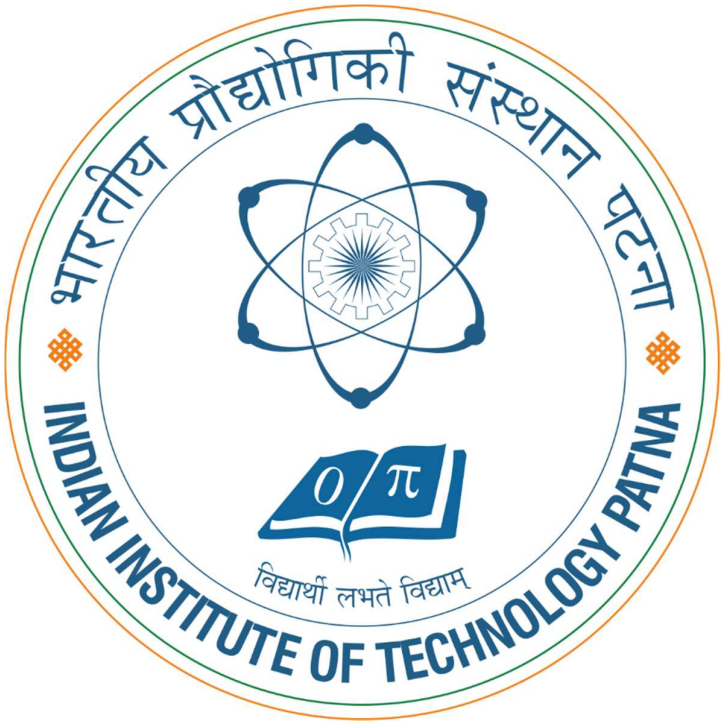
Lite-Mono:A Lightweight CNN and Transformer Architecture for Self-Supervised Monocular Depth Estimation.



Design Lab Report 2023-24

By

Meet Patel [2101EE44]

Supervisor-Dr. Rajib K Jha

Table of Contents:

Objective:-

Depth maps are crucial in robotics, autonomous driving, and augmented reality. Given the expense of depth sensors, research focuses on inferring depth from images. Lite-Mono, a hybrid architecture blending CNNs and Transformers, employs CDC for local feature extraction and LGFI for long-range global information using self-attention. This innovative approach enhances depth estimation accuracy, particularly in resource-constrained settings, offering improved performance and efficiency.

Motivation:-

Shallow networks offer a solution to reduce model size, while CDC's dilated convolutions efficiently expand input observation without adding extra parameters. This enhancement significantly improves local features. However, for capturing long-range information, the reliance on Transformers becomes crucial in providing global representation. Despite their effectiveness, Transformers face challenges in lightweight models due to the linear complexity of MHSA. To overcome this limitation, LGFI introduces cross-covariance attention along feature channels, offering an efficient alternative that bypasses the computational burden of spatial attention computation.

Design Methodology:-

Our model consists of mainly two blocks:- 1.Depth encoder and 2.Depth decoder.
Depth encoder consists of two hybrid blocks one is CDC(Consecutive Dilated Convolutions) and LGFI(Local Global Features Interaction).
The proposed CDC module utilizes dilated convolutions to extract multi-scale local features. Different from using a parallel dilated convolution module only in the last layer of the network we insert several consecutive dilated convolutions with different dilation rates into each stage for adequate multi-scale contexts aggregation.
Considering an input feature X our CDC module outputs X as follows

$$\hat{X} = X + Linear_G\left(Linear\left(BN\left(DDWConv_r\ (X)\right)\right)\right)$$

The LGFI block computes the cross-covariance attention for the input X using the formula:-

$$\tilde{X} = Attention(Q, K, V) + X$$

Where $Attention(Q, K, V) = V.Softmax(Q^T.K)$

Then, the non-linearity of the features can be increased:

$$\hat{X} = X + Linear_G\left(Linear\left(LN(\tilde{X})\right)\right)$$

Our depth decoder consists of a up-sampling block followed by a prediction head to output the depth map at full,$\frac{1}{2}$ and $\frac{1}{4}$ resolution.These are then concatenated and send to image reconstruction block which constructs a depth map from it.

Designed system diagram and photographs:-

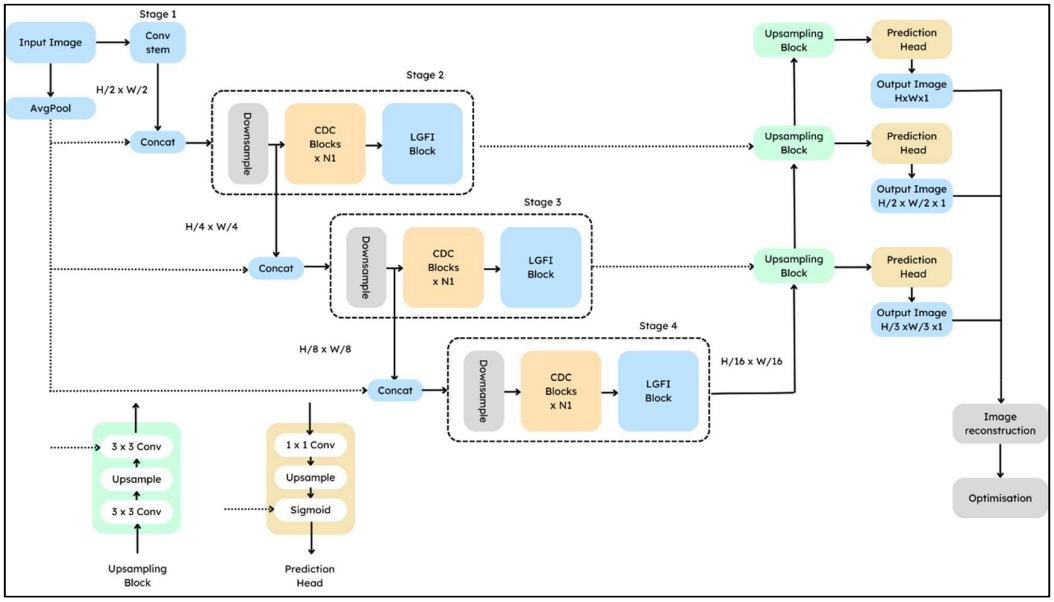


Figure1:Flowchart of complete lite-mono model

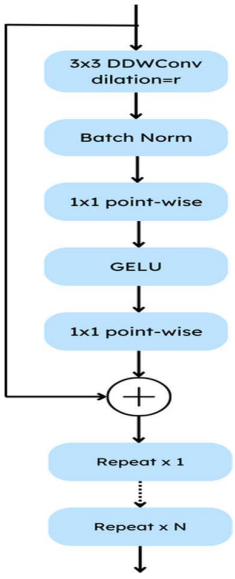


Figure 2:CDC block

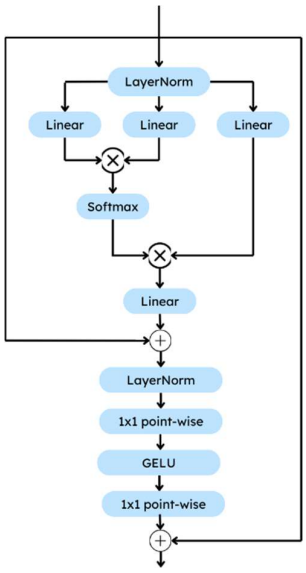


Figure 3:LGFI block

Results and Discussion:-

Figure 4:Input Image(left) and Output depth map(right)

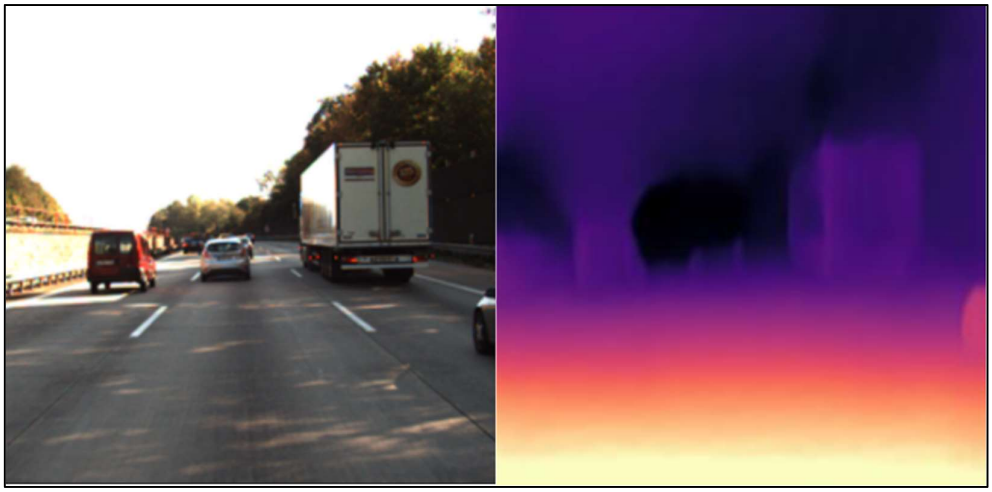


Figure 5:Input Image(left) and Output depth map(right)

Lighter areas denote closer objects in a depth map while Darker areas signify farther objects.

Thus, Our model achieves a good trade-off between model complexity and inference speed.

Future scope of extension:-

- Improving model accuracy by introducing some more layers and fine-tuning them.
- Improve model to detect object more clearly by using some object detection algorithm.

References:-

[1] Zhang, Ning and Nex, Francesco and Vosselman, George and Kerle, Norman. Lite-Mono: A Lightweight CNN and Transformer Architecture for Self-Supervised Monocular Depth Estimation.In CVPR 2023.

[2] Zhongkai Zhou, Xinnan Fan, Pengfei Shi, and Yuanxue Xin. R-msfm: Recurrent multi-scale feature modulation for monocular depth estimating. In ICCV, 2021.

[3] Chaoyang Wang, Jose Miguel Buenaposada, Rui Zhu, and ´ Simon Lucey. Learning depth from monocular videos using direct methods. In CVPR, 2018.