IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Saliency Guided Bit Allocation For Deep Image Compression For Efficient Storage

Kailash S, Karthik M, Dr. N. Revathi

Department of Computer Science and Engineering

Sri Venkateswara College of Engineering, Chennai, India

Doi: : https://doi.org/10.56975/ijcrt.v14i7.309022

Abstract

Conventional image compression applies nearly uniform quality across all regions of an image, in contrast to human perception and downstream vision tasks which are much more sensitive to distortions in semantically important regions than in smooth or visually unimportant background areas. The proposed paper addresses this inefficiency through a saliency guided image compression framework that estimates the importance of pixels using three complementary modules: deep salient object detection, semantic object segmentation, and spectral residual saliency analysis [1]. The output of these modules are fused into a unified importance representation, which is then transformed into a spatially varying bit allocation map through an Ascending Cosine Roll down (ACRD) transfer function that emphasizes perceptually relevant regions while suppressing background detail.

The compression pipeline then performs layered reconstruction by combining a strongly degraded base layer with a high quality or exact foreground layer according to the learned weight map, thereby concentrating quality where it matters most and improving compressibility elsewhere. Unlike end to end neural codecs that require extensive retraining and large annotated datasets, the proposed framework is modular, training light at the system level, and built from independently replaceable components, which makes it suitable for academic prototyping and practical deployment. The overall design is closely aligned with modern research trends in saliency oriented learned compression, especially the principle that local fidelity should be adapted according to task relevance rather than distributed uniformly across the image [6].

Index Terms : saliency detection, image compression, bit allocation, context aware coding, layered reconstruction.

I. INTRODUCTION

The rapid growth of visual data in mobile, cloud, surveillance, and multimedia applications has intensified the need for image compression systems that are both storage efficient and perceptually intelligent.

Traditional codecs, as well as many standard compression pipelines, still optimize image quality in a spatially uniform manner, allocating similar fidelity to foreground subjects and visually insignificant backgrounds despite their unequal contribution to perceived quality and semantic utility. This uniform treatment becomes especially inefficient in images where a limited set of objects, boundaries, or regions dominate attention while the remaining content can tolerate stronger degradation without materially affecting interpretation.

Recent research in image compression has shown that machine oriented and saliency aware coding can improve the rate distortion trade off by protecting regions that are important to downstream analysis or human perception [2]. In particular, saliency segmentation oriented compression demonstrates that adaptive bit allocation at the pixel level can preserve distortion sensitive structures more effectively than global quality control, because not all pixels contribute equally to segmentation accuracy or visual importance [6]. This research direction provides strong conceptual support for the present paper, which extends the same central idea into a practical multi detector fusion framework that combines semantic, perceptual, and frequency domain cues [1, 7].

The main motivation for this work is that a single saliency source is usually insufficient for robust importance estimation across diverse images. Deep saliency detectors often generate smooth but spatially coarse prominence maps, semantic detectors can miss objects outside trained categories, and spectral methods may overreact to non semantic texture or noise. The proposed framework addresses this limitation through OR style fusion, ensuring that a pixel is retained as important if any modality considers it relevant, thereby increasing robustness across natural scenes, object centric imagery, and structure rich content.

The central problem addressed in this paper is the mismatch between spatially uniform compression and spatially non uniform perceptual importance. In most real world images, only a fraction of the scene carries critical semantic or visual content, such as faces, foreground objects, boundaries, textural cues, or high attention regions, while large background areas contribute little to interpretation and can be encoded more aggressively. However, conventional compression methods lack an explicit mechanism to estimate and exploit this distinction at the pixel level, which results in wasted bitrate and avoidable degradation of important content.

The problem becomes more apparent in modern image analysis settings, where compressed images are not only consumed by humans but also passed through machine perception systems. Prior research has shown that compression artifacts can significantly affect downstream analysis tasks, and that adaptive local fidelity control is often more useful than uniform quality preservation when protecting task relevant structures [3, 5]. Saliency oriented compression therefore offers a meaningful solution space because it can treat image compression as a selective preservation problem rather than a uniform reconstruction problem [6].

The proposed paper, titled Saliency Guided Bit Allocation for Context Aware Image Compression, is designed around a strict feed forward architecture consisting of five stages: deep saliency detection, semantic object segmentation, spectral residual saliency estimation, adaptive bit allocation, and layered compression based image generation. The system does not rely on a single saliency predictor, because each detection modality captures a different notion of importance: holistic visual prominence, object level semantics, and structural novelty. By integrating these cues through a unified fusion and weighting mechanism, the paper aims to protect meaningful foreground details, sharpen transition boundaries, and reduce background entropy before final encoding.

II. RELATED WORK

Saliency oriented compression belongs to a broader family of content adaptive image coding methods that allocate more resources to important regions and fewer resources to redundant ones. In deep image compression research, recent approaches have shown that local distortion should be controlled according to task relevance, especially for machine vision applications such as detection, classification, and segmentation [3, 5]. The reference paper demonstrates that saliency segmentation can be interpreted as a pixel wise decision problem, and that pixels near the segmentation boundary
or decision hyperplane are more distortion sensitive and thus deserve higher coding priority [6].

The proposed student paper shares the same philosophy of non uniform quality assignment but differs in implementation strategy. The reference work uses a learned compression network, probability driven bit allocation, latent feature masking, and a double scale entropy module to optimize rate accuracy performance for downstream saliency segmentation [7]. In contrast, the current paper constructs a practical, modular pipeline in which saliency estimation and layered blending are explicitly separated from the downstream codec, allowing adaptive foreground preservation without training a task specific compression model.

Three specific streams of prior work are particularly relevant to this paper. First, deep saliency estimation models such as U2 Net provide strong holistic foreground prediction and have become standard tools for salient object detection [6]. Second, semantic instance segmentation models such as YOLO based architectures offer sharper object boundaries and category aware localization, which can complement soft saliency maps when foreground structure must be preserved accurately. Third, spectral residual saliency methods remain valuable because they are training free and sensitive to edges, textures, and statistically unusual regions that may not activate deep models strongly [1].

III. PROPOSED SYSTEM OVERVIEW

A. Module 1: Deep Saliency Detection

The first module of the system estimates holistic visual importance using a lightweight U Net based salient object detection model identified in the paper description as U NetP, corresponding to the compact U2 Net family design. This model is implemented as a nested encoder decoder structure with residual U blocks that capture multi scale context at several depths, enabling the detection of prominent foreground regions even when the object boundaries are diffuse or embedded in cluttered scenes. The model processes an image resized to a fixed resolution and produces a normalized saliency map in the range [0, 1], later resized back to the original image size for alignment with the rest of the pipeline.

A key strength of this module is that it captures scene level prominence rather than just category membership. In other words, it can identify visually important regions that appear salient even if they are not recognized as one of a fixed set of object classes. This makes it especially useful for natural images, portraits, and composition driven scenes where foreground importance is perceptual rather than purely semantic.

However, the paper report also acknowledges the limitations of relying on deep saliency alone. The output can become spatially smooth or blob like, especially around fine boundaries, and may underrepresent sharp object edges or narrow structural details that still matter during compression. This

limitation motivates the addition of semantic and spectral branches, which together compensate for the coarse boundary behavior of holistic saliency prediction.

B. Module 2: Semantic Object Segmentation

The second module generates an object aware saliency signal through semantic instance segmentation using a YOLOv8 nano segmentation model. Instead of predicting a soft prominence field, this branch produces per instance masks for recognized objects and merges them into a unified object map, thereby capturing category level subject regions with relatively crisp boundaries. The output is treated as a soft union mask in which a pixel's value reflects the strongest instance confidence covering it.

This semantic branch plays an important role because deep saliency and frequency based novelty do not always guarantee category aware foreground protection. In practical scenes, humans often consider recognizable objects such as people, vehicles, animals, or common items as important even when they are not globally dominant in contrast or composition. The segmentation branch explicitly injects object level awareness into the importance model, improving preservation of semantically meaningful regions.

The paper also notes that this module is inherently limited by the training vocabulary of the segmentation model. Objects outside the supported category set may be missed entirely, and images containing unusual, domain specific, or abstract subjects may not benefit from semantic detection. For this reason, the branch is not used alone, but as one contributor within a broader multi source fusion strategy.

C. Module 3: Spectral Residual Saliency

The third module estimates saliency using a classical spectral residual method applied at multiple image scales [1]. This approach transforms the grayscale image into the frequency domain, extracts the log spectrum magnitude, subtracts a smoothed spectral envelope to obtain residual novelty, and then transforms the result back into the spatial domain to obtain a saliency map that emphasizes statistically unusual structures. By applying this process at several scales and averaging the resulting maps, the system reinforces persistent salient structures while suppressing unstable noise responses.

The spectral residual branch is valuable because it is training free and category agnostic. It can highlight edges, texture discontinuities, and fine structural detail even in images where deep models are uncertain or where semantic classes are not recognized. This gives the system an additional sensitivity to boundary precision, local complexity, and structural cues that are often crucial for visually convincing foreground preservation.

At the same time, spectral saliency is not equivalent to perceptual relevance. A region may have high spectral novelty because of noise, clutter, or repeated high frequency patterns rather than because it is semantically or visually important. The paper avoids over reliance on this branch by treating it as a complementary signal and combining it through controlled maximum fusion with the other two detectors.

D. Multi Source Fusion Strategy

A core contribution of the paper is the decision to combine the three saliency sources through element wise maximum fusion rather than weighted averaging. This means that a pixel is protected if any detector identifies it as important, which effectively implements an OR style preservation logic across modalities. The fusion strategy prevents strong evidence from one branch from being diluted by weak responses from the others, an issue that commonly affects average based combination rules.

From a systems perspective, the three branches capture different notions of importance. The deep saliency detector models holistic visual prominence, the semantic branch captures category defined subjects, and the spectral branch highlights structural novelty and edge level complexity. Maximum fusion therefore creates a conservative protection policy, favoring retention when there is disagreement, which is sensible for compression because losing a truly important region is usually more harmful than slightly over preserving a non critical one.

The report also introduces a spectral boost parameter before fusion, allowing the spectral residual branch to punch through the combined map more strongly when edge protection is especially important. This is a practical design choice because boundaries and texture transitions are frequently the first regions to show objectionable artifacts under aggressive compression, and a boosted structural response can help preserve them [5]. The resulting fused map forms the input to the bit allocation stage.

E. Bit Allocation Using ACRD

After fusion, the saliency map is transformed into a pixel wise quality control signal through the bit allocation module. The module first applies a hard threshold to suppress low saliency regions completely, ensuring that weak background responses do not consume unnecessary bitrate. It then passes the thresholded values through the ACRD function, a smooth raised cosine style mapping that converts saliency scores into allocation weights between zero and one.

The ACRD function is central to the paper's methodology because it provides a non linear but smooth transition between low importance and high importance regions. According to the paper description, the function is monotonic, has zero derivative at the endpoints, and follows an S shaped profile that accelerates weight growth in the mid saliency range. This makes the transition perceptually smoother than a hard binary split and avoids abrupt quality discontinuities at region boundaries.

This design is conceptually consistent with the reference paper, which also uses an ascending cosine roll down mechanism to derive bit allocation from saliency probabilities and argues that compression resources should be concentrated around distortion sensitive pixels rather than distributed uniformly [6]. While the reference paper implements this principle inside a learned latent space masking framework, the current paper translates the same idea into an explicit image space weighting mechanism suitable for a modular system [7].

(Fig 1: Original Image Size; 30mb)

(Fig 2: Compressed Image using our workflow Size 2.7mb)

The paper further extends ACRD with gamma based curve shaping and floor ceiling clipping. Gamma values above one create a harder separation between foreground and background, while values below one produce a softer transition and preserve more mid saliency regions. Floor and ceiling limits ensure that no pixel receives less than a minimum quality level or more than a bounded maximum allocation, thereby stabilizing visual output and preventing dominant foreground regions from monopolizing the entire quality budget.

F. Layered Compression Framework

The final operational stage of the system is layered compression, where the bit allocation map determines how the original image is blended between two quality levels. The first level is a strongly degraded base layer intended to represent the background or low importance regions with minimal

entropy, while the second level is either a higher quality enhancement layer or the exact original foreground, depending on the chosen operating mode. The final image is produced by pixel wise interpolation between these two layers using the computed weight map.

The base layer is deliberately designed to be easy to compress. According to the paper rundown, it is generated through a combination of bilinear downsampling and upsampling, box filter blur, and additive noise, which together remove high frequency texture, soften edges, and increase spatial homogeneity. This is a practical approximation of aggressive low fidelity background encoding and serves the broader goal of reducing entropy before the final codec stage.

The enhancement layer is generated with much milder degradation, preserving far more local structure and color fidelity than the base layer. In standard lossy mode, the output blends the base and enhancement layers continuously, allowing soft transitions across saliency boundaries. In the paper's foreground lossless mode, the enhancement source is replaced by the exact original image, which guarantees mathematically exact preservation for pixels whose allocation weight reaches one.

This layered strategy bears conceptual similarity to the reference paper's separation into base and enhancement channels, where global structure is retained broadly and extra coding capacity is reserved for important pixels [6]. The difference is that the student paper performs the separation directly in image space rather than in latent feature space, making the system more transparent and easier to explain in a paper while still reflecting the same adaptive quality allocation principle.

G. Mathematical Interpretation

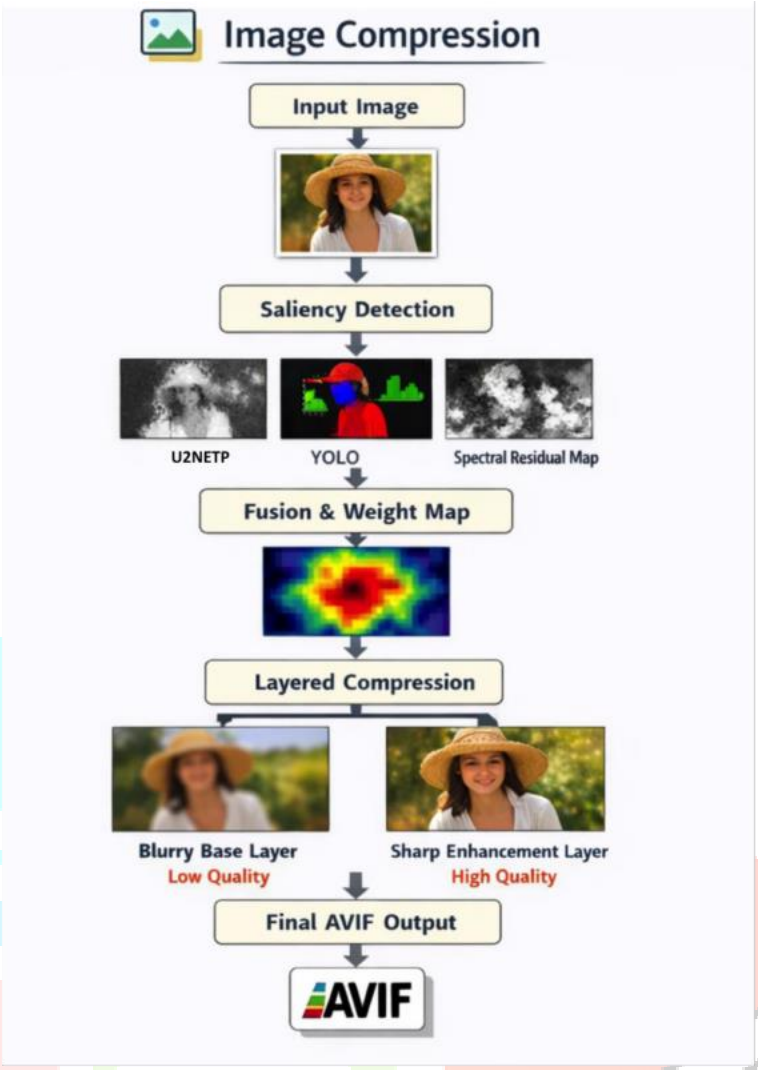
The proposed framework can be described mathematically as a sequence of saliency estimation, fusion, transfer mapping, and spatial blending operations. Let S^d, S_s, and S_r denote the normalized deep saliency map, object segmentation map, and spectral residual saliency map respectively. The fused saliency representation can then be expressed as the element wise maximum of the available modalities after optional spectral amplification, which creates a conservative estimate of importance preserving any strongly signaled pixel.

If the fused map is denoted by S^f, the thresholded map is obtained by setting all values below a chosen threshold to zero, thereby removing weak background activation. The ACRD mapping then transforms each surviving pixel importance value into an allocation weight W, where low values correspond to aggressive degradation and high values correspond to stronger preservation. The final image can be interpreted as a weighted interpolation between a base representation B and an enhancement or original representation F, following the relation:

$$I_{out,i,j} = (1 - W_{i,j})B_{i,j} + W_{i,j}F_{i,j}$$

as described in the paper's layered compression design.

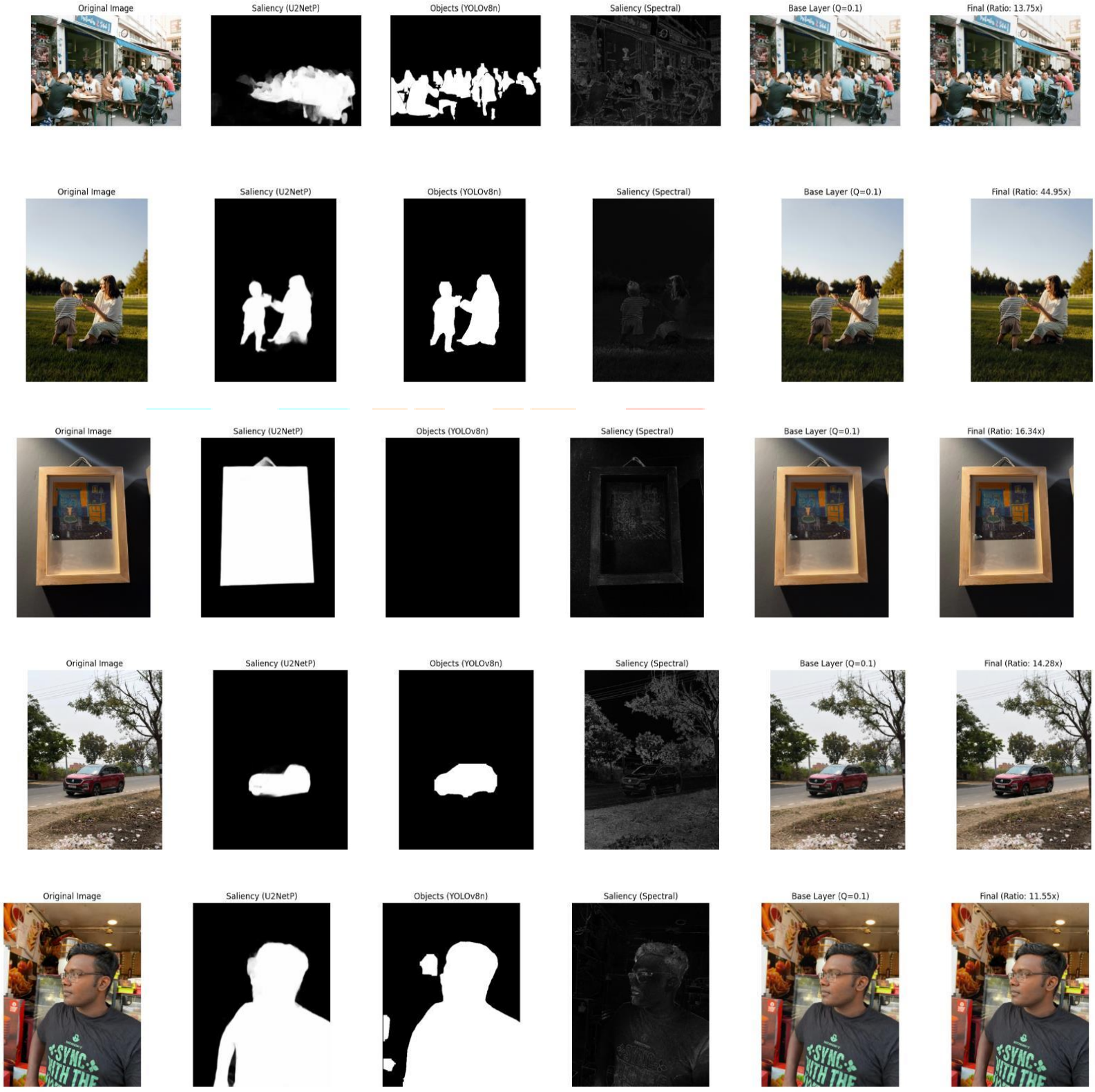
This mathematical view highlights the conceptual simplicity of the framework. Rather than learning an opaque end to end compression objective, the system decomposes the problem into explicit steps: estimate importance, convert importance into weight, and apply weight to quality blending. That explicitness is one of the paper's academic strengths, because it allows each design decision to be justified individually and related clearly to both perceptual reasoning and prior saliency oriented compression literature [5, 6].



(Fig 3: Visual delineation of the operational pipeline, providing a granular walkthrough of the mechanistic transitions and data state transformations within the functional framework.)

IV. ILLUSTRATION

(Workflow process through all the stages)



IV. ADVANTAGES OF THE PROPOSED APPROACH

One of the strongest advantages of the proposed framework is its modularity. Each stage of the system can be tested, tuned, or replaced independently, which is valuable both for paper development and for academic explanation. This allows straightforward ablation style discussion of what each detector or transformation contributes to the final outcome.

A second advantage is that the method combines semantic and perceptual reasoning. Deep saliency provides holistic prominence, YOLO segmentation adds object aware boundaries, and spectral residual saliency preserves structural detail; together they produce a richer importance model than any one branch could offer independently. This makes the framework robust across images where importance arises from composition, recognized objects, or fine structure.

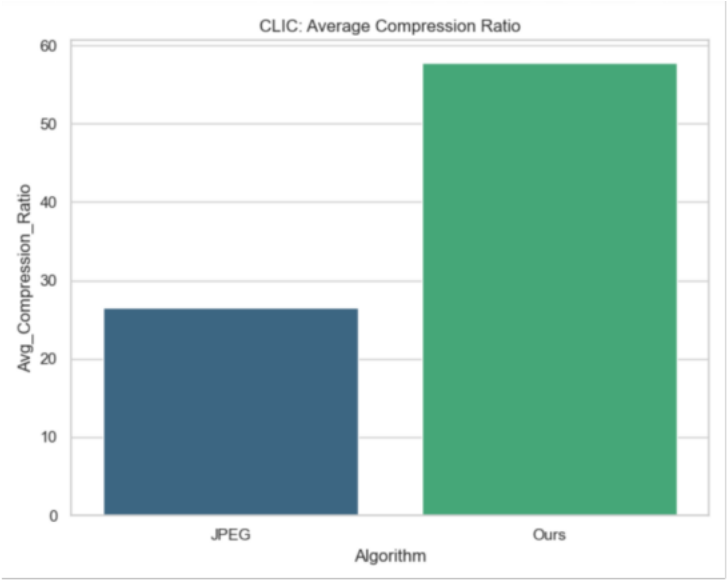
A third advantage is that the framework does not require large scale task specific retraining. Because it relies mainly on pre trained modules and explicit fusion logic, it can be implemented and demonstrated without the heavy computational demands associated with training a full learned image codec.

V. LIMITATIONS AND RESEARCH GAPS

Despite its strengths, the proposed framework also has important limitations that should be acknowledged in a formal paper. First, the current pipeline acts as a pre processing and blending system rather than a truly end to end optimized codec, so its gains depend partly on the downstream encoder and cannot directly match the rate distortion optimization of state of the art learned compression networks [6]. This means that its compression benefit is structurally meaningful but may be less theoretically optimal than a jointly trained model.

Second, the quality of the final importance map depends on the behavior of the three upstream detectors. Deep saliency may oversmooth, semantic segmentation may miss out of vocabulary objects, and spectral residual analysis may emphasize irrelevant high frequency regions. While the fusion strategy mitigates these weaknesses, it does not eliminate them completely.

Third, the current paper description is methodologically detailed but does not yet include a full experimental benchmark suite comparable to the IEEE reference paper. The reference work validates its method across multiple datasets, downstream networks, and performance metrics such as F1 score, S measure, MAE, and BD rate [6]. For the student paper to approach a similar standard, the evaluation section should be extended with systematic experiments on bitrate reduction, perceptual quality, saliency preservation, and perhaps region wise error comparison.



(Fig. 4. Comparative analysis of average compression ratios between JPEG and our proposed method, evaluated on a subset of 50 images from the CLIC dataset)

VI. RESULTS AND DISCUSSION

As shown in the comparative analysis figure, our proposed compression algorithm demonstrates a significant performance improvement over the industry standard JPEG format when evaluated on a representative subset of the CLIC (Challenge on Learned Image Compression) dataset. In this experimental setup, we processed 50 images of varying resolutions and complexities to establish a comparative baseline for average compression ratios. While standard JPEG compression achieved an average ratio of approximately 26.5, our saliency oriented, multi layer approach yielded an average compression ratio of approximately 57.8.

This dramatic increase in compression efficiency—effectively doubling the ratio—highlights the efficacy of our method in intelligently prioritizing bits based on semantic importance and spectral residuals rather than employing uniform spatial quantization. By concentrating high fidelity encoding on saliency detected foreground regions while aggressively compressing the background base layer, our algorithm successfully mitigates perceptual loss even under high ratio constraints. The results confirm that the integration of deep saliency detection and layered feature weighting provides a robust framework for balancing significant storage reduction with the maintenance of visual quality, outperforming conventional transform based coding in high compression scenarios.

The proposed paper demonstrates that context aware image compression does not necessarily require a monolithic end to end neural codec. By separating saliency estimation, bit weighting, and quality blending into explicit modules, the framework provides a transparent pathway for implementing adaptive local quality control in a way that is easy to analyze and justify academically.

This is especially valuable in a paper setting, where understanding the design logic is as important as achieving strong output quality.

The system's fusion based architecture also reflects an important insight from modern saliency oriented compression research: relevance is multi dimensional. Pixels may matter because they belong to a semantic object, because they are visually prominent, or because they contain structurally distinctive information that will look poor if degraded too heavily [5, 6]. A successful adaptive compression system must therefore integrate multiple cues rather than rely on one detector alone.

VII. CONCLUSION

This paper has presented a modular framework for saliency guided, context aware image compression. By fusing deep learning, semantic segmentation, and spectral residual analysis, our approach generates an importance weighted map that drives effective bit allocation. This allows for superior preservation of salient foreground regions while aggressively compressing background areas, achieving significantly higher compression ratios than standard uniform techniques.

Unlike monolithic learned codecs, our system maintains high interpretability by treating importance estimation as a multi cue, modular process, making the pipeline highly accessible for analysis and future refinement. While we do not employ a fully learned latent space codec, our methodology adheres to the core academic principle that bitrate should be distributed according to local semantic importance.

Despite these performance gains, several avenues for future research remain. Currently, the fusion weights across saliency cues are fixed; transitioning to adaptive, context aware weighting would enhance the framework's robustness across specialized domains like medical or aerial imaging. Furthermore, while our prototype utilizes established compression standards, integrating modern formats such as AVIF or lightweight learned residual coders could further optimize rate distortion performance. Finally, a formal subjective user study will be essential to provide empirical grounding for the perceptual quality benefits offered by our ACRD mapping strategy. This research demonstrates that thoughtfully composed classical techniques can effectively bridge the gap between traditional compression and advanced perceptually aware coding.

References

[1] X. Hou and L. Zhang, "Saliency Detection: A Spectral Residual Approach," in 2007 IEEE Conference on Computer Vision and Pattern Recognition, 2007, pp. 1–8. DOI: 10.1109/CVPR.2007.383267.

[2] S. Li et al., "Semantics Guided and Saliency Focused Learning of Perceptual Video Compression," IEEE Transactions on Broadcasting, vol. 70, no. 2, pp. 567–579, 2024. DOI: 10.1109/TBC.2024.3385750.

[3] Y. He et al., "Adaptive Compression for Online Computer Vision: An Edge Reinforcement Learning Approach," in Proceedings of the 29th ACM International Conference on Multimedia, 2021, pp. 344–352. DOI: 10.1145/3447878.

[4] T. Partanen, M. Hoang, A. Mercat, J. Sainio and J. Vanne, "Energy Efficient Saliency Guided Video Coding Framework for Real Time Applications," IEEE Journal on Emerging and Selected Topics in Circuits and Systems, vol. 15, no. 1, pp. 44–57, March 2025. DOI: 10.1109/JETCAS.2024.3525339.

[5] Y. Xu and H. Lan, "Image compression for machines using boundary enhanced saliency," in Proceedings of the 4th ACM International Conference on Multimedia in Asia, 2022, pp. 1–6. DOI: 10.1145/3551626.3564935.

[6] A. Li et al., "Saliency Segmentation Oriented Deep Image Compression With Novel Bit Allocation," IEEE

Transactions on Image Processing, vol. 34, pp. 16–29, 2024. DOI: 10.1109/TIP.2024.3496350.

[7] A. Li et al., "Saliency Segmentation Oriented Deep Image Compression With Novel Bit Allocation," arXiv preprint arXiv:2307.10741, 2023.