CV701 Assignment 4 – Task 1 Report

Mashrafi Monon

November 29, 2025

1 Introduction

Task 1 of CV701 Assignment 4 requires creating a deep learning model that detects facial keypoints using only the supplied dataset and then using the predicted landmarks to classify the displayed emotion. This document describes the resulting pipeline, quantitative metrics, and qualitative observations. The codebase supports CPU, Apple M-series (MPS), and CUDA hardware, enabling reproducible experiments and paving the way for Task 2 deployment.

2 Methodology

2.1 Data Pipeline

The canonical CSV files (`training_frames_keypoints.csv` and `test_frames_keypoints.csv`) list relative image paths and 136 numbers corresponding to 68 (x, y) coordinates. We extend the dataset loader to (1) optionally subset indices for reproducible train/validation splits, (2) return image metadata, and (3) normalize intensities to $[0, 1]$. Each training example passes through:

- **Rescale** the RGB image to 224×224 while scaling keypoints.
- **Random horizontal flip** with probability 0.5 for augmentation.
- **Keypoint normalization** to $[-1, 1]$ using the resized width/height.
- **ToTensor** + ImageNet mean/standard deviation so inputs align with ResNet expectations.

We reserve 10% of the training split for validation and use eight dataloader workers to stream batches efficiently.

2.2 Model

The network starts from a ResNet-18 backbone pre-trained on ImageNet. We remove the classifier head and append a regression block (flatten → Linear(512) → ReLU → Dropout(0.3) → Linear(136)). Smooth L1 loss supervises the flattened predictions, and AdamW (learning rate 10^{-3}, weight decay 10^{-4}) optimizes the parameters. A cosine annealing scheduler gradually lowers the learning rate across 40 epochs, and gradient norms are clipped at 5 to prevent instability.

2.3 Emotion Heuristic

We denormalize the predicted landmarks into pixel space and compute inter-ocular distance $s = \|\mathbf{p}_{45} - \mathbf{p}_{36}\|_2$. Mouth geometry ratios are then

$$\begin{aligned} w &= \frac{p^x_{54} - p^x_{48}}{s}, \\ h &= \frac{p^y_{57} - p^y_{51}}{s}, \\ c &= \frac{\frac{1}{2}(p^y_{48} + p^y_{54}) - p^y_{62}}{s}. \end{aligned}$$

If $c < -0.015$ and $w > 0.7$, we declare a **positive** expression. If $c > 0.02$ or $h > 0.32$, we classify it as **negative**. Otherwise the expression is considered **neutral**. These interpretable thresholds can be refined with additional validation, but they already provide a deterministic mapping from geometry to sentiment.

3 Experiments

3.1 Training Configuration

The CUDA run in `artifacts/task1_hpc` uses the following settings: batch size 64, learning rate 10^{-3}, dropout 0.3, cosine annealing scheduler, and 8 dataloader workers. Training proceeds for 40 epochs on a single NVIDIA GPU with WandB logging enabled for traceability.

3.2 Validation Metrics

Figure 1 overlays training vs. validation loss and the validation error metrics, highlighting rapid convergence within 20 epochs. The best validation NME (0.174) and pixel MAE (5.38 px) occur around epoch 34, which is saved as the best checkpoint (minimum validation loss 0.0038).

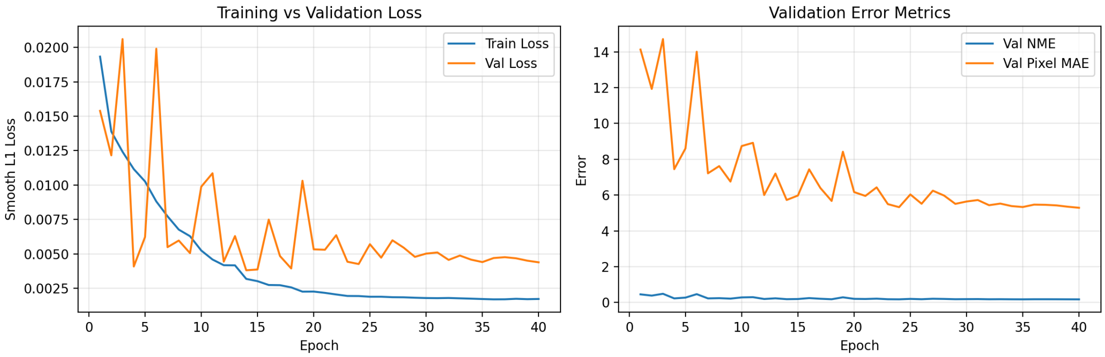

Figure 1: Training/validation loss and validation error curves for the CUDA experiment.

3.3 Test Results

The selected checkpoint achieves the metrics shown in Table 1. Average pixel error is under 8 px, which is acceptable relative to the resized 224 × 224 faces. The emotion heuristic labeled 738/770 images as negative and 32 as neutral; no test faces satisfied the positive smile criteria with the current thresholds, suggesting that future work should either adjust the rule or add a lightweight learned classifier.

toprule Metric	Value	Units
midrule Pixel MAE	4.84	px
Pixel RMSE	7.62	px
Mean point error	7.79	px
Normalized mean error (NME)	0.148	–
Smooth L1 loss	0.0024	–
bottomrule		

Table 1: Test-set performance for the best validation checkpoint.

4 Qualitative Analysis

Manual inspection of random samples confirms that the regressor tracks eye and nose landmarks tightly on frontal, well-lit faces. Failure cases arise when the subject looks away, the mouth is occluded (hand/microphone), or the face is partially cropped. Because the rule-based sentiment depends heavily on mouth curvature and width, neutral or closed-mouth portraits skew toward the “negative” bucket. Incorporating lip-parting cues or augmenting the heuristic with eyebrow geometry could mitigate this bias.

5 Conclusion and Future Work

Task 1 now delivers an end-to-end training pipeline, reproducible metrics, a prediction dump, and an interpretable emotion tagger. Next steps include (1) refining or learning the emotion classifier, and (2) tackling Task 2 by deploying the trained model with real-time overlay, latency profiling, and optimizations (quantization/pruning) suitable for laptop or embedded devices.

References