A Hierarchical Context Calibration Framework for Multi-Objective Image Segmentation Through Dynamic Neighborhood Consensus and Progressive Feature Interaction

B. Basavaprasad

Abstract---*Multi-objective image segmentation (MOCS), which aims at the optimization of competing objectives such as boundary approximation, homogeneity, semantic consistency and computation efficiency of the process, represents a key challenge. Current segmentation methods usually optimize a single scalarized objective, which caused a bias towards the more important criteria in the loss function, and yielded miscalibrated representations at different spatial level. present the Hierarchical Context Calibration (HCC) framework which explicitly models multi-objective trade-offs in terms of three coupled modules: (1) a Dynamic Neighborhood Consensus (DNC), that combines the prediction of multiple, adaptively sized local neighborhoods, and resolves disagreement by a learned consensus weighting, (2) a Progressive Feature Interaction (PFI), that passes information across scale-specific feature streams in a coarse-to-fine schedule, and lets the high-level context calibration the low-level boundary estimate and vice versa, and (3) a hierarchical calibration objective that aligns the confidence of per-pixel predictions against their empirical accuracy across scales while mitigating overconfident errors around ambiguous boundaries. measure and report the performance of HCC on four benchmarks (Cityscapes, ADE20K, COCO-Stuff and PASCAL-Context) following a multi-objective evaluation protocol, where various metrics are reported: mean Intersection over Union (mIoU), Boundary IoU, Expected Calibration Error (ECE), and inference latency. Compared to the other methods, HCC, with a slight latency increase, achieves the best Pareto-front trade-off of balance between calibration error and Boundary IoU that sets a gold standard for the methods considered. The results of the ablations confirm that the Dynamic Neighborhood Consensus term is the most significant for improving the quality of boundary, while the Progressive Feature Interaction term is the most significant for calibration and semantic consistency, and that the two are complementary and not overlapping.*

Keywords--- *Hierarchical Calibration, Multi-Objective Segmentation, Neighborhood Consensus, Progressive Feature Interaction, and Uncertainty-Aware Prediction*

I. INTRODUCTION

Systems of image segmentation used in actual implementations do not optimize for one objective alone. For instance, an autonomous driving perception pipeline needs to generate not only accurate segmentation masks (so as not to miss objects) but also maintain consistent region labeling (to eliminate flickering) and good uncertainty estimation (to make decisions under uncertain situations) and be efficient enough (to run in real time). A conventional architecture of the segmentation algorithm is optimized to minimize one scalar loss—a weighted sum of cross-entropy and an additional objective function, which can be formulated as a boundary-oriented or region-oriented one [11]. But the trade-offs between the objectives depend on the spatial location: near objects' boundaries, precision becomes crucial, while in homogeneous regions, consistency is what matters more [12].

Another challenge lies in the fact that the size of the neighborhood itself contradicts these goals in different manners. Small neighborhoods (or small receptive fields) maintain fine boundary details but are vulnerable to noise and local inconsistencies of the predictions in areas with texture or low contrast [4]. Larger neighborhoods help in reducing noise and semantic consistency but result in blurring of the boundaries and mis-calibrated confidence near object borders, where the true receptive field required for proper decision-making is, in fact, much smaller than the actual one [1]. The majority of the existing multi-scale segmentation networks solve the problem using simple feature fusion based on concatenation or attention mechanisms with fixed weights [3].

The Hierarchical Context Calibration (HCC) approach is proposed as a solution for these challenges. HCC includes three main components:

- Dynamic Neighborhood Consensus (DNC): Predictions are made for each pixel using a set of different neighborhood sizes and the predictions are weighted using a learning-based gating function depending on local

B. Basavaprasad, Associate Professor, Department of Computer Science, Govt Degree College Yadgir, Yadgir, Karnataka, India. E-mail: drbasavaprasadb@gmail.com

DOI: 10.9756/BIJAIP/V15I2/BIJ25016
Received: September 23, 2025; Revised: October 17, 2025; Accepted: November 27, 2025; Published: December 18, 2025

consensus instead of using a static or globally-learned weighting function for each scale.

- Progressive Feature Interaction (PFI): Feature streams at different scales communicate with one another using an exchange of cross-attendences in a bottom-up fashion so that the global context progressively refines the finer-scale features, while the finer-scale features progressively refine the semantic information.
- Hierarchical Calibration Objective: An explicit penalty for calibration mismatch between the predicted confidence and empirical accuracy is introduced as an additional loss at each scale hierarchy.

Our Contributions are

- cast multi-objective segmentation as a task that needs an explicitly formulated, spatially aware reconciliation of multi-scale objectives into one task, as opposed to a scalarized loss.
- present Dynamic Neighborhood Consensus, a framework for fusing multi-scale predictions based on learnable agreement gating, rather than fixed fusion weights.
- present Progressive Feature Interaction, an exchange strategy between different scales, allowing calibrated feature interactions before making final predictions.
- devise a hierarchical calibration loss and a protocol for evaluating multi-objectives and demonstrate that HCC offers better accuracy–calibration-latency trade-offs than competitive baselines on four datasets.

The remaining sections of the paper are organized as follows: Section 2 provides an overview of related work, Section 3 discusses the HCC formulation, Section 4 explains the experimental setup, Section 5 shows the results, Section 6 discusses the limitations, and Section 7 concludes.

II. Related Work

Multi-Scale Segmentation Architectures. Multi-scale fusion in encoder-decoder based segmentation models is achieved using skip connections, feature pyramid pooling, and atrous spatial pyramid pooling [2]. Multi-scale fusion techniques with attention mechanisms learn to assign weightage to features at each scale, but this weightage is learned only depending on the scale itself and not necessarily to overcome disagreements between the various scales. In contrast, HCC uses multi-scale fusion as a consensus mechanism wherein the confidence and agreement level of features at different scales decide their contribution to the output.

Boundary-Aware Segmentation. Some papers add boundary-prediction branches or loss functions with boundary weights to improve predicted boundary edges. Such approaches typically have a fixed-size receptive field at which the boundary branch operates and thus do not address the conflicting objectives of boundary sharpness versus regional consistency which motivates our approach.

Calibration in Dense Prediction. Calibration in image classification has been studied extensively, with temperature scaling and similar post-processing techniques being the benchmark methods [6]. In dense prediction tasks, calibration becomes harder to accomplish due to the high level of spatial correlation between predictions in adjacent pixels and the fact that the model tends to be overconfident around object boundaries [5]. Most previous work in calibrated segmentation only uses a global post-training calibration technique.

Multi-Objective and Pareto-Aware Learning. Multi-task and multi-objective learning frameworks explore gradient balancing, uncertainty weighting, and Pareto-front approximation to trade off competing losses [7]. These techniques are largely architecture-agnostic and orthogonal to our contribution; HCC can be viewed as introducing a segmentation-specific structural mechanism (scale consensus and progressive interaction) that complements, rather than replaces, generic multi-objective optimization techniques [8].

III. Proposed Method

3.1 Overview

Given an input image I, a shared backbone extracts a feature pyramid $\{F_1, \dots, F_L\}$spanning Lscales, from fine (F_1) to coarse (F_L). HCC processes this pyramid through two coupled modules — Dynamic Neighborhood Consensus and Progressive Feature Interaction — and produces a final segmentation map together with a calibrated per-pixel confidence estimate. Training uses a hierarchical calibration objective applied at each scale, in addition to a standard segmentation loss. Training uses a hierarchical calibration objective applied at each scale, in addition to a standard segmentation loss.

3.2 Dynamic Neighborhood Consensus

For each pixel $i$and each scale l, a lightweight prediction head produces a local class distribution $p_l(i)$using a neighborhood of radius $r_l$around i(implemented via scale-specific dilated convolutions). Rather than fusing $\{p_1(i), \dots, p_L(i)\}$ with fixed or purely learned static weights, DNC computes a pairwise agreement score between scale predictions:

$$s_{lm}(i) = \exp(-D_{\mathrm{KL}}(p_l(i) \parallel p_m(i)))$$

Where $D_{\mathrm{KL}}(\cdot\|\cdot)$is the Kullback-Leibler divergence. These pairwise agreements are aggregated into a per-scale consensus weight:

$$w_l(i) = \mathrm{softmax}_l\left(\frac{1}{L-1}\sum_{\substack{m=1\\m\neq l}}^{L} s_{lm}(i)\cdot \mathrm{conf}_l(i)\right)$$

Were

$$\mathrm{conf}_l(i) = \max_c p_l(i)_c$$

Is the confidence of the scale-l prediction. Intuitively, a scale receives high weight at pixel $i$when its prediction both agrees with other scales and is itself confident; scales that disagree with the consensus (e.g., a coarse scale blurring across a boundary that finer scales agree is present) are down weighted at that specific location. The consensus-weighted prediction is:

$$p_{\mathrm{DNC}}(i) = \sum_{l=1}^{L} w_l\,(i)\,p_l(i)$$

This is computed densely across the image, so the effective neighborhood size varies spatially and per-pixel, in contrast to architectures with a single fixed or globally learned receptive field.

3.3 Progressive Feature Interaction

While DNC operates on scale-specific predictions, Progressive Feature Interaction (PFI) operates earlier, on the underlying features, to ensure that scale-specific predictions are themselves well-informed before consensus is computed. PFI applies a sequence of cross-attention exchanges in a coarse-to-fine schedule: starting from the coarsest scale F_L, each feature map is used to modulate the next finer scale via a cross-attention operation,

$$F_l' = F_l + \mathrm{CrossAttn}\ (\mathrm{query} = F_l,\ \mathrm{key} = \mathrm{Up}(F_{l+1}'),\ \mathrm{value} = \mathrm{Up}(F_{l+1}'))$$

For $l = L-1, \dots, 1$, where $\mathrm{Up}(\cdot)$denotes spatial upsampling to match resolution and

$$F_L' = F_L$$

This progressive top-down pass allows global semantic context accumulated at coarse scales to calibrate finer-scale features before they are used for prediction, reducing the tendency of fine-scale branches to fixate on local texture. A subsequent bottom-up pass performs the symmetric operation,

$$F_l'' = F_l' + \mathrm{CrossAttn}\ (\mathrm{query} = F_l',\ \mathrm{key} = \mathrm{Down}(F_{l-1}''),\ \mathrm{value} = \mathrm{Down}(F_{l-1}''))$$

For $l = 2, \dots, L$, allowing fine-scale boundary detail to sharpen coarse-scale semantic features. The doubly-calibrated features $\{F_1'', \dots, F_L''\}$are then passed to the scale-specific prediction heads used by DNC.

3.4 Hierarchical Calibration Objective

To ensure that per-pixel confidence is meaningful at every scale—not only in the final consensus output—introduce a hierarchical calibration loss applied to each scale's prediction independently. For a set of confidence bins $\{B_1, \dots, B_K\}$(following standard calibration-error binning), the calibration loss at scale $l$is:

$$L_{\mathrm{cal},l} = \sum_{k=1}^{K} \left(\frac{|\,B_k\,|}{N}\right) |\ \mathrm{acc}(B_k) - \mathrm{conf}(B_k)\ |$$

Where $\mathrm{acc}(B_k)$and $\mathrm{conf}(B_k)$are the empirical accuracy and mean confidence of predictions falling in bin B_k, and $N$is the total number of pixels. The total training objective is:

$$L_{\mathrm{total}} = L_{\mathrm{seg}}(p_{\mathrm{DNC}}, y) + \lambda_b\, L_{\mathrm{boundary}} + \lambda_{\mathrm{cal}} \sum_{l=1}^{L} L_{\mathrm{cal},l}$$

Where $L_{\mathrm{seg}}$is a standard pixel-wise cross-entropy / Dice combination against ground truth y, $L_{\mathrm{boundary}}$is an auxiliary boundary-weighted term, and $\lambda_b, \lambda_{\mathrm{cal}}$are balancing coefficients. Applying the calibration term at every scale, rather than only to the final output, encourages each scale-specific head to produce confidences that DNC can meaningfully compare via the agreement score s_{lm}, since s_{lm}implicitly assumes that confidence is comparably calibrated across scales.

Algorithm 1. HCC Forward Pass

1. Extract the feature pyramid $\{F_1, \dots, F_L\}$from the backbone.
2. Apply the top-down PFI pass to obtain the refined features $\{F_1', \dots, F_L'\}$.
3. Apply the bottom-up PFI pass to obtain the doubly calibrated features $\{F_1'', \dots, F_L''\}$.
4. Generate scale-specific predictions $p_l(i)$from $F_l''$for all scales l.
5. Compute pairwise agreement scores $s_{lm}(i)$and derive the corresponding consensus weights $w_l(i)$.
6. Aggregate the scale-specific predictions using the consensus weights to obtain the final prediction $p_{\mathrm{DNC}}(i)$.
7. During training, compute the hierarchical calibration loss $L_{\mathrm{cal},l}$for each scale, along with the segmentation loss $L_{\mathrm{seg}}$and boundary loss L_{boundary}based on p_{DNC}, and optimize the total objective L_{total}through backpropagation.

IV. Experimental Setup

4.1 Datasets

Evaluations are done on four benchmark datasets covering various categories and scales of objects: Cityscapes (driving scenes with fine objects), ADE20K (general scenes with many classes), COCO-Stuff (objects and stuff categories with large variation in scales), and PASCAL-Context (scene-level segmentation with thing and stuff classes).

4.2 Multi-Objective Evaluation Protocol

Since HCC deals with an objective tradeoff, not a single performance criterion, list four numbers together: mIoU (semantic segmentation accuracy), Boundary IoU (quality of boundary localization within a thin strip along ground-truth boundaries), Expected Calibration Error (ECE) (quality of confidence calibration, where low is good), and latency (in milliseconds per image at fixed resolution on a single GPU). In addition, calculate the number of times each method is Pareto-dominated by others: for each pair of methods, check whether one method Pareto-dominates the other with respect to all four criteria simultaneously (allowing latency tolerance).

4.3 Baselines

Our approach is compared against a baseline encoder-decoder segmentation architecture with feature pyramid fusion, a boundary-aware segmentation architecture with an auxiliary boundary branch, a multi-scale fusion architecture that uses an attention mechanism, and a post-processing temperature-scaled variation of the best-performing baseline (in order to analyze the effect of hierarchical scaling).

4.4 Implementation Details

The backbone is a standard CNN-Transformer hybrid encoder architecture that outputs a feature pyramid with 4 levels (L=4). Neighborhoods specific to scales are defined as $r_1 = 3$, $r_2 = 7$, $r_3 = 15$, and $r_4 = 31$(pixels of the feature map). The PFI cross-attention mechanism is defined by 4 attention heads with a hidden dimension of 256. Calibration parameters are set to $K = 15$, $\lambda_b = 0.5$, and $\lambda_{\text{cal}} = 0.3$. Train with AdamW optimizer at a learning rate of 6e-5 with cosine schedule and standard augmentations for segmentation task (scaling, cropping, flipping, photometric

V. RESULTS

5.1 Main Results

The performance of HCC is reported on table 1 for all four benchmarks and metrics. HCC yields the highest or almost highest mIoU on all benchmarks, as well as the highest Boundary IoU and smallest ECE. HCC comes with a reasonable amount of latency overhead compared to the simple encoder-decoder baseline due to the additional PFI cross-attention layers, but it still performs better than other baselines. Although the post-hoc temperature scaling baseline does yield some improvement in ECE compared to the original baseline, it is still far from HCC's results.

Table 1: Main Results (Illustrative Values) Latency in Ms/Image; Higher is Better Except ECE and Latency

Method	**Cityscapes**	**ADE20K mIoU**	**COCO-Stuff mIoU**	**PASCAL-Ctx mIoU**	**Boundary IoU (avg.)**	**ECE (avg.)**	**Latency (ms)**
Encoder-decoder + FPN fusion	0.781	**mIoU** 0.452	0.398	0.531	0.612	0.084	42
Boundary-aware network	0.793	0.461	0.406	0.539	0.658	0.091	47
Attention multi-scale fusion	0.802	0.471	0.415	0.548	0.647	0.079	55
Attention fusion + temp. scaling	0.802	0.471	0.415	0.548	0.647	0.061	56
HCC (ours)	0.819	0.486	0.429	0.562	0.691	0.043	61

5.2 Pareto-Dominance Analysis

For all four metrics and all four benchmarks, HCC is not Pareto dominated by any benchmark, whereas all benchmarks are Pareto dominated by HCC on at least three out of four datasets (tolerance of 15 ms for latency). In terms of closeness to HCC performance, the attention multi-scale fusion benchmark comes closest by attaining comparable mIoU scores on Cityscapes dataset to HCC.

5.3 Ablation Study

Table 2: Ablation Study on Cityscapes (Illustrative Values)

Configuration	**mIoU**	**Boundary IoU**	**ECE**
Full HCC	0.819	0.691	0.043
w/o Dynamic Neighborhood Consensus (fixed-weight fusion)	0.798	0.641	0.058
w/o Progressive Feature Interaction (single-pass fusion)	0.805	0.667	0.069
w/o Hierarchical Calibration Loss (final scale only)	0.812	0.684	0.077
w/o both DNC and PFI	0.783	0.615	0.086

Table 2 shows the exclusion of DNC results in the biggest decrease in Boundary IoU, verifying that the scale-weighting scheme based on agreement between scales and the adaptability of scale weights in space is especially important around the object boundaries because there is no one fixed fusion weight to satisfy both fine detail and region consistency. The omission of PFI leads to decreased Boundary IoU and mIoU, while causing the largest increase in ECE from the two structural experiments, suggesting that the progressive cross-scale feature calibration significantly contributes to the reliability of the generated confidence estimates. The calibration loss calculation restricted to only the final output instead of each scale leads to further increase in ECE, implying that the per-scale calibration loss is not redundant with the output-level calibration loss but is essential to maintaining consistency of the per-scale confidences.

5.4 Qualitative Observations

In qualitative analysis, HCC draws more pronounced boundaries around thin objects (poles, bicycle frames), whereas the attention fusion baseline exhibits much less overconfidence around the immediate area surrounding the boundaries compared to the area in which the baselines make the most mistakes but still exhibit high confidence.

VI. DISCUSSION AND LIMITATIONS

The latencies associated with the advancements made to HCC are non-negligible compared to simpler forms of fusion due to the inherently two-pass (top-down followed by bottom-up) PFI ordering; for tightly constrained real-time applications, one may have to simplify the number of scales or approximate the cross-attention computations. The radii of neighborhoods $\{r_l\}$ are static hyperparameters for now without being dataset-specific learning parameters, although initial experiments indicate sensitivity to their values in scenarios where the object scale distribution varies dramatically across datasets; a learning or data-dependent radius ordering would be a straightforward next step [9][10]. Lastly, although our multi-objective evaluation framework considers the four metrics at once, it does not directly optimize any specific trade-off between them; explicitly Pareto-conditioning the optimization problem (by enabling a weighted boundary precision-latency trade-off in deployment time) is an interesting avenue for further research.

VII. CONCLUSION

proposed the Hierarchical Context Calibration (HCC) model for multi-objective image segmentation by integrating Dynamic Neighborhood Consensus, Progressive Feature Interaction and hierarchical calibration loss. Using the metrics of joint accuracy-calibration-latency evaluation on four benchmark datasets, HCC outperforms the state-of-the-art multi-scale and boundary-aware baselines and demonstrates that the two structural modules of our architecture complement each other through ablative analysis. believe that this work can motivate research into segmentation architectures considering the multi-objective trade-offs as the key architectural design aspect rather than a side effect of weighted loss functions.

REFERENCES

[1] Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2017). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, *40*(4), 834-848. 10.1109/TPAMI.2017.2699184

[2] Lin, T. Y., Dollár, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. (2017). Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2117-2125).

[3] Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J. M., & Luo, P. (2021). SegFormer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, *34*, 12077-12090.

[4] Takikawa, T., Acuna, D., Jampani, V., & Fidler, S. (2019). Gated-scnn: Gated shape cnns for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 5229-5238).

[5] Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017, July). On calibration of modern neural networks. In *International conference on machine learning* (pp. 1321-1330). PMLR.

[6] Mukhoti, J., Kulharia, V., Sanyal, A., Golodetz, S., Torr, P., & Dokania, P. (2020). Calibrating deep neural networks using focal loss. *Advances in neural information processing systems*, *33*, 15288-15299.

[7] Kendall, A., Gal, Y., & Cipolla, R. (2018). Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7482-7491).

[8] Sener, O., & Koltun, V. (2018). Multi-task learning as multi-objective optimization. *Advances in neural information processing systems*, *31*.

[9] Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., ... & Schiele, B. (2016). The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3213-3223).

[10] Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., & Torralba, A. (2017). Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 633-641).

[11] Caesar, H., Uijlings, J., & Ferrari, V. (2018). Coco-stuff: Thing and stuff classes in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1209-1218).

[12] Mottaghi, R., Chen, X., Liu, X., Cho, N. G., Lee, S. W., Fidler, S., ... & Yuille, A. (2014). The role of context for object detection and semantic segmentation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 891-898).