Data Collection and Preprocessing for YOLO, Faster-RCNN and RetinaNet on the Stanford Dogs Dataset

Okeoghene Tracy Erivwode,Kayode Sunday Ajayi ,Usoh Esther Emem

***Abstract*—We document the data collection and preprocessing pipeline of data used in three instances: to train a YOLOv8 based object detector, to train a Faster R-CNN based object detector, and to train a RetinaNet based object detector, all on the Stanford Dogs dataset. The dataset contains over twenty thousand dog images across 120 breeds with Pascal VOC XML annotations. We convert annotations to YOLO format, standardize images for model input, and perform exploratory data analysis (EDA) to characterize class balance, image resolutions, and bounding-box geometry. The analysis informs design choices for augmentation, input size, and sampling.**

***Index Terms*—Object detection, YOLO, Stanford Dogs, Pascal VOC, Dataset curation, Preprocessing, EDA**

I. Introduction

Fine-grained dog breed recognition and localization present a challenging benchmark task due to high intra-class variation and significant inter-class similarity. In this work, we train and evaluate three object detection architectures; YOLOv8, Faster R-CNN, and RetinaNet, on the Stanford Dogs dataset, utilizing Ultralytics and PyTorch based tooling for implementation and analysis. This report details the dataset composition, ethical and licensing considerations, preprocessing workflow, and empirical characteristics revealed through exploratory data analysis (EDA). These findings guide the data preparation and augmentation strategies employed across all three models.

II. Dataset Description

A. Source and Licensing

We use the Stanford Dogs dataset (Kaggle mirror by Jessica Li). The dataset comprises images and Pascal VOC XML annotations organized by breed. Licensing is Creative Commons Attribution-ShareAlike 3.0.

B. Composition

The dataset contains 120 dog breeds with images typically ranging from a few hundred to several thousand pixels in width/height. Each image has at least one annotated dog instance. Our EDA indicates a single dominant object per image and moderate variation in resolution.

C. Directory Structure

A typical layout is `images/` and `annotations/`, each with 120 subfolders (one per breed). XML files follow VOC schema with object class labels and bounding boxes.

D. Ethical Considerations

No personally identifiable information is intentionally targeted; however, backgrounds may contain people or private locations. We avoid publishing sensitive metadata and follow dataset licensing requirements. Potential class imbalance across breeds is acknowledged and mitigated during training.

III. Preprocessing Pipeline

A. Annotation Parsing and Conversion

Pascal VOC XML annotations are parsed to extract $(\text{name}, \text{xmin}, \text{ymin}, \text{xmax}, \text{ymax})$ and corresponding image dimensions (W, H). Each bounding box is converted to model-specific formats. For YOLOv8, annotations are normalized to $(c, x/W, y/H, w/W, h/H)$, where (x, y) represents the box center. For Faster R-CNN and RetinaNet, annotations are preserved in absolute pixel coordinates and stored in COCO-style JSON dictionaries containing image metadata, category IDs, and bounding-box coordinates in $(x_{\min}, y_{\min}, w, h)$ format. All annotation files are cross-validated to ensure class label consistency across the three model pipelines.

B. Image Preparation

All models share the same base image preprocessing workflow. Images are read and resized while preserving aspect ratio. For YOLOv8, letterboxing is applied to match the model's fixed input dimension (typically 640×640), with automatic padding handled by the Ultralytics training framework. For Faster R-CNN and RetinaNet (implemented via PyTorch or Detectron2), images are rescaled such that the shorter side is 800 pixels (with a maximum longer side of 1333 pixels) following common COCO training conventions. Pixel normalization is performed using ImageNet mean and standard deviation statistics for all three models. Data augmentation is applied with framework-specific utilities. YOLOv8 leverages built-in augmentations including random horizontal flip, photometric distortion, mosaic, and mixup. Faster R-CNN employs random flip, scale jittering, and color perturbations using torchvision transforms. RetinaNet incorporates similar geometric and photometric augmentations but typically omits mosaic and mixup to maintain consistency with its focal-loss training regime. All preprocessed images and annotations are cached for reproducibility and to ensure identical dataset splits across the three model training pipelines.

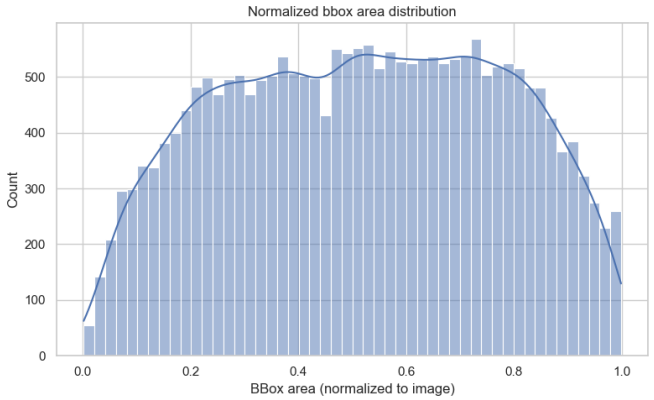


Fig. 1: Normalized bounding-box area distribution (box area divided by image area). The mass between 0.3 and 0.8 indicates dogs occupy a substantial portion of the frame, guiding input-size and anchor choices.

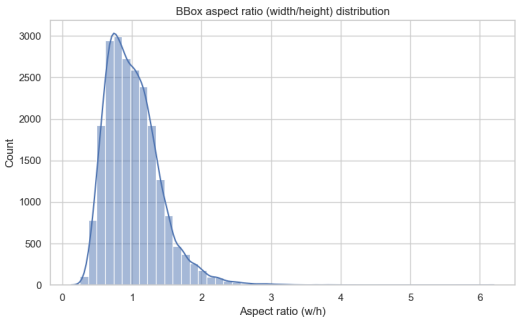


Fig. 2: Bounding-box aspect ratio (w/h). The mode near 1.0 suggests roughly square boxes are common, but a long right tail indicates elongated poses; augmentation should include aspect-preserving scaling.

C. Data Splits and Reproducibility

We create train/validation/test splits with stratification by breed where feasible. Random seeds are fixed and package versions are recorded to ensure reproducibility.

IV. Exploratory Data Analysis

We summarize four key EDA results using the user's generated figures.

A. Qualitative Samples

Annotated exemplars illustrate label quality across breeds and scales (Fig. 5).

V. Implications for Modeling

Given the sizeable box areas and near-unimodal aspect ratios, default YOLO anchors (or anchor-free variants) perform well with moderate input sizes (e.g., 640). Class imbalance and occasional extreme aspect ratios motivate mixup/mosaic, class-aware sampling, and scale jitter. High-resolution outliers suggest enabling multi-scale training. For Faster R-CNN, the

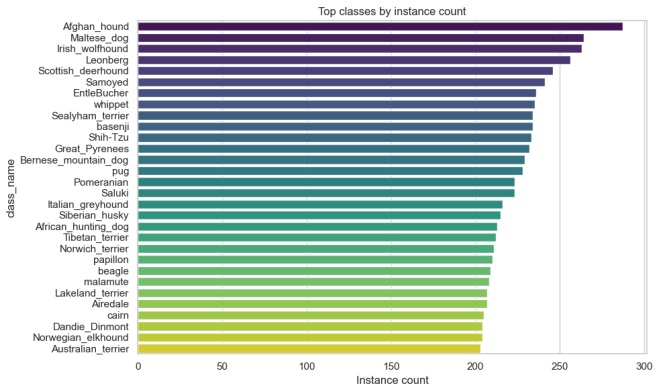


Fig. 3: Top classes by instance count (subset of breeds). The distribution is slightly imbalanced across breeds; rare classes may benefit from sampling or targeted augmentation.

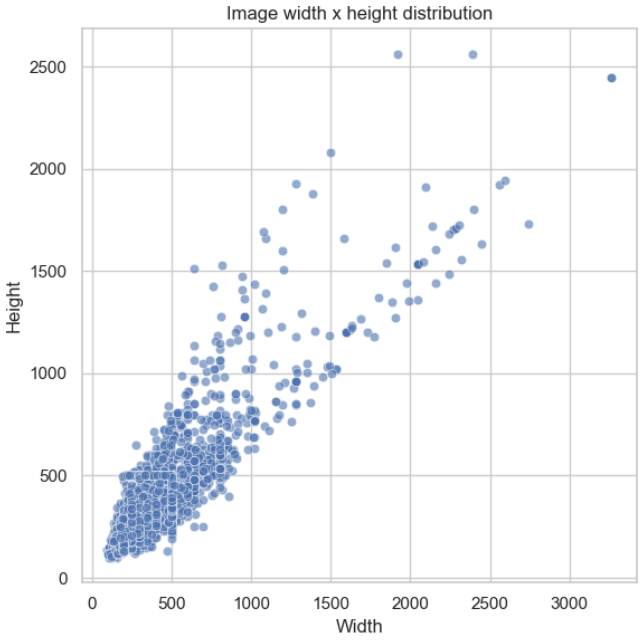
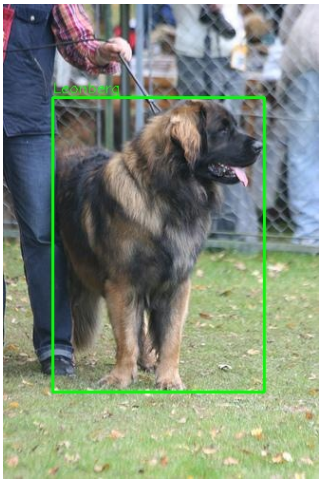
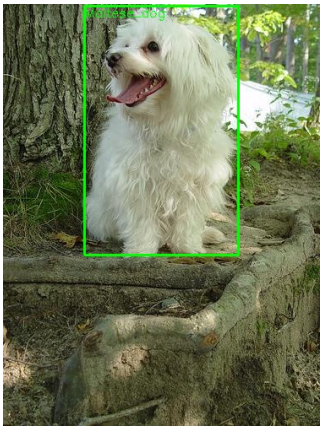


Fig. 4: Image width vs. height scatter. Resolutions vary from a few hundred up to several thousand pixels; letterboxing avoids geometric distortion during resizing.

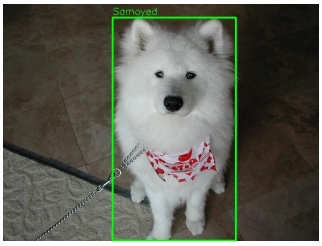
generally square bounding boxes align well with its Region Proposal Network (RPN) default anchor configurations, though incorporating additional aspect ratios (e.g., 1:2 and 2:1) may better capture elongated instances identified in the EDA. The moderate class imbalance implies that balanced sampling or focal loss weighting could stabilize training. Given the variable image resolutions, employing image pyramids or adaptive resizing can help the RPN maintain proposal recall across scales. For RetinaNet, the dense prediction strategy benefits from the observed aspect-ratio consistency, allowing default anchor priors to remain effective. However, the dataset's mild long-tail class distribution justifies the use of its focal loss



(a) Leonberg



(b) Maltese dog



(c) Samoyed

(d) Scottish deerhound

Fig. 5: Sample images with bounding boxes overlaid, demonstrating variability in pose, scale, and background.

formulation to mitigate foreground–background imbalance. Multi-scale training is also advantageous here, ensuring robust feature-map coverage for both small and large dog instances present in the dataset.

.

VI. Conclusion

We presented a reproducible dataset description and preprocessing workflow for training YOLOv8, Faster R-CNN, and RetinaNet on the Stanford Dogs dataset. The accompanying exploratory data analysis (EDA) provided quantitative insights into class balance, image resolution, and bounding-box geometry, informing model-specific choices in anchor design, input scaling, and augmentation strategies.

References

[1] A. Khosla, N. Jayadevaprakash, B. Yao, and L. Fei-Fei, "Novel Dataset for Fine-Grained Image Categorization," in *First Workshop on Fine-Grained Visual Categorization*, IEEE Conference on Computer Vision and Pattern Recognition, 2011.

[2] Stanford Dogs Dataset (Kaggle mirror by J. Li), https://www.kaggle.com/datasets/jessicali9530/stanford-dogs-dataset.

[3] Ultralytics YOLO documentation, https://docs.ultralytics.com/.