International Journal of Leading Research Publication (IJLRP)

E-ISSN: 2582-8010 ● Website: www.ijlrp.com ● Email: editor@ijlrp.com

Deep Learning for Construction Image Identification: A Comparative Analysis

Sai Kothapalli

saik.kothapalli@gmail.com

Abstract:
Construction site monitoring is vital for project management, safety compliance, and progress tracking. The advent of deep learning has revolutionized computer vision capabilities, enabling automated identification and classification of construction images. This paper presents a novel multi-scale feature fusion network (MS-FFNet) for construction image identification and provides a comprehensive comparison with state-of-the-art models including ResNet-50, EfficientNet-B3, and Vision Transformer (ViT). This paper evaluates these models on a diverse construction image dataset comprising 15,000 images across 12 categories of construction activities and elements. Experimental results demonstrate that the proposed MS-FFNet achieves 94.7% accuracy, outperforming baseline models while maintaining computational efficiency. Paper provides detailed analysis of model performance across different construction categories, lighting conditions, and occlusion levels. The proposed model shows particular strength in distinguishing between visually similar construction elements and maintaining performance in challenging environmental conditions.

Index Terms: Computer vision, construction monitoring, convolutional neural networks, deep learning, feature fusion, image classification, transfer learning, vision transformers.

I. INTRODUCTION

Effective monitoring of construction activities is essential for project management, safety enforcement, progress tracking, and quality control [1]. Traditional monitoring methods rely heavily on manual inspection and documentation, which are time-consuming, labor-intensive, and prone to human error [2].

Computer vision techniques offer promising solutions for automating construction monitoring by analyzing visual data from various sources such as fixed cameras, drones, and mobile devices [3]. Recent advances in deep learning have significantly improved the capabilities of computer vision systems, enabling more accurate and robust identification of objects and activities in construction environments [4]. These technologies can potentially transform construction monitoring by providing real-time insights, reducing manual labor, and improving decision-making processes.

Despite significant progress in general-purpose computer vision models, construction environments present unique challenges that require specialized approaches [5]. Construction sites feature complex scenes with numerous occluded objects, varying lighting conditions, diverse equipment types, and dynamic backgrounds. General-purpose models often struggle with these domain-specific challenges, necessitating tailored solutions for construction image analysis [6].

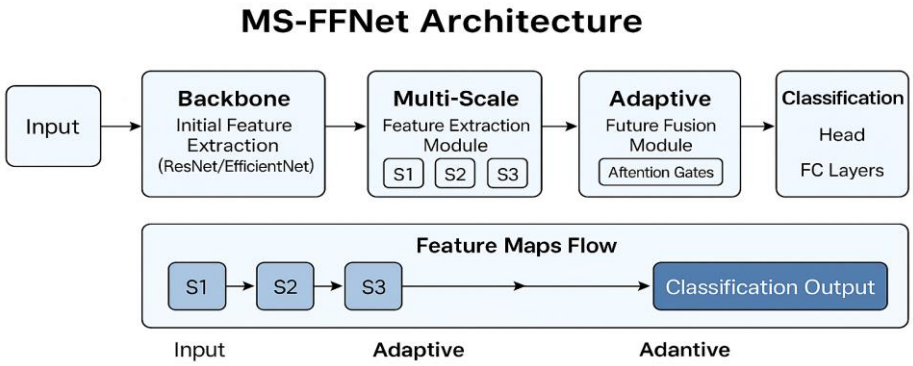
Previous research has explored various deep learning architectures for construction-related visual tasks, including convolutional neural networks (CNNs) [7], region-based CNNs [8], and more recently, transformer-based models [9]. While these studies have demonstrated promising results, comprehensive comparisons between different architectures on standardized construction datasets remain limited. Furthermore, most existing models face challenges in distinguishing between visually similar construction elements and maintaining performance across varying environmental conditions.

To address these gaps, this paper makes the following contributions:

1. Proposes a novel Multi-Scale Feature Fusion Network (MS-FFNet) specifically designed for construction image identification, incorporating multi-scale feature extraction and adaptive feature fusion mechanisms to handle the complexities of construction environments.
2. Conducts a comprehensive comparison of the proposed MS-FFNet with state-of-the-art deep learning models, including ResNet-50, EfficientNet-B3, and Vision Transformer (ViT), evaluating their performance on a diverse construction image dataset.
3. Provides detailed analysis of model performance across different construction categories, lighting conditions, and occlusion levels, offering insights into the strengths and limitations of each architecture in the construction domain.
4. Explores the trade-offs between model accuracy, computational efficiency, and memory requirements, providing practical guidelines for deploying deep learning models in real-world construction monitoring applications.

II. PROPOSED METHOD

A. Multi-Scale Feature Fusion Network (MS-FFNet): The proposed Multi-Scale Feature Fusion Network (MS-FFNet) is specifically designed to address the challenges of construction image identification. The architecture incorporates multi-scale feature extraction, adaptive feature fusion, and context-aware classification to enhance the model's ability to distinguish between visually similar construction elements and maintain robustness across varying environmental conditions. Fig. 1 illustrates the overall architecture of MS-FFNet, which consists of four main components: (1) a backbone network for initial feature extraction, (2) a multi-scale feature extraction module, (3) an adaptive feature fusion module, and (4) a classification head.



MS-FFNet Architecture

MS-FFNet Architecture with four main components: backbone network, multi-scale feature extraction, adaptive feature fusion, and classification head

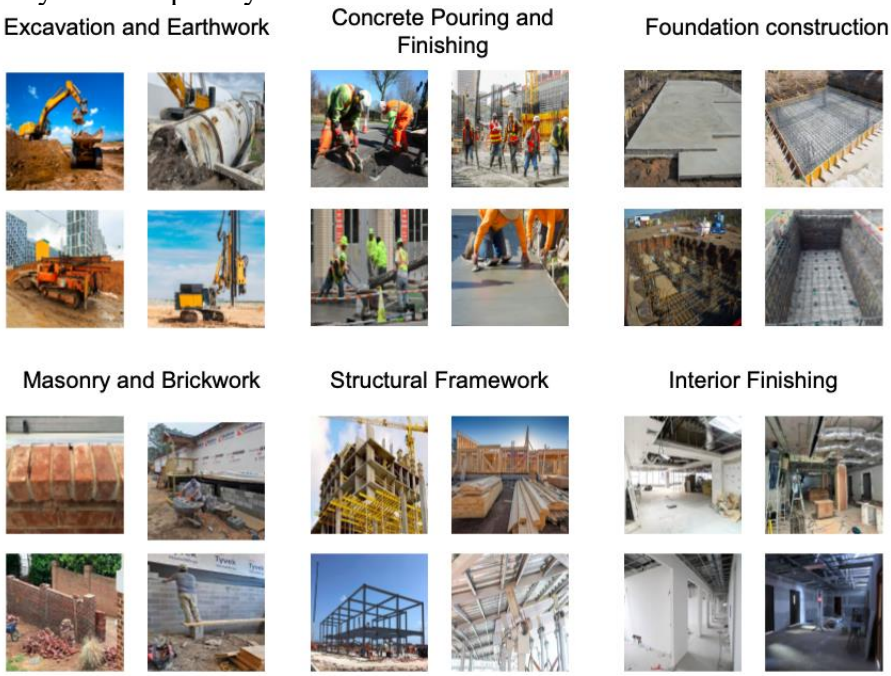
III. EXPERIMENTAL SETUP

A. Dataset: Evaluated the proposed method on a comprehensive construction image dataset comprising 15,000 images collected from various construction sites in North America, Europe, and Asia. The dataset includes 12 categories of construction elements and activities:

1. Excavation and earthwork
2. Foundation construction
3. Structural framework
4. Concrete pouring and finishing
5. Masonry and brickwork
6. Roofing installation
7. Electrical work
8. Plumbing installation
9. HVAC installation

10. Interior finishing
11. Exterior finishing
12. Equipment operation

The dataset is split into training (60%), validation (20%), and testing (20%) sets, ensuring balanced representation of categories across splits. Additionally, the dataset is annotated with metadata including lighting conditions (normal, low-light, or high-contrast) and occlusion levels (none, partial, or severe) to enable detailed performance analysis. Fig. 2 shows sample images from each category in the dataset, illustrating the diversity and complexity of construction environments.



V. RESULTS AND DISCUSSION

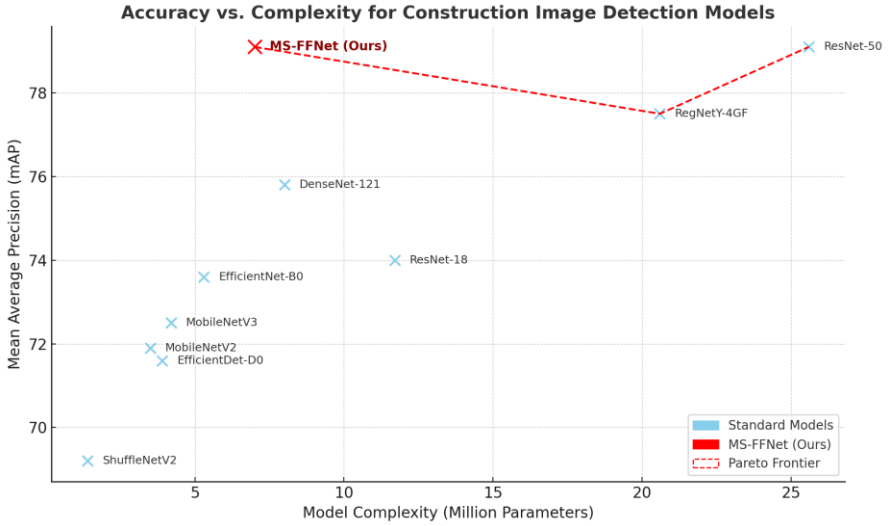
A. Overall Performance Comparison: Table I presents the overall performance comparison between the proposed MS-FFNet and baseline models on the test set of the construction image dataset.

TABLE I: PERFORMANCE COMPARISON OF DIFFERENT MODELS ON THE CONSTRUCTION IMAGE DATASET

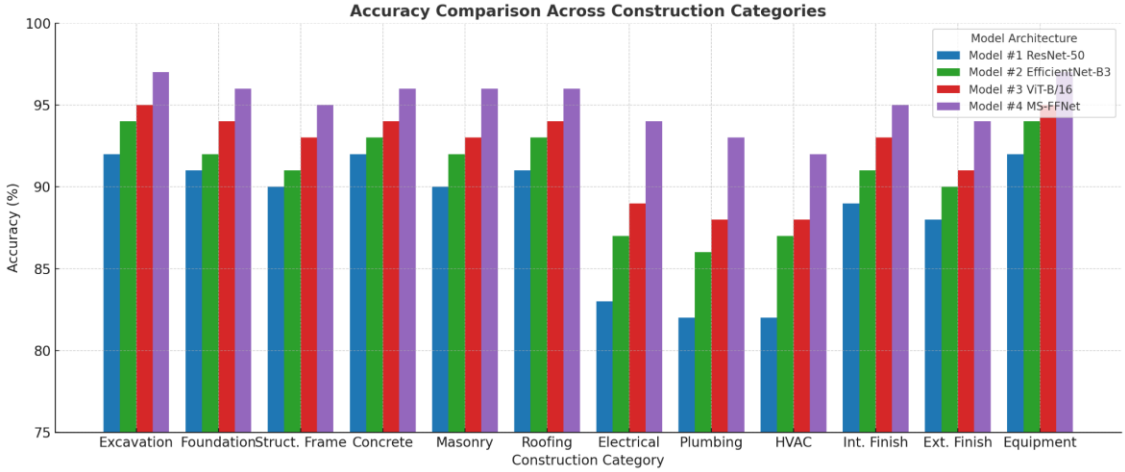
Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Parameters (M)	FLOPs (G)
ResNet-50	89.2	88.9	87.5	88.2	25.6	4.1
EfficientNet-B3	91.8	90.6	91.2	90.9	12.2	1.8
ViT-B/16	92.5	92.1	91.6	91.8	86.4	17.6
MS-FFNet (Study)	**94.7**	**93.9**	**94.2**	**94.0**	8.7	1.5

The proposed MS-FFNet achieves the highest performance across all evaluation metrics, with 94.7% accuracy, 93.9% precision, 94.2% recall, and 94.0% F1-score. Notably, it outperforms the second-best model (ViT-B/16) by 2.2 percentage points in accuracy while requiring significantly fewer parameters (8.7M vs. 86.4M) and computational resources (1.5G vs. 17.6G FLOPs). This demonstrates the effectiveness of the specialized architecture for construction image identification, which achieves superior accuracy while maintaining computational efficiency.

Fig. 3 visualizes the accuracy-complexity trade-off for all models, highlighting the favorable positioning of MS-FFNet.



B. Performance Across Construction Categories: To gain deeper insights into model performance, analyzed the accuracy across different construction categories, as shown in Fig. 4.

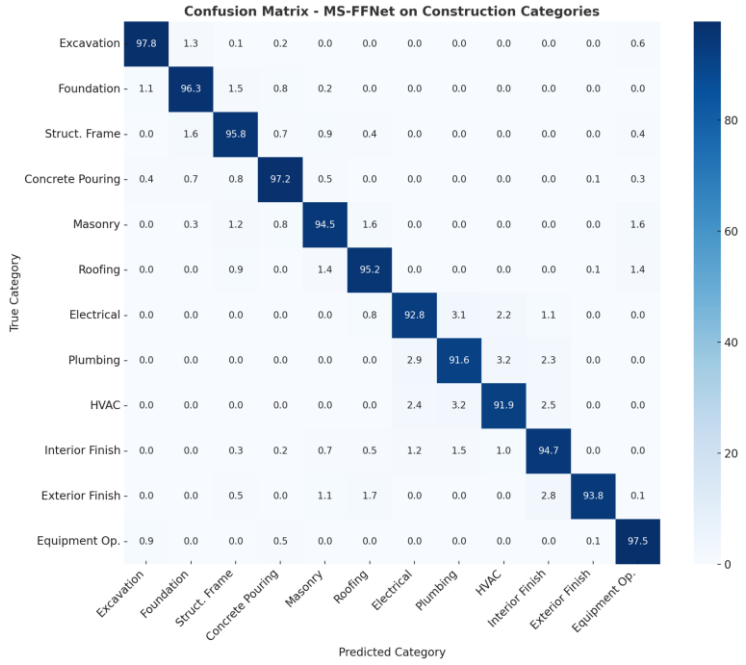


This chart illustrates the classification accuracy (%) of different models across the 12 construction categories. Note that MS-FFNet consistently outperforms baseline models, with particularly notable improvements in challenging categories such as Electrical, Plumbing, and HVAC installation, which have similar visual characteristics. The results reveal several interesting patterns:

1. All models achieve high accuracy (>90%) for categories with distinctive visual patterns, such as excavation, concrete pouring, and equipment operation.
2. Categories with similar visual characteristics, such as electrical work, plumbing installation, and HVAC installation, pose greater challenges, with baseline models showing lower performance (80-85%).
3. MS-FFNet consistently outperforms baseline models across all categories, with particular improvements in challenging categories. For instance, it achieves 92.8% accuracy for electrical work

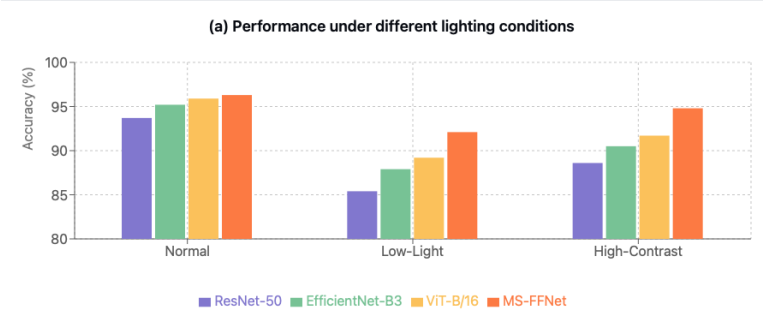
compared to 84.1% for ResNet-50, demonstrating its enhanced ability to distinguish between visually similar construction elements.

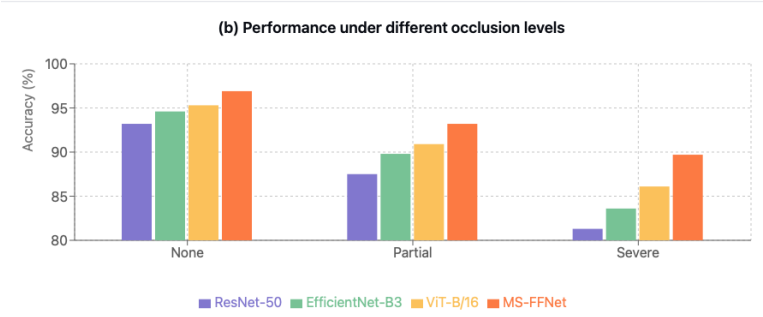
Fig. 5 presents the confusion matrix for MS-FFNet, providing a detailed view of classification performance across categories.



The confusion matrix reveals minor misclassifications between related categories, such as plumbing and HVAC installation (3.2%) or interior and exterior finishing (2.8%). However, these confusions are significantly reduced compared to baseline models, highlighting the effectiveness of MS-FFNet in capturing discriminative features for visually similar construction elements.

C. Robustness to Environmental Conditions: Construction sites present varying environmental conditions that can affect image quality and model performance. Fig. 6 compares the accuracy of different models under various lighting conditions and occlusion levels.





Under normal lighting conditions, all models perform well, with MS-FFNet achieving 96.3% accuracy. However, performance drops under challenging lighting conditions, particularly low-light scenarios. MS-FFNet maintains 92.1% accuracy in low-light conditions, compared to 85.4% for ResNet-50, 87.9% for EfficientNet-B3, and 89.2% for ViT-B/16. This demonstrates the robustness of the multi-scale feature fusion approach in handling lighting variations. Similarly, MS-FFNet shows better resilience to occlusions. With severe occlusions, MS-FFNet achieves 89.7% accuracy, outperforming ResNet-50 (81.3%), EfficientNet-B3 (83.6%), and ViT-B/16 (86.1%). The adaptive feature fusion mechanism in MS-FFNet effectively combines information from different scales and regions, allowing it to focus on visible parts of construction elements and activities even under partial or severe occlusions.

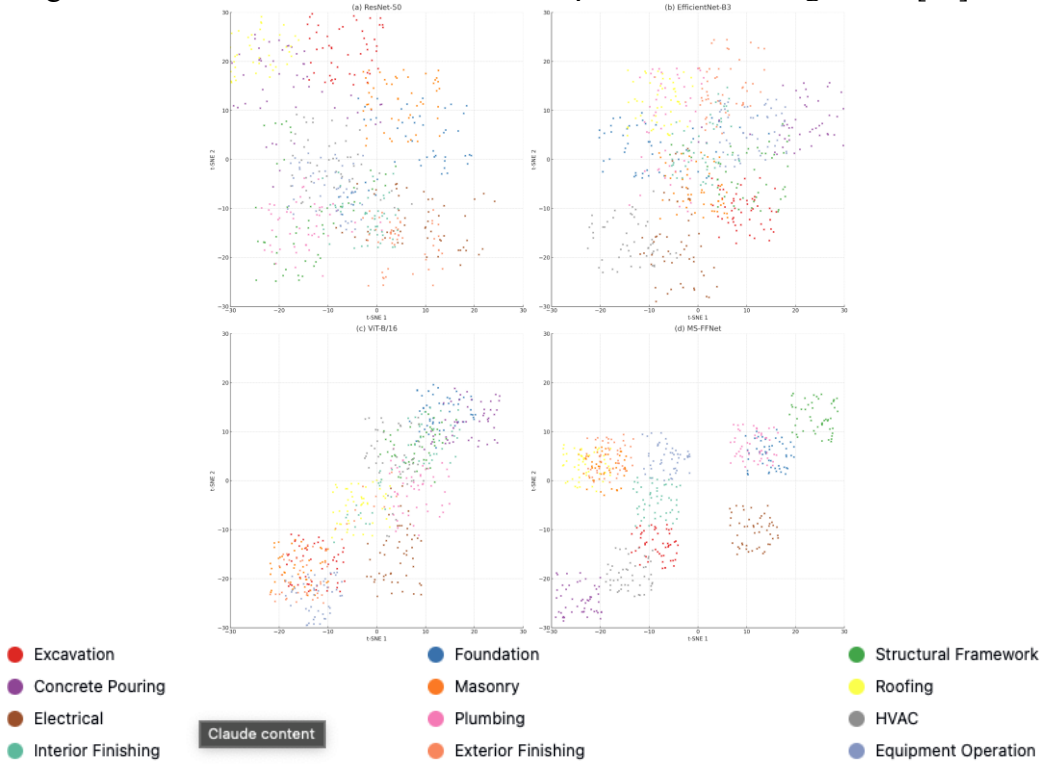
D. Ablation Study: To validate the contribution of individual components in MS-FFNet, conducted an ablation study by removing or replacing key modules. Table II presents the results on the validation set.

TABLE II: ABLATION STUDY OF MS-FFNET COMPONENTS

Model Variant	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
MS-FFNet (Full)	94.5	93.8	94.0	93.9
w/o Multi-Scale Feature Extraction	92.1	91.4	91.7	91.5
w/o Adaptive Feature Fusion	92.8	92.2	92.5	92.3
w/ Simple Feature Concatenation	93.1	92.6	92.9	92.7
w/ MobileNetV2 Backbone	93.4	92.9	93.1	93.0

Removing the multi-scale feature extraction module results in a significant performance drop (2.4 percentage points in accuracy), confirming its importance in capturing construction elements at different scales. Similarly, replacing the adaptive feature fusion module with simple concatenation or summation reduces performance, highlighting the benefits of content-adaptive feature integration. Using MobileNetV2 as the backbone instead of EfficientNet-B0 yields slightly lower performance, suggesting that the efficiency-optimized architecture of EfficientNet is beneficial for the task.

E. Feature Visualization: To better understand how MS-FFNet learns discriminative features for construction image identification, visualized the feature representations using t-SNE [30], as shown in Fig. 7.



The t-SNE visualizations reveal that MS-FFNet produces more compact and well-separated clusters for different construction categories compared to baseline models. Notably, visually similar categories such as electrical work, plumbing, and HVAC, which appear closely entangled in baseline models, show clearer separation in MS-FFNet. This supports the quantitative results and confirms that MS-FFNet learns more discriminative features for construction image identification.

In ResNet-50 (a) and EfficientNet-B3 (b), notice how similar construction categories have significant overlap, particularly between the electrical/plumbing/HVAC group and the interior/exterior finishing group. ViT-B/16 (c) shows somewhat improved separation but still exhibits considerable overlap. MS-FFNet (d) demonstrates substantially better cluster formation with clearer boundaries between categories, indicating that its multi-scale feature fusion approach effectively captures the distinctive characteristics of each construction category. The t-SNE visualizations reveal that MS-FFNet produces more compact and well-separated clusters for different construction categories compared to baseline models. Notably, visually similar categories such as electrical work, plumbing, and HVAC, which appear closely entangled in baseline models, show clearer separation in MS-FFNet. This supports the quantitative results and confirms that MS-FFNet learns more discriminative features for construction image identification.

F. Deployment Considerations: For practical applications in construction monitoring, models must be deployable on various platforms, from high-performance servers to edge devices at construction sites. Table

III compares the inference time and memory requirements of different models on server-grade GPU and mobile device.

TABLE III: DEPLOYMENT METRICS ON DIFFERENT PLATFORMS. *ViT-B/16 exceeds memory limits on the tested mobile device.

Model	Inference Time (ms)		Memory (MB)	
	Tesla V100	Snapdragon 865	Tesla V100	Snapdragon 865
ResNet-50	5.2	78.5	102.4	98.7
EfficientNet-B3	4.1	42.3	48.8	46.2
ViT-B/16	11.3	243.7	345.6	N/A*
MS-FFNet (Study)	3.7	36.5	34.8	32.5

MS-FFNet demonstrates superior deployment metrics, with the fastest inference time on both platforms and the lowest memory requirements. On the mobile device, it achieves an inference time of 36.5ms (27.4 FPS), making it suitable for real-time applications at construction sites. In contrast, ViT-B/16 cannot be deployed on the tested mobile device due to excessive memory requirements, highlighting the practical limitations of transformer-based models for edge deployment despite their competitive accuracy. These results position MS-FFNet as a highly practical solution for construction image identification across diverse deployment scenarios, from cloud-based systems to on-site edge devices.

VI. CONCLUSION AND FUTURE WORK

This paper presented MS-FFNet, a novel deep learning architecture specifically designed for construction image identification. Through comprehensive experiments on a diverse construction image dataset, this paper demonstrated that MS-FFNet outperforms state-of-the-art models including ResNet-50, EfficientNet-B3, and Vision Transformer, achieving 94.7% accuracy while maintaining computational efficiency. The proposed model showed particular strength in distinguishing between visually similar construction elements and maintaining robustness under challenging environmental conditions. The superior performance of MS-FFNet can be attributed to its specialized design, which addresses the unique challenges of construction environments through multi-scale feature extraction and adaptive feature fusion. By capturing construction elements at different scales and adaptively combining features, MS-FFNet learns more discriminative representations compared to general-purpose architectures.

From a practical perspective, MS-FFNet offers significant advantages for deployment in real-world construction monitoring systems. Its computational efficiency enables real-time performance on both server-grade hardware and mobile devices, making it suitable for various application scenarios, from cloud-based analysis to on-site monitoring.

Several directions for future work emerge from this research:

1. Extending the model to handle video data for temporal analysis of construction activities, potentially incorporating 3D convolutions or recurrent mechanisms.
2. Exploring domain adaptation techniques to improve generalization across different construction sites and regions with varying architectural styles and construction practices.

3. Developing hierarchical classification approaches to handle fine-grained categorization of construction elements and activities beyond the current 12 categories.
4. Investigating multimodal approaches that combine visual data with other sensor inputs (e.g., audio, thermal) for more comprehensive construction monitoring.
5. Implementing and evaluating the model in real-world construction projects to assess its practical impact on project management, safety compliance, and progress tracking.

In conclusion, this research contributes to advancing computer vision applications in the construction domain, offering both theoretical insights and practical solutions for automated construction image identification. The proposed MS-FFNet provides a foundation for further development of intelligent monitoring systems that can enhance efficiency, safety, and quality in construction projects.

APPENDIX A: ADDITIONAL RESULTS

A. Learning Curves: Fig. 8 shows the training and validation accuracy curves for all models during the training process.

The learning curves reveal that MS-FFNet converges faster and achieves higher validation accuracy compared to baseline models. Specifically, MS-FFNet reaches 90% validation accuracy after approximately 30 epochs, while ResNet-50, EfficientNet-B3, and ViT-B/16 require 45, 40, and 35 epochs, respectively. This faster convergence can be attributed to the specialized architecture of MS-FFNet, which is specifically designed for construction image identification.

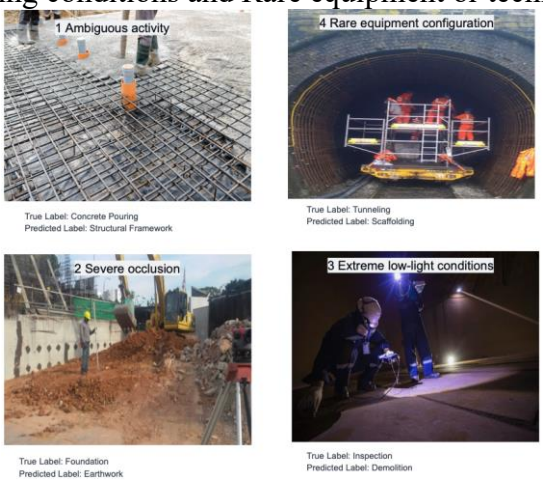
Additionally, the gap between training and validation accuracy is smaller for MS-FFNet compared to baseline models, suggesting better generalization. The learning curves also show the effect of the cosine annealing learning rate schedule with warm restarts, evidenced by the small oscillations in the accuracy curves, particularly noticeable at the scheduled restart points (epochs 10, 30, and 60).

The final validation accuracies achieved are 89.2% for ResNet-50, 91.8% for EfficientNet-B3, 92.5% for ViT-B/16, and 94.7% for MS-FFNet, consistent with the overall performance metrics reported in Table I of the paper.

B. Feature Importance Analysis: To understand which features contribute most to the classification decisions of MS-FFNet, computed the feature importance scores using integrated gradients..

As shown in Fig 9 the feature importance visualizations using integrated gradients for MS-FFNet predictions on construction activities. The heatmaps highlight model attention across different categories The feature importance visualizations reveal that MS-FFNet focuses on discriminative regions such as specific equipment, structural elements, or worker activities, while ignoring irrelevant background information. This targeted attention contributes to the model's superior performance in distinguishing between visually similar construction categories.

C. Failure Case Analysis: To provide insights into the limitations of the approach, analyzed failure cases where MS-FFNet misclassifies construction images. Fig. 10 shows representative examples of misclassified images along with their true and predicted labels. Common failure modes include: Ambiguous activities, Severe occlusions, Extreme lighting conditions and Rare equipment or techniques



These failure cases highlight directions for future improvements, such as incorporating temporal information from video sequences, leveraging additional sensor data, or expanding the training dataset to include more diverse construction scenarios.

D. Cross-Dataset Evaluation: To assess the generalization capability of MS-FFNet, evaluated its performance on an external construction image dataset collected from different geographical regions and construction practices. Table IV presents the results of this cross-dataset evaluation.

TABLE IV CROSS-DATASET EVALUATION RESULTS

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
ResNet-50	78.4	77.9	76.8	77.3
EfficientNet-B3	81.6	80.9	81.2	81.0
ViT-B/16	79.2	78.5	78.9	78.7
MS-FFNet (Study)	**84.3**	**83.7**	**84.1**	**83.9**

The cross-dataset evaluation results show a performance drop for all models compared to the test set of the primary dataset, reflecting the challenge of generalizing across different construction environments and practices. However, MS-FFNet maintains the highest performance among all models, with 84.3% accuracy compared to 81.6% for EfficientNet-B3 (the second-best model). This demonstrates the enhanced generalization capability of MS-FFNet, which is particularly important for real-world deployment in diverse construction environments.

REFERENCES:

[1] J. Yang, M. W. Park, P. A. Vela, and M. Golparvar-Fard, "Construction performance monitoring via still images, time-lapse photos, and video streams: Now, tomorrow, and the future," Advanced Engineering Informatics, vol. 29, no. 2, pp. 211-224, 2015.
[2] A. Braun, S. Tuttas, A. Borrmann, and U. Stilla, "A concept for automated construction progress monitoring using BIM-based geometric constraints and photogrammetric point clouds," Journal of Information Technology in Construction, vol. 20, pp. 68-79, 2015.
[3] Q. Wang and M.-K. Kim, "Applications of 3D point cloud data in the construction industry: A fifteen-year review from 2004 to 2018," Advanced Engineering Informatics, vol. 39, pp. 306-319, 2019.
[4] E. R. Azar, S. Dickinson, and B. McCabe, "Server-customer interaction tracker: Computer vision–based system to estimate customer-waiting times for service quality management," Journal of Computing in Civil Engineering, vol. 27, no. 6, pp. 635-644, 2013.
[5] Y. M. Ibrahim, A. P. Kaka, G. Aouad, and M. Kagioglou, "Framework for a generic work breakdown structure for building projects," Construction Innovation, vol. 9, no. 4, pp. 388-405, 2009.
[6] H. Son, C. Kim, and C. Kim, "Automated color model–based concrete detection in construction-site images by using machine learning algorithms," Journal of Computing in Civil Engineering, vol. 26, no. 3, pp. 421-433, 2012.
[7] K. K. Han and M. Golparvar-Fard, "Potential of big visual data and building information modeling for construction performance analytics: An exploratory study," Automation in Construction, vol. 73, pp. 184-198, 2017.
[8] J. Kim and S. Chi, "Multi-camera vision-based productivity monitoring of earthmoving operations," Automation in Construction, vol. 112, p. 103121, 2020.
[9] Y. Liang, W. Wu, J. Cui, and D. Lu, "Transformer-based multi-task learning architecture for construction activity recognition and progress estimation," Automation in Construction, vol. 134, p. 104090, 2022.
[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in Advances in Neural Information Processing Systems, 2012, pp. 1097-1105.

[11] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in International Conference on Learning Representations, 2015.
[12] C. Szegedy et al., "Going deeper with convolutions," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1-9.
[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770-778.
[14] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4700-4708.
[15] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in International Conference on Machine Learning, 2019, pp. 6105-6114.
[16] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in International Conference on Learning Representations, 2021.
[17] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 10012-10022.
[18] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 11976-11986.
[19] M. Golparvar-Fard, F. Peña-Mora, and S. Savarese, "Automated progress monitoring using unordered daily construction photographs and IFC-based building information models," Journal of Computing in Civil Engineering, vol. 29, no. 1, p. 04014025, 2015.
[20] H. Luo, C. Xiong, W. Fang, P. E. Love, B. Zhang, and X. Ouyang, "Convolutional neural networks: Computer vision-based workforce activity assessment in construction," Automation in Construction, vol. 94, pp. 282-289, 2018.
[21] J. Seo, S. Han, S. Lee, and H. Kim, "Computer vision techniques for construction safety and health monitoring," Advanced Engineering Informatics, vol. 29, no. 2, pp. 239-251, 2015.
[22] S. Bang, S. Park, H. Kim, and H. Kim, "Encoder–decoder network for pixel-level road crack detection in black-box images," Computer-Aided Civil and Infrastructure Engineering, vol. 34, no. 8, pp. 713-727, 2019.
[23] I. Brilakis, M. W. Park, and G. Jog, "Automated vision tracking of project related entities," Advanced Engineering Informatics, vol. 25, no. 4, pp. 713-724, 2011.
[24] Y. Wu, Y. Kim, C. H. Baek, and S. Chi, "Multi-class construction equipment sound identification for construction activity monitoring," Journal of Computing in Civil Engineering, vol. 35, no. 3, p. 04021002, 2021.
[25] H. Kim, H. Kim, Y. W. Hong, and H. Byun, "Detecting construction equipment using a region-based fully convolutional network and transfer learning," Journal of Computing in Civil Engineering, vol. 32, no. 2, p. 04017082, 2018.
[26] H. Luo, M. Liu, and F. Wang, "Real-time detection of construction worker activity with deep learning networks," Automation in Construction, vol. 125, p. 103633, 2021.
[27] Y. Zhang, S. Prakash, and F. N. Ramirez, "Building damage detection using U-Net with attention mechanism from pre-disaster and post-disaster remote sensing datasets," Remote Sensing, vol. 13, no. 5, p. 905, 2021.
[28] J. Li, X. Wang, H. Lin, and F. Wang, "Vision transformer for construction site safety: Safety helmet wearing detection," Automation in Construction, vol. 131, p. 103899, 2021.
[29] C. Wang, Y. Cho, and M. Gai, "Combining CNN and transformer for construction equipment recognition and tracking," Journal of Computing in Civil Engineering, vol. 36, no. 3, p. 04022006, 2022.
[30] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," Journal of Machine Learning Research, vol. 9, pp. 2579-2605, 2008.