SeeNN: Leveraging Multimodal Deep Learning for In-Flight Long-Range Atmospheric Visibility Estimation in Aviation Safety

Taha Bouhsine *, Giuseppina Carannante.†, Nidhal C. Bouaynaya.‡
Electrical and Computer Engineering Department, Henery M.Rowan College of Engineering, Rowan University, Glassboro, New Jersey, 08028

Soufiane Idbraim §
IRF-SIC Laboratory, Computer Science Department, Faculty of Sciences Agadir, Ibn Zohr University, Agadir, Morocco

Phuong Tran, Grant Morfit, Maggie Mayfield, Charles Cliff Johnson
William J. Hughes Technical Center, Federal Aviation Administration, Atlantic City, NJ, USA

Deep learning (DL) models have attained state-of-the-art performance in numerous fields. Nevertheless, for certain real-world applications, existing models encounter diverse challenges, ranging from a lack of generability to new data to issues of scalability and overfitting. In this context, integrating information extracted from different modalities holds promise as a potential solution to alleviate these challenges. This paper introduces SeeNN (`https://github.com/skywolfmo/seeNN-paper`), a multimodal deep-learning framework for long-range atmospheric visibility estimation. Using multimodal deep learning, SeeNN fuses various modalities to estimate long-range atmospheric visibility. These modalities include RGB imagery, Edge Map, Entropy Map, Depth Map, and Normal Surface Map. Results show that in contrast to single-modality RGB, which achieves only 87.92% accuracy, multimodal deep learning models achieve an accuracy of over 96%. This significant improvement highlights the potential of multimodal approaches to enhance the accuracy and reliability of atmospheric visibility estimation, which is crucial for improving safety in applications such as aviation, maritime navigation, and autonomous vehicles. By addressing challenges such as data variability, environmental factors, and the inherent complexity of atmospheric conditions, SeeNN contributes to more reliable and robust visibility estimation systems, thereby enhancing safety and operational efficiency in critical environments.

I. Introduction

Atmospheric visibility [1] is a critical factor in aviation safety [2–6], directly impacting a pilot's ability to navigate and make critical decisions. The tragic accident involving professional basketball player Kobe Bryant in January 2020 highlights the severe consequences of visibility-related issues. The National Transportation Safety Board (NTSB) report concluded that the pilot's decision to continue visual flight rules (VFR) into instrument meteorological conditions (IMC) led to spatial disorientation and the subsequent crash. This incident underscores the critical need for accurate, real-time visibility estimation in flight.

Visibility estimation in aviation is particularly challenging due to several factors. Currently, pilots rely heavily on their prior knowledge of landmarks and terrain features to gauge visibility [7]. This dependence on pre-existing knowledge makes automatic estimation complex, as systems must account for varying geographical contexts. Additionally, conditions such as flying inside clouds or encountering rapidly changing weather patterns further complicate the problem, requiring robust and adaptable solutions.

While deep learning methods have shown great promise in solving complex problems, they face challenges such as overfitting, generalization issues, and potential biases. These limitations are particularly evident in single-modality approaches, especially those relying solely on RGB images for visibility estimation across diverse flight conditions. RGB

[*] Graduate Research Fellow

[†] Postdoctoral Fellow

[‡] Associate Dean for Research & Graduate Studies and Professor of Electrical & Computer Engineering

[§] Computer Science Professor and Head of IRF-SIC Laboratory data alone often fails to capture crucial atmospheric properties or account for factors like glare, low-light conditions, or rapid weather changes [8–16].

To address these challenges, multimodal deep learning techniques have emerged as a promising solution. By integrating information from diverse modalities, these approaches enhance model capabilities and overcome the limitations of single-modality systems [17–20]. Each modality captures a distinct type of information, often inaccessible through a single modality approach, resulting in a more comprehensive understanding of the environment and more accurate predictions.

The significance of multimodal deep learning solutions in visibility estimation is increasingly recognized [9, 21–29]. These models can overcome limitations that plague single-modality systems [30]. The integration of multimodal deep learning (Table 1) significantly enhances the robustness, safety, and reliability of deep learning solutions, making them particularly well-suited for critical real-world applications where reliability is crucial [31, 32].

Table 1 Literature Review of Modalities Used in the Literature for On-Ground Atmospheric Visibility Estimation

Modality	**[25]**	**[33]**	**[24]**	**[26]**	**[27]**	**[28]**	**[29]**
Depth Map			X				
Transmission Map	X		X		X		
Disparity Map	X						
Entropy						X	X
Edge Detection					X		
Contrast Computation					X		
Koschmieder Law					X		X
FFT		X					
Spectral Filter		X					
Dark Channel Prior				X	X		X

Despite advances in deep learning for visibility estimation, there remains a significant gap in addressing in-flight visibility. This gap is largely attributable to the scarcity of comprehensive in-flight visibility datasets, which are crucial for training and validating deep learning models in aviation contexts [9]. Existing datasets concentrate on short-range visibility, offering a limited spectrum of sceneries, and land covers, and are mainly focused on ground-level atmospheric visibility. Moreover, they are predominantly confined to ground-level altitudes, neglecting the variability and complexity introduced by different elevation viewpoints [9]. This restriction in dataset diversity hampers the development of more universally applicable and robust in-flight visibility estimation models.

In this work, we propose a multimodal framework for training visibility estimation systems to enhance the accuracy, trustworthiness, and robustness of DL models for atmospheric visibility estimation. We demonstrate how integrating diverse modalities can significantly mitigate the limitations inherent in unimodal RGB approaches, paving the way for more versatile and reliable DL applications in fields where environmental variability is critical. We also address the gap in publicly available datasets by creating a comprehensive dataset, capturing the visibility degradation across different land covers and altitudes.

In detail, the contributions of this work are as follows:

- We provide a meticulously curated dataset as a benchmark for visibility estimation and related challenges such as dehazing and visibility restoration [34]. This dataset, named SeeSet V1, will be made available to the community alongside the code at `https://github.com/skywolfmo/seeNN-paper`. The data, acquired from the X-Plane 11 flight simulator, encompasses a wide array of images captured under varied visibility conditions and at multiple altitudes, ranging from ground level to 2,000 feet Above Ground Level (AGL). The dataset's comprehensiveness, spanning a wide range of visibility scenarios at multiple altitudes, establishes a robust foundation for training and evaluating visibility estimation approaches and other in-flight visibility restoration and image dehazing methods.
- We have developed a multimodality fusion framework for estimating atmospheric visibility. This framework is used to train and validate multimodality deep learning models. Our results demonstrate that the multimodality deep learning models outperform the single-modality RGB model in terms of accuracy.

II. Methodology

We introduce a novel framework incorporating multiple modalities for building atmospheric visibility estimation solutions, as well as constructing a dataset that includes both ground-level and elevated altitude conditions.

A. SeeSet V1 Dataset

To overcome the limitations encountered in the previous studies and to include a broader range of real-world scenarios, we collect a novel aerial imagery dataset SeeSet v1. This dataset was carefully curated to incorporate dynamic views capturing sceneries from multiple locations, encompassing ground-based and aerial perspectives.

This section presents a detailed description of the data collection and labeling process (II.A.1). In II.A.2, we present the extraction techniques used to generate the additional image modalities.

1. Dataset Collection Process

To generate our synthetic dataset, we utilize an FAA-approved flight simulator. This advanced simulator facilitates the controlled acquisition of images, showcasing diverse viewpoints and a range of visibility degradations. The process, as illustrated in Figure 1, starts at ground-level altitude. We systematically increased visibility in incremental steps, extending up to 100 miles. Subsequently, when the visibility reaches the limit, we elevate the viewpoint's altitude and then reset the visibility to 0, continuing this procedure up to a maximum of 2000 feet AGL.

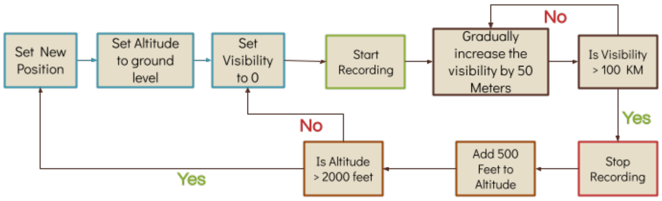


Fig. 1 Automatic Dataset Collection Process using X-Plane 11

The collected images are automatically labeled into five discrete bins, each tailored to specific FAA requirements. This categorization is based on visibility conditions and regulations relevant to both ground-based and aerial environments. The designated bins serve as the basis for the five labels utilized in training our DL models. We report the classes (bins) specifications and the corresponding counts in Table 2.

Table 2 Visibility Categories and Images Count

Category	Visibility in miles	Visibility in meters	Count
4	≥ 5 miles	≥ 8046.72m	67002
3	3 to 5 miles	4828.03m to 8046.72m	19584
2	1 to 3 miles	1609.34m to 4828.03m	19648
1	0.5 miles to 1 mile	804.672m to 1609.34m	4928
0	≤ 0.5 mile	≤ 804.672m	4938
			116100

2. Modalities

Monocular Depth Estimation:

In our approach, we used the Omnidata toolkit to extract depth maps from monocular images [35, 36]. This toolkit provides a comprehensive and scalable method for depth estimation, essential for understanding the spatial arrangement in a scene. The generated depth maps offer a pixel-wise measurement of distance from the viewpoint, aiding in the accurate representation of the three-dimensional structure of the scene.

While depth maps offer valuable information, the models employed to generate them exhibit a notable limitation when applied to our data. The training methodology involves masking the sky and exclusively considering the ground

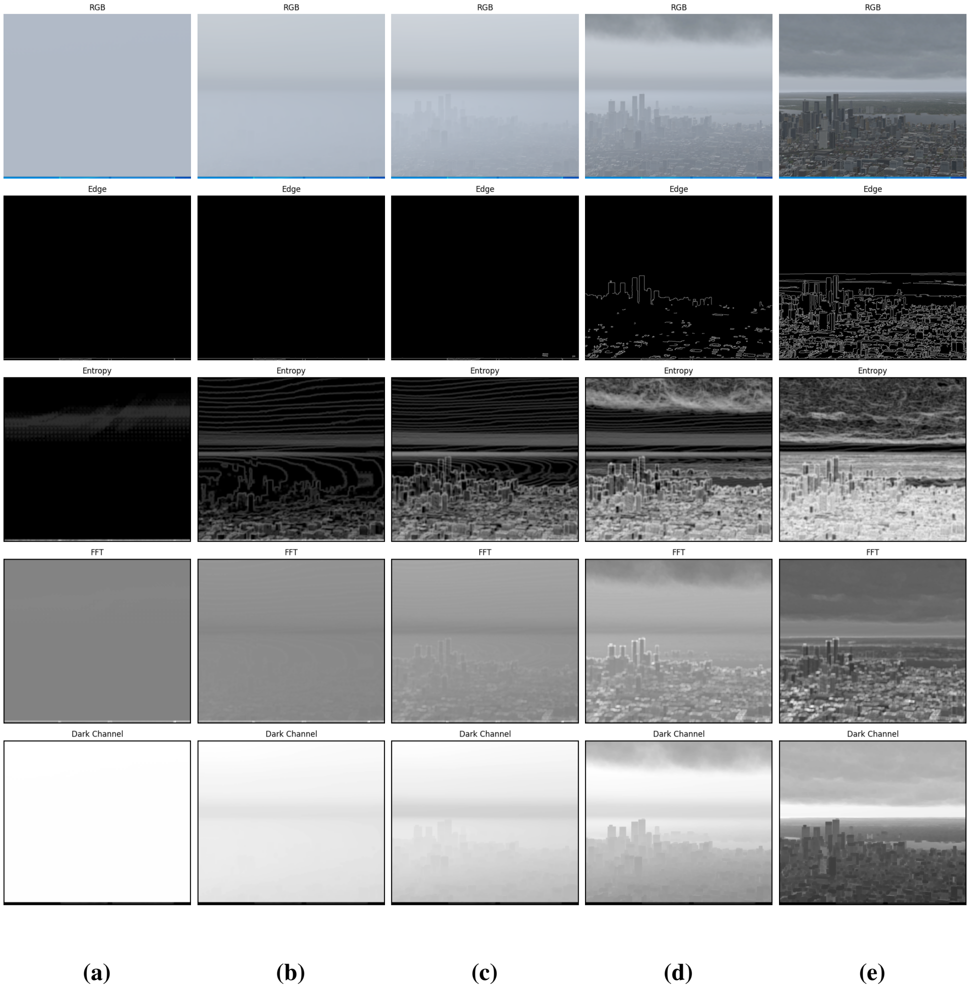


Fig. 2 We show the impact of visibility on the multiple modalities for the 6N7 Sealane 01 View. Each row shows one modality: RGB, edge map, entropy map, FFT magnitude, and dark channel prior. Each column refers to a visibility bin: (a) visibility of < 0.5 mile, (b) visibility range $(0.5, 1]$ miles, (c) visibility range $(1, 3]$ miles, (d) visibility range $(3, 5]$ miles, and (e) Visibility of > 5 miles.

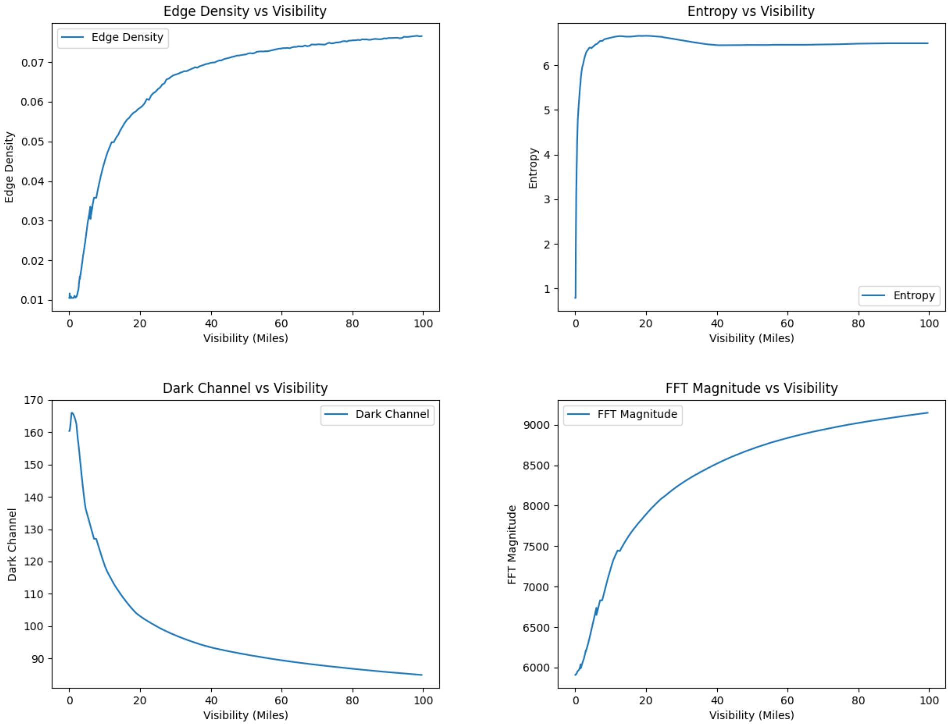


Fig. 3 Impact of Visibility Degradation on Edge Density (a), Entropy map (b), Dark Channel Prior (c), and FFT Magnitude (d) vs Visibility in Miles

before depth estimation. This approach may pose challenges with certain images in our dataset, as they are collected at varying altitudes.

Normal Surface Estimation:

Alongside depth maps, we also used the Omnidata toolkit for normal surface estimation [35]. This modality provides information about the orientation of surfaces in the image, which is crucial for understanding the geometric properties of the scene. Unlike depth estimation, this modality estimator considers both sky and ground details.

Entropy Map:

We incorporate an image entropy map as a modality to enhance the model's sensitivity to changes in visibility, especially in low-visibility conditions. The entropy map quantifies the amount of information present in different regions of an image.

Edge Detection:

Edge detection is another key modality well-suited for scenarios involving long-range visibility where defining objects and scene boundaries is critical. By highlighting the contours and edges within the images, this modality aids in delineating shapes and structures, thus providing a clear distinction between different objects and features in the scene.

In Figures 2, 3, we show the impact of visibility degradation on various modalities of the same scenery. Each row shows one modality, while each column refers to a visibility bin.

B. Fusing Modalities

Numerous methods have been proposed in the literature for fusing different modalities and multiple stream networks [37–39]. These methods range from simple concatenation of different inputs from the input space to fusion at different levels of the model architecture.

Early fusion [40], where we concatenate or preprocess all input streams in the input space, then feed it to a single feature extractor which extracts as much information as possible before feeding it to a decoder, classifier head, or projection head. While this method is the simplest method one can use to fuse different modalities, it has many limits where the feature-extracted model might learn to ignore most of the modalities from the early layers and be dominated

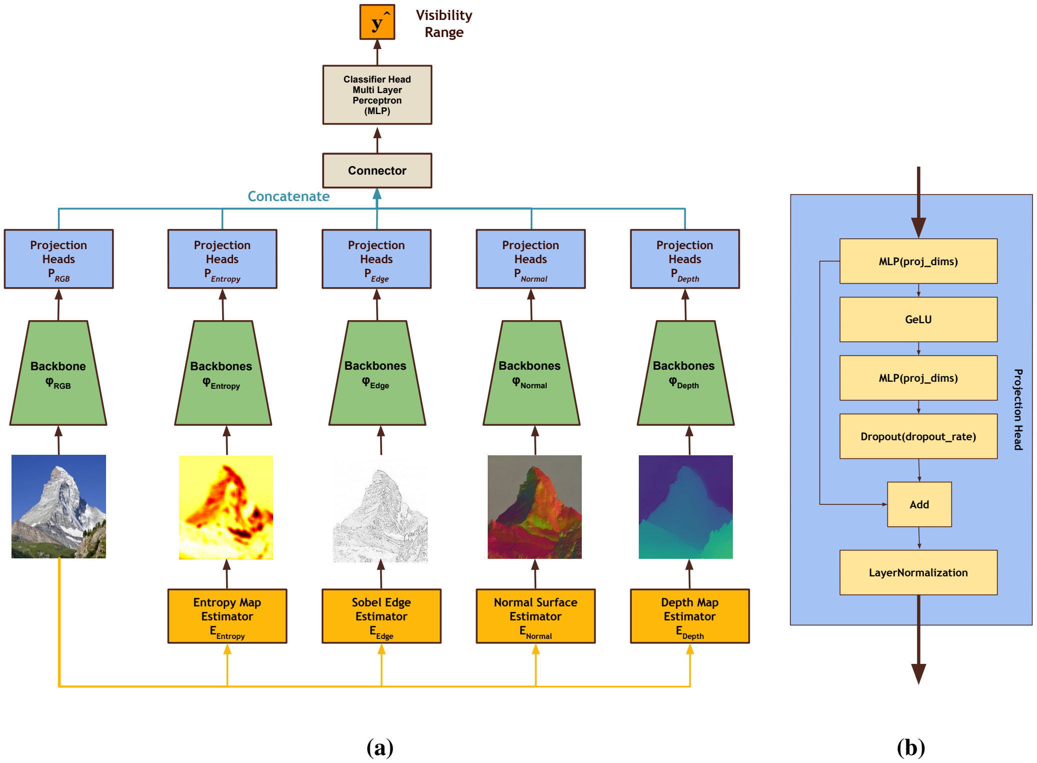

y^
Visibility Range
Classifier Head Multi Layer Perceptron (MLP)
Connector
Concatenate
Projection Heads P_{RGB}
Projection Heads $P_{Entropy}$
Projection Heads P_{Edge}
Projection Heads P_{Normal}
Projection Heads P_{Depth}
Backbone φ_{RGB}
Backbones $\varphi_{Entropy}$
Backbones φ_{Edge}
Backbones φ_{Normal}
Backbones φ_{Depth}
Entropy Map Estimator $E_{Entropy}$
Sobel Edge Estimator E_{Edge}
Normal Surface Estimator E_{Normal}
Depth Map Estimator E_{Depth}
MLP(proj_dims)
GeLU
MLP(proj_dims)
Dropout(dropout_rate)
Add
LayerNormalization
Projection Head
(a)
(b)

Fig. 4 (a) SeeNN Framework: The framework first extracts features (entropy map, surface normals map, edge map, depth map) from the input image. Separate encoders $\phi_m(\cdot)$ ($\phi_m(\cdot)$ denotes modality encoders) process these features followed by a projection head (b), followed by fusion of these features through a Connector and prediction via a classifier $\hat{\mathbf{y}}$. (b) Projection Head: The input vector is transformed by an MLP (Multi-Layer Perceptron) with a non-linear activation function (GeLU) and dropout for regularization.

by only one modality.

Another approach, widely used in the literature, involves feeding the different modalities into separate encoder layers before fusing all the extracted embeddings [40]. In this approach, the model learns to extract useful features from each modality before they are fused, thus preventing one modality from dominating the feature space. But one of the biggest benefits of using this type of architecture is the ability to use recent advancements in representation learning, where each modality is processed through its own encoder, then using either contrastive learning to pre-train all the encoders to align together, or using the unsupervised representation learning where fusion happens in the between the encoder and decoder layer or right from the get-go.

Late fusion is another type of fusion [40], where each modality is passed through its own encoder, decoder, or any number of layers until the decision layer (e.g., binary classification). Fusion is then performed either by voting between different models or by taking the mean of the different decisions.

1. Multimodal Fusion Methods

In the literature, various techniques such as Tensor Fusion [41], Low-Rank Fusion [42], and attention [43] have been proposed Ȧlthough each method has its own advantages and disadvantages, self-attention has emerged as the cornerstone for many recent large-scale methods. While it requires more training data, its computation is significantly low compared to methods like tensor fusion.

2. SeeNN Multimodal Fusion Framework

The proposed SeeNN framework (Figure 4) integrates multimodal DL techniques to process the images alongside the multiple modalities. First, each input RGB image I undergoes a series of transformations through modality estimators to produce a depth map $E_d(I)$, a normal surface map $E_n(I)$, an edge detection map $E_e(I)$, and an entropy map $E_s(I)$. Each of these modalities captures distinct characteristics of the input, providing diverse perspectives on the image's content.

Let m denote one of the modalities (generated depth map $depth$, normal surface $normal$, entropy map $entropy$, edge map $edge$, and RGB image rgb). We implement different backbone models $\Phi_m(\cdot)$ for each modality input X_m. We use DenseNet121 [44] as the architecture for all Φ_m and we feed the resulted embedding from each encoder to a projection head P_m which consists of an MLP with a non-linear activation function (GeLU) and dropout for regularization, followed by a layer normalization which is a crucial step to align the feature representations, mitigating the risk of dominance by any single modality due to varying magnitudes of feature values, then we obtain a feature vector F_m.

We apply this process to RGB image X_{rgb}, depth X_{depth}, normal surface X_{normal}, entropy $X_{entropy}$, and edge map X_{edge} to obtain F_{rgb}, F_{depth}, F_{normal}, $F_{entropy}$, F_{edge}, respectively.

Following projection heads, the SeeNN framework concatenates these embeddings into a single, comprehensive feature vector F. The concatenation is represented as $F = [F_{rgb}, F_{depth}; F_{normal}; F_{entropy}; F_{edge}]$, which is then fed to the connector C that is responsible for fusing these modalities. We then apply an MLP classifier head to get the final prediction $\hat{y}$.

When it comes to the connector, we primarily used two methods. The first method involves passing the flattened F directly to the MLP, which involves a simple fusion of the different features. The second method utilizes an attention block to perform self-attention on F, followed by flattening the output and feeding it to the MLP head.

3. Experimental Setup

For our study, we use our collected dataset, SeeSet v1 II.A, made of 320 distinct views collected across 20 locations with different land covers, each with visibility ranging from 0 to 100 miles. We split the dataset into two subsets using the holdout approach, where we select all the views from specific locations, and hid them from the model during validation, to ensure that our model doesn't overfit the sceneries and vegetation and learn to estimate the visibility based on the degradation of the image [8]. This resulted in $100,350$ instances for training and $15,750$ for validation. Each image in the dataset is preprocessed to an input shape of 224×224 pixels.

We leverage the Omnidata models to preprocess RGB images and export the estimated Depth Map and the Normal Surface [35], which is based on the DPT-Hybrid architecture [36] and is similar to the approach taken in the literature to generate pseudo labels from RGB to pre-train multimodal models [45, 46]. For the other modalities, i.e., edge map and entropy map, the RGB images are initially processed through the hand-made estimators (Figure 4).

We train all models for 100 epochs and we use the Adam optimizer with a learning rate of 0.001. For all models, we set the batch size to 32.

III. Results and Discussion

Our research categorizes atmospheric visibility into five distinct categories, as summarized in Table 2. This categorization is based on the visibility range in miles, which was set according to the requirements of the FAA. Resulting in a comprehensive dataset of 116,100 images spanning various altitudes, land covers, and sceneries.

The training of a single modality RGB model did not achieve high accuracy on the different classes of our problem, with an overall accuracy of 87.92%. This demonstrates the limitations of such models, especially when tested on unseen views. Unlike previous works that were tested on a limited number of views and with a data split that might have caused data leakage between the training set and test set, our model may have given false results during evaluation as it had already seen a similar image during training. To prevent the same problem in our experiments, we used the holdout method as we mentioned in the Experimental Setup, where we hide certain locations totally from the model during the training.

Unlike the single modality RGB model, results from the multimodality models (Table 3, Figure 5) show a big improvement in the accuracy of prediction in the validation set, compared with 87.92% from the RGB model, when we fuse different modalities, we notice a leap of 10% in accuracy. For instance, When we combine the embedding extracted from the RGB model with the embedding extracted from the depth map, the RGB-Depth model achieves a high accuracy of 96.53% using a simple concatenate connector.

The confusion matrix and results table highlight the efficacy of multimodal deep learning in atmospheric visibility estimation for most of the categories (Figure 5). The overall performance is good, but the multimodality approach still struggles with label 3, which represents the visibility range from 3 to 5 miles. We've observed that all different

Table 3 Results comparison of using different modalities,

Connector	RGB	Entropy	Edge	Depth	Normal Surface	# trainable param	val acc.
Unimodality	✓					7M	87.92
Concatenate	✓	✓				14M	96.4
	✓		✓			14M	96.53
	✓			✓		14M	**97.57**
	✓			✓	✓	21M	97.14
	✓	✓	✓	✓	✓	38M	96.3
Self-Attention	✓		✓			14M	96.86
	✓			✓		14M	96.31
	✓			✓	✓	21M	97.47
	✓	✓	✓	✓	✓	38M	**97.63**

combinations of modalities have difficulty with this class specifically, and most of them incorrectly predict it as class 4. Future work could focus on improving our solution for this specific class. The combination of RGB and Depth alone resulted in improved results in this specific class. This suggests that future research could focus on testing different modalities to achieve better results and reducing the number of modalities used to estimate visibility. This would help to reduce the computation cost of estimating visibility without negatively impacting the accuracy of the models.

Another perspective that should be considered when deciding which modality you want to use for the visibility estimation is the computation cost required to run the inference, while the model that merged all the available modalities gave us the best results, it still requires passing the input to all the modality estimators and then to the backbones, which makes it require more resources. and when planning to deploy such models, you will start facing the limits to what your hardware can give, so you need to consider that, especially for embedded devices.

While our dataset addressed the gap in the availability of publicly accessible multi-view datasets for atmospheric visibility, one of its limitations is the lack of diversity in landscapes and land covers. While experiments were conducted to ensure that there was no data leakage and that the model was tested on unseen locations, there is still a need to improve the quality of such a dataset by using the latest simulator technologies such as Microsoft Flight Simulator, X-Plane 12 that provides near real-world simulations for visibility degradation, which will help tackle the lack of available datasets.

Future research should focus on diversifying the dataset, incorporating a wider range of atmospheric conditions and scenarios. This expansion is not just about quantity, but also about variety, ensuring the SeeNN framework is tested against different visibility situations. Adding different land covers will enrich the dataset, making the model more adaptable to varying geographical locations and environmental conditions.

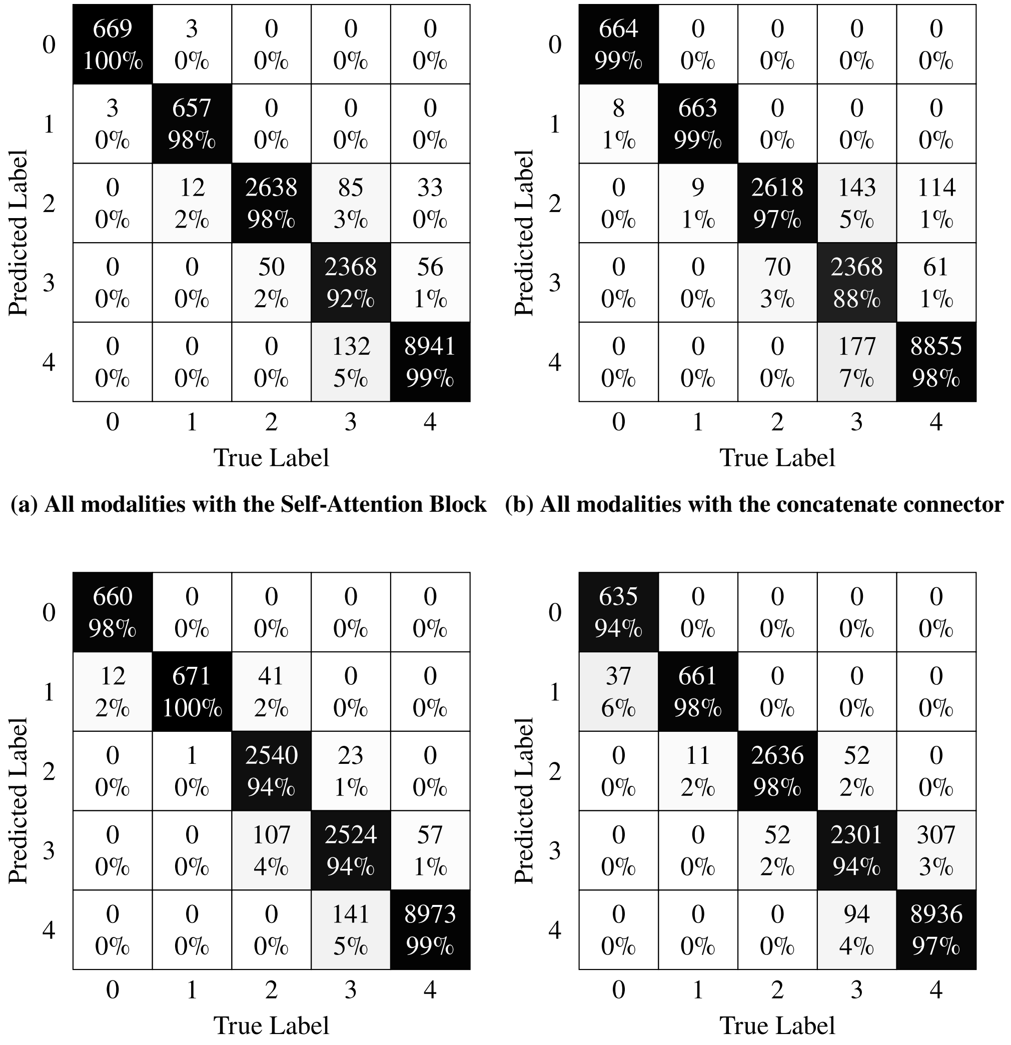
Another future work will be the use of the different pretraining techniques used in the literature to improve representation extracted from the different images, as similar to our work, most of the multimodal works make the use of the advancement of task-specific models to generate pseudo labels that are used to pre-train the model in an unsupervised manner, removing the need of requiring labeled data.

Furthermore, we found the need to understand how the different visibility degradations impact the quality of features extracted features by the different architectures, successfully understanding this point will help us in improving the trustworthiness of such models in real-world situations and not only reliable for sandbox situations.

IV. Conclusion

In this work, we have presented a novel multimodal approach to atmospheric visibility estimation, focusing on challenging in-flight scenarios through the application of advanced deep-learning architectures. Our primary contributions are twofold:

- We propose SeeNN, a multimodal fusion framework that integrates RGB imagery with entropy maps, edge maps, depth information, and normal surface maps. Through rigorous experimentation, we demonstrate that this multimodal approach significantly outperforms single-modality baselines, including traditional RGB-based models. The superior performance of SeeNN underscores the efficacy of leveraging diverse data modalities in addressing the complex task of visibility estimation.



(a) All modalities with the Self-Attention Block

(b) All modalities with the concatenate connector

(c) RGB and Depth Map Model with the Concatenate Connector

(d) RGB and Depth Map Model with the Self-Attention Block

Fig. 5 Confusion Matrix of the top 4 multimodality models

- We introduce a comprehensive, open-source benchmark dataset for atmospheric visibility estimation. This dataset is distinguished by its diversity, encompassing a wide range of altitudes, land cover types, and visibility conditions. It represents a valuable resource for the research community, enabling robust evaluation and comparison of visibility estimation algorithms.

Our empirical results indicate that the proposed multimodal deep learning framework offers substantial improvements in estimation accuracy compared to the single-modality RGB method. The release of our benchmark dataset addresses a critical gap in the field, providing a standardized platform for algorithm development and evaluation. We anticipate that this resource will facilitate rapid progress in the domain, spurring the development of increasingly sophisticated multimodal deep learning techniques for atmospheric visibility estimation.

Future work may explore the integration of additional modalities, the application of more advanced fusion techniques, or the extension of our approach to related problems in atmospheric science. Moreover, the potential for transfer learning and domain adaptation in this context remains an open and promising avenue for investigation.

In conclusion, our work contributes to the growing body of research at the intersection of deep learning and atmospheric science, offering both methodological advancements and resources for the broader research community. As the field continues to evolve, we believe that multimodal approaches like SeeNN will play an increasingly crucial role in addressing complex environmental perception tasks, with far-reaching implications for aviation safety and beyond.

Acknowledgments

References

[1] Malm, W., *Visibility: The Seeing of Near and Distant Landscape Features*, 2016.

[2] Kulesa, G., "WEATHER AND AVIATION: HOW DOES WEATHER AFFECT THE SAFETY AND OPERATIONS OF AIRPORTS AND AVIATION, AND HOW DOES FAA WORK TO MANAGE WEATHER-RELATED EFFECTS?" 2003. URL `https://api.semanticscholar.org/CorpusID:108023423`.

[3] Fultz, A. J., and Ashley, W. S., "Fatal weather-related general aviation accidents in the United States," *Physical Geography*, Vol. 37, No. 5, 2016, pp. 291–312.

[4] Long, T., "Analysis of weather-related accident and incident data associated with Section 14 CFR Part 91 Operations," *The Collegiate Aviation Review International*, Vol. 40, No. 1, 2022.

[5] Fujita, T. T., and Caracena, F., "An analysis of three weather-related aircraft accidents," *Bulletin of the American Meteorological Society*, Vol. 58, No. 11, 1977, pp. 1164–1181.

[6] Ramee, C., Speirs, A., Payan, A. P., and Mavris, D., "Analysis of weather-related helicopter accidents and incidents in the United States," *AIAA Aviation 2021 Forum*, 2021, p. 2954.

[7] Ahlstrom, U., Racine, N., and Hallman, K., "Assessments of Flight and Weather Conditions during General Aviation Operations," Tech. Rep. DOT/FAA/TC-19/33, Federal Aviation Administration, William J. Hughes Technical Center, Atlantic City International Airport, NJ, 2019. Available from the Federal Aviation Administration William J. Hughes Technical Center: `https://actlibrary.tc.faa.gov`.

[8] Bouhsine, T., Idbraim, S., Bouaynaya, N. C., Alfergani, H., Ouadil, K. A., and Johnson, C. C., "Atmospheric Visibility Image-Based System for Instrument Meteorological Conditions Estimation: A Deep Learning Approach," *Proc. 2022 9th International Conference on Wireless Networks and Mobile Communications (WINCOM)*, Rabat, Morocco, 2022, pp. 1–6. [Online]. Available: https://doi.org/10.1109/WINCOM55661.2022.9966454.

[9] Ait Ouadil, K., Idbraim, S., Bouhsine, T., et al., "Atmospheric visibility estimation: a review of deep learning approach," *Multimedia Tools and Applications*, 2023. [Online]. Available: https://doi.org/10.1007/s11042-023-16855-z.

[10] Li, S., Fu, H., and Lo, W., "Meteorological Visibility Evaluation on Webcam Weather Image Using Deep Learning Features," 2017. https://doi.org/10.7763/IJCTE.2017.V9.1186.

[11] Chaabani, H., Werghi, N., Kamoun, F., Taha, B., Outay, F., and Yasar, A.-U.-H., "Estimating meteorological visibility range under foggy weather conditions: A deep learning approach," *Procedia Computer Science*, Vol. 141, 2018, pp. 478–483. https://doi.org/10.1016/j.procs.2018.10.139, URL https://www.sciencedirect.com/science/article/pii/S1877050918317885.

[12] Palvanov, A., and Im Cho, Y., “DHCNN for Visibility Estimation in Foggy Weather Conditions,” *2018 Joint 10th International Conference on Soft Computing and Intelligent Systems (SCIS) and 19th International Symposium on Advanced Intelligent Systems (ISIS)*, 2018, pp. 240–243. https://doi.org/10.1109/SCIS-ISIS.2018.00050.

[13] Choi, Y., Choe, H.-G., Choi, J. Y., Kim, K. T., Kim, J.-B., and Kim, N.-I., “Automatic Sea Fog Detection and Estimation of Visibility Distance on CCTV,” *Journal of Coastal Research*, , No. 85 (10085), 2018, pp. 881–885. https://doi.org/10.2112/SI85-177.1, URL https://doi.org/10.2112/SI85-177.1.

[14] You, Y., Lu, C., Wang, W., and Tang, C.-K., “Relative CNN-RNN: Learning Relative Atmospheric Visibility From Images,” *IEEE transactions on image processing: a publication of the IEEE Signal Processing Society*, Vol. 28, No. 1, 2019, pp. 45–55. https://doi.org/10.1109/TIP.2018.2857219.

[15] Li, Q., Tang, S., Peng, X., and Ma, Q., “A Method of Visibility Detection Based on the Transfer Learning,” *Journal of Atmospheric and Oceanic Technology*, Vol. 36, 2019. https://doi.org/10.1175/JTECH-D-19-0025.1.

[16] Outay, F., Taha, B., Chaabani, H., Kamoun, F., Werghi, N., and Yasar, A. U.-H., “Estimating ambient visibility in the presence of fog: a deep convolutional neural network approach,” *Personal and Ubiquitous Computing*, Vol. 25, No. 1, 2021, pp. 51–62. https://doi.org/10.1007/s00779-019-01334-w, URL https://doi.org/10.1007/s00779-019-01334-w.

[17] Liu, K., Li, Y., Xu, N., and Natarajan, P., “Learn to Combine Modalities in Multimodal Deep Learning,” , 2018. [Online]. Available: https://arxiv.org/abs/1805.11730.

[18] Castanedo, F., Ursino, D., and Takama, Y., “A Review of Data Fusion Techniques,” *The Scientific World Journal*, Vol. 2013, 2013, p. Article ID 704504. [Online]. Available: https://doi.org/10.1155/2013/704504.

[19] Molino-Minero-Re, E., Aguileta, A. A., Brena, R. F., and Garcia-Ceja, E., “Improved Accuracy in Predicting the Best Sensor Fusion Architecture for Multiple Domains,” *Sensors*, Vol. 21, No. 7007, 2021. [Online]. Available: https://doi.org/10.3390/s21217007.

[20] Blasch, E., Pham, T., Chong, C. Y., Koch, W., Leung, H., Braines, D., and Abdelzaher, T., “Machine Learning/Artificial Intelligence for Sensor Data Fusion-Opportunities and Challenges,” *IEEE Aerospace and Electronic Systems Magazine*, Vol. 36, No. 7, 2021, pp. 80–93. [Online]. Available: https://doi.org/10.1109/MAES.2020.3049030.

[21] Palvanov, A., and Cho, Y. I., “VisNet: Deep Convolutional Neural Networks for Forecasting Atmospheric Visibility,” *Sensors*, Vol. 19, No. 6, 2019, p. 1343. https://doi.org/10.3390/s19061343, URL https://www.mdpi.com/1424-8220/19/6/1343.

[22] Department of Computer Science, Chu Hai College of Higher Education, 80 Castle Peak Road, Castle Peak Bay, Tuen Mun, N.T. Hong Kong, Lo, W. L., Zhu, M., and Fu, H., “Meteorology Visibility Estimation by Using Multi-Support Vector Regression Method,” *Journal of Advances in Information Technology*, 2020, pp. 40–47. https://doi.org/10.12720/jait.11.2.40-47, URL http://www.jait.us/index.php?m=content&c=index&a=show&catid=198&id=1091.

[23] Li, J., Lo, W. L., Fu, H., and Chung, H. S. H., “A Transfer Learning Method for Meteorological Visibility Estimation Based on Feature Fusion Method,” *Applied Sciences*, Vol. 11, No. 3, 2021. https://doi.org/10.3390/app11030997, URL https://www.mdpi.com/2076-3417/11/3/997.

[24] Zhang, F., Yu, T., Li, Z., Wang, K., Chen, Y., Huang, Y., and Kuang, Q., “Deep Quantified Visibility Estimation for Traffic Image,” *Atmosphere*, Vol. 14, No. 1, 2023, p. 61. https://doi.org/10.3390/atmos14010061.

[25] You, J., Jia, S., Pei, X., and Yao, D., “DMRVisNet: Deep Multihead Regression Network for Pixel-Wise Visibility Estimation Under Foggy Weather,” *IEEE Transactions on Intelligent Transportation Systems*, Vol. 23, No. 11, 2022, pp. 22354–22366. https://doi.org/10.1109/TITS.2022.3180229.

[26] Chen, X.-H., and Li, Z., “Dark Channel Based Visibility Measuring from Daytime Scene Videos,” *Journal of Internet Technology*, Vol. 23, No. 3, 2022, pp. 593–599.

[27] Wauben, W., and Roth, M., “Exploration of Fog Detection and Visibility Estimation from Camera Images,” *WMO Technical Conference on Meteorological and Environmental Instruments and Methods of Observation, CIMO TECO*, 2016, pp. 1–14.

[28] Cheng, X., Liu, G., Hedman, A., Wang, K., and Li, H., “Expressway Visibility Estimation Based on Image Entropy and Piecewise Stationary Time Series Analysis,” , 2018. [Online]. Available: arXiv:1804.04601.

[29] Zhou, H., Dai, M., Shi, D., Meng, Y., Peng, B., and Chen, T., “Research on visibility detection model optimization based on dark channel prior and image entropy and visibility development trend prediction,” *IOP Conference Series: Earth and Environmental Science*, Vol. 826, 2021, p. 012031. https://doi.org/10.1088/1755-1315/826/1/012031.

[30] Huang, Y., Du, C., Xue, Z., Chen, X., Zhao, H., and Huang, L., “What Makes Multi-modal Learning Better than Single (Provably),” , 2021. URL https://arxiv.org/abs/2106.04538.

[31] Chen, X.-W., and Lin, X., “Big Data Deep Learning: Challenges and Perspectives,” *IEEE Access*, Vol. 2, 2014, pp. 514–525. https://doi.org/10.1109/ACCESS.2014.2325029.

[32] Liang, P. P., “Foundations of Multisensory Artificial Intelligence,” , 2024. URL https://arxiv.org/abs/2404.18976.

[33] Palvanov, A., and Cho, Y. I., “VisNet: Deep Convolutional Neural Networks for Forecasting Atmospheric Visibility,” *Sensors*, Vol. 19, No. 6, 2019, p. 1343. https://doi.org/10.3390/s19061343.

[34] Gui, J., Cong, X., Cao, Y., Ren, W., Zhang, J., Zhang, J., Cao, J., and Tao, D., “A comprehensive survey and taxonomy on single image dehazing based on deep learning,” *ACM Computing Surveys*, Vol. 55, No. 13s, 2023, pp. 1–37.

[35] Eftekhar, A., Sax, A., Bachmann, R., Malik, J., and Zamir, A., “Omnidata: A Scalable Pipeline for Making Multi-Task Mid-Level Vision Datasets from 3D Scans,” , 2021. [Online]. Available: https://arxiv.org/abs/2110.04994.

[36] Ranftl, R., Bochkovskiy, A., and Koltun, V., “Vision Transformers for Dense Prediction,” , 2021.

[37] Akkus, C., Chu, L., Djakovic, V., Jauch-Walser, S., Koch, P., Loss, G., Marquardt, C., Moldovan, M., Sauter, N., Schneider, M., Schulte, R., Urbanczyk, K., Goschenhofer, J., Heumann, C., Hvingelby, R., Schalk, D., and Aßenmacher, M., “Multimodal Deep Learning,” , 2023.

[38] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I., “Learning Transferable Visual Models From Natural Language Supervision,” , 2021.

[39] Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q. V., Sung, Y., Li, Z., and Duerig, T., “Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision,” , 2021.

[40] Huang, S.-C., Pareek, A., Seyyedi, S., Banerjee, I., and Lungren, M. P., “Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines,” *npj Digital Medicine*, Vol. 3, No. 1, 2020, p. 136. https://doi.org/10.1038/s41746-020-00341-z, URL https://doi.org/10.1038/s41746-020-00341-z.

[41] Zadeh, A., Chen, M., Poria, S., Cambria, E., and Morency, L.-P., “Tensor Fusion Network for Multimodal Sentiment Analysis,” , 2017. URL https://arxiv.org/abs/1707.07250.

[42] Liu, Z., Shen, Y., Lakshminarasimhan, V. B., Liang, P. P., Zadeh, A., and Morency, L.-P., “Efficient Low-rank Multimodal Fusion with Modality-Specific Factors,” , 2018. URL https://arxiv.org/abs/1806.00064.

[43] Nagrani, A., Yang, S., Arnab, A., Jansen, A., Schmid, C., and Sun, C., “Attention Bottlenecks for Multimodal Fusion,” *Advances in Neural Information Processing Systems*, Vol. 34, edited by M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Curran Associates, Inc., 2021, pp. 14200–14213. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/76ba9f564ebbc35b1014ac498fafadd0-Paper.pdf.

[44] Huang, G., Liu, Z., van der Maaten, L., and Weinberger, K. Q., “Densely Connected Convolutional Networks,” , Jan. 2018. https://doi.org/10.48550/arXiv.1608.06993, URL http://arxiv.org/abs/1608.06993.

[45] Bachmann, R., Mizrahi, D., Atanov, A., and Zamir, A., “MultiMAE: Multi-modal Multi-task Masked Autoencoders,” , 2022.

[46] Wang, X., Chen, G., Qian, G., Gao, P., Wei, X.-Y., Wang, Y., Tian, Y., and Gao, W., “Large-scale Multi-Modal Pre-trained Models: A Comprehensive Survey,” , 2024.