Silicon Sight

Automated Semiconductor Defect Detection using Vision Transformers

Team SILICON SIGHT

Abstract

Silicon Sight is an AI-powered quality control system designed to detect and classify microscopic defects in semiconductor wafers. Leveraging state-of-the-art Self-Supervised Vision Transformers (DinoV2), the system achieves **100% defect recall** and operates with real-time latency (~29ms) on edge hardware. This document outlines the problem context, technical approach, dataset architecture, and final model performance.

1 1. Problem Understanding

In the semiconductor manufacturing industry, yield rate is the primary driver of profitability. A single 300mm wafer can contain hundreds of chips worth tens of thousands of dollars. Defects such as micro-scratches, particle contamination, or topological faults (bridges/opens) can render chips non-functional.

Current Limitations:

- **Manual Inspection:** Relying on human operators using optical microscopes is slow, subjective, and prone to fatigue-induced errors.
- **Traditional AOI:** Automated Optical Inspection systems based on rule-based computer vision struggle with complex, non-linear defect patterns and require frequent recalibration for new chip designs.

Objective: To build a robust, Deep Learning-based classifier capable of distinguishing between "Clean" wafers and 6 specific defect types with zero false negatives (1.0 Recall).

2 2. Technical Approach

We adopted a **Data-Centric AI** approach combined with **Transfer Learning**.

2.1 Algorithm Selection: Vision Transformer (ViT)

Instead of standard Convolutional Neural Networks (CNNs), we utilized a **Vision Transformer (ViT)** architecture.

- **Global Context:** Unlike CNNs which look at local pixel neighborhoods, ViTs use Self-Attention mechanisms to understand the entire image simultaneously. This is critical for detecting structural defects like long cracks or distributed particle contamination.
- **Model Backbone:** We utilized **DinoV2-Base** (by Meta AI). DinoV2 is trained using Self-Supervised Learning (SSL) on 142 million images. This pre-training allows the model to understand geometry and texture without requiring massive labeled semiconductor datasets.

2.2 Deployment Strategy

To ensure industrial applicability, the trained PyTorch model was exported to **ONNX (Open Neural Network Exchange)** format. This optimization allows the model to run efficiently on edge devices (like factory cameras or ARM-based processors) without heavy server dependencies.

3 3. Dataset Plan

We constructed a hybrid dataset aggregating high-quality industrial samples to represent the full spectrum of wafer defects.

3.1 Data Source & Mapping

The dataset combines images from the **DeepPCB** (Logic/Topological defects) and **NEU Surface Defect** (Texture/Material defects) databases, re-mapped to semiconductor industry standard terms.

Target Class (Semiconductor)	Source Defect Type	Description
Clean	DeepPCB Template	Defect-free reference wafers.
Bridge	Short / Rolled-in Scale	Unintended electrical connections.
Opens	Open Circuit	Broken electrical connections.
Cracks	Crazing	Micro-fractures in the die.
CMP Residue	Spur / Patches	Chemical/Material residue.
Particles	Inclusion	Dust or foreign contamination.
Scratch	Scratches	Physical surface damage.

Table 1: Defect Class Ontology

3.2 Folder Structure (Submission Format)

The dataset is organized into a strictly stratified split to prevent data leakage. The single `.zip` file follows this structure:

```
dataset/
  train/                  (70% of data)
      clean/
      other/
          bridge/
          opens/
          cracks/
          cmp_residue/
          particles/
          scratch/
  validation/             (15% of data)
      [Same structure as train]
  test/                   (15% of data)
       [Same structure as train]
```

Total Images: 3,301 images (Balanced for evaluation).

4 4. Model Plan

4.1 Architecture Details

- **Model Name:** `facebook/dinov2-base`
- **Input Resolution:** 224 × 224 pixels (RGB).
- **Output Layer:** Linear Classification Head (7 Classes: 1 Clean + 6 Defect).
- **Parameters:** ~86 Million parameters.

4.2 Training Configuration

- **Loss Function:** Cross Entropy Loss (optimized for multi-class classification).
- **Optimizer:** AdamW (Adaptive Moment Estimation with Weight Decay).
- **Learning Rate:** $2e^{-5}$ (Low rate to preserve pre-trained feature extraction capabilities).
- **Batch Size:** 16.
- **Epochs:** 5 (Convergence achieved early due to strong pre-training).
- **Platform:** Training performed on **Apple M4 Neural Engine (MPS Acceleration)**.

5 5. Model Results

The model was evaluated on the held-out **Test Set** (15% of total data). The results demonstrate "Diamond Standard" performance suitable for production deployment.

5.1 Performance Metrics

Metric	Result
Accuracy	**100.00%**
Precision	1.0000
Recall	**1.0000** (Critical: No missed defects)
F1-Score	1.0000
MCC (Matthews Correlation)	1.0000

Table 2: Evaluation on Unseen Test Data

5.2 Operational Metrics

- **Inference Platform:** Apple M4 Chip (Metal Performance Shaders).
- **Average Latency: 29.12 ms** per wafer.
- **Throughput: 34 Frames Per Second (FPS)**.
- **Model Size (ONNX):** ~330 MB.

5.3 Robustness (Stress Test)

To validate reliability, we performed a synthetic stress test by injecting Gaussian noise into test images. The model maintained **>99% accuracy** even at noise levels of 0.1 standard deviations, proving it relies on structural features rather than pixel-perfect inputs.

6 6. GitHub Repository

The complete source code, including training scripts, the ONNX export utility, and the real-time simulation demo, is available at:

`https://github.com/Unwaver-afk/Deeptech_NPTEL`

The repository includes:

- **README.md:** Full documentation and usage guide.
- **src/**: Modular source code for training and inference.
- **demo/**: Real-time production line simulator.
- **onnx/**: The exported edge-ready model.