A Multimodal Fusion Architecture and Dataset: Advancing Camera and Geophone Integration for Smarter Infrastructure

MATTHEW Y. TAKARA and KATHERINE A. FLANIGAN

ABSTRACT

Computer vision has become integral to the operation and management of complex systems, improving functionality by reducing the reliance on physical contact sensors. Multimodal data fusion techniques further enhance these systems by integrating diverse data sources to uncover complex patterns in the environment. While RGB images are often fused with other imagery types (e.g., Lidar, infrared) to enhance system performance under challenging conditions such as poor weather or occlusions, civil infrastructure applications present distinct challenges. These challenges stem from limited robustness under non-ideal conditions, privacy concerns, and resource constraints—factors that demand the integration of heterogeneous data from both vision-based and non-vision-based sensors. Low-dimensional data, such as time-series data collected in situ from physical sensors, can provide valuable information without compromising privacy or overburdening computational infrastructure. However, this data generally contains less information, making it more difficult to train robust models and underscoring the need for a deeper understanding of the tradeoffs among information value, scalability, and the level of privacy offered by different data types. In this paper, we develop a time-synchronized dataset that includes RGB images and geophone vibration data collected in situ. We then evaluate baseline models comparing vehicle classification performance of unimodal models to that of a spatial-attention fusion model for vehicle classification on a noisy urban road. Our preliminary fusion model shows improvements in classification accuracy over the image- or vibration-based models alone, laying the foundation for broader integration of diverse vision and non-vision modalities.

INTRODUCTION

Recent advances in artificial intelligence (AI) have fueled the development of smart city technologies, enabling data-driven insights across urban infrastructure systems [1]. These capabilities depend on increasingly diverse sources of data, including high-resolution cameras, in situ sensors, crowd-sourced observations, and contextual

Matthew Y. Takara[1], Katherine A. Flanigan, Ph.D[2] (Corresponding author). Email: {mtakara[1], kflaniga[2]}@andrew.cmu.edu, Department of Civil and Environmental Engineering, Carnegie Mellon University, Pittsburgh, PA, USA

TABLE I. Comparison of existing multimodal datasets and corresponding analyses. Note: There are numerous additional datasets that do not fulfill any of the three pillars.

Author (Year)	Modalities	Addresses Privacy	Addresses Performance	Addresses Efficiency
Naphade, et al. (2023) [9]	Text, Camera, Synthetic Data	Redacted license plates and faces	✓	✗
Alibeigi, et al. (2023) [10]	Lidar, GPS, Camera, Inertial measurement unit	Facial blurring and replacement with synthetic data	✓	✗
Liu, et al. (2023) [11]	Camera, Lidar	✗	✓	✓
Franchi, et al. (2024) [12]	Camera, Depth, Infrared	Option to remove personal data	✓	✗
El Ahmar, et al. (2023) [13]	Camera, Thermal	✗	✓	✓
Ma, et al. (2024) [14]	Camera, Lidar, Fisheye	Blurring faces	✓	✗

databases such as weather feeds. While vision-based sensing—particularly from ubiquitous traffic and security cameras—offers rich, high-resolution information and has driven major advances in traffic monitoring [2], structural health assessment [3], and public safety [4], it also presents notable challenges. These include high computational, network, and storage infrastructure demands [5] and persistent public concerns about surveillance and privacy [6]. As a result, there is growing recognition of the need to complement vision data with other sensing modalities that may be less invasive, more efficient, or better suited to large-scale deployment.

One promising response to these challenges is multimodal fusion—the integration of multiple, heterogeneous data streams to improve robustness, accuracy, and adaptability in real-world decision making. Multimodal fusion can capitalize on the complementary strengths of different sensor types: while cameras provide visual detail, in situ sensors like geophones detect vibrations with much lower bandwidth and without capturing personally identifiable information. For example, a passing vehicle can be observed visually via a camera, or seismically via a geophone [7]. Each modality encodes different features of the same event. While prior research has extensively studied multimodal fusion within the visual domain—e.g., combining RGB, depth, and LiDAR for autonomous driving [8]—less attention has been paid to fusing vision with non-vision data.

This work aims to fill that gap by introducing a multimodal fusion architecture and accompanying dataset that combines low quality vision (RGB cameras), in situ time-series data (geophones and laser vibrometry), contextual data (weather), and sparse spatiotemporal data (bus locations). While this full dataset lays the foundation for broader multimodal integration, this preliminary paper focuses on fusing a subset of the data from geophones with camera-based vision data. To the best of our knowledge, the fusion of camera-based data with geophone signals has not been explored in the existing literature. Our framework is grounded in the recognition that real-world sensing must strike a balance across three key pillars: performance, efficiency, and responsibility (e.g., meeting privacy constraints). Existing approaches typically optimize one or two of these dimensions in isolation; for instance, maximizing accuracy without considering effi-

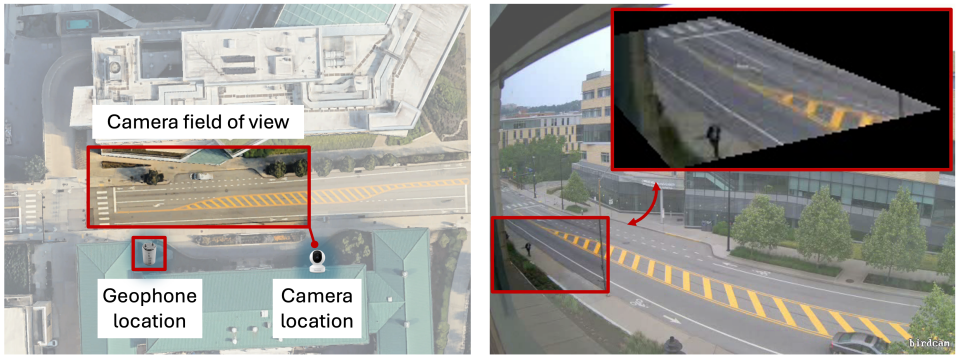

Figure 1. Map of camera and geophone placements (left), alongside an example camera view with its corresponding masked and zoomed-in version (right).

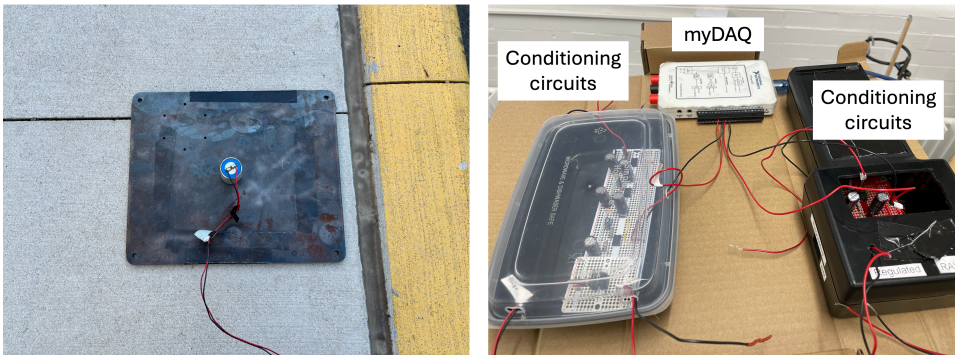

Figure 2. Uncovered and mounted geophone (left) and signal conditioning circuits connected to a National Instruments myDAQ data acquisition device (right).

ciency or privacy implications (Table I). In contrast, we propose a holistic approach that aims to evaluate trade-offs across these dimensions and empower stakeholders to tailor systems to their contextual needs, technical constraints, and community values.

Our approach is informed by the principle of data minimization—collecting only what is necessary to achieve a task with acceptable performance [15]. Under the broader "privacy by design" paradigm, reducing the collection of privacy-sensitive data reduces potential harms from security breaches and public mistrust [16]. For instance, geophone signals can capture critical traffic-related information while producing a fraction of the data volume of 1080p video, without exposing identifying visual details. While encryption, anonymization, and other post-hoc techniques can mitigate risk, upstream design decisions—such as what sensors to deploy and how to fuse their outputs—represent a more proactive means of reducing privacy risk while supporting scalability.

To this end, we present a preliminary study focused on fusing camera data with geophone signals. This represents a critical first step toward developing scalable, privacy-preserving, and context-sensitive monitoring systems for smart infrastructure. We demonstrate that by leveraging the complementary information from visual data and lightweight, privacy preserving in situ sources, performance can be improved through multimodal fusion. In doing so, we aim to contribute a scientifically grounded, human-centered approach to multimodal sensing in the smart city landscape.

DATA COLLECTION

A road segment on the Carnegie Mellon University campus (Pittsburgh, PA) was selected as the data collection site. This section, situated between two university buildings, was chosen for its consistent traffic flow with heterogeneous vehicle types (e.g., passenger vehicles, trucks, busses), accessibility for sensor installation, and minimal disruption to both vehicular and pedestrian traffic. The observed roadway is a two-lane street with eastbound and westbound lanes, a left-turn pocket in the westbound direction, dedicated bike lanes, and sidewalks on both sides. It also includes a loading zone adjacent to a university building and a signalized intersection, offering a variety of real-world traffic conditions. Figure 1 provides an overview of these features.

To capture multimodal data, we deployed a Reolink E1 smart-home RGB camera and a ground-mounted geophone sensing system. The camera was positioned on the third floor of a nearby university building, overlooking the roadway (Figure 1). Video was recorded at 10 frames per second (FPS) at 480p resolution and stored locally on a micro SD card. Ground velocity was collected in real time and synchronized with the camera using a geophone sampled at 250 Hz. The geophone was mounted on a steel plate fixed to the pavement to improve coupling with the pavement and covered with a protective box to shield it from environmental conditions [17]. The geophone signal was routed through an adjustable-gain amplification circuit (referred to herein as the conditioning circuits) and recorded using a National Instruments myDAQ data acquisition device (Figure 2).

The dataset was generated over four days (June 06—09, 2023) during peak traffic (8:00-–10:00AM and 3:00–5:00PM), resulting in eight data collection sessions. In total, approximately 13.5 hours of usable multimodal data were gathered. Following

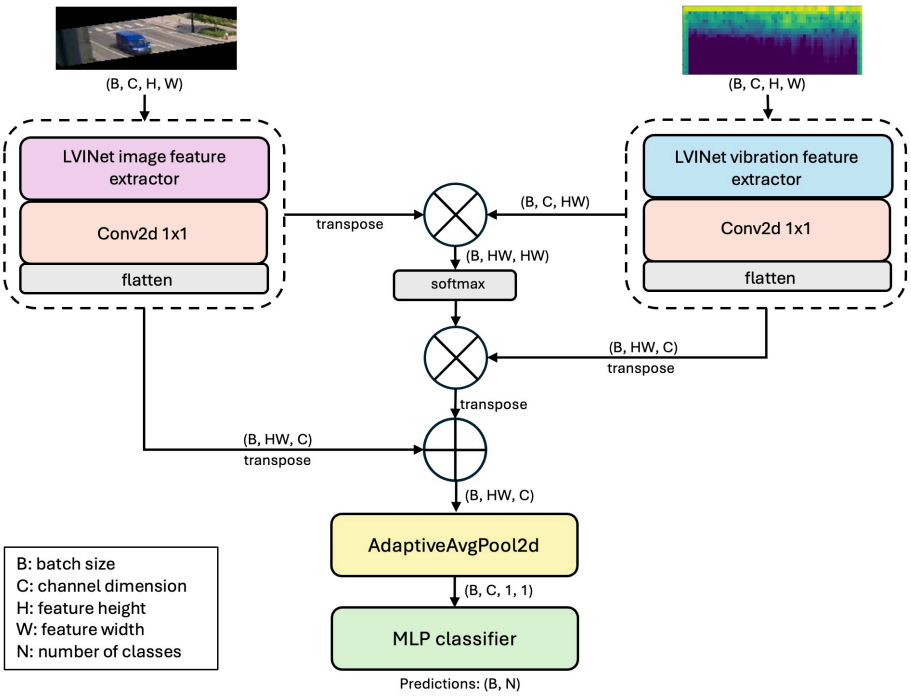

Figure 3. Fusion network structure. Two instances of the LVINet backbone are used to generate feature maps for images and vibration spectrograms, respectively. They are then passed to a simple spatial attention module, and finally through a fully connected classifier layer.

post-processing, the dataset included 104,915 video frames at 480×640 resolution and approximately 1.8 GB of geophone data sampled at 250 Hz. Minor timestamp misalignments between the geophone and camera data streams were corrected during preprocessing to account for latency differences.

DATA POST-PROCESSING, LABELING, AND PREPARATION

Ground truth labels were generated through a two-stage process. First, the image data was downsampled from 10 FPS to 2.5 FPS to reduce labeling costs and computational load. These frames were then passed through a pretrained YOLOv8 object detection model [18] to identify vehicles within each frame. Finally, all model-generated labels were reviewed and corrected manually to ensure labeling accuracy.

To facilitate targeted analysis of vehicle classification tasks, we applied additional preprocessing to the images and ground truth labels. We performed image background masking and cropping to 144×72 pixels. This isolates and retains only the relevant portions of the roadway, minimizes irrelevant background information, and improves alignment between the visual and vibration-based data. We then created a subset of the ground truth consisting of only examples where at most one vehicle is in either westbound or eastbound lane of the image. This restriction was imposed to simplify the classification problem and ensure label consistency at the work's preliminary stages. After downsampling to improve balance, class distributions for this subsets are summarized as car (train: 3596; validation: 1200; test: 1198), truck (train: 1,798; validation: 600; test: 599), and bus (train: 2,110; validation: 703; test: 703).

Geophone signals were preprocessed to correct for variability introduced by hardware and environmental factors. Different gain values were applied during each data collection session to adjust for electromechanical sensitivity and ambient noise conditions. As part of preprocessing, these gain differences were accounted for by normalizing the raw signals accordingly. All geophone time series data were then min-max scaled to the range $[0, 1]$ to ensure consistency across sessions and compatibility with downstream fusion models. We generated mel spectrograms [19] from 3-second (750-point) windows centered on each event (1.5 seconds before and after). These were then resized to match the cropped image dimensions.

EXPERIMENTAL SETUP AND MODEL EVALUATION

Multimodal fusion is challenging due to the mismatch in resolution and dimension of different types of data. Additionally, heterogeneous data fusion methods can be complex and non-intuitive [20]. A common application that fuses vision and non-vision is in audio-visual fusion, which can have applicatioins such as caption generation and action recognition [21]. We tested 3 different models: LVINet [22] for image data only; pretrained MobileNetV2 [23] for vibration data only; and a custom spatial-attention fusion model that uses LVINet to generate feature maps for fusing image and vibration data (Figure 3). MobileNetV2 was chosen for its lightweight applications used in resource constrained or edge environments and LVINet was chosen for its demonstrated use in similar multimodal applications fusing visual and audio data.

TABLE II. Training configuration and hardware specifications.

Parameter	**Value**
Loss function	Cross-entropy with class weighting
Optimizer	Adam
Learning rate	0.001
Batch size	32
Number of workers	8
Number of epochs	100
Training/Validation/Test split	60% / 20% / 20%
CPU	AMD Threadripper 7960X
GPU	NVIDIA RTX 4090 (24GB)

TABLE III. Comparison of image, vibration, and fusion model performance.

Metric	**Image**	**Vibration**	**Image + Vibration**	**Δ (Hybrid - Image)**
Loss	0.3072	0.6557	**0.2511**	**-0.0562**
F1 Weighted	0.8946	0.7355	**0.9158**	**+2.12%**
F1 (Car)	0.9040	0.7569	**0.9230**	**+2.10%**
Precision (Car)	0.8887	0.7932	**0.9357**	**+4.70%**
Recall (Car)	0.9199	0.7237	0.9107	-1.00%
F1 (Bus)	0.9496	0.7620	**0.9598**	**+1.02%**
Precision (Bus)	0.9352	0.7588	**0.9511**	**+1.59%**
Recall (Bus)	0.9644	0.7653	0.9687	**+0.43%**
F1 (Truck)	0.8113	0.6615	**0.8496**	**+3.75%**
Precision (Truck)	0.8598	0.6146	0.8366	-2.71%
Recall (Truck)	0.7679	0.7162	**0.8631**	**+9.52%**

To simulate lower quality and less reliable camera setups, we applied lossy compression, Gaussian blur, and Gaussian noise to images. Additionally, we applied Gaussian noise, and random scaling, temporal shifting, and time-frequency masking to spectrogram data to improve robustness during training. Weighted cross-entropy loss was used to address class imbalance. Additional training specifications are shown in Table II.

RESULTS AND DISCUSSION

Precision, recall, and F1 scores were calculated for each class in the test set and a weighted F1 score was used to measure overall performance (see Table III). The pretrained MobileNetV2 trained on vibration data alone performed worse than LVINet trained on images alone. This was expected due to the lower information content of vibration data. Our fusion network trained on both image and vibration data improved F1 score from the image-only network by 2.12% (89.46% to 91.58%). While not all performance metrics showed improvement, per-class F1 score increased (2.10% for cars, 1.02% for buses, and 3.75% for trucks). We found that the fusion model improved recall for trucks by 9.52% (76.69% to 86.31%) and precision for cars by 4.7% (88.87% to 93.57%). This suggests that the model improved at identifying instances of trucks, reducing the number of missed detections, and reducing the number of false positives for cars. Visually, there is greater similarity in size/shape between cars and trucks compared to buses, which are easy to distinguish from other classes of vehicles. This is reflected

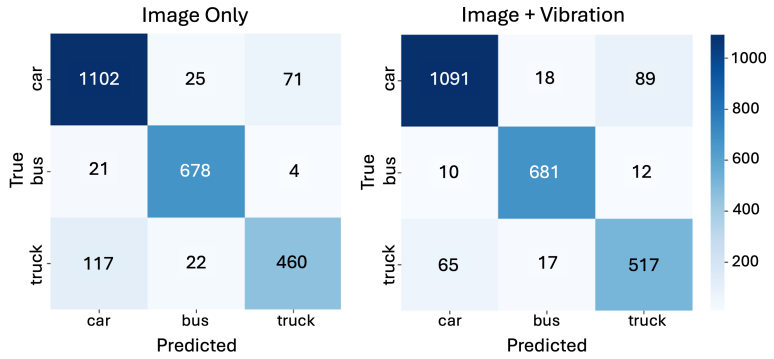

Figure 4. Confusion matrices for image-only test set (left) and fusion test set (right).

in the confusion matrices (Figure 4), where trucks most commonly misclassified as cars. Since cars and trucks vary more in weight, they likely introduce greater variation in road vibrations, which may explain the model's improved performance on these classes.

CONCLUSION

In this work, we present a foundational approach for fusing vision and in situ time series data by training a model to predict vehicle classes using data from RGB cameras and geophones simultaneously. We find that overall, our preliminary fusion model out performs the baseline models trained on a single modality. By fusing information-rich RGB images and low-resolution, but privacy-preserving in situ data, we set the stage to explore a more holistic approach to intelligent monitoring systems that take advantage of diverse, but complementary, sources of information. Future work should investigate vision and non-vision data fusion across the full heterogeneous dataset, in addition to quantifying tradeoffs between value of information and privacy.

ACKNOWLEDGEMENTS

This work is supported by the Pennsylvania Infrastructure Technology Alliance.

REFERENCES

1. Alahi, M. E., A. Sukkuea, F. W. Tina, A. Nag, W. Kurdthongmee, K. Suwannarat, and S. C. Mukhopadhyay. 2023. "Integration of IoT-enabled technologies and artificial intelligence (AI) for smart city scenario: Recent advancements and future trends," *Sensors*, 23(11):5206.
2. Dilek, E. and M. Dener. 2023. "Computer vision applications in intelligent transportation systems: A survey," *Sensors*, 23(6):2938.
3. Dong, C.-Z. and F. N. Catbas. 2021. "A review of computer vision–based structural health monitoring at local and global levels," *Structural Health Monitoring*, 20(2):692–743.
4. Myagmar-Ochir, Y. and W. Kim. 2023. "A survey of video surveillance systems in smart city," *Electronics*, 12(17):3567.
5. Ibrahim, M. R., J. Haworth, and T. Cheng. 2020. "Understanding cities with machine eyes: A review of deep computer vision in urban analytics," *Cities*, 96:102481.
6. Ahmad, K., M. Maabreh, M. Ghaly, K. Khan, J. Qadir, and A. Al-Fuqaha. 2022. "Developing future human-centered smart cities: Critical analysis of smart city security, data management, and ethical challenges," *Computer Science Review*, 43:100452.

7. Ahmad, A. B. and T. Tsuji. 2021. "Traffic Monitoring System Based on Deep Learning and Seismometer Data," *Applied Sciences*, 11(10):4590.
8. Caesar, H., V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom. 2020. "nuScenes: A multimodal dataset for autonomous driving," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11621–11631.
9. Naphade, M., S. Wang, D. C. Anastasiu, Z. Tang, M.-C. Chang, Y. Yao, L. Zheng, M. S. Rahman, M. S. Arya, A. Sharma, Q. Feng, V. Ablavsky, S. Sclaroff, P. Chakraborty, S. Prajapati, A. Li, S. Li, K. Kunadharaju, S. Jiang, and R. Chellappa. 2023. "The 7th AI city challenge," *arXiv preprint arXiv:2304.07500*.
10. Alibeigi, M., W. Ljungbergh, A. Tonderski, G. Hess, A. Lilja, C. Lindström, D. Motorniuk, J. Fu, J. Widahl, and C. Petersson. 2023. "Zenseact open dataset: A large-scale and diverse multimodal dataset for autonomous driving," Paris, France, pp. 20121–20131.
11. Liu, Y., R. Chen, X. Li, L. Kong, Y. Yang, Z. Xia, Y. Bai, X. Zhu, Y. Ma, Y. Li, Y. Qiao, and Y. Hou. 2023. "UniSeg: A unified multi-modal LiDAR segmentation network and the OpenPCSeg codebase," pp. 21662–21673.
12. Franchi, G., M. Hariat, X. Yu, N. Belkhir, A. Manzanera, and D. Filliat. 2024. "InfraParis: A multi-modal and multi-task autonomous driving dataset," pp. 2973–2983.
13. El Ahmar, W., Y. Massoud, D. Kolhatkar, H. AlGhamdi, M. Alja'Afreh, R. Hammoud, and R. Laganiere. 2023. "Enhanced thermal-RGB fusion for robust object detection," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 365–374.
14. Ma, C., L. Qiao, C. Zhu, K. Liu, Z. Kong, Q. Li, X. Zhou, Y. Kan, and W. Wu. 2024. "HoloVIC: Large-scale dataset and benchmark for multi-sensor holographic intersection and vehicle-infrastructure cooperative," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22129–22138.
15. Ganesh, P., C. Tran, R. Shokri, and F. Fioretto. 2024. "The data minimization principle in machine learning," *arXiv preprint arXiv:2405.19471*.
16. Xu, R., N. Baracaldo, and J. Joshi. 2021. "Privacy-preserving machine learning: Methods, challenges and directions," *arXiv preprint arXiv:2108.04417*.
17. Carcione, J. M., H. S. Almalki, and A. N. Qadrouh. 2016. "Geophone-ground coupling with flat bases: Geophone-ground coupling with flat bases," *Geophysical Prospecting*, 64(2):255–267, ISSN 00168025, doi:10.1111/1365-2478.12263.
18. Jocher, G., J. Qiu, and A. Chaurasia. 2023, "Ultralytics YOLO," .
19. Zhang, T., G. Feng, J. Liang, and T. An. 2021. "Acoustic scene classification based on Mel spectrogram decomposition and model merging," *Applied Acoustics*, 182:108258, ISSN 0003-682X, doi:10.1016/j.apacoust.2021.108258.
20. Fadhel, M. A., A. M. Duhaim, A. Saihood, A. Sewify, M. N. A. Al-Hamadani, A. S. Albahri, L. Alzubaidi, A. Gupta, S. Mirjalili, and Y. Gu. 2024. "Comprehensive systematic review of information fusion methods in smart cities and urban environments," *Information Fusion*, 107:102317, ISSN 1566-2535, doi:10.1016/j.inffus.2024.102317.
21. Hori, C., T. Hori, T.-Y. Lee, Z. Zhang, B. Harsham, J. R. Hershey, T. K. Marks, and K. Sumi. 2017. "Attention-based multimodal fusion for video description," in *IEEE international conference on computer vision*, pp. 4193–4202.
22. Shaikh, M. B., D. Chai, S. M. S. Islam, and N. Akhtar. 2024. "Multimodal fusion for audio-image and video action recognition," *Neural Computing and Applications*, 36(10):5499–5513.
23. Sandler, M., A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen. 2018. "Mobilenetv2: Inverted residuals and linear bottlenecks," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4510–4520.