SegFormer [34]. We selected the former as it poses a strong CNN-based baseline, which was additionally employed successfully in the context of few-shot segmentation of historical documents [6,5]. For the latter, we wanted to contrast CNN baselines with a transformer-based architecture. With respect to our model, as more data is available, i.e., when training with 100% and 10%, the number of epochs is reduced to 30, and for runs with very few examples, i.e., $k \leq 5$, results are averaged over 10 runs.

First, we evaluate our model on railways with its results denoted in Table 3, which indicates the strong performance of our simple yet effective approach as it outperforms the SOTA in all few-shot settings. Notably, it achieves a +5%

Table 3: Comparing the segmentation performance of SAM variants against a U-Net and our approach, based on RADIO, on the railway dataset.

| **Railways** | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Method** | Full (5872) | | 10% (587) | | 1% (58) | | 10-shot | | 5-shot | | 1-shot | |
| | F1 | IoU | F1 | IoU | F1 | IoU | F1 | IoU | F1 | IoU | F1 | IoU |
| U-Net [24] | 96.2 | 92.1 | 94.9 | 90.2 | 86.6 | 72.0 | 82.2 | 70.5 | 72.3 | 54.7 | 14.4 | 9.5 |
| DeepLabV3+ [3] | 96.6 | 92.9 | 95.6 | 91.0 | 92.8 | 85.0 | 83.2 | 72.2 | 75.8 | 62.1 | 28.1 | 16.4 |
| SegFormer [34] | 96.6 | **93.0** | 96.0 | 91.8 | 94.1 | 87.6 | 87.8 | 78.0 | 81.5 | 69.3 | 36.4 | 23.8 |
| SAMed* | 92.0 | 86.3 | 91.6 | 85.7 | 91.8 | 86.0 | 84.6 | 75.4 | - | - | - | - |
| MapSAM | 94.1* | 89.5* | 93.6* | 88.7* | 92.1* | 86.5* | 87.2* | 78.5* | 75.2 | 64.3 | 40.3 | 34.6 |
| Ours | **96.7** | 92.0 | **96.4** | **92.5** | **94.9** | **90.5** | **89.4** | **82.0** | **86.1** | **77.0** | **52.4** | **41.5** |

improvement in the 10-shot setting and a 20% gain in the more challenging 5-shot setting. Second, our approach also yields the best performance in all few-shot settings for the vineyard dataset. Analyzing the results in Table 4, we observe an improvement of +13% in the 10-shot setting and +23% in the 5-shot setting.

Table 4: Comparing the segmentation performance of SAM variants against a U-Net and our approach, based on RADIO, on the vineyard dataset.

| **Vineyards** | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Method** | Full (613) | | 10-shot | | 5-shot | | 1-shot | |
| | F1 | IoU | F1 | IoU | F1 | IoU | F1 | IoU |
| U-Net [24] | 80.3 | 69.3 | 56.0 | 39.5 | 51.3 | 36.5 | 29.3 | 18.6 |
| DeepLabV3+ [3] | 83.9 | 72.9 | 67.9 | 52.7 | 62.3 | 44.1 | 34.4 | 21.4 |
| SegFormer [34] | 83.2 | 73.7 | 76.3 | 63.5 | 72.7 | 58.7 | 35.5 | 23.9 |
| SAMed* | 82.8 | **74.9** | 72.0 | 61.5 | - | - | - | - |
| MapSAM | 82.8* | 74.3* | 70.5* | 60.0* | 64.7 | 51.6 | 32.9 | 21.9 |
| Ours | **86.2** | 74.4 | **78.5** | **67.9** | **75.2** | **63.3** | **47.6** | **34.6** |

**ICDAR2021** For this dataset, not only are the F1 and mIoU computed. According to the competition details, instance segmentation performance is evaluated using the COCO Panoptic Quality ($PQ$) metric (cf. [1]): first, connected components are extracted from the predicted label mask, from which Segmentation

*results are taken from [33].

Quality ($SQ$), the mIoU of matched regions, and Recognition Quality ($RQ$), the F-score of detected regions, is computed; the final score is defined as:

$$PQ = SQ \times RQ, \tag{7}$$

where a predicted shape is considered a match if its IoU with a reference shape exceeds 0.5; higher is better. Our performance on the first task of the ICDAR 2021 competition is denoted in Table 5. Our simple approach yields an average $PQ$ of 67.3%, placing us *second* in the rankings. In terms of IoU and F1-Score, we achieve 77.1 for the former and 82.7 for the latter on the test set.

We find this result striking, given that our model has not been tuned in any way to the shape-aware evaluation criteria of the $PQ$, exemplifying the generalization ability of our methodology.

Table 5: Our results on the ICDAR 2021 challenge on building segmentation.

**Per-Map Results**

| Map | PQ | SQ | RQ |
|---|---|---|---|
| 301 | 71.3 | 93.3 | 76.4 |
| 302 | 64.2 | 93.2 | 68.9 |
| 303 | 66.4 | 93.4 | 71.1 |
| Ours | 67.3 | 93.3 | 72.1 |

**Mean Comparisons**

| Method | PQ |
|---|---|
| 1st Place | 74.1 |
| **Ours** | **67.3** |
| 2nd Place | 62.6 |
| 3rd Place | 44.0 |
| U-Net [24] | 14.4 |

## 5 Conclusion

In this work, we addressed the challenges posed by the diverse visual representations and limited annotated data of historical maps, which are rich sources of historical information. We proposed a simple yet effective approach for few-shot segmentation of historical maps, leveraging the rich semantic embeddings of large vision foundation models combined with parameter-efficient fine-tuning. Our method outperforms the state-of-the-art on the Siegfried benchmark dataset in vineyard and railway segmentation, achieving +5% and +13% relative improvements in mIoU in 10-shot scenarios and around +20% in the more challenging 5-shot setting. Additionally, it demonstrates strong performance on the ICDAR 2021 competition dataset, attaining a mean PQ of 67.3% for building block segmentation, despite not being optimized for this shape-sensitive metric, underscoring its generalizability. Notably, our approach maintains high performance even in extremely low-data regimes while requiring only 689k trainable parameters – just 0.21% of the total model size. In summary, our approach enables precise segmentation of diverse historical maps while drastically reducing the need for manual annotations, advancing automated processing and analysis in the field. Future work could explore test-time fine-tuning with the nearest neighbor example, as suggested by Sun et al. [9], and incorporate test-time augmentation techniques proposed by Kim et al. [14].

## References


1. Chazalon, J., Carlinet, E., Chen, Y., Perret, J., Duménieu, B., Mallet, C., Géraud, T., Nguyen, V., Nguyen, N., Baloun, J., Lenc, L., Král, P.: ICDAR 2021 Competition on Historical Map Segmentation. In: Lladós, J., Lopresti, D., Uchida, S. (eds.) Document Analysis and Recognition – ICDAR 2021. pp. 693–707. Springer International Publishing, Cham (2021)
2. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. IEEE Transactions on Pattern Analysis and Machine Intelligence **40**(4), 834–848 (Apr 2018). https://doi.org/10.1109/tpami.2017.2699184
3. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation, p. 833–851. Springer International Publishing (2018)
4. Cheng, B., Schwing, A., Kirillov, A.: Per-pixel classification is not all you need for semantic segmentation. In: Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., Vaughan, J.W. (eds.) Advances in Neural Information Processing Systems. vol. 34, pp. 17864–17875. Curran Associates, Inc. (2021)
5. De Nardin, A., Zottin, S., Paier, M., Foresti, G.L., Colombi, E., Piciarelli, C.: Efficient few-shot learning for pixel-precise handwritten document layout analysis. In: 2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). IEEE (Jan 2023)
6. De Nardin, A., Zottin, S., Piciarelli, C., Colombi, E., Foresti, G.L.: A one-shot learning approach to document layout segmentation of ancient arabic manuscripts. In: 2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). IEEE (Jan 2024)
7. Ding, H., Zhang, H., Jiang, X.: Self-regularized prototypical network for few-shot semantic segmentation. Pattern Recognition **133**, 109018 (Jan 2023)
8. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2021)
9. Hardt, M., Sun, Y.: Test-time training on nearest neighbors for large language models. In: The Twelfth International Conference on Learning Representations (2024)
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). p. 770–778. IEEE (Jun 2016)
11. He, W., Zhang, Y., Zhuo, W., Shen, L., Yang, J., Deng, S., Sun, L.: APSeg: Auto-prompt network for cross-domain few-shot semantic segmentation. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). p. 23762–23772. IEEE (Jun 2024)
12. Hu, E.J., yelong shen, Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. In: International Conference on Learning Representations (2022)
13. Hyeon-Woo, N., Ye-Bin, M., Oh, T.H.: FedPara: Low-rank Hadamard product for communication-efficient federated learning. In: International Conference on Learning Representations (2022)

14. Kim, Y., Oh, J., Kim, S., Yun, S.Y.: How to Fine-tune Models with Few Samples: Update, Data Augmentation, and Test-time Augmentation (Aug 2022)
15. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollar, P., Girshick, R.: Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 4015–4026 (October 2023)
16. Li, Y., Mao, H., Girshick, R., He, K.: Exploring Plain Vision Transformer Backbones for Object Detection, p. 280–296. Springer Nature Switzerland (2022)
17. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollar, P.: Focal loss for dense object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence **42**(2), 318–327 (Feb 2020)
18. Liu, Y., Zhu, M., Li, H., Chen, H., Wang, X., Shen, C.: Matcher: Segment anything with one shot using all-purpose feature matching. In: The Twelfth International Conference on Learning Representations (2024)
19. Mao, Y., Huang, K., Guan, C., Bao, G., Mo, F., Xu, J.: DoRA: Enhancing parameter-efficient fine-tuning with dynamic rank distribution. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). p. 11662–11675. Association for Computational Linguistics (2024)
20. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y., Li, S.W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: DINOv2: Learning robust visual features without supervision. Transactions on Machine Learning Research (2024)
21. Perera, R., Halgamuge, S.: Discriminative sample-guided and parameter-efficient feature space adaptation for cross-domain few-shot learning. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). p. 23794–23804. IEEE (Jun 2024)
22. Petitpierre, R., Kaplan, F., di Lenardo, I.: Generic semantic segmentation of historical maps. In: Proceedings of the Workshop on Computational Humanities Research (CHR 2021). CEUR, Amsterdam, NL (2021)
23. Ranzinger, M., Heinrich, G., Kautz, J., Molchanov, P.: AM-RADIO: Agglomerative vision foundation model reduce all domains into one. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). p. 12490–12500. IEEE (Jun 2024)
24. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation, p. 234–241. Springer International Publishing (2015)
25. Smith, L.N., Topin, N.: Super-Convergence: Very Fast Training of Neural Networks Using Large Learning Rates (May 2018)
26. Snell, J., Swersky, K., Zemel, R.: Prototypical networks for few-shot learning. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc. (2017)
27. Sterzinger, R., Brenner, S., Sablatnig, R.: Drawing the Line: Deep Segmentation for Extracting Art from Ancient Etruscan Mirrors. In: 2024 ICDAR International Conference on Document Analysis and Recognition, submitted (2024)
28. Sterzinger, R., Stippel, C., Sablatnig, R.: Fusing Forces: Deep-Human-Guided Refinement of Segmentation Masks, p. 154–169. Springer Nature Switzerland (Dec 2024)

29. Sudre, C.H., Li, W., Vercauteren, T., Ourselin, S., Jorge Cardoso, M.: Generalised Dice Overlap as a Deep Learning Loss Function for Highly Unbalanced Segmentations, p. 240–248. Springer International Publishing (2017)
30. Sun, Y., Chen, J., Zhang, S., Zhang, X., Chen, Q., Zhang, G., Ding, E., Wang, J., Li, Z.: VRP-SAM: SAM with visual reference prompt. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). p. 23565–23574. IEEE (Jun 2024)
31. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: Attention is all you need. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc. (2017)
32. Xia, X., Jiao, C., Hurni, L.: Contrastive pretraining for railway detection: Unveiling historical maps with transformers. In: Proceedings of the 6th ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery. p. 30–33. SIGSPATIAL '23, ACM (Nov 2023)
33. Xia, X., Zhang, D., Song, W., Huang, W., Hurni, L.: MapSAM: Adapting segment anything model for automated feature detection in historical maps. GIScience & Remote Sensing **62**(1) (Apr 2025)
34. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: Segformer: Simple and efficient design for semantic segmentation with transformers. In: Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., Vaughan, J.W. (eds.) Advances in Neural Information Processing Systems. vol. 34, pp. 12077–12090. Curran Associates, Inc. (2021)
35. Yun, S., Han, D., Chun, S., Oh, S.J., Yoo, Y., Choe, J.: CutMix: Regularization Strategy to Train Strong Classifiers With Localizable Features. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE (Oct 2019)
36. Zhang, H., Cissé, M., Dauphin, Y.N., Lopez-Paz, D.: Mixup: Beyond empirical risk minimization. In: 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings (2018)
37. Zhang, R., Jiang, Z., Guo, Z., Yan, S., Pan, J., Dong, H., Qiao, Y., Gao, P., Li, H.: Personalize segment anything model with one shot. In: The Twelfth International Conference on Learning Representations (2024)