Multimodal Foundation Models as Router Models for High-Resolution Aerial Image Segmentation

Cooper Li
Montgomery Blair High School
cligeog@umd.edu

Zhihao Wang
University of Maryland
zhwang1@umd.edu

Yiqun Xie
University of Maryland
xie@umd.edu

***Abstract*—Advances in remote sensing with higher-resolution images provide new opportunities for monitoring tasks in forestry, ecology, agriculture, energy, etc. However, applications at large scale often face the challenge of spatial variability, where patterns vary over geographic regions. Existing methods often assume a fixed, known spatial partitioning as input, which is mostly unavailable in practice, or exhaust models over locations that are computationally prohibitive for large-scale Earth monitoring. We propose an adaptive router approach, which leverages large, pretrained multimodal foundation models as a model router to automatically select scenario-specialized segmentation models for each input image. We evaluate our method using high-resolution tree mortality mapping, which has been recognized as a critical gap in large-scale carbon monitoring. The new method shows over 30% improvement in polygon F1 score over baselines including vision foundation models.**

***Index Terms*—Multimodal Foundation Models, Router Models, Ecological Monitoring, Segmentation, Aerial Imagery**

I. Introduction

As remote sensing technologies continue to advance–with ever-growing resolution and scale–satellite and aerial images have enabled valuable monitoring capabilities for important applications, such as forestry, agriculture, ecology, energy, water, etc. In particular, higher resolution images, together with the revolution of AI, have provided new opportunities to collect fine-grained information that was previously infeasible via lower-resolution platforms.

However, broad-scale applications of machine learning methods to remote sensing images face the challenge of spatial variability. Varying terrains, geographical context and spatial patterns make it challenging for a single model to perform well and stably in different geographic regions [1].

Several directions in related work have tried to mitigate the challenge: (i) predefined partitions, which are often used for study sites far away from each other with separate models trained for different locations [2]. However, in real-world applications, the partitions are often unknown as inputs, and users need to deal with a mixture of scenarios over large areas. and (ii) geographically-weighted models, which by design generates a separate model at every location to address variability [3]. Due to the high computational cost, this is only applicable to simpler models (e.g., linear models) or studies with limited locations (e.g., a hundred). (iii) adaptive partitioning [4], which can automatically learn spatial partitions to build local models. However, these methods require integration of a deep learning model architecture into the partitioning framework, which needs additional designs for different types of models, resulting in significant engineering complexity.

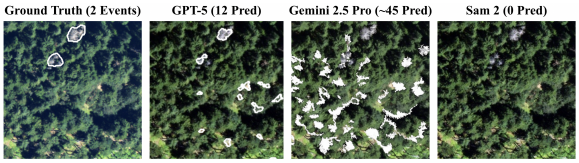

Fig. 1. Limited performance of foundation models for remote sensing image segmentation.

With the success of ChatGPT, various general-purpose foundation models have been developed. However, these models are often less effective for specialized remote sensing tasks because of a gap between natural and remote sensing images [5], [6]. For example, SAM2 [7], a foundation model for segmentation, shows limited ability in segmenting tree mortality events according to Fig. 1. We propose a new way to better leverage the ability of large multimodal foundation models (MMFM) such as GPT4o and Gemini to improve remote sensing image segmentation under spatial variability. Specifically, our approach utilizes MMFM as model routers to select the most suitable segmentation model (e.g., vision transformers like SegFormer) trained for different scenarios. This easy-to-implement process can be directly applied to each of the test images during inference and does not require changes to the segmentation model architectures.

We evaluate our method using high-resolution tree mortality mapping, which has been recognized as a critical gap in large-scale carbon monitoring. The new method shows over 30% improvement in polygon F1 score over baselines including vision foundation models.

II. Problem Formulation

A. Task Setup

The general problem is formulated as follows:

- **Input:** High-resolution remote sensing images at different locations.
- **Output:** Pixel-level classification (i.e., segmentation masks) for each image.
- **Goal:** Maximizing model performance (e.g., F1 score).

2375-9259/25/$31.00 ©2025 IEEE
DOI 10.1109/ICDMW69685.2025.00385

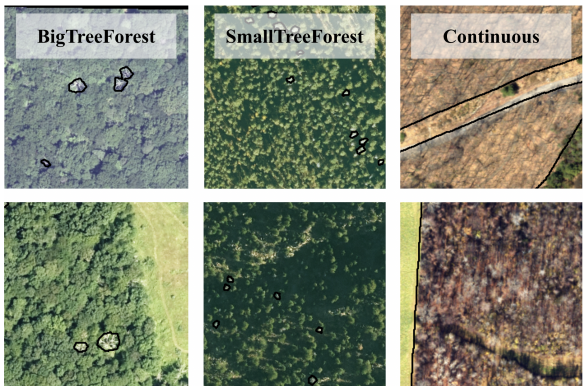

Fig. 2. Comparison between the three different scenarios.

A desired method should perform well across differing regions despite spatial variability.

B. A Concrete Application Example

To better illustrate our method throughout this study, we will use an important monitoring task as a concrete example: tree mortality event monitoring using high-resolution remote sensing images. Existing tree mortality methods mostly focus on large-patch wipe-outs (i.e., tree deaths in a large contiguous area) using moderate resolution images, such as Landsat at 30m resolution. However, recent studies found that these products cannot identify scattered tree death events. As pointed out by a recent Nature Communications paper [8]. This gap is critical: 60% of dead trees in California occurred in small groups. Such scattered tree deaths contribute substantially to forest degradation, and they can only be observed using high-resolution data (e.g., sub-meter level).

In the following text, we will use this application as an example to introduce our method, testing on both large-patch wipe-outs and scattered tree death events. Our inputs will be high-resolution NAIP images (0.6m resolution), and the output will be segmentation masks of tree mortality events.

III. Method

A. MARS Framework

1) General design: The goal of the framework is to develop an easy-to-implement and model-agnostic approach to improve a model's performance under spatial variability, expressed as different scenarios. Specifically, we propose a Multimodal Adaptive Router for Segmenting remote sensing images (MARS) as shown in Figure 3. First, to set up MARS, we summarize a set of different scenarios from data collected from broad geographic regions, where the characteristics or patterns of the target events differ across the scenarios. Second, we perform scenario-based training (e.g., using regular-size models such as U-Net or vision transformers such as SegFormer), which can use different strategies, e.g., first train a global model and then finetune it toward different scenarios, or independently train separate models for different scenarios. Finally, for inference in the testing stage, we use MMFM as a model router, which automatically determines the scenario of an input image and uses the corresponding scenario-based model to generate pixel-level classification results.

TABLE I
Comparison of router model prompts.

Type	Prompt
Simple	You are to classify each image into one of the 3 categories: BigTreeForest, SmallTreeForest, Continuous. Return only the category name, and nothing else.
Descriptive	You are to classify each image into one of the 3 categories: BigTreeForest, SmallTreeForest, Continuous. BigTreeForest refers to forested regions where branch separation is visible...

2) Example in tree mortality event mapping: As a concrete example, we categorize the scenarios into the following three groups as visualized in Fig. 2, where the first two correspond to fine-granularity tree mortality and the last one corresponds to contiguous wipe-outs:

- **BigTreeForest**: forested regions with scattered tree deaths, where trees are larger, so branch separation is visible among dead trees.
- **SmallTreeForest**: forested regions with scattered tree deaths, where trees are smaller and branch separation is not visible among dead trees.
- **Continuous**: forested regions with continuous dead-tree wipe-outs in large patches.

While in each individual scene there are a larger number of dead trees in the continuous scenario, scattered tree deaths are substantially more prevalent, whereas large-patch wipe-outs happen mainly in very specific locations (e.g., due to large fires or deforestation). Thus, mapping scattered tree deaths has been shown to have a substantial contribution to tree loss while simultaneously requiring higher-resolution sensing platforms [8]. Our primary application-side focus is on improving the monitoring of scattered tree deaths.

B. Router Models

The router model determines the most suitable model (i.e., the most suitable scenario) for each input image. We investigate multimodal foundation models that integrate visual and textual signals as a means to perform the routing. As we aim to make the process easy to implement, we consider two widely-used and pretrained MMFMs–Gemini 2.5 Flash Lite and GPT-4o–in both zero-shot and fine-tuned regimes. For zero-shot, Table I illustrates the prompts we use to determine the scenario using MMFM. The prompt contains: (1) The input image for segmentation; (2) A text description as guidance. Specifically, we consider two types of text prompts:

- Simple: A minimal scenario name-only prompt that lists the categories and asks for a single choice;
- Descriptive: A more detailed definition prompt that includes brief class descriptions.

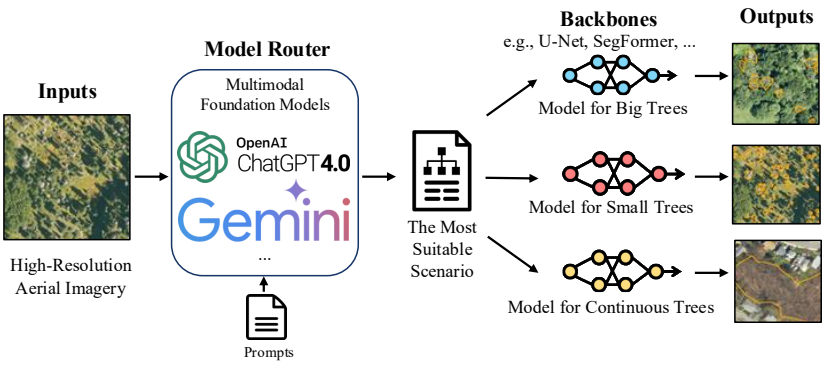

Fig. 3. Example MARS routing and inference pipeline for dead tree segmentation.

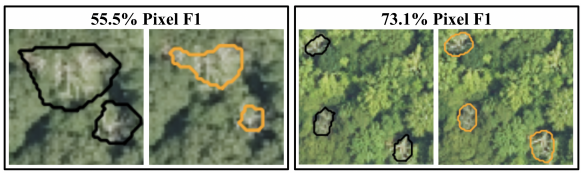

Fig. 4. Pixel metric limitations. **Orange** is predicted, **black** is ground truth.

Additionally, we consider task-specific fine-tuning, aiming for additional gains beyond direct prompting. In this case, we use scenario labels and associated training images to finetune the parameters of pretrained MMFMs.

C. Segmentation Models

Once the MMFM-based router determines the scenario for each image, the trained segmentation model for the scenario is used to produce the segmentation. We consider both standard-size architectures and transformer-based foundation models:

- **U-Net**: Encoder-decoder with skip connections; trained from scratch as a localized, non-pretrained baseline [9].
- **DeepLabV3+**: ResNet-50 backbone with atrous spatial pyramid pooling for multi-scale context; ImageNet-pretrained, with adapted input and singlelogit output [10].
- **SegFormer**: Hierarchical Transformer with a lightweight decoder; initialized from ADE20K, adapted to multi-band inputs via a learnable projection and a single-logit head [11].
- **Mask2Former**: Transformer framework with a Swin-Tiny backbone and masked attention; ADE20Kpretrained, using a class-agnostic mask head for binary output [12].

IV. Experiments

A. Dataset

We use the NAIP imagery at 0.6m spatial resolution for tree mortality event mapping. To make the dataset more representative, we sampled 75 sites over forested regions across the contiguous United States. We manually annotate the dead trees using Google Earth Engine. A tree was labeled dead if it was visually distinct from surrounding trees, displayed well-defined bare branches, and/or showed consistent discoloration (e.g. gray tones). We also consulted auxiliary visualization to assist the labeling (e.g., different color composites such as the false-color composite which is commonly used to highlight vegetation activities).

TABLE II
Comparison of router models.

Model	Prompt	Accuracy (%)
GPT4o	Simple	33.3
GPT4o	Description	33.3
GPT4o$_{FT}$	Simple	33.3
GPT4o$_{FT}$	Description	40.0
Gemini	Simple	66.7
Gemini	Description	73.3
Gemini$_{FT}$	Simple	**100.0**
Gemini$_{FT}$	Description	80.0

The images and labels at the sites are split into equal-size patches of size 224×224 pixels. In total, we extracted about 400 patches with an even split over the 3 scenarios. Each patch comprises RGB, NIR, and NDVI bands as features: $\mathbf{X} \in \mathbb{R}^{N\times 224\times 224\times 5}$. The corresponding binary labels indicate dead-tree pixels: $\mathbf{Y} \in \{0,1\}^{N\times 224\times 224\times 1}$.

The data are split into fixed train/validation/test sets using 70%/10%/20% across all scenarios. The training uses the Adam optimizer with an initial learning rate of 1×10^{-4} and runs for up to 100 epochs with early stopping based on validation loss. The objective is $\mathcal{L} = \mathcal{L}_{\text{BCE}} + \lambda_{\text{Dice}}\mathcal{L}_{\text{Dice}}$ (binary cross-entropy with logits, optional positive-class weight, plus Dice). Losses exclude no-data pixels via masking.

B. Candidate Methods

For each of the four model backbones (Sec. III-C), we consider two frameworks:

- **All scenario**: A single model trained on the entire dataset with all scenarios. We select the best of three runs based on validation performance on all scenarios. These all-

TABLE III
PERFORMANCE ACROSS MODELS AND PIPELINES FOR ALL SCENARIOS (NUMBERS IN PARENTHESES SHOW DIFFERENCES).

Model	Pipeline	Polygon F1 (%)	Polygon Precision (%)	Polygon Recall (%)
BigTreeForest				
U-Net	All-Scenario	18.4	**66.7**	10.7
	Gemini MARS	**57.3 (38.9 ↑)**	44.8 (21.8 ↓)	**79.3 (68.6 ↑)**
Deeplabv3+	All-Scenario	0.0	0.0	0.0
	Gemini MARS	**65.3 (65.3 ↑)**	**66.0 (66.0 ↑)**	**64.6 (64.6 ↑)**
Mask2Former	All-Scenario	51.5	42.5	65.4
	Gemini MARS	**71.4 (19.8 ↑)**	**64.4 (21.9 ↑)**	**80.0 (14.6 ↑)**
SegFormer	All-Scenario	0.0	0.0	0.0
	Gemini MARS	**67.0 (67.0 ↑)**	**72.1 (72.1 ↑)**	**62.6 (62.6 ↑)**
SmallTreeForest				
U-Net	All-Scenario	4.4	9.1	2.7
	Gemini MARS	**69.9 (65.5 ↑)**	**64.3 (55.2 ↑)**	**76.6 (73.9 ↑)**
Deeplabv3+	All-Scenario	0.0	0.0	0.0
	Gemini MARS	**65.2 (65.2 ↑)**	**67.2 (67.2 ↑)**	**63.2 (63.2 ↑)**
Mask2Former	All-Scenario	40.8	36.0	**47.1**
	Gemini MARS	**52.8 (12.0 ↑)**	**56.6 (20.6 ↑)**	49.1 (2.0 ↑)
SegFormer	All-Scenario	0.0	0.0	0.0
	Gemini MARS	**74.2 (74.2 ↑)**	**77.1 (77.1 ↑)**	**71.4 (71.4 ↑)**
Continuous				
U-Net	All-Scenario	**95.8**	**93.9**	**97.9**
	Gemini MARS	93.5 (2.3 ↓)	93.5 (0.3 ↓)	93.5 (4.4 ↓)
Deeplabv3+	All-Scenario	94.4	**97.7**	91.3
	Gemini MARS	**95.8 (1.4 ↑)**	97.1 (0.5 ↓)	**94.4 (3.1 ↑)**
Mask2Former	All-Scenario	**81.2**	69.6	**97.5**
	Gemini MARS	80.5 (0.8 ↓)	**70.2 (0.6 ↑)**	94.3 (3.2 ↓)
SegFormer	All-Scenario	**98.4**	**100.0**	**96.8**
	Gemini MARS	96.3 (2.1 ↓)	**100.0 (0.0 ↔)**	92.9 (3.9 ↓)

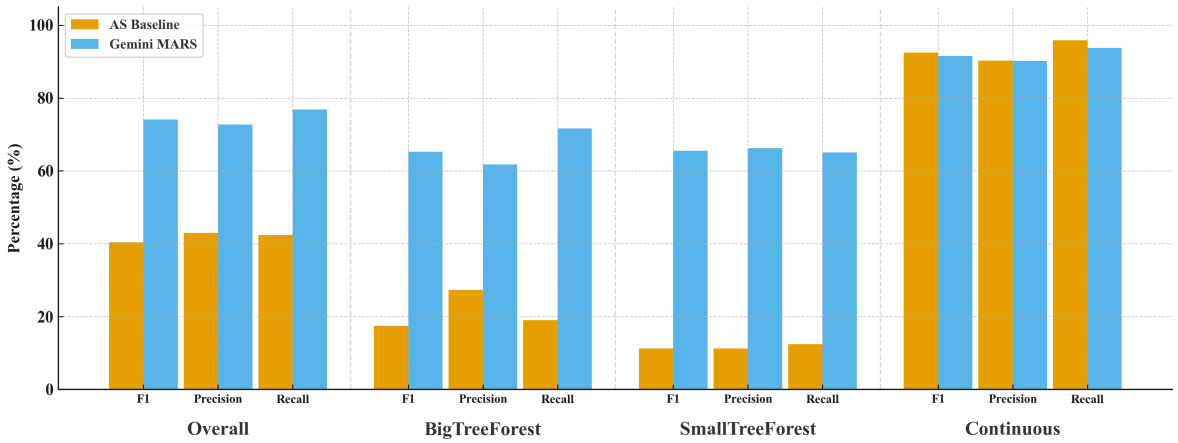

Fig. 5. Comparison of average metrics across backbone segmentation models between All-Scenario (AS) and Gemini MARS.

scenario models (i.e., U-Net, DeepLabV3+, SegFormer, Mask2Former) serve as the baseline models.

- **MARS**: This is the proposed approach where MMFMs are used as routers to identify the scenario and corresponding model for segmentation given an input image.

C. Evaluation metrics

1) Segmentation: We frame tree-mortality mapping as a segmentation rather than object detection challenge. At 0.6m resolution, individual dead trees in dense stands are rarely separable—their crowns and branches merge in continuous mortality patches, making object boundaries ill-posed. In addition, the visual signature of tree mortality is mostly determined by local-neighborhood patterns and should not depend on the total area or shape of a continuous large-patch of dead trees.

On the other hand, while local features should determine and locate dead trees, for evaluation, we found that pixel-level scores are highly sensitive and may not best reflect the success or failure of detection for each tree mortality event. Fig. 4 shows examples where the boundaries in black are the ground truth and those in orange are predicted. Though the mapping is already quite accurate, the direct pixel-level F1

score is still low, which is not ideal to reflect success/failure cases. Additionally, the boundaries of tree crowns often have relatively lower contrast compared to other objects (e.g., thin branches, shadows between branches), making it easier for small boundary misalignments to disproportionately lower pixel-level scores. Thus, to better capture the quality, we convert the masks into polygons and evaluate the scores at the polygon level (i.e., binarization of polygon-level match based on overlaps). This is also helpful as our primary goal is to uncover scattered tree death events, expanding beyond traditional large-patch wipe-out mapping with lower-resolution images. As each individual event is often smaller in area, they are more susceptible to the small misalignments without the polygon-based evaluation metrics. Specifically, we consider precision, recall and F1 scores at the polygon level. A predicted polygon is considered a true positive if it overlaps with any ground-truth polygon; unmatched predictions or ground truths count as false positives or false negatives.

2) Model routing: We also evaluate the performance for the MMFM-based model routing step. The performance is reported as top-1 accuracy—the fraction of images whose predicted scenario exactly matches the ground truth.

D. Results

Table II shows router model performances across different combinations of MMFM and prompting/finetuning strategies. As introduced in Sec. III-B, we considered GPT4o and Gemini with two prompting strategies (simple and descriptive versions) and two finetuning strategies (zero-shot and fine-tuned). Overall, finetuned Gemini with simple prompting received the best performance, so we chose it as the MMFM-based router in the experiments for segmentation. When no finetuning is performed, Gemini with description-expanded prompting received the best accuracy at 73.3%. For both models, finetuning showed improvements potentially because the pretraining data in both MMFMs are more oriented toward general images for vision-related problems instead of remote sensing images.

Table III and Figure 5 compare our method to the all-scenario (AS) baselines. The statistics in Fig. 5 are averaged over all the backbone segmentation models targeting each given scenario; for "Overall," we averaged over all the backbones from all three scenarios to generate the aggregated numbers. MARS performs significantly better in most of the comparisons, especially the BigTreeForest and SmallTreeForest scenarios that are our focus (Sec. III-A2), and roughly the same as AS in the Continuous scenario. For the overall average F1, MARS scores over 30% higher than AS.

V. Conclusion

We presented a new segmentation pipeline (MARS) leveraging multimodal foundation-models as a model router to automatically select the best scenario-based segmentation models (e.g., vision transformers) to address the spatial variability challenge. MARS is easy to implement and does not require re-configuration of backbone models, making it convenient to use in different applications. We also considered different prompting and finetuning strategies and different foundation models for the model router. Additionally, we validated MARS using an important real-world application on tree mortality mapping, which has been identified as a critical knowledge gap in existing large-scale ecological monitoring applications. Using publicly available NAIP imagery at high resolution, the approach showed strong potential in monitoring fine-granularity but widespread tree deaths over large geographic areas. Averaged across all scenarios, MARS significantly improved the F1 scores over the AS models. Our future work will further examine the MARS design with a larger dataset and on other applications that may require textual context alongside visual information.

Acknowledgement

This work is supported in part by the US NSF under Grant No. 2126474, 2147195, 2425844, and 2530610; NASA under grant 80NSSC25K0013 and 80NSSC25K7221; Google's AI for Social Good Impact Scholars program; and the Zaratan cluster at the University of Maryland.

References

[1] Y. Xie, E. He, X. Jia, H. Bao, X. Zhou, R. Ghosh, and P. Ravirathinam, "A statistically-guided deep network transformation and moderation framework for data with spatial heterogeneity," in *2021 IEEE international conference on data mining (ICDM)*. IEEE, 2021, pp. 767–776.

[2] J. Gupta, C. Molnar *et al.*, "Spatial variability aware deep neural networks (svann): A general approach," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 12, no. 6, pp. 1–21, 2021.

[3] A. S. Fotheringham, C. Brunsdon, and M. Charlton, "Geographically weighted regression," *The Sage handbook of spatial analysis*, vol. 1, pp. 243–254, 2009.

[4] Y. Xie, A. N. Nhu, X.-P. Song, X. Jia, S. Skakun, H. Li, and Z. Wang, "Accounting for spatial variability with geo-aware random forest: A case study for us major crop mapping," *Remote Sensing of Environment*, vol. 319, p. 114585, mar 2025. [Online]. Available: https://doi.org/10.1016/j.rse.2024.114585

[5] Y. Xie, Z. Wang, G. Mai, Y. Li, X. Jia, S. Gao, and S. Wang, "Geo-foundation models: Reality, gaps and opportunities," in *Proceedings of the 31st acm international conference on advances in geographic information systems*, 2023, pp. 1–4.

[6] Y. Xie, Z. Wang, W. Chen, Z. Li, X. Jia, Y. Li, R. Wang, K. Chai, R. Li, and S. Skakun, "When are foundation models effective? understanding the suitability for pixel-level classification using multispectral imagery," *arXiv preprint arXiv:2404.11797*, 2024.

[7] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson, E. Mintun, J. Pan, K. V. Alwala, N. Carion, C.-Y. Wu, R. Girshick, P. Dollár, and C. Feichtenhofer, "Sam 2: Segment anything in images and videos," *arXiv preprint arXiv:2408.00714*, 2024. [Online]. Available: https://arxiv.org/abs/2408.00714

[8] Y. Cheng, S. Oehmcke, M. Brandt *et al.*, "Scattered tree death contributes to substantial forest loss in california," *Nature Communications*, vol. 15, p. 641, 2024.

[9] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv. (MICCAI)*, 2015, pp. 234–241.

[10] L. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv:1706.05587*, 2017.

[11] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," *Adv. Neural Inf. Process. Syst.*, vol. 34, pp. 12 077–12 090, 2021.

[12] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention mask transformer for universal image segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 1290–1299.