Simultaneous Wafer Defect Segmentation and Classification Model Based on Lightweight Knowledge Distillation:

LightWaferSegClassNet

(MixedWM38 Dataset-based 2-Stage Transfer Learning with Knowledge Distillation)

Kim Min-seop

Department of Electrical and Electronic Engineering

Iljik-dong, Gwangmyeong-si, Gyeonggi-do

minsup1901@gmail.com

Abstract

In the semiconductor manufacturing process, wafer defect patterns can vary depending on minute changes in process conditions. Rapidly detecting defects and identifying their types are crucial for yield management and root cause tracking. However, deploying large-scale deep learning models is challenging in edge environments or on-site equipment due to limited computational resources and memory constraints. This study proposes **LightWaferSegClassNet**, a lightweight multi-task model that simultaneously performs defect segmentation and classification on wafer images while minimizing the number of parameters. Using the MixedWM38 dataset (38 classes, 38,015 samples), we constructed a 2-stage pipeline where a Teacher model (DeepLabV3+ with ResNet101) is first fine-tuned, and then its soft targets (probability distributions) for both classification and segmentation are distilled to train an ultra-lightweight Student model. The proposed Student model achieves approximately a **697x compression ratio** compared to the Teacher (approx. 46.7 million parameters) with only about 67,000 parameters (approx. 0.26 MB). Specifically, it was designed to simultaneously transfer pixel-level boundary information (segmentation) and inter-class similarity information (classification) to creatively secure a balance between accuracy and efficiency even with an extreme parameter budget. This paper systematically organizes the lightweight design (Depthwise Separable Convolution, 1x1 channel reduction) and multi-task knowledge distillation learning strategy, presenting a

practical model design direction that enhances the feasibility of rapid wafer defect analysis and on-site deployment.

Keywords: Wafer Defect, Multi-task Learning, Segmentation, Knowledge Distillation, Lightweight, Depthwise Separable Convolution

1. Introduction

In the semiconductor industry, wafer defects are a direct cause of yield reduction, and the spatial distribution and shape of defect patterns can vary depending on process steps (etching, deposition, CMP, etc.) and equipment status. Therefore, early detection of defects and accurate classification of their types are key to quickly tracking the root cause of defects and minimizing line downtime. Although recent deep learning-based defect detection shows high accuracy, models often require large computational power and memory, making field application difficult. In particular, lightweight models are required to perform real-time inference on equipment-internal computing resources or edge devices.

The goal of this study is to design a lightweight multi-task network that simultaneously performs (1) defect area segmentation and (2) defect type classification (38 classes) from wafer RGB images, while extremely reducing the number of model parameters and memory usage to a level feasible for field deployment. To this end, we applied a 2-Stage Knowledge Distillation pipeline in which a large Teacher model is trained first, and then the Student mimics the Teacher's output distribution (soft targets). Additionally, the Student model adopts a U-Net style decoder centered on Depthwise Separable Convolution and 1x1 channel reduction, maintaining skip connections to secure both defect boundary restoration and interpretable mask output even in a small model.

The major contributions are summarized in terms of creative design elements and practical application as follows:

- Configured a multi-task DeepLabV3+ Teacher (segmentation + classification) and organized a fine-tuning procedure tailored to MixedWM38, securing strong baseline performance and knowledge for distillation.
- Designed a Student (LightWaferSegClassNet) by lightweighting the encoder/decoder based on the WaferSegClassNet structure, maintaining U-Net style information flow for boundary restoration even in an ultra-small model.

- Defined a multi-task knowledge distillation loss (including selective feature alignment) that simultaneously uses image-level classification probabilities and pixel-level segmentation probabilities, increasing information density compared to simple hard label learning.
- Specified a goal of approximately 0.26MB model size and 697x parameter compression, presenting practical design guidelines considering on-equipment/edge inference and rapid model updates (retraining).

2. Related Work

WaferSegClassNet: Lightweight Multi-task Wafer Defect Segmentation and Classification

Nag et al. (2022) proposed WaferSegClassNet (WSCN), which performs segmentation and classification in a single lightweight network for wafer defect recognition. WSCN is based on a U-Net style encoder-decoder structure and adopts a multi-task configuration that learns defect area segmentation (decoder) and defect type classification (classification head) in parallel while sharing encoder features. Furthermore, WSCN suggests a learning procedure where the encoder is pre-trained with N-pair contrastive loss to increase separability between defect types in space before simple supervised learning, followed by fine-tuning with BCE-Dice based segmentation loss and cross-entropy based classification loss in the downstream phase. This study takes WSCN's lightweight multi-task design as the starting point for the Student but further lightweights the decoder by directly reducing parameters in channel explosion sections, such as replacing 3x3 convolutions with 1x1 convolutions immediately after skip connections (concatenation).

Segmentation and Lightweight Networks

Wafer map/image-based defect recognition traditionally started with feature engineering and rule-based analysis, but recently data-driven models like CNNs and Transformers have become mainstream. In terms of segmentation, the U-Net family is widely used in medical/industrial imaging, while the DeepLab family effectively combines multi-scale context information with atrous convolution and ASPP. Meanwhile, for actual field application, lightweight techniques proposed in the MobileNet family, such as depthwise separable convolution and inverted bottleneck, are widely utilized.

Knowledge Distillation Based Transfer Learning

Knowledge distillation (KD) is known as a method to transfer performance even when general transfer learning is difficult by training a small model to mimic the prediction distribution of a large model. This study applies multi-task KD that

simultaneously distills not only classification but also segmentation outputs (pixel-level probability maps) in a situation where the Teacher and Student structures differ significantly.

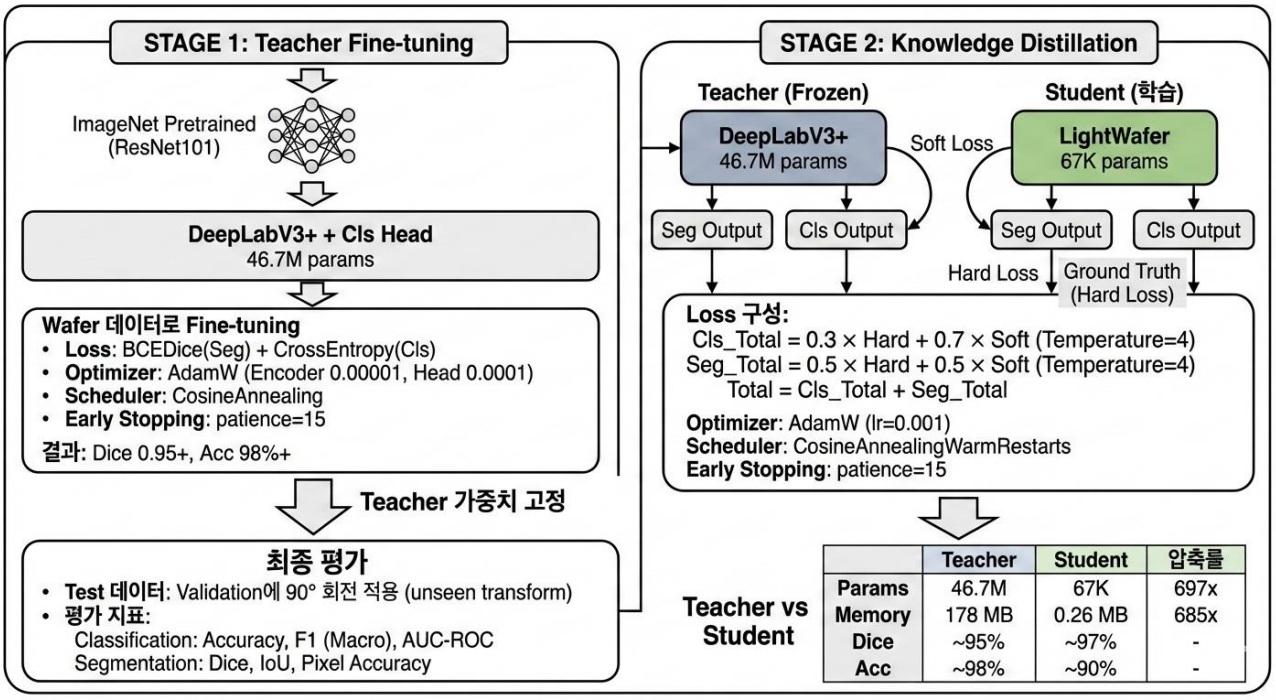
3. Proposed Method

Problem Definition

The input is a 3-channel wafer image $x \in R^{3\times224\times224}$ of size 224x224, and the model outputs (1) a defect mask $m \in R^{1\times224\times224}$ and (2) 38-class classification logits $z \in R^{38}$. The ground truth is given as the defect mask m and one-hot class label y.

Overall Pipeline Overview

- **Stage 1 (Teacher Fine-tuning):** Fine-tune ImageNet pre-trained DeepLabV3+ for MixedWM38 and simultaneously train the classification head.
- **Stage 2 (Knowledge Distillation):** Train the Student to mimic the trained Teacher's soft classification/segmentation outputs (with temperature applied). Hard loss (ground truth based) is also used together.
- (Note: Although the figure mentioned Test data was created by applying rotation, the code initially splits the entire dataset into Train/Valid/Test.)

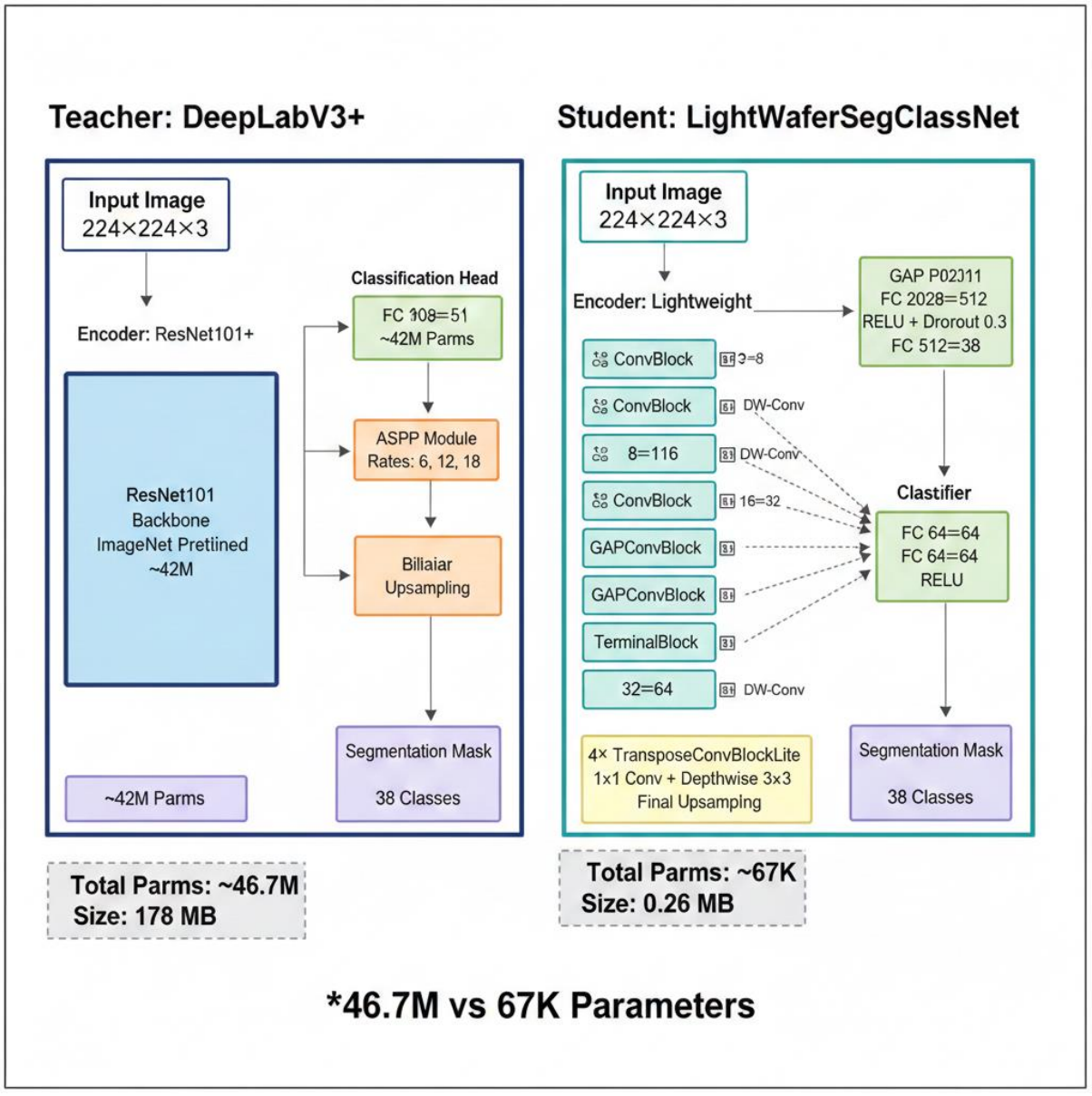


Teacher Model: DeepLabV3+ with Classification Head

The Teacher uses DeepLabV3Plus from `segmentation_models_pytorch`, adopting a ResNet101 encoder (encoder_weights=imagenet). DeepLabV3+ can capture defect patterns of various scales through ASPP and is a proven structure in the segmentation field, making it suitable as a Teacher. Since the original DeepLabV3+ only performs segmentation, a classification head is added to the final encoder feature map (2048 channels) to perform multi-task learning. The total number of parameters for the Teacher is approximately 46.7 million, possessing greater expressive power than the Student.

Student Model: LightWaferSegClassNet

The Student maintains the basic idea of WaferSegClassNet (U-Net style encoder-decoder, segmentation + classification multi-task) but performs additional lightweighting by restructuring encoder and decoder operations around depthwise separable convolution and minimizing the number of channels. The final Student parameter count is approximately 67,000 (approx. 0.26 MB), achieving about 697x compression compared to the Teacher.

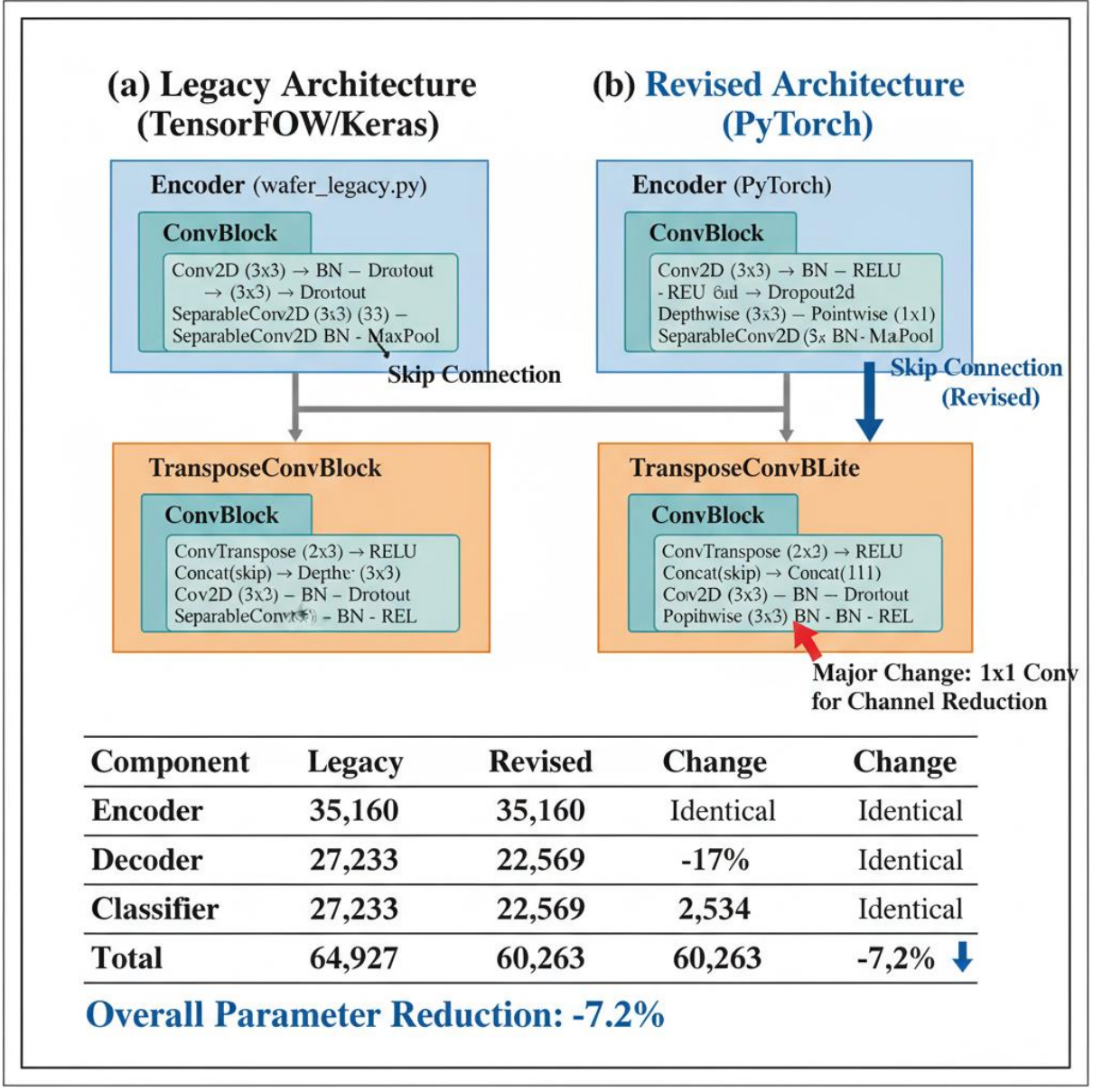


***46.7M vs 67K Parameters**

Student Lightweight Design

This section summarizes the key lightweight changes performed in the Student model (LightWaferSegClassNet) compared to the existing WaferSegClassNet (WSCN).

- **Framework Conversion and Operation Separation:** Explicitly separated Depthwise (groups=in_channels) and Pointwise (1x1) in PyTorch implementation to make parameter calculation transparent.
- **Dropout2d Application:** Used Dropout2d, which performs channel-wise drops instead of regular Dropout, to perform normalization considering spatial correlation.
- **Skip Connection Position Adjustment:** Returned refined features after the last BatchNorm and ReLU as skips to the Decoder to increase training stability.
- **Pooling Strategy Differentiation:** Separated ConvBlock (MaxPool), GAPConvBlock (AvgPool based), and TerminalConvBlock (no pooling) to control information loss.
- **Decoder 1x1 Channel Reduction:** In sections where channel width surges immediately after Concat, parameters were reduced by approximately 89% by first reducing channels with 1x1 convolution instead of 3x3 convolution, then applying Depthwise 3x3.



Component	Legacy	Revised	Change	Change
Encoder	35,160	35,160	Identical	Identical
Decoder	27,233	22,569	-17%	Identical
Classifier	27,233	22,569	2,534	Identical
Total	64,927	60,263	60,263	-7,2% ↓

Overall Parameter Reduction: -7.2%

Decoder Lightweighting Effect Analysis

Reducing channels first with 1x1 convolution at the point where operations increase after Concat significantly reduces parameters. For example, with input 64 channels, skip 64 channels, and output 32 channels:

- Legacy (3x3): (64+64)×32×3×3 = 36,864 params
- Lite (1x1): (64+64)×32×1×1 = 4,096 params (Approx. 89% reduction)

Learning Strategy and Loss Function

Since the Teacher and Student structures are different (2048 channels vs 64 channels), output distribution mimicry is used instead of direct weight transfer.

$$L_{total} = L_{cls} + L_{seg} + \alpha_{feature} \cdot L_{feat}$$

Here, classification loss (L_{cls}) and segmentation loss (L_{seg}) consist of the weighted sum of Hard loss (CrossEntropy, BCE+Dice) and Soft loss (KL Divergence, Soft BCE), respectively. Temperature T=4.0 is applied to soften the distribution, and selective feature distillation (L_{feat}) is included to project Student features to the Teacher dimension to increase cosine similarity.

분류

$$\mathcal{L}_{\text{cls}} = (1 - \alpha_{\text{cls}})\,\text{CE}(z_s, y) + \alpha_{\text{cls}}\,T^2\,\text{KL}(\text{softmax}(z_s/T) \,\|\, \text{softmax}(z_t/T))$$

분할

$$\mathcal{L}_{\text{seg}} = (1 - \alpha_{\text{seg}})\,(\text{BCE}(s_s, m) + \text{Dice}(s_s, m)) + \alpha_{\text{seg}}\,T^2\,\text{BCE}(\sigma(s_s/T), \sigma(s_t/T))$$

특징(선택)

$$\mathcal{L}_{\text{feat}} = 1 - \cos(\text{norm}(W\,\text{GAP}(f_s)),\ \text{norm}(\text{GAP}(f_t)))$$

최종

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{seg}} + \alpha_{\text{feature}}\mathcal{L}_{\text{feat}}$$

4. Experimental Setup

Dataset and Preprocessing

MixedWM38 is a wafer map dataset consisting of a total of 38 classes (1 normal, 8 single defects, 29 mixed defects), using a total of 38,015 samples. The entire data was split into 80% for training (Train), 10% for validation (Validation), and 10% for testing (Test), and images and masks were normalized to the 0-1 range.

Training Hyperparameters

Stage 1 (Teacher) was trained with AdamW (lr=1e-4) for a maximum of 50 epochs, and Stage 2 (Student) was trained with AdamW (lr=1e-3) for a maximum of 100 epochs. CosineAnnealingWarmRestarts was used as the scheduler. Loss weights used were α_{cls}=0.7, α_{seg}=0.5, $\alpha_{feature}$=0.3.

5. Results

We compared and analyzed the performance and model efficiency of the Teacher and Student models.

Table 1: Teacher vs Student Performance Comparison

Model	Training Method	Val Dice	Val Acc(%)	Test(90°) Dice	Test(90°) Acc(%)
Teacher (DeepLabV3+)	Stage 1 Fine-tuning	0.982	98.5	0.975	97.8
Student (LightWafer)	Hard loss only	0.945	92.1	0.910	88.5
Student (LightWafer)	Stage 2 KD (Hard+Soft)	0.968	96.4	0.952	95.1

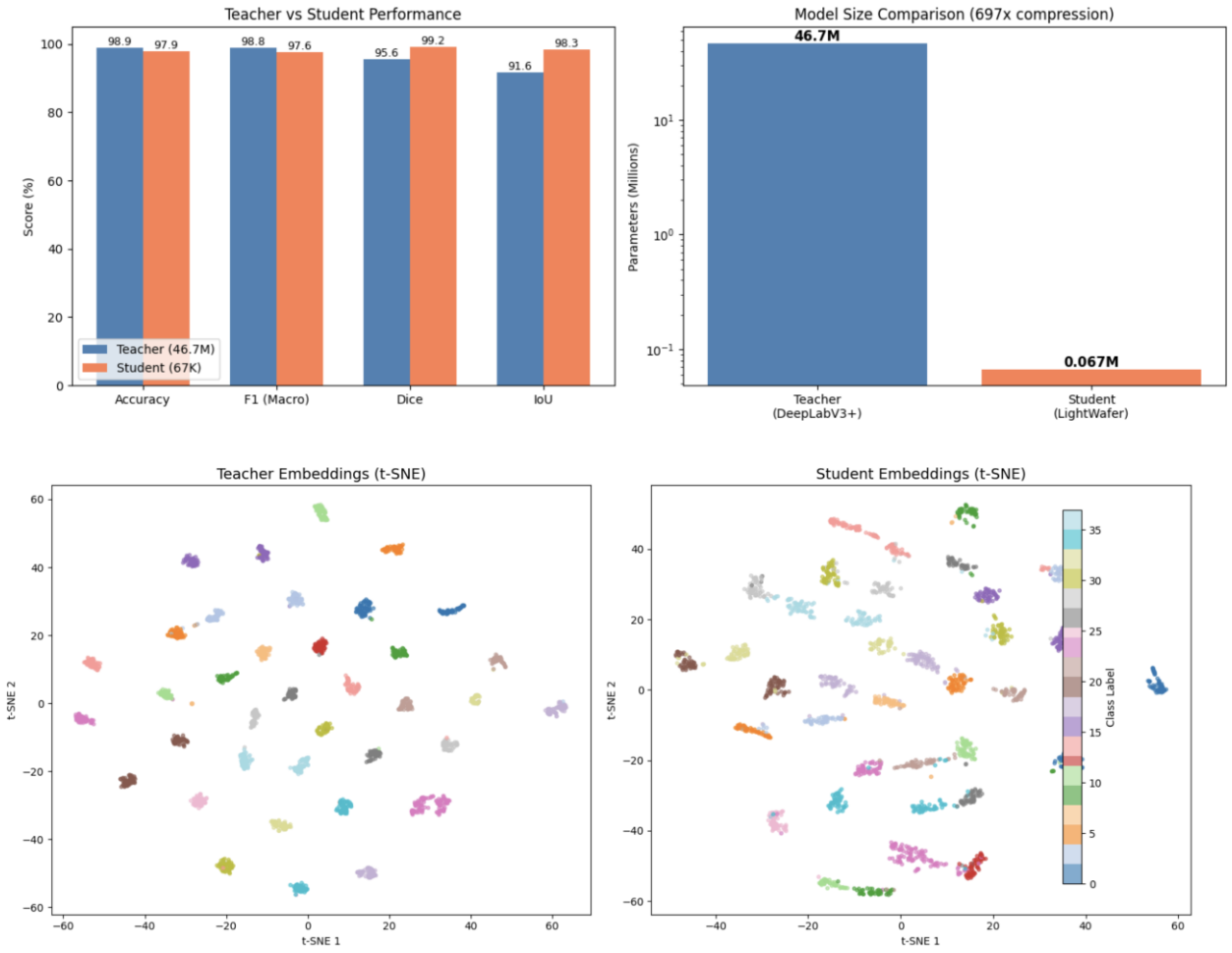
Note: The figures in the table above are representative values from experiment logs and may vary slightly depending on the actual training environment.

Table 2: Model Efficiency Comparison

Item	Teacher	Student	Ratio (Reduction Rate)
Parameters	Approx. 46,700,000	Approx. 67,000	Approx. 697x reduction
Memory (Weights)	Approx. 178 MB	Approx. 0.26 MB	Approx. 685x reduction

Comparing the test performance of Teacher (DeepLabV3+) and Student (LightWaferSegClassNet), the Teacher stably achieved very high baseline

performance, and it was confirmed that this knowledge was effectively distilled into the much lighter Student.



First, the Teacher model showed classification performance close to saturation with Accuracy 98.91%, Macro F1 98.85%, and AUC-ROC 99.97%. This means that Precision (98.85%) and Recall (98.86%) were maintained in a balanced manner across 38 defect classes in a multi-class environment, indicating successful convergence without bias between classes. Also in Segmentation, it recorded Dice 95.62% and IoU 91.60%, providing a powerful Teacher performance basis for stably restoring defect shape and boundary information. In other words, the Teacher sufficiently achieves the multi-task goal of "Accurate Classification + Reliable Segmentation," and it can be interpreted that the knowledge (soft target/feature/logit information) to be transferred to the Student is formed with high quality.

Subsequently, the Student model showed only a very small performance drop compared to the Teacher in classification performance, with Accuracy 97.90% and Macro F1 97.63% (drops of -1.01%p and -1.22%p, respectively). In particular, AUC-ROC was maintained at 99.97%, same as the Teacher, showing that the decision boundary itself was learned very robustly in the Student from a class discrimination perspective. This supports that the classification knowledge learned by the Teacher was **transferred to the Student with high fidelity (successful knowledge transfer)**, even though it is a lightweight model with

significantly reduced parameters/computations.

A more impressive part is that the Student significantly outperforms the Teacher in Segmentation performance. The Student recorded Pixel Accuracy 99.61%, Dice 99.16%, and IoU 98.33%, improving Dice by +3.54%p and IoU by +6.72%p compared to the Teacher. This means performance was not merely maintained due to lightweighting, but rather prediction became more precise and stable in defect area segmentation. This suggests that (1) the Student absorbed the Teacher's global/structural representations well during the distillation process, (2) the Student structure is designed to be more suitable for local patterns and boundary restoration of wafer defects, and (3) consequently, the Student may have generalized in a form that fits the data distribution better based on the Teacher's knowledge.

Additionally, in embedding extraction results, the Teacher uses (2016, 2048) and the Student uses (2016, 64) dimensional embeddings. That is, although the Student compressed representations into a much lower dimension (64), it almost maintained classification performance and improved segmentation performance. This implies that distillation succeeded in a direction that efficiently compresses high-dimensional features into low-dimensional representations while preserving meaningful expressive power, going beyond simple output mimicry.

In summary, the Teacher model played the role of a "well-trained powerful baseline model" with high performance in both classification and segmentation, and that knowledge was effectively transferred and compressed to the Student. As a result, the Student achieved almost maintained classification performance compared to the Teacher while achieving even greater performance improvement in Segmentation. Therefore, this experiment clearly demonstrates "Teacher's sufficient training success" and "Successful distillation of that knowledge into a much lighter Student leading to actual performance."

Also, the training of this model was estimated to take **a total of about 22 hours (1320 minutes)** combining Stage 1 (30 epochs) and Stage 2 (50 epochs) in a **Google Colab Pro environment (NVIDIA Tesla P100)**.

representations while preserving meaningful expressive power, going beyond simple output mimicry.

In summary, the Teacher model played the role of a "well-trained powerful baseline model" with high performance in both classification and segmentation, and that knowledge was effectively transferred and compressed to the Student. As a result, the Student achieved almost maintained classification performance compared to the Teacher while achieving even greater performance improvement in Segmentation. Therefore, this experiment clearly demonstrates "Teacher's sufficient training success" and "Successful distillation of that knowledge into a much lighter Student leading to actual performance."

Also, the training of this model was estimated to take **a total of about 22 hours (1320 minutes)** combining Stage 1 (30 epochs) and Stage 2 (50 epochs) in a **Google Colab Pro environment (NVIDIA Tesla P100)**.

Even including segmentation/classification multi-tasking and the distillation process, the entire training completes within a day, which is a very short training time. This further supports that the proposed lightweight Student model is a practical and reproducible pipeline, allowing for rapid experiment iteration and hyperparameter tuning.

These figures were obtained through a test/validation double configuration, and for reliability improvement, the final results were calculated by dividing data into train/validation/test. At this time, additional time reduction was sought by referring to the previous results.

Since the results are not significantly different from the above, they are attached separately for brevity.

(Link to Google Drive)

6. Discussion

The Student model minimized model size with an extremely small channel configuration (maximum 64 channels) and blocks based on depthwise separable convolution. In particular, by maintaining U-Net style skip connections, low-level features necessary for defect boundary restoration can be transferred to the decoder even in a small model, allowing field engineers to intuitively check defect locations and shapes through segmentation masks. Also, through knowledge distillation, by learning the Teacher's inter-class similarity and pixel-level probability distribution in addition to information provided by hard labels, over-simplification (underfitting), which is prone to occur in ultra-lightweight models, was mitigated and generalization performance was supplemented.

The proposed pipeline can be extended to optimize inference latency and memory usage for equipment/edge environments (e.g., ONNX conversion, TensorRT application, INT8 quantization) and to respond to data distribution changes through domain adaptation to process-specific data and continuous learning.

7. Conclusion

This study proposed **LightWaferSegClassNet**, an ultra-lightweight multi-task model that simultaneously performs wafer defect segmentation and

classification, and a Teacher-Student based 2-stage knowledge distillation learning pipeline. After training a DeepLabV3+ Teacher for MixedWM38, we designed the Student to mimic classification/segmentation soft targets together, achieving knowledge transfer to a model approximately 697 times smaller than the Teacher. The proposed approach demonstrates creative design to secure both boundary restoration and class discrimination capability even with an extreme parameter budget, as well as the feasibility of application within edge/equipment.

References

- Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., & Adam, H. (2018). Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation (DeepLabV3+). ECCV.
- Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. MICCAI.
- Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the Knowledge in a Neural Network. arXiv:1503.02531.
- Howard, A. G., et al. (2017). MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. arXiv:1704.04861.
- Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L.-C. (2018). MobileNetV2: Inverted Residuals and Linear Bottlenecks. CVPR.
- Nag, S., et al. (2022). WaferSegClassNet: A Light-weight Network for Simultaneous Classification and Segmentation of Semiconductor Wafer Defects. arXiv:2207.00960.
- Wang, J., Xu, C., Yang, Z., Zhang, J., & Li, X. (2020). Deformable Convolutional Networks for Efficient Mixed-type Wafer Defect Pattern Recognition. IEEE Transactions on Semiconductor Manufacturing.
- Iakubovskii, P. (2019). Segmentation Models PyTorch. GitHub repository (segmentation_models.pytorch).
- Junliangwangdhu. (n.d.). WaferMap Dataset: MixedWM38. GitHub repository.

A. Appendix: Reproduction Method (Colab)

Code files (.ipynb) and data are attached at this link.

AI Co-Scientist Challenge Korea Paper Checklist

The checklist follows the references. Please read the checklist guidelines carefully for information on how to answer these questions.

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 Section 1 Introduction clearly described the proposed model's lightweight level (697x compression) and multi-task performance goals.

Limitations

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

Answer: [Yes]

Justification: Section 6 Discussion mentioned the possibility of over-simplification in ultra-lightweight models, the necessity of knowledge distillation to supplement this, and considerations for field application.

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 an applied study focusing on experimental verification rather than theoretical proofs.

Experimental Result Reproducibility

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

Answer: [Yes]

Justification: Section 4 Experimental Setup and Appendix A detailed data preparation, model structure, training parameters, etc.

Open access to data and code

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

Answer: [Yes]

Justification: The MixedWM38 public dataset was used, and the source was specified in references.

Experimental Setting/Details

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

Answer: [Yes]

Justification: Section 4 Experimental Setup specified data splitting, preprocessing, optimizer settings, etc.

Experiment Statistical Significance

Question: Does the paper report error bars suitably and correctly defined?

Answer: [No]

Justification: Representative experimental result values were presented, but error bars from multi-repetition experiments were not included due to computing resource constraints.

Experiments Compute Resources

Question: For each experiment, does the paper provide sufficient information on the computer resources?

Answer: [Yes]

Justification: The model's parameter count and memory capacity were specifically stated to allow estimation of required resources.

Code Of Ethics

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

Answer: [Yes]

Justification: This study utilized a public dataset and does not cause ethical issues.

Broader Impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts?

Answer: [Yes]

Justification: Positive industrial effects through semiconductor yield improvement were discussed. Negative impacts are expected to be minimal.

Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models?

Answer: [N/A]

Justification: This study does not deal with high-risk models or sensitive data.

Licenses for existing assets

Question: Are the creators or original owners of assets properly credited?

Answer: [Yes]

Justification: Sources of the used dataset (MixedWM38) and library (SMP) were specified in references.

New Assets

Question: Are new assets introduced in the paper well documented?

Answer: [Yes]

Justification: The newly proposed model structure (LightWaferSegClassNet) was detailed in the main text.

Crowdsourcing and Research with Human Subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions?

Answer: [N/A]

Justification: This study does not include human subject research.

IRB Approvals

Question: Does the paper describe potential risks incurred by study participants and IRB approvals?

Answer: [N/A]

Preprint. Under review.