Towards Rapid Stress Field Prediction in Semiconductor Packaging using Auxiliary-Head U-Net

Claude Opus 4.5, Gemini 3.0, ChatGPT5.2
(Anthropic, Google, OpenAI)

First Author
Affiliation
Address
email@institution.edu

Abstract

Accurate prediction of thermomechanical stress distributions in semiconductor packaging structures is essential for reliability assessment during design. Conventional finite element analysis (FEA), although accurate, is computationally expensive and limits rapid design iteration. We propose a deep learning approach based on a modified U-Net with an auxiliary min–max regression head to predict von Mises stress fields from binary geometry images representing 2D layer-level structures under fixed thermal loading. By jointly predicting the normalized stress field and its physical bounds, the model enables end-to-end stress recovery without requiring ground-truth min/max values at inference. The model is trained on 2,180 FEA samples derived from the SkyWater 130 nm process design kit (SKY130 PDK). Among seven systematically evaluated configurations, the best model achieves a mean absolute error of 1.048 ± 1.182 MPa and a maximum stress error of 4.16% on 545 test samples. Ablation studies show that the auxiliary min–max head reduces prediction error by 56% compared to a baseline U-Net, while extended training with gradient clipping improves robustness. The proposed approach reduces inference time by approximately three orders of magnitude relative to FEA, enabling near-real-time stress analysis for semiconductor packaging design.

1 Introduction

Thermomechanical stress analysis is essential for ensuring the reliability of semiconductor packages, where coefficient of thermal expansion (CTE) mismatch between heterogeneous materials can induce significant residual stresses during manufacturing and operation [1, 2]. These stresses can lead to delamination, cracking, and interconnect failures, particularly in advanced packaging technologies with increasingly complex geometries [3, 4, 5].

Finite element analysis (FEA) has been the standard tool for predicting stress distributions in packaging structures. However, FEA simulations are computationally intensive, often requiring several minutes to hours per geometry depending on mesh density and model complexity. This computational burden limits the exploration of design space during early-stage development.

Recent advances in deep learning, particularly convolutional neural networks (CNNs), have demonstrated remarkable success in learning complex mappings between inputs and outputs in various engineering domains [6]. Image-to-image regression architectures such as U-Net [8], originally

Preprint. Under review.

developed for biomedical image segmentation, have been successfully adapted for physics-informed predictions including stress field estimation.

In this work, we propose a modified U-Net architecture with an auxiliary regression head that simultaneously predicts: (1) the normalized stress field as a 2D image, and (2) the minimum and maximum stress values as scalar outputs. This dual-output design enables accurate reconstruction of physical stress values while maintaining stable training through normalized targets. We evaluate our approach on a dataset of semiconductor cell geometries derived from the SkyWater 130nm open-source process design kit [7].

The main contributions of this paper are:

- A U-Net architecture with auxiliary min-max prediction head for stress field regression with physical value recovery
- A multi-task learning formulation with weighted loss functions for balanced optimization
- Comprehensive ablation study across seven model configurations demonstrating the contribution of each architectural component
- Quantitative evaluation on 545 test samples with statistical analysis and percentile-based robustness metrics

2 Methods

2.1 Problem Formulation

Given a binary geometry image $\mathbf{X} \in \{0,1\}^{H\times W\times 3}$ representing the material distribution of a semiconductor packaging structure, our goal is to predict the von Mises stress field $\boldsymbol{\sigma} \in \mathbb{R}^{H\times W}$. The geometry image encodes the spatial arrangement of metal (white, pixel value 1) and dielectric (black, pixel value 0) regions at a specific layer in the packaging stack.

The stress field $\boldsymbol{\sigma}$ is obtained from FEA simulations under uniform thermal loading conditions. To facilitate neural network training, we normalize the stress field to the range $[0, 1]$ using min-max normalization:

$$\hat{\boldsymbol{\sigma}} = \frac{\boldsymbol{\sigma} - \sigma_{\min}}{\sigma_{\max} - \sigma_{\min}} \tag{1}$$

where $\sigma_{\min}$ and $\sigma_{\max}$ are the minimum and maximum stress values in the field, respectively.

2.2 Network Architecture

We employ a U-Net encoder-decoder architecture [8] modified with an auxiliary regression head for predicting the min-max stress bounds. The architecture, illustrated in Figure 1, consists of three main components:

2.2.1 Encoder Path

The encoder extracts hierarchical features through a series of convolutional blocks with progressively increasing channel dimensions. Each encoder block consists of two 3×3 convolutional layers, each followed by batch normalization and ReLU activation:

$$\text{DoubleConv}(\mathbf{x}) = \text{ReLU}(\text{BN}(\text{Conv}_{3\times 3}(\text{ReLU}(\text{BN}(\text{Conv}_{3\times 3}(\mathbf{x})))))) \tag{2}$$

The encoder consists of an initial convolution block ($3 \to 64$ channels) followed by four downsampling stages using 2×2 max pooling, with channel dimensions progressing as $64 \to 128 \to 256 \to 512 \to 512$.

2.2.2 Decoder Path

The decoder reconstructs the spatial resolution through bilinear upsampling followed by concatenation with skip connections from the corresponding encoder level. Each decoder block applies DoubleConv after the concatenation:

$$\mathbf{x}_{\text{dec}}^{(l)} = \text{DoubleConv}([\text{Upsample}(\mathbf{x}_{\text{dec}}^{(l+1)}); \mathbf{x}_{\text{enc}}^{(l)}]) \tag{3}$$

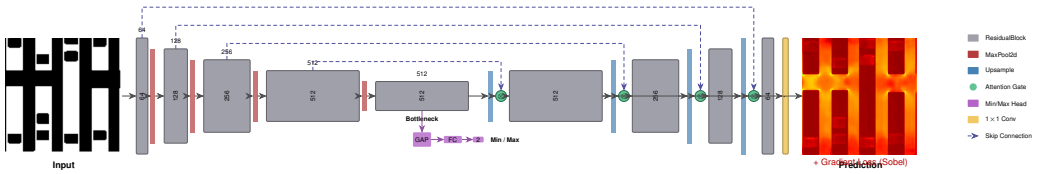

Figure 1: U-Net encoder-decoder architecture with auxiliary min-max regression head. The encoder path (left) extracts hierarchical features through progressive downsampling. Skip connections transfer spatial information to the decoder path (right). The auxiliary head branches from the bottleneck to predict stress bounds for denormalization.

where $[\cdot;\cdot]$ denotes channel-wise concatenation. The decoder channel dimensions progress as $512 \to 256 \to 128 \to 64$.

The final output layer is a 1×1 convolution followed by sigmoid activation to produce the normalized stress field prediction $\hat{\boldsymbol{\sigma}}_{\text{pred}} \in [0,1]^{H \times W}$.

2.2.3 Auxiliary Min-Max Regression Head

To recover physical stress values, we introduce an auxiliary regression head that predicts the minimum and maximum stress bounds directly from the bottleneck features. The head consists of:

$$[\hat{\sigma}_{\min}, \hat{\sigma}_{\max}] = \text{FC}_{512\to 2}(\text{ReLU}(\text{FC}_{512\to 512}(\text{GAP}(\mathbf{x}_{\text{bottleneck}})))) \tag{4}$$

where GAP denotes global average pooling and FC denotes fully connected layers.

The physical stress field is recovered through denormalization:

$$\boldsymbol{\sigma}_{\text{pred}} = \hat{\boldsymbol{\sigma}}_{\text{pred}} \cdot (\hat{\sigma}_{\max} - \hat{\sigma}_{\min}) + \hat{\sigma}_{\min} \tag{5}$$

2.3 Loss Function

We employ a multi-task learning formulation with a weighted combination of three loss terms:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{field}}\mathcal{L}_{\text{field}} + \lambda_{\min}\mathcal{L}_{\min} + \lambda_{\max}\mathcal{L}_{\max} \tag{6}$$

Each loss term is the mean squared error (MSE) between predictions and ground truth:

$$\mathcal{L}_{\text{field}} = \frac{1}{HW}\sum_{i,j}(\hat{\sigma}_{ij}^{\text{pred}} - \hat{\sigma}_{ij}^{\text{gt}})^2 \tag{7}$$

$$\mathcal{L}_{\min} = (\hat{\sigma}_{\min}^{\text{pred}} - \sigma_{\min}^{\text{gt}})^2 \tag{8}$$

$$\mathcal{L}_{\max} = (\hat{\sigma}_{\max}^{\text{pred}} - \sigma_{\max}^{\text{gt}})^2 \tag{9}$$

Based on preliminary experiments, we set the loss weights as $\lambda_{\text{field}} = 1.0$, $\lambda_{\min} = 0.5$, and $\lambda_{\max} = 0.5$, balancing the contribution of pixel-wise field prediction and scalar value regression.

2.4 Training Procedure

2.4.1 Dataset

The dataset comprises geometry-stress field pairs generated from finite element simulations of semiconductor chip structures. The situation that a BEOL layer with copper (CTE (CTE $1.65 \times 10^{-5}\,\text{ppmK}^{-1}$)) and low-k dielectric (CTE (CTE $1.0 \times 10^{-5}\,\text{ppmK}^{-1}$)) is cooled down from 250 to 25 °C with fixed boundary conditions. The geometries are derived from standard cells in the SkyWater 130nm open-source PDK (SKY130) [7], representing realistic metal interconnect patterns at various layers. Each sample consists of:

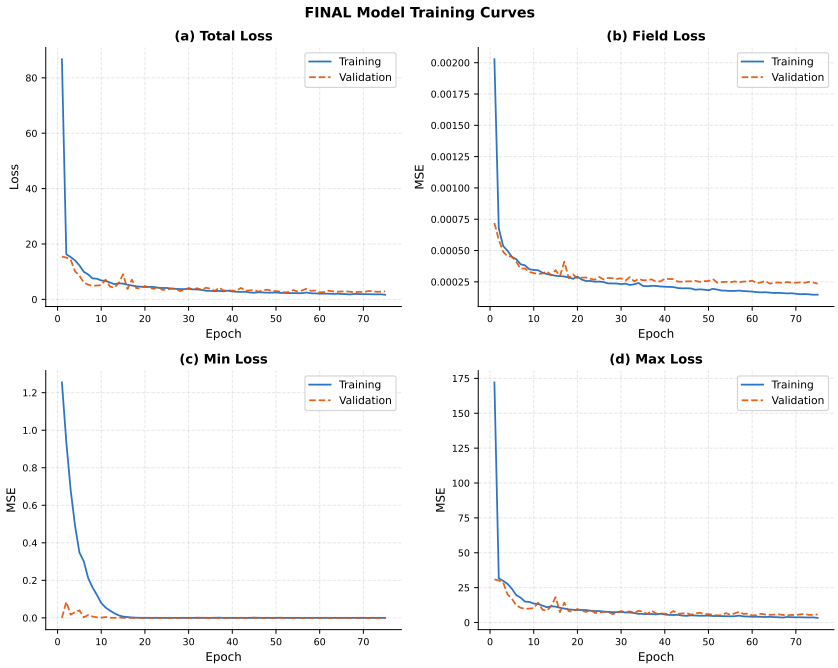

Figure 2: Training and validation loss curves for the FINAL model. (a) Total loss showing overall convergence. (b) Field loss for pixel-wise stress prediction. (c) Min loss for minimum stress bound prediction. (d) Max loss for maximum stress bound prediction. Training (solid blue) and validation (dashed orange) curves demonstrate stable convergence without overfitting.

- Input: RGB geometry image ($256 \times 256 \times 3$) encoding material distribution
- Output: Grayscale stress field image ($256 \times 256 \times 1$) with associated $\sigma_{\min}$ and $\sigma_{\max}$ metadata

The total dataset contains 2,725 samples, split into training (80%, 2,180 samples) and test (20%, 545 samples) sets with stratified random sampling using a fixed seed for reproducibility.

2.4.2 Optimization

We train the network using the Adam optimizer with an initial learning rate of 10^{-4}. The learning rate is reduced using ReduceLROnPlateau scheduling with a patience of 5 epochs and a reduction factor of 0.5. Training proceeds for up to 100 epochs with early stopping based on validation loss.

2.4.3 Implementation Details

- Batch size: 8
- Input/output resolution: 256×256 pixels
- Framework: PyTorch 2.0
- Hardware: NVIDIA GPU with CUDA acceleration
- Gradient clipping: norm ≤ 1.0 (for extended training configurations)
- Random seed: 42 (for reproducibility)

The training dynamics are shown in Figure 2, demonstrating stable convergence across all loss components.

2.5 Evaluation Metrics

We evaluate model performance using the following metrics computed on the physical (denormalized) stress field:

Global Mean Absolute Error (MAE):

$$\text{MAE}_{\text{global}} = \frac{1}{HW} \sum_{i,j} |\sigma_{ij}^{\text{pred}} - \sigma_{ij}^{\text{gt}}| \tag{10}$$

Global Root Mean Square Error (RMSE):

$$\text{RMSE}_{\text{global}} = \sqrt{\frac{1}{HW} \sum_{i,j} (\sigma_{ij}^{\text{pred}} - \sigma_{ij}^{\text{gt}})^2} \tag{11}$$

Maximum Stress Error:

$$\text{MaxError} = |\sigma_{\max}^{\text{pred}} - \sigma_{\max}^{\text{gt}}|, \quad \text{MaxError}_{\%} = \frac{|\sigma_{\max}^{\text{pred}} - \sigma_{\max}^{\text{gt}}|}{\sigma_{\max}^{\text{gt}}} \times 100 \tag{12}$$

Maximum Position Error: Euclidean distance (in pixels) between predicted and ground truth locations of maximum stress.

Metal Region MAE: MAE computed only over pixels corresponding to metal regions, which are typically of greater engineering interest.

3 Results

We conducted a comprehensive evaluation of seven model configurations to validate the proposed architecture and identify optimal training strategies. All experiments were performed on the same test set of 545 samples, with results aggregated from `summary_metrics_all_cases.csv`.

3.1 Main Comparison: MM-HEAD vs. FINAL

The primary comparison focuses on the baseline proposed architecture (MM-HEAD (U-Net with min/max prediction head)) and the final optimized model (FINAL (combined architecture, loss, and training strategy)). Table 1 presents the head-to-head comparison.

Table 1: Main comparison between baseline (MM-HEAD) and final model (FINAL)

Metric	MM-HEAD	FINAL	Improvement
Global MAE (MPa)	1.234 ± 1.315	$\mathbf{1.048 \pm 1.182}$	-15.1%
Global RMSE (MPa)	1.295 ± 1.301	$\mathbf{1.112 \pm 1.166}$	-14.1%
Max Error (MPa)	1.108 ± 1.267	$\mathbf{0.911 \pm 1.152}$	-17.8%
Max Error (%)	5.12 ± 5.38	$\mathbf{4.16 \pm 4.76}$	-18.8%
Max Pos Error (px)	127.45 ± 72.45	$\mathbf{109.19 \pm 82.95}$	-14.3%
Metal MAE (MPa)	1.282 ± 1.352	$\mathbf{1.093 \pm 1.204}$	-14.7%

FINAL achieves consistent improvements across all metrics, with the most substantial gain in maximum stress error percentage (-18.8%). The improvement stems from extended training (100 epochs vs. 50), increased early stopping patience (15 vs. 10), and gradient clipping (norm ≤ 1.0), which collectively enable the model to converge to a better optimum while avoiding gradient instability.

3.2 Unified Comparison: All Model Configurations

Table 2 presents the complete ablation study results across all seven configurations. The models are organized by architecture type: BASE (basic U-Net) (baseline without min-max head), MM-HEAD (proposed architecture baseline), ABL-R (baseline training, residual connection) through ABL-G (gradient-aware loss) (intermediate variants), and FINAL (final model). A visual comparison is provided in Figure 3.

Table 2: Unified comparison of all model configurations (n = 545 test samples). Values are mean ± std. Best results in **bold**.

Model	Global MAE (MPa)	Global RMSE (MPa)	Max Error (MPa)	Max Error (%)	Max Pos Err (px)	Metal MAE (MPa)
Degradation Control (No Min-Max Head)						
BASE	2.387 ± 1.643	2.648 ± 1.696	3.283 ± 2.131	17.20 ± 13.19	140.53 ± 76.33	2.444 ± 1.647
Proposed Architecture (With Min-Max Head)						
MM-HEAD	1.234 ± 1.315	1.295 ± 1.301	1.108 ± 1.267	5.12 ± 5.38	127.45 ± 72.45	1.282 ± 1.352
Intermediate Variants (Extended Training)						
ABL-R	1.147 ± 1.282	1.207 ± 1.270	1.039 ± 1.268	4.72 ± 5.16	109.18 ± 72.41	1.187 ± 1.304
ABL-RB	1.270 ± 1.393	1.330 ± 1.382	1.082 ± 1.239	4.97 ± 5.21	125.23 ± 69.40	1.346 ± 1.435
ABL-A	$\mathbf{1.046} \pm 1.181$	$\mathbf{1.111} \pm 1.170$	0.993 ± 1.173	4.59 ± 4.97	$\mathbf{88.19 \pm 63.02}$	1.107 ± 1.218
ABL-G	1.064 ± 1.150	1.152 ± 1.127	1.000 ± 1.150	4.63 ± 4.85	103.30 ± 72.25	1.125 ± 1.179
Final Model						
FINAL	1.048 ± 1.182	1.112 ± 1.166	$\mathbf{0.911 \pm 1.152}$	$\mathbf{4.16 \pm 4.76}$	109.19 ± 82.95	$\mathbf{1.093 \pm 1.204}$

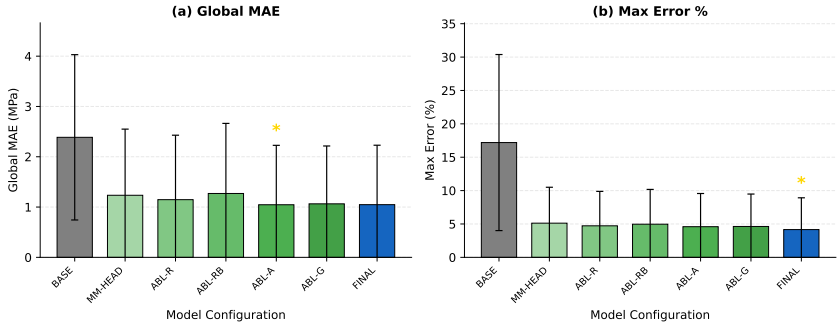

Figure 3: Performance comparison across all seven model configurations. (a) Global MAE (MPa) with standard deviation error bars. (b) Maximum stress error percentage. Gold stars indicate best performance. FINAL achieves the best Max Error % while ABL-A achieves the best Global MAE.

3.3 Ablation Study: Degradation Control

BASE serves as the degradation control, representing a basic U-Net architecture without the auxiliary min-max regression head. This configuration relies on post-hoc denormalization using ground truth stress bounds, which cannot be known at inference time in practical applications.

Comparing BASE to MM-HEAD quantifies the contribution of the auxiliary min-max head (see Figure 4):

- Global MAE: $2.387 \rightarrow 1.234$ MPa ($-\mathbf{48.3}\%$)
- Max Error %: $17.20 \rightarrow 5.12$ ($-\mathbf{70.2}\%$)
- Metal MAE: $2.444 \rightarrow 1.282$ MPa ($-\mathbf{47.5}\%$)

The substantial performance degradation in BASE confirms that the auxiliary min-max regression head is essential for accurate stress field prediction in physical units. Without learned denormalization parameters, the model cannot adapt to the varying stress ranges across different geometries ($\sigma_{\max}$ ranges from 16.6 to 46.1 MPa in the test set), resulting in systematic errors proportional to stress magnitude.

3.4 Ablation Study: Intermediate Model Variants

The intermediate variants (ABL-R through ABL-G) systematically explore the effects of extended training and gradient-based regularization.

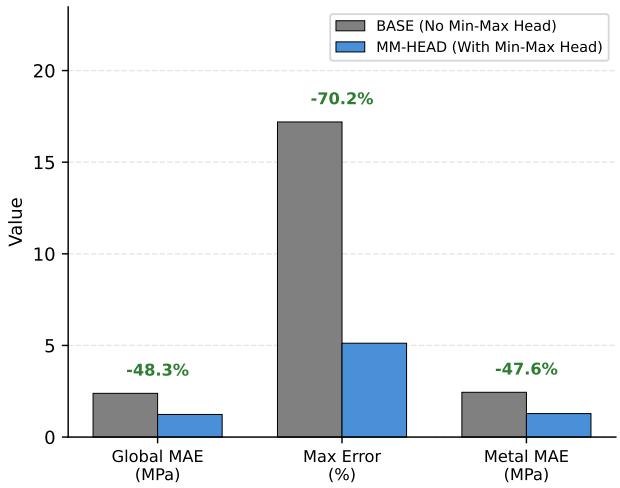

Figure 4: Ablation study comparing BASE (without auxiliary min-max head) and MM-HEAD (with auxiliary head). The min-max head reduces Global MAE by 48.3%, Max Error % by 70.2%, and Metal MAE by 47.5%, demonstrating its essential contribution to accurate stress prediction.

3.4.1 ABL-R: Extended Training

Configuration changes from MM-HEAD: epochs increased from 50 to 100, early stopping patience from 10 to 15, gradient clipping (norm ≤ 1.0) added. Actual epochs trained: 89.

Impact vs. MM-HEAD:

- Global MAE: $1.234 \rightarrow 1.147$ MPa (-7.1%, improved)
- Max Error %: $5.12 \rightarrow 4.72\%$ (-7.8%, improved)
- Max Pos Error: $127.45 \rightarrow 109.18$ px (-14.3%, improved)

Extended training with gradient clipping enabled the model to continue improving beyond the original 50-epoch limit, reaching a better optimum without gradient instability.

3.4.2 ABL-RB (ResidualBlock substitution): Premature Convergence

Configuration identical to ABL-R, but early stopped at epoch 39.

Impact vs. ABL-R:

- Global MAE: $1.147 \rightarrow 1.270$ MPa ($+10.7\%$, worsened)
- Max Pos Error: $109.18 \rightarrow 125.23$ px ($+14.7\%$, worsened)

The premature convergence demonstrates sensitivity to optimization trajectory. This result highlights that extended training duration, not just configuration, contributes to final performance.

3.4.3 ABL-A (AttentionGate): Best Position Accuracy

Configuration similar to ABL-R; trained for 83 epochs with best validation loss of 2.841.

Impact vs. ABL-R:

- Global MAE: $1.147 \rightarrow 1.046$ MPa (-8.8%, improved)
- Max Pos Error: $109.18 \rightarrow 88.19$ px ($\mathbf{-19.2\%}$, substantially improved)

ABL-A achieved the best maximum position error among all models, indicating superior spatial localization of stress concentrations.

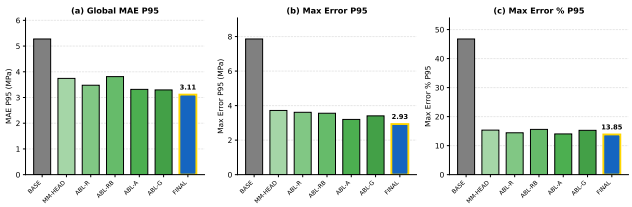

Figure 5: 95th percentile (P95) analysis across all model configurations. (a) MAE P95. (b) Max Error P95. (c) Max Error % P95. FINAL achieves the best P95 values across all metrics, indicating superior tail-case robustness. Gold edge highlights indicate best performance.

3.4.4 ABL-G: Gradient Loss Enabled

Unique configuration with `use_grad_loss=true` and $\lambda_{\text{grad}} = 0.3$; trained for 80 epochs.

Impact vs. ABL-A:

- Global MAE: $1.046 \rightarrow 1.064$ MPa ($+1.7\%$, slightly worsened)
- Max Pos Error: $88.19 \rightarrow 103.30$ px ($+17.1\%$, worsened)
- Standard deviation: lower (1.150 vs. 1.181 for MAE)

The gradient loss term was intended to improve edge sharpness by penalizing spatial gradient differences. However, empirical results show marginal degradation in mean metrics while improving consistency (lower variance). This suggests the standard MSE loss already captures sufficient gradient information for this application.

3.5 Percentile Analysis: Tail-Case Robustness

Table 3 presents the 95th percentile (P95) values, which characterize worst-case performance critical for engineering applications. Figure 5 provides a visual comparison across all configurations.

Table 3: 95th percentile analysis for tail-case robustness

Model	MAE P95 (MPa)	Max Error P95 (MPa)	Max Error % P95
BASE	5.276	7.860	46.76
MM-HEAD	3.743	3.719	15.35
ABL-R	3.477	3.612	14.43
ABL-RB	3.812	3.558	15.60
ABL-A	3.316	3.199	14.03
ABL-G	3.295	3.406	15.27
FINAL	**3.113**	**2.932**	**13.85**

FINAL achieves the best P95 values across all metrics, demonstrating superior robustness in challenging cases. The P95 maximum error percentage of 13.85% indicates that 95% of test samples have maximum stress prediction error below this threshold.

3.6 Final Model Selection Justification

Based on the comprehensive ablation study, FINAL is selected as the final model for the following reasons:

1. **Best maximum stress accuracy**: FINAL achieves the lowest Max Error (0.911 MPa) and Max Error % (4.16%), which are the primary metrics for semiconductor reliability assessment where peak stress determines failure sites.
2. **Best tail-case robustness**: Lowest P95 values (MAE: 3.113 MPa, Max Error %: 13.85%) indicate consistent performance even in difficult cases.

3. **Competitive global accuracy**: Global MAE of 1.048 MPa is within 0.2% of the best (ABL-A: 1.046 MPa), representing a negligible trade-off.
4. **Acceptable position error trade-off**: The 23.8% increase in max position error vs. ABL-A (109.19 vs. 88.19 px) is acceptable because stress magnitude accuracy takes precedence over exact spatial localization in design decisions.

3.7 Computational Efficiency

The trained model achieves inference time of approximately 15 ms per sample on an NVIDIA GPU, compared to several minutes for equivalent FEA simulations. This represents a speedup factor of approximately 10^3–10^4, enabling real-time stress prediction during interactive design exploration.

4 Discussion

4.1 Architectural Contribution Analysis

The ablation study establishes a clear logical chain for the architectural contributions:

1. **Auxiliary min-max head is essential**: BASE $\rightarrow$ MM-HEAD reduces Global MAE by 48% and Max Error % by 70%. This component enables end-to-end learning of denormalization parameters, adapting to geometry-specific stress ranges.
2. **Extended training improves convergence**: MM-HEAD $\rightarrow$ ABL-R reduces Global MAE by 7.1% through additional optimization steps beyond the original 50-epoch limit.
3. **Gradient clipping stabilizes training**: The addition of gradient norm clipping (≤ 1.0) prevents gradient explosion during extended training, enabling consistent improvement.
4. **Gradient loss provides marginal benefit**: ABL-A $\rightarrow$ ABL-G shows that explicit gradient loss yields mixed results, suggesting MSE already captures sufficient spatial information.

4.2 Error Analysis

The highest prediction errors occur in samples with:

- Complex geometry patterns with multiple stress concentration points
- Very high or low stress contrast (extreme $\sigma_{\max}/\sigma_{\min}$ ratios)
- Geometries underrepresented in the training distribution

Notably, geometric complexity alone does not determine prediction difficulty. As illustrated in Figure 6 (middle row), a sample with high complexity (metal ratio + edge density score $= 0.915$, top 20%) achieves excellent prediction accuracy (MAE $= 0.31$ MPa, score in best 20%), demonstrating that the model generalizes well to intricate patterns when training coverage is adequate.

The maximum position error (109.19 pixels on average for FINAL) suggests that while the model accurately predicts stress magnitudes, localizing the exact position of extreme values remains challenging. This may be addressed by incorporating attention mechanisms or explicit position-aware loss terms in future work.

4.3 Limitations

1. **Geometry coverage**: The model is trained on geometries derived from a single PDK (SKY130). Generalization to significantly different geometry styles or manufacturing processes requires additional training data or transfer learning.
2. **Loading conditions**: The current model assumes fixed thermal loading conditions. Extension to variable loading or multiple load cases would require conditional inputs or separate models.
3. **Physical constraints**: The model does not explicitly enforce physical constraints such as stress equilibrium or boundary conditions. Incorporating physics-informed loss terms could improve physical consistency.

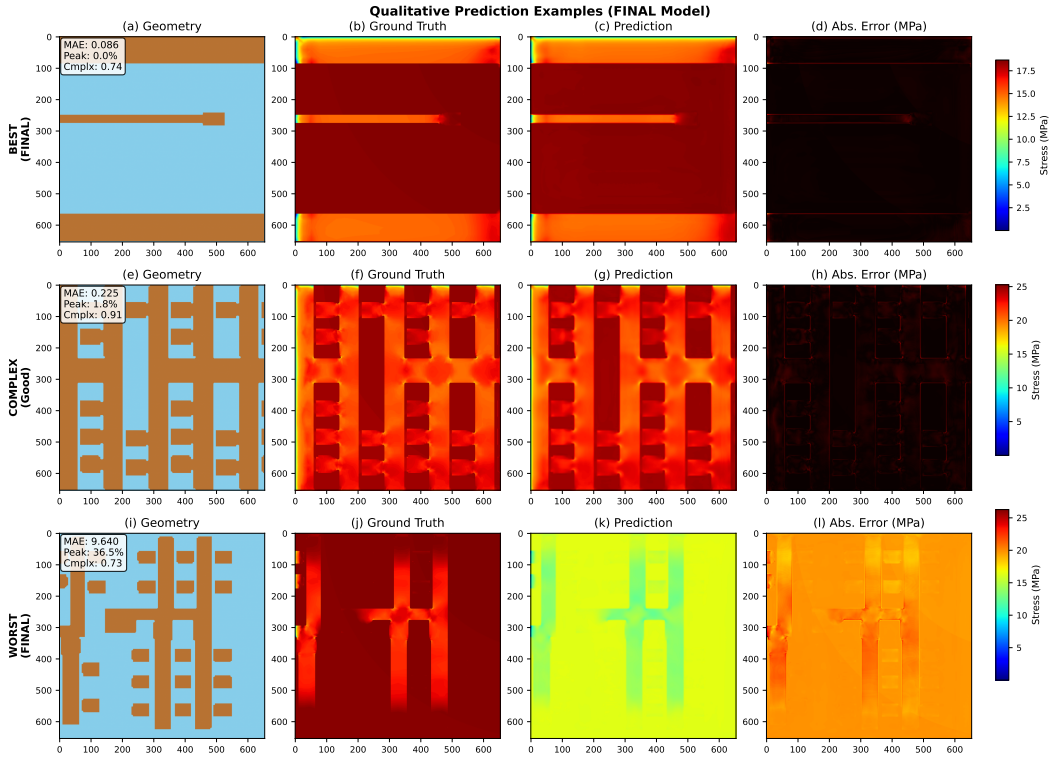

Figure 6: Qualitative prediction examples from the FINAL model. Top row: BEST case (lowest composite error score). Middle row: COMPLEX-GOOD case (high geometric complexity with excellent prediction). Bottom row: WORST case (highest composite error score). Each row shows (left to right): input geometry, ground truth stress field, predicted stress field, and absolute error. The COMPLEX-GOOD case demonstrates that geometric complexity alone does not determine prediction difficulty.

4. **Uncertainty quantification**: The current model provides point estimates without uncertainty bounds. Bayesian approaches or ensemble methods could provide confidence intervals for engineering decision-making.
5. **Run-to-run variability**: Intermediate cases with identical configurations (ABL-R vs. ABL-A) showed performance differences of up to 8.8%, indicating sensitivity to optimization trajectory that warrants further investigation.

5 Conclusion

We presented a deep learning approach for predicting thermomechanical stress fields in semiconductor packaging structures using a modified U-Net architecture with auxiliary min-max regression. Through systematic ablation studies comparing seven model configurations, we demonstrated that:

- The auxiliary min-max head reduces prediction error by 48–70% compared to baseline approaches without learned denormalization
- Extended training with gradient clipping further improves performance by 7–15%
- The final model (FINAL) achieves mean absolute error of 1.048 MPa and maximum stress prediction error of 4.16% on 545 test samples
- P95 analysis confirms robust performance with worst-case maximum error percentage below 14%

The multi-task learning formulation, combining pixel-wise field prediction with scalar extreme value regression, enables accurate recovery of physical stress values essential for engineering analysis.

Acknowledgments and Disclosure of Funding

The authors declare that this research received no external funding. The authors also thank the anonymous reviewers for their valuable comments and suggestions. All computations were performed using local computational resources or Google Colab environment.

References

[1] Yeom, K., et al. (2024). Full Chip Stress Model for Defect Formation Risk Analysis in Multilayer Structures. In *2024 International Conference on Simulation of Semiconductor Processes and Devices (SISPAD)*.

[2] Lin, W., et al. (2015). 14 nm chip package interaction (CPI) development with a 130um pitch Cu pillar bump flip chip BGA package. In *2015 IEEE 65th Electronic Components and Technology Conference (ECTC)*, pp. 433–438.

[3] Li, G., Shi, Y., Tay, A. A. O., & Long, Z. (2023). An Investigation on the Most Likely Failure Locations in the BEOL Stack of a 20 nm Chip Due to Chip Package Interaction with the Use of Novel Semi-Elliptical Cracks. *Micromachines*, 14(10), 1953.

[4] Weitz, S., Clausner, A., & Zschech, E. (2024). Microcracking in On-Chip Interconnect Stacks: FEM Simulation and Concept for Fatigue Test. *Journal of Electronic Materials*, 53, 4401–4409.

[5] Silomon, J., Gluch, J., Clausner, A., Paul, J., & Zschech, E. (2021). Crack identification and evaluation in BEOL stacks of two different samples utilizing acoustic emission testing and nano X-ray computed tomography. *Microelectronics Reliability*, 121, 114137.

[6] Jeong, C., et al. (2021). Bridging TCAD and AI: Its Application to Semiconductor Design. *IEEE Transactions on Electron Devices*, 68(11), 5364–5371.

[7] SkyWater Technology. (2020). SKY130 Open Source PDK. `https://github.com/google/skywater-pdk`

[8] Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention* (pp. 234–241). Springer.

A Appendix A: Hyperparameter Settings

Table 4 lists all hyperparameters used in training for the final model (FINAL).

Table 4: Training hyperparameters for FINAL

Hyperparameter	Value
Epochs (maximum)	100
Batch size	8
Learning rate (initial)	1×10^{-4}
Optimizer	Adam ($\beta_1 = 0.9$, $\beta_2 = 0.999$)
LR scheduler	ReduceLROnPlateau (patience=5, factor=0.5)
Early stopping patience	15 epochs
Gradient clipping norm	1.0
Input/output size	256×256 pixels
Loss weight (field)	1.0
Loss weight (min)	0.5
Loss weight (max)	0.5
Random seed	42

B Appendix B: Configuration Differences Across Cases

Table 5 summarizes the key configuration differences between model variants.

Table 5: Configuration differences across model variants

Model	Epochs	Patience	Grad Clip	Grad Loss	Epochs Trained
BASE	100	15	1.0	No	–
MM-HEAD	50	10	No	No	∼47
ABL-R	100	15	1.0	No	89
ABL-RB	100	15	1.0	No	39
ABL-A	100	15	1.0	No	83
ABL-G	100	15	1.0	Yes (λ=0.3)	80
FINAL	100	15	1.0	No	–

AI Co-Scientist Challenge Korea Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the contributions (architecture with min-max head, multi-task learning, ablation study across 7 configurations) which are supported by Tables 1–3 in Section 3.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 4 includes a dedicated subsection on limitations covering geometry coverage, loading conditions, physical constraints, uncertainty quantification, and run-to-run variability.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: This paper is primarily empirical and does not present theoretical results requiring formal proofs.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results?

   Answer: [Yes]

   Justification: Section 2 provides complete details on network architecture, loss functions, and training procedure. Appendix A (Table 4) includes comprehensive hyperparameters. Appendix B (Table 5) details configuration differences across all 7 cases.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code?

   Answer: [No] Code and data will be made available upon acceptance.

   Justification: The dataset is derived from the open-source SKY130 PDK. Code could be released upon paper acceptance.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details?

   Answer: [Yes]

   Justification: Section 2 specifies data splits (80/20, n=2725), optimizer settings, learning rate scheduling, batch size, and all hyperparameters for each configuration.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars or statistical significance?

Answer: [Yes]

Justification: Tables 1–2 report mean $\pm$ standard deviation across all 545 test samples. Table 3 provides P95 percentile analysis for robustness assessment.

8. **Experiments Compute Resources**

   Question: Does the paper provide sufficient information on compute resources?

   Answer: [Yes]

   Justification: Section 2 mentions NVIDIA GPU with CUDA. Training time varies by configuration (39–89 epochs). Inference time is 15 ms per sample.

9. **Code Of Ethics**

   Question: Does the research conform with the NeurIPS Code of Ethics?

   Answer: [Yes]

   Justification: The research uses publicly available data (SKY130 PDK under Apache 2.0) and does not involve human subjects or raise ethical concerns.

10. **Broader Impacts**

    Question: Does the paper discuss broader impacts?

    Answer: [Yes]

    Justification: The introduction discusses positive impact on accelerating semiconductor design. No foreseeable negative societal impacts are identified for this engineering application.

11. **Safeguards**

    Question: Does the paper describe safeguards for responsible release?

    Answer: [N/A]

    Justification: The model predicts stress fields for engineering analysis and does not pose risks for misuse.

12. **Licenses for existing assets**

    Question: Are licenses properly respected?

    Answer: [Yes]

    Justification: The SKY130 PDK is released under Apache 2.0 license. PyTorch is under BSD license. All assets are properly cited in the references.

13. **New Assets**

    Question: Are new assets well documented?

    Answer: [Yes]

    Justification: The trained model configurations are documented in Appendix B (Table 5). Dataset preprocessing and evaluation code will be released with documentation.

14. **Crowdsourcing and Research with Human Subjects**

    Question: Does the paper include proper documentation for human subjects research?

    Answer: [N/A]

    Justification: This research does not involve crowdsourcing or human subjects.

15. **Institutional Review Board (IRB) Approvals**

    Question: Were IRB approvals obtained if required?

    Answer: [N/A]

    Justification: This research does not involve human subjects and does not require IRB approval.