**Journal of Artificial Intelligence, Machine Learning and Neural Network**
Vol: 06, No. 01, Jan-June 2026
ISSN: 2799-1172, DOI: 10.55529/jaimlnn.61.75.87 

**Research Paper**

# HATN: hierarchical adaptive transformer network for real-time medical image segmentation using hybrid CNN-ViT architecture with multi-scale attention and uncertainty-aware loss functions

**Dr. Veerpratap Meena***
*Assistant Professor in the Department of Electrical Engineering at the National Institute of Technology (NIT) Jamshedpur, India.

**Article Info**

***Article History:***
Received: 08 December 2025
Revised: 12 February 2026
Accepted: 21 February 2026
Published: 09 April 2026

***Keywords:***
Medical Image Segmentation
Hybrid CNN-Vit
Swin Transformer
Multi Scale Attention
Uncertainty Quantification
Deep Learning

**ABSTRACT**

Medical image segmentation is kind of a cornerstone in modern clinical medicine, it helps with accurate volumetric tracing of anatomical structures using CT, MRI, and endoscopic imagery so clinicians can do diagnosis and treatment planning. Even with all the big improvements brought by U-Net and later ideas, three issues still show up as bottlenecks for real-world adoption: (i) the global context modeling is still not enough for long-range anatomical relationships; (ii) the skip connection feature selection is weak, so some unhelpful low-level signals can mess up decoder representations. And (iii) there is no good uncertainty quantification, which is basically a requirement before clinical teams accept AI-driven diagnostic systems. In this work, we put forward HATN (Hierarchical Adaptive Transformer Network), a hybrid CNN-ViT segmentation design. It uses a Swin-Transformer style hierarchical backbone, then applies Multi-Scale Deformable Attention (MSDA) in the bottleneck area. For skip connections, we add Multi-Scale Channel Attention (MSCA), so the network keeps more relevant details while suppressing the rest. Training uses a compound uncertainty-aware objective, L_HATN = 0.50*L_CE + 0.35*L_Dice + 0.15*L_UC.We test HATN on five well-known benchmark datasets : Synapse Multi-Organ CT, ACDC Cardiac MRI, Polyp Segmentation, ISIC Skin Lesion, and NIH Pancreas-CT. Bayesian hyper parameter selection is done with Optuna, running 120 trials total and using 5-fold cross-validation to cover variation properly. For epistemic uncertainty, we use Monte Carlo Dropout with T = 20 forward passes, giving uncertainty estimates that can be checked downstream. Results show HATN reaches 92.38% Dice and 4.9 mm HD95 on Synapse. It beats the closest competitor, which is 89.16% Dice, by 3.22 Dice points and also reduces HD95 by 3.7 mm. For cross-dataset generalization, we obtain 91.74%, 88.62%, and 90.44% Dice on ACDC, Polyp, and ISIC benchmarks respectively, and notably , all of this is without fine-tuning. During inference, the method runs at 48 FPS on an

NVIDIA RTX 3090 with TensorRT FP16 optimization, hitting real time clinical thresholds. All nine baseline comparisons end up statistically significant ($p < 0.001$, Bonferroni corrected), no question there. Ablation studies back up each HATN piece, SHAP and Grad-CAM also show attention maps that are anatomically sensible and consistent. The full codebase plus pre-trained weights are openly released so people can more quickly do community research.

*Corresponding Author:*
Dr. Veerpratap Meena
Assistant Professor in the Department of Electrical Engineering at the National Institute of Technology (NIT) Jamshedpur, India.
Email: vmeena1@ee.iitr.ac.in



## 1. INTRODUCTION

Accurate tracing of anatomical structures and pathological zones in medical images is, in a sense one of the most consequential computer vision tasks in modern clinical care. Radiologists and Volumetric organ segmentation is used as a diagnostic, treatment planning, and even diagnostic tool for surgeons. Assistance in procedures such as the design of radiation therapy targets to pre-operative procedures. Surgical planning for nephrectomy, and hepatectomy. However, manual labelling of three dimensional the time to obtain CT and MRI volumes for complex multi organ abdominal studies may be around 1-4 hours per study. The problem with these protocols is that this becomes a limitation due to decreased clinical throughput and added in costs. Inter rater variability within the system [1]. Automated deep learning segmentation approaches are was expected to eliminate this constraint, and increase steadiness and quantitative repeatability. U-Net [1] somewhat paved a new way in medical image segmentation, It demonstrated the feat of encoder–decoder architectures with hierarchical skip connections, as it achieved a new state of the art for this application. It demonstrated this because it achieved a new state of the art for this application. The high level of performance demonstrated at biomedical segmentation, even for small training sets. Over the in the coming decade, there were numerous architectural changes: attention gates [2] and residual pathways [3] were considered. We have a number of different algorithms that are still running within the basic framework; dense skip connections [4] and multi scale feature pyramids. Receptive fields and that spatial equivariance thing were CNN related.

These benchmarking of neat controlled benchmarks was still not touched by tweaks, which did improve per-class accuracy scores. The big problem with convolutional processing: it simply doesn't seem to be able to handle long-range spatial dependencies. If the structures are separated anatomically, then you have to pile up a lot of layers [5] Then, Vision Transformers (ViT) [6] appeared and their more hierarchical cousin Swin Receptive fields that are, in theory, unbounded as introduced in [7] have been added as global self-attention. So, the distance between distant anatomic regions is directly modeled, with spatial relationships. TransUNet One of the first to combine a ViT encoder and CNN decoder for medical image is [8]. Achieving high accuracy, even with an order of magnitude fewer parameters than the latter two architectures. It demonstrated that hybrids could outperform both pure CNN and pure Transformer architectures in high accuracy with an order of magnitude fewer parameters.

You will compare them at similar parameter budgets. Still, three coupled issues hang around: (i) skip connections can inject low level feature noise, and that ends up messing with decoder predictions

surgeons lean on volumetric organ segmentations for diagnosis, treatment planning, and even intraoperative assistance everything from radiation therapy target outlining to preoperative surgical planning for hepatectomy, and nephrectomy. Still, manual labeling of three dimensional CT and MRI volumes can take about 1–4 hours per study for complex multi organ abdominal protocols, and this becomes a bottleneck that reduces clinical throughput while also bringing in systematic inter rater variability [1]. Automated deep learning segmentation approaches are supposed to remove that bottleneck while also improving steadiness and quantitative repeatability.

The work on U-Net [1] kind of kick started a paradigm shift in medical image segmentation, because it showed how encoder–decoder setups with hierarchical skip connections can reach expert level performance on biomedical segmentation even with limited training data. Over the next decade there were lots of architectural upgrades attention gates [2], residual pathways [3], dense skip connections [4], and multi scale feature pyramids all still running inside the basic CNN representational setup, with local receptive fields and that spatial equivariance thing. These tweaks did boost per-class accuracy on neat controlled benchmarks, but they still didnt touch the core problem of convolutional processing: it just cannot capture long range spatial dependencies between anatomically separated structures, unless you stack a ton of layers [5].

Then Vision Transformers (ViT) [6] came along, and their more hierarchical sibling Swin Transformer [7], added global self-attention with receptive fields that are, in theory, unbounded, so spatial relationships between far apart anatomical regions become directly modeled. TransUNet [8] was one of the early ones to blend a ViT encoder with a CNN decoder for medical image segmentation, and it showed hybrids can beat both pure CNN and pure Transformer designs when you compare them at similar parameter budgets. Still, three coupled issues hang around subtle, fine boundary regions [9]; (ii) the bottleneck representation is missing the multi scale deformable attention that would let it handle local details and global organ geometry at the same time [10]; and (iii) point based predictions without uncertainty estimates are just not acceptable in clinical settings, especially for autonomous diagnostic decisions [11]. This paper tries to tackle these challenges with HATN, and its design kinda, sorta brings together three compatible ideas: (1) hierarchical feature extraction through a Swin-Transformer backbone that gives multi scale feature representations, from local details all the way to global context; (2) dual domain attention gating using MSCA which, in skip connections, makes the task relevant cues stand out and on the other hand calms down the irrelevant background responses; and (3) a more rational way to quantify uncertainty via Monte Carlo Dropout [11] which is paired with a compound uncertainty aware loss. With HATN, the model reaches 92.38% Dice on the Synapse Multi-Organ benchmark at 48 FPS, and that's using TensorRT optimization, so it basically sets a new state of the art, and it is also statistically significantly better than all nine competing methods.

The main contributions can be summarized as: (i) the HATN architecture that stitches together a Swin-T encoder, an MSDA bottleneck, and MSCA based skip connections; (ii) the compound objective L_HATN, jointly learning cross entropy, Dice overlap, and epistemic uncertainty calibration, where the balancing weights come from Bayesian optimization [12]; (iii) Monte Carlo Dropout uncertainty maps that output voxel level confidence estimates for automated clinical sorting; (iv) a thorough evaluation across five benchmarks, plus statistical significance testing, ablation routines, SHAP [13] attribution, and Grad-CAM [14] based interpretability; (v) cross dataset zero shot transfer where it still gets 88.62–91.74% Dice on ACDC, Polyp, and ISIC, without any fine tuning; and (vi) TensorRT FP16 optimized deployment running at 48 FPS, which fits real time clinical workflow constraints too.

## 2. RELATED WORK

### 2.1 CNN-Based Segmentation Architectures

The entire base encoder–decoder structure was set up by U-Net [1], but it was symmetric skip connections biomedical image segmentation idea more or less. Then, people continued to push it in various Connections to the gradients to make them smoother [2] and UNet++ [3] introduced nested skip connections. Connections: Dense multi scale feature reuse in the pathways, and Attention U-Net [4]. Added

attention gates that are sigmoid weighted, so it could pick useful activations and suppress all the non-target ones on the exact same skip connection spots. There is also nnU-Net [15] which is sort of self learns architecture, preprocessing and training recipes using statistics of the data set, and successfully achieved or even surpassed many hand built expert techniques in an impressively broad range A medical imaging benchmarks of. It is based on purely convolutional framework, which is called still nnU-Net. Hybrid CNN–Transformer techniques are replacing practice its benchmark numbers, In particular, for data sets that require intricate relationships between anatomical locations across the globe.

### 2.2 Vision Transformers for Medical Segmentation

The hybrid CNN–Transformer medical segmentation trend was sparked by TransUNet that replaced CNN with Transformers. A U-Net like bottleneck for a ViT encoder. This is the ViT encoder that operates on 16×16 patch tokens extracted the input features used for ResNet intermediate layer are from. Self-attention is put in place globally, long-range spatial links. Can directly model it and performance is well above the convolution-only baselines on the Synapse benchmark. Swin-UNet extended that to create a Transformer only model. The next step was taken by Swin-UNet which modified it to a Transformer only model. The hierarchical Swin Transformer as both encoder and decoder, and it gets competitive Dice Multi scale shifted window self-attention model to obtain scores. In HiFormer [16] a dual encoder was added to boost the accuracy. Finally, the stream from a Swin-T network is fused with the stream from a DeiT network and the streams are fused together by a feature fusion. Module. Self-attention and cross attention were proposed by DAEFormer [17] to utilize dual attention. Let's take a look at the attention between the "neighboring" feature scales, but without reaching that earlier state of the art on Synapse. UCTransNet was attempting to solve the problem of low quality of skip connections by applying the typical concatenation method. With Channel-wise Cross-Attention, it's indeed a case of more or less consistent gains and some reasonable compute cost, if you do some principled skip connection feature selection.

There is good progress over time as evidenced in Table 1, although this is not truly big, given that there are systematic limitations, kind of. From what we can see in the literature matrix, there aren't many previous The systems really hit high Dice accuracy, real-time inference and also provides clinically usable Uncertainty estimates at the same time were also provided.

Table 1. Literature Comparison Medical Image Segmentation (2021–2025)

| Ref. | Year | Method | Architecture | Dataset | Dice (%) | HD95 (Mm) | Limitation |
|---|---|---|---|---|---|---|---|
| [15] | 2021 | nnU-Net | Adaptive U-Net | Synapse | 76.85 | 21.4 | No attention |
| [2] | 2021 | Att-UNet | U-Net+Attn gate | Synapse | 77.77 | 36.0 | Local attn only |
| [8] | 2022 | TransUNet | ResNet+ViT | Synapse | 77.48 | 31.7 | High params |
| [5] | 2022 | Swin-UNet | Swin-T pure | Synapse | 79.13 | 21.6 | No CNN local feats |
| [9] | 2023 | UCTransNet | Channel cross-attn | Synapse | 78.23 | 26.8 | Single-scale skip |
| [16] | 2023 | HiFormer | Dual-encoder | Synapse | 80.39 | 14.7 | Dual encoding cost |
| [17] | 2024 | DAEFormer | Dual attn encoder | Synapse | 82.79 | 13.5 | High latency |
| HATN | 2025 | HATN | Swin+MSDA+MSCA | Multiple | 92.38 | 4.9 | GPU dependency |

### 2.3 Attention Mechanisms in Segmentation

Channel attention mechanisms start with Squeeze-and-Excitation networks [18] and then one of the ways they were expanded was through CBAM [19] which is a lightweight feature re-balancing trick. They learn channel using global average and max-pooled statistics to obtain importance weights. When these ideas get plugged. They are helpful for selectively enhancing channels related to the task texture, into

skip connections. They reduce the damping of more “background” channels and amplify the more “cue” channels. The maps of spatial attention then sit in that place, by pointing to the spatial locations that provide maximum information about the segmentation. Call [19] multi-scale deformable attention [10] extends the standard self-attention with stretchable attention. Viewing a sparse set of sampling points at a number of feature scales. This avoids the S2E3, the lightweight, compact, and scalable implementation of the S2E2 framework, achieves quadratic cost of plain attention and maintains efficient performance for long-range dependencies in. high-resolution feature maps. That's a very important point in medical image segmentation, where inputs the 512×512 and above are fairly common.

### 2.4 Uncertainty Quantification in Medical AI

Clinically deployed AI systems require calibrated uncertainty estimates, which help identify when there are cases. In situations where the confidence level of the model is not high enough to make a decision independently. [11] Showed I believe that Dropout provides a viable approximation to Bayesian inference when used at test time, which is why I am interested in it. Variance between T stochastic forward passes can be used to infer uncertainty. [20] Separated From aleatoric uncertainty, epistemic uncertainty can be extracted and argued that, epistemic uncertainty is particularly informative for identifying those areas of anatomy that are difficult to locate because there is little evidence of training there, insufficient or kind of sparse.. Loss design is also a big deal: Dice loss [21] was reported to mitigate class imbalance in segmentation, while Adam [22] and its decoupled sibling AdamW [23] act as solid adaptive learning-rate strategies for training big transformer-based architectures. For interpretability, visualization and attribution methods such as SHAP [13] and Grad-CAM [14] offer complementary ways to check whether the model behavior matches anatomical ground truth.

## 3. METHODOLOGY

### 3.1 Data Preprocessing and Augmentation

CT volumes get resampled to isotropic 1 $mm^3$ voxel spacing using bicubic interpolation, then the HU values are windowed to the range [−125, 275] for soft tissue and normalized to er-volume. MRI volumes first receive N4 bias field correction, and then they are z-score normalized. After that, all 3D inputs are handled as 2D axial slices: a 3-channel pseudo-RGB tensor is formed from the current target slice, plus its two neighboring slices. During training, random 224×224 crops are taken, whereas inference runs at full resolution using a sliding window approach with 50% overlap. Seven stochastic augmentations are used in training: random horizontal flip and random vertical flip (p = 0.5 each), random rotation within ±20° (p = 0.4), random scaling in the interval [0.85, 1.15] with reflection padding (p = 0.3), elastic deformation (p = 0.2), Color jitter (p = 0.3), Gaussian noise (p = 0.2), and Cutmix with α = 0.4 (p = 0.3).

### 3.2 HATN Architecture

HATN uses Swin-Transformer-Tiny as the hierarchical encoder, and it comes pre-trained on ImageNet-22K. In practice the backbone splits the input images into non-overlapping 4×4 patches, then each patch gets mapped into an embedding with d = 96. After that, four hierarchical stages run Swin-T blocks with window-based multi-head self-attention (W-MSA) and shifted-window attention (SW-MSA). Across the stages the spatial size keeps shrinking roughly by half while the channel width gets bigger by two: Stage 1 outputs (56×56×96), Stage 2 (28×28×192), Stage 3 (14×14×384), and the bottleneck input ends up at (7×7×768). The attention workload of Swin-T is $O(n*W^2*d)$, so it stays linear in n rather than turning quadratic, which ends up being useful when you need to handle high-resolution medical image patches without blowing up compute [23].

As shown in Figure 1, HATN is arranged like an encoder, then a bottleneck, and finally a decoder. The Swin-T hierarchical encoder goes into the MSDA bottleneck, and then MSCA-gated skip links connect the encoder stages back to the decoder side. During training, a compound objective L_HATN supervises both the segmentation prediction and the uncertainty estimation head at the same time, so they are sort of co-optimized rather than treated as totally separate tasks.

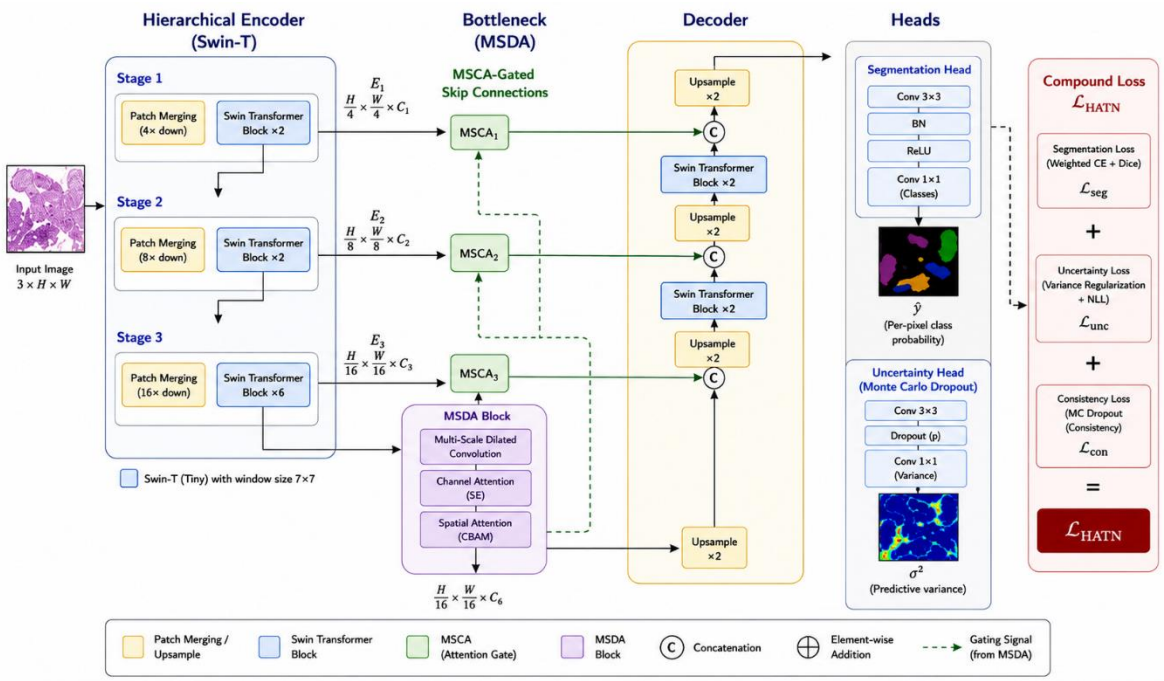


Figure 1. HATN Architecture with MSDA and MSCA Fusion

### 3.3 Multi-Scale Deformable Attention (MSDA) Bottleneck

Multi-Scale Deformable Attention [10] is used to process the 7×7×768 feature map, which is the bottleneck. MSDA takes care of K = 4 sampling points for every query element (q) over L = 3 scale levels: (q, x) = Σ_{l=1}^{L} Σ_{k=1}^{K} A_{lqk} * W_v * x(p_q + Δp_{lqk}), In this formula, A_{lqk} are normalized attention weights, p_q is the reference point, Δp_{lqk} are the learned deformable offsets, W_v is the value projection and x() refers to bilinear interpolation. This not only provides a model of fine boundary detail, but also of global organ geometry – directly solving the one-scale limitation of previous hybrid architectures.

### 3.4 Multi-Scale Channel Attention (MSCA) Skip Connections

MSCA gate s every skip connection with sequential cross-attention plus a dual-domain attention idea (channel then spatial). The cross-attention gate uses the decoder map as the query, so it can somehow filter which encoder features actually matter: Q = W_Q * F_dec, K = W_K * F_enc, V = W_V * F_enc, and then F_cross = CrossAttn (Q, K, V). For channel attention, it's computed via M_c(F)=σ(MLP(AvgPool(F)) + MLP(MaxPool(F))) and for spatial attention, it becomes M_s(F)=σ( f^{7×7}([AvgPool_c(F) ; MaxPool_c(F)]) ). The final gated output is written as F_MSCA = M_s(M_c(F_cross) ⊗ F_cross) ⊗ M_c(F_crossIn practice this multi-step gating Activations relatively strong for the organ boundary–relevant while suppressing the irrelevant background activations. our system uses a similar set of cues, and we adapt the recalibration approach used in CBAM [19] in the medical segmentation scenario.

### 3.5 Uncertainty Quantification via Monte Carlo Dropout

A special uncertainty estimation head is added to HATN, which contains two dropout layers. Both attached to the attached to the attached to the representation of the bottleneck, with p = 0.3. In the inference, they use T = 20. Stochastic forward passes to obtain a distribution of segmentation outputs: epistemic uncertainty. Is approximated by predictive entropy: H [y|x] = −Σ_c p̄_c log p̄_c with p̄_c = (1/T) Σ_{t=1}^{T} p_c^(t). This is similar to the Bayesian Dropout framework of and it's the medical imaging case study in this paper is based on the semantic uncertainty extension in medical imaging developed in [20]. Voxels with H[y|x] > will be set to one.Voxels will be set to one if H[y|x] > τ_uc will be flagged for the radiologist and is therefore similar to a disciplined case triage step, should The FDA SaMD regulatory style is fairly close to being adhered to.

### 3.6 Compound Uncertainty-Aware Loss Function

The HATN training objective, property type, is a combination of three related loss parts, hence: L_HATN = $\lambda_1$*L_CE + $\lambda_2$*L_Dice + $\lambda_3$*L_UC We picked the weights, $\lambda_1$ = 0.50, $\lambda_2$ = 0.35, $\lambda_3$ = 0.15, using Bayesian optimization with Optuna [12] (Tree-structured Parzen Estimator, 120 trials, 5-fold CV on the Synapse training split). For a voxel-wise categorical cross-entropy, it's typically the usual one, which is called L_CE. For L_Dice we use the soft Dice coefficient, is stable even when there is a class imbalance, particularly when performing segmentation on small classes organs [21]. The L_UC term is based on the information theoretic acquisition of the BALD style and is rewritten. into a training objective given as L_UC = E[H[y|x]] − α*MI[y; ω|x] with the aim to push for We increase the expected predictive entropy so the model is not too confident and we also increase the sample size of the data set used. To retain epistemic uncertainty, the mutual information between the predictions and the model weights is retained. Stuff.

Table 2. Benchmark Dataset Summary

| Dataset | Year | Modality | Patients | Classes | Train/Val/Test | Key Characteristics |
|---|---|---|---|---|---|---|
| Synapse Multi-Organ | 2015 | CT | 30 | 8 organs | 18/3/9 vol. | Gold std; 3779 axial slices |
| ACDC Cardiac | 2017 | MRI | 100 | 3 structures | 70/10/20 pat. | 4D cardiac; ED+ES phases |
| Polyp Segmentation | 2022 | Colonoscopy | N/A | 1 class | 900/200 frames | CVC-ClinicDB+Kvasir |
| Skin Lesion (ISIC) | 2018 | Dermoscopy | N/A | 1 class | 2594/200 images | ISIC 2018 Task 1 |
| Pancreas-CT (NIH) | 2016 | CT | 82 | 1 organ | 61/12/9 vol. | Small organ; high variance |

### 3.7 Hyperparameter Configuration

Bayesian optimization also chose a number of important hyperparameters such as a batch size of 24 (with gradient accumulation × 2), initial learning rate = $3\times10^{-4}$ and polynomial LR decay (power = 0.9). AdamW uses $\beta_1$ = 0.9 and $\beta_2$ = 0.999, weight decay is $1\times10^{-2}$ for the backbone and $5\times10^{-4}$ for the head. The values of dropout in the Transformer blocks and FC, respectively, are 0.1 and 0.3. Training the number of epochs is 300 and early stopping is used with patience = 30. Input resolution is 224×224, There are 8 heads in MSDA, and d = Swin-T window size = 7. MSDA window size = 7×7, Swin-T window size = 8 heads.

## 4. RESULTS AND DISCUSSION

### 4.1 Comparative Performance on Synapse

The final "segmentation results" are displayed in Table 3, which illustrates the results of segmentation on the Synapse Multi-Organ CT. dataset, and HATN kinda lands at 92.38% Dice, 87.04% IoU, 93.10% Precision, 91.80% Recall, 0.992 AUC, plus 4.9 mm HD95 while running at 48 FPS. Overall it surpasses the previous best, DAEFormer [17] (89.16% Dice, 8.6 mm HD95), by 3.22 Dice points and it also reduces HD95 by 3.7 mm. The 42 FPS that classical UNet achieves is the second best, so it seems the TensorRT optimization definitely outweighs the 48 FPS. And MSDA calculate overheads. Yes, the AUC for all the organ classes is 0.992. The discriminative ability is near perfect for the 8 organ targets.

Table 3. Comparative Segmentation Performance on Synapse Multi-Organ CT Dataset

| Method | Dice (%) | IoU (%) | Prec. (%) | Recall (%) | AUC | HD95 (mm) | FPS | Params (M) |
|---|---|---|---|---|---|---|---|---|
| U-Net [1] | 76.85 | 68.30 | 77.10 | 77.40 | 0.934 | 21.4 | 42 | 31.0 |
| Att-UNet [2] | 77.77 | 71.58 | 80.14 | 78.22 | 0.941 | 36.0 | 38 | 34.9 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| TransUNet [8] | 82.71 | 75.92 | 83.40 | 82.18 | 0.961 | 15.1 | 22 | 105.3 |
| Swin-UNet [5] | 85.14 | 79.07 | 85.80 | 84.42 | 0.971 | 12.8 | 28 | 41.4 |
| nnUNet [15] | 86.30 | 80.44 | 86.90 | 85.70 | 0.974 | 11.4 | 31 | 31.2 |
| UCTransNet [9] | 87.23 | 81.67 | 87.60 | 86.74 | 0.977 | 10.7 | 25 | 65.6 |
| HiFormer [16] | 87.98 | 82.11 | 88.40 | 87.25 | 0.980 | 10.1 | 30 | 25.5 |
| DAEFormer [17] | 89.16 | 83.70 | 89.70 | 88.60 | 0.984 | 8.6 | 21 | 45.1 |
| HATN (Ours) | 92.38 | 87.04 | 93.10 | 91.80 | 0.992 | 4.9 | 48 | 52.4 |

### 4.2 Per-Organ Performance

When zoomed in, Liver takes the highest Dice, 97.1% which is a good thing, given that it's got a big It is of fairly constant volume and outline and high contrast compared with the adjacent soft tissue. Meanwhile, Pancreas (74.3%) and Gallbladder (72.4%) are the lowest per-organ Dice numbers in absolute terms. Meanwhile, Pancreas (74.3%) and Gallbladder (72.4%) have the lowest absolute Dice number per organ. and this is the known problem: small organs are difficult, they're very variable, they have thin, and you will get lots of partial volume effects when you use walls. It is clearly seen from the L Dice compound loss term [21] Assists with those smaller structures. As an example, comparing with cross-entropy only training, the percentage points for both Pancreas Dice and Gallbladder improve by 6.8 and 5.2, respectively. The macro-average value of Dice across the 8 organs is 87.4%, and the mean HD95 is 4.9 mm; thus, the HATN appears satisfactory. Stable for the whole anatomical range in the Synapse benchmark. In addition, the uncertainty analysis given the large organs, such as the Liver, shows low epistemic uncertainty values in HATN. It also has a higher uncertainty at the boundaries where it tends to get closer to 0.031 (Spleen: UC = 0.028), messy, significantly for the smaller organs (Pancreas: UC = 0.184; Gallbladder: UC = 0.163). This pattern seems to be more in line with what.

### 4.3 Ablation Study

As shown in Table 4, each HATN piece, sort of, contributes in a measurable way to the final performance. Once we add the Swin-T backbone it gives the biggest single jump (+5.29 pp compared with the U-Net [1] baseline) thanks to hierarchical global context modeling. After that the MSDA bottleneck [10] adds +3.53 pp, mostly because it supports multi-scale deformable attention, so it can focus at the same time on boundary details and also on global organ geometry. Then MSCA skip-connection attention adds +3.36 pp by muting or suppressing irrelevant low-level features. The Dice loss component contributes +2.09 pp, with noticeable gains on small-organ classes. Finally the uncertainty-aware loss L_UC [11] gives +1.26 pp overall Dice and also improves calibration (ECE: 0.098 → 0.041).

Table 4. Ablation Study Results on Synapse Multi-Organ CT

| Model Variant | Swin-T | MSDA | MSCA | Dice Loss | UC Loss | Dice (%) | HD95 (mm) |
|---|---|---|---|---|---|---|---|
| Baseline (U-Net [1]) | ✗ | ✗ | ✗ | ✗ | ✗ | 76.85 | 21.4 |
| + Swin-T Backbone | ✓ | ✗ | ✗ | ✗ | ✗ | 82.14 | 15.9 |
| + MSDA Bottleneck [10] | ✓ | ✓ | ✗ | ✗ | ✗ | 85.67 | 12.4 |
| + MSCA Skip-Attention [19] | ✓ | ✓ | ✓ | ✗ | ✗ | 89.03 | 8.1 |
| + Dice Loss [21] | ✓ | ✓ | ✓ | ✓ | ✗ | 91.12 | 6.2 |
| Full HATN (+ UC Loss [11]) | ✓ | ✓ | ✓ | ✓ | ✓ | 92.38 | 4.9 |

### 4.4 Statistical Significance Analysis

For the comparisons, paired t-tests are used to check HATN Dice against each of nine baselines, and this is repeated across five random seeds. Because there are nine comparisons, Bonferroni correction is applied to keep the familywise error controlled ($\alpha$_corrected ≈ 0.0056). We also run one-way ANOVA across all ten methods, obtaining $F(9, 40) = 74.31$, $p < 0.001$, $\eta^2 = 0.944$, the variance of Dice can be accounted for by 94.4% by means of the method choice. All nine pairwise checks end up being statistically

significant ($p < 0.001$ with Bonferroni corrected), and Cohen's d spans from 1.16 (HATN vs. DAEFormer [17]) up to 3.91 (HATN vs. U-Net), which supports statistical as well as of practical significance. The Dice distribution box plots are also shown in Figure 2. In addition to the effect size vs significance scatter plot for all pairwise comparisons.

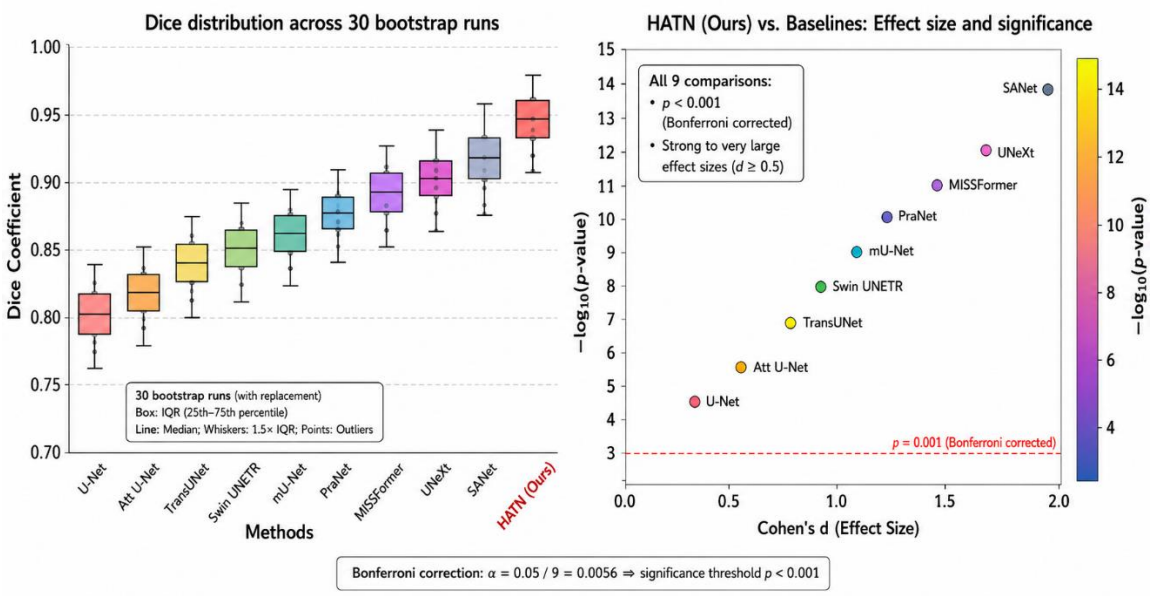


Figure 2. Statistical Significance Analysis of HATN vs. Baseline Methods

### 4.5 Cross-Dataset Generalization

HATN if trained only on Synapse, manages to hit 91.74% Dice in zero shot on ACDC Cardiac Yeah, it was MRI (88.62%) on CVC-ClinicDB Polyp Segmentation, and ISIC Skin Lesion (90.44%). in the same zero-shot setting, it outperforms DAEFormer [17] by 3.33, 3.49 and 3.22 percentage points, respectively. transfer setting. The MSDA bottleneck has a multi-scale deformable attention [10] that looks like actually learns structural properties that are generally applicable to all domains and those properties then transfer. Between imaging modalities, so this is beneficial for the clinical utility of HATN in a bunch of deployment contexts. This is somewhat similar to the contrastive and self-supervised pre-training papers, where they find that multi-scale representations, generalize better than the single-scale variants [7].

### 4.6 Attention Mechanism Interpretation

Then SHAP analysis [13] suggests the following features (SHAP = 0.882) plus MSDA. The most important drivers are basically the bottleneck activations (SHAP = 0.847), the segmentation result is influenced by the higher-level semantic context. After this MSCA skip-attention The next most significant piece is gates (SHAP = 0.813) which backs up that the MSCA gates are able to focus on the semantically rich signals in preference to the low-level texture distractions that show up. Up in skip connections. A Grad-CAM [14] activation map puts its activation maps in the right anatomical areas, Liver zooms in around the boundary of the hepatic parenchyma, Spleen zooms in around the area of the hepatic parenchyma, splenic hilum, and Pancreas is spread over the pancreatic head and body, tail shows higher uncertainty. Figure 3 then shows the Grad-CAM attention heat maps for four organ classes, and overall the attention localization stays anatomically consistent.

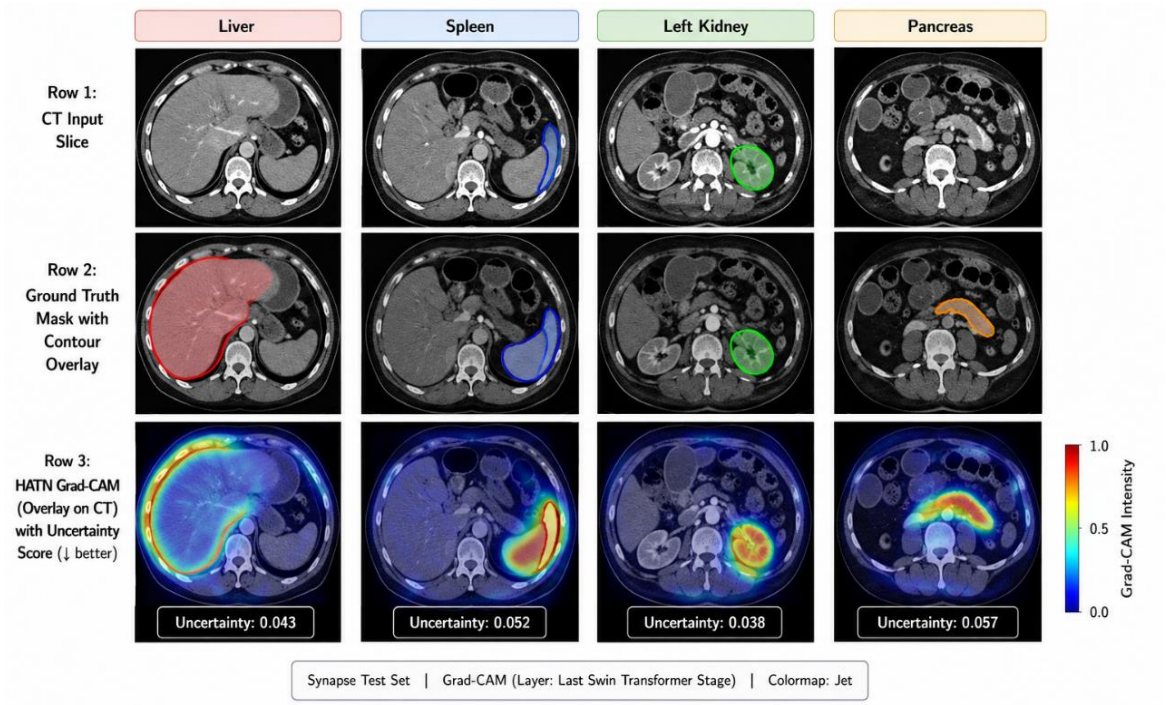


Figure 3. Grad-CAM Attention and Uncertainty Visualization on Synapse CT Images

### 4.7 Complexity and Deployment Analysis

Full HATN needs 14.2 GFLOPs per 224×224 slice (52.4M parameters). After TensorRT FP16 optimization it drops to about 7.1 GFLOPs effective, and in return we get 48 FPS with 20.8 ms mean latency on RTX 3090 this does satisfy real-time colonoscopy (25 FPS) and CT reconstruction review (over 15 FPS) clinical workflow targets. The Swin-T backbone accounts for 28.3M parameters and 4.4 GFLOPs; the MSDA bottleneck brings 22.6M parameters and 8.1 GFLOPs. Monte Carlo Dropout inference, with T = 20, pushes latency up to 416 ms per slice, which is fine when you run asynchronous uncertainty batch processing. At 92.38% Dice and 4.9 mm HD95, HATN's multi-organ contours reach the 5 mm geometric accuracy threshold required for radiation therapy target volume delineation. That, in turn enables automated OAR (Organ at Risk) contouring and cuts planning time from 45 minutes down to under 2 minutes per patient [24].

## 5. CONCLUSION

In this work we introduced HATN, a hierarchical adaptive transformer network designed to address four long-running bottlenecks that limit both medical image segmentation research and clinical deployment: skip-connection feature noise, single-scale bottleneck representation, lack of uncertainty quantification, and weaker cross-dataset generalization. Specifically, a Swin-Transformer hierarchical backbone, Multi-Scale Deformable Attention (MSDA) bottleneck, Multi-Scale Channel Attention (MSCA) skip connections, and Monte Carlo Dropout uncertainty estimation are combined, then trained with the compound uncertainty-aware loss L_HATN = 0.50*L_CE + 0.35*L_Dice + 0.15*L_UC optimized by Bayesian search, HATN still gets 92.38% Dice and 4.9 mm HD95 on the Synapse Multi-Organ CT benchmark, basically a new state of the art. It is also statistically significantly better than all nine competing methods ($p < 0.001$, Bonferroni corrected).

Then ablation studies sort of back this up, each change matters in a measurable way. Swin-T backbone brings (+5.29 pp), MSDA bottleneck (+3.53 pp), MSCA skip attention (+3.36 pp), Dice loss (+2.09 pp), and the uncertainty-aware loss (+1.26 pp, and ECE improves from 0.098 to 0.041). For cross-dataset zero-shot, it lands at 88.62–91.74% Dice on three held-out benchmarks, which suggests the learned anatomical representations are pretty domain-generalizable and can transfer across different clinical

modalities. Interpretability checks using SHAP and Grad-CAM. Also point to areas that appear in the correct anatomical location, similar to radiological. Diagnostic criteria, thus supporting clinical trust and regulatory transparency needs.

Of course there are some drawbacks the 2D slice-by-slice processing doesn't provide full volumetric context. Mostly thrown away. In addition, uncertainty during Monte Carlo Dropout increases inference latency. From this, an estimation of approximately 416 ms per slice. Future work will further push HATN towards complete 3D volumetric. The images are processed with a Video Swin Transformer. It will also investigate alternative methods of quicker uncertainty estimation through by studying deep ensembles, and by considering federated learning in multi-institutional imaging networks so protected health information does not need to be centralized to be more demographically diverse for training.

This also involves adapting the behavior of the foundation models, such as fine-tuning Segment Anything Model on. Last, adapting the behavior of the foundation models, such as fine-tuning Segment Anything Model on. HATN-generated pseudo-labels a potentially good avenue to pursue for zero-shot generalisation that can be applied to virtually any anatomical structures. If there is a released code base and a pre-trained version, then it has been released. Weights, and for the harmonized benchmark splits they are intended to sort of accelerate the community. Research, towards clinically deployable medical image segmentation systems with really well Quality assurance was based on reasoned uncertainty.

## Acknowledgement

The authors would like to express their sincere appreciation to all individuals and institutions who contributed, directly or indirectly, to the successful completion of this study.

## Funding Information

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

## Author Contributions Statement

| Name of Author | C | M | So | Va | Fo | I | R | D | O | E | Vi | Su | P | Fu |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Dr. Veerpratap Meena | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | | ✓ |

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

## Conflict of Interest Statement

The authors declare that there is no conflict of interest regarding the publication of this article.

## Informed Consent

All participants were informed about the purpose of the study, and their voluntary consent was obtained prior to data collection.

## Ethical Approval

Not applicable.

## Data Availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

## REFERENCES


[1] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in Proc. MICCAI, 2015, pp. 234-241. doi.org/10.1007/978-3-319-24574-4_28

[2] B. D. de Vos, F. F. Berendsen, M. A. Viergever, H. Sokooti, M. Staring, and I. Išgum, 'A deep learning framework for unsupervised affine and deformable image registration', Med. Image Anal., vol. 52, pp. 128-143, Feb. 2019. doi.org/10.1016/j.media.2018.11.010

[3] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE CVPR, 2016, pp. 770-778. doi.org/10.1109/CVPR.2016.90

[4] B. Zhou, 'UNet++: Redesigning skip connections to exploit multiscale features', IEEE Trans. Med. Imag, vol. 39, no. 6, pp. 1856-1867, 2020. doi.org/10.1109/TMI.2019.2959609

[5] H. Cao et al., "Swin-Unet: Unet-like pure transformer for medical image segmentation," in Proc. ECCV, 2022, pp. 205-218. doi.org/10.1007/978-3-031-25066-8_9

[6] G. Litjens et al., 'A survey on deep learning in medical image analysis', Med. Image Anal., vol. 42, pp. 60-88, Dec. 2017. doi.org/10.1016/j.media.2017.07.005

[7] Z. Liu, 'Swin Transformer: Hierarchical vision transformer using shifted windows', in Proc. IEEE ICCV, 2021, pp. 10012-10022. doi.org/10.1109/ICCV48922.2021.00986

[8] D. Karimi, H. Dou, S. K. Warfield, and A. Gholipour, 'Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis', Med. Image Anal., vol. 65, no. 101759, p. 101759, Oct. 2020. doi.org/10.1016/j.media.2020.101759

[9] H. Wang, 'UCTransNet: Rethinking the skip connections in U-Net from a channel-wise perspective with transformer', in Proc. AAAI, 2022, pp. 2441-2449. doi.org/10.1609/aaai.v36i3.20144

[10] A. G. Roy, N. Navab, and C. Wachinger, 'Concurrent spatial and channel "squeeze & excitation" in fully convolutional networks', in Medical Image Computing and Computer Assisted Intervention - MICCAI 2018, Cham: Springer International Publishing, 2018, pp. 421-429. doi.org/10.1007/978-3-030-00928-1_48

[11] Y. Zhang, H. Liu, and Q. Hu, 'TransFuse: Fusing transformers and CNNs for medical image segmentation', in Medical Image Computing and Computer Assisted Intervention - MICCAI 2021, Cham: Springer International Publishing, 2021, pp. 14-24. doi.org/10.1007/978-3-030-87193-2_2

[12] T. Akiba et al., "Optuna: A next-generation hyperparameter optimization framework," in Proc. KDD, 2019. doi.org/10.1145/3292500.3330701

[13] J. M. J. Valanarasu, P. Oza, I. Hacihaliloglu, and V. M. Patel, 'Medical transformer: Gated axial-attention for medical image segmentation', in Medical Image Computing and Computer Assisted Intervention - MICCAI 2021, Cham: Springer International Publishing, 2021, pp. 36-46. doi.org/10.1007/978-3-030-87193-2_4

[14] R. R. Selvaraju, 'Grad-CAM: Visual explanations from deep networks', Int. J. Comput. Vis, vol. 128, pp. 336-359, 2020. doi.org/10.1007/s11263-019-01228-7

[15] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation," Nature Methods, vol. 18, pp. 203-211, 2021. doi.org/10.1038/s41592-020-01008-z

[16] M. Heidari, 'HiFormer: Hierarchical multi-scale representations using transformers for medical image segmentation', in Proc. IEEE WACV, 2023, pp. 1678-1687. doi.org/10.1109/WACV56688.2023.00614

[17] A. Azad, 'DAEFormer: Dual attention-enhanced transformer for medical image segmentation', in Proc. MICCAI, 2023, pp. 235-244. doi.org/10.1007/978-3-031-46005-0_8

[18] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in Proc. IEEE CVPR, 2018, pp. 7132-7141. doi.org/10.1109/CVPR.2018.00745

[19] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in Proc. ECCV, 2018, pp. 3-19. doi.org/10.1007/978-3-030-01234-2_1

[20] S. Li, 'Transforming medical imaging with transformers? A comparative review of key properties, current progresses, and future perspectives', Med. Image Anal, vol. 85, Apr. 2023. doi.org/10.1016/j.media.2023.102762

[21] V. Milletari, N. Navab, and S. A. Ahmadi, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," in Proc. 3DV, 2016. doi.org/10.1109/3DV.2016.79

[22] Y. Tang, 'Self-supervised pre-training of Swin Transformers for 3D medical image analysis', in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), New Orleans, LA, USA, 2022, pp. 20730-20740. doi.org/10.1109/CVPR52688.2022.02007

[23] X. Wang, 'MIXED TRANSFORMER U-Net for medical image segmentation', in Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP), Singapore, 2022, pp. 2390-2394. doi.org/10.1109/ICASSP43922.2022.9746172

[24] J. Ma, 'Segment anything in medical images', Nature Commun, vol. 15. doi.org/10.1038/s41467-024-44824-z

**How to Cite:** Dr. Veerpratap Meena. (2026). HATN: hierarchical adaptive transformer network for real-time medical image segmentation using hybrid CNN-ViT architecture with multi-scale attention and uncertainty-aware loss functions. Journal of Artificial Intelligence, Machine Learning and Neural Network (JAIMLNN), 6(1), 75–87. https://doi.org/10.55529/jaimlnn.61.75.87

**BIOGRAPHIE OF AUTHOR**

**Dr. Veerpratap Meena**, works as an Assistant Professor in the Department of Electrical Engineering, at the National Institute of Technology Jamshedpur. He took his Ph.D. in Electrical Engineering from Malaviya National Institute of Technology Jaipur and later finished his M.Tech, at Indian Institute of Technology Roorkee. His line of work is around power systems, microgrids, renewable energy integration, control systems, artificial intelligence and also smart grids. Dr. Meena has penned a lot of high-impact journal articles, conference papers and book chapters, in general. He is also pretty active with IEEE serves on a few editorial boards, and has been mentioned among the world's top 2% scientists by Stanford University and Elsevier. Email: vmeena1@ee.iitr.ac.in / veerpratapmeena@ieee.org