BiSeSAM for Road Segmentation
CIL ETHZ Road Segmentation Kaggle Challenge 2024

Jannek Ulm Douglas Orsini-Rosenberg Raoul van Doren Paul Ellsiepen

Kaggle Team: Long way to go

Abstract

Vision Transformers (ViTs) have recently started to appear on top of vision task leaderboards. We extend ideas and models from cutting-edge vision transformers to present **BiSeSAM** - an efficient and novel road segmentation model. **BiSeSAM** embeds the generalized power of Meta's Segment Anything (SAM) image encoder in a custom architecture, with several plug-in decoders. Trained with our manually generated dataset, **BiSeSAM** achieves on-par performance with state-of-the-art models, and out-performs them in some configurations.

1. Introduction

Since they have been released, CNNs have been the dominating architecture for most Computer Vision Tasks. Recently, transformer models have gained popularity, and have started to out-perform CNNs for classical vision problems. Inspired by this trend, we explore the performance of Vision Transformers (ViTs [2]) for tackling the road segmentation problem, i.e. the binary semantic segmentation of satellite images.

We harness the power of Meta's Segment Anything Model – SAM [4]) which we adapt and extend for this specific problem. We propose a variety of model architectures **(BiSeSAM)** that build on SAM's image encoder.

Our contributions are as follows:

1. Proposing a new model **BiSeSAM**: SAM image encoder with various custom decoders (MLP, Conv, Spatially Aware, Skip Connection).
2. Generating a dataset of $>$12k images and corresponding groundtruth road masks.
3. Evaluating different versions of BiSeSAM and comparing them to baselines.

The main idea of BiSeSAM is to combine a "strong" image encoder (SAM) - trained and purposed for generalized segmentation - with our task-specific decoders. We choose SAM as a cutting-edge segmentation model and leverage its great image embedding capabilities. SAM has been trained on 256 A100 GPUS for 68 hours [4]. We use the pre-trained SAM encoder as the base for BiSeSAM and embed it in a custom model architecture, including: MLPs, transposed Convolutions and Skip Connections.

Our approach yields on-par performance with our implementations of baseline models (UNet,UNet++), and achieves the 1st place in the CIL Kaggle competition. Thus, our work follows the trend of Vision Transformers enabling performance improvements and innovation.

2. Related Work

Vision Transformer(s). Recently, several works leverage transformers for vision tasks. As the classic transformer architecture operates on 1-dimensional sequences [9], a major design challenge is how to present image data to the encoder. The pioneering work, ViT [2], simply appends a positional embedding to a linear projection of image patches[1]. Already being data-hungry by design, vision transformers furthermore observe the difficulty of an extremely irregular information density in most data (including our domain of aerial images), leading to a slower learning process of relevant semantic components [3]. We implicitly address this with our usage of SAM's image encoder which was pre-trained on the large, and model-in-the-loop generated, SA-1B dataset [4].

Segformer. As a specific adjustment for image segmentation, Segformer also uses residual ("skip") connections from several layers (i.e. different focus points) in the encoder as input to the decoder [10]. The ViT paper already identified that early layers in encoder tend to attend locally[2], similarly to early layers in CNN, while higher layers almost exclusively "pay attention" globally. This availability of global and localized information is particularly valuable for image (and road) segmentation - local information is essential for identifying (potential) element boundaries while precise labeling and boundary decisions rely on (more) global information [8].

Our work also makes use of skip connections for some of the custom decoders, while ensuring that sufficient semantic knowledge is represented in the image embedding by using the pre-trained SAM encoder. In contrast to spatial resolution changes and patch merging employed by Segformer [10], the SAM image encoder uses a fixed spatial resolution with global and local information sharing through window attention.

[1] For a visualization, see [2; 3].

[2] The analogous concept to receptive fields in CNN-based architectures. This can be derived from the attention weights at the various layers - with multi-head self-attention, even more data points are available.

3. BiSeSAM - Model Architecture

3.1. META - Segment Anything Model

SAM is a vision model created by Meta for prompt-based image segmentation and consists of three parts: an image encoder, a prompt encoder and a mask decoder. The image encoder and prompt encoder create embeddings separately. Both are used as an input to the mask decoder, which then produces segmentation masks.

Image Encoder. SAM utilizes a Vision Transformer (ViT) that is based on a masked auto-encoder (MAE) [3] pre-trained model. The encoder first breaks down the input image (1024x1024 pixels with 3 channels) into 64x64 non-overlapping patches (each 16x16 pixels) and then maps each patch to a 768-dimensional embedding space through a convolutional layer. Positional embeddings are added to preserve spatial information. The 12-ViT blocks handle the sequence of patch embeddings. ViT Blocks 2,5,8 and 11 use global attention w.r.t. the other patches, while the other blocks have a window size of 14 [4]. The output of the last transformer block is compressed by a convolutional "neck" layer, producing a final output of shape (64,64,256).

Prompt Encoder. The prompt encoder transforms prompts into embedding vectors. Prompts can be points, boxes and masks.

Mask Decoder. The mask decoder merges the embeddings from the image encoder and prompt encoder to create segmentation masks. It employs a transformer decoder block with self attention and cross-attention mechanisms along with a dynamic mask prediction head that suggests multiple masks for different prompts.

While SAM's mask decoder is powerful for various prompt-based tasks, it requires explicit prompts to generate masks. Since road segmentation only requires road masks, there is no need for a generic mask decoder. Therefore we propose using SAM's ViT image encoder and replacing the prompt encoder + mask decoder with different specialized decoders, as shown in Fig 1. This way, we can still benefit from the feature extraction capabilities of the pre-trained ViT image encoder.

3.2. Decoders

While the SAM mask decoder uses a transformer decoder block, we refrained from trying to train such an architecture from scratch. While CNN's have an inherently large inductive bias, the transformer architecture needs vast amount data and compute to achieve competitive results [1; 2]. Hence we propose and compare a variety of other Decoders: a standard MLP, a Convolutional Decoder, a MLP with Spatial awareness, and a MLP with skip connections, that uses intermediate results from the image encoder in addition to the final image embedding.

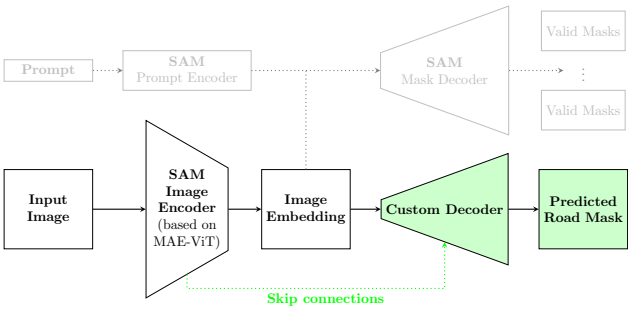

Figure 1. Proposed Model Architecture: *(Original SAM Model architecture in grey, proposed novelties in green)*

3.2.1. Vanilla MLP Decoder

This decoders uses a simple fully connected multi-layer perceptron (MLP). It consists of 6 layers with LeakyReLU activation functions that maps the ViT's 256-dim image embedding for each 16x16 patch directly to the corresponding patch in the segmentation mask output.

3.2.2. Transposed Convolution Decoder

Additionally, we experimented with a convolutional decoder proposed by [5]. It includes a series of transposed 2d convolution layers (over the 256-channel 64x64 encoded image). Upsampling is performed to increase the spatial resolution of the ViT embeddings while learning to generate segmentation boundaries.

3.2.3. Spatially Aware MLP Decoder

To improve performance, we extend the standard MLP Decoder with enhanced contextual awareness to generate the mask for each path: In addition to the patch currently being decoded, the decoder also receives the 4 neighboring patches as inputs in Spatial-Aware-S, as can be seen in Fig 2. We also implement a full version (Spatial Aware-F), where the decoder incorporates the 8 surrounding patches. We believe this benefits the model's performance since road masks overlap with the neighboring patches, allowing the decoder to generate better transitions between patches.

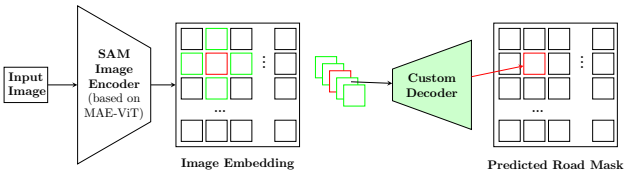

Figure 2. Spatially Aware Decoder

3.2.4. Skip Connection MLP Decoder

For this decoder we employ skip connections for each patch from intermediate states from within the image encoder (756-dim). The intermediate patch representations (between two ViT-blocks) are concatenated together with the final (256-dim) patch embedding to a patch embedding of form (3x756+256-dim). This richer representation is then used by a patch wise 6-layer leaky ReLu MLP to produce the patch's mask prediction. The skip connections use the outputs of global attention blocks 2,5 and 8 as input. The intermediate embedding states after global attention blocks are represented by green boxes in Fig 3.

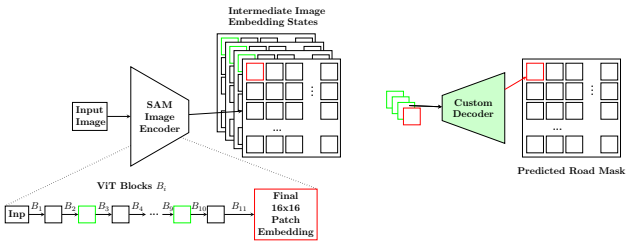

Figure 3. Skip Connection MLP Decoder

3.3. Ensemble

We also used a simple majority vote ensemble that combines the outputs of each BiSeSAM model (differing in the decoder) to predict the final output.

4. Methods

4.1. Data

The Kaggle dataset contains 144 RGB images and their masks. We refer to this dataset as **kaggle**.

To improve the quality of our training we build our own more extensive dataset, using the Google Maps API. We determine the required scale for the images and download an additional 12,990 satellite images and their corresponding road masks from the following U.S. cities: Boston, New York City, Philadelphia, and Austin. Images with less than 3% road pixels were removed. We refer to this secondary dataset with a total of 11500 samples as **gmaps**.

Moreover, we perform data augmentation by randomly rotating the images along their axes. For preprocessing, we rely on SAM's standard preprocessing pipeline, which normalizes pixel values and ensures the inputs are square. All pictures in the dataset have size 400x400. To make them compatible with the SAM image encoder, we upsample to 1024x1024 to train BiSeSAM. Our dataset has a class imbalance of 85% non-road to 15% road pixels. For more details, see Appendix C.

4.2. Training

All models are trained using the Adam Optimizer with a decaying learning rate and a batch size of 5. We train 14 epochs on **gmaps** and 15 finetune epochs on **kaggle**. The layers of BiSeSAM are gradually unfrozen, for more details see Appendix A. Note that due to limited resources we could only afford to tune hyperparameters w.r.t. the loss function and learning rate, and could complete the full training pipeline for each model only once.

4.3. Loss Function

We test various combinations of loss functions including weighted BCE and Dice Loss. All tested functions perform similarly. Further research [11; 6] suggests that the combination $loss = BCE + Dice$ is robust in general, hence, we choose this loss. The convex nature of BCE helps the complex non-convex loss landscape of Dice loss. A quantification of these results can be found in Appendix D.

Furthermore, we explored Focal loss as a promising segmentation loss function [11]. However, it did not yield promising results in preliminary testing. We hypothesize that this is due to the nature of our dataset[3].

4.4. Baselines

As our baseline models we implement the traditional state-of-the-art architectures UNet and UNet++[4]. Both are initialized with two different encoder backbones, `efficientnet-b5` and `resnet34`, resulting in four separate baseline models. As initial model checkpoint, we use pre-trained ImageNet weights. The optimizer and training data is equivalent to BiSeSAM. For more details on training, see B.

5. Results

In the following, we present our findings to compare and evaluate various decoders.

Training all decoders according to the schedule defined in section 4.2, we obtain the results displayed in table 5.

To measure performance we report the mean F1-Score of our algorithms on a validation split of **kaggle**, as well as the public Kaggle F1-score achieved by the algorithm's submission. We compare the baseline implementations (UNet, UNet++) with BiSeSAM in its various decoder configurations.

[3]Focal loss performs well for exceptionally unbalanced datasets - in our dataset we observe "only" an imbalance of 85% negatively to 15% positively labeled pixels.

[4]More information about the architecture of UNet and UNet++ can be found in Ronneberger's paper "U-Net: Convolutional Networks for Biomedical Image Segmentation" [7] and Zhou's paper "UNet++: A Nested U-Net Architecture for Medical Image Segmentation" [12].

Table 1. Model Results

Model	F1 - Test	F1 - Kaggle
Unet - ResNet	tbd	tbd
Unet++ - ResNet	tbd	tbd
Unet - EfficientNet	94.00	93.43
Unet++ - EfficientNet	93.61	93.27
BiSeSAM - Conv	94.11	93.33
BiSeSAM - MLP	94.06	93.25
BiSeSAM - Spatial Aware-S	94.01	93.21
BiSeSAM - Spatial Aware-F	94.09	93.15
BiSeSAM - Skip Connect	94.07	93.26
All-BiSeSAM - Majority Ens.	**94.19**	**93.63**

The experimental results demonstrate that the BiSeSAM models generally perform equally well, independent of the choice of decoder. They all performed better than the UNet and UNet++ baseline based on `resnet34`. However, combining UNet and UNet++ with the `efficientnet-b5` backbone always performed better than the BiSeSAM models. Additionally, the ensemble using a majority voting performed better than all the other individuals models, including the baseline models.

Figure 4. Qualitative Analysis. MLP-BiSeSAM on unseen Worcester sample: TP (Green), FP (Red), FN (Blue)

6. Discussion

In terms of the building blocks of our model, the SAM image encoder appears to constitute the largest contributor to the (good) performance of `BiSeSAM`. As shown in table 5, different choices for the decoder do not lead to significant changes in performance. In the scope of this project, we could not complete enough training pipeline iterations to perform a meaningful statistical analysis of the different decoders as their single trial F1 scores range in the same tenth of a percentage point. As initially assumed when choosing the SAM image encoder, the nonetheless comparable performance to state-of-the-art baseline models can be explained by SAM's image embedding capabilities - the result of catalyzing an optimized ViT modification with a huge, and smartly chosen dataset [4].

The performance of UNet and UNet++ seems on par with our transformer-based approach (i.e. not significantly worse), at least for our training and test data. On the one hand, we attribute this to the established observation that the more traditional CNN-based approaches perform better with low(er) amounts of data as Transformers need *more training*, learning the semantic structuring of the input domain to inter alia compensate for their lack of translation equivariance [2] or lack of 2D locality awareness. On the other hand, it is simply expected that state-of-the-art image classifiers perform well on a relatively standard image segmentation task.

7. Summary

In this work, we showed that transformers are likely to remain the most promising approach for image segmentation problems. Our results further indicate that the amount and quality of data are essential for the performance of vision transformers: 1. SAM as a generalizing image encoder fed with vast amounts of data is powerful enough to enable state-of-the-art performance in our specialized road segmentation problem. 2. more specialized CNN-based approaches in UNet and UNet++ out-perform the non-ensemble versions of BiSeSAM on the same data.

We are convinced that our results do not constitute the limits of the potential of transformers - combining more data and compute, crafting and fine-tuning more specialized encoders, augmenting transformers with sophisticated pre- and post-processing steps, or exploring innovative ideas in the application of 1D performers to 2D domains (such as images) are all promising future advances for binary image segmentation that make an interesting near future for the vision research community.

References

[1] Deininger, L., Stimpel, B., Yuce, A., Abbasi-Sureshjani, S., Schönenberger, S., Ocampo, P., Korski, K., and Gaire, F. A comparative study between vision transformers and cnns in digital pathology, 2022. URL https://arxiv.org/abs/2206.00389.

[2] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

[3] He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R. Masked autoencoders are scalable vision learners. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15979–15988, 2022. doi: 10.1109/CVPR52688.2022.01553.

[4] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W., Dollár, P., and Girshick, R. B. Segment anything. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pp. 3992–4003. IEEE, 2023. doi: 10.1109/ICCV51070.2023.00371. URL https://doi.org/10.1109/ICCV51070.2023.00371.

[5] Nalkar, Y. R. Promptless-taskspecific-finetuning of metaai sam, 2024. URL https://www.kaggle.com/code/yogendrayatnalkar/promptless-taskspecific-finetuning-of-metaai-sam.

[6] Rajput, V. Robustness of different loss functions and their impact on networks learning capability, 2021. URL https://arxiv.org/abs/2110.08322.

[7] Ronneberger, O., Fischer, P., and Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Navab, N., Hornegger, J., III, W. M. W., and Frangi, A. F. (eds.), *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015 - 18th International Conference Munich, Germany, October 5 - 9, 2015, Proceedings, Part III*, volume 9351 of *Lecture Notes in Computer Science*, pp. 234–241. Springer, 2015. doi: 10.1007/978-3-319-24574-4_28. URL https://doi.org/10.1007/978-3-319-24574-4_28.

[8] Strudel, R., Garcia, R., Laptev, I., and Schmid, C. Segmenter: Transformer for semantic segmentation. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7242–7252, 2021. doi: 10.1109/ICCV48922.2021.00717.

[9] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

[10] Xie, E., Wang, W., Yu, Z., Anandkumar, A., Álvarez, J. M., and Luo, P. Segformer: Simple and efficient design for semantic segmentation with transformers. In Ranzato, M., Beygelzimer, A., Dauphin, Y. N., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 12077–12090, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/64f1f27bf1b4ec22924fd0acb550c235-Abstract.html.

[11] Xu, H., He, H., Zhang, Y., Ma, L., and Li, J. A comparative study of loss functions for road segmentation in remotely sensed road datasets. *Int. J. Appl. Earth Obs. Geoinformation*, 116:103159, 2023. doi: 10.1016/J.JAG.2022.103159. URL https://doi.org/10.1016/j.jag.2022.103159.

[12] Zhou, Z., Siddiquee, M. M. R., Tajbakhsh, N., and Liang, J. Unet++: A nested u-net architecture for medical image segmentation. In Stoyanov, D., Taylor, Z., Carneiro, G., Syeda-Mahmood, T. F., Martel, A. L., Maier-Hein, L., Tavares, J. M. R. S., Bradley, A. P., Papa, J. P., Belagiannis, V., Nascimento, J. C., Lu, Z., Conjeti, S., Moradi, M., Greenspan, H., and Madabhushi, A. (eds.), *Deep Learning in Medical Image Analysis - and - Multimodal Learning for Clinical Decision Support - 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings*, volume 11045 of *Lecture Notes in Computer Science*, pp. 3–11. Springer, 2018. doi: 10.1007/978-3-030-00889-5_1. URL https://doi.org/10.1007/978-3-030-00889-5_1.

A. Training Procedure - BiSeSAM

Adam Optimizer with $(\beta_1, \beta_2) = (0.9, 0.999)$, no weight decay and the following learning rate/epoch procedure:

All models were trained with a Nvidia RTX 3090 GPU, training one model takes $\sim$ 9 hours.

———— Training Procedure: pro model 8h training_set_whole = 11.5k images

batch_size = 5 SAM encoder has 176 layers (weights) num_layers_front, num_layers_back 3x epoch 0-25, lr = 0.001 3x epoch 0-65 lr = 0.0001 4x epoch 15-85 lr = [0.0001, 0.0001/2, 0.0001/ 3, 0.0001 /4] 4x epoch 25-105 lr = [0.0001, 0.0001/2, 0.0001/ 3, 0.0001 /4]

Report -¿ kaggle data [loss, f1, iou]

————

pro model 5min Finetuning:

training with original_split_0.8 validation with original_split_0.2

10x epoch 25-105 lr = [0.0001]

report -¿ of valid split: [loss, f1, iou]

————-

B. Training Procedure - Baselines

This here trains Unet and Unet ++

C. Data

Austin 1370,Boston 6500, NYC 1870, Philadelphia 3250 = 12990 total

D. Hyperparameter Tuning - Loss Function

Table 2. Loss Functions

Loss Function	Indicative F1 - Kaggle
BCE	93.00
BCE + Dice	93.08