IEEE International Conference on Multimedia Information Processing and Retrieval (MIPR) 2024

Predicting Risk from Dashcam Footage: 2nd AVA Challenge @ IEEE MIPR 2024

Team: **ServerDown**

Outline

1. Introduction
2. Objectives
3. Related Works
4. Dataset
5. Methodology
6. Training and Experimental Setup
7. Result and Performance Analysis
8. Challenges and Future Works
9 . Reference

Task: Predict the Risk of an Impending Car Accident to the Recording Vehicle

- **Video Classification** task to assess whether the dashcam-equipped vehicle is at risk of an accident.
- **Vision-based Sequence Classification** problem.

- Major Differences from Other Video Classification Tasks:
  - Not directly prediction of accident from the video
  - Has to predict the risk to the recorded vehicle of an imminent accident.
  - If there is an accident already happened in front of the recording car and it **has significant effect for the recording vehicle** then prediction will be 1
  - If the accident is severe but the **recording vehicle does not have a significant reaction** then the prediction will be 0

Task: Predict the Risk of an Impending Car Accident to the Recording Vehicle

Though accident/risky situation on the scene but not risky for the recording vehicle

Literature Review

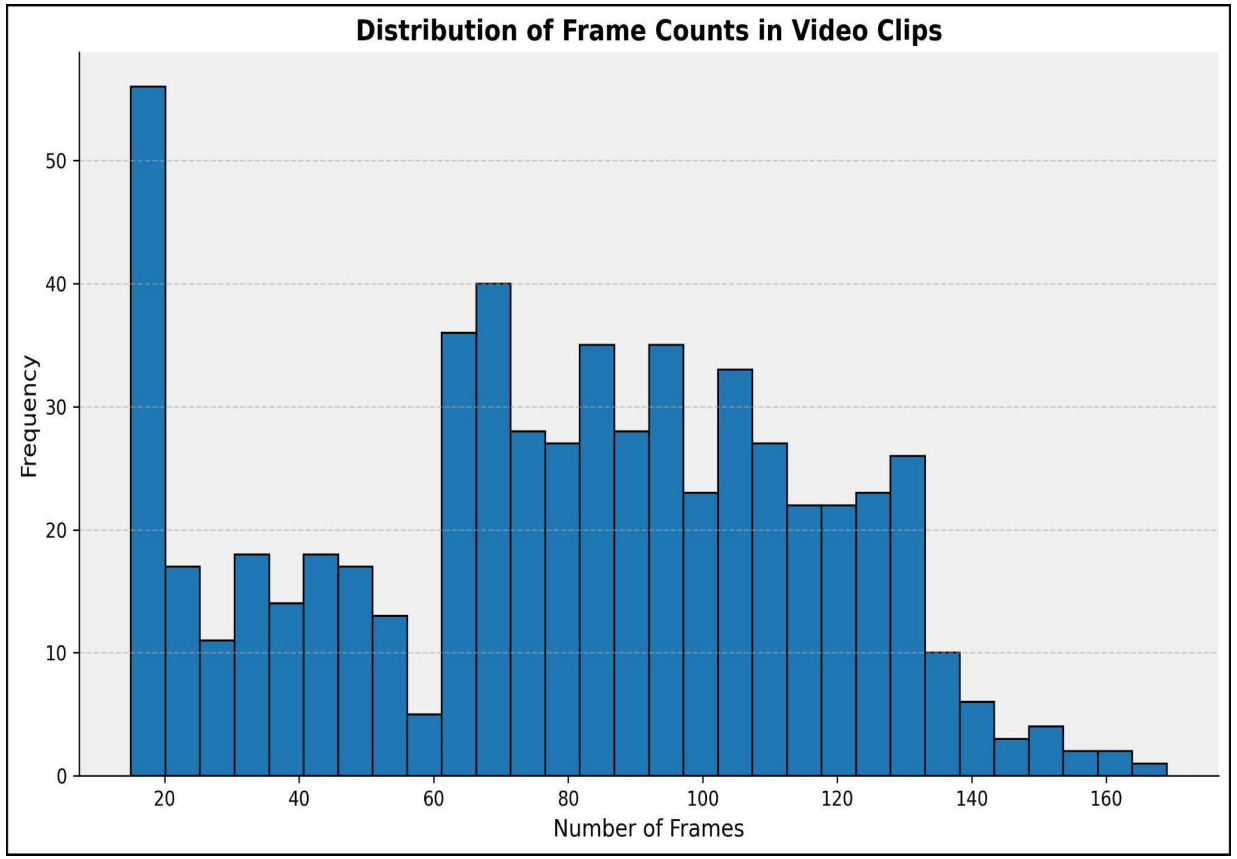
- Facial Expression Recognition in Videos: An CNN-LSTM based Model for Video Classification [8]
  - **Convolutional Neural Network (CNN)**-based feature extractor
  - **Recurrent Neural Network (RNN)** to encode the temporal information

- Can Spatiotemporal 3D CNNs Retrace the History of 2D CNNs and ImageNet [7]
  - 3D CNN to extract both spatial and temporal information
  - Analysis of deep or shallow network to examine its effect on time axis

- ViViT: A Video Vision Transformer [5]
  - Frames are considered as spatio-temporal tokens
  - Attention mechanism

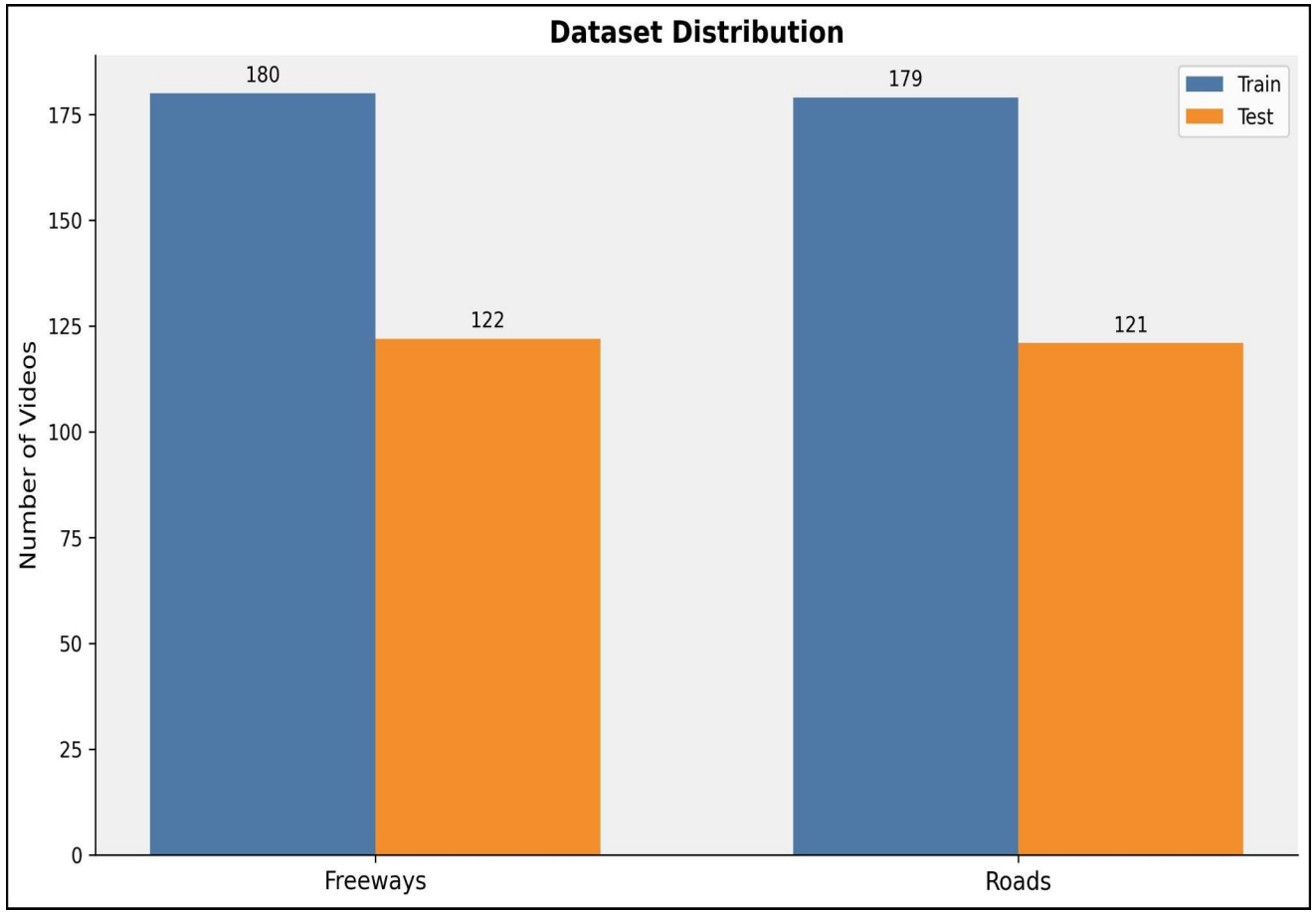
Dataset

- **602** annotated video clips
- 2 labels.
  - 0 - Low Risk
  - 1 - High Risk
- Different situations: **Roads**, **Freeways**
    - Day and Night time videos
- Video clips was provides in continuous image frames
  - Maximum #frames for a video : **169**
  - Minimum #frames for a video : **15**
- Oversampling: Padding with Empty Frames
- Undersampling: Final Frames



Dataset

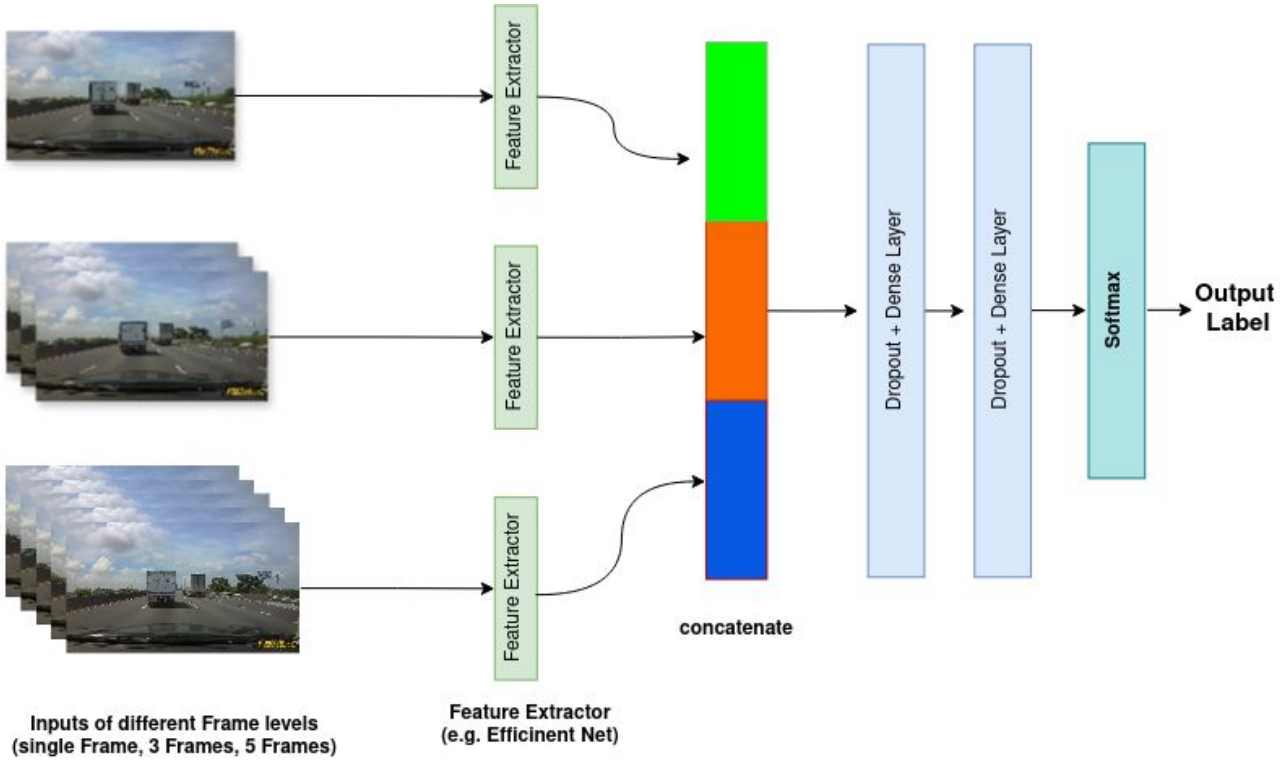
- Dataset distribution
  - Train Clips: **359**
    - Freeways: 180
    - Roads: 179
  - Test Clips: **243**
    - Freeways: 122
    - Roads: 121



First approach: Frame Level Classification

- Architecture
  - Pretrained Convolutional Neural Network (CNN) Backbone as **Feature Extractor**
    - EfficientNetV2M, EfficientNetB5 [1]
    - ConvNext [2]
    - ResNext [3]
  - Dropout + GAP Layer(s)

- Multiple Final Frames Used
  - Only Last Frame
  - Final 3 frames
  - Final 5 frames
  - Combination of All of These

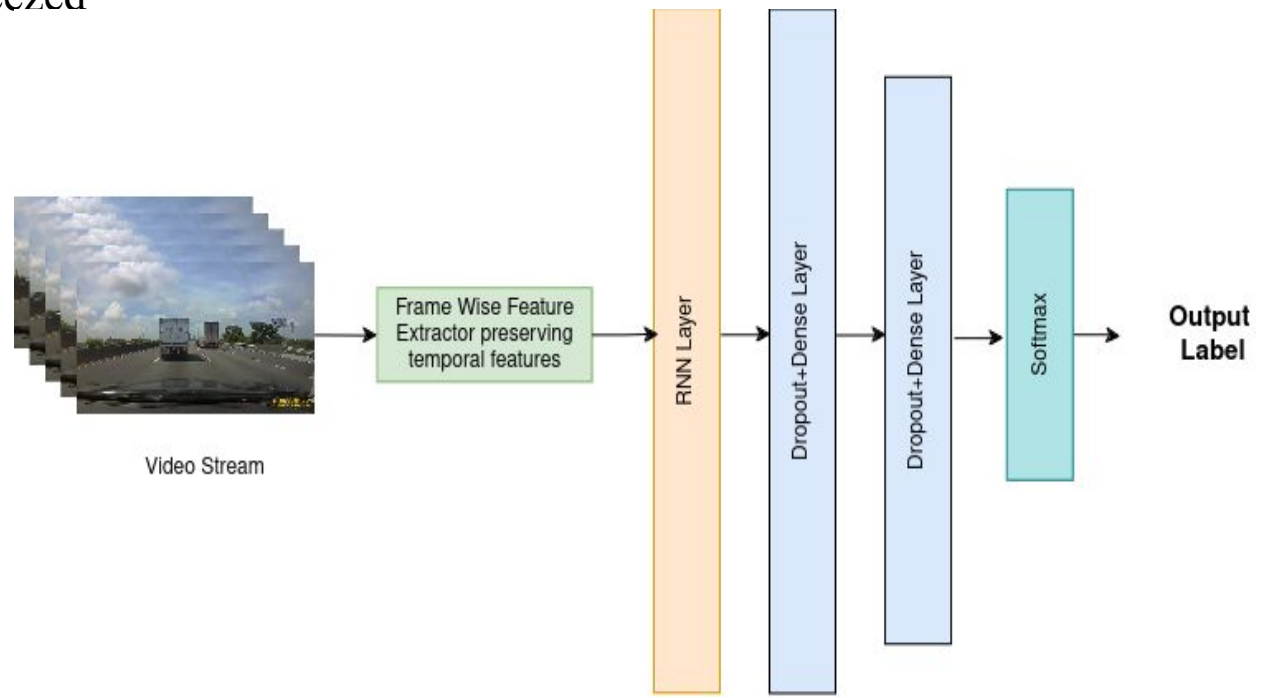


Second approach: Pretrained CNN Encoder with RNN

- Architecture
  - Pretrained Convolutional Neural Network (CNN) Backbone as **Feature Extractor**
    - EfficientNetV2, EfficientNetB5
    - Feature Extractor Layers were Freezed
  - RNN Layer
  - Dropout + Dense Layer(s)

- Different RNN architectures was utilized
  - LSTM
  - GRU

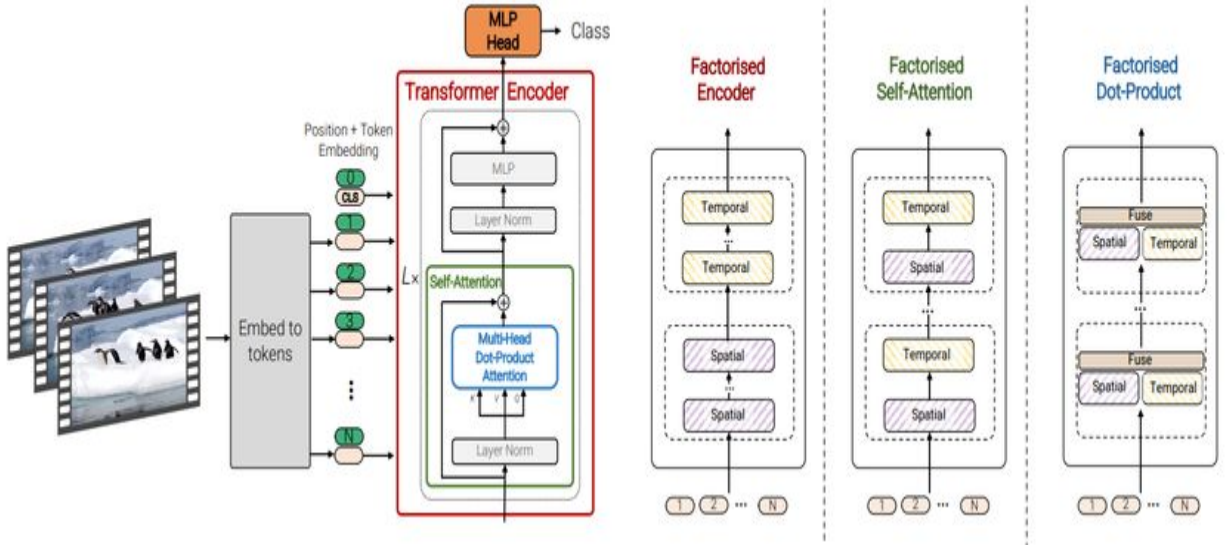
- Able to Experimented with large number of frames. (**64-128** frames)



Third approach: Video Vision Transformer Model

- Architecture
  - Used Pretrained State of The Art Vision Transformer-based Models:
    - ViVit [5]
    - VidSwin [4]
  - Fine-tune final layers with the provided dataset

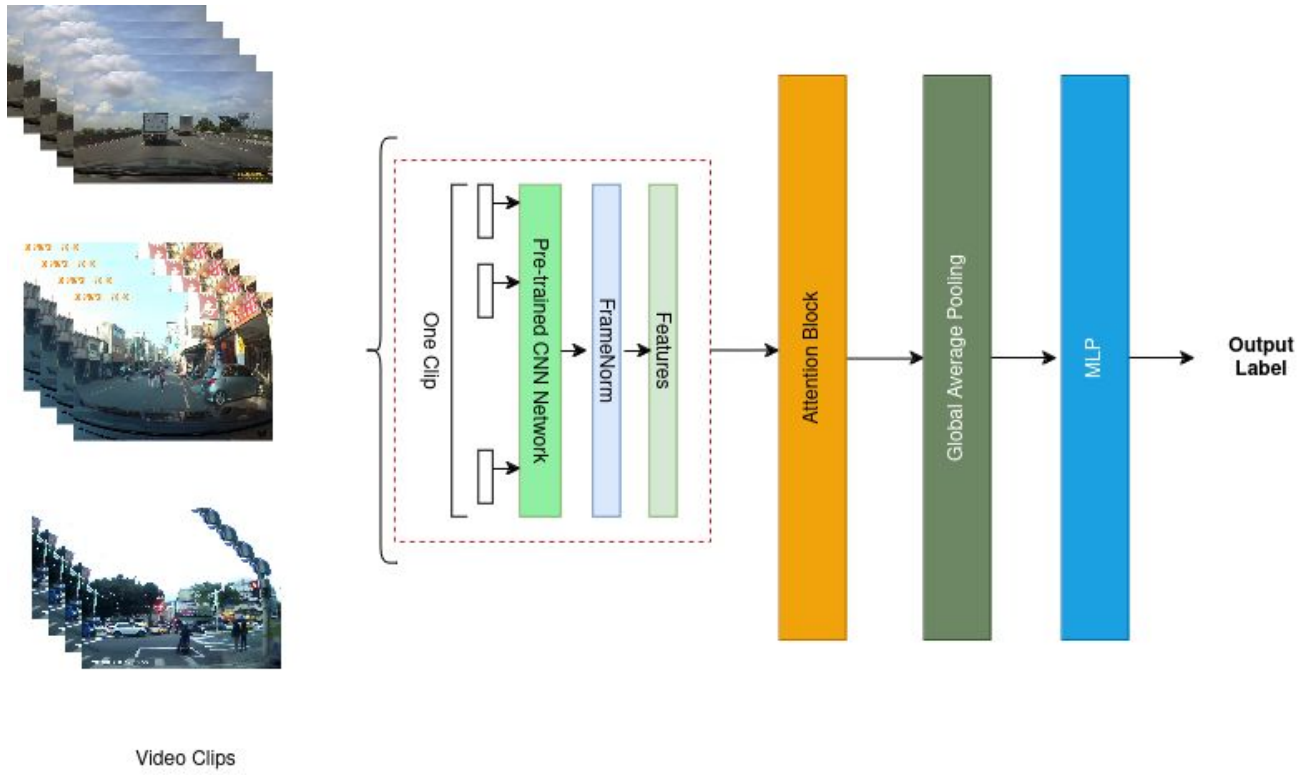
- Experimented with different number of frames. (**16 to 24** frames)



Fourth approach: End-to-End CNN-Transformer

- Architecture
  - Imagenet-Pretrained EfficientNetV2 Backbone as **Feature Extractor** (trainable layers)
  - Frame Norm (Frame wise feature normalization)
  - Transformer Block after FE
  - Finally, MLP with Dropout
    Global Average Pooling for
    the Output Labels

- Experimented with different number of frames. (**16 to 24** frames)



Training and Experimental Setup

- Epoch : 30
- Batch Size : Varied Based on Approach
- Image Size: 224 x 224 x 3
- Loss Function: Cross Entropy Loss
- Optimizer: Adam with ReduceLROnPlateau, SGD with Different Learning Rate Schedules
- Regularization Techniques: BatchNorm, Dropout, Augmentations
- Data Split: Split train data to 80:20 for training and validation

Result and Performance Analysis

Methods Used	Public ROC	Private ROC
Per-Frame Considering Last Three Frames	0.6873	0.7428
VidSwin	0.6619	0.7118
Ensemble (Average) (CNN-Transformer and Per Frame)	0.7459	0.7005
End-to-End CNN Transformer	0.6905	0.6820
Ensemble (Weighted) (CNN-Transformer and Per Frame) (Selected Submission on Kaggle)	0.7560	0.6576
Pretrained-CNN + RNN	0.6280	0.6571

Table 1: Experiment Results

Challenges

- Video with different lighting condition makes it more difficult especially **very low lighting** condition at **night**
- A significant challenge is when an **accident occurs outside the immediate path of the recording vehicle**, leading to misclassification as risky despite no immediate threat
- **Limited data** and model complexity leading to overfitting issues
- **Variable frame rates (FPS)** affected on the sampling methods and also on the lengths of sequences

Future Works

- Grad-CAM-like **visualization tools** to analyze attention weights and interpret model decisions
- **Graph Neural Networks (GNNs)** to detect and track objects, analyze relative distances, and model spatial-temporal dependencies
- **Monocular Depth Estimation**: Using Depth Maps to gauge the distances of the surrounding objects and enhance scene understanding
- Develop **domain adaptation** methodologies like domain adversarial training and meta-learning

Future Works

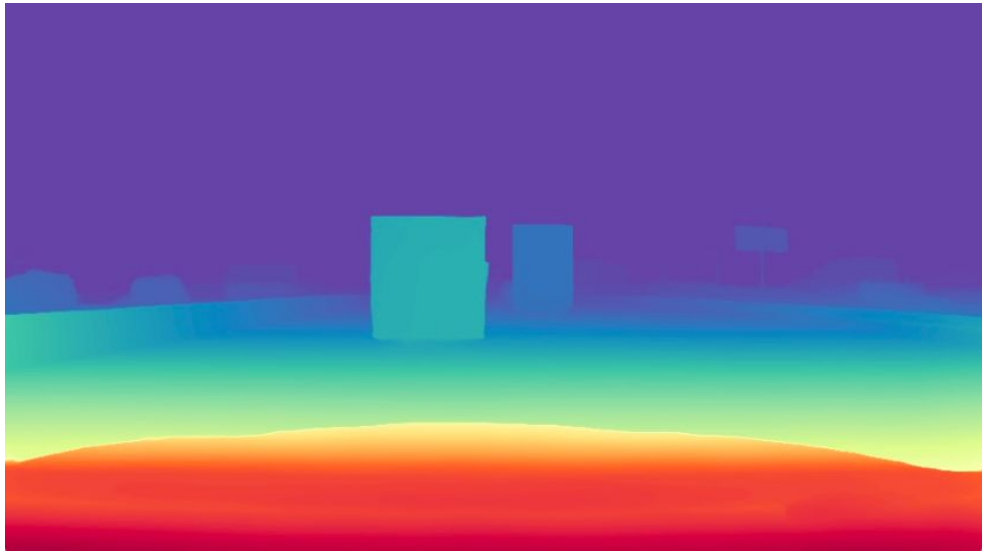


Figure: Original Frame (left) and Generated Depth Map (right) using Depth-Anything [6]

References

- [1] M. Tan and Q. Le, "Efficientnetv2: Smaller models and faster training," in International conference on machine learning. PMLR, 2021, pp.10 096–10 106.
- [2] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 11 976–11 986.
- [3] S. Xie, R. Girshick, P. Dollar, Z. Tu, and K. He, "Aggregated residual´ transformations for deep neural networks," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 1492–1500.
- [4] Z. Liu, J. Ning, Y. Cao, Y. Wei, Z. Zhang, S. Lin, and H. Hu, "Video swin transformer," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 3202–3211.
- [5] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lucič, and C. Schmid,´ "Vivit: A video vision transformer," in Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 6836–6846.

References

- [6] Yang L, Kang B, Huang Z, Xu X, Feng J, Zhao H. Depth anything: Unleashing the power of large-scale unlabeled data. InProceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2024 (pp. 10371-10381).

- [7] Hara K, Kataoka H, Satoh Y. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet?. InProceedings of the IEEE conference on Computer Vision and Pattern Recognition 2018 (pp. 6546-6555)

- [8] Abdullah M, Ahmad M, Han D. Facial expression recognition in videos: An CNN-LSTM based model for video classification. In2020 International conference on electronics, information, and communication (ICEIC) 2020 Jan 19 (pp. 1-3). IEEE.


THANK
YOU