Evaluating Multimodal AI for Winter Road Surface Conditions Monitoring: Does Metadata Improve Classification Accuracy?

Yong Lee, Mingjian Wu, and Tae J. Kwon

Abstract— **Harsh winter weather conditions create hazardous driving environments and impose substantial costs on winter road maintenance (WRM). Road Weather Information Systems (RWIS), one of the most critical highway infrastructure sensor suites, provide essential road surface conditions (RSC) measurements and imagery data during winter; however, RSC monitoring remains heavily dependent on manual interpretation of its data, which is time-consuming and limits its real-time applicability. This dependency not only reduces efficiency but also underutilizes the full potential of available data. While multimodal artificial intelligence (AI) techniques have been applied in various domains, their potential benefits for RWIS data have not been explored. To address this gap, this study investigates whether multimodal AI, specifically convolutional neural networks (CNNs) and vision transformers (ViT) combined with a dual cross-attention fusion mechanism, can improve RSC classification compared to vision-only models. To evaluate the feasibility and added value of this approach, we trained and tested four different models on a balanced dataset comprising 5,136 training images and 2,200 testing images collected over multiple winter seasons in the state of Iowa, US. The results indicate that the CNN image-only model achieved 81.4% accuracy, 79.8% balanced accuracy, and a macro F1-score of 77.6%, while its multimodal counterpart improved these metrics to 90.2%, 84.0%, and 86.3%, respectively. Similarly, the ViT image-only model recorded 88.1% accuracy, 76.0% balanced accuracy, and 78.0% macro F1, compared to 91.5%, 85.6%, and 88.0% for the ViT fusion model. While additional testing with larger datasets is warranted, these findings provide strong empirical evidence that integrating metadata with imagery significantly improves RSC classification accuracy across varying road weather conditions, thereby enhancing real-time assessments and contributing to more effective WRM operations and improved road safety for all drivers during winter months.**

INTRODUCTION

Harsh winter weather in northern regions of the United States and Canada poses significant challenges to road safety and mobility. Heavy snowfall and ice not only contribute to 16% of fatal motor vehicle crashes in the U.S. between 1994 and 2012 [1] but also reduce traffic volumes and road capacity by as much as 56% and 40%, respectively [2], [3], [4]. In accordance, state highway agencies invest billions of dollars annually in winter road maintenance (WRM); however, limited resources and the complex task of prioritizing WRM necessitate technological solutions that offer timely, accurate assessments of road surface conditions.

To tackle such challenges, agencies often adopt Road Weather Information Systems (RWIS), which are integrated networks of sensors, cameras, and data loggers deployed along roadways to continuously monitor meteorological and surface conditions. These systems deliver critical real-time data, including high-resolution images, temperature, humidity, wind parameters, and pavement condition measurements that support proactive maintenance and enhanced road safety. Despite the extensive data collected by RWIS, real-time evaluation of road surface condition (RSC) still relies heavily on manual interpretation, which is time consuming, labor-intensive, and undermines operational efficiency. Recent advances in computer vision using deep learning, particularly using convolutional neural networks (CNNs) and vision transformers (ViT) for image-based analysis [5], [6], [7], [8], [9], [10], have improved efficiency, but integrating the wealth of supplementary metadata remains underexplored. Some studies have explored multimodal approaches outside deep-learning frameworks [11], [12], demonstrating that integrating imagery with underutilized data streams can capture complex environmental interactions and improve prediction reliability.

Our study addresses this gap by evaluating the feasibility and added value of multimodal AI for RSC classification. Specifically, we compare CNN-based and ViT-based architectures within a unified multimodal framework to determine whether integrating sensor metadata with imagery improves classification performance over vision-only models. The findings of this study have the potential to enhance prediction accuracy and support data-driven decision-making in WRM operations by providing more accurate and reliable real-time assessments of winter road conditions.

*Research supported by Faculty of Engineering of University of Alberta and Aurora Program.

Yong Lee is with the Department of Electrical and Computer Engineering, University of Alberta, Edmonton, AB, Canada, T6G 2W2 (e-mail: yongwook@ualberta.ca).

Mingjian Wu is with the Department of Civil and Environmental Engineering, University of Alberta, Edmonton, AB, Canada, T6G 2W2 (e-mail: mingjian.wu@ualberta.ca).

Tae J. Kwon is with the Department of Civil and Environmental Engineering, University of Alberta, Edmonton, AB, Canada, T6G 2W2 (corresponding author; phone: +1(780) 492-6121; e-mail: tjkwon@ualberta.ca).

METHODOLOGY

A. CNN and Transformers

Our approach begins with two baseline models that exclusively process visual data. The first baseline is a CNN model based on a state-of-the-art ConvNeXt architecture [13] pretrained on ImageNet. This model captures spatial hierarchies by automatically learning to detect textures, edges, and other critical features [14]. In contrast, the second baseline employs a Vision Transformer (ViT) that segments each image into fixed size patches, embeds these patches into a latent space, and processes them using multiheaded self-attention layers [15]. The ViT model is designed to capture long-range dependencies and complex visual patterns, making it particularly effective for recognizing nuanced RSC variations. These image-only models, chosen primarily for their proven effectiveness in visual feature extraction and classification, establish the performance baseline and set the stage for evaluating the gains from integrating additional modalities.

B. Multimodal Fusion Methods

To fully exploit the heterogeneous data collected by RWIS, our fusion models integrate image features with complementary meteorological and pavement sensor data. Rather than simply concatenating these inputs, our proposed architecture employs a dual cross-attention mechanism. Meteorological parameters (e.g., air temperature, dew point, wind speed, wind direction, wind gust) and pavement sensor readings (indicating surface condition and temperature) are normalized and encoded via multilayer perceptrons. In the dual cross-attention scheme, queries from one modality interact with keys and values from the other, allowing the model to dynamically weigh the most relevant features based on contextual cues as shown in **Figure 1.** This approach, inspired by prior work in multimodal fusion [16], [17], addresses the underutilization of metadata and leads to improved overall accuracy and better balanced predictions across RSC classes.

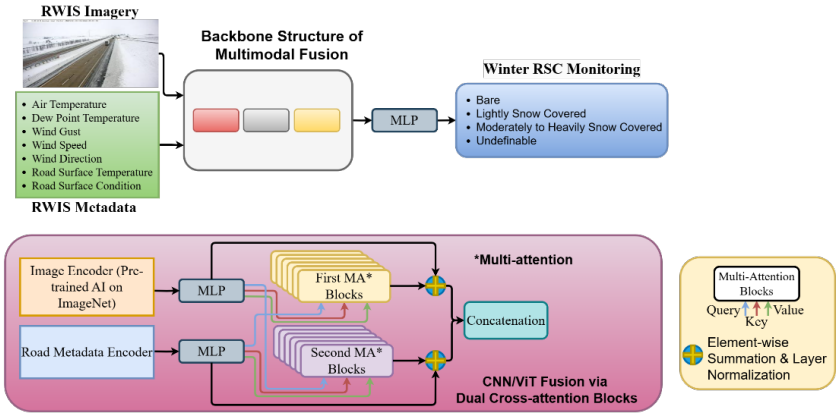

Figure 1. Overview of Multimodal Fusion Architectures

CASE STUDY

A. Study Area and Data

The empirical evaluation used a comprehensive dataset from RWIS stations along Interstates 35 and 80 in Iowa, U.S. These data were collected over four winter seasons (December, January, and February from 2021 to 2024), ensuring images were captured under adequate lighting conditions during daytime. To maximize coverage, four strategically deployed cameras captured road conditions concurrently with atmospheric and pavement sensors. Each image was paired with its corresponding meteorological and pavement data to ensure consistency using a strict time threshold of 30 minutes. To simulate real-world scenarios and validate our model in actual use cases, the data were split chronologically: the winter seasons from December 2021 to February 2023 served as the training set, while those from December 2023 to February 2024 served as the testing set. This approach ensures that the model is evaluated on future, unseen conditions, improving its generalizability and robustness to temporal variations in road weather patterns. The final dataset comprises 7,336 image-metadata pairs with 5,136 used for training and 2,200 for testing for testing. Each image was manually labeled into one of four RSC categories: Bare, Lightly Snow Covered, Moderately to Heavily Snow Covered, and Undefinable. To address class imbalance, we down-sampled the dataset before training. First, we determined the number of samples per class and identified the majority class alongside the count of the

second most frequent class. For any majority class exceeding this count, we randomly selected a subset equal in size to the second most frequent class. Finally, we combined these sampled indices with all indices from the other classes to form a balanced dataset. This extensive dataset provides rich contextual information for both visual and sensor modalities, forming a robust basis for evaluating the fusion strategy.

B. Results

Performance evaluation was based on overall accuracy, balanced accuracy, and macro F1-score. As shown in **Table 1**, multimodal fusion models consistently outperformed their image-only counterparts across all metrics. The CNN dual cross-attention model achieved over 5% higher balanced accuracy and nearly 9% higher macro F1-score than the CNN image-only model, while the ViT fusion model improved balanced accuracy by nearly 10% points over its vision-only counterpart. These gains suggest that integrating meteorological and surface sensor data provides critical contextual information, whereby allowing the model to better differentiate between road surface conditions.

Table 1. Performance Comparison of Different Models

Model Architecture	Accuracy	Balanced Accuracy	F1 Macro
CNN Image Only	81.4%	79.8%	77.6%
CNN Dual Cross-Attention	90.2%	84.0%	86.3%
ViT Image Only	88.1%	76.0%	78.0%
ViT Dual Cross-Attention	91.5%	85.6%	88.0%

Confusion matrices in **Figure 2** further highlight that fusion-based models reduce misclassification rates particularly for lightly snow-covered and moderately to heavily snow-covered categories, where image-only models often struggle likely due to visual similarities. By incorporating supplementary metadata, multimodal AI enhances classification consistency across all RSC categories, leading to more reliable predictions. These findings further strengthen the broader argument that leveraging fusion methods enables a more robust and generalizable approach to winter road condition monitoring. Beyond reducing reliance on manual interpretation, this integration improves the accuracy of automated RSC assessments, thereby supporting more effective and data-driven WRM decision-making.

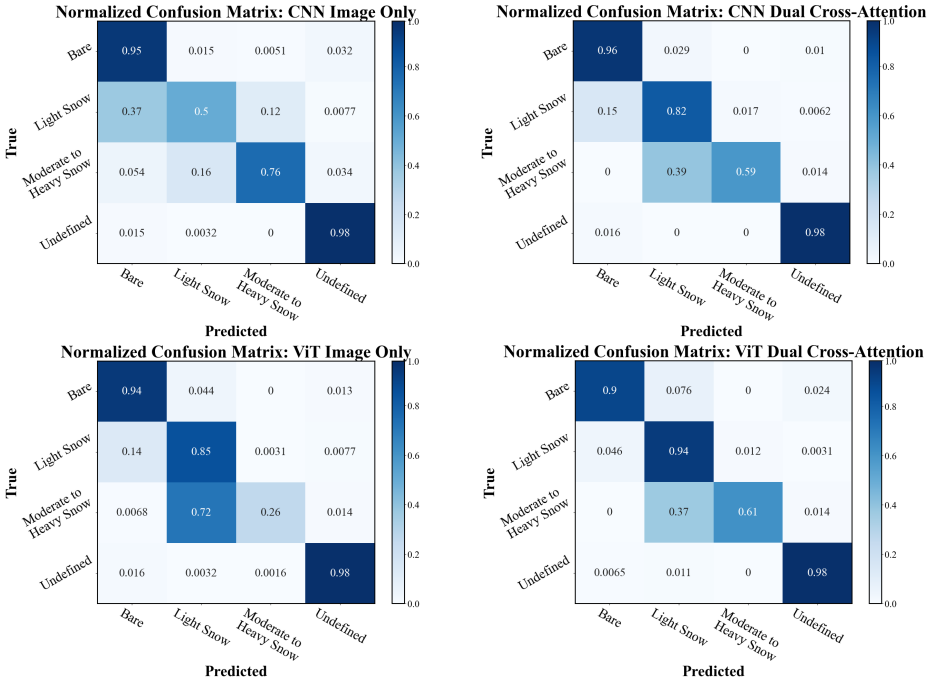

Figure 2. Comparison of Various Fusion Architectures' Normalized Confusion Matrix

CONCLUSION

This study demonstrates that integrating RWIS imagery with meteorological and pavement sensor data via a dual cross-attention fusion mechanism provides a marked improvement in RSC classification compared to image-only methods. The fusion models, particularly those based on ViT, achieve higher overall accuracies and, more importantly, better-balanced class predictions. These improvements not only enhance real-time road conditions assessments but also demonstrate the effectiveness of multimodal AI in dynamically integrating visual and sensor data. This advancement allows more precise and data-driven WRM decisions, which in turn can lead to more effective and efficient operational efficiency and enhanced roadway safety during inclement weather events. Future research will explore incorporating friction data from RWIS to provide a more objective measure of road safety further validating and refining the current approach while contributing to a more comprehensive framework for winter RSC monitoring.

ACKNOWLEDGMENT

The authors would like to thank the Iowa Department of Transportation for providing the data used to complete this study. We would also like to thank the Aurora Program (www.aurora-program.org) – an international research consortium for advancing road weather information systems technology, for funding this research project.

REFERENCES

[1] S. Saha, P. Schramm, A. Nolan, and J. Hess, "Adverse weather conditions and fatal motor vehicle crashes in the United States, 1994-2012," *Environ. Health*, vol. 15, no. 1, p. 104, Dec. 2016, doi: 10.1186/s12940-016-0189-x.

[2] T. J. Kwon, L. Fu, and C. Jiang, "Effect of Winter Weather and Road Surface Conditions on Macroscopic Traffic Parameters," *Transp. Res. Rec.*, vol. 2329, no. 1, pp. 54–62, Jan. 2013, doi: 10.3141/2329-07.

[3] M. Cools, E. Moons, and G. Wets, "Assessing the impact of weather on traffic intensity," *Weather Clim. Soc.*, vol. 2, no. 1, pp. 60–68, 2010.

[4] S. Datla and S. Sharma, "Impact of cold and snow on temporal and spatial variations of highway traffic volumes," *J. Transp. Geogr.*, vol. 16, no. 5, pp. 358–372, Sep. 2008, doi: 10.1016/j.jtrangeo.2007.12.003.

[5] Q. Xie and T. J. Kwon, "Development of a Highly Transferable Urban Winter Road Surface Classification Model: A Deep Learning Approach," *Transp. Res. Rec. J. Transp. Res. Board*, vol. 2676, no. 10, pp. 445–459, Oct. 2022, doi: 10.1177/03611981221090235.

[6] M. Wu and T. J. Kwon, "An Automatic Architecture Designing Approach of Convolutional Neural Networks for Road Surface Conditions Image Recognition: Tradeoff between Accuracy and Efficiency," *J. Sens.*, vol. 2022, pp. 1–11, Jul. 2022, doi: 10.1155/2022/3325282.

[7] R. Ojala and E. Alamikkotervo, "Road Surface Friction Estimation for Winter Conditions Utilising General Visual Features," *ArXiv Prepr. ArXiv240416578*, 2024, Accessed: Jun. 19, 2024. [Online]. Available: https://arxiv.org/abs/2404.16578

[8] A. Abdelraouf, M. Abdel-Aty, and Y. Wu, "Using Vision Transformers for Spatial-Context-Aware Rain and Road Surface Condition Detection on Freeways," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 10, pp. 18546–18556, Oct. 2022, doi: 10.1109/TITS.2022.3150715.

[9] G. Yu and C. Zhang, "A Transformer-based Video Segmentation Method for Road Weather Conditions," in *2024 International Symposium on Intelligent Robotics and Systems (ISoIRS)*, IEEE, 2024, pp. 299–304. Accessed: Oct. 12, 2024. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10649691/?casa_token=3bL2rjm6OD8AAAAA:scBMqlf0s_IrXD50UbXFyq6JLiE67bzZ_Eo6lRxoE6jqeiyBs5mV9DqigCnTfTpHNqDsFoyQcg

[10] Z. Bai, Y. Wang, A. Zhang, H. Wei, and G. Pan, "Road Surface Condition Monitoring in Extreme Weather Using a Feature-Learning Enhanced Mask–RCNN," *J. Transp. Eng. Part B Pavements*, vol. 150, no. 3, p. 04024030, Sep. 2024, doi: 10.1061/JPEODX.PVENG-1503.

[11] S. Yang and C. Lei, "Research on the Classification Method of Complex Snow and Ice Cover on Highway Pavement Based on Image-Meteorology-Temperature Fusion," *IEEE Sens. J.*, vol. 24, no. 2, pp. 1784–1791, Jan. 2024, doi: 10.1109/JSEN.2023.3336667.

[12] J. Carrillo and M. Crowley, "Integration of roadside camera images and weather data for monitoring winter road surface conditions," 2019.

[13] 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. Accessed: Oct. 01, 2024. [Online]. Available: http://openaccess.thecvf.com/content/CVPR2022/html/Liu_A_ConvNet_for_the_2020s_CVPR_2022_paper.html

[14] C. Ma, J.-B. Huang, X. Yang, and M.-H. Yang, "Hierarchical convolutional features for visual tracking," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 3074–3082. Accessed: Oct. 01, 2024. [Online]. Available: https://www.cv-foundation.org/openaccess/content_iccv_2015/html/Ma_Hierarchical_Convolutional_Features_ICCV_2015_paper.html

[15] A. DOSOVITSKIY, "An image is worth 16x16 words: Transformers for image recognition at scale," *ArXiv Prepr. ArXiv201011929*, 2020.

[16] Y. Moroto, K. Maeda, R. Togo, T. Ogawa, and M. Haseyama, "Multimodal Transformer Model Using Time-Series Data to Classify Winter Road Surface Conditions," *Sensors*, vol. 24, no. 11, p. 3440, 2024.

[17] G. Cai, Y. Zhu, Y. Wu, X. Jiang, J. Ye, and D. Yang, "A multimodal transformer to fuse images and metadata for skin disease classification," *Vis. Comput.*, vol. 39, no. 7, pp. 2781–2793, Jul. 2023, doi: 10.1007/s00371-022-02492-4.