SAM-based Segmentation of Multi-Class Bridge Components from Diverse Real-Scene Inspection Images

Abstract

Traditional bridge inspection methods rely on manual visual inspection, which is time-consuming, labor-intensive, and potentially dangerous. Automated inspection approaches, which use unmanned aerial vehicles (UAVs) and computer vision, aim to address this issue. However, three knowledge gaps remain. First, although considerable research has been conducted on defect detection and segmentation in bridge inspection, there has been limited focus on segmenting and characterizing specific bridge components that contain defects. Such segmentation provides essential contextual information for understanding the importance of defects for maintenance decision making. Second, existing bridge component recognition approaches face challenges in generalizing across various scenarios, especially in close-range inspections where contextual information is often missing. Third, current developments in the foundation models in the computer vision, such as the segment anything model (SAM), remain unexplored for bridge component segmentation from inspection images due to its lack of domain-specific knowledge and unable to assign semantic labels to multiple segmented components. To address these limitations, this paper proposes a SAM-based image segmentation method for multi-class bridge component segmentation from diverse bridge inspection images. This method leverages the SAM architecture and pre-training from Segment Anything 1 Billion (SA-1B) to enhance feature extraction and improve generalizability. The method also integrates a U-Net decoder to address the challenges of multi-class bridge component segmentation. The proposed method was trained and tested end-to-end on seven classes based on the FHWA's Bridge Inspector's Reference Manual. The results demonstrate promising performance, indicating the potential of this SAM-based approach for efficient and accurate bridge component segmentation.

Keywords
Bridge inspection, computer vision, image segmentation

1 Introduction

Bridges are crucial transportation infrastructures connecting regions, which require extensive inspection to ensure safety and functionality. However, the traditional manual inspection method is time-consuming, labor-intensive, and potentially hazardous. Recent advancements in automated bridge inspection leverage unmanned aerial vehicles (UAVs) and computer vision techniques to enhance the inspection process. These techniques enable efficient collection and analysis of bridge images, allowing for detailed bridge condition assessment. Numerous research efforts have focused on extracting defect information from civil infrastructure inspection images using deep learning-based methods (Spencer et al. 2019; Bai and Sezen 2021, Zhang et al. 2023), including image classification (Cha et al. 2017; Xu et al. 2019; Amirkhani et al. 2024), object detection, and semantic segmentation (Xu et al. 2022). For example, Cha et al. (2018) employed fast region-based convolutional network (Fast R-CNN), a region-based object detection algorithm, to identify concrete cracks, corrosion, and steel delamination. Truong et al. (2023) used UAVs with high-resolution cameras and CNN-based method to identify cracks in concrete bridges. Gao et al. (2023) applied a few-shot learning approach with ProtoNet and transfer learning to detect bridge damage. Song et al. (2024) proposed a lightweight

CNN-based method to segment cracks from bridge images and achieved a high (93.3%) intersection over union (IoU) score.

Despite the significance of these efforts, three primary knowledge gaps remain. First, there is a lack of research on segmenting and characterizing bridge components that may contain defects (Wang and El-Gohary 2024), which is important to link the context of the defects for informed maintenance decision making. Second, there is a lack of pixel-level annotated bridge component segmentation datasets that comply with the FHWA's inspection manual, which limits the development and evaluation of computer vision algorithms for bridge component segmentation (Liu and El-Gohary 2020; Wang and El-Gohary 2024). Furthermore, existing bridge component recognition approaches (Narazaki et al. 2020; Narazaki et al. 2021; Bianchi and Hebdon 2022; Yu and Nishio 2022; Flotzinger et al 2024) are limited in generalizability when applied to diverse bridge types and background scenes, especially for close-range bridge inspection images, where lack some context information (Wang and El-Gohary 2024). These factors introduce challenges and variations for the bridge component segmentation task. Third, while foundational models in computer vision, such as the segment anything model (SAM), represent a significant advancement in general-purpose segmentation tasks, they remain largely unexplored within the bridge inspection domain. Despite their capabilities, large vision foundation models lack the domain-specific training required to accurately interpret bridge component features and cannot assign semantic labels to segmented components, limiting their utility in real-world bridge inspection applications. In addition, multi-class segmentation of bridge components demands that the model differentiate among various visually similar components, each representing a distinct class. Bridge images contain different components with small visual differences, which poses more challenges for SAM to differentiate.

To address these gaps, this paper proposes a segment anything model (SAM)-based image segmentation method to segment bridge components, which leverages SAM and the extensive pre-training of Segment Anything 1 Billion (SA-1B) to improve feature extraction and generalizability, and integrates a U-Net decoder to address the challenges of multi-class bridge component segmentation. The proposed method was trained and tested on seven classes, according to the Federal Highway Administration (FHWA) Bridge Inspector's Reference Manual (Hartle et al. 2002).

2 Literature Review

2.1 Image Segmentation

Unlike other tasks in computer vision (e.g., image classification, object detection), image segmentation poses a greater challenge due to its requirement for pixel-level accuracy, meaning that each pixel must belong to one and only one class. Therefore, annotating a segmentation dataset is demanding, time-consuming, and costly. With advancements in deep learning, several powerful models have been developed to solve segmentation tasks. U-Net (Ronneberger et al. 2015) is a U-shaped encoder-decoder network architecture that comprises four encoder blocks and four decoder blocks connected through a bridge. This model is particularly effective when working with limited amounts of data and delivers precise segmentation results. DeepLabv3+ (Chen et al. 2018) employs an encoder-decoder structure along with atrous convolutions to complex features, enabling it to perform semantic segmentation on images with high resolution. Mask R-CNN (He et al. 2017) utilizes a backbone to extract feature maps and employs region proposal networks to scan these maps for regions that are likely to contain objects. Additionally, it includes a mask head to predict a segmentation mask for each object. Vision transformer (ViT) (Dosovitskiy et al. 2020) introduces a novel approach by applying transformers to image segmentation, where images are divided into patches that are processed sequentially to capture the global context of the image.

2.2 SAM

The SAM model (Kirillov et al. 2023), developed by Meta, marks a significant advancement in the field of computer vision. This state-of-the-art instance segmentation model demonstrates a remarkable capacity for executing complex image segmentation tasks with unmatched accuracy and flexibility. SAM's innovative design enables it to adapt to new image distributions and tasks without any prior training, a feature referred to as zero-shot transfer. It was pre-trained on a large dataset, the Segment Anything 1 Billion (SA-1B), which contains 11M images and 1.1 billion masks. Based on the generalization capabilities of SAM, several studies explored its application in the construction domain. For instance, Ye et al. (2024) introduced two novel SAM-based instance segmentation methods aimed at automating masonry crack detection. Similarly, Ahmadi et al. (2024) applied SAM for crack detection in concrete.

2.3 Transfer Learning

Transfer learning is a widely adopted approach in deep learning, enabling a model trained on a large-scale dataset in a source domain to serve as the foundation for a model developed on a downstream task within a target domain (Iman et al. 2023). This process leverages the broad, generalized data representations learned from extensive datasets, such as ImageNet (Deng et al. 2009), to facilitate effective model adaptation in more specialized domains. It is commonly utilized in deep learning to tackle challenges associated with limited data availability, as deep learning models typically require substantial amounts of data for training, which can be both costly and difficult to obtain. Transfer learning can be classified into two main categories: pre-training and finetuning. Pre-training involves initializing the model with weights from a previously trained, general-purpose model by providing foundational knowledge into a new domain (Hou et al. 2020). Finetuning involves refining the model's parameters with labeled data from the target domain, allowing it to adapt more precisely to the specific features of a task (Wang et al. 2023). This refinement significantly enhances the model's performance for that task compared to a general-purpose pre-trained model.

3 Proposed SAM-based Method for Bridge Component Segmentation

This paper proposes a SAM-based image segmentation method for recognizing and segmenting bridge components from bridge inspection images. The proposed method includes three primary steps: (1) data collection and annotation: the dataset was collected and annotated for developing and evaluating segmentation models that can perform multi-class recognition of seven bridge components; (2) bridge component segmentation: this paper employs a transfer learning strategy using a SAM model integrated with a U-Net decoder, pre-trained on the extensive Segment Anything 1 Billion (SA-1B) dataset; and (3) evaluation.

3.1 Data Collection and Preprocessing

The study utilized image data collected from the Washington Department of Transportation (WSDOT) and the Ohio Department of Transportation (ODOT) through web scraping systems developed by the author. After filtering out low-resolution images and those that do not contain classes of interest, the combined dataset contains 1,000 bridge inspection images. The images feature complex backgrounds and varied environmental contexts, which introduce additional challenges for segmentation. The dataset also includes a range of shot types, from wide-angle views that capture entire bridge spans to close-up images focusing on specific components. This diversity helps in training and evaluating the model's ability to generalize across different bridge structures, backgrounds, and image scales. To meet the input size requirements of the SAM-based segmentation method, all images were resized to a uniform resolution of 1024×1024 pixels.

3.2 Bridge Component Categories and Data Annotation

The FHWA's Bridge Inspector's Reference Manual (Hartle et al. 2002) provides detailed and standardized instructions for inspecting and evaluating bridges. It covers eight types of bridge components, including backgrounds, bearings, abutments, decks, piers/bents, primary superstructure members, and secondary superstructure members. Typical primary members are responsible for carrying primary live loads from vehicles. They include girders, floorbeams, stringers, trusses, spandrel girders, spandrel columns or bents, arch ribs, rib chord bracing, hangers, frame girders, frame legs, frame knees, and pin and hanger links (Hartle et al. 2002). In contrast, secondary members do not typically bear traffic loads directly. They include elements such as diaphragms, cross or X-bracing, lateral bracing, sway-portal bracing, and assemblies (e.g., through bolts, pin caps, nuts, cotter pins on small assemblies, spacer washers, doubler plates) (Hartle et al. 2002). The images utilized in this study were annotated at a per-pixel level using the Labelme tool (Russell et al. 2008). Figure 1 shows examples of these annotated images.

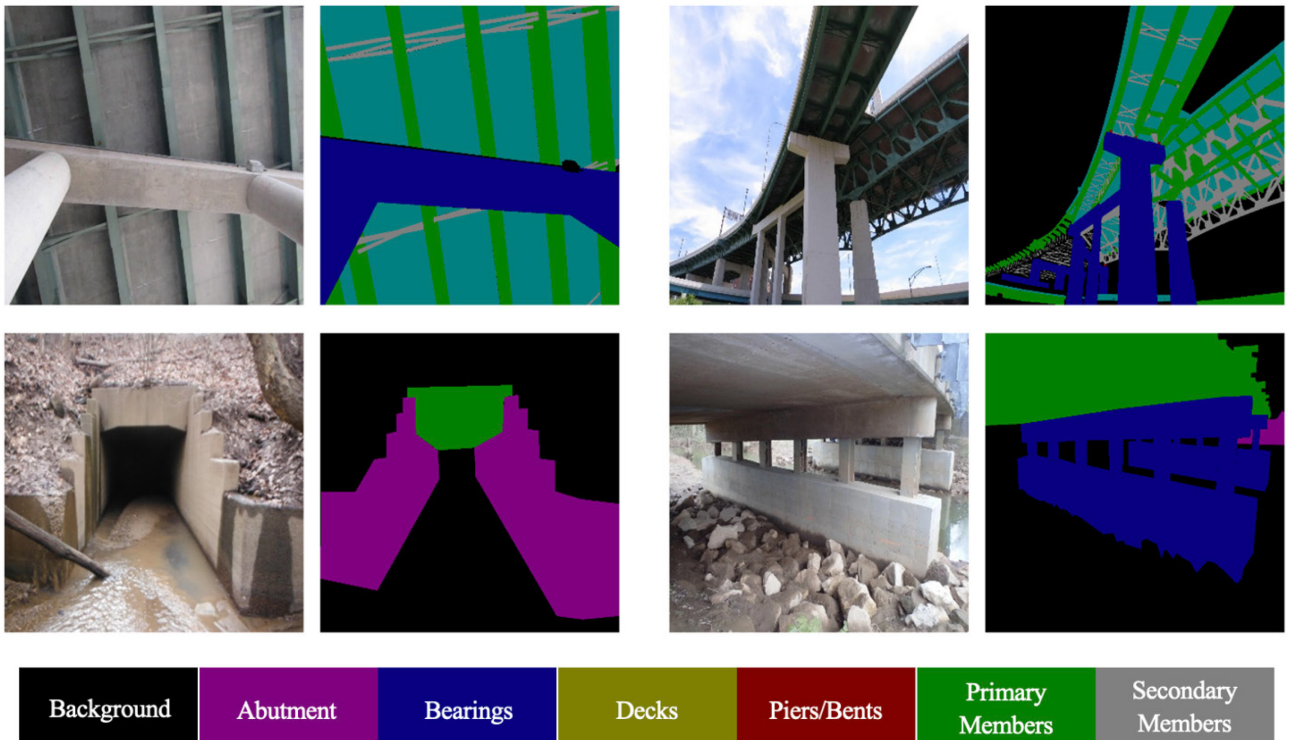

Figure 1. Examples of original images and annotated bridge components.

3.3 Model Architecture

In this study, the SAM was adapted with a custom U-Net decoder to address challenges in bridge component segmentation, where SAM's original prompt-driven architecture falls short. The original SAM architecture comprises three components: an image encoder, a prompt encoder, and a mask decoder. The image encoder, which serves as the core of the model, employs a base ViT model for enhanced scalability. This base ViT model consists of 12 transformer layers, each containing a multi-head self-attention block and a multilayer perceptron (MLP) block that incorporates layer normalization (Ma et al. 2024). The output from this encoder is a feature embedding that is a 16× downscaled representation of the original image. The model takes input with a resolution of 1024×1024×3, and transforms it into a dense embedding of size 64×64×256.

While SAM's prompt encoder and mask decoder are effective for general segmentation, they lack domain-specific understanding of bridge components and are unable to assign semantic labels independently, limiting their effectiveness in inspection tasks that demand class-specific pixel-level prediction. To address this, a U-Net decoder was selected for its hierarchical upsampling structure, for

restoring spatial detail and resolving class boundaries in pixel-level segmentation. The custom U-Net decoder consists of four bilinear upsampling layers progressively increasing resolution. Each upsampling layer is followed by convolution, layer normalization, and ReLU activation to refine features, stabilize training, and introduce non-linearity as SAM's feature embeddings are expanded. Additionally, dropout (with a drop rate of 0.2) is applied to improve generalization, and the final output is a 1024×1024×7 tensor, where each channel represents a distinct bridge component class. This design enhances SAM's capability, making it suitable for multi-class bridge inspection images.

3.4 Training Strategy

Transfer learning was employed, with the model's optimal performance achieved by freezing parameter layers. In the parameter layer freezing approach, the entire encoder layer was frozen, except for the encoder neck, which consists of several convolutional layers and normalization techniques for feature extraction. The weights of the encoder neck and decoder parameters were updated during the training process.

Minimizing focal loss was set as the training objective to handle class imbalance issues encountered in bridge inspection images. Focal loss modulates the standard cross-entropy loss by emphasizing harder-to-classify examples, thereby reducing the relative loss contribution from well-classified examples. This approach helps the model focus on misclassified examples, which is useful in datasets where certain bridge component classes may be underrepresented. The focal loss function applies a scaling factor of $(1 - p)^{\gamma}$ to the cross-entropy loss, where p is the predicted probability for the true class. α is a weighting factor that adjusts the importance of the minority classes, and γ is a focusing parameter that adjusts the rate at which easy examples are down-weighted (Ross and Dollár 2017). In this implementation, α is set to 0.75, to provide additional weight to less frequent classes. γ is set to 5, making the model more sensitive to difficult, low-confidence predictions.

The formula for focal loss used in training is as follows:

$$Focal\ Loss = -\alpha \times (1 - p)^{\gamma} \times \mathrm{y} \times \log(p) \tag{1}$$

where y represents the target label, and p is the predicted probability, smoothed with a small constant (1e-8) to avoid undefined behavior in log operations. This setup ensures robust performance on class-imbalanced datasets, improving the model's ability to segment and classify various bridge components accurately.

3.5 Model Evaluation

The performance of bridge component segmentation was assessed by comparing it with the ground truth labeling masks using four metrics: precision (P), recall (R), F-1 measure, and Intersection-over-Union (IoU). Precision is the proportion of correctly segmented and classified pixels to all predicted pixels. Recall is calculated as the ratio of the number of segmented and classified pixels to the total number of pixels that should be segmented and classified. The harmonic mean of them is known as the F-1 measure, also called the Dice score. To address data imbalance among different classes, performance was assessed using a macro average of accuracy, recall, and F-1 measure in order to prevent the majority classes (those with more occurrences) from misrepresenting the results.

Intersection over Union (IoU), also known as the Jaccard index, is a widely used metric to assess the effectiveness of semantic segmentation. It calculates the ratio of the overlapping pixels between the ground truth and predicted masks to the total number of pixels in both masks. The IoU value ranges from 0 to 1, with 0 indicating no overlap and 1 indicating complete overlap. The mean Intersection over Union (mIoU) is calculated as the macro average of the IoU values for each class.

4 Preliminary Experimental Results and Discussion

The experiments were carried out on the University of Illinois Urbana-Champaign's National Center for Supercomputing Applications (NCSA) Delta GPU through advanced cyberinfrastructure coordination ecosystem: services & support (ACCESS), which consists of a single AMD 64-core 2.45 GHz Milan processor, and one NVIDIA A100 GPU with 40 GB HBM2 RAM (Boerner et al. 2023). The random seed and PyTorch manual seed were both set to 0. The cuDNN.deterministic option was enabled to ensure that cuDNN uses deterministic convolution algorithms while cuDNN.benchmark was disabled to prevent the dynamic selection of cuDNN. The dataset was then divided into training and testing sets using a 9:1 ratio, resulting in 900 training and 100 testing images. Additionally, the Adam optimizer was employed with a learning rate of 0.0001. The model was trained for 25 epochs.

The performance results are summarized in Table 1. The results demonstrate that the proposed SAM-based method performed better on certain categories, such as bearing, primary members, and secondary members. These classes showed higher precision, recall, F-1 score, and IoU, with precision scores over 85%. Specifically, the background class achieved a recall of 94.0%, while bearing achieved a relatively balanced precision and recall, resulting in an F-1 score of 80.9% and an IoU of 68.0%. This indicates that the model accurately captured these components, with sufficient distinction between the classes. However, the model struggled with other classes, such as abutment, deck, and pier/bent, where precision was relatively high, but recall was considerably lower. For example, abutment showed a high precision of 89.6% but a low recall of 33.6%, resulting in an F-1 score of 48.9% and an IoU of 30.9%. Similarly, the deck showed a precision of 75.7% but a recall of only 48.7%, with an F-1 score of 59.3% and an IoU of 40.3%. This suggests that while the model accurately identified instances within these classes, it tended to under-segment and miss some true positives. This issue may be due to the visual similarities (both concrete materials), which are particularly challenging for components like abutments and pier/bent. Overall, the mean F-1 score is 66.2%, indicating a moderate balance between precision and recall across classes. The mean IoU of 49.7% reflects a fair overlap between the predicted masks and ground truth masks, suggesting that while the model was effective for many classes, it could benefit from further adaptation and enhancement to improve recall for under-segmented classes and performance in the segmentation of smaller, more visually similar bridge components.

Table 1. Model performance on bridge component segmentation.

Class	Proposed SAM-based method			
	Precision	**Recall**	**F-1**	**IoU**
Background	77.8%	94.0%	85.1%	75.9%
Abutment	89.6%	33.6%	48.9%	30.9%
Bearing	87.2%	75.4%	80.9%	68.0%
Deck	75.7%	48.7%	59.3%	40.3%
Pier/bent	90.1%	39.7%	55.1%	33.7%
Primary member	86.9%	63.2%	73.2%	55.2%
Secondary member	91.4%	45.5%	60.8%	43.9%
Mean	85.5%	57.2%	66.2%	49.7%

Figure 2 shows visual examples of the segmentation results of the proposed SAM-based method for bridge component segmentation. For the first example, the predicted segmentation mask accurately identifies the primary structure components, with piers/bents and bearings, although missing some true positives, and misclassifies concrete deck as abutment. For the example in the second row, the segmentation mask successfully captures bearing and secondary structure members. The third row shows a failure case, illustrating areas where the segmentation model struggles. The predicted segmentation mask in the center demonstrates a significant misclassification by recognizing the pier

as an abutment and blending into other classes (e.g., primary member) due to visual similarities (both concrete). Besides that, secondary structural members are visible but partially occluded, and thus, some secondary components are incorrectly labeled in the prediction. This failure case suggests that the model encounters challenges when components are visually similar (e.g., same material), thin, or partially occluded, leading to confusion between classes. Overall, this example demonstrates the model's strength in segmenting larger, clearly defined components while highlighting limitations in accurately distinguishing occluded or visually similar bridge components.

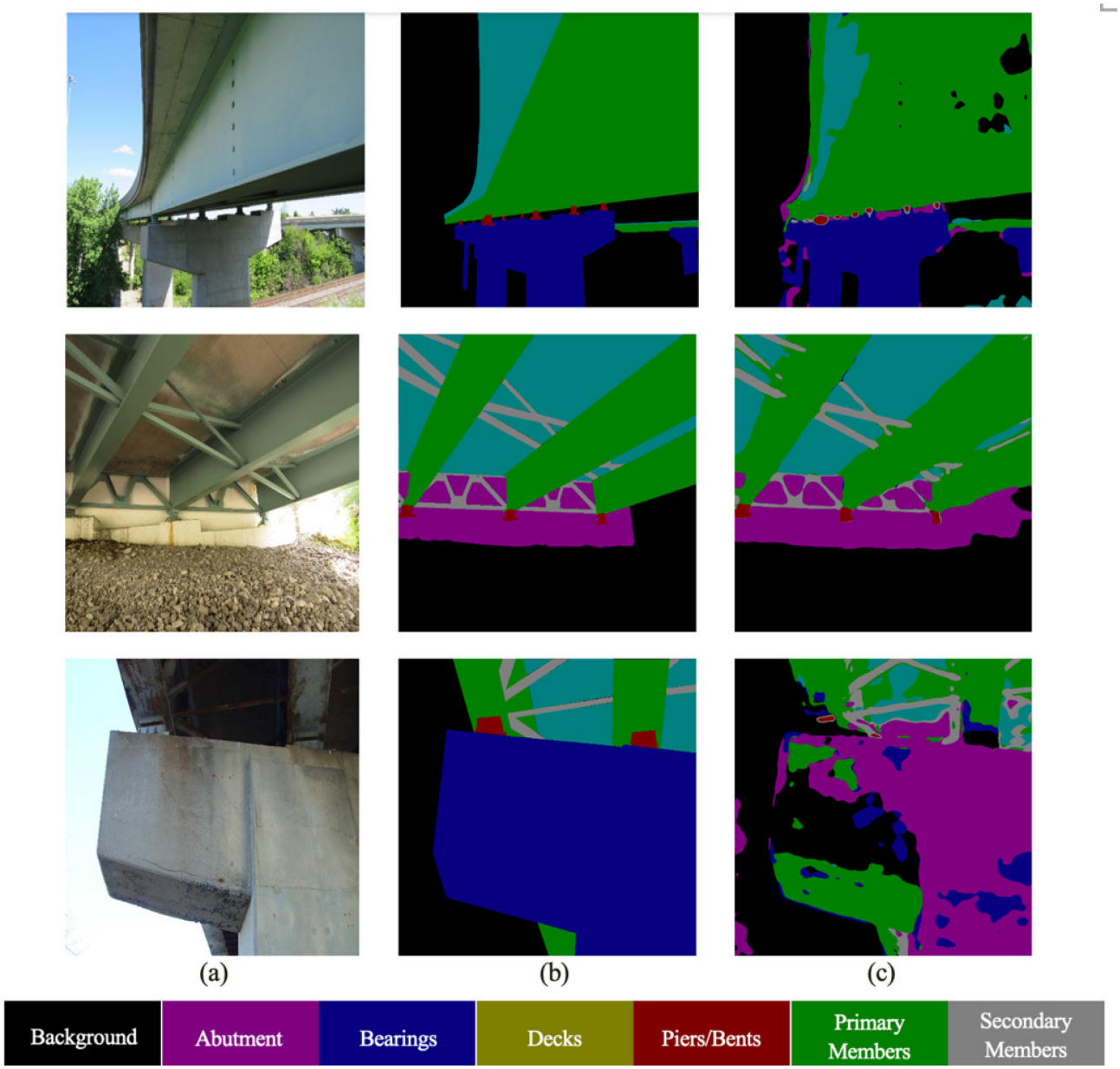

Figure. 2. Visualization of proposed method on some test images: (a) Original Image; (b) Ground Truth Mask; (c) Predicted Mask.

5 Conclusions and Further Work

In this paper, the authors proposed a SAM-based image segmentation method to recognize and segment bridge components, which leverages extensive pre-training of the SAM to improve feature extraction and generalizability. To address the challenges of multi-class bridge component segmentation, a U-Net decoder was integrated, providing detailed spatial reconstruction for component

segmentation. A total of 1,000 bridge inspection images were collected from WSDOT and ODOT and precisely annotated. A transfer learning technique was used to transfer visual knowledge from Segment Anything 1 Billion (SA-1B), a large-scale dataset, to the domain-specific bridge component segmentation problem. The proposed method achieved a mean precision, recall, F-1 measure, and IoU of 85.5%, 57.2%, 66.2%, and 49.7%, respectively, which indicates promising performance in semantic segmentation of bridge components, on average. However, it has some limitations that need to be addressed in future research. First, the dataset used for training and testing the network is relatively small. Second, the model suffered from differentiating certain visual similarity classes (e.g., decks, piers/bents, and abutments). Third, while SAM demonstrated strong performance in general segmentation, it still limits performance and has room to improve in the civil infrastructure domain. Moreover, the SAM-based method failed to recognize very thin objects, such as joints, from the bridge inspection images due to the small occupancy of their pixels, and they often resemble adjacent components (e.g., decks).

To address these limitations, the authors plan to include additional bridge inspection images from diverse sources or regions and employ data augmentation techniques to enhance the model's generalizability. Furthermore, the authors plan to explore more network architectures and techniques to effectively handle imbalanced images, which can be modified and tailored to bridge inspection images, thereby achieving enhanced accuracy and robustness.

6 Acknowledgements

The authors would like to thank the National Science Foundation (NSF). This paper is based on work supported by NSF under Grant No. 2305883. This work used NCSA Delta GPU at the University of Illinois Urbana-Champaign through allocation CIV230015 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, which is supported by National Science Foundation grants #2138259, #2138286, #2138307, #2137603, and #2138296. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of NSF.

7 References

Ahmadi, M., A. G. Lonbar, H. K. Naeini, A. T. Beris, M. Nouri, A. S. Javidi, and A. Sharifi. 2024. Application of Segment Anything Model for Civil Infrastructure Defect Assessment. arXiv.

Amirkhani, D., Allili, M. S., Hebbache, L., Hammouche, N., and Lapointe, J. F. (2024). Visual Concrete Bridge Defect Classification and Detection Using Deep Learning: A Systematic Review. *IEEE Trans. Intell. Transp. Syst.*

Bai, M. and Sezen, H. 2021. Detecting cracks and spalling automatically in extreme events by end-to-end deep learning frameworks. *Proc., ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci., XXIV ISPRS Congress, International Society for Photogrammetry and Remote Sensing.*

Bianchi, E. and Hebdon, M. 2022. Visual Structural Inspection Datasets. *Autom. Constr.*, 139, 104299.

Boerner, T., S. Deems, T. R. Furlani, S. L. Knuth, and J. Towns. (2023). "ACCESS: Advancing Innovation: NSF's Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support." In *Proc., Practice and Experience in Advanced Research Computing (PEARC '23), July 23–27, 2023, Portland, OR, USA. ACM, New York, NY, USA*, 4 pages.

Cha, Y.-J., Choi, W. and Büyüköztürk, O. 2017. Deep Learning-Based Crack Damage Detection Using Convolutional Neural Networks. *Comput.-Aided Civ. Infrastruct. Eng.*, 32(5), 361-378.

Cha, Y.-J., Choi, W., Suh, G., Mahmoudkhani, S. and Büyüköztürk, O. 2018. Autonomous Structural Visual Inspection Using Region-Based Deep Learning for Detecting Multiple Damage Types. *Comput.-Aided Civ. Infrastruct. Eng.*, 33(9), 731-747.

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F. and Adam, H. 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. *Proc., European conference on computer vision (ECCV)*.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K. and Fei-Fei, L. 2009. ImageNet: A large-scale hierarchical image database. *Proc., 2009 IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*

Dosovitskiy, A., L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv preprint arXiv: 2010.11929*.

Flotzinger, J., Rösch, P. J., and Braml, T. 2024. dacl10k: benchmark for semantic bridge damage segmentation. *Proc., IEEE/CVF Winter Conference on Applications of Computer Vision* (pp. 8626-8635).

Gao, Y., Li, H., and Fu, W., 2023. Few-shot learning for image-based bridge damage detection. *Eng. Appl. Artif. Intell.*, 134, 107078.

Hartle, R. A., Ryan, T. W., Mann, E., Danovich, L. J., Sosko, W. B., & Bouscher, J. W. 2002. Bridge Inspector's Reference Manual: Volume 1 and Volume 2. United States Department of Transportation. https://rosap.ntl.bts.gov/view/dot/54492 [Accessed 13Nov.2024]

He, K., G. Gkioxari, P. Dollár, and R. Girshick. 2017. Mask R-CNN. *Proc. IEEE Int. Conf. Comput. Vision.*

Hou, S., B. Dong, H. Wang, and G. Wu. 2020. Inspection of surface defects on stay cables using a robot and transfer learning. *Autom. Constr.*, 119: 103382.

Iman, M., H. R. Arabnia, and K. Rasheed. 2023. A Review of Deep Transfer Learning and Recent Advancements. Technol., 11 (2): 40.

Kirillov, A., E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick. 2023. Segment Anything. Preprint, submitted Apr 5, 2023.

Liu, P. C.-Y. and El-Gohary, N. 2020. "Semantic Image Retrieval and Clustering for Supporting Domain-Specific Bridge Component and Defect Classification." *Proc., Construction Research Congress 2020: Infrastructure Systems and Sustainability*, American Society of Civil Engineers Reston, VA.

Ma, J., Y. He, F. Li, L. Han, C. You, and B. Wang. 2024. Segment anything in medical images. *Nat Commun*, 15 (1): 654. Nature Publishing Group.

Narazaki, Y., Hoskere, V., Hoang, T. A., Fujino, Y., Sakurai, A. and Spencer, B. F. 2020. Vision - Based Automated Bridge Component Recognition with High - Level Scene Consistency. *Comput.-Aided Civ. Infrastruct. Eng.*, 35(5), 465-482.

Narazaki, Y., Hoskere, V., Yoshida, K., Spencer, B. F. and Fujino, Y. 2021. Synthetic Environments for Vision-Based Structural Condition Assessment of Japanese High-Speed Railway Viaducts. *Mech. Syst. Signal Process.*, 160, 107850.

Ohio Department of Transportation. State of Ohio Bridge Photos. https://brphotos.dot.state.oh.us/

Russell, B. C., Torralba, A., Murphy, K. P. and Freeman, W. T. 2008. "LabelMe: a database and web-based tool for image annotation." *Int. J. Comput. Vis., 77(1), 157-173.*

Ronneberger, O., P. Fischer, and T. Brox. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. *Med. Image Comput. Comput.-Assisted Intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, *Springer*.

Ross, T.Y. and Dollár, G.K.H.P., 2017, July. Focal loss for dense object detection. Proc., *IEEE conference on computer vision and pattern recognition* (pp. 2980-2988).

Song, F., Sun, Y. and Yuan, G., 2024. Autonomous identification of bridge concrete cracks using unmanned aircraft images and improved lightweight deep convolutional networks. *Struct. Control Health Monit.*, *2024*(1), p.7857012.

Spencer, B. F., Hoskere, V. and Narazaki, Y. 2019. Advances in Computer Vision-Based Civil Infrastructure Inspection and Monitoring. *Engineering*, 5(2), 199-222.

Truong, C. T., Dang, M. Q., Pham, T. P., Do, P. V., and Tran, H. Q., 2023. A novel automated crack identification method for concrete bridge structure using an unmanned aerial vehicle, In: *IOP Conference Series: Materials Science and Engineering*, August 1, Tran Phu Bridge. IOP Publishing, 1289(1), 012037.

Wang, S., and El-Gohary, N. 2024. Semantic Segmentation of Bridge Components from Various Real Scene Inspection Images. *Proc., 2024 ASCE CI & Construction Research Congress (CRC) Joint Conference*, Des Moines, IA, March 20-23, 2024.

Wang, T., and V. J. L. Gan. 2023. Automated joint 3D reconstruction and visual inspection for buildings using computer vision and transfer learning. *Autom. Constr.*,149: 104810.

Xu, X., Zhao, M., Shi, P., Ren, R., He, X., Wei, X. and Yang, H. 2022. Crack Detection and Comparison Study Based on Faster R-CNN and Mask R-CNN. *Sensors*, 22(3), 1215.

Xu, Y., Bao, Y., Chen, J., Zuo, W. and Li, H. 2019. Surface fatigue crack identification in steel box girder of bridges by a deep fusion convolutional neural network based on consumer-grade camera images. *Struct. Health Monit.*, 18(3), 653-674.

Ye, Z., L. Lovell, A. Faramarzi, and J. Ninić. 2024. Sam-based instance segmentation models for the automation of structural damage detection. *Adv. Eng. Inf.*, 62: 102826.

Yu, W. and Nishio, M. 2022. Multilevel Structural Components Detection and Segmentation toward Computer Vision-Based Bridge Inspection. *Sensors*, 22(9), 3502.

Zhang, C., Karim, M. M., and Qin, R. 2023. A multitask deep learning model for parsing bridge elements and segmenting defect in bridge inspection images. *Transp. Res. Rec.*, *2677*(7), 693-704.