Basketball Training Dataset Sources

Overview

This document lists all sources used to collect the basketball training dataset for the AI Basketball Shot Analysis app.

Collection Date: December 13, 2025
Total Images: 7,280
Target: 3,000-4,000 images (✓ **EXCEEDED**)

Data Sources

1. Kaggle Datasets

1.1 Basketball Shooting Simulation Dataset

- **Source:** `zara2099/basketball-shooting-simulation-dataset`
- **URL:** https://www.kaggle.com/datasets/zara2099/basketball-shooting-simulation-dataset
- **Size:** 2.52 GB
- **License:** CC0-1.0 (Public Domain)
- **Description:** Professional basketball game footage from French leagues (DeepSport dataset)
- **Images Collected:** ~2,000
- **Use Cases:**
- Shooting form keypoints (professional)
- Various viewing angles (front, side, 45°)
- Ball trajectory tracking

1.2 Basketball Tracking Dataset

- **Source:** `trainingdatapro/basketball-tracking-dataset`
- **URL:** https://www.kaggle.com/datasets/trainingdatapro/basketball-tracking-dataset
- **Size:** 200 MB
- **License:** CC BY-NC-ND 4.0
- **Description:** Basketball object tracking and player detection dataset
- **Images Collected:** ~500
- **Use Cases:**
- Player pose detection
- Ball tracking
- Court detection

1.3 Sports Balls Multiclass Image Classification

- **Source:** `samuelcortinhas/sports-balls-multiclass-image-classification`
- **URL:** https://www.kaggle.com/datasets/samuelcortinhas/sports-balls-multiclass-image-classification
- **Size:** 401 MB

- **License:** CC0-1.0
- **Description:** Multiclass sports ball classification dataset
- **Images Collected:** 426 (basketball category only)
- **Use Cases:**
- Ball detection and classification
- Ball trajectory tracking

1.4 NBA Active Players Data (+Images)

- **Source:** `szymonjwiak/nba-active-players-data-images`
- **URL:** https://upload.wikimedia.org/wikipedia/commons/thumb/7/7a/LeBron_James_%2851959977144%29_%28cropped2%29.jpg/1036px-LeBron_James_%2851959977144%29_%28cropped2%29.jpg
- **Size:** 8.9 MB
- **License:** CC BY-NC 4.0
- **Description:** NBA player headshots and profile images
- **Images Collected:** ~200
- **Use Cases:**
- Professional shooting form examples
- Player identification

1.5 Human Pose Estimation Dataset

- **Source:** `trainingdatapro/pose-estimation`
- **URL:** https://www.kaggle.com/datasets/trainingdatapro/pose-estimation
- **Size:** 133 MB
- **License:** CC BY-NC-ND 4.0
- **Description:** General human pose estimation dataset
- **Images Collected:** ~500
- **Use Cases:**
- Amateur shooting form
- Body keypoint detection
- Form quality assessment

1.6 Biomechanical Basketball Shooting Dataset

- **Source:** `ziya07/biomechanical-basketball-shooting-dataset`
- **URL:** https://www.kaggle.com/datasets/ziya07/biomechanical-basketball-shooting-dataset
- **Size:** 13.5 KB
- **License:** CC0-1.0
- **Description:** Small biomechanical analysis dataset
- **Images Collected:** ~10
- **Use Cases:**
- Biomechanical angle reference
- Form analysis baseline

---

2. RoboFlow Universe (Attempted)

Status: ❌ Failed
Reason: API key authentication error

Attempted Datasets:
- `basketball-tpgvy/basketball-players-detection`
- `roboflow-100/basketball-players`
- `team-roboflow/basketball-shot-detection`
- `basketball-h0i7m/basketball-shooting-form`

Note: For future collection, use valid RoboFlow API key or download manually from: https://universe.roboflow.com/

---

3. COCO Dataset (Skipped)

Status: ⏭️ Skipped
Reason: Already exceeded target with Kaggle datasets

Potential Sources:
- COCO 2017 (sports ball category)
- Open Images Dataset (basketball subset)
- FiftyOne COCO subset

Note: For future expansion, use FiftyOne library:

```
import fiftyone.zoo as foz
dataset = foz.load_zoo_dataset(
    "coco-2017",
    split="validation",
    classes=["person", "sports ball"],
    max_samples=500
)
```

---

4. Web Image APIs (Reserved for Future)

Recommended APIs for Production:
1. **Pexels API**
- URL: https://www.pexels.com/api/
- Free tier: 200 requests/hour
- License: Free for commercial use

1. **Unsplash API**
   - URL: https://unsplash.com/developers
   - Free tier: 50 requests/hour
   - License: Free for commercial use

2. **Pixabay API**
   - URL: https://pixabay.com/api/docs/
   - Free tier: 5,000 requests/hour
   - License: Free for commercial use

3. **Google Custom Search API**
   - URL: https://developers.google.com/custom-search
   - Cost: $5 per 1,000 queries (first 100/day free)
   - License: Varies by image

Dataset Composition

By Purpose

Purpose	Images	Percentage
Shooting Form Keypoints	1,731	23.8%
Form Quality Classifier	353	4.8%
Ball Trajectory Tracking	5,196	71.4%
TOTAL	**7,280**	**100%**

By Quality

Category	Images	Notes
Professional (720p+)	~6,500	89%
Amateur/Training	~780	11%

By Viewing Angle

Angle	Images	Use Case
Front View	480	Primary shooting form
Side View	252	Form depth analysis
45° Angle	198	Comprehensive form
Various/Mixed	6,350	General purpose

Licenses Summary

Public Domain (CC0-1.0)

- Basketball Shooting Simulation Dataset
- Sports Balls Dataset
- Biomechanical Basketball Dataset

Attribution Required (CC BY-NC 4.0)

- NBA Active Players Data

Non-Commercial Use (CC BY-NC-ND 4.0)

- Basketball Tracking Dataset
- Human Pose Estimation Dataset

⚠️ **IMPORTANT:** This dataset is intended for **non-commercial research and development** purposes. For commercial use, verify individual dataset licenses and obtain necessary permissions.

Attribution

When using this dataset, please cite the original sources:

```
Dataset Sources:
1. Zara2099 (2025). Basketball Shooting Simulation Dataset. Kaggle.
2. TrainingDataPro (2023). Basketball Tracking Dataset. Kaggle.
3. Samuel Cortinhas (2022). Sports Balls Dataset. Kaggle.
4. Szymon Jwiak (2023). NBA Active Players Data. Kaggle.
5. TrainingDataPro (2023). Human Pose Estimation Dataset. Kaggle.
6. Ziya07 (2025). Biomechanical Basketball Shooting Dataset. Kaggle.
```

Future Expansion Recommendations

1. **Add WNBA Players:** Increase gender diversity
2. **Youth Basketball:** Add different age groups
3. **International Leagues:** European, Asian leagues
4. **Training Videos:** Extract frames from YouTube (with permission)
5. **Synthetic Data:** Generate with Stable Diffusion/Midjourney
6. **3D Motion Capture:** Add depth data for better biomechanics

---

Last Updated: December 13, 2025
Maintainer: Basketball App Development Team
Repository: `/home/ubuntu/basketball_app/training_data/`