Fashion Datasets Analysis for CuratorAI

1. DeepFashion Datasets & Commercial Licensing

Concern: DeepFashion datasets may not be appropriate for commercial use despite being technically superior.

Response:

You're correct about the licensing concern. DeepFashion datasets are primarily released for academic research purposes. For a commercial platform like CuratorAI, we need to either:

- Obtain explicit commercial licensing (if available)
- Use datasets with permissive commercial licenses
- Build our own proprietary dataset over time (if necessary for scalability but not feasible for initial launch)

Recommendation: I recommend we verify the exact license terms of each DeepFashion variant, but plan to proceed with commercially-viable alternatives for our initial launch.

---

2. Multiple Datasets Strategy & Migration Risks

Concern: Feasibility of using multiple datasets and risks of switching datasets after initial training with DeepFashion.

Response:

Using multiple datasets is not only possible but often recommended for production ML systems. Here's the strategic approach:

Training Phase: - We can combine multiple datasets to improve model generalization and reduce bias - Diverse data sources help the model handle real-world variations better

Transfer Learning: - If we prototype with DeepFashion (non-commercial), we can use transfer learning to adapt the model to our production dataset with minimal performance loss - Pre-trained weights can be fine-tuned on commercially-licensed data

Hybrid Approach: - Train base models on permissive datasets - Fine-tune on domain-specific data - Implement continuous learning pipelines

Risk Mitigation:

The risk you identified about switching datasets is valid—model performance depends heavily on training data similarity to production data.

Mitigation Strategies: - Use datasets with similar image characteristics (resolution, angles, styling) - Plan for a fine-tuning phase with sufficient

commercially-licensed data - Implement A/B testing to validate performance after dataset migration - Maintain dataset similarity metrics during transition

---

3. Kaggle Fashion Images Assessment

Concern: Kaggle Fashion Images appears favorable but may create additional work.

Response:

Kaggle Fashion datasets are indeed a solid choice. The additional work typically involves:

Data Preparation: - Data cleaning and quality assurance - Annotation/labeling if metadata is incomplete - Standardization of image formats and resolutions - Potential augmentation to increase dataset size

Benefits: - Many Kaggle datasets come with permissive licenses - Diverse, real-world images align well with commercial needs - Active community support and documentation - Regular updates and improvements

Work Estimation: - Initial data pipeline setup: 2-3 weeks - Quality assurance and validation: 1-2 weeks - Ongoing maintenance: Minimal after setup

The additional effort is justified by the commercial viability and quality of results.

---

4. Fashion MNIST Evaluation

Concern: Fashion MNIST may not provide the reality experience users expect.

Response:

Agreed. Fashion MNIST consists of 28×28 grayscale images of simplified clothing items—it's excellent for algorithm prototyping but insufficient for a production fashion AI platform.

Limitations: - Low resolution (28×28 pixels) - Grayscale only (no color information) - Simplified representations - Limited attribute detail (no texture, patterns, materials)

User Expectations: - High-resolution, realistic imagery - Detailed visual attributes (color, texture, patterns) - Professional-quality product representation - Realistic styling and context

Verdict: Fashion MNIST would not meet production requirements for CuratorAI.

Recommendation

I support your recommendation for **Kaggle Fashion Images as the primary dataset**, supplemented by:

1. Additional Open-Source Datasets

- Certain subsets from OpenImages with commercial-friendly licenses
- iMaterialist Fashion (check latest license terms)
- ASOS, H&M, or other retailer-released datasets

2. Synthetic Data Generation

- For specific underrepresented categories
- To augment edge cases and rare items
- To increase dataset diversity

3. User-Generated Content

- With proper consent and licensing
- As we scale, to continuously improve model accuracy
- Feedback loop for real-world performance optimization

Dataset Selection Criteria

Prioritize datasets with:

Commercial use permissions - Explicit licensing for commercial deployment **High-resolution real-world images** - Minimum 512×512, preferably 1024×1024+ **Rich metadata** - Categories, attributes, descriptions, tags **Diversity** - Styles, demographics, fashion domains, seasonal variations **Data quality** - Professional photography, consistent formatting **Scalability** - Large enough for deep learning (100K+ images minimum) **Update frequency** - Active maintenance and version updates

---

Dataset Comparison Matrix

Dataset	Size	License	Resolution	Metadata Quality	Commercial Use	Suitability Score
DeepFashion	800K+ images	Academic Only	High (varies)	Excellent	No	6/10 (Technical: 9/10, Legal: 2/10)

Dataset	Size	License	Resolution	Metadata Quality	Commercial Use	Suitability Score
Kaggle Fashion Product Images	44K+ images	CC0/Open	High (2400×1800)	Good	Yes	8/10
Fashion MNIST	70K images	MIT License	Low (28×28)	Basic	Yes	3/10 (Prototype only)
iMaterialist Fashion	1M+ images	Custom (verify)	Medium-High	Excellent	Check Terms	7/10
OpenImages (Fashion Subset)	100K+ images	CC BY 4.0	Varies	Good	Yes	7/10
ASOS/H&M Public Sets	Varies	Custom	High	Excellent	Check Terms	7/10

Scoring Criteria (Out of 10):

- **License Compliance:** Commercial viability (3 points)
- **Data Quality:** Resolution, clarity, professionalism (2 points)
- **Metadata Richness:** Annotations, attributes, descriptions (2 points)
- **Scale & Diversity:** Size and variety (2 points)
- **Maintenance:** Active updates and community (1 point)

Technical Implementation Strategy

Phase 1: Foundation

- Set up data pipeline with Kaggle Fashion Images
- Implement data augmentation strategies
- Establish baseline model performance metrics
- Validate commercial licensing compliance

Phase 2: Enhancement

- Integrate additional commercially-licensed datasets
- Implement transfer learning from research models (if applicable)
- Fine-tune on production data
- A/B test model performance

Phase 3: Optimization

- Incorporate user-generated content pipeline
- Implement continuous learning mechanisms
- Synthetic data generation for edge cases
- Performance monitoring and optimization

Phase 4: Scale

- Proprietary dataset expansion
- Advanced model architectures
- Multi-modal learning (text, image, metadata)
- Real-time adaptation to fashion trends

Risk Assessment & Mitigation

Risk 1: Dataset License Violations

Impact: High (Legal, Financial) **Mitigation:** - Legal review of all dataset licenses - Maintain license compliance documentation - Regular audits of data sources

Risk 2: Model Performance Degradation During Dataset Migration

Impact: Medium (Technical, User Experience) **Mitigation:** - Transfer learning strategies - Gradual migration with A/B testing - Performance benchmarking at each stage

Risk 3: Insufficient Data Diversity

Impact: Medium (Model Accuracy, User Satisfaction) **Mitigation:** - Multi-dataset approach - Synthetic data generation - User-generated content integration

Risk 4: Data Quality Issues

Impact: Medium (Model Accuracy) **Mitigation:** - Automated quality assurance pipelines - Manual review sampling - Continuous monitoring and cleanup

---

Budget Implications

Dataset Acquisition & Licensing

- Open-source datasets: $0
- Commercial dataset licenses: $5,000 - $50,000 (estimated, varies by vendor)

Data Engineering

- Pipeline development: 160-240 hours
-

Quality assurance: 80-120 hours

Conclusion

The Kaggle Fashion Images dataset represents the optimal balance between commercial viability, data quality, and implementation effort for CuratorAI's initial deployment. By supplementing with additional commercially-licensed datasets and implementing a robust transfer learning strategy, we can achieve production-grade performance while maintaining full legal compliance.

The multi-dataset approach, combined with continuous learning mechanisms, positions CuratorAI for long-term success and scalability in the competitive fashion AI market.

Appendix A: Dataset License Summary

Kaggle Fashion Product Images

- **License:** CC0 1.0 Universal (Public Domain)
- **Commercial Use:** Permitted
- **Attribution Required:** No
- **Modifications Allowed:** Yes

OpenImages (Fashion Subset)

- **License:** CC BY 4.0
- **Commercial Use:** Permitted
- **Attribution Required:** Yes
- **Modifications Allowed:** Yes

Fashion MNIST

- **License:** MIT License
- **Commercial Use:** Permitted

- **Attribution Required:** Yes (minimal)
- **Modifications Allowed:** Yes

DeepFashion

- **License:** Academic/Research Use
- **Commercial Use:** Requires special permission
- **Attribution Required:** Yes
- **Modifications Allowed:** Research purposes only

Appendix B: Recommended Tools & Frameworks

Data Pipeline

- **Apache Airflow** - Workflow orchestration
- **DVC (Data Version Control)** - Dataset versioning
- **Great Expectations** - Data quality validation

Model Training

- **PyTorch/TensorFlow** - Deep learning frameworks
- **Hugging Face Transformers** - Pre-trained models
- **Weights & Biases** - Experiment tracking

Deployment

- **Docker/Kubernetes** - Containerization
- **TensorFlow Serving/TorchServe** - Model serving
- **AWS SageMaker/GCP Vertex AI** - MLOps platform

Prepared by: Samuel Ssekizinvu **For:** CuratorAI Development Team **Document Version:** 1.0 **Last Updated:** October 6, 2025