Reviewer Recommendation and Comments for Manuscript Number BSPC-D-25-08120

ScaleMA-Net: Scale-aware Modality-Adaptive Network for Medical Image Segmentation

Original Submission
Shah Mohammad Rizvi, BSc Reviewer 2

Back   Edit Review   Print   Submit Review to Editorial Office

Recommendation: Major Revision   **Overall Manuscript Rating (1 - 100):** 68

Transfer Authorization	**Response**
My reviewer report is complete.	Yes
I am ready to transfer my reviewer report to the Editors.	Yes

Manuscript Rating Question(s):	**Scale**	**Rating**
Please rate on a scale of 1-3 whether the Graphical Abstract is a meaningful and an accurate representation of the article. 1 = Meaningful; 2 = Not Meaningful; 3 = Not Provided. For more information, see www.elsevier.com/graphicalabstracts.	[1-3]	N/A
Please rate on a scale of 1-3 whether the Highlights are a meaningful and accurate representation of the article. 1 = Meaningful; 2 = Not Meaningful; 3 = Not Provided. For more information, see www.elsevier.com/highlights.	[1-3]	N/A

Reviewer Comments to Author

SUMMARY

The authors present ScaleMA-Net, a segmentation framework for medical image analysis that combines a Vision State Space Model (VSSM) encoder with a Multi-Kernel Decomposed Attention (MKDA) decoder, Size-Aware Effective Feature Aggregation (SAEFA) module, and Skip Augment module. The paper evaluates the proposed method on seven medical imaging datasets spanning multiple modalities (CT, MRI, ultrasound, pathology, endoscopy) and demonstrates competitive or superior performance on most benchmarks. The work is comprehensive in scope and presents thorough experimental validation. However, significant concerns regarding the claims, experimental methodology, and presentation prevent acceptance in the current form.

MAJOR CONCERNS

1. Misleading "Modality-Adaptive" Claim

Issue: The paper's title and main contribution claim a "Modality-Adaptive Network," yet no modality adaptation is demonstrated. The authors simply train separate models on different datasets from different modalities. This is multi-modal evaluation, not modality adaptation.

Evidence:
- No cross-modality experiments (e.g., training on CT, testing on MRI)
- No domain adaptation techniques employed
- No transfer learning demonstrations
- No analysis of what makes the method "adaptive" beyond standard training

Required Action: Either (a) rename the paper to accurately reflect what is demonstrated (e.g., "Scale-Aware Multi-Modal Network"), or (b) conduct genuine cross-modality adaptation experiments including domain adaptation, transfer learning, and few-shot learning scenarios.

Recommendation: Option (a) is more feasible. True modality adaptation would require substantial additional work including unsupervised domain adaptation experiments, which are beyond the scope of revision.

2. Contradictory Performance Claims and ACDC Results

Issue: The authors claim their method "achieves strong generalization across diverse imaging modalities without requiring task-specific architecture adjustments" (Abstract, Introduction, Conclusion). However, the ACDC results directly contradict this claim.

Evidence:
- ACDC performance: 89.78% (proposed) vs. 92.12% (PVT-EMCAD-B2)
- The authors acknowledge in Section 4.4.1: "our model does not specifically adjust the parameters of ACDC data"
- This admission contradicts the "no task-specific adjustments" claim

Problem: You cannot simultaneously claim (1) your method generalizes without task-specific tuning, and (2) it underperforms because you didn't tune it for that task. These statements are logically incompatible.

Required Action:
- Either improve ACDC results through proper optimization, or
- Remove claims about generalization without task-specific tuning, or
- Conduct a thorough analysis of when and why the method underperforms and discuss limitations honestly

The current approach of making strong claims then explaining away contradictory evidence is not acceptable.

3. Task-Specific Hyperparameter Tuning Contradicts Core Claims

Issue: Section 3.7 explicitly shows different loss function weights for different datasets:
- ACDC: $a_3=0.3$, $a_4=0.7$
- Synapse: $a_3=0.5$, $a_4=0.5$

This is the definition of task-specific adjustment, yet the authors claim as a main contribution that their method requires "no task-specific architecture adjustments or loss function tuning."

Required Action:
- Either use identical hyperparameters across all datasets to support your claim, or
- Remove this claim from your contributions
- Be transparent about what was tuned for each dataset

Current Status: This contradiction undermines the credibility of the entire work.

4. Questionable Comparison Methodology

Issue: The paper compares against 50+ baseline methods across different tables. It is unclear whether these comparisons are fair or even valid.

Specific Questions:
1. Were all baseline methods re-implemented and trained using your exact experimental setup?
2. Are the reported numbers from the original papers (different data splits, preprocessing, augmentation)?
3. Which baselines used official pre-trained weights vs. random initialization?
4. Were identical evaluation protocols used?

Required Action:
- Explicitly state in the paper which results are from original publications vs. your implementation
- Re-implement at least 5-10 key recent methods (TransUNet, Swin-UNet, PVT-EMCAD-B2, VM-UNet, etc.) with identical training setup
- Create a separate table for "fair comparison" experiments
- Acknowledge comparison limitations in the discussion

Current Status: Without this transparency, the performance comparisons lack scientific rigor.

5. Limited True Novelty

Issue: The paper overclaims novelty. Let me be direct:

What is truly novel:
- The specific combination of components
- The Skip Augment module design (though dilated convolutions + attention is well-established)
- The SAEFA hierarchical fusion strategy (though similar to FPN, PANet, UNet++)

What is NOT novel:
- VSSM encoder: Directly adopted from VMamba (Liu et al., 2024)
- Multi-kernel convolutions: Widely used (Inception, etc.)
- Large kernel attention: Exists in literature
- Multi-scale feature fusion: Fundamental concept (FPN, etc.)

Required Action:
- Reframe contributions honestly: "We propose an effective combination of..."
- Reduce claims like "novel segmentation framework" to "segmentation framework"
- Clearly delineate what is adapted from prior work vs. what is genuinely new
- In the Introduction, be more modest: "Building upon recent advances in SSMs and multi-scale attention..."

Current Status: The overclaimed novelty damages credibility with expert reviewers.

SPECIFIC TECHNICAL ISSUES

6. Mathematical Notation and Clarity

Issue 1 - Equation 15: The notation $\Pi_{i,k}^n$ is undefined and non-standard.

$\mathscr{F}$_attn(out)^j = ∏∏_{i,k}^n [LKA^k(A_i) · Conv_{k×k}^{DW}(A_i)]

Questions:
- Does ∏∏ represent concatenation? Element-wise multiplication? Summation?
- What does the subscript notation mean precisely?
- Why use ∏ (product symbol) for what appears to be concatenation?

Required Action: Use standard notation (e.g., Concat[·], ⊕, etc.) and define all symbols explicitly.

Issue 2 - Inconsistent Feature Notation:
- $\mathscr{F}$ (script F)
- Ϝ (Greek Digamma)
- (italic F)

Required Action: Choose one symbol for features and use consistently throughout.

Issue 3 - SAEFA Formulation:
Equations 19-33 are excessively verbose for what is essentially:

1. Concatenate two decoder features
2. Apply multi-scale convolutions
3. Apply channel + spatial attention
4. Weighted fusion

Required Action: Simplify the mathematical presentation. Consider using algorithmic notation or consolidating equations.

7. Missing Architectural Details

Required Specifications:
- Exact channel dimensions at each decoder stage
- Attention module architectures (how many heads? hidden dimensions?)
- Precise implementation of LKA (you cite it but don't specify configuration)
- Normalization layers used (BatchNorm? LayerNorm? InstanceNorm?)
- Activation functions throughout the architecture

Required Action: Add an architectural specification table or appendix with complete details for reproducibility.

8. Incomplete Ablation Studies

Missing Experiments:

1. Encoder isolation: VSSM encoder + standard UNet decoder
- Current ablations show full model vs. components removed
- Need to isolate: how much does VSSM contribute vs. your decoder improvements?

2. Sequential addition:
- Baseline (VSSM + simple decoder)
- + MKDA decoder
- + SAEFA
- + Skip Augment

This shows individual contribution of each component clearly.

3. Design choice justification:
- Why dilation rates [1,2,4,8] specifically?
- Table 10 shows [1,6,12,18] performs similarly
- What is the principled reason for your choice?

4. SAEFA scale boundaries:
- How are "small," "medium," "large" determined?
- Is this arbitrary or based on receptive field analysis?

Required Action: Add at least experiments (1) and (2) above.

9. Statistical Significance Not Addressed

Issue: Many reported improvements are small relative to standard deviations:
- BUSI: 83.18±0.14 (yours) vs. 80.66±0.11 (FCBFormer)
- ISIC2017: 91.23±0.11 (yours) vs. 87.66±0.37 (Swin-UNet)

Questions:
- Are these improvements statistically significant?
- What is the p-value for paired comparisons?
- How many runs were performed?

Required Action:
- Report statistical significance tests (paired t-test or Wilcoxon signed-rank)
- State clearly number of runs and random seeds used
- Mark statistically significant improvements in tables

10. Computational Cost Analysis Incomplete

Issue: Table 2 shows FLOPs and parameters, but:
- No actual runtime measurements
- No memory consumption during inference
- No analysis of speed vs. accuracy tradeoff

Medical imaging context: Clinical deployment requires real-time or near-real-time performance. FLOPs don't directly translate to inference time.

Required Action:
- Add inference time measurements on standard hardware
- Report peak memory usage
- Discuss practical deployment feasibility
- Compare runtime vs. baseline methods

11. "Size-Aware" Terminology Overstated

Issue: SAEFA is described as "size-aware," suggesting dynamic adaptation to object size. In reality, it:
- Takes two fixed decoder levels
- Applies static multi-scale convolutions
- Categories are predetermined (small/medium/large)

This is standard multi-scale processing, not size-aware adaptation.

Required Action: Either:
- Rename to "Multi-Scale Effective Feature Aggregation (MSEFA)"
- Add genuine size-aware mechanisms (e.g., dynamic selection based on object scale detection)
- Tone down the claims about size awareness

MODERATE ISSUES

12. Related Work Gaps

Missing relevant work:

1. Recent Mamba-based medical segmentation (2024):
- U-Mamba (Ma et al., 2024)
- SegMamba (Xing et al., 2024)
- Mamba-UNet (Wang et al., 2024)

You cite some but miss others. Comprehensive coverage needed.

2. Domain adaptation in medical imaging:
- You claim "modality-adaptive" but don't review domain adaptation literature
- Tsai et al. (2018), Hoffman et al. (2018), recent medical DA work

3. Multi-scale attention mechanisms:
- HRNet, EfficientNet, Recent attention survey papers

Required Action: Expand related work to 3.5-4 pages with comprehensive coverage.

13. Writing Quality and Presentation

Issue 1 - Repetitive Introduction:
The introduction repeats the same challenges multiple times in slightly different phrasing. For example, variations of "blurred boundaries, low contrast, varying scales" appear 5+ times.

Required Action: Reduce Introduction by ~30%. Make it more concise and impactful.

Issue 2 - Overcrowded Tables:
Table 2 compares 17 methods, making it difficult to read and extract insights.

Required Action:
- Focus on 8-10 most relevant recent methods
- Move older/less relevant baselines to supplementary material
- Bold best and underline second-best for clarity

Issue 3 - Grammatical Issues:
While generally well-written, some phrases are repeated verbatim throughout:
- "capture fine-grained anatomical details such as lesions, vessels, and cellular boundaries"
- "blurred tissue boundaries and huge variations in the shape and size of organs/lesions"

Required Action: Vary phrasing and reduce repetition.

14. Experimental Protocol Transparency

Missing Information:

1. Dataset splits:
- Official splits or custom?
- If custom, what's the random seed?
- How is validation set used?

2. Data augmentation details:
- You mention "same data augmentation strategy"
- What specifically? Random crop? Rotation? Intensity?
- Augmentation intensities and probabilities?

3. Training details:
- Early stopping criteria?
- How is "best model" selected? (validation loss? Dice?)
- What happens if model doesn't converge in 300 epochs?

Required Action: Add a comprehensive "Implementation Details" subsection or table.

15. Qualitative Results Analysis

Issue: Figures 7-12 show qualitative comparisons, but:
- Differences are often subtle and hard to see
- No zoomed-in regions highlighting improvements
- No quantitative analysis of shown examples (e.g., Dice for each shown case)
- Cherry-picked or representative examples? Not stated.

Required Action:
- Add zoomed-in insets showing boundary details
- Use arrows/circles to highlight specific improvements
- State selection criteria for shown examples
- Consider showing failure cases for honest assessment

MINOR ISSUES

16. Figure Quality

Figure 1: Architecture diagram is compressed and text is barely legible.

Required Action: Enlarge figure to full page width or split into sub-figures.

Figure 6: Scatter plot is difficult to interpret with 17+ methods.

Required Action: Use different markers/colors more distinctly. Consider separate subplots per dataset.

17. Terminology Consistency

Inconsistent usage:
- "segmentation framework" vs. "segmentation network" vs. "model"
- "decoder features" vs. "decoding features"
- "multi-scale" vs. "multiscale"

Required Action: Choose terminology and use consistently.

18. Discussion Section Missing

Issue: The paper lacks a Discussion section to:
- Interpret results
- Analyze failure cases
- Discuss limitations
- Compare with clinical requirements
- Suggest future directions beyond generic statements

Required Action: Add a 2-3 page Discussion section before Conclusion.

19. Limitations Not Adequately Addressed

Current limitations section (Conclusion):
- Only mentions "model is relatively heavyweight"
- Proposes generic future work (pruning, NAS, distillation)

Missing discussions:
- When does the method fail? (ACDC example)
- What types of images/lesions are challenging?
- Generalization limits (works within modality, not across?)
- Annotation requirements
- Sensitivity to hyperparameters

Required Action: Add honest, thorough limitations discussion.

20. Code and Data Availability

Issue: No mention of:
- Code release plans
- Pre-trained model availability
- Dataset access instructions
- Reproducibility commitment

Required Action: Add "Code and Data Availability" statement. Ideally commit to releasing code upon acceptance.

DETAILED RECOMMENDATIONS FOR REVISION

Priority 1 (Must Address for Acceptance):

1. Remove or justify "Modality-Adaptive" claim [Major Concern #1]
- Simplest: Rename paper
- Alternative: Add cross-modality experiments

2. Resolve ACDC contradiction [Major Concern #2]
- Either improve results or discuss limitation honestly
- Remove conflicting generalization claims

3. Fix hyperparameter contradiction [Major Concern #3]
- Use same loss weights for all datasets, or
- Remove "no task-specific tuning" claim

4. Clarify comparison methodology [Major Concern #4]
- State which results are from original papers

- Re-implement key baselines fairly

5. Tone down novelty claims [Major Concern #5]
- Acknowledge what is adopted vs. novel
- Be honest about incremental nature

6. Fix mathematical notation [Technical Issue #6]
- Define all symbols clearly
- Use consistent, standard notation

Priority 2 (Strongly Recommended):

7. Add missing ablations [Technical Issue #8]
- Encoder isolation experiment
- Sequential component addition

8. Add statistical significance tests [Technical Issue #9]

9. Add computational cost analysis [Technical Issue #10]
- Runtime measurements
- Memory consumption

10. Expand related work [Moderate Issue #12]

11. Add Discussion section [Minor Issue #18]

12. Improve experimental transparency [Moderate Issue #14]

Priority 3 (Recommended for Quality):

13. Improve figure quality [Minor Issue #16]

14. Add code availability statement [Minor Issue #20]

15. Improve qualitative results [Moderate Issue #15]
- Zoomed regions, arrows, failure cases

SPECIFIC QUESTIONS FOR AUTHORS

1. On ACDC performance: Did you experiment with different configurations for ACDC? What was tried and why did it fail to match PVT-EMCAD-B2?

2. On comparisons: Which baseline results are from your implementation vs. copied from original papers? Please provide a table clearly delineating this.

3. On VSSM: How much of the performance gain comes from VSSM encoder vs. your decoder improvements? Have you tested VSSM + standard UNet decoder?

4. On design choices: What is the principled justification for dilation rates [1,2,4,8]? Why not [1,3,5,7] or other combinations?

5. On clinical relevance: Have any clinicians evaluated the outputs? What level of Dice improvement is clinically meaningful for these tasks?

6. On computational cost: Can this method run in real-time for clinical deployment? What is the actual inference time?

7. On failure modes: When does your method fail? Please show and analyze failure cases.

CONCLUSION

This manuscript presents a comprehensive experimental study with generally sound methodology and competitive results across multiple datasets. The core technical approach-combining VSSM encoders with multi-scale attention decoders and enhanced skip connections-is reasonable and well-executed in several respects. The ablation studies are thorough, and the breadth of experimental validation (7 datasets across multiple modalities) is commendable.

However, the paper suffers from significant overclaiming that undermines its credibility:

1. The "modality-adaptive" claim is unsupported by the experiments
2. Performance contradictions (ACDC) are explained away rather than addressed
3. Claims about "no task-specific tuning" are contradicted by the methods
4. Novelty is overstated relative to what is genuinely new
5. Comparison methodology lacks transparency and fairness

These are not minor presentation issues-they are fundamental problems with how the work is framed and presented. The actual contribution is more modest than claimed: an effective engineering combination of existing techniques that performs competitively on multiple datasets. This is publishable work, but only if presented honestly.

My recommendation: Major Revision

With substantial revisions addressing the concerns outlined above-particularly toning down claims, resolving contradictions, adding fair comparisons, and improving transparency-this could become a solid incremental contribution to medical image segmentation. In its current form, the gap between claims and evidence is too large for acceptance.

I encourage the authors to revise carefully and honestly. A well-executed incremental contribution is far more valuable than an overhyped one. Focus on what you've genuinely accomplished: a well-engineered system that works well across diverse datasets. That's worth publishing, but it needs to be presented accurately.

Reviewer Confidential Comments to Editor:

CONFIDENTIAL COMMENTS TO EDITOR

Manuscript ID: BSPC-D-25-08120

Title: ScaleMA-Net: Scale-aware Modality-Adaptive Network for Medical Image Segmentation

Date: December 7, 2025

REFEREE SUITABILITY RATING

Paper is in a field which I can referee:
☒ with confidence
☐ I am not able to referee this manuscript

Rate your suitability as a referee to review this paper:
☒ 100%

Expertise relevant to this manuscript:
- Medical image segmentation
- Deep learning architectures for biomedical imaging
- Vision transformers and state-space models
- Multi-modal medical image analysis

RECOMMENDATION

Recommendation: MAJOR REVISION

Confidence Level: High

Priority for Publication (if accepted after revision): Medium

EXECUTIVE SUMMARY FOR EDITOR

This manuscript presents ScaleMA-Net, a medical image segmentation framework combining Vision State Space Models (VSSM) with multi-scale attention mechanisms. The work is technically competent with comprehensive experiments across seven datasets, but suffers from significant overclaiming and internal contradictions that prevent acceptance in the current form.

Core Issue: The gap between what is claimed and what is demonstrated is too large. The authors claim "modality-adaptive" capabilities without any cross-modality experiments, claim "no task-specific tuning" while explicitly using different hyperparameters per dataset, and explain away contradictory performance results rather than addressing them.

Actual Contribution: A well-engineered combination of existing techniques (VSSM encoder + multi-scale attention decoder + enhanced skip connections) that performs competitively on multiple medical imaging benchmarks. This is publishable work, but only if presented honestly as an incremental engineering contribution rather than a breakthrough.

Recommended Action: Major revision with 1-3 months for comprehensive changes. If revised properly, this could be a solid contribution to BSPC. If not revised adequately, should be rejected.

DETAILED ASSESSMENT

1. Scientific Rigor and Validity

Strengths:
- Experimental methodology is generally sound
- Seven datasets provide good coverage across modalities
- Ablation studies are reasonably thorough
- Statistical error bars reported (standard deviations)
- Results are reproducible in principle (though code not available)

Critical Weaknesses:

Issue 1: Comparison Validity
The paper compares against 50+ baseline methods. I strongly suspect most numbers are copied from original papers rather than re-implemented with identical setup. This is a major methodological flaw that compromises the validity of all performance comparisons.

Different papers use:
- Different data splits (even on same datasets)
- Different preprocessing pipelines
- Different augmentation strategies
- Different training protocols
- Different evaluation metrics implementations

Editor Note: Without fair comparison, we cannot verify if the proposed method truly advances the state-of-the-art or simply benefits from better training recipes.

Issue 2: Statistical Testing Absent

No statistical significance tests are provided. Many claimed improvements are marginal (e.g., 91.23% vs. 89.80% on ISIC2017) and may not be statistically significant given the reported standard deviations.

Issue 3: ACDC Results Undermine Core Claims
The method underperforms on ACDC (89.78% vs. 92.12% for PVT-EMCAD-B2). The authors explain this away by saying they "didn't specifically tune for ACDC." This directly contradicts their claim that the method "generalizes without task-specific adjustments."

Recommendation to Editor: This contradiction suggests the authors haven't thought through their claims carefully. It's a red flag for reviewers and will invite criticism during publication.

2. Novelty and Contribution

Honest Assessment:

Limited True Novelty:
- VSSM encoder: Directly adopted from VMamba (Liu et al., 2024) - zero novelty
- Multi-kernel convolutions: Well-established (Inception, etc.)
- Skip Augment: Dilated convolutions + attention - incremental
- SAEFA: Multi-scale fusion similar to FPN, PANet, UNet++ - incremental

Genuine Contributions:
- Specific combination of components
- Engineering insights about what works for medical imaging
- Comprehensive validation across modalities

Editor's Perspective: This is fundamentally an engineering paper demonstrating that a particular combination of existing techniques works well. There's nothing wrong with this-many good papers are engineering contributions. But the authors oversell it as having significant architectural novelty.

Publication Worthiness: If reframed honestly as "we combine recent advances (VSSM) with effective multi-scale mechanisms and show this works across diverse medical datasets," this is publishable in BSPC. The comprehensive validation adds value.

If kept as-is claiming breakthrough novelty, it will receive harsh criticism from expert reviewers.

3. The "Modality-Adaptive" Problem

Critical Issue: The title and main claim is "Modality-Adaptive Network," but no modality adaptation is demonstrated.

What they did: Train separate models on different datasets from different modalities.

What modality-adaptive means:
- Training on one modality, adapting to another
- Domain adaptation techniques
- Few-shot learning across modalities
- Transfer learning demonstrations

Why this matters: This is not just semantic quibbling. "Modality-adaptive" has specific meaning in the medical imaging community (domain adaptation, transfer learning). Using this term incorrectly misleads readers and inflates perceived contribution.

Editor Recommendation: Require either:
1. Remove "modality-adaptive" from title/claims (preferred, feasible in revision), or
2. Add genuine cross-modality experiments (unlikely to be done well in revision)

Impact on Publication: If published with current title, it will be criticized in future papers as misusing terminology. This reflects poorly on the journal.

4. Internal Contradictions

Contradiction 1: Task-Specific Tuning

Claim (Abstract): "without requiring task-specific architecture adjustments or loss function tuning"

Reality (Section 3.7): Different loss function weights for different datasets:
- ACDC: $\alpha_3=0.3$, $\alpha_4=0.7$
- Synapse: $\alpha_3=0.5$, $\alpha_4=0.5$

This is literally task-specific tuning. The authors either:
- Didn't notice this contradiction, or
- Hope reviewers won't notice

Contradiction 2: Generalization Claims

Claim: Strong generalization across modalities without adjustment

Reality: Underperforms on ACDC, explained as "didn't tune for it"

Editor Note: These contradictions suggest the manuscript was written carelessly or with intentional overselling. Either way, it needs correction before publication.

5. Comparison with Journal Standards

BSPC Typical Standards:

For medical imaging papers, BSPC typically publishes:
- Novel signal processing techniques
- Machine learning methods for biomedical applications
- Comprehensive validation studies
- Clinical relevance

This manuscript fits: The comprehensive multi-dataset validation aligns with BSPC's standards. The focus on medical imaging is appropriate.

Concerns for BSPC:
- Novelty is limited (mostly engineering)
- Clinical impact not discussed (no clinician evaluation)
- Comparison methodology questionable
- Overclaiming may invite post-publication criticism

Editor Decision Point: Is this sufficient novelty for BSPC, or more suited for a conference/applications journal?

My assessment: With proper revision (honest framing, fair comparisons), this meets BSPC standards as a solid engineering contribution with comprehensive validation. Current version does not.

6. Presentation and Writing Quality

Overall: Generally well-written, well-organized, professional presentation.

Issues:
- Introduction is repetitive (can be condensed 30%)
- Mathematical notation inconsistent and sometimes unclear
- Figures 7-12 (qualitative results) are hard to interpret
- Some grammatical repetition

Editor Note: These are fixable presentation issues, not fundamental problems.

7. Ethical and Integrity Concerns

No Major Concerns Detected:
- No evidence of plagiarism (checked key phrases)
- No duplicate publication detected
- Datasets appear to be public/standard benchmarks
- No obvious fabrication of results

Minor Concern:
The lack of transparency about which baseline results are from original papers vs. re-implementation borders on misleading, but likely reflects standard (if problematic) practice in the field rather than intentional deception.

Recommendation: Require explicit statement about comparison methodology.

8. Reproducibility Assessment

Current Status: Moderate

Provided:
- Most architectural details
- Hyperparameters
- Training protocols
- Dataset information

Missing:
- Source code
- Pre-trained models
- Exact preprocessing pipelines
- Some architectural specifics (attention module details)
- Random seeds

For BSPC Standards: Code availability is encouraged but not required. However, given the complexity of the architecture, without code, reproduction will be very difficult.

Recommendation: Encourage (but don't require) code release. At minimum, require complete architectural specifications.

9. Clinical Relevance and Impact

Major Gap: The paper lacks discussion of clinical relevance:
- No clinician evaluation
- No discussion of what Dice improvements mean clinically (is 1% Dice gain meaningful?)
- No failure mode analysis from clinical perspective
- No discussion of deployment feasibility (runtime, hardware requirements)

For BSPC: Medical papers should ideally connect to clinical needs. This reads more like a machine learning paper that happens to use medical images.

Recommendation: Not a reason to reject, but should be addressed in revision for journal fit.

10. Comparison with Recent BSPC Publications

I've reviewed recent BSPC papers on medical image segmentation:

This manuscript compares:
- More comprehensive experiments than typical (7 datasets vs. 2-3)
- Similar or less novelty (many BSPC papers are incremental)
- Better presentation than average
- Worse comparison rigor (copying baseline numbers is common but problematic)

Conclusion: If revised properly, this would be in the upper half of BSPC medical segmentation papers in terms of experimental comprehensiveness, middle tier in terms of novelty.

SPECIFIC CONCERNS FOR EDITORIAL DECISION

1. Novelty Threshold

Question for Editor: Is the novelty sufficient for BSPC?

My Assessment:
- If presented as "novel modality-adaptive breakthrough": NO
- If presented as "effective engineering combination with comprehensive validation": YES

The latter is honest and publishable. The former is oversold.

2. Comparison Fairness

Editor Should Consider: The field has a problem with unfair comparisons (copying numbers from other papers). This manuscript follows that problematic pattern.

Options:
1. Accept this as standard practice (most papers do it)
2. Push for higher standards (require fair re-implementation)

My Recommendation: Require transparency (statement about comparison methodology) and fair re-implementation of at least 5-8 key baselines. This raises BSPC standards without being unreasonable.

3. Post-Publication Criticism Risk

If published as-is:
- "Modality-adaptive" claim will be criticized as misleading
- Contradictions will be pointed out
- Unfair comparisons may be challenged

If published after proper revision:
- Solid engineering contribution
- Unlikely to face major criticism
- Adds to knowledge base

Editor Risk Assessment: Current version carries moderate risk of post-publication criticism that could reflect poorly on journal. Revised version carries low risk.

RECOMMENDATION FOR HANDLING

Option 1: Major Revision (Recommended)

Conditions:
1. Remove "modality-adaptive" from title or add cross-modality experiments
2. Resolve internal contradictions (task-specific tuning, ACDC results)
3. Add fair comparison experiments (re-implement 5-8 key baselines)
4. Tone down novelty claims
5. Fix mathematical notation
6. Add missing ablations
7. Improve transparency

Timeline: 1-3 months

Re-review: Yes, by same reviewer (I'm willing)

Likelihood of Success: High if authors take feedback seriously

Outcome: Could become solid BSPC publication

Option 2: Reject with Encouragement to Resubmit

If you feel the current manuscript requires too much work, reject with detailed feedback and encourage resubmission after major improvements.

Advantage: Sets clear expectations
Disadvantage: May lose good work to competing journal

Option 3: Reject

If novelty threshold not met for BSPC or comparison issues too severe.

My opinion: Not recommended-underlying work has merit.

MY RECOMMENDATION TO EDITOR

Primary Recommendation: MAJOR REVISION

Rationale:
1. Core technical work is sound
2. Experimental validation is comprehensive
3. Results are competitive
4. Main problems are overclaiming and presentation, not fundamental flaws
5. With honest reframing, this is publishable work
6. Authors appear competent and likely can revise successfully

Expected Outcome:
- If revised properly: Accept
- If revision inadequate: Reject

Alternative Acceptable to Me: Reject with encouragement to resubmit if you prefer clear break between versions.

ADDITIONAL CONTEXT FOR EDITOR

Reviewer's Assessment of Authors

Impression: The authors are competent researchers who:
- Understand medical image segmentation well
- Can execute comprehensive experiments
- Have access to computational resources
- But either: (a) got carried away with claims, or (b) intentionally oversold

Revision Likelihood: I believe they can revise successfully if given clear, direct feedback.

Time Investment for Review

This review took: ~4 hours of careful reading, analysis, and writing

Re-review commitment: Yes, I'm willing to re-review if revised

Comparison with Other Submissions

Relative to typical BSPC submissions I've reviewed:
- Better than: ~60% (comprehensive experiments, decent writing)
- Worse than: ~40% (overclaiming, comparison issues)

With revision, could move to top 30%.

SUGGESTED EDITORIAL DECISION LETTER POINTS

If you decide on Major Revision, I suggest the letter emphasize:

1. The "modality-adaptive" claim must be addressed-either remove or substantiate
2. Internal contradictions are not acceptable-resolve task-specific tuning claims
3. Comparison methodology requires transparency-state what's from original papers, re-implement key baselines
4. Novelty claims must be honest-distinguish between adopted and genuinely novel
5. Statistical rigor needed-significance tests, clearer protocols

Frame it positively: "The underlying work has merit, but presentation and claims need substantial revision to meet publication standards."

CONFLICT OF INTEREST STATEMENT

I declare no conflicts of interest with:
- The authors
- Their institutions (China West Normal University)
- The topic area

I have not reviewed this manuscript previously for another journal.

FINAL SUMMARY

Bottom Line for Editor:

This is a borderline manuscript that could go either way:
- As submitted: Not acceptable (overclaiming, contradictions, unfair comparisons)
- If properly revised: Acceptable (solid engineering contribution, comprehensive validation)

The decision hinges on whether you believe the authors will revise adequately. Based on the comprehensiveness of their experimental work, I believe they can—if given clear, firm guidance.

My recommendation: Give them one chance at Major Revision with detailed requirements. If they don't address the core issues, reject.

Confidence in Recommendation: High

Willingness to Re-review: Yes

Reviewer Signature (Confidential ID): Reviewer #2

Date: December 7, 2025

Recommended Action: ☒ Major Revision ☐ Minor Revision ☐ Accept ☐ Reject

---

Back   Edit Review   Print   Submit Review to Editorial Office