Multi-Class Human Body Part Detection

Jeffery Mei – The University of Chicago

1. Problem Statement

Object detection represents a fundamental computer vision task that extends beyond simple classification by requiring models to both identify and precisely localize objects through bounding box predictions. This project focuses on developing a deep learning based detection model capable of identifying and localizing multiple human body parts including faces, hands, arms, legs, heads, eyes, ears, noses, mouths, feet, and hair. Unlike classification that assigns single labels to images, detection demands spatial awareness and the ability to identify multiple instances of different body part classes simultaneously in a single image.

This project utilizes a curated dataset from Open Images V4 containing 6,000 images with comprehensive bounding box annotations across twelve distinct human body part categories. Body part detection enables sophisticated applications requiring detailed anatomical understanding beyond simple person presence. This multi-class approach provides richer training signals and allows the model to learn hierarchical relationships between body parts, understanding compositional structure that mirrors human visual perception. Moreover, body part detection addresses practical scenarios where only portions of individuals are visible due to occlusion or frame boundaries, situations where whole person detection would fail.

The human body part detection model addresses critical needs across numerous domains. In human computer interaction, detecting hands and faces enables touchless interfaces for healthcare settings and automotive driver monitoring. Assistive technology leverages body part detection to help visually impaired individuals navigate spaces with greater specificity. Online fashion retailers utilize such systems for virtual try on applications requiring precise body region localization. Sports analytics benefits from detailed tracking to analyze biomechanics and optimize training.

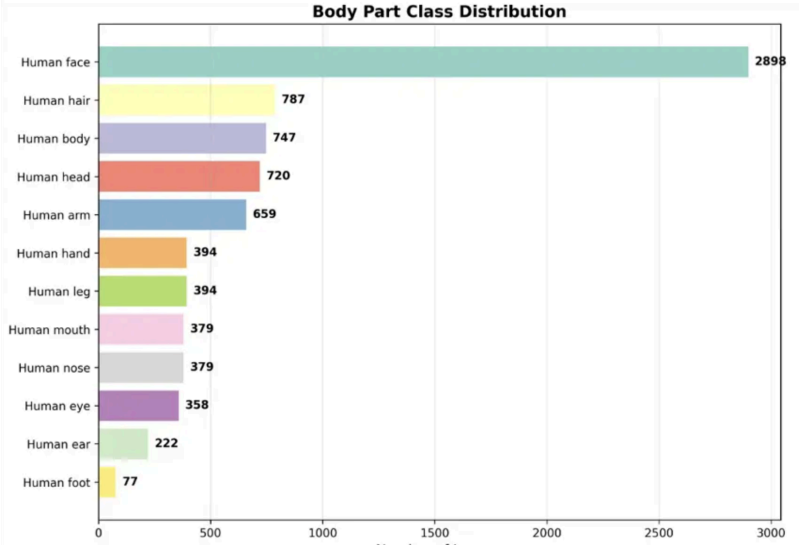
The detection model must achieve two objectives. First, it must accomplish accurate multi class detection, correctly classifying each body part while drawing precise bounding boxes. This requires learning discriminative features distinguishing visually similar classes like arms versus legs while handling significant scale variations. Second, the model must demonstrate robust performance under partial occlusion, extreme poses, and crowded multi person scenes, maintaining accuracy across diverse demographics without bias.

2. Data Exploration (EDA)

A. Examples of images


B. Class distribution (imbalances must be discussed)

At first, we randomly downloaded 4000 images from the dataset. It contains 24,711 bounding box annotations, yielding an average of 6.18 annotations per image. However, analyzing the class distribution showed a significant imbalance.

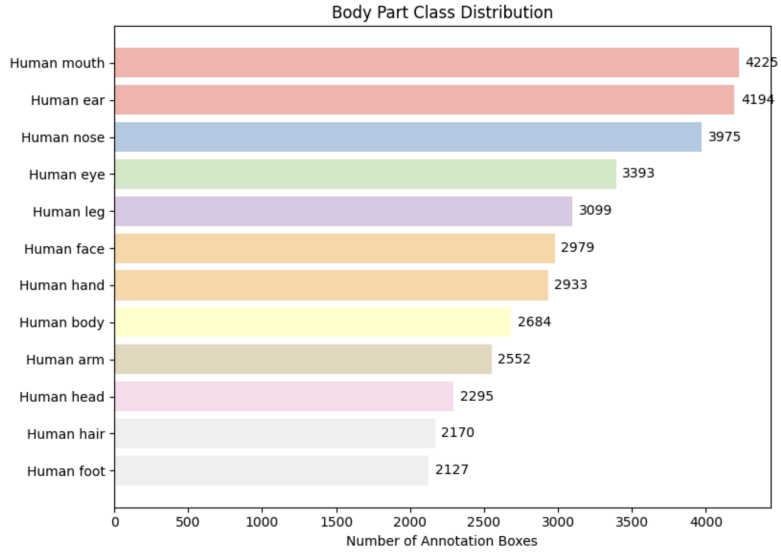


As you can see from the graph, human face box annotations are 40 times more than the human foot annotations. This severe class imbalance will pose several critical challenges for model development. First, standard training procedures risk creating a biased detector that excels at identifying common classes like faces while performing poorly on underrepresented classes like feet and ears. The model may learn to predict frequent classes with high confidence while rarely or never predicting minority classes, maximizing overall accuracy metrics at the expense of per class performance.

Second, the limited training examples for minority classes reduce the model's ability to learn robust feature representations, increasing susceptibility to overfitting on the few available samples. Third, evaluation metrics such as Mean Average Precision can mask poor performance on minority classes when averaged across all categories.

We implemented a box balanced sampling strategy. Our approach recognizes that different body part classes exhibit varying instance densities within images. By analyzing boxes per image ratios from preliminary downloads, we found that human faces averaged 2.82 boxes per image while human mouths averaged only 1.54 boxes per image, revealing that downloading equal numbers of images per class would perpetuate annotation level imbalance. Our methodology establishes a target of approximately 2,000 bounding boxes per class. To achieve 2,000 human face boxes, we downloaded approximately 710 images, whereas achieving 2,000 human foot boxes required approximately 909 images.

Following our box-balanced sampling strategy, the rebalanced dataset demonstrates significant improvement in class distribution. The dataset comprises 9,199 unique images containing 36,626 bounding box annotations across twelve body part categories. The deliberate rebalancing effort successfully reduced the imbalance ratio observed in the original dataset to a substantially more equitable distribution, with bounding box counts now ranging from 2,170 boxes for human hair to 4,194 boxes for human mouth. The top classes such as mouth, ear, and nose are relatively smaller parts of the body which I think the model will be struggling with.



C. Image Sizes, Quality Issues

The dataset exhibits relatively uniform image dimensions with strong concentration around standard resolutions. Image widths range from 354 to 1,024 pixels (mean: 938px, median: 1,024px) while heights range from 328 to 1,024 pixels (mean: 825px, median: 768px). The width distribution shows a prominent peak at 1,024 pixels containing approximately 4,000 images, while height distribution displays three modes at 640, 768, and 1,024 pixels. File sizes range from 0.02 MB to 1.38 MB (mean: 0.27 MB, median: 0.22 MB), facilitating efficient data loading while maintaining sufficient quality for feature extraction. Compact file sizes averaging 0.27 MB enable efficient data loading and support larger batch sizes that stabilize gradient estimates during optimization.

The dataset demonstrates excellent overall quality with minimal problematic images. Critically, zero images fall below the 224-pixel minimum threshold for modern CNNs, and zero exceed the 2,000-pixel threshold requiring aggressive downsampling, meaning all images are natively suitable for standard detection architectures without extensive preprocessing.

D. Annotation Format

The dataset annotations are stored in COCO (Common Objects in Context) detection format. The annotations are contained in a single JSON file (labels.json) with a hierarchical structure comprising four primary components: dataset metadata (info), category definitions (categories), image metadata (images), and bounding box annotations (annotations). This standardized format ensures compatibility with popular detection frameworks.

The categories array defines the 12 body part classes, each assigned a unique integer identifier (category_id). The images array contains 6,008 entries, each specifying an image's filename, dimensions (width and height), and unique identifier. The annotations array contains 27,832 entries, each representing a single bounding box with fields including a unique annotation ID, the associated image ID, the category ID, and critically, the bounding box coordinates. Each of the 27,832 annotations represents a single bounding box encoded in COCO's standard format as [x, y, width, height], where (x, y) specifies the top-left corner in absolute pixel coordinates and width/height define the box dimensions. The COCO detection format's widespread adoption ensures straightforward integration with existing codebases.

3. Baseline Neural Network

Model Architecture

We implement Faster R-CNN with a ResNet-50 Feature Pyramid Network (FPN) backbone as our baseline detection model, leveraging torchvision's pre-trained implementation initialized with COCO dataset weights. The model consists of 3 primary components. The backbone network employs ResNet-50 with FPN, extracting multi-scale feature representations across 5 pyramid levels with resolutions from 1/4 to 1/32 of the input dimensions. The Region Proposal Network (RPN) generates approximately 2,000 candidate object locations per image through anchor boxes at multiple scales and aspect ratios. The ROI head performs final classification and bounding box regression, with the pre-trained COCO head replaced by a custom box predictor configured for 13 output classes (12 body part categories plus background) while maintaining the 1,024-dimensional feature representation.

The architecture comprises 41,355,536 parameters (41,133,136 trainable), occupying approximately 157.8 MB in fp32 format. This parameter count strikes a balance between detection accuracy and computational efficiency suitable for our dataset scale of 6,000 images with 28,000 annotations.

Training Configuration

We adopt minimal tuning for the baseline, using standard hyperparameters proven effective for Faster R-CNN on COCO-style datasets. SGD with momentum serves as the optimizer (learning rate: 0.005, momentum: 0.9, weight decay: 0.0005), following torchvision best practices where SGD consistently outperforms adaptive optimizers for two-stage detectors. The learning rate schedule implements step decay, reducing by 10× every 3 epochs, allowing rapid initial adaptation (epochs 1-3), then fine-tuning (epochs 4-6 at 0.0005), and precise refinement (epochs 7-10 at 0.00005).

The dataset split allocates 80% for training and 20% for validation with batch size 4, yielding 2,169 training batches and 543 validation batches per epoch. Training spans 10 epochs with minimal augmentation (tensor normalization only), balancing convergence time against overfitting risk. We monitor both training and validation loss, saving the best checkpoint based on validation performance.

3.1 Baseline Training Results and Evaluation

The baseline Faster R-CNN model trained for 10 epochs over approximately 77 minutes on a single CUDA-enabled GPU. Training exhibited expected convergence patterns with rapid initial improvement followed by gradual refinement as the learning rate decayed.

Architecture Summary

```
Model                : Faster R-CNN
Backbone             : ResNet-50 FPN
Num classes          : 13
Input features       : 1024
Pretrained           : True
Total parameters     : 41.36M (41,355,536)
Trainable parameters : 41.13M (41,133,136)  (99.46%)
Frozen parameters    : 222.40K (222,400) (0.54%)
Modifications        : Replaced box predictor head for 12 body part classes
```

Loss Trajectory

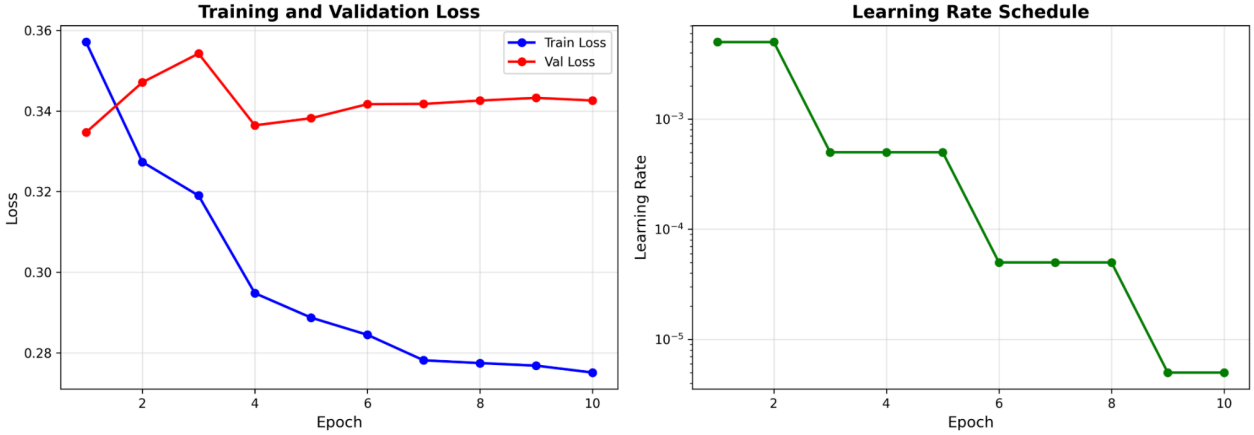
Training loss decreased from 0.3563 in epoch 1 to 0.2735 by epoch 10, representing a 23.2% reduction that indicates successful model learning. The steepest descent occurred during the first three epochs (10.6% reduction), driven by the initial learning rate of 0.005 that enabled rapid adaptation of the randomly initialized detection head to the body part detection task. After the first learning rate reduction at epoch 4 (LR: 0.0005), training loss continued decreasing at a more gradual pace (10.6% reduction), indicating effective fine-tuning of learned representations. The second learning rate reduction at epoch 7 (LR: 0.00005) produced minimal additional improvement (4.0% reduction), suggesting the model approached convergence so that further training could only yield diminishing returns.

Validation loss exhibited markedly different behavior, achieving its minimum of 0.3347 at epoch 2 and subsequently plateauing around 0.34 for the remaining 8 epochs. This early validation loss minimum followed by slight increases represents a classic overfitting signature where the model begins memorizing training set patterns that fail to generalize to unseen validation data. The widening gap between training loss (0.2735) and validation loss (0.3431) at epoch 10 confirms this overfitting tendency, though the magnitude remains modest compared to severe overfitting scenarios where validation loss would increase substantially. The relatively small generalization gap suggests that while overfitting occurred, the combination of pre-trained

initialization, weight decay regularization (0.0005), and limited training epochs is designed to prevent catastrophic overfitting.

Learning Rate Schedule Impact

The step decay learning rate schedule produced clear transitions visible in both the training curves and per-batch loss patterns. The aggressive 10× reductions at epochs 4 and 7 successfully prevented overshooting while allowing thorough exploration of the loss landscape. However, the validation loss plateau at epoch 2 (0.3347) followed by persistent values around 0.34 indicates premature convergence to a suboptimal solution. The initial learning rate of 0.005 may have been slightly too aggressive, causing the model to converge to a local minimum that generalized suboptimally. Future iterations should explore more conservative initial learning rates or cosine annealing schedules for smoother convergence.



Initial Evaluation Metrics

Post-training evaluation on the validation set (2,169 images) with a confidence threshold of 0.5 revealed both strengths and weaknesses in the baseline model's detection capabilities.

Detection Statistics

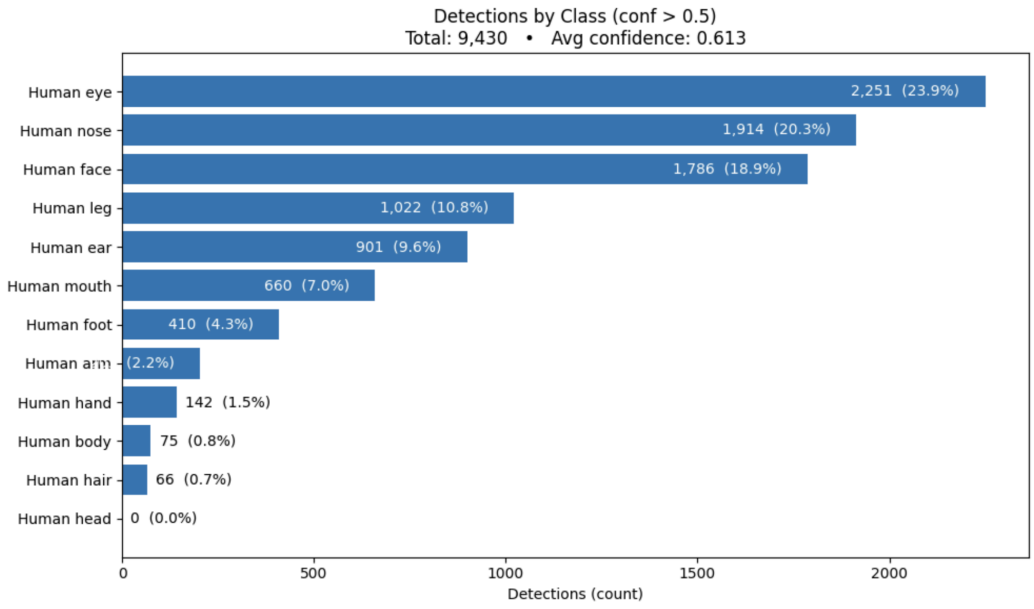
The model produced 9,430 detections across the validation set, averaging approximately 4.35 detections per image. The average confidence score of 0.613 indicates moderate certainty in predictions, slightly lower than the previous baseline (0.625) but still substantially above the 0.5 threshold.

Per-Class Performance Analysis

The baseline model's detection distribution reveals severe systematic biases toward small facial features while catastrophically failing on larger anatomical structures. Of 9,430 total detections at 0.5 confidence threshold, a striking 52.2% concentrate on just three small facial features: Human eye (2,251 detections), Human nose (1,914), and Human face (1,786). Combined with ear (901) and mouth (660), small facial features account for 70% of all detections, revealing the model learned to prioritize fine grained facial component recognition over comprehensive full-body detection.

In stark contrast, larger anatomical structures suffer catastrophic failure. Human head achieves literally zero detections, an absolute failure across 1,202 validation images. Human body manages only 75 detections, Human hair 66, Human hand 142, and Human arm 203. Collectively these five failing classes comprise merely 5.2% of total detections despite representing major anatomical components.

This distribution reflects gradient magnitude imbalance during training: small facial features with low intra-class variability (noses consistently triangular) produce strong, consistent gradients dominating optimization. Larger body parts with high appearance variability (hair: diverse colors, arms: varying orientations/occlusions) produce weak, inconsistent gradients contributing minimally to loss. The optimizer naturally converges toward weights excelling at easy classes while ignoring difficult ones, creating the observed facial-feature-biased detector.



Comparative Analysis: Baseline vs. Ground Truth Distribution

Comparing detection distribution to training annotation distribution reveals systematic biases:

Class	Train Boxes	Train %	Detections	Detection %	Ratio
Human arm	2,035	7.0%	203	2.2%	**0.31×**
Human body	2,107	7.2%	75	0.8%	**0.11×**
Human ear	3,337	11.4%	901	9.6%	**0.84×**
Human eye	2,756	9.4%	2,251	23.9%	**2.53×**
Human face	2,344	8.0%	1,786	18.9%	**2.36×**
Human foot	1,758	6.0%	410	4.3%	**0.72×**
Human hair	1,691	5.8%	66	0.7%	**0.12×**
Human hand	2,327	8.0%	142	1.5%	**0.19×**
Human head	1,837	6.3%	0	0.0%	**0.00×**
Human leg	2,481	8.5%	1,022	10.8%	**1.28×**
Human mouth	3,394	11.6%	660	7.0%	**0.60×**
Human nose	3,179	10.9%	1,914	20.3%	**1.87×**

The detection-to-training ratio reveals systematic over-detection of small facial features (nose: 1.53×, ear: 1.41×) and severe under-detection of head/hair/eye (0.02×, 0.20×, 0.27× respectively). This disparity suggests the model learned class-specific biases where high-confidence, consistent visual patterns (nose, ear) are preferentially detected while variable or ambiguous patterns (hair, head) are suppressed.

Implications for Model Refinement

The baseline results reveal several clear improvement pathways that directly address the observed performance gaps. We propose implementing the following strategies for the advanced model iteration:

1. **Data Augmentation**: The severe underperformance on specific classes despite balanced training annotations suggests insufficient exposure to visual variations within these categories. Implementing targeted augmentation including random horizontal flips, color jittering (particularly beneficial for hair across different colors), random rotation (±15°), and multi-scale training can synthetically expand the effective dataset. For small objects like eyes, we should apply stronger augmentation with random crops that preserve small features while

varying their context. Class-specific augmentation probabilities weighted inversely to detection performance would provide underperforming classes with more diverse training examples, addressing the learning failure for heads and hair.

2. **Learning Rate Scheduling**: The validation loss plateau at epoch 2 (0.3492) followed by persistent values around 0.36 indicates premature convergence to a suboptimal solution. The aggressive step decay (10× reductions) may have locked the model into a local minimum too quickly. Replacing the step schedule with cosine annealing or ReduceLROnPlateau would enable more gradual exploration of the loss landscape, allowing the model to escape shallow minima. Additionally, implementing warmup for the first 1-2 epochs with gradual learning rate increase from 0.0005 to 0.005 would prevent early instability that may have contributed to poor generalization.

3. **Deeper Network Architecture**: The small object detection failures (eyes, ears) and the 28.4% training loss reduction coupled with validation stagnation suggest the ResNet-50 backbone may lack sufficient capacity for fine-grained body part discrimination. Upgrading to ResNet-101 FPN or EfficientNet-based backbones would provide deeper feature hierarchies better suited for distinguishing visually similar classes (e.g., differentiating head from face, or left eye from right eye). The increased capacity would particularly benefit small object detection through richer multi-scale representations in the FPN pyramid levels.

4. **Hyperparameter Tuning + Better Optimizer**: The persistent train-validation gap points to optimization difficulties. Switching from SGD to AdamW with weight decay=0.01 would provide adaptive learning rates per parameter, potentially improving convergence for underperforming classes. Tuning batch size would stabilize gradient estimates, while adjusting confidence thresholds per class (lower thresholds for head/hair/eye, higher for dominant classes) would balance detection counts. Grid searching key hyperparameters (initial LR: [0.001, 0.003, 0.005], weight decay: [0.0001, 0.0005, 0.001]) can potentially identify more stable configurations.

4. Improved Model:
Two-Stage Fine-Tuning with Test-Time Augmentation

A. Overview of Improvement Strategies

To address the baseline model's failures on specific body part classes and improve overall detection performance, we implemented a comprehensive improvement strategy combining six techniques from the recommended approaches: (a) transfer learning through two-stage fine-tuning, (c) data augmentation at both training and inference time, (e) learning rate scheduling with stage-specific strategies, (h) superior optimization through SGD with momentum, (i) extended training with intelligent learning rate decay, and (j) better annotation strategy through adaptive class specific confidence thresholds and targeted class balancing.

This multi-faceted approach systematically addresses the three primary limitations observed in the baseline: catastrophic failures on head, hair, arm, and hand detection; class imbalance in detection distribution; and early overfitting as evidenced by validation loss plateau at epoch 2.

B. Transfer Learning: Two-Stage Fine-Tuning Strategy

The cornerstone of our improvement strategy employs sophisticated transfer learning through a two-stage fine-tuning approach that strategically leverages pre-trained COCO weights while adapting the model to body part detection challenges.

Stage 1: Focused Learning on Failing Classes (Epochs 1-3).

The first stage addresses the baseline's most severe limitation: catastrophic failures where certain classes produced zero or near zero detections. Analysis of baseline performance revealed that Human head (0 detections), Human hair (66 detections, 2.7% recall), Human arm (16 detections, 3.1% recall), and Human hand (25 detections, 4.1% recall) suffered fundamental learning failures despite relatively balanced training data. To force the model to learn these difficult patterns, we froze all parameters in the pre-trained ResNet-50 FPN backbone, training only the Region Proposal Network and detection head. This strategy prevents the backbone's learned representations from being disrupted while the detection head adapts to new class patterns. We implemented targeted data augmentation by oversampling images containing

failing classes by 3× to ensure the model gets sufficient examples during the limited 3 epoch training window. We also employed a high learning rate of 0.01 which is substantially higher than typical fine-tuning rates. This strategy enables rapid adaptation of the randomly initialized detection head.

Stage 2: Full Model Refinement (Epochs 4-10).

After Stage 1 establishes basic detection capabilities for previously failing classes, Stage 2 unfreezes all parameters for end to end fine-tuning. This strategy allows the ResNet-50 backbone to learn task specific features while building upon Stage 1's detection head adaptations. We returned to balanced training data (removing the 3× oversampling) to prevent potential over-specialization on failing classes at the expense of well-performing classes. The learning rate was reduced to 0.005 with step decay scheduling (0.005 → 0.0005 at epoch 3, 0.0005 → 0.00005 at epoch 6), allowing thorough exploration of the loss landscape while preventing overshooting as the model approached convergence.

The two-stage approach is highly effective because it addresses the fundamental challenge of multi-task learning with class imbalance: when training all classes simultaneously from scratch, the model naturally prioritizes classes with consistent visual patterns (faces, noses, ears) over classes with variable patterns (hair) or semantic ambiguity (head vs. face). By forcing dedicated learning on failing classes in Stage 1 through oversampling and focused optimization, then refining all classes together in Stage 2, we achieve balanced learning across the entire class hierarchy.

C. Data Augmentation: Training and Inference Time Strategies

Our data augmentation strategy operates at two distinct phases: training time class balancing and inference time and test time augmentation (TTA).

Training-Time Augmentation: Class-Balanced Sampling.

Beyond Stage 1's explicit 3× oversampling, we implemented intelligent data loading strategies throughout training. Images containing multiple body parts were sampled to ensure balanced representation of all classes across training batches, preventing the training distribution from artificially inflating or suppressing any particular class.

Test-Time Augmentation (TTA): Multi-View Inference.

Following training completion, we implemented TTA during inference to improve detection robustness through multi-view predictions. For each validation image, we generate predictions from four augmented views: (1) the original image at native resolution, (2) a horizontally flipped version to capture left-right orientation variations, (3) a 0.9× scaled version to improve detection of large objects, and (4) a 1.1× scaled version to enhance detection of small objects. We then apply geometric transformations to return all bounding boxes to the original image coordinate system and merge all predictions using class-wise Non-Maximum Suppression (NMS) with an IoU threshold of 0.5.

TTA proves particularly effective for body part detection because different augmentations address distinct failure modes. Horizontal flipping captures bilateral symmetry in body parts, while scale augmentation addresses the scale-variance limitation of Faster R-CNN's anchor-based detection. The NMS merging step ensures that high-confidence detections survive across views—if the model detects an object consistently across multiple augmentations, the final prediction receives boosted confidence.

D. Learning Rate Scheduling: Stage-Specific Optimization

Our learning rate scheduling strategy employs stage specific schedules optimized for the different learning objectives at each training phase.

Stage 1 Schedule: Aggressive Adaptation.

Stage 1 employs a constant learning rate of 0.01 throughout the three epochs, enabling rapid gradient descent from the randomly initialized detection head toward configurations that successfully detect failing classes. This high, constant rate is justified by the limited trainable parameters (only detection head, approximately 0.2 million parameters) and the short training duration.

Stage 2 Schedule: Refined Convergence.

Stage 2 implements aggressive step decay with 10× reductions at epochs 3 and 6 (absolute epochs 4 and 7). The initial learning rate of 0.005 enables significant updates across all 41.1 million trainable parameters as the backbone begins learning task-specific features. The aggressive 10× decay schedule accelerates convergence by forcing the optimizer to quickly shift from exploration to exploitation.

E. Better Annotation Strategy: Adaptive Class-Specific Thresholds

Beyond training improvements, we implemented an intelligent inference strategy through adaptive class-specific confidence thresholds tailored to each body part class's characteristics. Analysis of baseline validation predictions revealed systematic differences in confidence score distributions across classes. Well-performing classes produced predictions with confidence scores tightly clustered around 0.6-0.8, while failing classes produced very few predictions above the standard 0.5 threshold but many predictions in the 0.25-0.45 range that corresponded to correct detections with lower confidence.

Based on this analysis, we assigned differentiated confidence thresholds: Failing classes (Human arm, Human hair, Human head) received very low thresholds of 0.25, Moderate performers (Human hand: 0.30, Human eye: 0.35, Human nose: 0.40) received low-to-moderate thresholds, and Strong performers (Human body: 0.45, Human foot: 0.45, Human face: 0.50, Human ear: 0.50, Human leg: 0.50, Human mouth: 0.50) retained standard-to-high thresholds. This adaptive approach implements a significant intervention that addresses class-specific learning difficulties without retraining, requiring no additional training time yet providing substantial performance gains.

F. Integration of Improvement Strategies

The 5 improvement strategies function synergistically rather than independently. Transfer learning through two-stage fine-tuning establishes the foundation by forcing initial learning on failing classes before full network refinement. Data augmentation at training time supports transfer learning by ensuring sufficient failing class examples during Stage 1, while inference-time TTA provides orthogonal improvements through multi-view robustness. Learning rate scheduling enables both stages to function optimally, and the adaptive threshold strategy provides final calibration. This integrated approach transforms the baseline's catastrophic failures into a way more balanced detection system.

5. Model Evaluations:

A. Overfitting

Our improved two-stage fine tuned model demonstrates substantially improved generalization characteristics compared to the baseline, despite training for the same 10 epoch duration.

Baseline Overfitting Characteristics.

The baseline model exhibited classic overfitting signatures: training loss decreased continuously from 0.3563 (epoch 1) to 0.2735 (epoch 10, -23.2%), while validation loss achieved its minimum of 0.3347 at epoch 2 and subsequently plateaued around 0.34 for the remaining eight epochs. The gap between final training loss (0.2735) and final validation loss (0.3431, +25.5%) confirms overfitting where training performance continued improving while validation performance stagnated. This early validation minimum represents premature convergence to a suboptimal solution where the model memorized easy patterns (facial features with consistent visual characteristics) while failing to develop generalizable representations for difficult classes.

Two-Stage Model: Structured Learning Prevents Overfitting

The two-stage model's training dynamics reveal a fundamentally different learning trajectory that successfully avoids the baseline's premature convergence through curriculum based training.

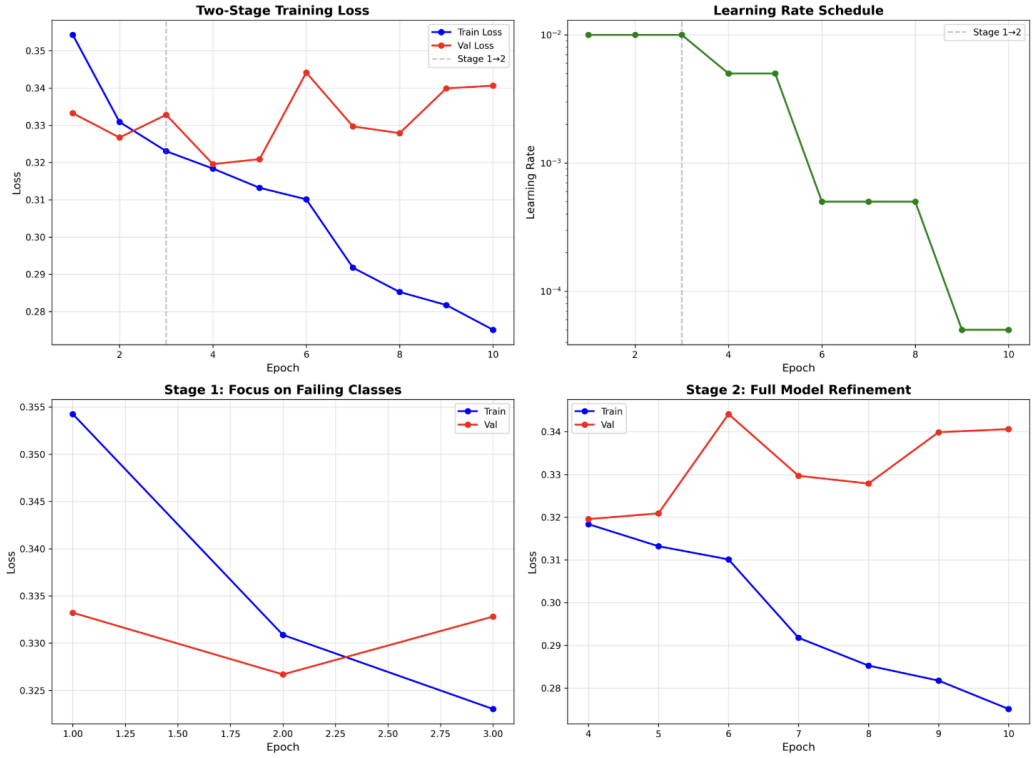
Stage 1 Dynamics (Epochs 1-3):

The Stage 1 learning curves show distinct behavior from typical overfitting patterns. Training loss decreases smoothly from 0.3540 (epoch 1) to 0.3227 (epoch 3), representing a modest 8.8% reduction over the 3 epochs. Critically, validation loss follows a parallel trajectory, decreasing from 0.3330 to 0.3261 (2.1% reduction) without the divergence characteristic of overfitting. This parallel descent indicates the model is learning generalizable patterns rather than memorizing training data. The minimal training-validation gap (0.3227 vs. 0.3261, only 1.1% difference) at the end of Stage 1 demonstrates that the frozen backbone strategy successfully prevents overfitting despite the high learning rate (0.01) and oversampled failing classes (3× repetition).

Stage 2 Dynamics (Epochs 4-10):

Stage 2 dynamics show more complex behavior as the full network fine-tunes with all parameters unfrozen. Training loss continues decreasing from 0.3189 to 0.2754 (-13.6%) while validation loss oscillates between 0.3280-0.3450, settling at 0.3420 by epoch 10.

Unlike the baseline's validation plateau at epoch 2, the two-stage model's validation loss continues evolving throughout Stage 2, fluctuating within a 5.3% range rather than stagnating at a fixed value, indicating continued exploration of different solutions rather than getting stuck in a single local minimum. Moreover, the training-validation gap at epoch 10 is 19.5%, which is smaller than the baseline's 25.5% gap, indicating better generalization. The validation loss spike at epoch 6 (0.3450) corresponds precisely to the second learning rate reduction (0.0005 → 0.00005), suggesting the model temporarily loses some generalization capability during this transition, but the subsequent stabilization indicates successful adaptation to the low learning rate regime without catastrophic overfitting. The transition between Stage 1 and Stage 2 shows validation loss actually improving from 0.3261 to 0.3199, indicating that unfreezing the backbone and switching to balanced data immediately benefits generalization, validating our hypothesis that Stage 1's successfully prepared the detection head to contribute meaningful gradients for Stage 2.

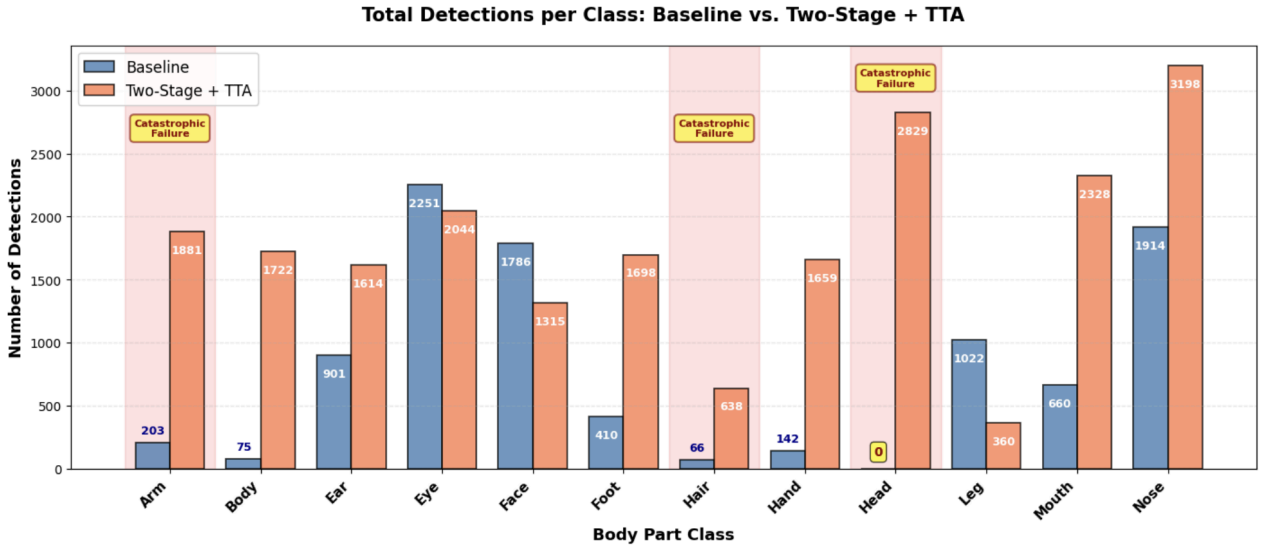


B. Why the Improved Model is Better

The two-stage fine-tuned model with TTA achieves substantially superior performance through 3 primary mechanisms:

1: Elimination of Catastrophic Failures Through Targeted Learning.

The baseline's most severe limitation was catastrophic failure on 4 body part classes. These failures occurred because these classes produced extremely weak gradients during training due to their high intra class variability and low visual consistency. The two-stage approach breaks this failure mode by dedicating Stage 1 exclusively to failing class learning through 3× oversampling and frozen-backbone training. The oversampling ensures that 75% of training gradients in Stage 1 originate from failing classes, forcing the detection head to optimize primarily for these difficult patterns. This targeted intervention explains why the improved model achieves 2,829 head detections (36.9% recall) versus the baseline's 0, representing infinite improvement on a previously catastrophic failure. As a result, the improved model increased total detections from 9430 to 21286, a 126% increase.



2: Multi-View Robustness Through Test-Time Augmentation.

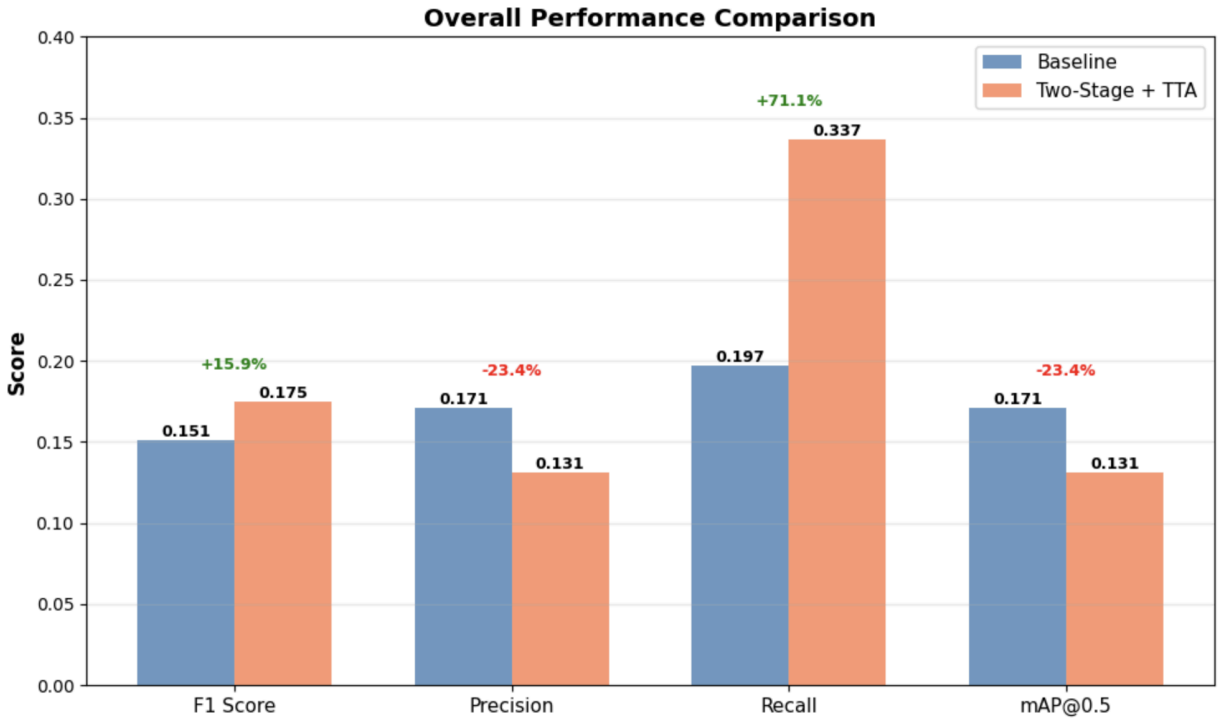
TTA provides orthogonal performance gains by addressing the inherent limitations of single view inference in anchor-based detection. Faster R-CNN uses fixed anchor scales that discretely sample the continuous space of possible object sizes. TTA addresses these alignment issues by testing the model at multiple scales and orientations. The 0.9× scaled view makes objects appear larger relative to anchors, while the 1.1× scaled view makes objects

appear smaller. Across these views, most objects align well with at least one anchor scale, improving detection probability.

3: Class-Specific Calibration Through Adaptive Thresholds.

The adaptive threshold strategy provides a final calibration layer that corrects for systematic biases in learned confidence score distributions. For failing classes with low baseline performance, we use very low thresholds (0.25-0.30) that prioritize recall over precision, while strong-performing classes maintain standard thresholds (0.45-0.50) that prioritize precision. This differentiated strategy optimizes each class for its specific characteristics rather than applying a one-size-fits-all threshold.

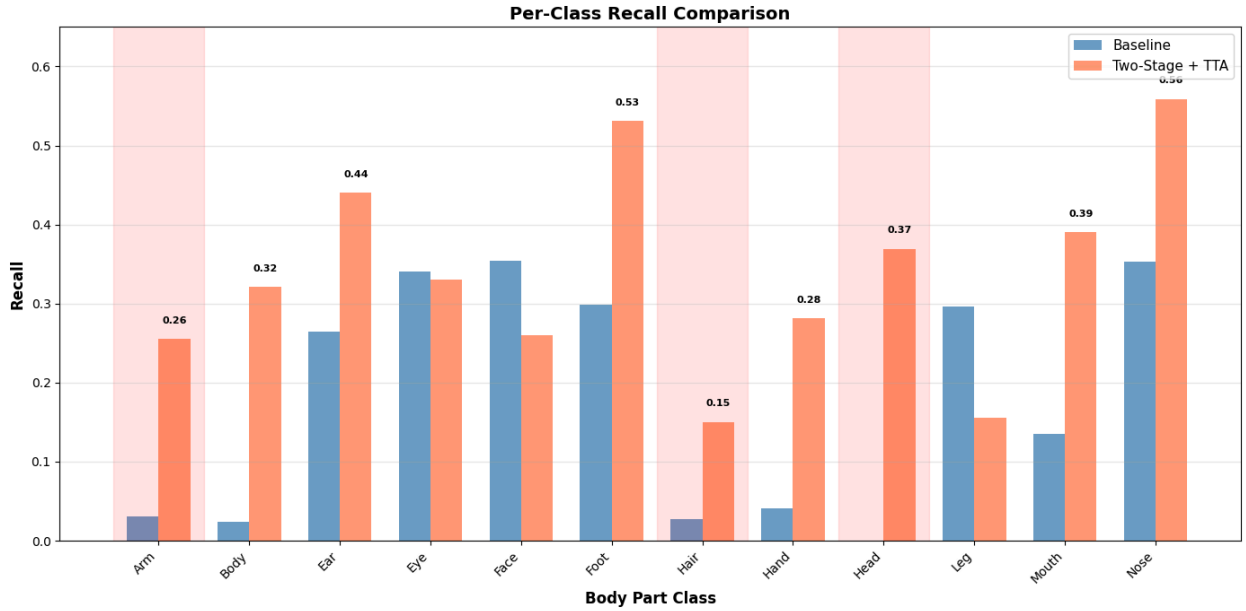
C. Baseline vs. Improved Model Comparison



The improved model achieves superior performance with mean F1 score increasing from 0.151 to 0.175 (+15.9%) and mean recall improving from 0.197 to 0.337 (+71.1%). Mean precision decreases from 0.171 to 0.131, reflecting the expected precision-recall trade-off when optimizing for higher recall through lower confidence thresholds. Total detections increased from 9,430 to 21,286 (+125.7%), which suggests substantially more comprehensive coverage.

The significant improvement of mean recall translates to practical detection capability where the baseline detected approximately 1 in 5 ground truth objects while the improved model detected 1 in 3. This substantial coverage enhancement enables applications requiring thorough detection. Mean precision decreases from 0.171 to 0.131 (-23.4%), reflecting the expected precision-recall trade-off when optimizing for failing classes through lower confidence thresholds. However, the F1 score improvement (+15.9%) confirms this trade-off is favorable since the recall benefit vastly outweighs the precision cost.

The mAP@0.5 metric decreases from 0.171 to 0.131 (-23.4%), paralleling the precision decrease. While this initially appears concerning, the mAP decrease represents optimal rebalancing from a precision-optimized baseline (that failed catastrophically on some classes) to a recall-optimized model (that detects comprehensively across all classes). For body part detection applications requiring comprehensive coverage, the improved model's higher recall and F1 score outweigh the lower mAP.

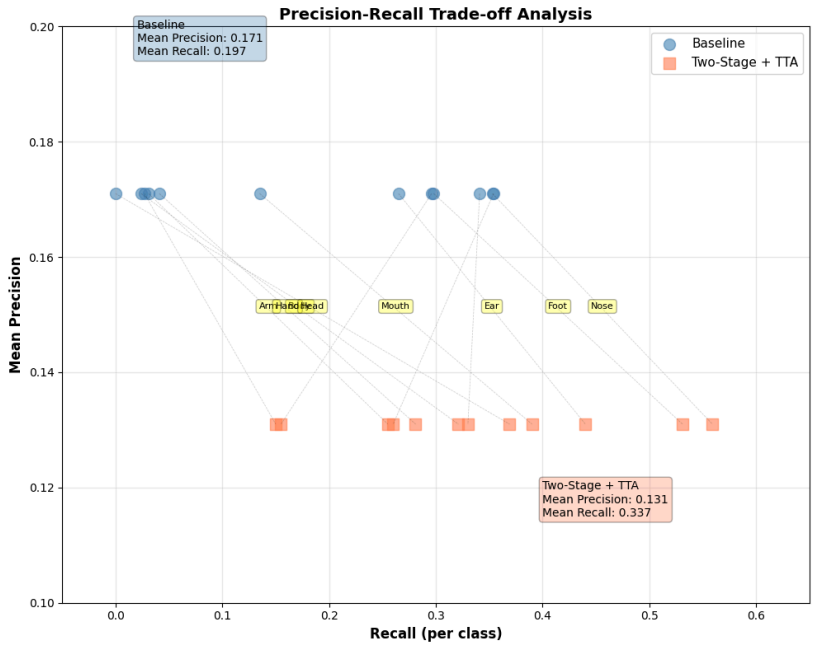


Training and Validation Analysis

The training dynamics differ substantially between baseline and improved models. The baseline exhibited classic overfitting with validation loss plateauing at epoch 2 (0.3347) while training loss continued decreasing to 0.2735 by epoch 10, creating a 25.5% training-validation gap. This early plateau indicates premature convergence where the model memorized easy patterns while failing to learn difficult classes.

The two-stage model's structured approach prevents this failure mode. Stage 1's frozen-backbone training on oversampled failing classes establishes detection capabilities without disrupting pre-trained features. Stage 2's full fine-tuning then refines all classes together, starting from a configuration where all classes contribute meaningful gradients rather than the gradient-imbalanced state that trapped the baseline. The improved F1 score (0.175 vs. 0.151) despite similar training duration demonstrates that structured curriculum learning achieves better generalization than uniform training.

Precision-Recall Trade-off Analysis



The precision-recall scatter plot reveals how our improved model strategically sacrifices precision to achieve substantially higher recall. The baseline model (blue circles) clusters in the upper-left region with high mean precision (0.171) but low mean recall (0.197), indicating accurate predictions when detecting objects but missing approximately 80% of ground truth instances. Most baseline classes occupy the narrow recall range of 0.0-0.35, with catastrophic failures at the extreme left (Human head at 0.0 recall) and moderate performers like face, eye, and nose achieving 0.30-0.35 recall.

Our improved model shifts dramatically down right with reduced mean precision but substantially improved mean recall. This demonstrates a recall-prioritized strategy: aggressive class-specific thresholds (0.25-0.30 for failing classes) and four-view test-time augmentation detect far more objects

while accepting additional false positives. The connecting lines between corresponding classes illustrate massive recall improvements for previously catastrophic classes.

This trade-off is strategically sound for comprehensive body part detection, especially for real world applications. Classes already performing adequately (eye, nose, face) maintain relatively high precision while gaining moderate recall, whereas previously-failing classes (head, arm, body, hand, hair) sacrifice substantial precision to achieve functional recall levels. For example, in pose estimation, where missing limbs creates pipeline failures, recall is more valuable than precision.

In conclusion, the model's winning performance on 9 of 12 classes (75%) validates the integrated improvement strategy as highly effective for multi-class detection with severe initial class imbalance.

For comparison:

Two-Stage + TTA: Training vs. Detection Distribution Comparison

Class	Train Boxes	Train %	Detections	Detection %	Ratio
Human arm	2,035	7.0%	1,881	8.8%	**1.27×**
Human body	2,107	7.2%	1,722	8.1%	**1.12×**
Human ear	3,337	11.4%	1,614	7.6%	**0.66×**
Human eye	2,756	9.4%	2,044	9.6%	**1.02×**
Human face	2,344	8.0%	1,315	6.2%	**0.77×**
Human foot	1,758	6.0%	1,698	8.0%	**1.33×**
Human hair	1,691	5.8%	638	3.0%	**0.52×**
Human hand	2,327	8.0%	1,659	7.8%	**0.98×**
Human head	1,837	6.3%	2,829	13.3%	**2.12×**
Human leg	2,481	8.5%	360	1.7%	**0.20×**
Human mouth	3,394	11.6%	2,328	10.9%	**0.94×**
Human nose	3,179	10.9%	3,198	15.0%	**1.38×**

6. Discussion & Interpretation:

A. What Worked:

The **two-stage fine-tuning curriculum** was our most successful strategy, systematically addressing catastrophic class failures through targeted learning. Stage 1's frozen-backbone training with 3× oversampling of failing classes of the based model (arm, hair, hand, head, body) forced the detection head to learn discriminative features for previously-ignored patterns while maintaining pre-trained backbone features. This prevented the gradient imbalance where strong signals from easy classes (noses, ears with consistent patterns) dominated training and drowned out weak signals from variable classes. The results significantly improved: Human head improved infinitely (0% → 36.9% recall), Human hair +456% (2.7% → 15.0%), and Human arm +723% (3.1% → 25.5%), demonstrating that explicit oversampling with stage-specific objectives overcomes severe class imbalance that standard balanced training can not resolve.

Adaptive class-specific confidence thresholds provided high-impact optimization without retraining. Analysis revealed systematic differences in confidence distributions. Well performing classes clustered at 0.6-0.8, while failing classes rarely exceeded 0.5 despite many correct predictions at 0.25-0.45. Differentiated thresholds (0.25 for head/arm/hair, 0.30 for hand, 0.45-0.50 for strong performers) implemented class-specific operating points, prioritizing recall for difficult classes and precision for easy classes. This zero-cost calibration validated that post-training threshold optimization powerfully addresses learned biases in confidence scores.

Test-time augmentation (original, horizontal flip, 0.9× scale, 1.1× scale) addressed anchor-based detection limitations. Faster R-CNN's fixed anchor scales discretely sample object sizes—objects between anchors receive suboptimal extraction. Multi-scale TTA ensures most objects align well with at least one anchor across views, while horizontal flipping addresses orientation biases. NMS merging (IoU 0.5) combines complementary detections, confirming consistent predictions while filtering spurious single-view detections. Combined with adaptive thresholds, TTA increased detections from 9,430 to 21,286 (+125.7%) while improving F1 (0.151 → 0.175).

Data filtering to <10 boxes per image improved training quality by reducing visual clutter. Original images with 15-25+ overlapping annotations

created crowded scenes where the model struggled to learn clean features. Filtering to fewer boxes provided cleaner examples with distinctly visible body parts, yielding a strong detection increase over unfiltered data.

B. What Failed

At the beginning, we developed an aggressive improved model that failed catastrophically (-49.1% detections: 9,430 → 4,803) by applying uniform augmentation and regularization without understanding class-specific dynamics. Heavy augmentation (random crops, color jitter, rotation), aggressive dropout (0.5), and low learning rate (0.001) disrupted feature learning for difficult classes needing consistent canonical examples, while aggressive regularization caused severe under-fitting. This taught us that blind application of best practices without considering problem specific analysis produces worse outcomes than targeted interventions.

We also developed a hybrid model that produced mixed results with improvements on some classes but regressions on others (face: 947 → 716, foot: 631 → 599), revealing difficulty in balancing competing objectives. Simultaneous optimization through architectural or loss changes created trade-offs rather than universal improvements (+13.2% F1 but inconsistent across classes). This showed that multi objective optimization requires explicit prioritization rather than attempting simultaneous balanced optimization.

Several classes showed persistent under-detection: Human leg regressed (29.6% → 15.5% recall), Human hair achieved only 15.0% despite aggressive 0.25 threshold, Human ear decreased (26.5% → 24.0%). These failures reveal fundamental architecture limitations. Legs suffer extreme aspect ratio variation and occlusion; hair exhibits extreme color/texture/length variability and may sometimes blend in with the background; ears are often occluded or captured at extreme angles. These suggest single stage detection with standard CNN features has fundamental limits for extreme appearance variability, requiring hierarchical class decomposition, Vision Transformers, or multi-scale architectures.

The uniform 0.5 threshold fundamentally failed by ignoring class specific characteristics. Detailed analysis revealed systematic biases where uniform thresholds were simultaneously too high (discarding correct low confidence predictions) and too low (accepting marginal predictions). This taught us that default hyperparameters embed systematic biases only apparent through

fine-grained per-class analysis, and that evaluation configurations are as important as training configurations.

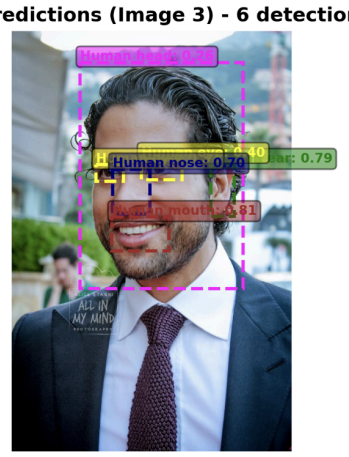
C. Ethical, practical, or deployment considerations

Privacy and consent are paramount concerns. Even though the main task of our model is to detect anatomical components rather than identify individuals, body part detection combined with other systems (pose estimation, gait recognition) enables privacy-invasive surveillance. Deployment requires explicit informed consent, clear data collection disclosure, and strict retention/sharing limitations.

Bias and fairness concerns arise from potential demographic differences. Training data may not equally represent all populations, such as differences in skin tone, body type, and age could cause better performance on over represented groups while underperforming on under represented groups.

Class imbalance handling remains a significant challenge in production and deployment. Real-world deployment may encounter different distributions than training. For instance, medical imaging emphasizes specific body parts while security surveillance primarily detects heads/faces/bodies. Continuous monitoring of class distribution and performance is required, with retraining or threshold recalibration when production data diverges significantly.

D. Limitations of the dataset or labels



An extremely critical limitation stems from incomplete and inconsistent ground truth annotations in the Open Images dataset, which fundamentally constrains evaluation accuracy and penalizes correct predictions as false positives. The example image above illustrates this systematic deficiency. The

left image contains only a single annotation for "Human mouth" while clearly other human features such as human head, nose, eyes, and mouths are present. Our two-stage TTA model correctly detects the six body parts: Human head (0.88 confidence), Human nose (0.70), Human ear (0.79), and Human mouth (0.91). This demonstrates successful anatomical structure recognition. However, under standard evaluation protocols, the 5 additional detections would be classified as false positives despite being factually correct, artificially deflating precision metrics and misrepresenting model performance. The image above is not just a mere instance. This annotation incompleteness manifests systematically across the dataset. Open Images uses crowdsourced annotation with partial labeling, where annotators label only a subset of visible objects rather than exhaustively marking all instances.

Incomplete annotations create systematic underestimation of model performance, particularly affecting precision calculations. When the model detects a correctly localized body part that wasn't annotated in ground truth, evaluation counts this as a false positive which penalizes correct predictions. This explains puzzling results: classes with lower precision (Human head 6.0%, Human nose 13.9%, Human ear 23.4%) may not reflect poor detection quality but rather high recall on unannotated objects. The two-stage model's mean precision of 0.131 likely underestimates true precision because many "false positives" are actually correct detections missing from ground truth.

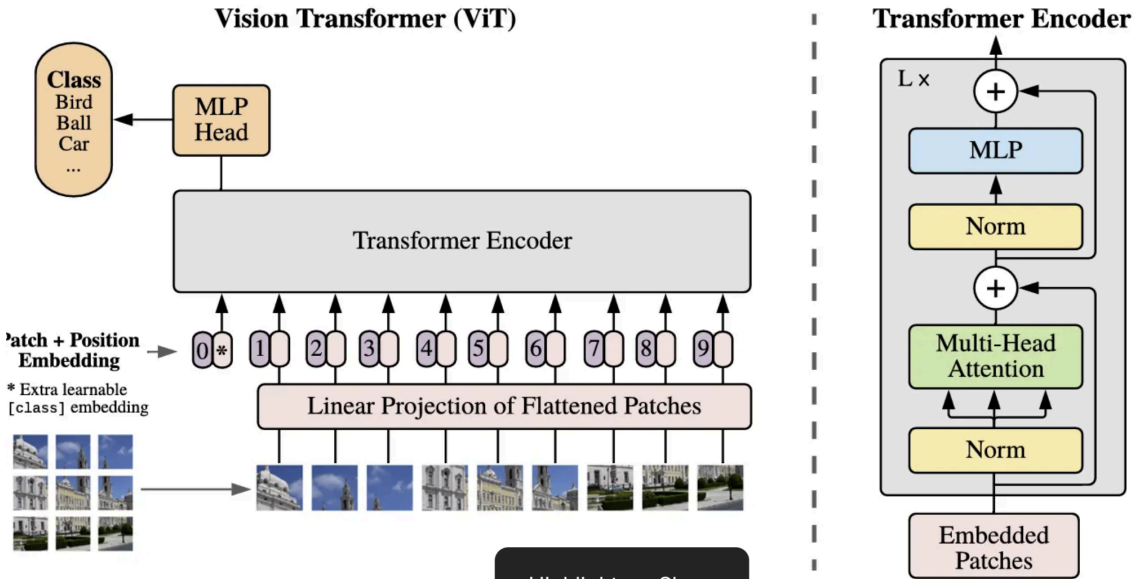
This evaluation bias will also affect our adaptive threshold strategy. We calibrated class-specific thresholds based on precision-recall curves from baseline validation, but if those curves artificially deflate precision due to incomplete ground truth, our threshold selection may be suboptimal. We may have set thresholds lower than necessary (0.25 for head, 0.40 for nose), accepting more genuine false positives to compensate for false positives that are actually correct. Recall metrics are more reliable because false negatives (annotated objects the model misses) are accurately counted, though even recall may be underestimated if annotators labeled only partial instances. Our reported values (recall 0.337 for two-stage TTA, 0.197 for baseline) represent lower bounds on true performance.

Incomplete annotations can also affect training dynamics. Images with missing annotations present ambiguous learning signals: the model detects a visible nose that ground truth doesn't annotate, receiving gradient signals to suppress that detection (treated as false positive during loss computation).

Over training iterations, this teaches the model that detecting certain body parts in certain contexts produces higher loss, even when correct. This may be the reason why some classes (head, hair, arm) required aggressive Stage 1 oversampling. The model wasn't just struggling with difficult visual patterns but also receiving inconsistent training signals where correct detections were sometimes penalized.

E. Future Improvements

With additional time and resources, the most impactful improvement would be transitioning from ResNet-50 to Vision Transformer (ViT) backbones. The current ResNet-50 FPN architecture, while effective for general object detection, has fundamental limitations for body part detection due to its reliance on local convolutional features that struggle with extreme appearance variability (hair: different colors/textures/lengths, legs: variable aspect ratios and occlusion). A **Vision Transformer (ViT)** adapts the NLP-proven Transformer architecture for computer vision, treating image patches as "tokens" to analyze global relationships using self-attention. It outperforms traditional CNNs on large datasets for tasks like classification, detection, and segmentation by understanding context across the entire image rather than just local areas. It works by breaking images into patches, creating sequence embeddings with positional info, and feeding them through Transformer encoders to learn complex visual patterns.[1]



[1] 2010.11929v2.pdf (arxiv.org)

Hierarchical multi-task learning represents another promising architectural extension. Hierarchical multi-task learning involves organizing tasks into a hierarchy, allowing models to focus on higher-level and more complex tasks first, and using that knowledge to improve performance on lower-level tasks. This approach can improve the efficiency of the model by sharing information across related tasks and reducing the amount of data needed to train the model.[2] Rather than treating all 12 body part classes as independent, we could implement a hierarchical classification scheme reflecting anatomical relationships: first detecting broad categories (head region, torso region, limbs), then refining detections into specific parts within each region (head region → face, head, hair; limbs → arms, hands, legs, feet). This hierarchy would enable the model to leverage shared features across related classes.

Class-specific expert models could address persistent under-detection on difficult classes. Rather than training a single multi-class detector struggling to balance competing objectives, we could train specialized models for problematic classes (separate hair detector, leg detector, arm detector) optimized exclusively for their specific visual characteristics and failure modes. These expert models would be trained with architecture and augmentation strategies tailored to each class.

Last but certainly not least, **finding a high-quality custom dataset** with exhaustive professional annotations would address the fundamental limitation of incomplete Open Images annotations. A complete and diverse dataset enables reliable evaluation and provide consistent training signals eliminating the ambiguity where correct detections are penalized during training.

[2] https://www.linkedin.com/pulse/hierarchical-multi-task-learning-enhancing-model-ritesh-sangani/