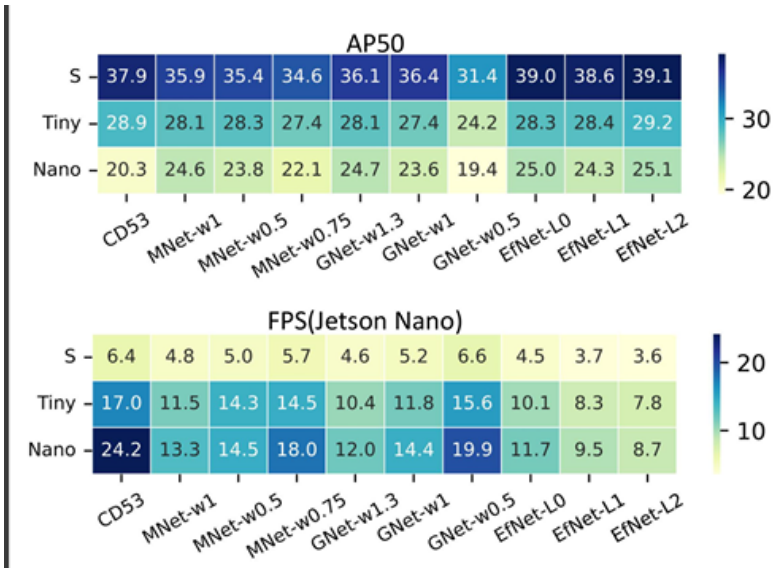
LWUAVDet :- IEEE INTERNET OF THINGS JOURNAL, VOL. 11, NO. 13, 1 JULY 2024

(Lightweight UAV Det)

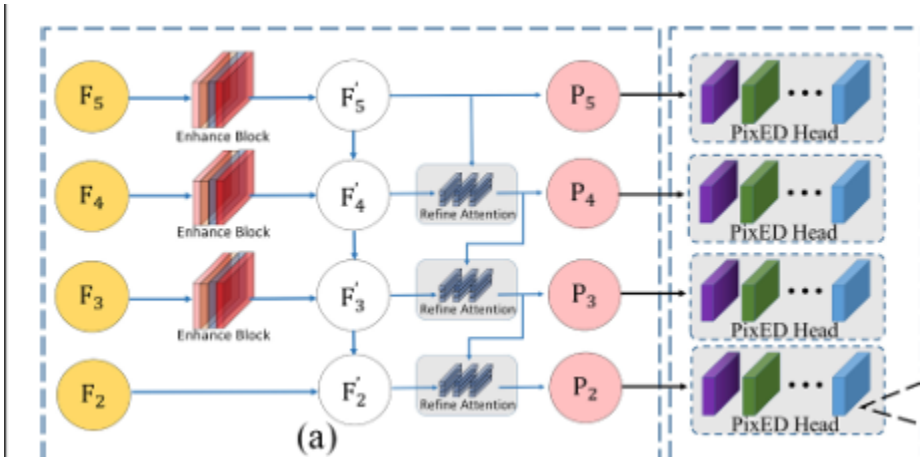
Backbone

Many backbones like MobileNetV2, EfficientNet-L2, GhostNet were evaluated on edge devices It was observed that EfficientNet was giving best detection accuracy while the FPS is lowest among backbones. CSPDarkNet53 was good in all, so it was decided to use it for feature extraction.

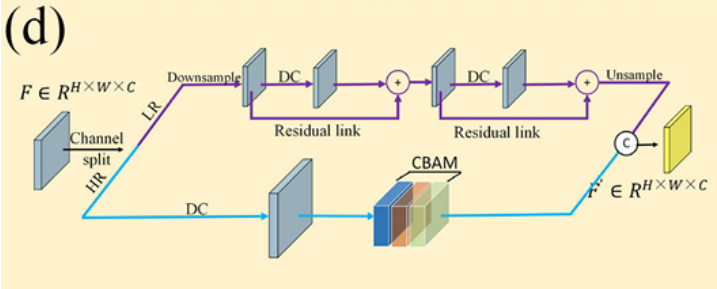


Neck

New architecture called E-FPN was introduced. It generates a 4- level feature pyramid to exchange wider information. It is a normal FPN but has Enhance blocks before getting the feature maps from CD53 and Refine Attention blocks before giving the output to the “Head” part of the model.



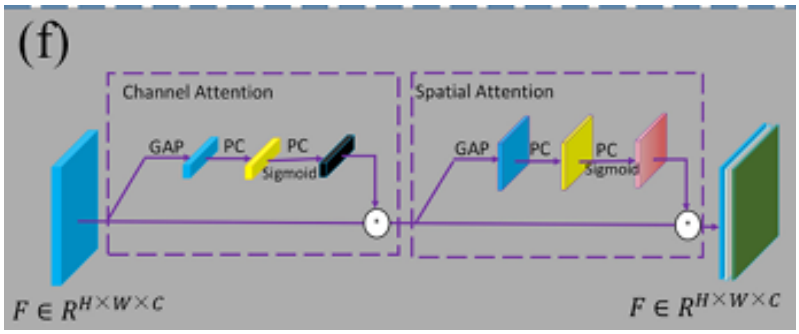
The Enhance block utilizes channel split and spatial sampling to divide feature maps into two sets of branches, i.e, High Resolution(HR) and Low Resolution(LR).



→ Enhance Block

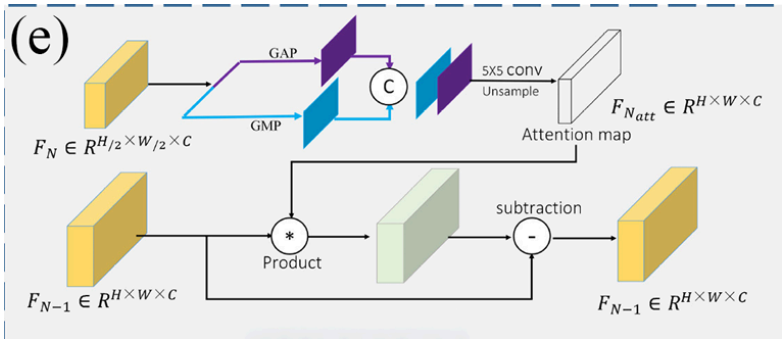
In the LR branch, downsampling is done to capture LR and low-frequency feature map, which reduces spatial redundancy and improves computational efficiency. Then Depthwise convolution is performed and residual link is added for following feature extraction.

We keep the HR branch as input for Depthwise Conv and then introduce CBAM. This has a Spatial Attention layer and a Channel Attention layer. This helps us to refine high-frequency features.



→ CBAM

The repetitive feature fusion of FPN can make tiny objects undetectable due to small are and shallow semantics in lower layers. So the paper proposes a Refine Attention block to eliminate the aliasing area of each stage in FPN, making semantic representation neater for tiny objects.

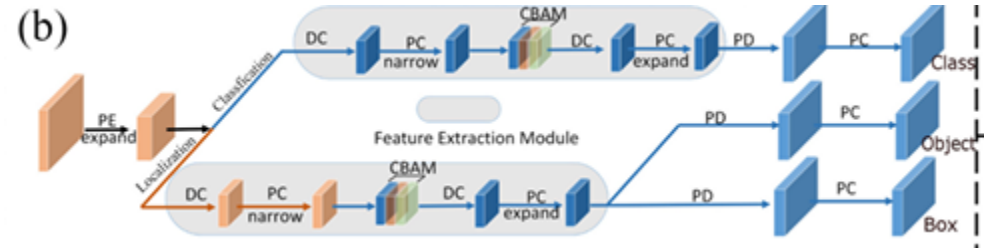


→ Refine Attention Block

In Refine Attention, the adjacent upper layer is influenced by the lower layer and the aliasing area is suppressed by utilizing the features in the lower layer. It first aggregates semantic

representation of the lower layer by using Global Average Pooling(GAP) and Global Maximum pooling(GMP) across channels, and then concatenate the two to generate layer-level attention map (F_{Natt}) using convolution with kernel size 5x5. Then it is upsampled to the same scale of F_{N-1} and element wise multiplication between them is computed. The product feature marks the object area belonging to the lower level in F_{N-1} , and the aliasing effect in the upper layer is weakened by subtraction.

PixED head(Pixel encode[PE], Pixel decode[PD]) is introduced to exchange dimensions between space and channels at both ends of the Head layer.



→ PixED

PE is used to reduce spatial computation and generate channel-rich features. PD decodes the channel dimensions back into spatial locations.

AuxHead is used during training to assist with feature representation and enhance the learning process through distillation.