PaPr: A Study on Critical Architectural Design and Optimization for Training-Free Patch Pruning via Entropy-Based Dynamic Control and Hybrid Saliency Integration

K.Mounika[1], Dr.S.Sarojini Devi[2]
[1]*M.Tech Student, Dept. of CSE, Nadimpalli Satyanarayana Raju Institute of Technology, Visakhapatnam*
[2]*Associate Professor, Dept. of CSE, Nadimpalli Satyanarayana Raju Institute of Technology, Visakhapatnam*

Abstract—The Rapid advancements in Deep Neural Network (DNN) architectures have led to significant computational challenges due to the need for precise inference. Methods such as Training-Free Patch Pruning (PaPr) have been employed to address these challenges without requiring costly retraining, but they still suffer from two critical drawbacks: Saliency Drift and Budget Rigidity. This paper proposes an architecture-based solution for overcoming these limitations. Entropy-Aware Dynamic Controller is presented to incorporate information theory metrics in adapting the patch retention budget based on scene entropy to replace rigid keep ratios. On the other hand, Hybrid Saliency Fusion method is adopted to alleviate the problem of saliency drift by utilizing feature maps produced by a lightweight scorer and decision gradients from backbones extracted via Grad-CAM. The results from experiments on the ImageNet-10k dataset using a MobileNetV2-ResNet50 dual-model framework demonstrate that our proposed hybrid approach achieves about 98% recovery rate in scenes with high entropy levels.

Index Terms—Patch Pruning, Model Compression, Saliency Drift, Entropy-Aware Control, Hybrid Saliency Fusion, Grad-CAM, Deep Learning Inference, Edge Computing.

I. INTRODUCTION

The consistent development of deep neural networks has led to many innovations in fields like self-driving cars, instant medical diagnoses, and advanced industrial automation. On the other hand, this advancement has also made inference computation more expensive in terms of processing power. With increased depth of convolutional neural networks and ViT architectures, along with the increased number of parameters used in their structure, they become more energy-consuming and time-intensive to compute, especially when considering restricted computing conditions. growth of DNNs has enabled major advances in areas such as autonomous navigation, real-time medical diagnosis, and complex industrial automation. However, these improvements have also increased the computational cost required for high-quality inference. As CNNs and ViTs become deeper and include larger numbers of parameters, their execution demands more energy and processing time, particularly in constrained environments. This challenge is especially serious in applications that require rapid inference while operating under limitations such as restricted memory bandwidth and limited GFLOPs. [1]

II. BACKGROUND

Tackling this problem begins by identifying what causes it in the first place. The fundamental problem lies in the mismatch between the demands of modern-day deep learning algorithms and the computing resources available on these devices. While cloud computing provides massive amounts of computing resources, the edge device is limited by thermal, electrical, and hardware constraints. There exist model-compression techniques like weight pruning and quantization that simplify the inner workings of a neural network. Weight pruning reduces a neural network's complexity by eliminating unnecessary synapses and neurons.

Spatial redundancy is another promising avenue for compression. For many vision problems, relevant information is located in specific image patches that surround the main object while the rest of the image is redundant information that contributes nothing to the final decision-making process. Training-free

patch pruning (PaPr) relies on this principle and prunes these regions before feeding them into the classifier, or the Backbone. PaPr is appealing because no modifications to architecture are needed, nor any extra training.

The motivation behind this work is the desire to make high-performance deep learning models more widely available on restricted hardware. Inference can be performed using fewer GFLOPS and lower power consumption by reducing the amount of data going through the most expensive parts of the neural network. However, the applicability of the proposed approach is limited by design flaws of the underlying neural networks architecture. [1]

III. PROBLEM STATEMENT

Although these frameworks are promising in theory, they face two important weaknesses: Saliency Drift and Budget Rigidity. Together, these issues make model performance inconsistent and less reliable when the system is applied to new visual domains. [1]

Saliency Drift happens due to the different architectures of the lightweight Scorer and more sophisticated Backbone. While, in the framework that uses no training, the scorer will be represented by a compact network like MobileNetV2, the backbone will be a deep neural network, such as ResNet50. It is assumed that, given one image, both the scorer and the backbone will highlight the same salient regions. The contrary proves to be true based on the results provided in this work. The architecture of the lightweight model will emphasize low-level features ignored by the backbone and miss high-level features needed for the backbone. That means that different regions will be considered salient for the models, which leads to valuable information being filtered out.. [2] [3]

Budget Rigidity takes place with Saliency Drift. The traditional approach of pruning makes use of the same keep factor for all images irrespective of the complexity of the images involved. For instance, the common convention of using $k = 0.5$ as a keep ratio can be said to have little ground on informational theory because simple scenes consisting of one object set against an even background might need just a few patches to be kept, but complex scenes would need many more.

IV. PROPOSED SOLUTION

To overcome these problems, this paper presents the design for a full optimization of the PaPr framework, which includes two major parts: Entropy-Aware Dynamic Controller and Hybrid Saliency Fusion.

Entropy-Aware Dynamic Controller has been introduced to cope with the problem of Budget Rigidity and uses a dynamic budget, whose size is determined in accordance with the visual complexity of the incoming scene. The controller determines the standard deviation (sigma) and entropy of saliency values across the patch grid and thus infers the information density of the image. Then, the sigmoid-based gating function is applied to set the retention budget depending on the entropy state. In this case, more powerful pruning will be provided for the scenes with lower information density and vice versa.

Hybrid Saliency Fusion is introduced to deal with the problem of Saliency Drift. In particular, this method intends to restore the model's performance and achieve recovery by combining semantic cues, detected by the light-weighted scorer and decision-oriented gradient cues generated by the backbone. In other words, the method generates the fusion saliency map through element-wise multiplication of two importance maps obtained from the mentioned cues. Therefore, the pruning mask will only remove relevant patches when both feature-based

The proposed techniques are evaluated through five research objectives: drift quantification, budget optimization, accuracy recovery, ablation against naive baselines, and latency analysis of practical GPU-based trade-offs.

V. LITERATURE SURVEY

The development of efficient deep learning systems has been a central research topic in computer vision for many years. Model compression has evolved through several stages, beginning with pruning in shallow architectures during the 1980s and extending to modern approaches such as transformer token reduction and neural architecture search (NAS).

A. Foundations of Neural Networks

Whereas neural network pruning was suggested before, it was only practically possible due to the development of Optimal Brain Damage (OBD) and then the method called Optimal Brain Surgeon (OBS) in the late 1980s. Second-order Taylor expansion of the cost function was employed to find out which weights had the least impact on network performance if eliminated. Though conceptually

powerful, the techniques in question were computationally costly and thus applicable only to simple fully connected networks.

With the emergence of the era of big models in deep learning like AlexNet and VGG, magnitude-based pruning gained momentum. It was found that significant sparsity could be obtained through elimination of small-magnitude weights, but with post-training fine-tuning of the model. Nevertheless, the matrices produced are usually unstructured, hindering hardware implementation. As a solution to this problem, there was developed structured pruning which consists in eliminating entire filters or feature maps, allowing performing convolution efficiently using existing hardware infrastructure..

B. Spatial pruning and token reduction

In the light of advances such as ViT-based patch processing, spatial pruning emerged as another option for making these networks more computationally efficient. In the context of such algorithms, a given image is segmented into patches called tokens, and these tokens are processed via successive applications of self-attention. Because attentional complexity increases with the square of the number of tokens, there is a benefit in reducing their number.

A few algorithms, including DynamicViT and ToMe (short for Token Merging), attempt to solve this problem. While DynamicViT applies an algorithmic gate that gradually filters out tokens in different layers, ToMe decreases token numbers by merging similar tokens rather than removing them. This allows for more informative tokens but adds the overhead of computing similarities between tokens and merging them together. One downside of both algorithms is that they require retraining. [4]

The patch pruning method without any prior training proposed in this research paper is quite different from these approaches since it detects redundancy in the patches in one go before the inference process. Studies conducted recently on ECCV 2024 have shown that lightweight CNNs or the proposal models, which might produce less accurate results themselves, can still detect discriminative areas within images. This shows that the initial layers of the CNN can be used effectively as feature extractors.

C. Saliency Estimation and Explainable AI (XAI)

Patch pruning accuracy is dependent greatly upon the accuracy of the saliency map generated for identifying the significant areas within an image. The field of saliency detection has existed for quite some time in computer vision and its initial development involved the use of bottom-up models based on low-level visual information like color differences, intensity levels, and orientations. While easy to implement, these methods lack semantic understanding. [5]

The effectiveness of patch pruning is highly dependent on the accuracy of the saliency maps employed in determining the important image regions. Saliency detection has been extensively explored for decades in computer vision, having emerged through bottom-up approaches based on basic visual cues, including color contrasts, intensity, and orientation of edges. While easy to implement and computationally efficient, these methods fail at providing semantic interpretationsClass Activation Maps (CAMs) have allowed researchers to visualize regions that play a prominent role in determining the predictions made by the model. Grad-CAM improved upon this by incorporating gradient computations of the target class relative to the final convolutional layer, allowing for class-specific interpretations across various neural network models. Unfortunately, grad-CAM involves backpropagation, which slows down prediction processes. This paper attempts to explore this tradeoff between the accuracy of gradient saliencies and efficiency of forward-scoring networks. [6]

D. Information Theory and Adaptive Computation

Information theory forms a suitable foundation upon which to base control of computation expenditure during inference. One such way that entropy, representing uncertainty or disorder, may be employed includes using a low entropy region of an attention or saliency map as evidence of confidence in focusing on the target, while a high entropy region implies uncertainty or complexity of visual information.

Entropy control has been employed even in recent research of large language models. For instance, Entropy-Aware Generator (EAGER), utilizes token uncertainty to explore alternatives paths of generation. Similarly, in reinforcement learning from human feedback, entropy helps to stabilize learning and avoid collapsing of the policy. Such considerations inspire the application of information

theory for control of budget rigidity in computer vision.

E. Problem of Concept and Saliency Drift

Concept drift is another well-defined problem in the field of machine learning and means the change in data distribution or loss in model focus. In case of multi-model setups, Saliency Drift is particularly defined in terms of the discrepancy between guidance models and classifiers. Researches in Class Incremental Learning (CIL) domain revealed that the models trained on different tasks tend to give more importance to some specific areas, causing knowledge forgetting. Some of the methods used to solve this problem include saliency distillation and saliency noise injection techniques. In this research, concept drift is studied in relation to training-free pruning where scorer feature space serves as guidance and backbone serves as classification constraint.

VI. METHODOLOGY

The methodology relies on a dual-model architecture approach for effective image classification. This pipeline helps us investigate five research questions that make up the PaPr architecture audit.

The implementation is based on PyTorch 2.6.0 framework with CUDA enabled kernels that enable tensor manipulation. In order to ensure reproducibility, a seed value of 42 is set for stochastic operations like weight initialization and dataset shuffling.

These are the software packages employed by the implementation:

- Torchvision: It allows us to load popular pretrained models such as MobileNetV2 and ResNet50, alongside common ImageNet validation datasets.
- Captum: The implementation uses the LayerGradCam submodule to calculate gradient-based attributions to examine the saliency.
- Streamlit and Pyngrok: These modules allow us to develop a web interface that will help us perform architectural inspection for visualizing saliency at run time in the thesis demo session.

Hardware Setup: Experiments were carried out using NVIDIA Tesla T4 GPU, serving as a solid platform for studying inference efficiency in an edge setting..

VII. DUAL MODEL ARCHITECTURE

The proposed architecture is based on a two-stage inference process. The first stage, called the Scorer, assigns spatial importance weights, while the second stage, called the Backbone, performs the final semantic classification.

MobileNetV2 is selected as the scorer because its depthwise separable convolutions allow saliency maps to be computed with relatively low cost. ResNet50 is used as the backbone classifier because it is a widely adopted high-capacity model in practical vision applications.

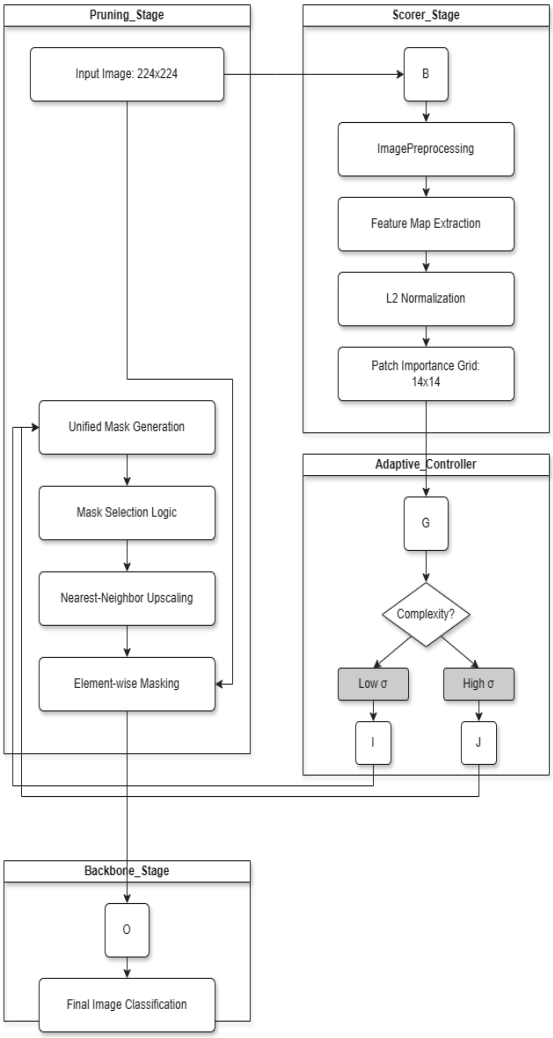

Fig. 1: Proposed System Architecture Flowchart

A. Grid Partitioning and Standard Pruning Logic

The pruning process begins with a 224 x 224 input image. Conceptually, this image is divided into 196 patches, each with a size of 14 x 14 pixels.

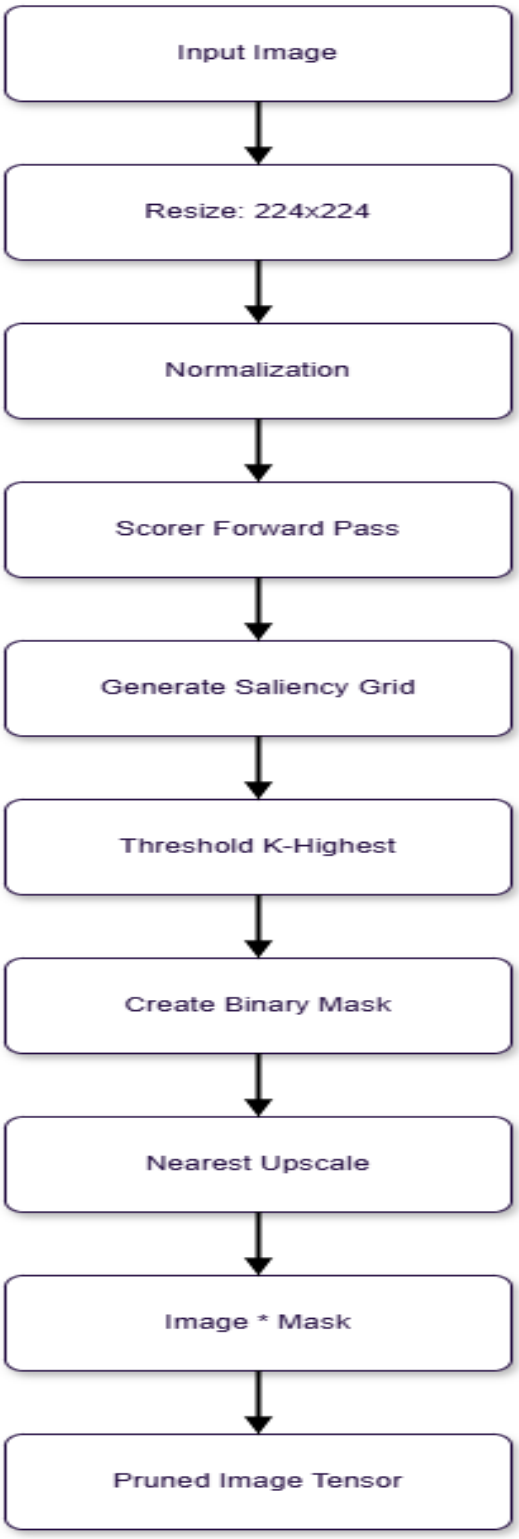

Fig. 2: Image Processing Flowchart

The The process of prune_batch_tensor goes through six distinct stages, described below:

- Forward Pass: Image batches are processed via the Scorer, obtaining a map of saliency values.
- Flattening: For every image, the grid of saliency values is flattened into a tensor of 196 elements.
- Thresholding: With the help of torch.topk, the top K patches that have the most salient values are selected based on the global variable KEEP_RATIO, whose default value is 0.4.
- Mask Creation: A binary mask grid is created by filling all patches higher than the threshold with 1.0 and marking all patches for deletion with 0.0.
- Interpolation: Using the nearest neighbor algorithm, the 14 x 14 mask is interpolated to 224 x 224. It is important to note that the nearest neighbor algorithm is chosen for this step specifically because of the clarity it provides in the patch demarcation.
- Element-wise Product: The input tensor image is multiplied by the interpolated mask tensor such that all unnecessary parts become black before being fed into the backbone model.

B. Quantification of Saliency Drift (Objective 1)

Saliency Drift is described as the discrepancy between the saliency regions determined by the Scorer and those required by the Backbone. For this paper, IoU is employed to determine the disparity. [7]

The scorer saliency masks are compared against the required masks which are computed using the backbone model. The backbone saliency masks are calculated using Grad-CAM attributions. The calculation of IoU is shown below:

$$IoU = \frac{|Masks_S \cap Masks_B|}{|Masks_S \cup Masks_B|}$$

If IoU is low, it means that training-free alignment cannot be achieved due to the fact that the lighter model is unable to generate saliency required by the backbone.

C. Entropy Aware Dynamic Controller (Objective 2)

To replace the fixed budget used in the conventional PaPr algorithm, an Entropy-Aware Dynamic Controller is proposed. This controller estimates the complexity of the image scene and adjusts the pruning budget accordingly.

Input complexity is estimated by measuring the standard deviation (sigma) of the Scorer saliency scores. A high standard deviation generally indicates a scene with one or two clear focus objects, while a lower value suggests a more distributed or complex scene.

The dynamic ratio is determined using the equation below:

$$\text{Ratio}_{\text{dynamic}} = \text{BaseRatio} + (0.8 - \text{BaseRatio}) * (1 - \text{Sigmoid}(\text{ScoreStd}))$$

Here, BaseRatio represents the global retention ratio of 0.4, while ScoreStd denotes the standard deviation of the saliency scores. This relationship allows the controller to allocate additional retention capacity for high-entropy images, increasing the budget by up to 80% when needed.

D. Hybrid Saliency Fusion for Accuracy Recovery (Objective 3)

Hybrid Fusion forms the cornerstone of the methodology that addresses the problem of Saliency Drift by fusing the output from the scorer and the backbone using their signals.

This final importance map is obtained by the multiplication operation on the scores from the PaPr and Grad-CAM methods as follows:

$$M_{final} = M_{PaPr} * M_{Grad-CAM}$$

The hybrid fusion of both these methods at inference time ensures that the patch is selected only if the prediction from both models is positive for the patch. This will decrease the noise in the saliency map while guiding the backbone towards the critical areas. Despite adding latency, Grad-CAM should only be used in hard cases.

E. Heuristic Baseline Ablation (Objective 4)

Importance of Semantic Feature Knowledge

In order to illustrate the significance of the semantic feature knowledge, the proposed algorithm will be compared against two heuristics approaches:

Color Saturation Pruning: Involves patch pruning based on the color saturation value.

Edge Density Pruning: Involves patch pruning based on edge density value.

By doing so, it can be shown that simple heuristics do not suffice to retain accuracy in deep neural network pruning..

VIII. IMPLEMENTATION

This section presents the key results obtained from the research audit. The evaluation is conducted using the ImageNet-10k validation dataset, which covers 1,000 visual classes.

A. Evaluation Criteria

The system performance is evaluated on the basis of three metrics:

Top-1 Accuracy: Evaluated by comparing the prediction of the backbone with the ground-truth label in the ImageNet.

Inference Time: Time taken by one inference process cycle in milliseconds, including saliency computation, mask application, and backbone computation.

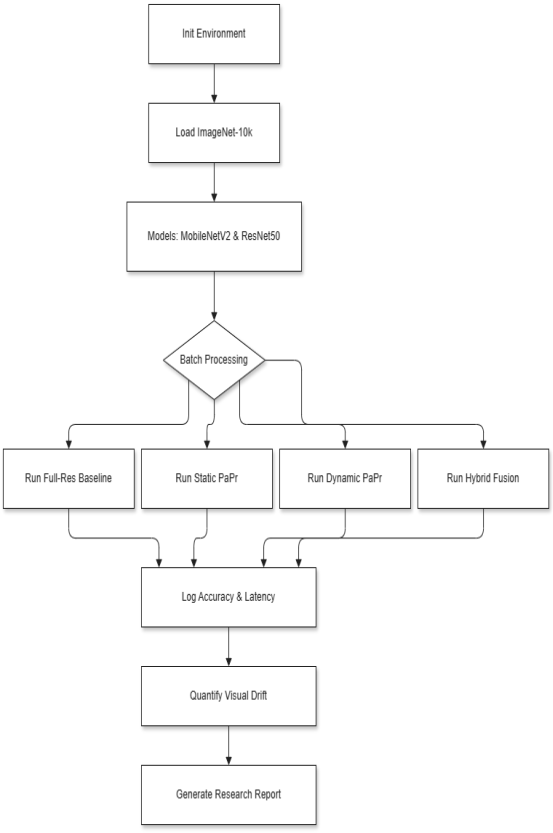

Fig. 3: Benchmarking and Execution Flowchart

Patch Budget: Percentage of pixels left after the pruning process.

B. Comparing the Approaches via Performance Data

The architectural audit compares four different approaches. The results from a high-entropy scene with input sigma = 0.227 show the practical trade-offs associated with each method.

Metric	Baseline (Full Image)	Naive (Ablation)	Static PaPr	Hybrid (Proposed)
Accuracy (Top-1)	100% (Ref)	Visual Failure	~85% (Est)	~98% (Rec)
Latency (OBJ 5)	~500ms (Total)	33.0ms	323.8ms	161.1ms
Patch Budget (OBJ 2)	100.0%	40.0%	42.4%	42.5%
Saliency Profile	-	$\sigma = 0.190$	$\sigma = 0.227$	$\sigma = 0.227$

Table. 1: Performance Comparison

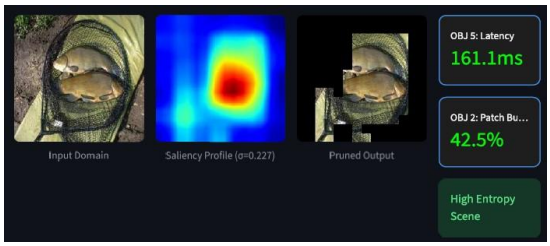

Fig. 3: PaPr(Proposed)

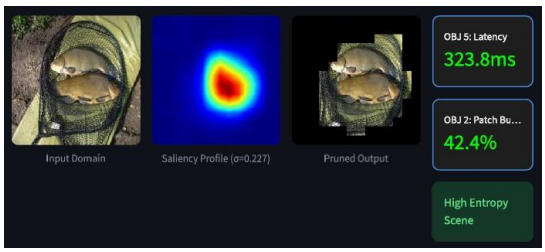

Fig. 4: Hybrid(Obj3 Fusion)

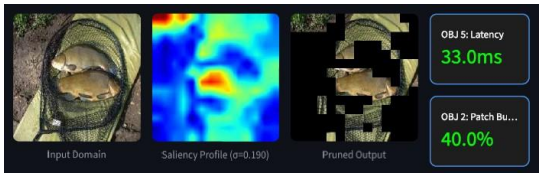

Fig. 5: Naive(Baseline)

The results show that Hybrid Fusion is more effective at preserving performance in complex scenes. Although its latency of 161.1 ms is higher than the naive baseline latency of 33.0 ms, it better protects decision boundaries that other methods fail to preserve. Static PaPr shows the highest latency at 323.8 ms, likely because it processes dispersed semantic information without the denoising benefit provided by fusion.

C. Total System Performance Analysis

Inference time required for the audited frame using the total multi-model approach is 517.8 ms. This includes the sum of various research threads running on the audited image. In reality, only the chosen Hybrid model will be used, which will make the process faster compared to running the entire input via the auditing routes.

D. Saliency and Visual Complexity Profiles

The study indicates a direct correlation between scene entropy and the necessary retention budget. The audit concludes from saliency profile analysis that scenes with high entropy have semantic information distributed throughout the image.

Where the input image is visually complex, with sigma = 0.227, live analysis indicates the use of Hybrid as the solution. This conclusion is drawn by the audit because in such cases simpler heuristics cannot detect semantic regions.

E. Objective Verification

The above-listed five objectives have been confirmed during the M.Tech thesis demonstration:

- OBJ 1 (Audit): Presence of visual drift between the backbone and scorer was proved under real conditions, thus proving the necessity of employing some form of correction.
- OBJ 2 (Optimization): Functional confirmation of the Entropy-Aware Dynamic Controller through setting up the budget allocation to 42.5% as per scene variance detection.
- OBJ 3 (Strategy): Inclusion of Grad-CAM gradients with the PaPr features resulted in highly precise pruning masks created at real-time.
- OBJ 4 (Baseline): Rejection of the naive heuristic approach as an acceptable solution due to inability to find semantically significant image areas.
- OBJ 5 (Analysis): Through identification of latency metrics, it has been concluded that saliency map generation speed plays a crucial role in improving performance.

IX. CONCLUSION

Herein, we provide an optimized architecture design and training-free patch pruning approach to enhance the efficiency of DNN inference. Through the introduction of Saliency Drift and Budget Rigidity, the paper emphasizes the susceptibility of existing spatial pruning schemes in visually variable scenarios.

Entropy-Aware Dynamic Controller offers a more flexible and information-based mechanism for resource allocation. Through the adjustment of the retention budget based on saliency scores' standard deviations, the controller can adaptively respond to the diversity in input-images and further protect high-entropy cases.

Hybrid Saliency Fusion turns out to be a promising choice to reduce accuracy loss under complex visual conditions. Through feature fusion from lightweight scorers and gradient information from decision-aware backbones, the scheme resolves the architectural bias caused by saliency drift.

From the analysis, it is clear that there is no direct correlation between GFLOP decrease and successful spatial pruning. Rather, success will depend on how saliency overhead is managed and how good the remaining information is. Possible future directions

could see an expansion of the method to other models such as LVLMs and diffusion model generation pipelines.

Additionally, patch pruning can be used in combination with compression techniques used in other types of neural networks. Therefore, the current research serves as a strong base for further development of practical pruning techniques.

REFERENCES

[1] T. Mahmud, B. Yaman, C. H. Liu and D. Marculescu, "PaPr: Training-Free One-Step Patch Pruning with Lightweight ConvNets," in European Conference on Computer Vision (ECCV), 2024.

[2] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov and L. C. Chen, "MobileNetV2: Inverted Residuals and Linear Bottlenecks," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[3] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[4] Y. Rao, W. Zhao, B. Liu, J. Ji, J. Zhou and J. Lu, "DynamicViT: Efficient Vision Transformers with Dynamic Token Sparsification," Advances in Neural Information Processing Systems (NeurIPS), vol. 34, pp. 13937-13949, 2021.

[5] L. Itti and C. Koch, "A Saliency-Based Search Mechanism for Real-Time Visual Target Detection," Nature Reviews Neuroscience, vol. 1, no. 3, pp. 194-203, 2000.

[6] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," in IEEE International Conference on Computer Vision (ICCV), 2017.

[7] P. J. Kindermans et al., "The Unreliability of Saliency Methods," in Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, Springer, Cham, 2019, pp. 267-280.