Foundation Model-Based Pipeline for 3D Damage Localization in Built Infrastructure

Zhiya Yang[1], Roberto de Lima-Hernandez[2], Maarten Vergauwen[3]

[1] Faculty of Engineering Technology, KU Leuven - 9000 Gent, Belgium - zhiya.yang@kuleuven.be
[2] Faculty of Engineering Technology, KU Leuven - 9000 Gent, Belgium - roberto.delimahernandez@kuleuven.be
[3] Faculty of Engineering Technology, KU Leuven - 9000 Gent, Belgium - maarten.vergauwen@kuleuven.be

Keywords: Damage Classification, Damage Segmentation, Damage Localization, Feature Extraction, Built Infrastructure.

Abstract

Accurate damage localization is essential for infrastructure inspection, but conventional segmentation methods rely on dense pixel-level annotations that are costly to obtain and difficult to scale. This paper presents a foundation model-based pipeline for data-efficient damage localization in built infrastructure. The proposed workflow combines DINOv3 features for image-level classification, Grad-CAM for weak localization, and the Segment Anything Model (SAM) for prompt-guided pixel-level segmentation. The resulting masks are further transferred into 3D space for spatially contextualized visualization.
The pipeline is evaluated on two case studies. On a subset of Sewer-ML, three representative sewer defect classes are used to compare pretrained backbones and to qualitatively assess downstream localization. The DINOv3-based classifier achieves a higher average F2-score than a Google ViT baseline, reaching about 0.72 versus 0.64. On a custom historic masonry dataset, the method is quantitatively evaluated for material-loss segmentation using manually annotated test masks. The proposed heatmap-guided prompting strategy achieves a mean Dice score of 0.69 and a mean IoU of 0.53, while the classification stage reaches an F2-score of 0.99. A proof-of-concept experiment further demonstrates that segmented damage regions can be visualized within a larger local 3D scene. Overall, the results show that the proposed foundation-model based pipeline can support data-efficient and spatially meaningful damage localization across different infrastructure domains.

1. Introduction

Accurate detection and localization of structural damage are essential for maintaining the safety and durability of built infrastructure. In recent years, artificial intelligence (AI) has been increasingly applied to automate damage detection, with substantial progress in both classification and segmentation. However, standard supervised approaches typically require dense pixel-level annotations, which are costly and time-consuming to produce. This limits their practical use in domains where annotated data are scarce, heterogeneous, or difficult to collect, such as underground infrastructure, heritage sites, and construction environments.

Moreover, many existing approaches operate solely in two dimensions. While 2D predictions may indicate the damage location in an image, they lack the spatial context needed for geometric documentation, comparison across views, or integration into digital representations of assets. For infrastructure inspection workflows, and especially for digital twin applications, it is desirable to localize damage accurately and relate them to three-dimensional geometry. Achieving this in a data-efficient manner remains challenging, particularly when dense segmentation annotations are unavailable. This gap hinders reliable, large-scale condition assessment and delays actionable maintenance decisions.

To address these limitations, this study proposes a data-efficient pipeline that combines pretrained foundation models with 3D reprojection, eliminating the need for segmentation annotations and providing spatially contextualized results for large-scale inspection. The proposed pipeline follows a multi-stage process that combines semantic understanding from classification with geometric precision from segmentation and spatial mapping. First, visual features are extracted from images using the self-supervised DINOv3 model (Siméoni et al., 2025). These features are then used by a lightweight classifier to determine the damage category. Next, Gradient-weighted Class Activation Mapping (Grad-CAM) (Selvaraju et al., 2017) is used to generate class-specific activation maps that localize the regions most relevant to each classification decision. These activation maps are subsequently used as prompts for the Segment Anything Model (SAM) (Kirillov et al., 2023) to obtain refined, pixel-level masks. Finally, the resulting masks can be mapped into 3D space by combining pixel-level segmentation using available geometric information and camera pose estimates, enabling spatially contextualized damage localization. Unlike conventional supervised approaches that rely on dense pixel-level segmentation annotations, the proposed pipeline localizes damage without the need to segment training data by leveraging pretrained foundation models. This approach ensures data efficiency while preserving semantic interpretability and geometric accuracy.

The proposed pipeline is evaluated through a combination of classification, segmentation, and 3D localization experiments. On the Sewer-ML dataset (Haurum and Moeslund, 2021a), the study investigates the effectiveness of pretrained backbones for multi-label damage classification and presents qualitative examples of heatmap-guided segmentation for representative sewer defect classes. To complement this, a historic masonry dataset acquired with a stereo camera is used to quantitatively evaluate the segmentation stage for material loss using manually annotated masks. In addition, a proof-of-concept experiment demonstrates how the segmented damage regions can be reprojected into 3D space for spatial visualization. In this way, the paper aims to show that the proposed pipeline can support data-efficient and spatially meaningful damage localization across different infrastructure domains.

This paper is structured as follows. Section 2 reviews related work on AI-based damage detection, weakly supervised localization, and spatial damage mapping. Section 3 presents the proposed pipeline, including feature extraction, classification, heatmap generation, segmentation, and 3D reprojection. Section 4 describes the datasets, implementation details, and evaluation metrics used in the experiments. Section 5 reports the experimental results on Sewer-ML and historic masonry datasets, including classification performance, segmentation evaluation, and the 3D proof-of-concept demonstration. Finally, Section 6 concludes the article with the main findings, limitations and the direction for future work.

2. Related Work

Computer-vision-based condition assessment has matured from heuristic feature engineering to deep learning pipelines that address classification and segmentation of surface and material defects (e.g., cracks, spalling, corrosion, efflorescence), often to support safety-critical inspection and maintenance decision making. The goal is to automatically and robustly convert inputs (e.g. images or videos) into actionable condition information (Spencer Jr et al., 2019). However, challenges remain, including viewpoint variation, illumination changes, and the difficulty of integrating model outputs into real inspection workflows in a way that supports practical decision making.

The growth of this field has been supported by the emergence of benchmark datasets across different infrastructure domains. In sewer inspection, Sewer-ML has provided a large-scale benchmark for multi-label defect classification and has helped standardize evaluation for image-based condition assessment (Haurum and Moeslund, 2021b). In adjacent areas such as concrete and masonry inspection, numerous crack and surface-defect datasets have enabled progress in damage detection, as summarized in recent reviews (Hamishebahar et al., 2022, König et al., 2022). However, compared with image-level or defect-level labels, pixel-level annotated segmentation datasets remain relatively scarce, especially for infrastructure defects with thin structures, diffuse material loss, or ambiguous boundaries on rough surfaces. Creating such annotations at high quality is labor-intensive and often subjective, which limits the scalability of fully supervised segmentation approaches. Reviews of crack analysis consistently show that segmentation is valuable because it supports measurement and quantification, but it is also sensitive to annotation quality and demands substantial manual labeling effort (Hamishebahar et al., 2022, König et al., 2022). This motivates the search for weaker forms of supervision that can still provide spatially meaningful outputs.

A major line of work in this direction is weakly supervised localization and segmentation. Class Activation Mapping (CAM) was originally introduced to reveal the image regions most responsible for a classifier’s decision (Zhou et al., 2016), and related variants have since become widely used as localization cues in weakly supervised semantic segmentation. Their appeal lies in the fact that they can be obtained from image-level labels alone. However, CAMs are typically spatially coarse, tend to focus on the most discriminative region rather than the full object extent, and therefore require additional refinement for segmentation. (Fan et al., 2020). In addition, classical weakly supervised semantic segmentation usually requires a multi-stage pipeline in which CAMs are refined into pseudo-labels and then used to train a dedicated segmentation network, as in affinity-based refinement (Ahn and Kwak, 2018), cross-image affinity learning (Fan et al., 2020), and self-supervised CAM regularization approaches (Wang et al., 2020). This increases pipeline complexity and still leaves performance strongly dependent on the quality of the initial activation maps.

Recent foundation models offer new opportunities to bridge the gap between weak localization and precise segmentation. Self-supervised visual encoders such as DINOv3 provide semantically rich feature representations that can support robust image-level recognition and localization with limited task-specific supervision (Siméoni et al., 2025). At the same time, the Segment Anything Model (SAM) has shown strong promptable segmentation capability across diverse image settings (Kirillov et al., 2023). The combination of activation maps with SAM is therefore emerging as a promising alternative to classical weakly supervised segmentation pipelines. A particularly relevant precedent is the “From SAM to CAMs” framework, which uses SAM to enhance CAM quality within a weakly supervised segmentation setting rather than relying only on post-processing (Kweon and Yoon, 2024). Similar ideas have also appeared in medical imaging, where CAM-guided SAM frameworks have been used to obtain lesion masks from image-level supervision in breast ultrasound, and more recent work has explored SAM-based weak supervision for medical image segmentation more broadly (Yue et al., 2026, Wang et al., 2025). These studies are closely related to the present work because they show that foundation-model priors can help convert coarse weak localization cues into sharper masks without fully supervised pixel annotations. However, most of them remain focused on 2D segmentation benchmarks and do not address infrastructure-oriented damage mapping in 3D space.

A further line of related work concerns the integration of image-based damage evidence into geometric models and digital twins. In heritage and infrastructure inspection, 3D point clouds and reconstructed surfaces are increasingly used to document deterioration because they provide geometric context for localization, measurement, and temporal comparison. Recent reviews on damage detection in heritage construction based on 3D point clouds highlight both the potential of these representations and the continuing need for more automated analysis pipelines (Sanchez-Aparicio et al., 2023). Likewise, recent digital-twin-oriented studies have shown how damage observations from inspection images can be mapped onto 3D models for large-scale assets (von Benzon and Chen, 2025). Nevertheless, many such workflows assume that reliable 2D detections or masks are already available, leaving open the question of how to obtain them efficiently when segmentation annotation is scarce. The present study is positioned at this intersection: it combines foundation-model-based classification and heatmap-guided segmentation with a 3D localization step, aiming to provide a data-efficient route from image-level labels to spatially contextualized damage mapping.

3. Methodology

This section presents the proposed pipeline for damage localization, which combines pretrained visual feature extraction, lightweight classification, heatmap-guided segmentation, and 3D localization. The workflow (Figure 1) is designed to reduce the dependence on dense pixel-level annotations while still producing spatially meaningful outputs. First, visual features are extracted from each image using DINOv3 (Siméoni et al.,

2025). These features are then used to train a lightweight classifier for damage recognition. Next, Grad-CAM (Selvaraju et al., 2017) is applied to the trained classifier to generate class-specific heatmaps that highlight the image regions most relevant to the prediction. The most activated regions are then converted into box prompts for SAM (Kirillov et al., 2023), which produces refined pixel-level masks. Finally, the segmented regions are reprojected into 3D space using available geometric information and camera pose estimates.

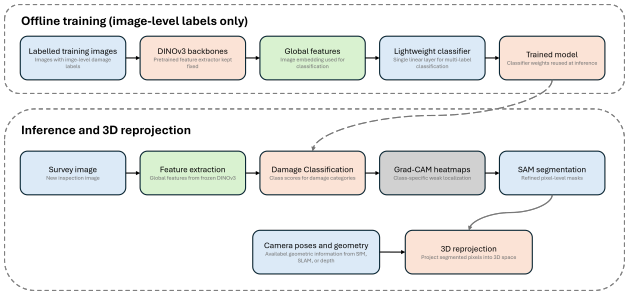


Figure 1. Proposed workflow.

3.1 Feature Extraction

Visual feature extraction is performed using a pretrained foundation model, which provide semantically rich image representations without requiring task-specific pretraining on the target datasets. In this study, the main focus is placed on DINOv3 due to its strong self-supervised representation learning capability. Given an input image, the backbone produces a set of feature embeddings that encode high-level semantic information. For classification, a global image representation is derived from the backbone output and used as input to the downstream classifier. By keeping the feature extractor fixed and training only a lightweight task-specific head, the pipeline remains data-efficient and reduces the risk of overfitting, especially in settings where annotated data are limited.

3.2 Classification

The extracted global feature embeddings are used as input to a lightweight classifier for damage recognition. Specifically, the classifier is implemented as a single linear layer, in which the number of input neurons corresponds to the dimensionality of the extracted feature vector and the number of output neurons corresponds to the target damage categories. In this way, each output neuron represents one damage type, and the classifier learns a direct mapping from the pretrained visual representation to the task-specific label space. Since multiple defects may co-occur within the same image, the classification stage is formulated as a multi-label problem. By training only this linear classification head while keeping the backbone frozen, the model preserves the general semantic knowledge encoded in the pretrained encoder while adapting efficiently to the inspection task. This design is also flexible, as the output dimension can be readily adjusted to accommodate different damage taxonomies across infrastructure domains. The predicted class scores then provide the semantic basis for the subsequent localization process, determining which class-specific heatmaps are generated and passed to the segmentation stage.

3.3 Heatmap Generation

To localize the image regions that contribute most strongly to each classification decision, Gradient-weighted Class Activation Mapping (Grad-CAM) (Selvaraju et al., 2017) is applied to the trained classifier. Grad-CAM generates a class-specific activation map by using the gradients of the target class score with respect to the activations of a selected intermediate layer, thereby highlighting the regions that contribute most to the prediction. These heatmaps provide weak localization cues derived from image-level supervision alone. In the proposed pipeline, they are not treated as final segmentation outputs, but rather as intermediate spatial priors that indicate where the damage is likely to be located. This allows the classification stage to be extended toward spatial localization while preserving interpretability and avoiding the need for dense segmentation supervision.

3.4 Segmentation

The class-specific heatmaps are subsequently used to guide the segmentation process. For each predicted damage class, the most activated regions in the heatmap are identified and converted into prompts for the Segment Anything Model. In the current implementation, bounding boxes derived from high-activation regions are used as prompts, allowing SAM to refine coarse localization cues into pixel-level masks with clearer object boundaries. In this way, the pipeline combines the semantic guidance of the classifier and Grad-CAM with the strong prompt-based segmentation capability of SAM. Manually annotated segmentation masks are not required during training. Instead, segmentation emerges from the interaction between weak localization cues and the pretrained segmentation model.

3.5 3D Reprojection

To extend the 2D segmentation results into a spatially meaningful inspection context, the predicted masks are mapped into 3D space using available geometric information and camera pose estimates. For a segmented image, the pixels belonging to the predicted damage region are associated with their corresponding 3D points. These segmented points can then be placed into a shared reference frame together with surrounding scene geometry, enabling damage visualization within a larger local context. In the present study, this stage is demonstrated as a proof of concept that shows how image-based localization can be transferred into 3D space for spatial interpretation. This 3D step is intended as a general component of the pipeline and can be implemented with different sources of geometric information, depending on the acquisition setup.

4. Experimental Setup

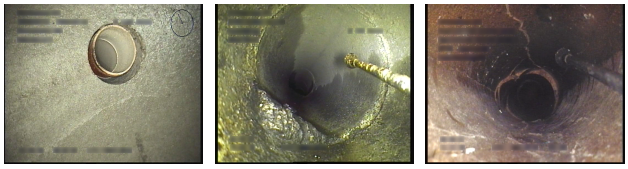
This section describes the datasets, implementation details, and evaluation metrics used to evaluate the proposed pipeline. The experiments were designed to evaluate different stages of the workflow across two infrastructure domains. On sewer inspection imagery, the focus is on multi-label damage classification and qualitative localization and segmentation. On historic masonry imagery, the focus is on segmentation performance for material loss and a proof-of-concept 3D localization experiment.

4.1 Datasets

Two datasets are used in this study. The first is a subset of the Sewer-ML dataset (Haurum and Moeslund, 2021a), which provides annotated sewer inspection imagery for multi-label defect classification. Following the scope of this work, three representative defect classes are selected (Figure 2): connection

(CO), displaced joint (FS), and cracks, breaks, and collapses (RB). For each class, approximately 5,000 images are used for training and 500 images for validation. Since Sewer-ML is formulated as a multi-label problem, the same image may belong to more than one defect class. This subset is used to evaluate the feature extraction and classification stages and to provide qualitative examples of heatmap-guided segmentation. No separate test split was constructed for this subset; therefore, the Sewer-ML case study is used primarily for validation-based comparison of backbone performance and qualitative assessment of the downstream localization pipeline.

The second dataset is a custom historic masonry dataset acquired at the historic wall of Arenberg Castle Park (Leuven, Belgium) using a ZED 2i stereo camera. For this case study, the task is formulated with the classes normal and material loss (Figure 3). The dataset contains 2,000 images per class for training, 250 per class for validation, and 250 per class for testing. To support quantitative evaluation of the segmentation stage, the material-loss images in the test set are manually annotated with pixel-level masks.



(a) CO (b) FS (c) RB

Figure 2. Sewerage examples.

(a) Material loss (b) Normal

Figure 3. Historic masonry examples.

4.2 Implementation Details

The proposed pipeline consists of feature extraction, classification, heatmap generation, segmentation, and 3D localization. Visual features are extracted from input images using pretrained Vision Transformer backbones, while keeping the backbone weights frozen throughout training. In the main experiments, DINOv3 ViT-B/16 (Siméoni et al., 2025) is used as the primary backbone, and Google ViT-B/16 (Dosovitskiy et al., 2021) is included as a comparison in the sewerage case study. Extracted global image embeddings are then used to train a lightweight linear classifier, whose input dimension matches the feature embedding dimension and whose output dimension corresponds to the number of target damage classes.

The classifier training setup differs between the two case studies. For the sewerage case study, the classification task is formulated as a multi-label problem over the three selected defect classes. The linear classification head is trained for 100 epochs using the AdamW optimizer, with a learning rate of 0.001, batch size 256, and binary cross-entropy-based loss. Sigmoid activation is applied to obtain class-wise probabilities, and a threshold of 0.4 is used to determine positive predictions. For the historic masonry case study, the same frozen-feature classification framework is adopted with two target classes, normal and material loss. In this case, the classifier is trained for 10 epochs using the same settings as the sewerage case study.

After classification, Grad-CAM is applied to the trained classifier to generate class-specific activation maps. For the DINOv3 backbone, the target layer corresponds to the norm1 layer of the last transformer block. For each predicted class, the most salient regions in the activation map are extracted and converted into prompts for the SAM. In this study, bounding boxes derived from the dominant activated regions are used as prompts to guide SAM toward refined pixel-level masks. Weak and spatially insignificant responses are removed, and the resulting boxes are slightly enlarged to retain relevant local context before segmentation. The SAM variant employed in this study is vit-h, using the official pretrained checkpoint.

For the historic masonry case study, a proof-of-concept 3D localization experiment is further performed. The segmented image regions are associated with corresponding stereo-derived 3D points, and the resulting segmented points are integrated into a larger local scene using camera pose estimates from nearby frames. This enables the localized material-loss regions to be visualized within their geometric context rather than as isolated 2D detections.

4.3 Evaluation Metrics

Different metrics are used to evaluate different stages of the pipeline. For classification, performance is assessed using the F2-score, which emphasizes recall and is therefore suitable for defect detection scenarios where missed damage is particularly undesirable. In the sewerage case study, classification performance is compared across backbone models using this metric.

For segmentation on the historic masonry dataset, performance is evaluated using Dice score and Intersection over Union (IoU), computed between the predicted masks and the manually annotated ground-truth masks. In addition, F2-score is used to assess the classification performance. The 3D localization stage is presented as a qualitative proof of concept, focusing on the spatial visualization of segmented damage regions within a local 3D scene rather than on a separate quantitative benchmark.

5. Results

5.1 Case Study: Sewerage

The sewerage case study focuses on the classification and weak localization stages of the proposed pipeline. Using the selected Sewer-ML subset, this experiment serves two purposes: first, to compare the effectiveness of different pretrained visual backbones for multi-label defect recognition, and second, to qualitatively assess how class-specific heatmaps can be translated into refined segmentation masks through SAM. Since no separate test split is used for this subset, the analysis is based on validation performance together with qualitative inspection of the downstream localization results.

To assess the influence of the visual backbone on classification performance, two pretrained Vision Transformer models are compared: DINOv3 ViT-B/16 and Google ViT-B/16. In both cases, global image features are extracted from the frozen backbone and used to train a linear classification head. Performance is indicated using the F2-score, which emphasizes recall and is therefore suitable for defect detection scenarios where missed damage is particularly undesirable. The results show that the classifier based on DINOv3 features consistently outperforms the classifier based on Google ViT features over the training

process. After 100 epochs, the DINOv3-based model achieves an average F2-score of approximately 0.72 across the three selected classes, whereas the Google ViT-based model reaches about 0.64 (Figure 4). These results indicate that DINOv3 provides more discriminative and task-relevant feature representations for sewer defect recognition in this setting.

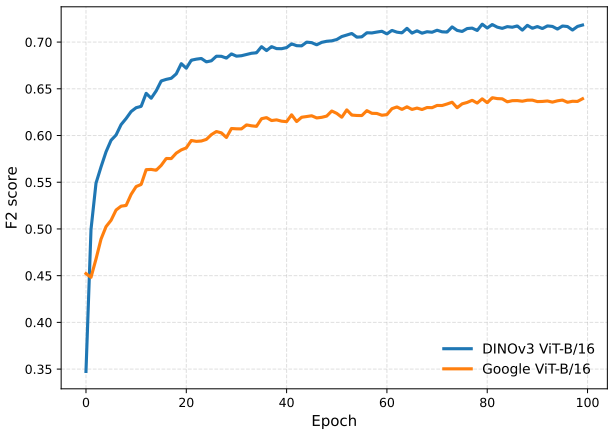
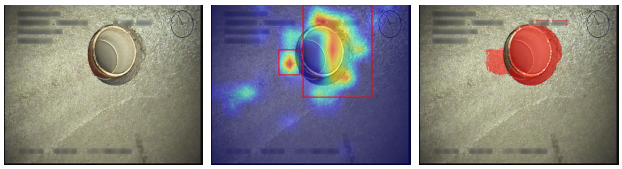


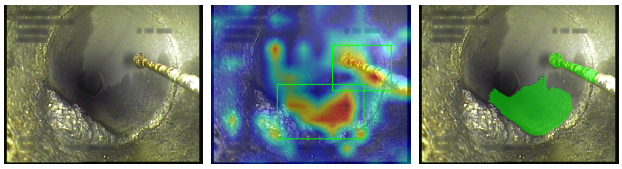
Figure 4. Comparison of backbone performance.

Beyond classification, the sewerage case study also illustrates how the predicted class labels are translated into segmentation outputs. For each predicted defect class, Grad-CAM is used to generate a class-specific activation map highlighting the image regions that contribute most strongly to the classification decision. These activation maps are then converted into bounding-box prompts for SAM, which refines the weak localization cues into pixel-level masks. Qualitative examples for CO, FS, and RB (Figure 5-7) show that the generated heatmaps generally concentrate on visually relevant defect regions and that the SAM outputs produce more spatially precise delineations than the raw activation maps alone.



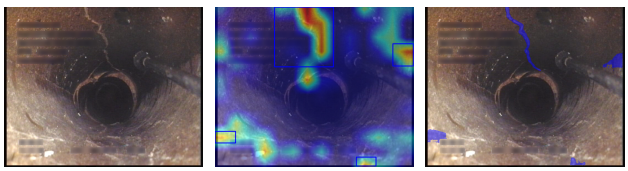
(a) Original (b) Heatmap (c) Segmentation

Figure 5. CO example.



(a) Original (b) Heatmap (c) Segmentation

Figure 6. FS example.



(a) Original (b) Heatmap (c) Segmentation

Figure 7. RB example.

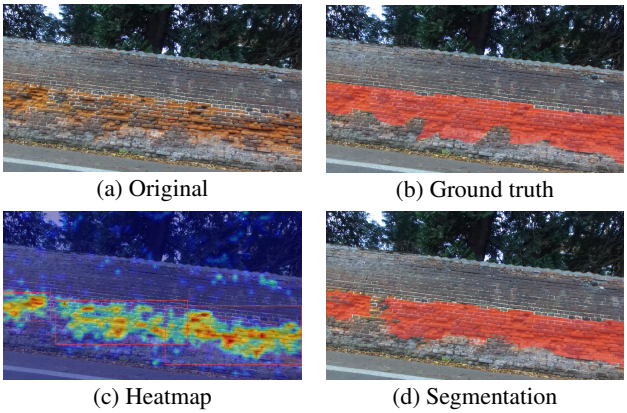
The qualitative results further suggest that the combination of DINOv3, Grad-CAM, and SAM provides a coherent route from image-level recognition to class-aware localization, even without segmentation annotations during training. While this case study does not provide a separate quantitative evaluation of segmentation accuracy, it demonstrates that the proposed pipeline can generate interpretable activation patterns and plausible defect masks for representative sewer defects. These observations support the use of Sewer-ML as a validation-based benchmark for the classification and weak localization stages of the workflow, while the quantitative segmentation assessment is deferred to the historic masonry case study.

5.2 Case Study: Historic Masonry

The historic masonry case study is used to evaluate the later stages of the proposed pipeline, namely segmentation and 3D localization, on a real built-heritage dataset. In contrast to the sewerage case study, which focuses primarily on validation-based backbone comparison and qualitative localization, this case study provides a quantitative assessment of segmentation performance using manually annotated pixel-level masks. The dataset was collected at the historic wall of Arenberg Castle Park using a ZED 2i stereo camera and is formulated with two target classes: normal and material loss.

For classification, the same frozen-feature framework is adopted as in the sewerage case study, using DINOv3 features and a linear classifier. The classifier is trained for 10 epochs and evaluated with the test set using the F2-score. The resulting classification performance reaches an F2-score of 0.99, indicating that the extracted DINOv3 features provide a reliable semantic basis for downstream localization in this binary masonry setting.

The segmentation stage is then evaluated on the material-loss images in the test set, for which pixel-level ground-truth masks are manually annotated. For each predicted material-loss image, Grad-CAM is first used to generate a class-specific activation map, and the dominant activated regions are converted into bounding-box prompts for SAM. Quantitative evaluation on the material loss test images yields a mean Dice score of 0.69 and a mean IoU of 0.53. These results show that the proposed foundation model-based strategy can produce meaningful pixel-level delineation of material-loss regions without requiring segmentation annotations during training.



(a) Original (b) Ground truth

(c) Heatmap (d) Segmentation

Figure 8. Representative historic masonry example.

A qualitative example of the full workflow is shown in Figure 8, including the original image, the manually annotated ground-truth mask, the Grad-CAM heatmap with derived bounding

boxes, and the final SAM segmentation. The comparison illustrates that the activation map successfully highlights the relevant damaged region, while SAM refines the coarse localization cue into a more spatially coherent mask. Although some discrepancies remain near irregular boundaries and low-contrast regions, the overall result demonstrates that the proposed pipeline is capable of recovering the main extent of visible material loss.

To extend the 2D result into a spatial context, a proof-of-concept 3D localization experiment is further conducted. The segmented pixels of the target image are associated with their corresponding stereo-derived 3D points, and these segmented points are then placed into a larger local scene using the poses of nearby frames. Figure 9 presents the resulting 3D visualization, in which the detected material-loss region is highlighted within the surrounding scene geometry. While this experiment is qualitative and limited in scale, it demonstrates the feasibility of transferring image-based damage localization into 3D space without a separate fully supervised segmentation model.

Figure 9. 3D Reprojection.

Overall, the historic masonry case study complements the sewerage experiments by providing quantitative evidence for the segmentation stage and an initial demonstration of 3D spatial integration. Together, these results support the main claim of the paper: that pretrained foundation models, combined with weak localization and prompt-guided segmentation, can provide a data-efficient route from image-level labels to spatially meaningful damage localization.

6. Conclusion

This paper presented a foundation model-based pipeline for data-efficient damage localization in built infrastructure. The proposed workflow combines pretrained DINOv3 features, a lightweight classifier, Grad-CAM heatmaps, and SAM-based prompt-guided segmentation, followed by 3D reprojection for spatial visualization. By relying on image-level supervision during training, the pipeline reduces the dependence on dense pixel-level annotations while still producing interpretable and spatially meaningful outputs.

The experiments across two infrastructure domains demonstrate the potential of this approach. On the Sewer-ML case study, DINOv3-based features achieved stronger multi-label classification performance than the Google ViT baseline and produced meaningful activation maps and plausible defect masks for representative sewer defects. On the historic masonry case study, the proposed heatmap-guided segmentation strategy achieved a mean Dice score of 0.69 and a mean IoU of 0.53 on manually annotated material-loss test images, while the classification stage reached an F2-score of 0.99. In addition, the proof-of-concept 3D experiment showed that the segmented regions can be transferred into a larger local scene and visualized in geometric context.

Overall, the results indicate that pretrained foundation models, combined with weak localization and prompt-guided segmentation, provide a promising route from image-level labels to spatially contextualized damage localization. At the same time, the present study has several limitations. The sewerage case study is based on validation data rather than a separate test set, and the 3D stage is demonstrated qualitatively rather than through a dedicated quantitative benchmark. Future work will therefore focus on broader multi-domain evaluation, more systematic comparison of heatmap generation and prompting strategies, and more rigorous assessment of 3D localization quality in operational inspection settings.

References

Ahn, J., Kwak, S., 2018. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4981–4990.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D. et al., 2021. An image is worth 16x16 words: Transformers for image recognition at scale. *International Conference on Learning Representation*.

Fan, J., Zhang, Z., Tan, T., Song, C., Xiao, J., 2020. Cian: Cross-image affinity net for weakly supervised semantic segmentation. *Proceedings of the AAAI conference on artificial intelligence*, 34number 07, 10762–10769.

Hamishebahar, Y., Guan, H., So, S., Jo, J., 2022. A comprehensive review of deep learning-based crack detection approaches. *Applied Sciences*, 12(3), 1374.

Haurum, J. B., Moeslund, T. B., 2021a. Sewer-ml: A multi-label sewer defect classification dataset and benchmark. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 13456–13467.

Haurum, J. B., Moeslund, T. B., 2021b. Sewer-ml: A multi-label sewer defect classification dataset and benchmark. *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, 13456–13467.

Kirillov, A., Mintun, E., Ravi, N., Mao, H. et al., 2023. Segment anything. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 4015–4026.

König, J., Jenkins, M., Mannion, M., Barrie, P., Morison, G., 2022. What’s cracking? A review and analysis of deep learning methods for structural crack segmentation, detection and quantification. *arXiv preprint arXiv:2202.03714*.

Kweon, H., Yoon, K.-J., 2024. From sam to cams: Exploring segment anything model for weakly supervised semantic segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19499–19509.

Sanchez-Aparicio, L. J., del Blanco-Garcia, F. L., Mencias-Carrizosa, D., Villanueva-Llaurado, P., Aira-Zunzunegui, J. R., Sanz-Arauz, D., Pierdicca, R., Pinilla-Melo, J., Garcia-Gago, J., 2023. Detection of damage in heritage constructions based on 3D point clouds. A systematic review. *Journal of Building Engineering*, 77, 107440.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 618–626.

Siméoni, O., Vo, H. V., Seitzer, M., Baldassarre, F. et al., 2025. DINOv3. arXiv preprint arXiv:2508.10104.

Spencer Jr, B. F., Hoskere, V., Narazaki, Y., 2019. Advances in computer vision-based civil infrastructure inspection and monitoring. *Engineering*, 5(2), 199–222.

von Benzon, H.-H., Chen, X., 2025. Mapping damages from inspection images to 3D digital twins of large-scale structures. *Engineering Reports*, 7(1), e12837.

Wang, H., Huai, L., Li, W., Qi, L., Jiang, X., Shi, Y., 2025. Weakmedsam: Weakly-supervised medical image segmentation via sam with sub-class exploration and prompt affinity mining. *IEEE Transactions on Medical Imaging*.

Wang, Y., Zhang, J., Kan, M., Shan, S., Chen, X., 2020. Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12275–12284.

Yue, X., Zhao, Q., Liu, X., Li, J., Bai, J., Song, C., Liu, S., Moreno, R., Yang, Z., Romero, S. E. et al., 2026. Morphology-enhanced CAM-guided SAM for weakly supervised breast lesion segmentation. *Biomedical Signal Processing and Control*, 116, 109509.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A., 2016. Learning deep features for discriminative localization. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2921–2929.