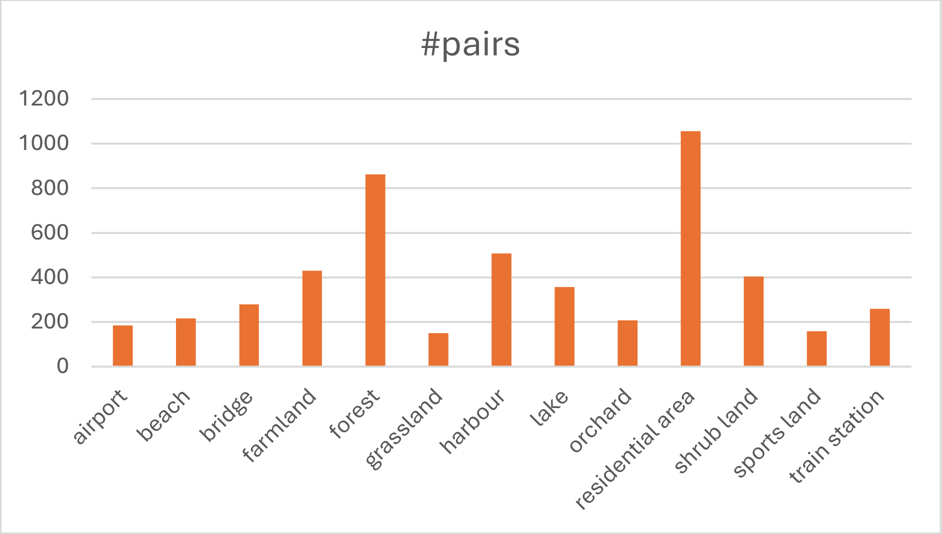
Dataset

AuDio Visual Aerial sceNe reCognition datasEt (ADVANCE)

The AuDio Visual Aerial sceNe reCognition datasEt (ADVANCE) is a comprehensive multimodal dataset specifically designed for aerial scene classification tasks. It combines both visual and audio data to enhance the recognition of diverse environmental scenes from an aerial perspective. The dataset comprises 5,075 paired samples, each consisting of an aerial image and its corresponding environmental audio recording. These pairs are categorized into 13 distinct classes capturing the unique auditory and visual characteristics of various locations such as airports, beaches, urban areas, forests, and more. Fig 1 shows the number of paired samples per class.



Methodology

ImageBind: A Multimodal Embedding Framework

ImageBind is a state-of-the-art multimodal model developed to generate unified embeddings from various data types, including images, audio, and text. The model achieves cross-modal alignment by leveraging modality-specific encoders, which map inputs into a shared embedding space by using a contrastive loss function. The goal of the contrastive loss is to minimize the distance between embeddings of semantically similar data points across different modalities while maximizing the distance between embeddings of dissimilar data points. This alignment allows ImageBind to effectively capture semantic similarities across different modalities, making it a powerful tool for retrieval and classification tasks.

Embedding Extraction

In our research, we employ the pretrained ImageBind model to extract embeddings from images and their corresponding audio counterparts. For each aerial scene, the corresponding image and audio data are processed through the ImageBind model to obtain 1024-dimensional embeddings. These embeddings serve as the foundational representations for our classification task. By utilizing ImageBind, we benefit from its robust ability to handle diverse data types, leading to more accurate and semantically meaningful results.

Multimodal approach

This study introduces a Transformer-based model leveraging a Dual Attention mechanism for the classification of aerial scenes using both image and audio data. The following sections outline the design and implementation of our model, which combines self-attention on individual modalities with cross-modal attention to integrate information across modalities.

1. Dual Attention Mechanism

The core of our approach is the Dual Attention mechanism, which processes image and audio embeddings in parallel and integrates them through a cross-modal attention operation. This mechanism enhances the model's ability to capture both intra-modal and inter-modal dependencies.

Architecture:

- **Self-Attention on Image and Audio Embeddings:**
    - The image and audio embeddings are first processed independently using separate Multihead Attention layers. These layers consist of 8 heads, where each head computes attention scores between the elements of the respective embeddings, allowing the model to capture relationships within each modality.
    - The output of each Multihead Attention operation is followed by a Layer Normalization step and a residual connection, which helps in maintaining the original embedding information while also incorporating the attention-derived context.
- **Cross-Modal Attention:**
    - After processing the embeddings independently, the model applies a Cross-Modal Attention mechanism. The image embeddings are used as queries, while the audio embeddings serve as keys and values. This allows the model to integrate complementary information from both modalities.
    - A Layer Normalization step and a residual connection are again applied to ensure stable learning and preserve important information from the original image embeddings.

The output of the Dual Attention mechanism is a set of combined embeddings that encapsulate both intra-modal and inter-modal interactions.

2. Transformer Classifier

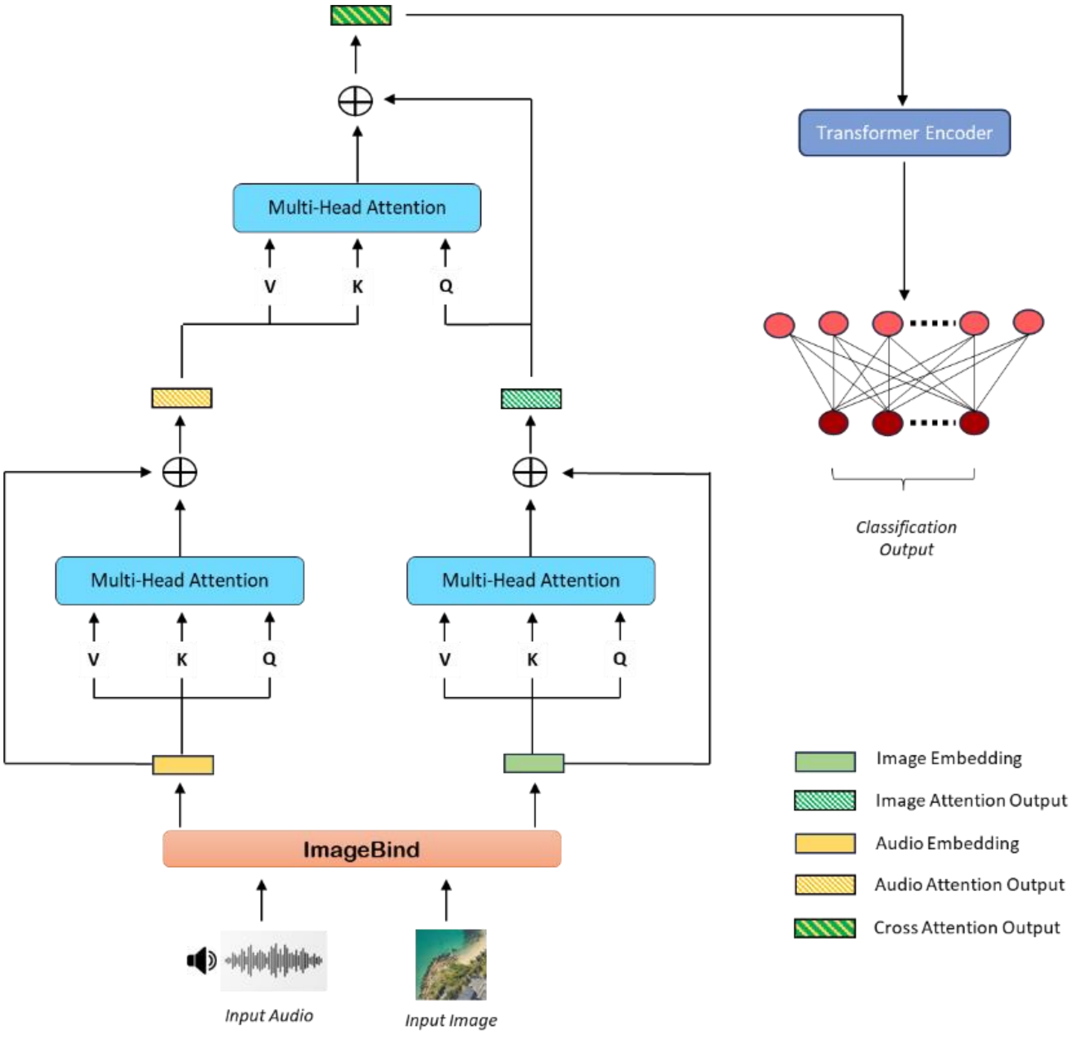
Following the Dual Attention mechanism, the model employs a Transformer-based architecture to further process and classify the combined embeddings.

Architecture:

- **Transformer Encoder:**
    - The combined embeddings are passed through a stack of 2 Transformer Encoder layers. Each encoder layer includes a Multihead Attention mechanism with 8 heads and a feedforward neural network.

  - The Multihead Attention within the encoder layers processes the combined embeddings to capture more complex dependencies across the entire input sequence.
  - The feedforward network, with a hidden dimension of 2048, applies dropout of 40% to prevent overfitting.
- **Global Average Pooling:**
  - The output of the Transformer Encoder is reduced to a fixed-size vector through global average pooling. This operation computes the mean of the sequence elements, yielding a global context vector that summarizes the combined embeddings.
- **Fully Connected Layer:**
  - The resulting global context vector is passed through a fully connected layer. The number of output units corresponds to the number of classes in the classification task.

Fig 2 shows the detailed architecture



Experiments

Implementation Details

The entire pipeline, including data preparation, model training, and evaluation, is implemented in Python using the PyTorch framework. The model is trained using the cross-entropy loss function, with AdamW as the optimizer, a learning rate of 1 x 10^-6, and a weight decay of 1 x 10^-5 to prevent overfitting. Training is conducted over 5 epochs, with the average loss per epoch calculated to monitor progress.

Aerial Scene Recognition

Approach	Precision	Recall	F1 score
ADVANCE	75.25	74.79	74.58
SoundingEarth	89.59	89.52	89.50
TFAVCNet	89.90	89.85	89.83
Ours	**95.39**	**95.30**	**95.02**