Multi-sensor Ensemble-guided Attention Network for Aerial Vehicle Perception Beyond Visible Spectrum

Alicja Kwasniewska
SiMa Technologies
226 Airport Parkway,
San Jose, CA 95110
alicja@sima.ai

Anastacia MacAllister, Rey Nicolas
General Atomics
14200 Kirkham Way,
Poway, CA 92064
anastacia.macallister@ga-asi.com
rey.nicolas@ga-asi.com

Javier Garza
Lockheed Martin
1 Lockheed Blvd,
Fort Worth TX 76108
javier.garza@lmco.com

Abstract

Researchers from different market domains have made significant developments in Artificial Intelligence (AI) enabling more advanced automated sensing systems and, thus, eliminating the need for the time-consuming manual analysis of data, which is prone to human errors. However, successful deployment of such systems in real world applications requires careful design and analysis of the proposed models. This work focuses on perception done on Unmanned Aerial Vehicles (UAV) using multi-task learning. There are multiple challenges when considering such platforms. First of all, they often operate in difficult and dynamic conditions affected by various factors, such as background noises, ego-noise of the motors and occluded views. At the same time, they require high performance local compute, co-designed with optimized software solutions that meet small size, weight, and power (SWaP) requirements. Therefore, the AI models designed for such systems should not introduce computational and memory overheads to allow for real time processing at the embedded edge. Taking this into account, this work proposes a novel neural network-based system that utilizes ensemble-guided modulations of audio path fused with the infrared (IR) visual embedding using the attention mechanism. The ensemble mechanism doesn't require spawning new ensemble members, but instead operates on FiLM (Feature-wise Linear Modulation) activation, making it suitable for resource-constraints embedded edge platforms. The performed experiments show that the proposed network outperforms a single FiLM network by 15% and is more robust to noise.

1. Introduction

Commercial industry and the Department of Defense (DoD) are investing billions of dollars into Artificial Intelligence (AI) and Machine Learning (ML) development. The market for AI/ML related technology is expected to grow thirty percent year over year to become a 1.3 trillion dollar market by 2030 [14,30]. A large part of this market is anticipated to focus on object detection and monitoring tasks such as aerial imagery analysis for disaster response [20, 40], wildfire detection [7, 33, 46], automatic target recognition (ATR) [16, 47], and self-driving cars [31]. AI/ML plays a pivotal role expanding this technology because of it's ability to help detect elements of interest without labor intensive human effort processing vast amounts of collected data. For example, its projected that there are more than 100 terabytes of imagery data collected every day by commercial organizations [26]. These large volumes of data are overwhelming analysts, meaning valuable information about disasters, troop movements, or environmental change are going undetected or ineffectively exploited. This exponentially increasing trove of data is directly driving growing interest in AI/ML since it will be required to ensure data is processed at the speed of relevance and decision making.

Often this type of AI/ML facilitated processing is done using deep machine learning techniques like convolutional neural networks (CNNs). These networks use imagery data and vast training data sets to detect objects or events of interest so action can be taken. However, imagery can only provide so much context and is limited by clear line of sight requirements, which often limits the technologies applicability in smoke or cloud occluded imagery or detecting targets in other occluded environments such as urban terrain. Consequently, AI/ML models can be relegated to addressing only elementary detection cases or non-ideal real world conditions might cause them to miss important events. However, as the market for AI/ML continues to expand and these models become integrated into critical systems, more robust means of detection need to be developed

to increase model resiliency [22, 27, 39]. One such way of increasing AI/ML detection resiliency involves fusing auxiliary pieces of information into the model besides imagery to provide added context. These deep learning fusion models have the potential to help provide more robust detection models, resilient to real world conditions. One such promising auxiliary source of information demonstrated in literature is audio. To date literature has shown the promise of this work, but has yet to address fundamental challenges such as data set drift that often occurs when deploying models. This is especially relevant for audio information since elements of the data such as background noise can change drastically based on the operational environment. For example, surveillance drones trained to detect troop movements in occluded urban environments using audio fusion could be sensitive to the background noise domain shift when deployed. As a result, additional development should focus on identifying how robust these fusion models are to data set drift.

Work in this paper begins to build more robust deep learning object detection models that are resilient to in field data shifts by using ensembles plus fused audio and video information. Contributions of this paper are: 1) verification of applicability of the drone data set with unsynchronized audio and visual signals to the task of IR and Audio Embedding Fusion in a transformer-based classification model; 2) improvement of the architecture with ensemble of vision-guided modulations of audio embeddings; 3) experiments are performed on noise introduced to audio, vision, and both modalities to verify robustness of the proposed ensemble-based solution. The code will be publicly available.

2. Related Work

Historically, detection of airborne platforms has been a function of early warning systems, such as the Airborne Warning and Control System (AWACS). These systems are installed on platforms such as the E-3 Sentry, a modified commercial airplane that includes a rotating radar dome. This allows the system to detect objects at a range of about 250 miles, and can be paired with an Identification Friend or Foe (IFF) system to differentiate between the objects detected [6].

While these systems have been useful and proven in the past, as there have been more advancements in sensor development, it is necessary to consider other potential modalities that could contribute to a more complete understanding of the operational space. Additionally, technological advancement in the field of unmanned aerial vehicles (UAVs) has led to more accessibility of such platforms to the general public, and potentially adversarial actors. Small UAVs can quickly become dangerous threats to engines of large platforms by causing Foreign Object Damage (FOD) incidents [21] or could be used for spying. Such potential threats need to be detectable for navigation and contingency planning. The ability to make sense of different types of data could provide better context.

Hengy et. al describes the usage of a heterogeneous sensor network composed of acoustic antennas, radar systems, and optical sensors to detect, localize, and classify UAVs. [10]. They found that each of these modalities complement each other with regard to obtaining accurate position information. Zhang et. al looked at using an end-to-end tracking framework for fusing the Red-Green-Blue (RGB) and Thermal Infrared (TIR) modalities [45]. They compared the performance of the fusion mechanism with single modality models, and found significant improvement. Hardejewicz, et. al describes the usage of fusion of non-coherent signals from independent radars working in different bands to detect small radar cross section (RCS) targets [2]. They compared the target detection probability of the fusion method with the target detection probability of single-band radars and found significant improvement. Diamantidou et. al proposed a general fusion neural network framework to merge features extracted from various single modality models and increase detection and classification accuracy [5]. They compared their approach with uni-modal approaches and found significant performance improvements. McCoy et. al developed an ensemble deep learning framework that includes hybrid synthetic and deep features for detection and classification of UAVs by combining acoustic, optical, and wireless radio frequency (RF) signals [24]. After performing experiments, they found that their proposed approach outperforms existing approaches for UAV detection.

However, the introduced ensemble solution used a combination of separate models, which leads to increased computational and memory overhead and is not aligned with SWaP requirements which are critical for UAVs. As such, the approach presented in this study introduces a multi-task learning solution with performance on par with ensembles, but without the need of introducing additional model members. The activations of the audio path are conditioned using the ensemble of IR features, which improves the robustness and accuracy of aerial vehicle classification. To the best of our knowledge such an approach has not been conducted in the attention-based sensor fusion networks before.

3. Methodology

In this section, the methodology used for multi-sensor aerial vehicle classification is described. At first, the data set generation and preparation approach is explained, followed by the detailed specification of the neural network used for the analysed task, including overview of the proposed ensemble-based vision-guided modulations of audio signals.

(a) (b) (c) (d)

Figure 1. Examples of IR frames used in the study, (a-b): drone, (c-d): helicopter

3.1. Data set generation

This work addresses the problem of limited classification capabilities and sensitivity to lighting conditions of visible spectrum-based systems by proposing the network for perception using IR and audio sensor fusion. The analysis is performed using the Drone Detection data set [35]. It contains 90 audio and 650 visual recordings (365 IR and 285 visible, ten seconds each), with a total of 203,328 manual annotations (done in the Matlab labeller application [34]) covering drone, helicopter, airplane, and bird categories. The IR camera used for data collection is the FLIR Breach PTQ-136 device with the Boson detector, capable of capturing frames with a resolution of 320×256 pixels in a Y16 16-bit grey scale format. The audio is recorded using the Boya BYMM1 mini cardioid directional microphone and contains the drone and helicopter sounds only. Since the visible spectrum sequences available in the data set are not used in this study, the details for the RGB sensor is omitted. The data was captured in 3 different locations using 3 drones varying in size and the offered performance. The farthest sensor to target distance for the recorded 'drone' sequences was 200m. In addition, the data has been extended with non-copyrighted material from YouTube to increase the number of samples in the helicopter and aircraft categories, a task difficult to capture by the authors.

The authors of the data set have also proposed a deep-learning based system for the aerial vehicle classification and detection task. However, the solution proposed along with the data set significantly differs from the network presented in this study, especially in the way the sensor fusion is implemented. The Drone Detection solution uses the weighted sum of classification scores produced by the YOLO detector [29] and the LSTM audio signal classificator [15], but all networks are run separately. Our approach performs a multi-modal temporal fusion of extracted embeddings in a single transformer block, what is described in detail in the next subsection.

The available data set was not designed for the joint fusion of visual and audio data. Originally, the audio recordings were provided independently of the frame sequences without the synchronization. Since the network proposed in this work uses a multi-modal temporal attention block for sensor fusion, the input data has to be represented by both modalities. However, post-recording synchronization of both inputs without information about data collection timestamp and location is not straightforward. Taking this into account, an interesting research question is whether the vision signal can be used for guiding audio embedding even though visual and auditory samples lack temporal synchronization and may not correspond to the same distance of the object to the sensor.

To verify this claim, a new data set was constructed from the available samples by merging available audio and IR samples and feeding them jointly into the proposed model. Given the differences in the number of visual and audio samples, some of the audio signals were merged with a few different image sequences. However, to ensure a fair evaluation process, the split between the train, validation, and test set was executed ensuring no overlap between audio recordings across these sets. The data set was cleaned from a few samples with no aerial vehicle in the view and was further augmented using random crop, rotation and flip. As mentioned before, the audio clips were covering only two categories (drone and helicopter), thus, the proposed network was designed as a binary classificator of these two classes. The remaining categories of IR sequences were skipped. Examples of the frames extracted from the final data set are presented in Fig. 1.

During the augmentation process, the up-sampling technique was also used to increase samples in the minority category (helicopter) and mitigate the class imbalance problem. The final data set consists of 224 drone and 214 helicopter 10-sec 30FPS recordings with both visual and auditory data in each sample (318, 60, and 60 in train, val, and test sets, respectively). There was no overlap in either visual or auditory information across sets.

3.2. Multi-sensor Aerial Vehicle Classification

Multi-level Attention Fusion Network (MAFNet) proposed by Brousmiche M. et al. [4] has been designed for audio-visual event recognition from the visible spectrum data. MAFNet has been proved to outperform models that

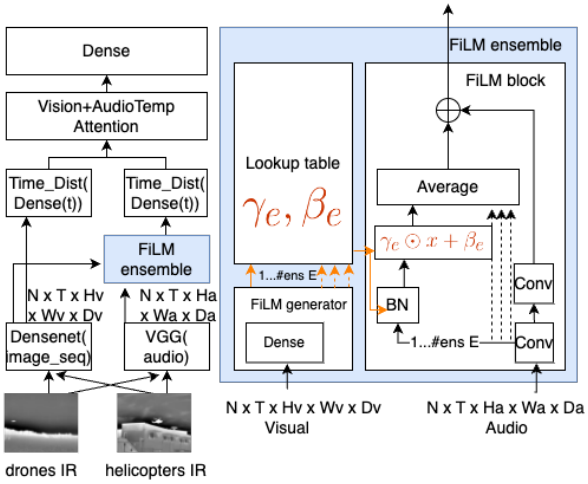


Figure 2. MENSA architecture extending MAFNet with ensemble-guided conditions of audio embeddings to improve its robustness to noisy data present in UAV platforms. Proposed block marked in blue - linear modulations of audio embeddings concatenated with visual features are fed to the multimodal temporal attention block, modulation parameters of the introduced FiLM ensemble are marked in orange.

rely on the visual input only in majority of the analysed scenarios. Taking into account the successful results of this network, the solution presented in this study is built on top of this architecture, improving its performance and robustness to the higher level of noise present in the Unmanned Aerial Vehicle platforms [23] by guiding the audio modality with ensemble of linear modulations instead of a single FiLM layer [28]. The proposed model is called MENSA - Multi-sensor Ensemble-guided Attention Network.

3.2.1 High level network overview

Fig.2 presents the overview of the MENSA architecture. Following [37], visual and audio embeddings are extracted from t non-overlapping 1-sec segments extracted from each input video using pre-trained MAFNet convolutional feature extractors. Thus, each video is represented by two representations of a size height H, by width W, by depth D: R_a audio and R_v vision, where $R_a = \{F_1^a ... F_T^a\}, F_t^a \in \mathbb{R}^{W_a x H_a x D_a}$ and $R_v = \{F_1^v ... F_T^v\}, F_t^v \in \mathbb{R}^{W_v x H_v x D_v}$. After that, the visual embedding is used to condition the audio representation by giving varying importance to different audio features. Contrary to the previous work [4], we do not use a single lateral connection between R_a and R_v, but rather introduce an ensemble of E linear modulations [38], allowing to capture more complex audio-vision dependencies. Audio signal features weighted by the γ and β parameters calculates as the function of the R_v are averaged and passed to the subsequent network blocks simultaneously with the R_v. These blocks ($Time\_Dist$) apply the linear mapping to every temporal slice of an input. The calculated features are reduced using another linear mapping layer to allow for adjusting feature selection during the model training, contrary to the MAFNet network that uses average pooling. To avoid overfitting of such topology with more parameters, the dropout is applied directly after the dense layers [32]. Next, the audio and visual features are concatenated and fed to the modality and temporal attention module. As analysed in previous studies, other more complex fusion methods have already been proposed, such as, multi-modal compact bilinear pooling (MCB) [8] or the multi-modal residual fusion (DMR) [37], but they have been proven to lead to a decline in performance in similar audio vision fusion classification tasks [4].

The modality and temporal attention module learns the attention scores for vision α_T^v, and audio α_T^a and weight both feature maps to generate the final fused representation of the input data. Finally, the constructed representation is fed to the fully connected layer followed by the softmax activation to obtain confidence scores for each class.

3.2.2 Ensemble-guided conditioning of audio

The Ensemble-guided conditioning of audio is explained first by introducing the concept of the Feature-wise Linear Modulation (FiLM) layer, proposed in [3].

The FiLM layer introduces a lateral connection between inputs from different sensor types and thus improves the prediction performance compared to the simple fusion [4]. Formally, the FiLM layer can be used to condition one path with parameters γ_m and β_m learnt as functions f and h of features extracted from the other path, where in our case m corresponds to either v vision or a audio. Thus, parameters used to condition audio γ_a and β_a, are computed as:

$$\gamma_a = f(R_v), \quad \beta_a = h(R_v) \tag{1}$$

Parameters used to condition vision (γ_v and β_v) would be computed is an equivalent way using the audio representation R_a. However, following [3], it has been shown that the best results are achieved when visual representation conditions the audio path. Thus, only the lateral connection from vision to audio is utilized in this work, and the connection in the opposite direction is not used. Therefore, for simplicity a sub-index is omitted from notations thereafter. Functions f and h are implemented as fully connected layers in our work to manipulate audio feature maps according to the visual input with the affine transformations. The output from the FiLM layer is defined as the Hadamard (element-wise) product:

$$FiLM(R_a|\gamma_a, \beta_a) = \gamma_a \odot R_a + \beta_a \tag{2}$$

This work also exploits a possible interaction between different sensor types, but goes one step further and introduces the FiLM Ensemble layer [38] for modulation of one modality with features from the other. The FiLM Ensemble [38] was proposed following the original FiLM implementation [3] to extend the idea of sensors decoupling due to the problem of uncertainty quantification. We argue that it's equally important in the multi-task learning used in the real world perception tasks on UAV platforms. Overconfident and uncalibrated prediction models are not able to behave robustly in the presence of noise and data imperfection. However, a simple ensemble of models lead to the computational overhead and increased memory utilization [41], which is not suitable for embedded edge platforms [18], especially UAVs for which low size, weight, and power (SWaP) requirements are crucial [17]. Taking this into account, the FiLM ensemble is introduced as the modulation of the network activations only. Specifically, the equation 3 is extended to compute E number of parameters:

$$\gamma_e = f_e(R_v), \quad \beta_e = h_e(R_v) \tag{3}$$

where $e \in \{1, ..., E\}$ denotes the number of ensemble.

Then, the FiLM layer is calculated per each ensemble:

$$FiLM(R_a|\gamma_e, \beta_e) = \gamma_e \odot R_a + \beta_e \tag{4}$$

and the final output of the introduced FiLM ensemble block is computed as:

$$\frac{1}{E}\sum_{e=1}^{E} y_e \tag{5}$$

where y_e is calculated during training for each ensemble member as $y_e = f_{\theta,\gamma_e,\beta_e}(\mathrm{R}_v), alltrainableparameters\theta$ are shared across ensembles except the computed FiLM parameters. All parameters are optimized together to optimize the loss function of the classification model. The loss function used in the study is the cross entropy loss.

By introducing such ensembles, we can obtain calibrated estimations of the model prediction uncertainties, which improves robustness to noise. Experimental analysis of this claim is presented in the following section.

3.2.3 Training techniques

As in [4], each clip is split into $T = 10$ segments. We also reuse ImageNet pretrained model DenseNet [12] for extraction of visual embeddings (size $7x7x1920$) without additional tuning of the model and the Audio set pretrained VGG model [11] for extraction of audio embeddings (size $12x8x512$) without additional tuning as well.

The number of filters in residuals and fully connected layers is 512, followed by the batch normalization with parameters calculated as a moving average. The network is trained with cross-entropy loss and adam optimizer. The learning rate parameter is selected using the grid search along with the number of ensemble, described in detail in the experimental analysis section. In addition to the standard training procedure, we also exploit additional well-known techniques that prevent overfitting, such as learning rate decay (0.9 each step), L1, L2 regularization (0.01) and dropout following dense layers (0.2). The training exit criteria is defined using the early stopping based on the validation accuracy is done with patience set to 50-100 epochs (searchable parameter).

In multi-modal training, both modalities have different speed of convergence. The visual path backpropagation step is randomly dropped, allowing the audio path to train longer [43].

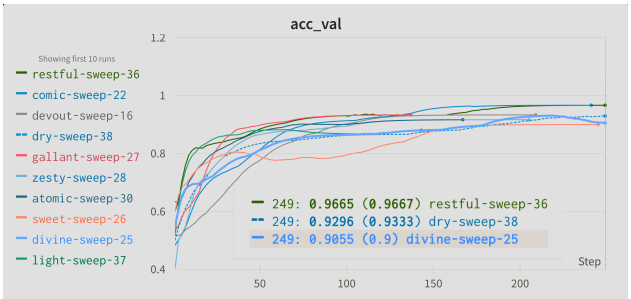


Figure 3. Validation accuracy for top 10 best runs, number at the end of the sweep name corresponds to the ensemble count.

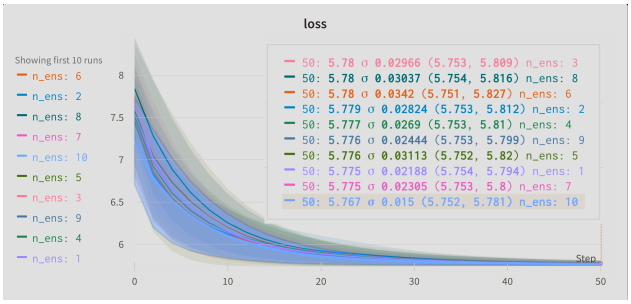


Figure 4. Cross Entropy Loss grouped by ensemble

4. Experimental results

The proposed network is evaluated with the publicly available Drone Detection data set, and pre-processed for the binary classification task addressed in this study. All hyperparameters, except the ones analyzed in the given experiment, remain unchanged. The model is trained on NVIDIA A100 GPU with 64GB memory using the Keras framework [9]. Grid search was performed using the sweep agent feature of the Weights and Biases framework [1].

Configuration, 10 best	Test accuracy, top1 [%]
$ens = 6, \quad lr = 9.00e-05$	96.67
$ens = 10, \quad lr = 5.00e-05$	93.33
$ens = 8, \quad lr = 9.00e-05$	94.99
$ens = 7, \quad lr = 5.00e-05$	94.99
$ens = 2, \quad lr = 5.00e-05$	93.33
$ens = 10, \quad lr = 7.00e-06$	91.66
$ens = 7, \quad lr = 9.00e-05$	91.66
$ens = 6, \quad lr = 5.00e-05$	91.66
$ens = 4, \quad lr = 9.00e-05$	91.66
$ens = 9, \quad lr = 5.00e-05$	89.99
$ens = 1, \quad lr = 5.00e-05$	80.00

Table 1. Accuracy achieved for different hyperparameter configurations. Proposed ensemble-based model outperforms the network with a single FiLM layer (ens=1 - original MAFNet model).

4.1. Experiments with ensemble count

The proposed MENSA model was trained with varying numbers of the ensembles and included multiple lateral connections and modulating audio features to validate the analysis. The increased number of ensembles also increased the number of network parameters. The learning rate was configurable during the search to allow for a slower optimization process and eliminate the risk of overfitting. Experimental analysis focused on selecting the best hyperparameters from the ensemble count in the range of $\{1 : 10 : 1\}$ and learning rate from the set of the following values: $\{1.00e-05, 3.00e-05, 5.00e-05, 7.00e-05, 9.00e-05\}$. Table 1 presents the 10 best configurations leading to the highest top1 accuracy achieved on the test set. The plot in Fig. 3 captures similar information but was obtained during training on the validation set. The loss function grouped by the ensemble is shown in Fig. 4. In all cases, the final loss is very close across different ensemble numbers ($mean = 5.7 \mp 0.03$). Test accuracy values show significant differences, proving the robustness of the proposed ensembles.

Fig. 5 shows the relationship between the test accuracy (top1 [%]) averaged across all runs for a given ensemble number and the ensemble count. Based on the achieved results, the increased number of ensembles can result in up to 10% performance gain when using 6-7 branches instead of the single FiLM layer as proposed in previous studies. However, the number of ensembles has to be carefully selected for a given problem since the large number of ensembles can introduce too many parameters. This creates a difficult convergence problem, and thus decreased accuracy (ensemble #10).

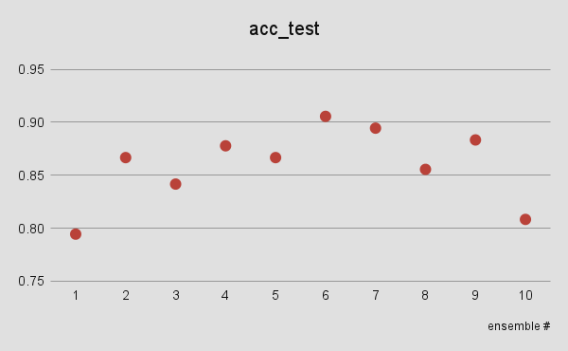
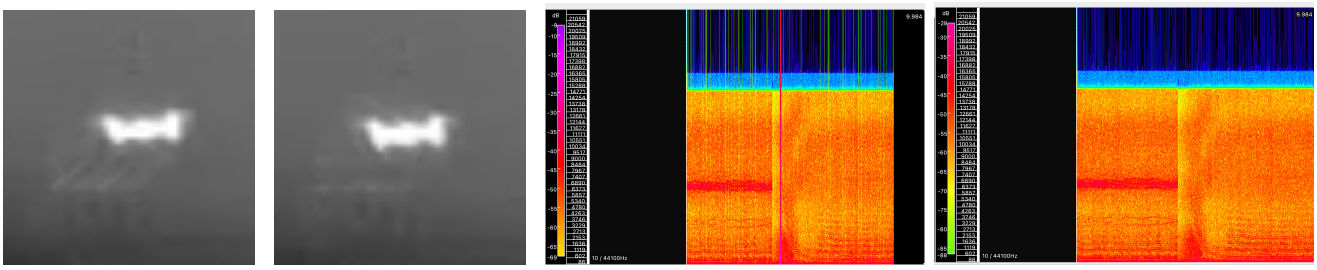


Figure 5. Average test accuracy vs number of ensemble

4.2. Experiments with noisy data

As proven in existing studies, one of the advantages of multiple classifier systems is an increased robustness to noise and other imperfections in data [25]. This feature is particularly important in the UAV domain addressed in this work since the ego-noise is a very common problem in such platforms [13]. This potentially leads to greater difficulties in the detection of other aerial vehicles. To verify whether the introduced ensemble-guided modulation of embeddings helps with mitigating impact of the noise on the classification capabilities, the sequences in the test set were modified and used for analysis of robustness of the MENSA to noisy data. Specifically, the ffmpeg tool was used to introduce a random noise to: a) audio channel, b) visual representation, c) both. In all scenarios the level of the noise was randomly selected from the range of 2500-5000 with the variable frequency of bytes in the packet being modified. Examples of the corrupted image and audio data are presented in Fig. 6. All model variants, previously trained on non-corrupted data, were tested using the test set with introduced noise. Table 2 shows the deviation in the test accuracy from the result obtained on the original test set for each analysed scenario (the smaller the value, the better, with the best result marked in blue and the second best marked in orange). It can be seen that all models, including ensemble 1, are much less sensitive to the audio noise. Yet, ensemble 2 allowed for a complete recovery of the performance, while ensemble 10 led to the biggest accuracy degradation. In case of the visual noise, the findings are much more interesting. Ensemble 1 (with a single FiLM layer as proposed in previous studies) resulted in more than a 30% accuracy drop. At the same time, ensemble 6 was almost completely prone to the introduced noise, leading to only 4% decrease in the accuracy. Similarly, for the combined noise, almost all ensembles were better than the single FiLM layer, proving the robustness of the proposed solution to noise and data imperfections, which are much more likely to occur in real-life scenarios.



(a) Image - random noise (b) Original image (c) Spectrogram - random noise (d) Spectrogram - original recording

Figure 6. Examples of data with introduced noise.

	Test accuracy decrease [%]		
Ensemble	Audio noise	Vision noise	Audio & Vision
MAFNet	0.69	31.72	26.57
2	0.00	23.72	15.38
3	0.63	19.44	22.77
4	0.63	11.17	20.25
5	0.96	4.17	9.61
6	0.61	10.15	13.49
7	1.86	16.62	21.73
8	0.64	20.75	16.23
9	0.62	19.41	24.52
10	2.57	9.83	15.46

Table 2. Decrease of the test accuracy achieved on noisy data compared to the results achieved on non-noisy data (the smaller the better - meaning the model is more robust to the distortion). Proposed ensemble-based models (ensemble 2 and more) outperforms the network with a single FiLM layer. Blue - the best, orange - the second best per each noise type.

5. Discussion

The presented study proposed the extended version of the multi-modal attention-based fusion network used for combining visual and auditory inputs in the aerial vehicle classification task. Specifically, the modulation of audio path guided by visual representation was done using the ensembles updating the activations of the FiLM block, instead of initializing separate ensemble members. This increases performance without significant computational cost or memory demand overheads.

Performed experiments proved that the introduced approach leads to better accuracy (15% gain) in aerial vehicle classification task on the publicly available multi-task learning Drone Detection data set even though both paths were lacking temporal synchronization. However, in-depth analysis showed that the number of ensembles leading to the increased accuracy saturates at some point, causing the accuracy to drop again after too many members are introduced. Thus, a careful evaluation is needed on a per case basis to identify the best configuration. This can be achieved using the sweep-based search of hyper-parameters, as accomplished in this work.

As discussed earlier, various noise sources often affect data collected on UAV platforms, which were considered in this study, thus, standard, usually uncalibrated neural networks may not be sufficient in creating robust predictions. Ensemble-based approach helps to mitigate this problem, but does not meet important SWaP requirements. Due to the characteristics and design of the ensemble method introduced in this study, the memory and computational overheads have not increased, which has a practical value for real world applications on UAV platforms. As shown in the performed experiments, the proposed ensemble-based MENSA model help with restoring classification accuracy on noisy data compared to the single FiLM-based network. This was particularly important in case of the vision noise, where accuracy may drop by as much as 30% if no ensemble is used.

Although preliminary results are satisfactory, further evaluation of the introduced technique is needed to prove its applicability and generalization capabilities. In future work, MENSA will be evaluated on other data sets. Since the number of publicly available multi-task aerial databases is limited, we may have to collect such samples ourselves or use synthetic data generators, e.g., diffusion models [44]. Similar experiments will be performed using different backbones for the feature extraction step of the pipeline. Additional modification of the network may lead to even better performance since heat flow in objects detected in thermal imagery can be more blurred than visible light spectrum images. As shown in the literature, increase of the receptive field, when dealing with IR data, improves accuracy of classification, object detection, and super resolution models [19]. We will explore such modification in future work, along with models used for improving quality of thermal data, e.g., denoising or deblurring [36]. Finally, the work will be also compared with single-modality solutions, e.g., object detection from visual data only, without the audio signal [42].

6. Conclusion

Multi-sensor models with ensemble modulations increase accuracy by at least 15% and are better at addressing noisy data, key variables of effective implementation of machine learning in any platform. Overlaying data collected from different types of sensors with the ability to reduce noise and excess data yields more precise situational awareness and decision making for operators and analysts. It will also extend dwell time over target areas of interest as operators will no longer have to waste additional time collecting over larger swaths of areas that are not germane to their intended collection efforts. This layering and effective noise reduction allows for greater speed and precision when searching for the proverbial needle in the haystack and will allow for a more complete and informed target picture for users to exploit.

References

[1] Lukas Biewald. Experiment tracking with weights and biases, 2020. Software available from wandb.com. 5

[2] T Brenner, J Hardejewicz, M Rupniewski, and M Nalecz. Signals and data fusion in a two-band radar. In *2012 13th International Radar Symposium*, pages 15–18. IEEE, 2012. 2

[3] Mathilde Brousmiche, Jean Rouat, and Stéphane Dupont. Audio-visual fusion and conditioning with neural networks for event recognition. In *2019 IEEE 29th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6. IEEE, 2019. 4, 5

[4] Mathilde Brousmiche, Jean Rouat, and Stéphane Dupont. Multi-level attention fusion network for audio-visual event recognition. *arXiv preprint arXiv:2106.06736*, 2021. 3, 4, 5

[5] Eleni Diamantidou, Antonios Lalas, Konstantinos Votis, and Dimitrios Tzovaras. Multimodal deep learning framework for enhanced accuracy of uav detection. In *Computer Vision Systems: 12th International Conference, ICVS 2019, Thessaloniki, Greece, September 23–25, 2019, Proceedings 12*, pages 768–777. Springer, 2019. 2

[6] US Air Force. E-3 sentry (awacs). *USAF Fact Sheet*, pages 92–59, 2015. 2

[7] Aaishwarya Gaikwad, Nishi Bhuta, Tejas Jadhav, Param Jangale, and Swati Shinde. A review on forest fire prediction techniques. In *2022 6th International Conference On Computing, Communication, Control And Automation (ICCUBEA*, pages 1–5, 2022. 1

[8] Yang Gao, Oscar Beijbom, Ning Zhang, and Trevor Darrell. Compact bilinear pooling. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 317–326, 2016. 4

[9] Antonio Gulli and Sujit Pal. *Deep learning with Keras*. Packt Publishing Ltd, 2017. 5

[10] Sebastien Hengy, Martin Laurenzis, Stéphane Schertzer, Alexander Hommes, Franck Kloeppel, Alex Shoykhetbrod, Thomas Geibig, Winfried Johannes, Oussama Rassy, and Frank Christnacher. Multimodal uav detection: Study of various intrusion scenarios. In *Electro-Optical Remote Sensing XI*, volume 10434, pages 203–212. SPIE, 2017. 2

[11] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. Cnn architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 131–135. IEEE, 2017. 5

[12] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 5

[13] Xabier Insausti, Bjørn O Hogstad, and Matthias Pätzold. Modelling and simulation of ego-noise of unmanned aerial vehicles. In *2020 IEEE 91st Vehicular Technology Conference (VTC2020-Spring)*, pages 1–5. IEEE, 2020. 6

[14] Fortune Business Insights. The artificial intelligence market. 2022. 1

[15] Sungho Jeon, Jong-Woo Shin, Young-Jun Lee, Woong-Hee Kim, YoungHyoun Kwon, and Hae-Yong Yang. Empirical study of drone sound detection in real-life environment with deep neural networks. In *2017 25th European Signal Processing Conference (EUSIPCO)*, pages 1858–1862. IEEE, 2017. 3

[16] Odysseas Kechagias-Stamatis and Nabil Aouf. Automatic target recognition on synthetic aperture radar imagery: A survey. *IEEE Aerospace and Electronic Systems Magazine*, 36(3):56–81, 2021. 1

[17] Alicja Kwasniewska, Onkar Chougule, Sneha Kondur, Sairam Alavuru, Rey Nicolas, David Gamba, Harsha Gupta, Dennis Chen, and Anastacia MacAllister. Ai-based rotation aware detection of aircraft and identification of key features for collision avoidance systems (sae paper 2022-01-0036). Technical report, SAE Technical Paper, 2022. 5

[18] Alicja Kwasniewska, Sharath Raghava, Carlos Davila, Mikael Sevenier, David Gamba, and Jacek Ruminski. Preferred benchmarking criteria for systematic taxonomy of embedded platforms (step) in human system interaction systems. In *2022 15th International Conference on Human System Interaction (HSI)*, pages 1–7. IEEE, 2022. 5

[19] Alicja Kwasniewska, Maciej Szankin, Jacek Ruminski, Anthony Sarah, and David Gamba. Improving accuracy of respiratory rate estimation by restoring high resolution features with transformers and recursive convolutional models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3857–3867, 2021. 7

[20] Jen D Lambert, Michelle Rose, Jeremy Ratcliff, Megan O'Connor, Tara Kedia, Sophia Oluic, Jeff Freeman, and Kaitlin Lovett. Ready for the next storm: Ai-enabled situational awareness in disaster response. Technical report, Johns Hopkins University Applied Physics Laboratory, 2021. 1

[21] Hu Liu, Mohd Hasrizam Che Man, and Kin Huat Low. Uav airborne collision to manned aircraft engine: Damage of fan blades and resultant thrust loss. *Aerospace Science and Technology*, 113:106645, 2021. 2

[22] Abdulrahman Mahmoud, Neeraj Aggarwal, Alex Nobbe, Jose Rodrigo Sanchez Vicarte, Sarita V Adve, Christo-

pher W Fletcher, Iuri Frosio, and Siva Kumar Sastry Hari. Pytorchfi: A runtime perturbation tool for dnns. In *2020 50th Annual IEEE/IFIP International Conference on Dependable Systems and Networks Workshops (DSN-W)*, pages 25–31. IEEE, 2020. 2

[23] Jose Martinez-Carranza and Caleb Rascon. A review on auditory perception for unmanned aerial vehicles. *Sensors*, 20(24):7276, 2020. 4

[24] James McCoy, Atul Rawal, Danda B Rawat, and Brian M Sadler. Ensemble deep learning for sustainable multimodal uav classification. *IEEE Transactions on Intelligent Transportation Systems*, 2022. 2

[25] Prem Melville, Nishit Shah, Lilyana Mihalkova, Raymond J Mooney, et al. Experiments on ensembles with missing and noisy data. *Multiple Classifier Systems*, 3077:293–302, 2004. 6

[26] Doug Mohney. Terabytes from space: Satellite imaging is filling data centers, Apr 2020. 1

[27] Kemal Oksuz, Baris Can Cam, Sinan Kalkan, and Emre Akbas. Imbalance problems in object detection: A review. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3388–3415, 2020. 2

[28] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. 4

[29] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. 3

[30] Zion Market Research. 422.37+ billion global artificial intelligence (ai) market size likely to grow at 39.4 percent cagr during 2022-2028. 2022. 1

[31] Teena Sharma, Benoit Debaque, Nicolas Duclos, Abdellah Chehri, Bruno Kinder, and Paul Fortier. Deep learning-based object detection and scene perception under bad weather conditions. *Electronics*, 11(4):563, 2022. 1

[32] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014. 4

[33] Ziheng Sun, Laura Sandoval, Robert Crystal-Ornelas, S Mostafa Mousavi, Jinbo Wang, Cindy Lin, Nicoleta Cristea, Daniel Tong, Wendy Hawley Carande, Xiaogang Ma, et al. A review of earth artificial intelligence. *Computers & Geosciences*, page 105034, 2022. 1

[34] Fredrik Svanström. Drone detection and classification using machine learning and sensor fusion, 2020. 3

[35] Fredrik Svanström, Cristofer Englund, and Fernando Alonso-Fernandez. Real-time drone detection and tracking with visible, thermal and acoustic sensors. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 7265–7272. IEEE, 2021. 3

[36] Maciej Szankin, Alicja Kwasniewska, and Jacek Ruminski. Influence of thermal imagery resolution on accuracy of deep learning based face recognition. In *2019 12th International Conference on Human System Interaction (HSI)*, pages 1–6. IEEE, 2019. 7

[37] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 247–263, 2018. 4

[38] Mehmet Ozgur Turkoglu, Alexander Becker, Hüseyin Anil Gündüz, Mina Rezaei, Bernd Bischl, Rodrigo Caye Daudt, Stefano D'Aronco, Jan Dirk Wegner, and Konrad Schindler. Film-ensemble: Probabilistic deep learning via feature-wise linear modulation. *arXiv preprint arXiv:2206.00050*, 2022. 4, 5

[39] Wenguan Wang, Qiuxia Lai, Huazhu Fu, Jianbing Shen, Haibin Ling, and Ruigang Yang. Salient object detection in the deep learning era: An in-depth survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(6):3239–3259, 2021. 2

[40] Jie Wei, Erik Blasch, Erika Ardiles-Cruz, Philip Morrone, and Alex Aved. Deep learning approach for data and computing efficient disaster mitigation in humanitarian assistance and disaster response applications. In *2022 IEEE International Humanitarian Technology Conference (IHTC)*, pages 79–85. IEEE, 2022. 1

[41] Matthew Whitehead and Larry S Yaeger. Multi-k machine learning ensembles. In *Midwest Artificial Intelligence and Cognitive Science Conference*, page 166, 2012. 5

[42] Xin Wu, Wei Li, Danfeng Hong, Ran Tao, and Qian Du. Deep learning for unmanned aerial vehicle-based object detection and tracking: A survey. *IEEE Geoscience and Remote Sensing Magazine*, 10(1):91–124, 2021. 7

[43] Fanyi Xiao, Yong Jae Lee, Kristen Grauman, Jitendra Malik, and Christoph Feichtenhofer. Audiovisual slowfast networks for video recognition. *arXiv preprint arXiv:2001.08740*, 2020. 5

[44] Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Yingxia Shao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models: A comprehensive survey of methods and applications. *arXiv preprint arXiv:2209.00796*, 2022. 7

[45] Lichao Zhang, Martin Danelljan, Abel Gonzalez-Garcia, Joost Van De Weijer, and Fahad Shahbaz Khan. Multimodal fusion for end-to-end rgb-t tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 2

[46] Yi Zhao, Jiale Ma, Xiaohui Li, and Jie Zhang. Saliency detection and deep learning-based wildfire identification in uav imagery. *Sensors*, 18(3):712, 2018. 1

[47] Zhong-Qiu Zhao, Peng Zheng, Shou-tao Xu, and Xindong Wu. Object detection with deep learning: A review. *IEEE transactions on neural networks and learning systems*, 30(11):3212–3232, 2019. 1