Journal of *Imaging*

MDPI

Article

Reducing Training Data Using Pre-Trained Foundation Models: A Case Study on Traffic Sign Segmentation Using the Segment Anything Model

Sofia Henninger [1,*], Maximilian Kellner [1,2], Benedikt Rombach [1] and Alexander Reiterer [1,2]

[1] Fraunhofer Institute for Physical Measurement Techniques IPM, 79110 Freiburg, Germany; maximilian.kellner@ipm.fraunhofer.de (M.K.); benedikt.rombach@ipm.fraunhofer.de (B.R.); alexander.reiterer@ipm.fraunhofer.de (A.R.)
[2] Department of Sustainable Systems Engineering INATECH, Albert Ludwigs University Freiburg, 79110 Freiburg, Germany
* Correspondence: sofia.henninger@ipm.fraunhofer.de

Abstract: The utilization of robust, pre-trained foundation models enables simple adaptation to specific ongoing tasks. In particular, the recently developed Segment Anything Model (SAM) has demonstrated impressive results in the context of semantic segmentation. Recognizing that data collection is generally time-consuming and costly, this research aims to determine whether the use of these foundation models can reduce the need for training data. To assess the models' behavior under conditions of reduced training data, five test datasets for semantic segmentation will be utilized. This study will concentrate on traffic sign segmentation to analyze the results in comparison to Mask R-CNN: the field's leading model. The findings indicate that SAM does not surpass the leading model for this specific task, regardless of the quantity of training data. Nevertheless, a knowledge-distilled student architecture derived from SAM exhibits no reduction in accuracy when trained on data that have been reduced by 95%.

Keywords: semantic segmentation; segment anything model; Mask R-CNN; training data reduction; traffic signs

check for updates

Citation: Henninger, S.; Kellner, M.; Rombach, B.; Reiterer, A. Reducing Training Data Using Pre-Trained Foundation Models: A Case Study on Traffic Sign Segmentation Using the Segment Anything Model. *J. Imaging* **2024**, *10*, 220. https://doi.org/10.3390/jimaging10090220

Academic Editor: Mrinal Mandal

Received: 19 July 2024
Revised: 2 September 2024
Accepted: 5 September 2024
Published: 7 September 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Traffic signs are an essential component of the infrastructure that enables the safe movement of all road users, including drivers, passengers, cyclists, and pedestrians. They provide unambiguous instructions and warnings that facilitate the prevention of accidents and reduce the risk of injuries or fatalities. It should be noted, however, that this only applies to traffic signs that are intact, correctly oriented, and undamaged. Due to the large number of traffic signs, manually inspecting, localizing, and constantly monitoring them is almost impossible. In Germany alone, for instance, the official traffic sign catalog lists a total of 1134 different traffic signs [1]. These include danger signs, regulatory signs, and direction signs, which are further complemented by traffic devices and additional signs. The estimated number of traffic signs on German roads is approximately 20 million [2]. This equates to one sign every 24 m along the entire street network of 830,000 km [3].

The automatic detection and segmentation of traffic signs has the potential to enhance the efficiency of the process of ensuring correctness. To solve these vision tasks, deep learning is the current state-of-the-art approach [4]. Traditionally, training robust models for such tasks requires an extensive amount of annotated data, especially in terms of the diversity of traffic signs. The collection of data is both time-consuming and costly. Recent advances in large foundation models such as BERT, DALL-E, and GPT [5] offer a promising solution to this challenge by using knowledge already acquired from these foundation models for such specific downstream tasks [5]. Thus, the need for large additional training

data can be minimized, because fine-tuning often requires only a limited amount of data or specific guidance. This study explores the application of pre-trained foundation models using *Meta*'s Segment Anything Model (SAM) [6]. SAM is a transformer-based foundation model developed for semantic segmentation. The use of the transformer model architecture enables more meaningful models to be created in less time through parallel training [5]. A large amount of training data utilized for pre-training the SAM improves the model's capabilities in the field of computer vision. The primary purpose of this study is to assess whether the SAM can maintain its performance despite being trained on a significantly reduced dataset, thereby lowering the barriers for subtasks and making them more accessible and cost-effective. The performance of the foundation model is evaluated in comparison to a leading architecture in the field of traffic sign segmentation.

The architecture of the SAM consists of an image encoder, prompt encoder, and mask decoder. Its heavyweight image encoder is based on a vision transformer (ViT) [7] architecture and extracts the input image features into image embeddings. The prediction depends on an input that is provided as a prompt. This prompt may take the form of one or more points, a bounding box, a mask, or text. The prompt is processed by the prompt encoder [6]. The lightweight mask decoder is designed for interactive real-time mask prediction. Several studies have modified the model to align with their specific research objectives. These include studies on remote sensing applications [8–10], shadows [11,12], food [13], web pages [14], oil spills [15], and camouflaged objects [16]. SAM fine-tuning has a particular focus on medical image segmentation [17–21] due to the limited availability of data in this field. In general, updating all parameters of the SAM is a time-consuming process. Consequently, numerous studies have focused their efforts on parameter-efficient fine-tuning [17]. They employed the use of adapter modules, which are positioned between the encoder transformer layers [11,18]. Subsequently, the adapter, prompt encoder, and mask decoder are fine-tuned on specific data. A significant limitation of the SAM is the dependency of segmentation on prompt input. For this particular task, it is necessary to adapt the SAM for the purpose of automatic traffic sign segmentation. One concept for automatic mask prediction is the use of the automatic mask generation pipeline [6], which employs a large grid of points as a prompt. However, the automatic mask generation pipeline is not suitable for productive use due to its long calculation time. For instance, in the context of SAM automation, there are approaches for developing a prompt generator, which generates prompts for the prompt encoder [10,22–24]. Another variant employs learnable embeddings as prompts [24].

In this work, the original SAM image encoder is combined with two distinct decoders to determine whether there is an influence of a decoder of the foundation model, which only contains 1% of all the model parameters. In addition to the original mask decoder (referred to as SAM-Fine-ViT$_H$), a decoder containing convolutional layers (referred to as SAM-Conv-ViT$_H$) is used. One significant drawback of the powerful SAM encoder is its long processing time, which is a consequence of the considerable number of parameters involved. Consequently, a parameter-distilled encoder version ViT$_T$ [25], created based on knowledge distillation to enhance performance, is employed. It is an opportune moment to consider whether the tiny version would yield comparable results, even with limited training data. The heavyweight image encoder has consequently been replaced with the tiny image encoder, while only the decoder has undergone fine-tuning. The decoder also contains convolutional layers (referred to as SAM-Conv-ViT$_T$).

Given the importance of traffic sign recognition for autonomous driving systems, as well as urban mobile mapping for maintenance planning or accident prevention, the field is widely covered in the scientific literature. However, in this field, the dominant approach is the use of convolutional neural networks (CNNs) [26–32]. In contrast to object detection, semantic segmentation of traffic signs is a relatively underrepresented field. References [33–38] demonstrate that the Faster R-CNN model is a widely used tool for the detection of traffic signs, with the model achieving highly satisfactory results. In addition to the Faster R-CNN model, its extension, the Mask R-CNN model is used, as it also

enables instance segmentation [39,40]. One of the few studies on sign segmentation uses a fully connected network [27]. Another approach [41] employs traffic sign recognition through segmentation with a specially developed SegU-Net, a fully connected network, that represents a combination of Seg-Net [42], and U-Net [43]. In 2022, an initial approach was implemented to investigate the use of ViT for traffic sign recognition [44]. It was found that the tested CNNs outperformed the ViT architecture. However, these were pre-trained compared to the transformer architectures. The publication points out the advantages of the ViT in terms of its shorter training duration. Another paper comparing the feature extraction of CNNs with that of ViT [45] demonstrates that the original ViT models are outperformed in the detection of advanced CNNs. Nevertheless, when a ViT is employed as a backbone model in the context of detection and compared to CNN backbone models, the ViT models demonstrate superior performance [46]. The three models, which are based on the SAM architecture, are evaluated and compared to the Mask R-CNN architecture as a well-established and leading framework for traffic sign segmentation.

In this study, we investigate the adaptability of foundation models to a specific task with minimal additional training data. Our contributions can be summarized as follows:

- A comparative evaluation between a state-of-the-art architecture provides for a specific task and a foundation model, focusing on their performance with reduced training data.
- Two distinct decoders applied to the same large-scale encoder are compared, examining their impact on model performance.
- The effects of knowledge distillation from foundation models are examined in the context of training data reduction.
- The results are evaluated against multiple benchmark datasets, ensuring the robustness, reliability, and generalizability of our findings.
- Model performance is measured based on the intersection over union (IoU), precision–recall, and their segmentation ability on an instance basis.

The main documentation is structured in two sections: Section 2 considers the structures of the four implemented architectures, while Section 3 presents the results of the implementation. In Section 2, the implementation details are provided, including descriptions of the data preprocessing steps, the hyperparameter selection process, and the training setup. The datasets used are described in detail in Section 3, followed by the presentation of segmentation results on different datasets. Additionally, the impact of training data reduction is analyzed, considering the IoU and precision–recall metric and presenting instance-level segmentation results. The results are interpreted in Section 4 and summarized in Section 5.

2. Methods

For implementation, the pre-trained ViT_H (huge) encoder is employed. The architectural structure enables the encoder to be used independently. During preprocessing, the image is scaled to 1024×1024 pixels and encoded into 64×64 with 256 feature maps [6]. For automated use, the prompt encoder is completely removed, so the prompt tokens that correspond to the output of the prompt encoder are replaced by learnable embeddings. The number of model parameters that are frozen and fine-tuned during training, along with their respective performance, is presented in Table 1. The performance is quantified in terms of floating-point operations per second (FLOPS) using an NVIDIA A100 SXM4 80GB GPU with an input tensor of dimensions $1024 \times 1024 \times 3$.

Table 1. The number of model parameters and the performance of the models, measured in FLOPS.

Model	All	Frozen	Fine-Tuned	FLOPS in 1/s
SAM-Fine-ViT$_H$	640,803,746	637,026,048	3,777,698	218.0
SAM-Conv-ViT$_H$	644,956,546	637,026,048	7,930,498	217.8
SAM-Conv-ViT$_T$	13,996,030	6,065,532	7,930,498	2413.8
Mask R-CNN	43,922,395	222,400	43,699,995	1439.6

For SAM-Fine-ViT$_H$, the originally designed lightweight decoder is utilized (Figure 1). It initially processes the embeddings with a transformer, and subsequently with 2D transposed convolutions, to increase the image size to 256 × 256. The transposed convolution employs a 2 × 2 kernel with a stride of two, which doubles the resolution and reduces the number of feature maps to a quarter. After a transposed convolution, layer normalization follows. The output of the up-scaling step is employed after a multilayer perceptron to generate the masks and the IoU token.

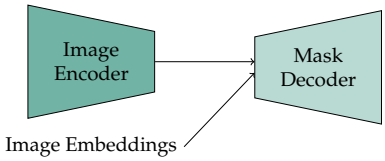


Figure 1. SAM-Fine-ViT: SAM architecture without prompt encoder.

The decoder of SAM-Conv-ViT corresponds to the final two up-sampling steps, which combine an up-sampling and convolution layer, of the ascending path of the U-Net implementation [43]. The initial step of the ViT patch-embedding process involves cropping image patches to a size of 16 × 16. Consequently, no encoding information is stored, and no skip connections can be used. Following the up-sampling, which doubles the size of the output and halves the number of feature maps, a subsequent double convolution is then applied (Figure 2). A simple parameter-efficient bilinear interpolation is employed for up-sampling. The convolution is performed with a 3 × 3 kernel, a stride of one, and a zero-padding of one, and it is supplemented by batch normalization and the rectified linear unit (ReLU) activation function. Finally, 64 feature maps with a resolution of 256 × 256 remain, which are then processed into a binary mask by a 1 × 1 convolution layer.

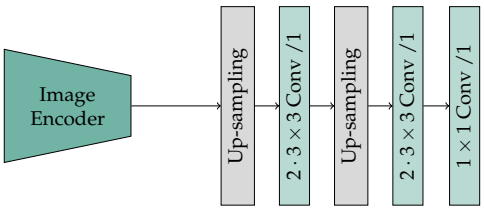


Figure 2. SAM-Conv-ViT$_{H/T}$: SAM image encoder with convolutional decoder.

In accordance with reference [47], the Mask R-CNN implementation (Figure 3) employs a ResNet50 backbone model for feature extraction, extended by a feature pyramid network (FPN) [48]. The FPN employs hierarchical processing of image information at four different resolutions. This approach is advantageous for the recognition of traffic signs, as these are often very small. In ResNet50, residual blocks derived from convolution operations with 1 × 1, 3 × 3, and 1 × 1 kernels with a stride of one were employed. To enhance the extraction of features, a ResNet50 pre-trained on the COCO dataset was utilized. The weights of the initial layer of a ResNet50 are frozen for the training process. This layer is typically responsible for processing fundamental features, such as edges and simple shapes, which are frequently useful in numerous image-recognition tasks. Based on the output of the encoder, a region proposal network (RPN) is used to predict potential bounding

boxes, from which smaller feature maps are formed. A small fully convolutional network (FCN) is added to the architecture as a decoder for the pixel-based prediction of the binary mask. Additionally, a fully connected (FC) layer has been included to provide the necessary classification capabilities.

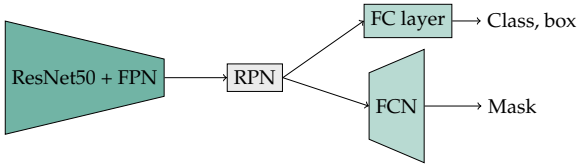


Figure 3. Mask R-CNN architecture.

Implementation Details

The decoders are subjected to fine-tuning on a dataset for traffic sign segmentation, while the encoder parameters (or a subset of them in the case of Mask R-CNN) are held constant throughout the training process. In the preprocessing step, the input images are scaled to the dimensions required by the architectures. Additionally, the red, green, and blue (RGB) values of the images are normalized with the average RGB and standard deviation (SD) values of images used for training the SAM and Mask R-CNN. The SAM encoder architecture requires an image size of 1024×1024, while the output mask of the decoder is 256×256. For Mask R-CNN, an input and output image size of 512×512 is employed.

To enhance the robustness and generalizability of architectures, the utilization of data augmentation for road images in the context of traffic sign recognition has a beneficial impact on the outcomes [38,39]. In this study, a series of functions, including brightness, contrast, rotation, distortion, blurring, and noise, were employed for data augmentation in the context of street scenes. The images have been adjusted to handle different lighting situations, varying camera settings and perspectives, and environmental influences such as motion blurring caused by high travel speeds. For the implementation, the data augmentation functions of torchvision, a library from PyTorch [49], and albumentation [50] are used.

The segmentation models are trained using an equal-weighted combination of the Dice and Binary Cross-Entropy loss functions. This approach is particularly well-suited for data sets with an imbalanced class distribution and for small objects [51]. Additionally, the Adam optimization algorithm [52] is employed. An initial learning rate of 0.1 (for all SAM models) and 0.001 (for Mask R-CNN) is scheduled based on the publication by Loshchilov et al. (2017) [53]. The approach employs the cosine annealing warm restarts function, which periodically repeats the descending part of the cosine function. The training phase was conducted using a maximum of five hundred epochs, but the process was terminated prematurely when the validation metric reached a stable point or began to decline. All models were trained on an NVIDIA A100 SXM4 80 GB GPU with a batch size of eight, which aligns with the maximum available resources of the working memory. The implementation is carried out in Python using the PyTorch [49] framework.

3. Results

The IoU metric utilized in the present study is based on the previously published paper on traffic sign instance segmentation and classification with data augmentation by Yoá et al. (2023) [39]. The IoU threshold value categorizes the segmented area as either a true positive (TP), a true negative (TN), a false positive (FP), or a false negative (FN) prediction. The results presented below demonstrate the IoU for a threshold of 0.75 for the class *traffic sign*.

3.1. Datasets

The segmentation results of the different architectures are evaluated using five datasets for traffic sign segmentation. A subset of images from each dataset is selected, containing the class *traffic sign*. Figure 4 provides an example of one image and mask per dataset. The size, mean, and SD of the images in Table 2 illustrate the considerable differences between the datasets.

The *Cityscapes Dataset* [54] comprises segmented urban images from 18 German cities, with 5000 images finely annotated and 20,000 images coarsely annotated. Only the finely annotated data are utilized in this study. Of the 5000 images, 3475 annotated images are publicly available. Of these, 94% (3281) contain the class *traffic sign*. The *Audi Autonomous Driving Dataset* (A2D2) [55] comprises 41,277 frames of semantic segmented images of highways, country roads, and cities in the south of Germany. A total of 1841 front–center images were utilized for model testing. A dataset with high diversity is the *Mapillary Vistas Dataset* [56]. The street-level photos were collected using different cameras from a rich community of contributors worldwide. There are 25,000 images within six continents. The official validation subset of 1726 images is used for testing purposes. A significant advantage of the Mapillary Vistas dataset is its annotation of traffic sign instances. A further dataset for autonomous driving is the *ApolloScape Dataset* [57], which contains 146,997 semantically annotated images from four regions in China. A total of 2269 selected examples were used for testing purposes. Furthermore, a large synthetic dataset is available in the form of *IDDA* [58]. This dataset imitates over 1M images of six US towns captured under three different weather conditions. For testing, 3075 images captured during clear daylight are utilized.



Figure 4. Image and mask examples of the datasets. From left to right: Cityscapes, A2D2, ApolloScape, IDDA, Mapillary Vistas.

Table 2. The number of available images per dataset, along with their image sizes, means, and SDs of RGB.

Dataset	Training	Testing	Size	Mean [r, g, b]	SD [r, g, b]
Cityscapes	2808	473	2048 × 1024	[73.6, 83.3, 72.9]	[45.3, 46.4, 45.6]
A2D2	-	1841	1920 × 1208	[129.4, 121.6, 119.2]	[47.6, 54.0, 60.1]
ApolloScape	-	2269	3384 × 2710	[125.7, 135.3, 141.0]	[94.3, 97.6, 97.2]
IDDA	-	3075	1920 × 1080	[90.7, 95.0, 103,8]	[50.6, 51.8, 57.4]
Mapillary Vistas	-	1726	arbitrary	[107.7, 117.7, 120.8]	[62.6, 65.9, 72.5]

3.2. Segmentation Results for Different Datasets

To assess the performance of the architectures, the models were tested on five distinct datasets (Section 3.1).

The segmentation results of different models across different test datasets are presented in both tabular form in Table 3 and visually in Figure 5. All four architectures trained on the Cityscapes dataset achieved their respective best results on the same test dataset. The models are particularly well adapted to the characteristics of this specific data set. The Mask R-CNN model achieved its top performance on Cityscapes. In general, the lowest segmentation results were observed on the IDDA test data, which are of a synthetic origin. This was particularly evident in the case of the Mask R-CNN model. Notably, the SAM-Conv-ViT$_H$ model exhibits the highest average performance (0.432), indicating that this model generalizes most effectively across the diverse test datasets. Nevertheless, the Mask R-CNN model exhibits the highest values in individual datasets (Cityscapes, ApolloScape, and Mapillary). In comparison with the SAM-Conv-ViT$_H$ model, it can be observed that the SAM-Conv-ViT$_T$ demonstrates minimal differences when applied to different datasets, but the overall results are lower than those obtained using the SAM-Conv-ViT$_H$ model. In overall consideration, the SAM-Fine-ViT$_H$ architecture achieved the lowest results.

Table 3. Results of the IoU for the class *traffic sign* for the models trained on the entire dataset (2808 images). For each dataset, the highest IoU is highlighted.

Dataset	SAM-Fine-ViT$_H$	SAM-Conv-ViT$_H$	SAM-Conv-ViT$_T$	Mask R-CNN
Cityscapes	0.436	0.592	0.370	**0.676**
A2D2	0.279	**0.499**	0.422	0.305
ApolloScape	0.259	0.390	0.401	**0.456**
IDDA	0.253	**0.360**	0.322	0.107
Mapillary Vistas	0.219	0.318	0.290	**0.342**
Mean	0.289	**0.432**	0.361	0.377
SD	0.086	0.098	**0.050**	0.195

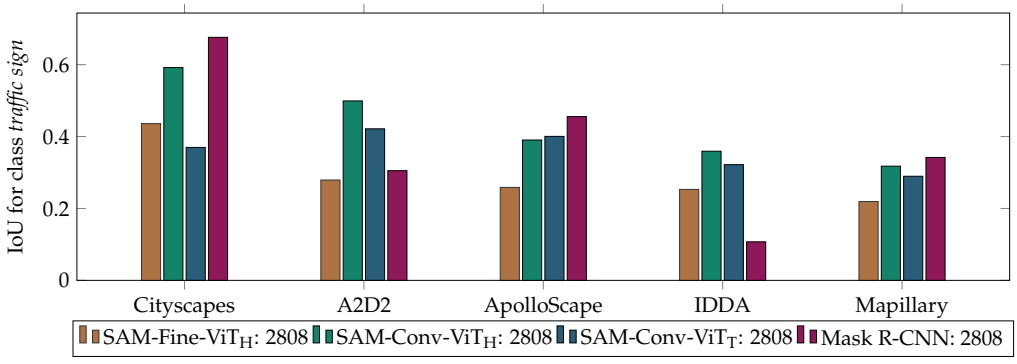


Figure 5. Results of the IoU for the class *traffic sign* for the models trained on the entire dataset (2808 images).

3.3. Reduction of Training Data

To evaluate the performance of the segmentation architectures on a limited amount of training data, the training was performed on a subset of 5.5% of the Cityscapes data. The impact of reducing the training dataset to 180 images is illustrated in Table 4, which also presents the difference in the results when training on the full dataset. Figure 6 illustrates the results in a visual format.

Table 4. Results of the IoU for the class *traffic sign* for the models trained on a smaller subset of data (180 images). The differences relate to the models that are trained on the entire dataset (2808 images). For each dataset, the highest IoU and the lowest difference are highlighted.

	SAM-Fine-ViT$_H$		SAM-Conv-ViT$_H$		SAM-Conv-ViT$_T$		Mask R-CNN	
Dataset	**Small**	**Diff**	**Small**	**Diff**	**Small**	**Diff**	**Small**	**Diff**
Cityscapes	0.242	−0.194	**0.428**	−0.164	0.290	**−0.080**	0.372	−0.304
A2D2	0.078	−0.201	0.349	−0.150	**0.370**	**−0.052**	0.194	−0.111
ApolloScape	0.118	−0.141	0.226	−0.164	**0.440**	**+0.039**	0.268	−0.188
IDDA	0.050	−0.203	0.228	−0.132	**0.267**	**−0.055**	0.192	+0.085
Mapillary Vistas	0.122	−0.097	0.263	−0.055	0.247	**−0.043**	**0.271**	−0.071
Mean	0.122	−0.167	0.299	−0.133	**0.323**	**−0.038**	0.259	−0.118

Notably, all models exhibited a significant decline in performance when trained on a smaller dataset. The mean accuracy loss (mean diff) was greatest for SAM-Fine-ViT$_H$ (−0.1672). It can be observed that the performance of SAM-Fine-ViT$_H$ exhibited a pronounced decline as a consequence of the reduction in the quantity of training data. This decline occurs to a greater extent than in the case of SAM-Conv-ViT$_H$, which employed the same encoder. The SAM-Conv-ViT$_T$ model demonstrates the most effective performance when trained on a reduced training set and is the most robust to the reduction in training data, exhibiting minimal impacts. The reduction in the accuracy of SAM-Conv-ViT$_H$ due to the reduction in the training data is greater than that of Mask R-CNN. However, the greatest loss of accuracy is observed on the Cityscapes dataset, particularly for Mask R-CNN.

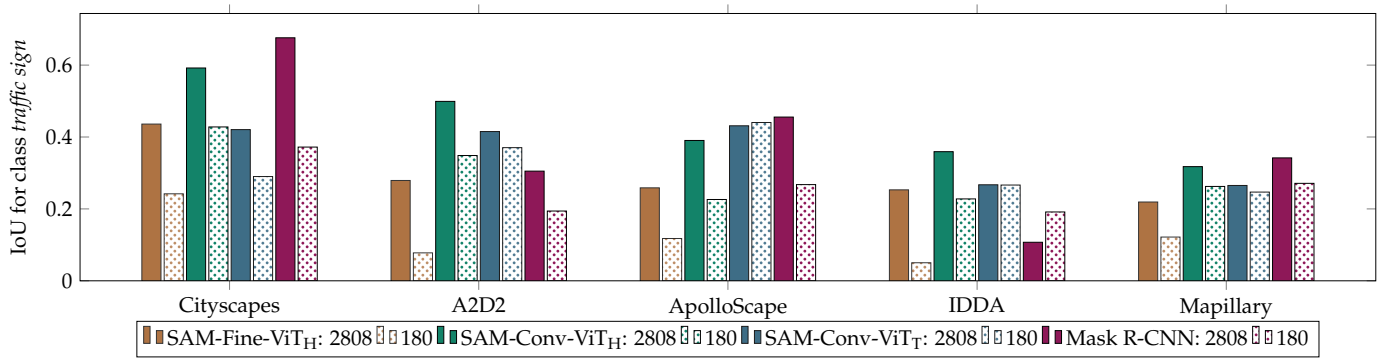


Figure 6. Results of the IoU for the class *traffic sign* for the models trained on the entire dataset (2808 images) and on a smaller subset (180 images).

For further analysis, the SAM-Conv-ViT$_H$ and Mask R-CNN were trained on 925 (32.9% of the total) and 90 (3.2% of the total) Cityscapes images (Figure 7). The smaller subsets comprise images of the larger ones. As expected, the two models generally show a decline in accuracy with the reduction of data. Mask R-CNN exhibits a more pronounced decline in accuracy when confronted with limited training data, whereas SAM-Conv-ViT$_H$ demonstrates a relatively stable and foreseeable reduction in accuracy. For very limited training data (90 images), the foundation model SAM-Conv-ViT$_H$ outperforms Mask R-CNN, SAM-Conv-ViT$_H$ has a decrease in IoU of −0.15, and Mask R-CNN has a decrease of −0.26 over all datasets, referring to the models trained on the entire dataset.

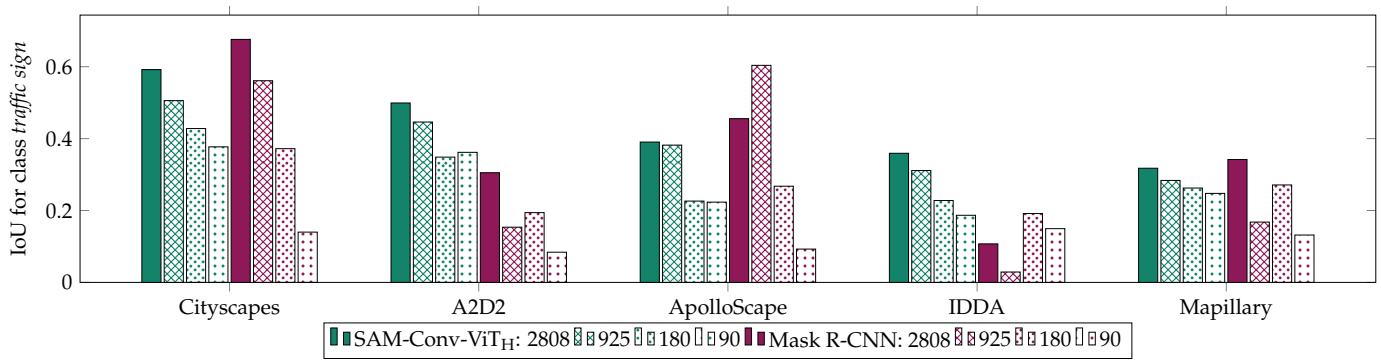


Figure 7. Results of the IoU for the class *traffic sign* for SAM-Conv-ViT$_H$ and Mask R-CNN trained on the entire dataset (2808 images) and on smaller subsets (925, 180, 90 images).

3.4. Precision–Recall

The precision–recall curve is employed as a metric to assess the segmentation outcomes across a range of threshold values (Figure 8). The observation is considered for the three datasets, Cityscapes, A2D2, and Mapillary Vistas, as the ground truth does not include annotations for the backsides of the signs, which affects the accuracy results.

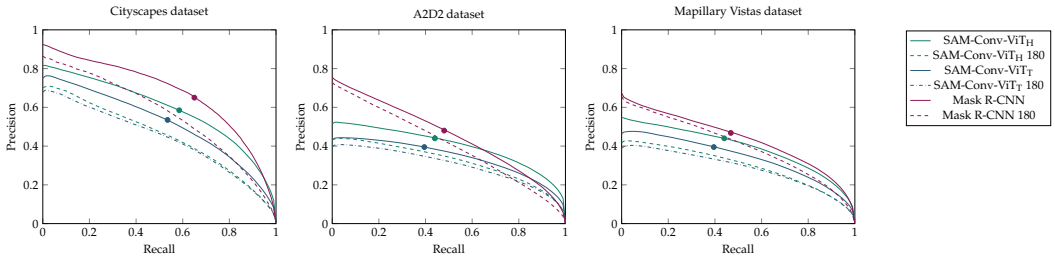


Figure 8. Precision–recall curve. The point where recall and precision are equal is marked.

As anticipated, the precision of all the models is observed to decline as recall increases. Mask R-CNN achieves the highest precision values relative to recall values for all three datasets, except for the high recall values observed in the A2D2 dataset. The graphs also demonstrate that SAM-Conv-ViT$_H$ achieves better results than SAM-Conv-ViT$_T$. The performance is quantified based on the area under the curve. The smallest reduction in precision was exhibited by SAM-Conv-ViT$_T$ (by -0.065) and Mask R-CNN (by -0.601) as a consequence of the decrease in the training data. In contrast, SAM-Conv-ViT$_H$ (by -0.105) displayed the greatest reduction. For lower thresholds, the discrepancies in precision between the datasets are more noticeable.

3.5. Sign Instances

Due to the varying sizes of the traffic signs, the segmentation results are considered for each sign instance separately. Therefore, the 9277 specific signs of the Mapillary Vistas Dataset (Section 3.1) were divided into five size categories (XS, S, M, L, XL) based on their pixel areas (Table 5). The number of detected signs with IoU values greater than 0.0 for a threshold of 0.75, irrespective of the segmentation results, is counted (det. ratio), and the IoU of these is calculated (IoU of det. signs). For instance-based results, the IoU is calculated using only the pixels in a bounding box around the specific sign instance. The bounding box is enlarged to about half the size of the actual sign. It should be noted that FP outside the bounding boxes will not be considered.

Table 5. Definition of instance sizes.

Size Category	Pixel Area A	No. of Instances
XS	$1 \leq A \leq 31$	1857
S	$32 \leq A \leq 73$	1854
M	$74 \leq A \leq 179$	1863
L	$180 \leq A \leq 521$	1850
XL	$522 \leq A \leq 74{,}343$	1853

When analyzing the segmentation results based on individual instances, it can be observed that only a relatively small number of signs are identified. The result indicates that, on average, only 32.04% of all sign instances were identified by SAM-Conv-ViT$_H$, 20.43% were identified by SAM-Conv-ViT$_T$ and 22.60% were identified by Mask R-CNN. Both the quantity (det. ratio) and the quality (IoU of det. signs) of recognized signs increase as the size of the objects increases (Table 6). This trend is recognized across all three models. It is important to note that the SAM-Conv-ViT$_H$ model consistently exhibits higher detection rates compared to the other two models across all size categories. Mask R-CNN demonstrates robust performance in terms of IoU, particularly in the L and XL categories. This performance was significantly better than the other two models. SAM-Conv-ViT$_H$ and SAM-Conv-ViT$_T$ have the strength to identify a greater number of signs, while Mask R-CNN is more adept at segmentation.

Table 6. Results based on instance size: the number of detected signs with IoU values greater than 0.0 is counted (det. ratio), and the IoU of these is calculated (IoU of det. signs). For each size category, the highest det. ratio and the highest IoU of det. signs are highlighted.

	SAM-Conv-ViT$_H$		**SAM-Conv-ViT$_T$**		**Mask R-CNN**	
Size Category	Det. ratio	IoU of det. signs	Det. ratio	IoU of det. signs	Det. ratio	IoU of det. signs
XS	**0.92%**	0.051	0.32%	0.055	0.81%	**0.060**
S	**3.83%**	0.089	1.08%	0.051	3.18%	**0.093**
M	**16.64%**	0.176	3.76%	0.102	12.18%	**0.270**
L	**53.51%**	0.323	26.59%	0.261	37.19%	**0.518**
XL	**85.48%**	0.522	75.61%	0.474	59.79%	**0.674**
	SAM-Conv-ViT$_H$ 180		**SAM-Conv-ViT$_T$ 180**		**Mask R-CNN 180**	
Size Category	Det. ratio	IoU of det. signs	Det. ratio	IoU of det. signs	Det. ratio	IoU of det. signs
XS	0.22%	0.032	0.27%	0.030	**0.43%**	**0.063**
S	0.65%	0.050	0.32%	0.058	**1.46%**	**0.173**
M	3.11%	0.130	2.09%	0.111	**5.53%**	**0.388**
L	**24.05%**	0.229	14.92%	0.212	16.92%	**0.515**
XL	**70.26%**	0.398	67.40%	0.432	42.80%	**0.684**

The reduction in training data has a minimal impact on the detection ratio and IoU of SAM-Conv-ViT$_T$ (Table 6). In the case of a high detection rate, SAM-Conv-ViT$_T$ and SAM-Conv-ViT$_H$ demonstrated superior performance. SAM-Conv-ViT$_H$ achieved the highest detection rate for the L and XL categories. The reduction in the Mask R-CNN detection ratio (by −17.0% for the XL category) is comparable to that of SAM-Conv-ViT$_H$ (−15.2% for the XL category). Notably, that the segmentation strength (as measured by IoU) of Mask R-CNN is not affected by the reduction in training data, in contrast to SAM-Conv-ViT$_H$.

4. Discussion

The segmentation of traffic signs is challenging due to their variability in shape, color, and small size. Across different model architectures, the number of detected signs is low. The smaller the size of the traffic signs, the lower the recognition rate. Among the models evaluated, the one with the greatest number of parameters, SAM-Conv-ViT$_H$, exhibited the highest overall performance, followed by Mask R-CNN and SAM-Conv-ViT$_T$. As expected, all the models demonstrated the highest prediction accuracy on the Cityscapes dataset due to its use for optimization. Mask R-CNN, which serves as the benchmark leading model, achieved the best results on the Cityscapes dataset. However, there are limitations to its adaptability to different datasets. In contrast, SAM-Conv-ViT$_H$ outperforms Mask R-CNN in accuracy across all datasets. Similarly, SAM-Conv-ViT$_T$ delivers less-impressive results overall but demonstrates the most consistent performance across different datasets. The models with a robust encoder indicate superior generalization capabilities. The analysis shows that SAM-Conv-ViT$_H$ demonstrates better segmentation accuracy than SAM-Fine-ViT$_H$. It can be concluded that the decoder of the foundation model has a notable impact, despite comprising only 1% of the parameters. The SAM-Fine decoder was initially developed as a lightweight decoder to enable real-time mask prediction based on a prompt. However, it became evident that this decoder was not suitable for the task of traffic sign segmentation.

With limited training data, Mask R-CNN, behind SAM-Fine-ViT$_H$, exhibits the most pronounced performance reduction. The results of the training data reduction indicate that SAM-Conv-ViT$_H$ demonstrates better robustness to smaller training datasets compared to Mask R-CNN. This is due to the greater number of pre-trained parameters, which are more relevant when training data are limited. SAM-Conv-ViT$_T$ benefits from knowledge distillation, showing superior generalization capabilities even with fewer parameters and demonstrating no decrease in accuracy when the training data are reduced. However, the precision–recall observation demonstrates noteworthy outcomes. Mask R-CNN exhibits a minimal reduction over the entire threshold spectrum. Contrary to expectations, SAM-Conv-ViT$_H$ shows the highest reduction over the precision–recall curve, suggesting that extensive pre-training is not essential for this task. The curve observation is more meaningful than considering a single IoU threshold at 0.75.

Instance-level observations indicate that SAM-Conv-ViT$_H$ is more effective at identifying sign instances, whereas the Mask R-CNN model demonstrates superior performance in terms of correct segmentation. A correlation is identified between the encoder and the quantity of recognized masks, as well as between the decoder and the quality of the masks. Although SAM-Conv-ViT$_T$ once again underperformed compared to SAM-Conv-ViT$_H$ and Mask R-CNN when trained on the entire dataset, it demonstrated a clear advantage in terms of data reduction, exhibiting no decrease in accuracy. The main issue with the foundation model, as with the others, is the low detection rate, particularly for smaller signs. The process of down-sampling employed by the encoder resulted in the loss of certain image details, particularly those of a smaller scale. The effective performance of Mask R-CNN can be explained using an FPN, which processes images at varying resolutions and predicts sharp-edged masks without interpolation. In contrast, SAM-Conv-ViT$_H$ relies on interpolation, leading to curved edges in the segmentation masks (Figure A1). The effects of interpolation on the quality of mask prediction are particularly pronounced in the case of small objects, such as traffic signs. To improve the segmentation ability of the SAM, a high-quality version of the SAM, HQ-SAM [59], has been developed. However, HQ-SAM does not outperform the original SAM in the case of small objects measuring less than 1024 pixels. For the Cityscapes dataset, for example, this means approximately 67% of the traffic signs fall into this category. Consequently, this area was not investigated further in this study.

5. Conclusions

The segmentation results from both the entire dataset and smaller datasets indicate that the foundation model does not clearly outperform the leading model, Mask R-CNN, in the specific field of traffic sign segmentation. This is particularly relevant when considering the drawbacks associated with the collection and annotation of extensive amounts of training data, as well as the lengthy training and processing times involved. To reduce the need for training data, our findings indicate that the knowledge-distilled model produces the most stable results. These impressive achievements demonstrate that quality is more important than quantity. There is significant redundant information in the dataset that does not advance the models. This example shows that more parameters or training data do not necessarily lead to better results. The main task is to obtain those with the most influence, with the aim of creating a generalized effect. However, applying knowledge distillation techniques to foundation models, which reduces the number of parameters and computational costs, is only applicable if a teacher model for knowledge distillation has already been established.

Author Contributions: Conceptualization, S.H. and M.K.; methodology, S.H. and M.K.; software, S.H.; validation, S.H., M.K. and B.R.; formal analysis, S.H.; investigation, S.H. and M.K.; resources, B.R. and A.R.; data curation, S.H.; writing—original draft preparation, S.H.; writing—review and editing, S.H., M.K., B.R. and A.R.; visualization, S.H.; supervision, M.K., B.R. and A.R.; project administration, M.K. and B.R.; funding acquisition, B.R. and A.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data have been created.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

CNN	convolutional neural network
FC	fully connected
FCN	fully convolutional network
FN	false negative
FP	false positive
FPN	feature pyramid network
IoU	intersection over union
ReLU	rectified linear unit
RGB	red, green, and blue
RPN	region proposal network
SAM	segment anything model
SD	standard deviation
TN	true negative
TP	true positive
ViT	vision transformer

Appendix A

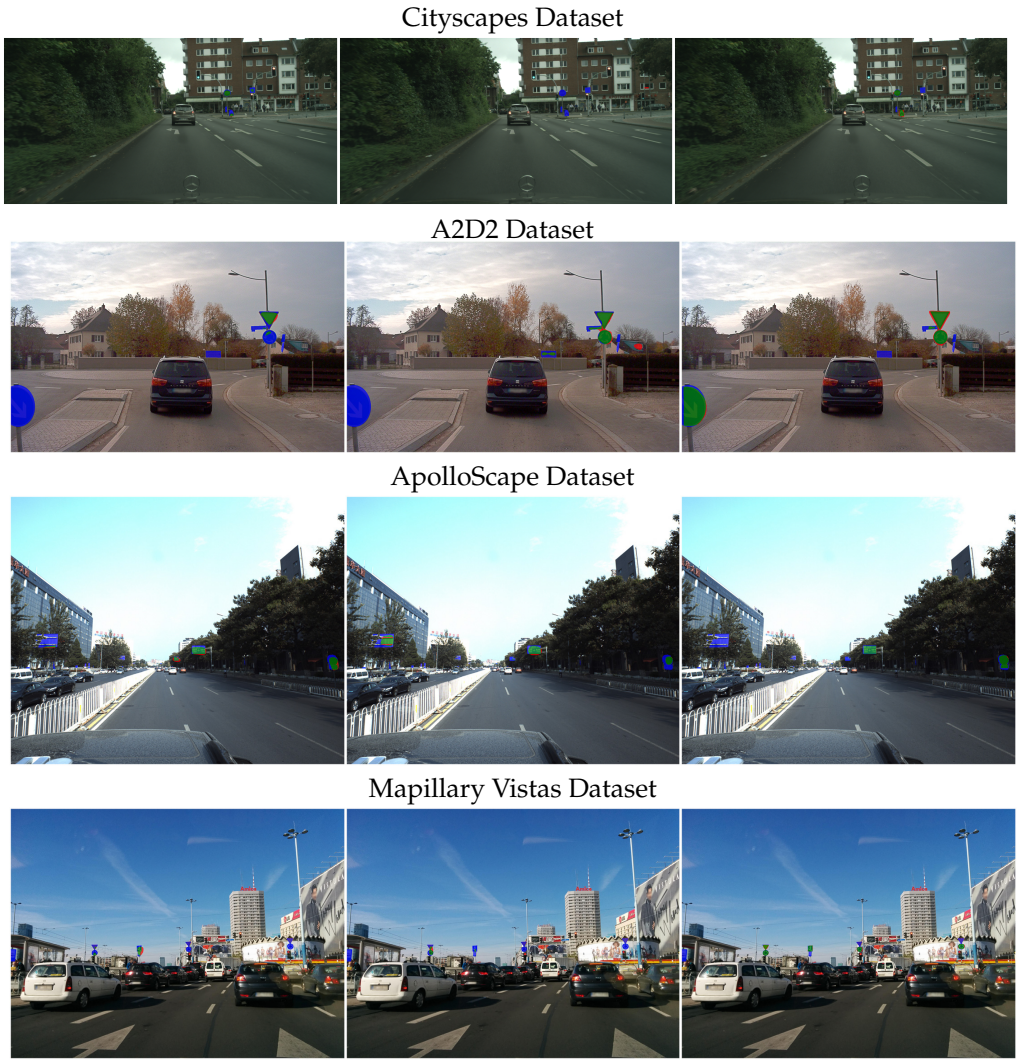


Figure A1. Examples of mask predictions for class *traffic sign* (green: TP, red: FP, blue: FN). From left to right: SAM-Conv-ViT$_H$, SAM-Conv-ViT$_T$, Mask R-CNN.

References

1. Verkehrszeichenkatalog. Available online: http://www.vzkat.de/2017/VzKat.htm (accessed on 4 January 2024).
2. Deutsche Verkehrszeichen nach StVO. Available online: https://www.verkehrszeichen-online.org/ (accessed on 4 January 2024).
3. Infrastruktur-Straßennetz. Available online: https://bmdv.bund.de/SharedDocs/DE/Artikel/G/infrastruktur-statistik.html (accessed on 12 July 2024).
4. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef] [PubMed]
5. Bommasani, R.; Hudson, D.A.; Adeli, E.; Altman, R.; Arora, S.; von Arx, S.; Bernstein, M.S.; Bohg, J.; Bosselut, A.; Brunskill, E.; et al. On the Opportunities and Risks of Foundation Models. *arXiv* **2021**, arXiv:2108.07258.
6. Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A.C.; Lo, W.Y.; et al. Segment Anything. *arXiv* **2023**, arXiv:2304.02643.
7. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16 × 16 Words: Transformers for Image Recognition at Scale. *arXiv* **2020**, arXiv:2010.11929.
8. Osco, L.; Wu, Q.; Lemos, E.; Gonçalves, W.; Ramos, A.P.; Li, J.; Marcato, J. The Segment Anything Model (SAM) for remote sensing applications: From zero to one shot. *Int. J. Appl. Earth Obs. Geoinf.* **2023**, *124*, 103540. [CrossRef]
9. Yan, Z.; Li, J.; Li, X.; Zhou, R.; Zhang, W.; Feng, Y.; Diao, W.; Fu, K.; Sun, X. RingMo-SAM: A Foundation Model for Segment Anything in Multimodal Remote-Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5625716. [CrossRef]
10. Chen, K.; Liu, C.; Chen, H.; Zhang, H.; Wenyuan, L.; Zou, Z.; Shi, Z. RSPrompter: Learning to Prompt for Remote Sensing Instance Segmentation Based on Visual Foundation Model. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 4701117. [CrossRef]

11. Chen, X.D.; Wu, W.; Yang, W.; Qin, H.; Wu, X.; Mao, X. Make Segment Anything Model Perfect on Shadow Detection. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 4410713. [CrossRef]
12. Chen, T.; Zhu, L.; Ding, C.; Cao, R.; Wang, Y.; Zhang, S.; Li, Z.; Sun, L.; Zang, Y.; Mao, P. SAM-Adapter: Adapting Segment Anything in Underperformed Scenes. In Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), Los Alamitos, CA, USA, 2–6 October 2023; pp. 3359–3367. [CrossRef]
13. Lan, X.; Lyu, J.; Jiang, H.; Dong, K.; Niu, Z.; Zhang, Y.; Xue, J. FoodSAM: Any Food Segmentation. *IEEE Trans. Multimed.* **2023**, 1–14. [CrossRef]
14. Ren, B.; Qian, Z.; Sun, Y.; Gao, C.; Zhang, C. WebSAM-Adapter: Adapting Segment Anything Model for Web Page Segmentation. In *Advances in Information Retrieval, 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10–14, 2022, Proceedings, Part I*; Goharian, N., Tonellotto, N., He, Y., Lipani, A., McDonald, G., Macdonald, C., Ounis, I., Eds.; Springer: Cham, Switzerland, 2024; pp. 439–454.
15. Wu, W.; Wong, M.S.; Yu, X.; Shi, G.; Kwok, C.Y.T.; Zou, K. Compositional Oil Spill Detection Based on Object Detector and Adapted Segment Anything Model From SAR Images. *IEEE Geosci. Remote Sens. Lett.* **2024**, *21*, 4007505. [CrossRef]
16. Hu, J.; Lin, J.; Cai, W.; Gong, S. Relax Image-Specific Prompt Requirement in SAM: A Single Generic Prompt for Segmenting Camouflaged Objects. *arXiv* **2023**, arXiv:2312.07374. [CrossRef]
17. Zhang, Y.; Shen, Z.; Jiao, R. Segment anything model for medical image segmentation: Current applications and future directions. *Comput. Biol. Med.* **2024**, *171*, 108238. [CrossRef]
18. Leng, T.; Zhang, Y.; Han, K.; Xie, X. Self-Sampling Meta SAM: Enhancing Few-shot Medical Image Segmentation with Meta-Learning. In Proceedings of the 2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 3–8 January 2024; pp. 7910–7920. [CrossRef]
19. Nakhaei, N.; Zhang, T.; Terzopoulos, D.; Hsu, W. Refining boundaries of the segment anything model in medical images using an active contour model. In *Proceedings of the Medical Imaging 2024: Computer-Aided Diagnosis*; Chen, W., Astley, S.M., Eds.; International Society for Optics and Photonics, SPIE: Bellingham, WA, USA, 2024; Volume 12927, p. 1292734. [CrossRef]
20. Giorgi, N. Kidney and Kidney Tumor Segmentation via Transfer Learning. In *Kidney and Kidney Tumor Segmentation, MICCAI 2023 Challenge, KiTS 2023, Held in Conjunction with MICCAI 2023, Vancouver, BC, Canada, October 8, 2023, Proceedings*; Heller, N., Wood, A., Isensee, F., Rädsch, T., Teipaul, R., Papanikolopoulos, N., Weight, C., Eds.; Springer: Cham, Switzerland, 2024; pp. 156–162.
21. Shi, P.; Qiu, J.; Abaxi, S.M.D.; Wei, H.; Lo, F.P.W.; Yuan, W. Generalist Vision Foundation Models for Medical Imaging: A Case Study of Segment Anything Model on Zero-Shot Medical Segmentation. *Diagnostics* **2023**, *13*, 1947. [CrossRef] [PubMed]
22. Na, S.; Guo, Y.; Jiang, F.; Ma, H.; Huang, J. Segment Any Cell: A SAM-based Auto-prompting Fine-tuning Framework for Nuclei Segmentation. *arXiv* **2024**, arXiv:2401.13220.
23. Wu, Q.; Zhang, Y.; Elbatel, M. Self-prompting Large Vision Models for Few-Shot Medical Image Segmentation. In *Domain Adaptation and Representation Transfer, 5th MICCAI Workshop, DART 2023, Held in Conjunction with MICCAI 2023, Vancouver, BC, Canada, October 12, 2023, Proceedings*; Springer: Cham, Switzerland, 2024; pp. 156–167.
24. Cui, R.; He, S.; Qiu, S. Adaptive Low Rank Adaptation of Segment Anything to Salient Object Detection. *arXiv* **2023**, arXiv:2308.05426.
25. Shu, H.; Li, W.; Tang, Y.; Zhang, Y.; Chen, Y.; Li, H.; Wang, Y.; Chen, X. TinySAM: Pushing the Envelope for Efficient Segment Anything Model. *arXiv* **2023**, arXiv:2312.13789
26. Zhu, Z.; Liang, D.; Zhang, S.; Huang, X.; Li, B.; Hu, S. Traffic-Sign Detection and Classification in the Wild. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 2110–2118. [CrossRef]
27. Timbuş, C.; Miclea, V.; Lemnaru, C. Semantic segmentation-based traffic sign detection and recognition using deep learning techniques. In Proceedings of the 2018 IEEE 14th International Conference on Intelligent Computer Communication and Processing (ICCP), Cluj-Napoca, Romania, 6–8 September 2018; pp. 325–331. [CrossRef]
28. Guan, H.; Yu, Y.; Peng, D.; Zang, Y.; Lu, J.; Li, A.; Li, J. A Convolutional Capsule Network for Traffic-Sign Recognition Using Mobile LiDAR Data with Digital Images. *IEEE Geosci. Remote Sens. Lett.* **2020**, *17*, 1067–1071. [CrossRef]
29. Saha, S.; Amit Kamran, S.; Shihab Sabbir, A. Total Recall: Understanding Traffic Signs Using Deep Convolutional Neural Network. In Proceedings of the 2018 21st International Conference of Computer and Information Technology (ICCIT), Dhaka, Bangladesh, 21–23 December 2018; pp. 1–6. [CrossRef]
30. Dhar, P.; Abedin, M.Z.; Biswas, T.; Datta, A. Traffic sign detection—A new approach and recognition using convolution neural network. In Proceedings of the 2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC), Dhaka, Bangladesh, 21–23 December 2017; pp. 416–419. [CrossRef]
31. Lee, H.S.; Kim, K. Simultaneous Traffic Sign Detection and Boundary Estimation Using Convolutional Neural Network. *IEEE Trans. Intell. Transp. Syst.* **2018**, *19*, 1652–1663. [CrossRef]
32. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. In Proceedings of the International Conference on Learning Representations, Computational and Biological Learning Society, San Diego, CA, USA, 7–9 May 2015; pp. 1–14.

33. Arcos-García, Á.; Álvarez García, J.A.; Soria-Morillo, L.M. Evaluation of deep neural networks for traffic sign detection systems. *Neurocomputing* **2018**, *316*, 332–344. [CrossRef]
34. Zhang, Z.; Zhou, X.; Chan, S.; Chen, S.; Liu, H. Faster R-CNN for Small Traffic Sign Detection. In *Computer Vision, Second CCF Chinese Conference, CCCV 2017, Tianjin, China, October 11–14, 2017, Proceedings, Part III*; Yang, J., Hu, Q., Cheng, M.M., Wang, L., Liu, Q., Bai, X., Meng, D., Eds.; Springer: Singapore, 2017; pp. 155–165.
35. Shao, F.; Wang, X.; Meng, F.; Zhu, J.; Wang, D.; Dai, J. Improved Faster R-CNN Traffic Sign Detection Based on a Second Region of Interest and Highly Possible Regions Proposal Network. *Sensors* **2019**, *19*, 2288. [CrossRef]
36. Cao, C.; Wang, B.; Zhang, W.; Zeng, X.; Yan, X.; Feng, Z.; Liu, Y.; Wu, Z. An Improved Faster R-CNN for Small Object Detection. *IEEE Access* **2019**, *7*, 106838–106846. [CrossRef]
37. Li, J.; Wang, Z. Real-Time Traffic Sign Recognition Based on Efficient CNNs in the Wild. *IEEE Trans. Intell. Transp. Syst.* **2019**, *20*, 975–984. [CrossRef]
38. Zhang, J.; Xie, Z.; Sun, J.; Zou, X.; Wang, J. A Cascaded R-CNN With Multiscale Attention and Imbalanced Samples for Traffic Sign Detection. *IEEE Access* **2020**, *8*, 29742–29754. [CrossRef]
39. Yao, J.; Chu, Y.; Xiang, X.; Huang, B.; Xiaoli, W. Research on detection and classification of traffic signs with data augmentation. *Multimed. Tools Appl.* **2023**, *82*, 38875–38899. [CrossRef]
40. Tabernik, D.; Skočaj, D. Deep Learning for Large-Scale Traffic-Sign Detection and Recognition. *IEEE Trans. Intell. Transp. Syst.* **2019**, *21*, 1427–1440. [CrossRef]
41. Kamal, U.; Tonmoy, T.I.; Das, S.; Hasan, M.K. Automatic Traffic Sign Detection and Recognition Using SegU-Net and a Modified Tversky Loss Function With L1-Constraint. *IEEE Trans. Intell. Transp. Syst.* **2020**, *21*, 1467–1479. [CrossRef]
42. Badrinarayanan, V.; Kendall, A.; Cipolla, R. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495. [CrossRef]
43. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015*; Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F., Eds.; Springer: Cham, Switzerland, 2015; pp. 234–241.
44. Wang, H. Traffic Sign Recognition with Vision Transformers. In Proceedings of the ICISDM'22: 6th International Conference on Information System and Data Mining, Silicon Valley, CA, USA, 27–29 May 2022; pp. 55–61. [CrossRef]
45. Manzari, O.N.; Kashiani, H.; Dehkordi, H.A.; Shokouhi, S.B. Robust transformer with locality inductive bias and feature normalization. *Eng. Sci. Technol. Int. J.* **2023**, *38*, 101320. h.2022.101320 [CrossRef]
46. Manzari, O.N.; Boudesh, A.; Shokouhi, S.B. Pyramid Transformer for Traffic Sign Detection. In Proceedings of the 2022 12th International Conference on Computer and Knowledge Engineering (ICCKE), Mashhad, Iran, 17–18 November 2022; pp. 112–116. [CrossRef]
47. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2980–2988. [CrossRef]
48. Lin, T.; Dollar, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 936–944. [CrossRef]
49. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. PyTorch: An imperative style, high-performance deep learning library. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*; Curran Associates Inc.: Red Hook, NY, USA, 2019.
50. Buslaev, A.; Iglovikov, V.I.; Khvedchenya, E.; Parinov, A.; Druzhinin, M.; Kalinin, A.A. Albumentations: Fast and Flexible Image Augmentations. *Information* **2020**, *11*, 125. [CrossRef]
51. Ma, J.; Chen, J.; Ng, M.; Huang, R.; Li, Y.; Li, C.; Yang, X.; Martel, A.L. Loss odyssey in medical image segmentation. *Med. Image Anal.* **2021**, *71*, 102035. [CrossRef]
52. Kingma, D.; Ba, J. Adam: A Method for Stochastic Optimization. In Proceedings of the International Conference on Learning Representations, Banff, AB, Canada, 14–16 April 2014.
53. Loshchilov, I.; Hutter, F. SGDR: Stochastic Gradient Descent with Warm Restarts. In Proceedings of the International Conference on Learning Representations, 2017, Toulon, France, 24–26 April 2017.
54. Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; Schiele, B. The Cityscapes Dataset for Semantic Urban Scene Understanding. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 3213–3223.
55. Geyer, J.; Kassahun, Y.; Mahmudi, M.; Ricou, X.; Durgesh, R.; Chung, A.S.; Hauswald, L.; Pham, V.H.; Mühlegg, M.; Dorn, S.; et al. A2D2: Audi Autonomous Driving Dataset. *arXiv* **2020**, arXiv:2004.06320
56. Neuhold, G.; Ollmann, T.; Bulò, S.R.; Kontschieder, P. The Mapillary Vistas Dataset for Semantic Understanding of Street Scenes. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 5000–5009.
57. Huang, X.; Cheng, X.; Geng, Q.; Cao, B.; Zhou, D.; Wang, P.; Lin, Y.; Yang, R. The ApolloScape Dataset for Autonomous Driving. *arXiv* **2018**, arXiv:1803.06184

58. Alberti, E.; Tavera, A.; Masone, C.; Caputo, B. IDDA: A Large-Scale Multi-Domain Dataset for Autonomous Driving. *IEEE Robot. Autom. Lett.* **2020**, *5*, 5526–5533. [CrossRef]
59. Ke, L.; Ye, M.; Danelljan, M.; Liu, Y.; Tai, Y.W.; Tang, C.K.; Yu, F. Segment anything in high quality. In Proceedings of the NIPS'23: 37th International Conference on Neural Information Processing Systems, New Orleans, LA, USA, 10–16 December 2023.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.