thesis outline

figures/tables

thesis content

Bachelor Thesis
Supervisor: Name of Advisor

This is the Title

This is the subtitle

Your Name

Contact: Your Name, youremail@tugraz.at

Institute of Visual Computing
Graz University of Technology, Austria

Graz, September 17, 2025

Abstract

Abstract plan:
- Problem & gap: Reliable dermoscopic classification is challenging due to severe class imbalance and the presence of out-of-distribution (OOD) inputs that can silently break models.
- Approach: Study a multi-stage pipeline (MLP1: skin vs non-skin -¿ MLP2: 8-class lesion type -¿ MLP3: benign vs malignant) and a direct final multi-class variant that predicts combined lesion x malignancy labels from the same frozen SAM features via MLP heads; OOD scoring is evaluated with MSP and ODIN.
- Data: In-distribution: ISIC 2018–2020 with unified metadata and train/val/test splits. External OOD: Places365 / DTD (textures) (far-OOD) to probe robustness.
- Evaluation: Headwise metrics (accuracy, weighted-F1), task-aware pipeline metrics (end-to-end accuracy/F1), confusion matrices; OOD metrics AUROC/AUPR for MSP/ODIN.
- Results (Exp1): Skin: ¿97% F1 (stable). Lesion: ∼51–64% accuracy, weighted-F1 ≈ 0.48–0.60 (macro-F1 much lower due to rare classes). Benign/Malignant: ∼77–79% accuracy, F1 ∼0.75–0.79. Task-aware pipeline: ∼30% overall accuracy/F1 due to error propagation. OOD: MSP/ODIN separate ISIC vs Places strongly (AUROC ≈ 0.98); ODIN-in-training yields no consistent gain.
- Takeaways: The direct final multi-class model matches the multi-stage pipeline while avoiding cascading errors; robustness benefits mainly come from simple MSP/ODIN scoring in far-OOD, not from training-time ODIN. Remaining gaps are driven by minority classes and near-OOD; background-removed inputs and harder OOD benchmarks are promising next steps.
- Keywords: Skin lesion classification; multi-task learning; out-of-distribution detection; SAM features; ISIC.

Abstract: One paragraph summary of
- Problem: difficulty of reliable skin lesion classification, risk of OOD.
- Approach: multi-stage classifier pipeline (skin detection -¿ lesion classification -¿ malignancy).
- Methods: SAM2 encoder features, MLP heads, dataset curation (ISIC, DermNet, DTD).
- Results: validation/test metrics, OOD performance, main findings.
- Contribution: task-aware design, background removal, OOD experiments.

Abstract: No figures/tables. Text-only summary.

Keywords: *Report, Technical report, template, IVC*

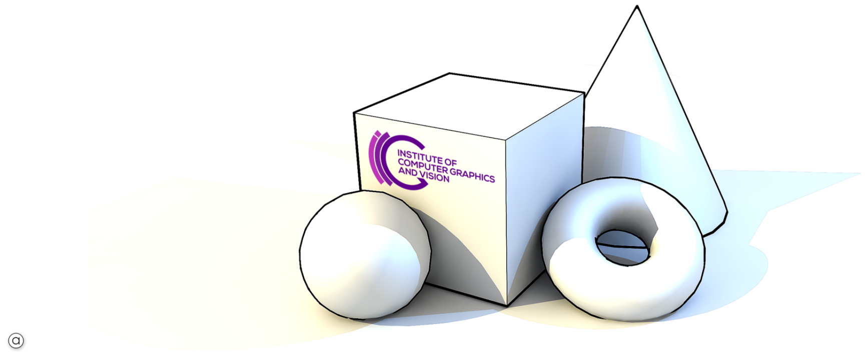

Figure 1: Overview. (a) Always provide a good caption in short, comprehensive sentences. The reader should understand your paper just by looking at the figures and the captions.

1 Introduction

Introduction: Motivation (medical relevance of skin cancer detection).
Problem statement: class imbalance, OOD, visual similarity.
Research questions:

- Can multi-stage classification improve reliability?
- How does task-aware evaluation compare to independent heads?
- What is the effect of background removal?
- How robust are models to OOD?

Contributions: bullet list.
Thesis structure: one sentence per chapter.

Introduction figures:
- Figure 1: Conceptual diagram of multi-stage pipeline (skin -¿ lesion -¿ bm).
- Figure 2: Example images showing skin vs. not-skin (illustrative).

Introduction — bullet outline (ready-to-expand):

Clinical context & motivation

- Skin cancer is common and impactful; early, reliable triage from dermoscopy images is critical.
- Automated support is attractive but must be reliable under real-world variability (devices, lighting, anatomy).

Problem framing

- Lesion diagnosis is inherently hierarchical: (1) plausibility check (skin vs. not-skin), (2) lesion type recognition (8 classes), (3) benign vs. malignant risk.
- Most ML systems treat this as a single flat task, ignoring hierarchy and error propagation.

Core challenges

- Severe class imbalance (few SCC/DF/VASC vs. many NV/MEL); majority classes dominate metrics.
- Near-OOD & label noise: non-skin or off-modality images mislabeled as skin; subtle domain shifts.
- Cascading errors in pipelines: early mistakes (skin/lesion) invalidate downstream decisions (BM).
- Calibration & OOD detection: high confidence on wrong inputs is unsafe.

Goal of this work

- Design & evaluate a dermoscopy classifier that respects hierarchical diagnosis and is robust to OOD and imbalance.

Research questions (RQs)

- RQ1: Does a multi-stage pipeline (skin -¿ lesion -¿ BM) improve reliability vs. independent heads?
- RQ2: Can a direct final multi-class model match/surpass the pipeline while avoiding error cascades?
- RQ3: How do MSP & ODIN perform for OOD detection here, and does ODIN-in-training help?
- RQ4: What is the effect of dataset strategies (background removal, oversampling) on minorities & pipeline?
- RQ5: Which evaluation protocol (headwise vs. task-aware/pipeline) better reflects clinical use?

Approach (high-level)

- Use frozen SAM features as a backbone; train light MLP heads on top.
- Two modelings: (i) multi-stage (MLP1 skin, MLP2 lesion, MLP3 BM) and (ii) final multi-class (combined lesion x BM).
- Reliability via OOD scoring (MSP, ODIN) and task-aware evaluation mirroring clinical flow.

Data & splits (summary)

- Unified ISIC 2018–2020 dermoscopy with single metadata (train/val/test).
- External sets (Places/DTD) for far-OOD stress tests.
- Features cached from the frozen backbone for fast, reproducible runs.

Evaluation plan

- Headwise: accuracy & weighted-F1 for skin/lesion/BM, plus confusion matrices.
- Task-aware pipeline: end-to-end accuracy/F1 on composed labels; pipeline confusion matrix.
- OOD: AUROC/AUPR for MSP/ODIN; qualitative failure modes.

Contributions

- Unified, reproducible pipeline for hierarchical dermoscopy on SAM features.
- Task-aware evaluation exposing error propagation vs. headwise scores.
- Direct final multi-class simplifying inference while matching pipeline performance.
- Comprehensive OOD study (MSP/ODIN), incl. ODIN-in-training vs. test-time-only.
- Ablations: oversampling, mixup, background-removed inputs; thorough confusion diagnostics.

Scope & limitations

- Dermoscopy still images; frozen features (no end-to-end finetuning).

2 Related Work

Related Work: medical imaging and skin lesion classification (CNN, transformers).
SAM2 and segmentation-based preprocessing.
Multi-task and hierarchical classification in ML.
OOD detection approaches (MSP, ODIN).
Data augmentation and balancing in medical ML.
Positioning: where this thesis fits.

Related Work tables/figures:
- Table 1: Comparison of existing approaches (method, dataset, accuracy, OOD).
- Optional figure: Positioning schematic (Venn: Segmentation, Multi-task, OOD).

3 Method

Datasets: ISIC 2018/2019/2020, DermNet, SD-198, DTD, ImageNet subset. Unified metadata (mapping diagnosis -¿ unified_diagnosis). Preprocessing: cleaning, resizing, background removal. Train/val/test splits (70/15/15).
Feature extraction: SAM2 encoder as frozen feature extractor; feature dimensions; caching to .pkl.
Model architectures: parallel multi-head model (skin/lesion/bm); task-aware pipeline evaluation logic; final multi-class ablation (11-class).
Training setup: losses (LMF, CE), task weights; oversampling vs natural; mixup and augmentations; optimizer, LR scheduler, epochs.
Evaluation setup: headwise metrics (accuracy, weighted F1); task-aware metrics (pipeline accuracy/F1); confusion matrices; OOD detection (MSP, ODIN).

Method figures/tables:
- Table 2: Dataset overview (images, classes, split sizes).
- Figure 3: Sample images per dataset (grid with labels).
- Figure 4: Class distribution histograms (lesion, benign/malignant).
- Figure 5: Image size/aspect ratio distribution.
- Figure 6: SAM2 encoder -¿ feature vector diagram.
- Figure 7: Background removal example (before/after).
- Figure 8: Parallel multi-head model block diagram.
- Figure 9: Task-aware pipeline flowchart (skin -¿ lesion -¿ bm).
- Figure 10: Final multi-class model (11-class head).
- Table 3: Hyperparameters (optimizer, LR, batch, epochs, loss weights).
- Figure 11 (optional): Training curriculum schedule.
- Figure 12: Headwise vs. task-aware evaluation scheme.
- Table 4: Metrics overview (accuracy, weighted F1, AUROC, AUPR).

4 Experiments

Baseline: parallel model with weighted F1.
Ablations: background removed vs original; final multi-class vs task-aware pipeline; oversampling vs no oversampling; mixup vs no mixup.
OOD experiments: MSP and ODIN; performance on DTD / ImageNet.
Task-aware evaluation: compare with headwise metrics; pipeline confusion matrices.

Experiments figures/tables:
- Table 5: Baseline results (skin/lesion/bm acc & F1). - Table 6: Ablation results (bg removal, multi-class, oversampling, mixup).
- Figure 13: Bar chart of lesion F1 across ablations.
- Table 7: OOD AUROC/AUPR for MSP and ODIN (skin vs. not-skin OOD).
- Figure 14: ROC curves MSP vs. ODIN.
- Table 8: Headwise vs. task-aware performance.
- Figure 15: Pipeline-level confusion matrix.

5 Results

Present per-experiment metrics in tables. Confusion matrices (lesion head and pipeline). OOD AUROC/AUPR comparisons.
Highlight: best configuration; trade-offs between simplicity (final multi-class) and interpretability (task-aware pipeline); oversampling impact (negative). Link results to research questions.

Results figures/tables:
- Figure 16: Lesion confusion matrix (parallel multi-head).
- Figure 17: Lesion confusion matrix (final multi-class).
- Figure 18: Pipeline confusion matrix (task-aware).
- Figure 19: Training/validation accuracy and loss curves.
- Figure 20: OOD AUROC curve.
- Table 9: Summary of all models (baseline, ablations, OOD).

6 Discussion

Interpretation: why certain methods helped/hurt.
Limitations: dataset biases; SAM2 frozen features; OOD labeling reliability.
Practical implications for medical usage.
Future work: trainable segmentation; larger datasets; clinical deployment with uncertainty estimation.

Discussion figures/tables:
- Figure 21: Error analysis examples (misclassified images with predicted vs. GT).
- Table 10: Common confusion pairs (e.g., NV vs. MEL).

7 Conclusion

Recap problem, approach, contributions. Main findings (weighted F1 better, oversampling harmful, task-aware adds interpretability). Final statement on feasibility of multi-stage lesion classification.

Conclusion: No figures/tables (optionally restate Figure 1 pipeline).

A Appendix (optional)

Additional confusion matrices. Detailed dataset stats. Training logs / configs.

Appendix figures/tables:
- Table A1: Full class counts per dataset and split.
- Figure A1: Additional confusion matrices (per dataset, per head).
- Figure A2: Extended ROC curves (all OOD datasets).
- Table A2: Full hyperparameter search log.

References