Analysis of Hand Segmentation in the Wild

Computer Vision Course Project

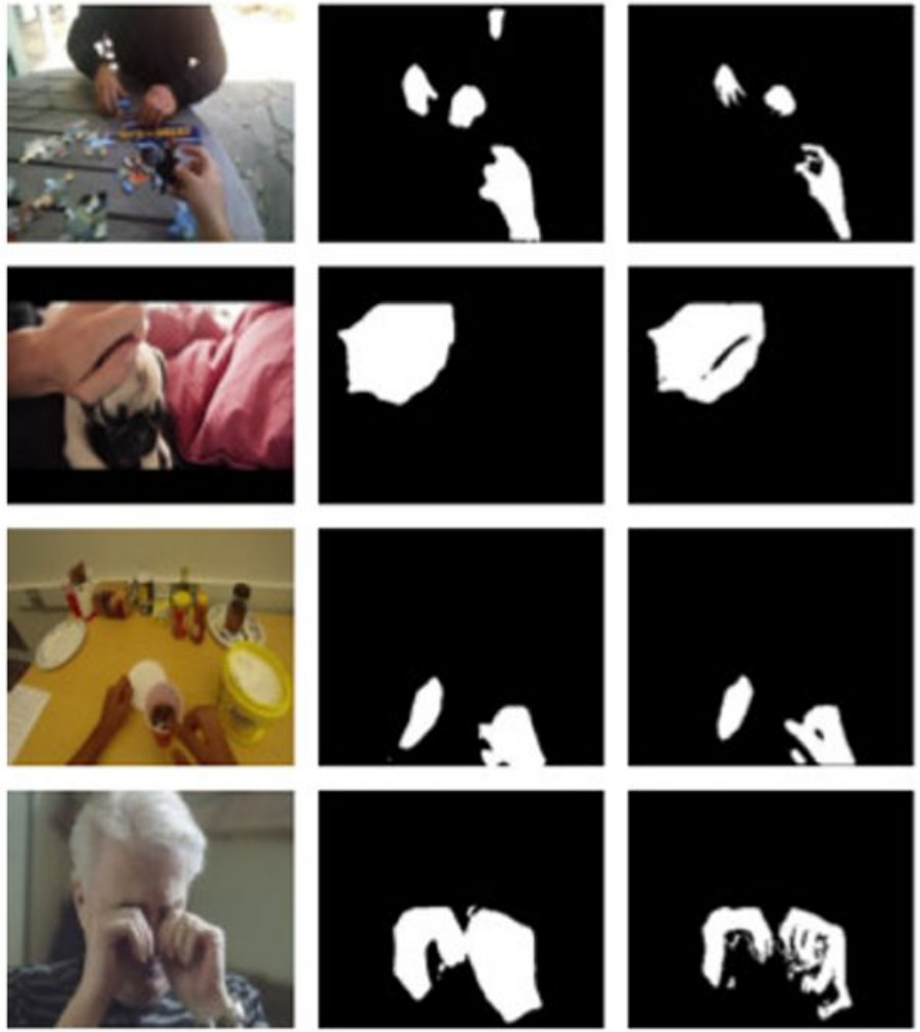
Team: Deepak Bhatter (B22EE022), Aditya Kumar (B22ME004), Mayank Agrawal (B22CS084)

Supervisor: Dr. Avinash Sharma

Introduction & Motivation

What is Hand Segmentation?

Egocentric (first-person) hand segmentation is a core problem for understanding human-computer interaction, AR, and robotics.



Why is it Difficult?

- Rapid viewpoint and illumination changes.
- Frequent occlusions from objects and motion blur.
- Ambiguity between hands and other skin-toned objects (like faces).
- No fixed cues, unlike in third-person video.

Our Project Pipeline

1. Baselines

Established initial performance with DeepLabV3 and a hybrid YOLOv8 + GrabCut model.

2. Pre-training

Replicated the paper's pre-training by training RefineNet-Res101 on the Pascal Parts dataset.

3. Fine-Tuning

Fine-tuned the model on four distinct hand datasets (EgoHands, EYTH, GTEA, HOF).

4. Analysis

Performed cross-dataset evaluation to test generalization and analyze results.

Baseline 1: DeepLabV3-ResNet50

Initial Experiment

Our first step was to test a standard, "naive" segmentation model (DeepLabV3 + ResNet) to understand the core challenges.

Key Failure

The model performed reasonably on simple frames but consistently misclassified faces as hands. This highlighted the critical problem of distinguishing between skin-toned regions based on context alone.

Baseline 2: YOLOv8 + GrabCut

A Hybrid Approach

We explored a hybrid method combining deep learning detection with classical segmentation:

- **Detect:** Use YOLOv8-Nano to draw a bounding box.
- **Seed:** Automatically create a small "foreground" seed in the box's center.
- **Segment:** Apply the GrabCut algorithm.

Key Failure

This method works well for high-contrast images but fails when the background has similar colors to the hand, as GrabCut is color-based.

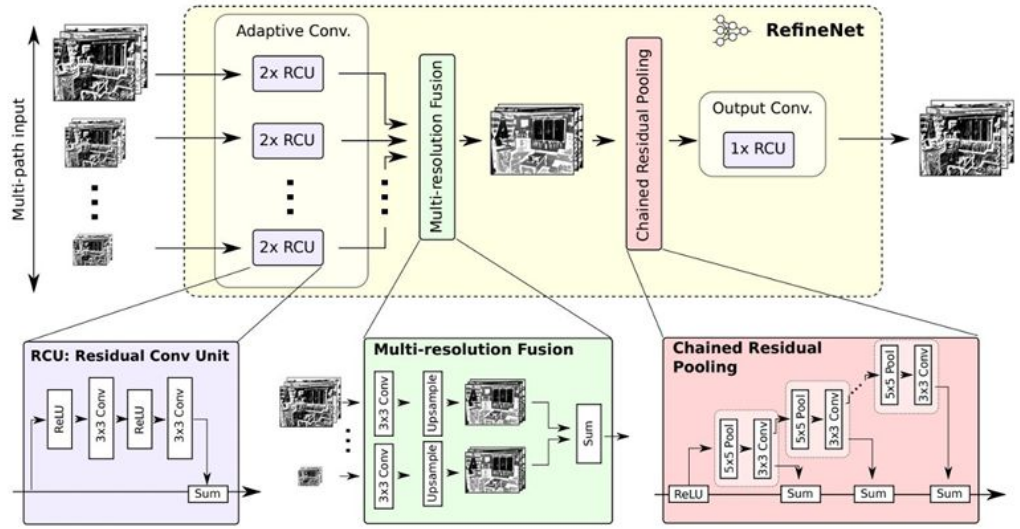
Core Model: RefineNet Architecture

Why RefineNet?

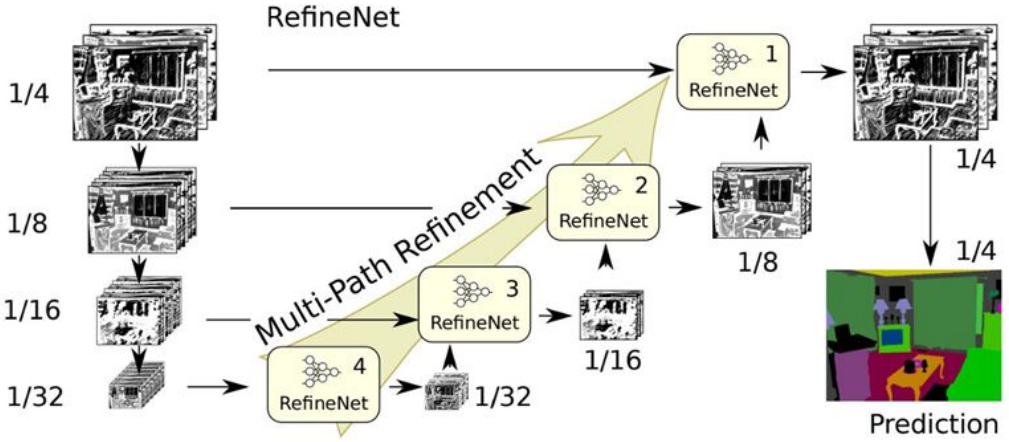
This is the main architecture from the paper. It's designed to fuse features from multiple resolutions, which is ideal for segmentation.

- **Coarse Features (Deep):** Provide context (e.g., "this is a person's arm").
- **Fine Features (Shallow):** Provide detail (e.g., "this is the edge of a finger").

RefineNet combines these pathways to get both high-level context and high-resolution detail simultaneously.



(a) Complete Refinement Network



(b) RefineNet Architecture

RefineNet Building Blocks

RCU

Residual Conv Unit: Refines feature maps using residual (identity) connections, similar to ResNet.

MRF

Multi-Resolution Fusion: Fuses feature maps from different scales by upsampling and summing them.

CRP

Chained Residual Pooling: Aggregates background context from large regions using a chain of pooling blocks.

Implementation: Step 1 (Pre-training)

The "From Scratch" Challenge

The paper's method relies on pre-training on the **Pascal Person-Part** dataset. However, public weights were not available.

Our Solution

We had to replicate this step manually:

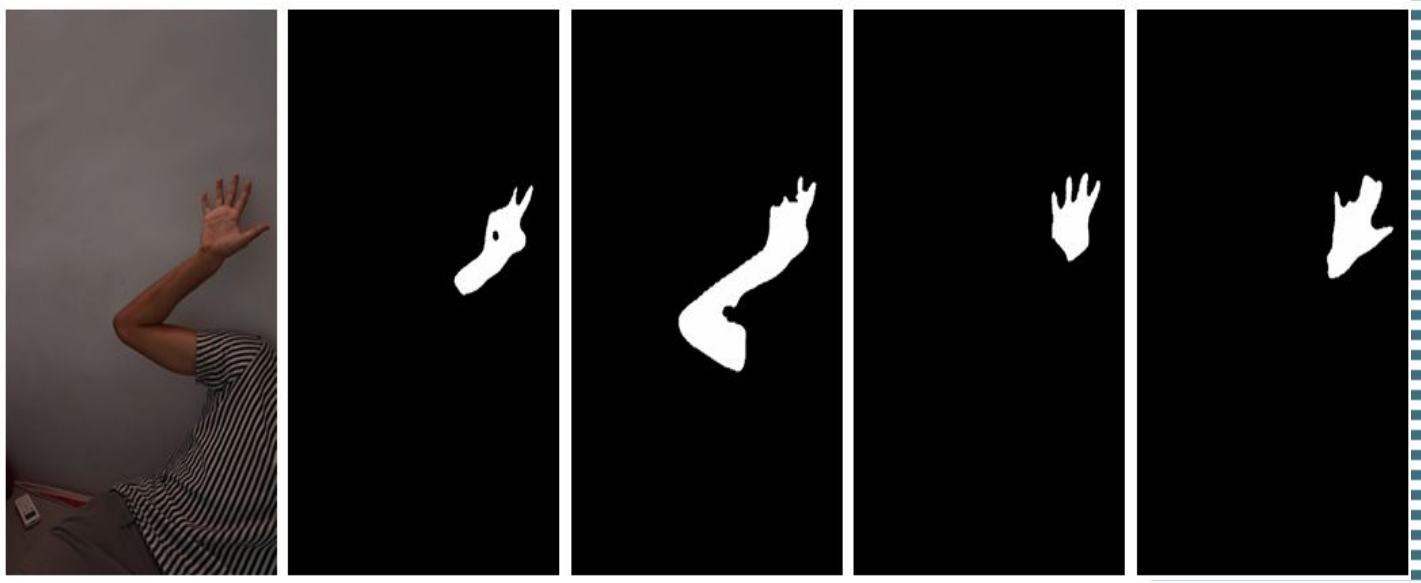
- The dataset annotations were in .mat (MATLAB) files.
- We wrote preprocessing scripts to convert these .mat files into standard .png segmentation masks.
- We then trained the RefineNet-Res101 model from scratch on this processed dataset.

Implementation: Step 2 (Fine-Tuning)

Specializing the Model

After pre-training on general human parts, we fine-tuned the model for the specific task of hand segmentation using four datasets:

- **EgoHands:** Egocentric video of two people interacting (puzzles, chess).
- **EYTH:** Diverse, "in-the-wild" videos from YouTube.
- **GTEA:** Constrained cooking activity dataset.
- **HOF:** Small dataset of faces occluded by hands.



Results: Quantitative (mIoU, Batch Size 16)

Trained on ↓ | Tested on →	EgoHands	EYTH	GTEA	HOF
EgoHands	**0.860**	0.219	**0.734**	0.415
EYTH	0.398	**0.603**	0.297	0.232
GTEA	0.414	0.163	**0.791**	0.118
HOF	0.448	0.262	0.539	**0.696**

Key Findings: 1. Models performed best on their own dataset (see diagonal). 2. **EgoHands** (diverse interactions) was the best all-rounder, generalizing well to GTEA. 3. **GTEA** (constrained) generalized very poorly to all other datasets.

Results: Qualitative Comparison

Visual Analysis

The quantitative results are reflected in the visual output. We can see the dataset bias clearly:

- **EgoHands Model:** Produces clean, precise masks, focusing only on the hand.
- **GTEA Model:** This model was trained on cooking data where arms are always visible. As a result, it incorrectly segments the entire arm as part of the "hand".
- **HOF Model:** Shows good performance on challenging skin-on-skin occlusion.

Conclusion & Q&A

Challenges & Conclusions

Challenges Faced:

- **Unavailable Weights:** Required manual pre-training on Pascal Parts, which was time-consuming.
- **High Computation Cost:** RefineNet is resource-intensive. Training was slow on T4, faster on A100/H200 GPUs.

Final Conclusions:

- Successfully replicated the core findings of the paper.
- Confirmed RefineNet is a highly effective architecture for this task.
- Demonstrated that **dataset diversity** (like in EgoHands) is more important for generalization than dataset size alone (like in GTEA).

Thank You!

Questions?