# Reflection Essay

## Introduction

Our project is based on few-shot viewpoint estimation which plays an important role in 3D scene understanding. Inspired by the paper *Few-Shot Object Detection and Viewpoint Estimation for Objects in the Wild* by Xiao et al. In our proposal we had proposed three modifications for our project.

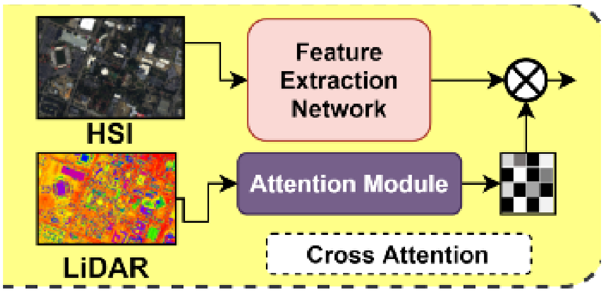
- Improving the way in which appearance-related features and geometric information are combined.
- Modifying the loss function.
- Trying a different meta-learning strategy based on MAML instead of fine-tuning based training for few-shot learning of viewpoint estimation.

## Project timeline:

### Improving Feature Aggregation

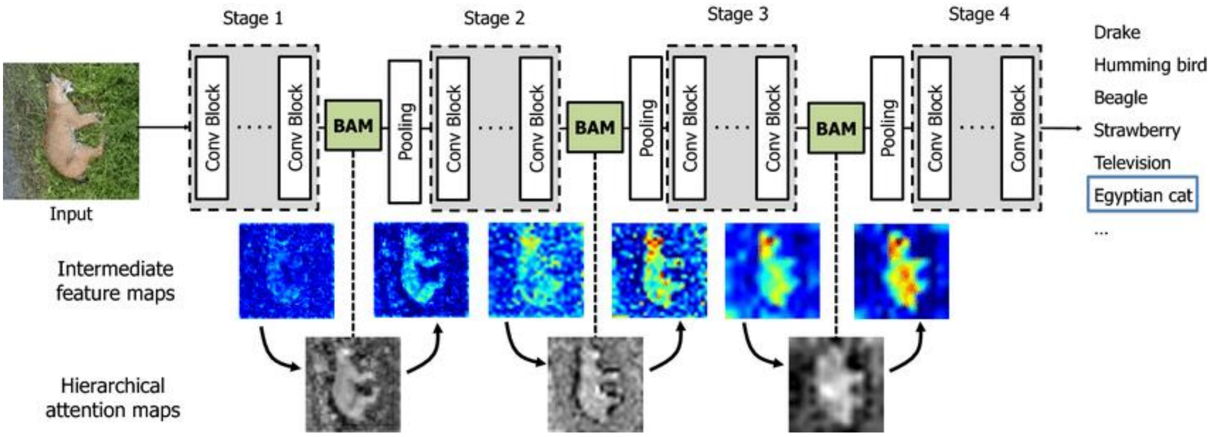
To improve the feature aggregation mechanism we had referred to the following paper: (http://repository.essex.ac.uk/31761/7/Chen-Gu2021_Article_CSA6DChannel-SpatialAttentionN.pdf). In our dataset there were limited point clouds available for every class which were used to obtain the geometric descriptor representative of the class. These were then aggregated with the query image features to obtain the viewpoint estimate. However, the technique mentioned in the above paper had one to one correspondence between the query image and point cloud. Having a one to one correspondence is highly beneficial as we can then combine the 3d maps of query image with the geometrical information provided by the point clouds using spatial and channel attention in 2D. This is a much more effective strategy to aggregate query and geometric features. This point cloud was obtained using the depth map of the query image. So, we wanted to use a pre-trained model to get depth maps which would then be converted to point clouds of the query image thereby, removing the dearth of point clouds available for each class. However, this was very computationally expensive and due to time limitations we had to drop this idea.

Hence, inspired by the idea of attention-based feature aggregation in 2d we tried to implement a 1-D cross-attention based feature aggregation model. We applied the same technique but in 1d by weighting the 1d query encoder output with the attention mask obtained by applying attention on Point Net model features as done in 2d below. The idea was to find the most important values from the 1-d Global shape descriptor and weight the query image features according to it to attain some form of fusion between appearance based and geometrical features. However, on testing this idea it didn't work so we dropped it.



Cross Attention Based Feature Aggregation

However, we thought that applying attention mechanism in the query encoder would still be effective as this will enhance the generated features to be more focused towards foreground object in the query images and less towards the background. This problem was also highlighted while discussing the results in the original paper and this inspired us to try the attention mechanisms to solve this problem. Therefore, we modified our query encoder which is a Resnet-18 model to Bottleneck attention(BAM) based Resnet model. BAM attempts to denoise low-level properties in the beginning. It becomes increasingly focused on the particular target as it progresses.



### Modifying Loss Function

The original paper estimates the viewpoint by first locating the angle bin which it belongs to and then finds the deviation with respect to the angle bin centre. Therefore, the loss function described in the paper consists of cross entropy loss for estimating the bins and regression loss for finding the deviation with respect to that bin. The problem with this loss formulation is that even if the error probability is minimized, it is desirable for the erroneous values to be close to the ground truth labels. In order to achieve this we decided to replace the classification part of the original loss function with a geometric aware loss. We hypothesized that this should help in obtaining better qualitative views due to smaller error values as the predicted angle will be mapped much closer to the ground truth bin as compared to the cross entropy based formulation.

### Generating Results

Our Final Results are obtained by combining the BAM based Resnet and Geometric aware loss. The model outperforms all the previous models to which the original paper model is compared with. However, it still lags a little behind the original paper model by a small amount. We attribute this to the fact that we were not able to proper hyperparameter tuning and using a large batch size as compared to that used in the original paper due to time limitations.

### Some Future Modifications to Try

We can also try applying attention mechanism in the Point Net model to obtain better geometrical features and we found many papers which have done the same. Moreover, since we were not able to implement the third modification which we had proposed related to meta learning training based on MAML. Therefore, this is another modification which we think can be tried as it has been proven beneficial by researchers of MetaView paper who implemented StarMap Algorithm using MAML(StarMap + M) in our results.