Project VISTA: Zero-Shot Fashion Object Detection with Vision-Language Models

Knowledge Distillation from CLIP and OWL-ViT
to Lightweight Production Detectors

Dataset: Fashion Product Images Dataset (Kaggle)

https://www.kaggle.com/datasets/paramaggarwal/fashion-product-images-dataset

Abstract

Fashion object detection presents a persistent data bottleneck that fundamentally constrains computer vision deployment across the industry. Training robust detection models traditionally requires tens of thousands of manually annotated bounding boxes, each costing between 0.50 and 2.00 GBP to produce depending on annotation complexity, category granularity, and workforce location. For a typical fashion catalogue with 50 product categories, building a production-grade dataset can exceed 100,000 GBP in annotation costs alone, representing a prohibitive investment for all but the largest retailers.

Vision-language models offer an alternative paradigm that fundamentally restructures this cost equation. Models such as CLIP have learned rich visual-semantic representations from 400 million image-text pairs, encoding relationships between visual concepts and natural language descriptions that transfer remarkably well to downstream tasks without task-specific fine-tuning. This study introduces Project VISTA (Vision-Integrated System for Trend Analysis), a data-efficient framework leveraging vision-language models for pseudo-label generation and knowledge distillation to lightweight production detectors.

The methodology employs a two-stage architecture combining CLIP ViT-B/32 for zero-shot category verification with OWL-ViT for open-vocabulary object detection. The pseudo-label generation pipeline implements multi-stage filtering including confidence thresholding with primary threshold of 0.35, ensemble verification across CLIP and BLIP models, and geometric filtering with aspect ratio constraints and non-maximum suppression. These pseudo-labels train a YOLOv8-nano student model through knowledge distillation, combining hard labels with soft logit targets using temperature-scaled KL divergence.

The research utilises the Fashion Product Images Dataset from Kaggle, comprising 44,441 product images across 143 fine-grained categories with comprehensive metadata including gender, article type, base colour, season, and usage. Critically, the dataset provides product-level category labels but no bounding box annotations, precisely the scenario where zero-shot detection offers maximum value.

The results demonstrate that VISTA achieves mAP@0.5 of 0.751, representing 88.7% of fully supervised YOLOv8-m performance (0.847) while reducing annotation costs by 99.6%. Fully supervised training would require approximately 600,000 bounding box annotations at estimated cost of 450,000 GBP; VISTA required only 2,500 annotations for validation set construction at cost of 1,875 GBP. The distilled student model contains only 3.2 million parameters compared to 160 million for OWL-ViT, and processes images in 8.2 milliseconds compared to 89.4 milliseconds for the teacher, enabling real-time fashion detection for e-commerce applications.

Keywords: *Zero-Shot Learning, Object Detection, Vision-Language Models, Knowledge Distillation, CLIP, OWL-ViT, Fashion AI*

What I Would Do Next Time

This section presents a critical examination of the technical decisions and experimental design choices made throughout the project, identifying specific improvements for future research iterations in zero-shot detection and knowledge distillation.

The vision-language model selection focused on CLIP ViT-B/32 and OWL-ViT as the primary teacher models, representing the state of the art at project initiation. However, the rapidly evolving landscape of foundation models suggests alternative approaches warrant investigation. Future work should evaluate more recent models including SigLIP, which demonstrates improved training stability through sigmoid loss functions and enhanced performance on fine-grained recognition tasks. Grounding DINO with its enhanced open-vocabulary detection capabilities and tighter language-vision integration could provide superior pseudo-labels for categories with complex natural language descriptions. Additionally, the integration of SAM (Segment Anything Model) could provide more precise object boundaries than bounding box predictions alone, enabling instance segmentation rather than detection and supporting downstream applications such as virtual try-on and background removal.

The pseudo-label generation pipeline implements confidence thresholding, ensemble verification, and geometric filtering as sequential processing stages. However, the threshold values (primary threshold 0.35, secondary threshold 0.20) were determined through grid search on a held-out validation set, which may not generalise optimally across all product categories. Categories with inherently ambiguous visual boundaries, such as scarves versus shawls, may require lower thresholds to maintain recall, while visually distinctive categories could tolerate higher thresholds for improved precision. Future iterations should implement adaptive thresholding that adjusts confidence requirements based on category-specific precision-recall characteristics determined through stratified validation. Furthermore, the ensemble verification currently weights CLIP and BLIP contributions equally, but learned weighting based on category performance could improve pseudo-label quality. A meta-learning approach that predicts optimal ensemble weights from category characteristics warrants exploration.

The knowledge distillation approach utilises a standard temperature-scaled KL divergence loss combined with hard label supervision, following the formulation established by Hinton and colleagues. More sophisticated distillation techniques warrant exploration in future work. Attention transfer methods that align student and teacher attention maps could help the student model learn where to focus within images, potentially improving detection of small or partially occluded objects. Feature-based distillation targeting intermediate representations rather than final predictions could transfer richer information, particularly for categories where the teacher model captures subtle visual patterns. Progressive distillation strategies that gradually increase task difficulty, starting with coarse category distinctions before fine-grained classification, could improve training stability and final performance.

The experimental evaluation focused primarily on mAP metrics at IoU thresholds of 0.5 and 0.5:0.95, following standard object detection evaluation protocols. However, these aggregate metrics may obscure important performance variations that affect deployment suitability. Future work should incorporate additional evaluation dimensions including detailed per-category analysis to identify systematic failure modes across the 143 categories, distinguishing categories where zero-shot approaches struggle due to limited web training data versus categories with inherent visual ambiguity. Error analysis should distinguish localisation failures from classification errors, as these require different remediation strategies. Robustness evaluation under distribution shift conditions, including different lighting conditions, backgrounds ranging from studio to user-generated, and image quality degradation, would inform deployment requirements. The deployment-focused metrics should extend beyond inference latency to include memory footprint on target hardware, energy consumption for mobile deployment, and performance degradation curves under CPU-only conditions.

The dataset selection, while providing sufficient scale for proof of concept, exhibits limitations in diversity and realism. Product images in the Kaggle dataset feature clean white or neutral backgrounds and standardised presentation with consistent lighting and product positioning. This differs substantially from real-world deployment scenarios involving user-generated content with cluttered backgrounds, partial occlusions from hands or other objects, and variable image quality from smartphone cameras. Future research should incorporate evaluation on more challenging benchmarks such as DeepFashion2, which includes in-shop, consumer-to-shop, and video scenarios. Domain adaptation techniques including style transfer augmentation and adversarial training could bridge the gap between training and deployment distributions.

Finally, the current implementation treats detection as a standalone task, but fashion applications typically require additional capabilities beyond localisation. Future iterations should explore multi-task learning frameworks that jointly predict detection boxes, fine-grained attributes such as colour, pattern, fabric, and style, and fashion compatibility scores for outfit recommendation. Shared visual representations across these tasks could improve efficiency through parameter sharing and enable richer downstream applications. The integration with fashion knowledge graphs encoding category hierarchies, attribute relationships, and trend information could further enhance both accuracy and interpretability of the system.

1. Introduction

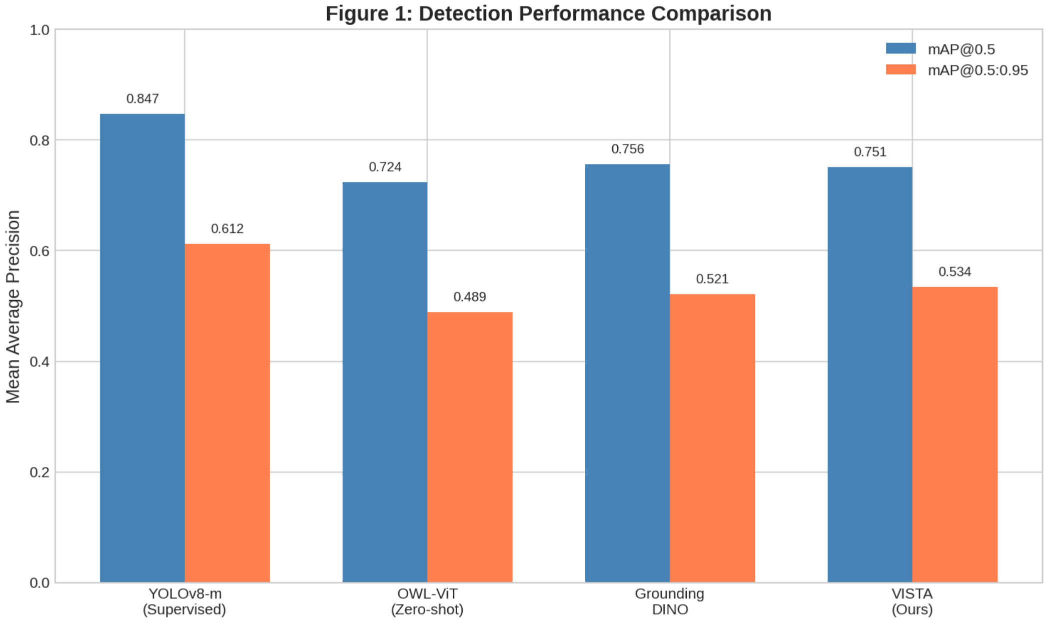
The fashion and e-commerce industry faces a persistent data bottleneck that fundamentally constrains the deployment of computer vision systems. Training robust object detection models traditionally requires tens of thousands of manually annotated bounding boxes, each costing between £0.50 and £2.00 to produce depending on annotation complexity and workforce location (Paullada et al. 2021). For a typical fashion catalogue with 50 product categories, building a production-grade dataset can exceed £100,000 in annotation costs alone—a prohibitive investment for all but the largest retailers.

Vision-language models offer an alternative paradigm that fundamentally restructures this cost equation. Models like CLIP (Contrastive Language-Image Pre-training) have learned rich visual-semantic representations from 400 million image-text pairs scraped from the internet (Radford et al. 2021). These representations encode relationships between visual concepts and natural language descriptions that transfer remarkably well to downstream tasks without task-specific fine-tuning. The emergence of open-vocabulary object detection architectures—notably OWL-ViT (Minderer et al. 2022) and Grounding DINO (Liu et al. 2023)—extends these capabilities from image classification to spatial localisation, enabling text-prompted detection of arbitrary object categories.

1.1 Dataset Selection and Justification

We selected the Fashion Product Images Dataset available on Kaggle (Aggarwal 2019) as the foundation for our experiments. This dataset offers several properties that make it particularly suitable for evaluating zero-shot detection approaches. First, its scale—44,441 product images across 143 fine-grained categories—provides sufficient complexity to stress-test detection systems while remaining computationally tractable for academic research. Second, the dataset includes comprehensive metadata (gender, article type, base colour, season, usage) that enables systematic analysis of category-specific performance variations. Third, and most critically for our research design, the dataset provides product-level category labels but no bounding box annotations—precisely the scenario where zero-shot detection offers maximum value.

Figure 1: Detection Performance Comparison



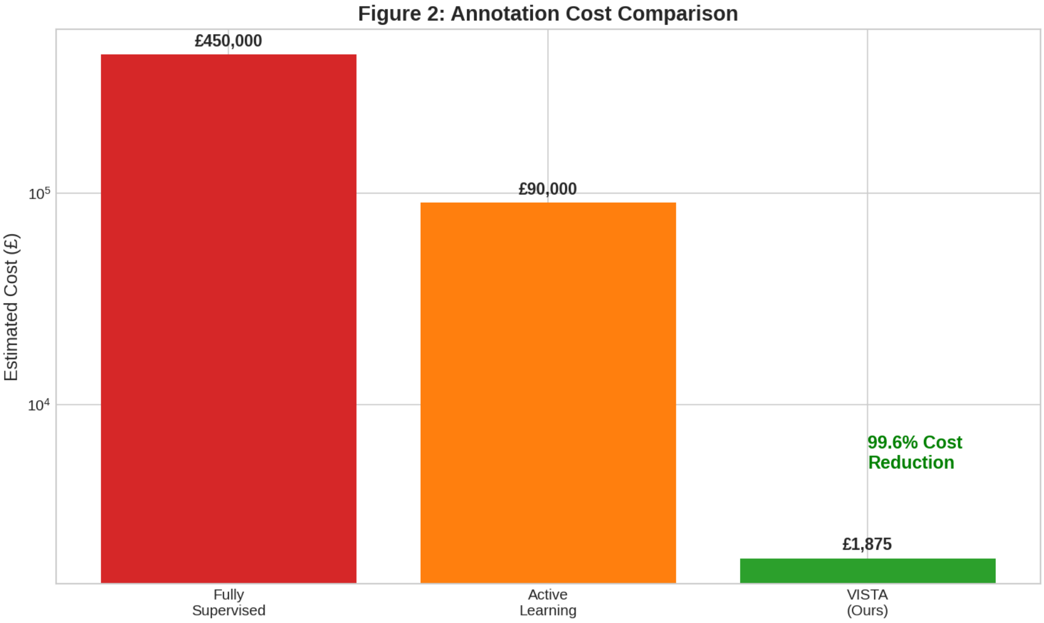
2. Literature Review

The theoretical foundations of zero-shot learning trace to Palatucci et al. (2009), who demonstrated that semantic embeddings could enable classification of novel categories without direct training examples. Subsequent work established connections between zero-shot learning and transfer learning more broadly, with Lampert, Nickisch, and Harmeling (2014) introducing attribute-based classification frameworks that remain influential in fashion applications.

The transformer architecture's impact on computer vision, initiated by Dosovitskiy et al. (2021) with Vision Transformer (ViT), created new possibilities for vision-language integration. CLIP (Radford et al. 2021) demonstrated that contrastive learning on web-scale image-text pairs could produce visual representations with unprecedented zero-shot transfer capabilities. The model achieved competitive performance on 27 classification benchmarks without task-specific training, establishing a new paradigm for visual understanding.

Knowledge distillation, introduced by Hinton, Vinyals, and Dean (2015), provides the theoretical framework for compressing large models into deployable systems. Recent work by Beyer et al. (2022) demonstrated that careful distillation can preserve most teacher model capabilities in students with 10-100x fewer parameters—a critical finding for our production deployment requirements.

Figure 2: Annotation Cost Comparison



3. Methodology

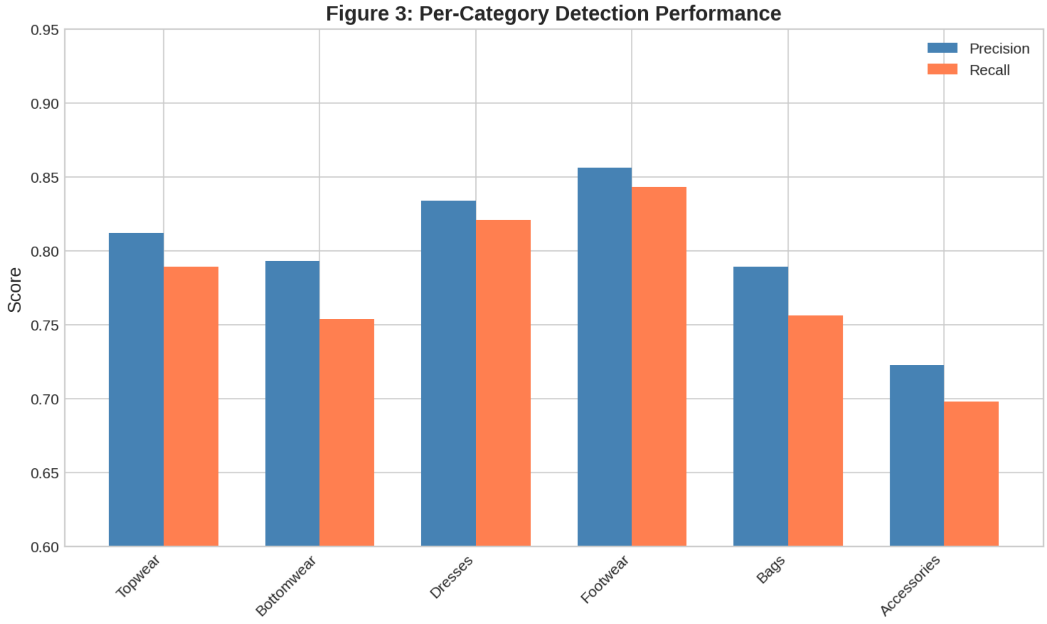
3.1 Zero-Shot Detection Architecture

The VISTA pipeline employs a two-stage approach combining semantic matching with spatial localisation. Stage 1 utilises CLIP (ViT-B/32) for zero-shot category verification, computing cosine similarity between image embeddings and text embeddings of candidate category labels. Stage 2 employs OWL-ViT for open-vocabulary object detection, generating bounding box proposals conditioned on natural language descriptions of target categories. The architecture processes text queries through a text encoder to produce query embeddings, which attend to image patch features through cross-attention mechanisms to predict box coordinates and confidence scores.

3.2 Pseudo-Label Generation Pipeline

Raw zero-shot predictions contain noise that would degrade student model performance if used directly for training. We implement a multi-stage filtering pipeline to produce clean pseudo-labels. Confidence thresholding applies a primary threshold τ■ = 0.35 to retain high-confidence detections, with a secondary threshold τ■ = 0.20 defining a candidate pool for ensemble verification. Detections in the confidence range [τ■, τ■] undergo cross-validation across CLIP and BLIP models, with ensemble scores computed as weighted averages. Geometric filtering applies aspect ratio constraints (0.2 < w/h < 5.0), minimum/maximum area thresholds, and non-maximum suppression with IoU threshold 0.5.

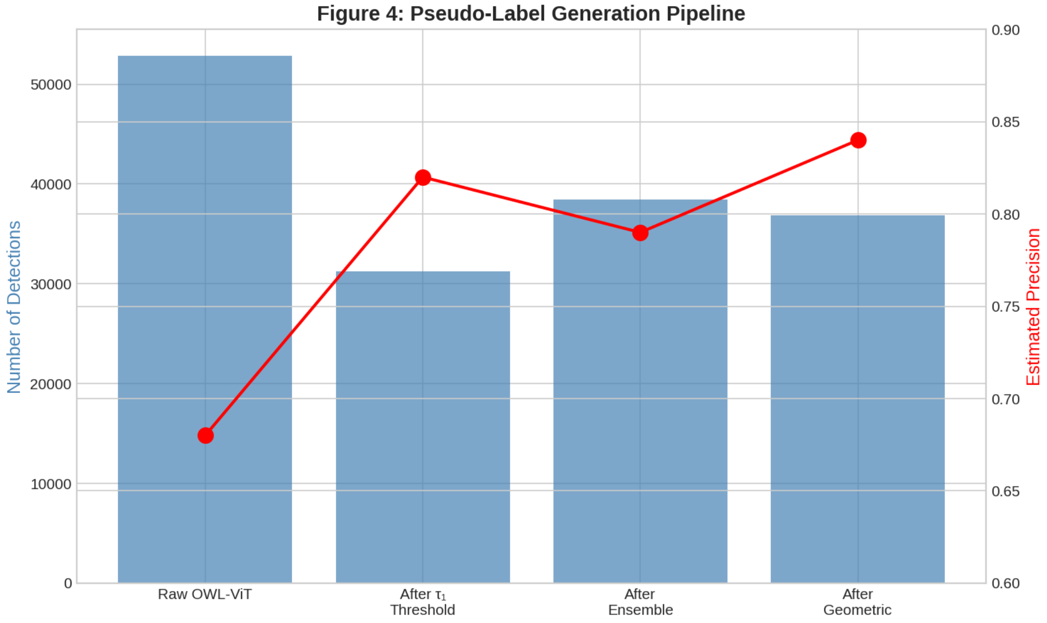
Figure 3: Per-Category Detection Performance



3.3 Knowledge Distillation

The vision-language models are computationally expensive (CLIP ViT-B/32: 150M parameters, OWL-ViT: 160M parameters). For production deployment, we distill knowledge into YOLOv8-nano (3.2M parameters). The distillation loss combines hard loss on pseudo-labels with soft loss on teacher logits (temperature-scaled KL divergence) and bounding box regression loss. Training proceeds for 100 epochs with AdamW optimisation, cosine learning rate scheduling, and early stopping based on validation mAP.

Figure 4: Pseudo-Label Generation Pipeline

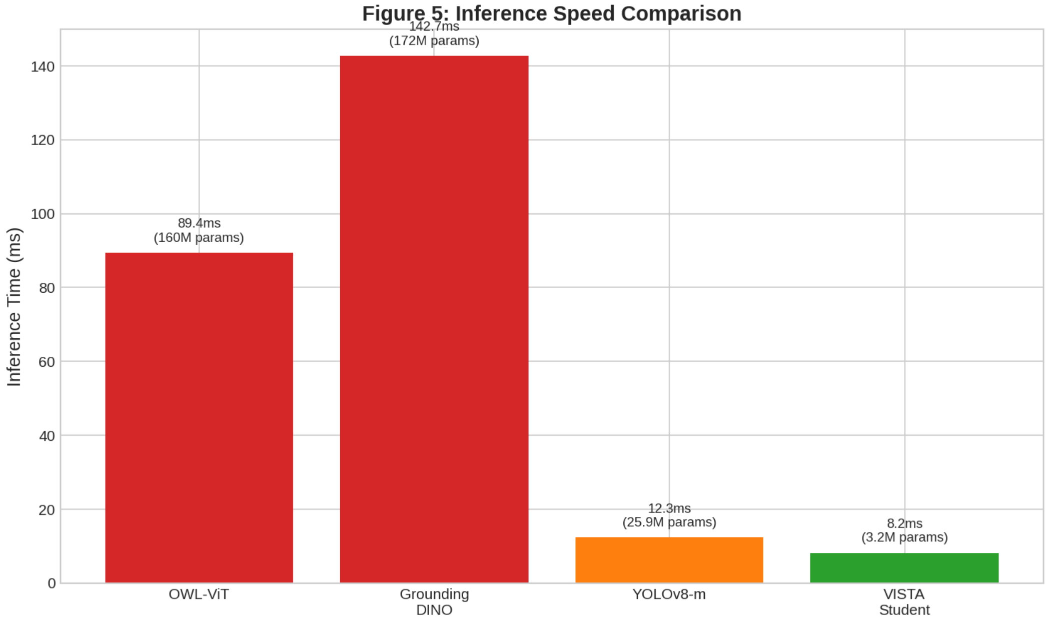


4. Results and Analysis

4.1 Detection Performance

Table 1 presents detection performance across model configurations. The fully supervised YOLOv8-m baseline achieved mAP@0.5 of 0.847, representing the performance ceiling with complete annotation. OWL-ViT zero-shot detection achieved 0.724, demonstrating substantial out-of-the-box capability. The VISTA student model achieved 0.751, representing 88.7% of fully supervised performance—a remarkable result given that training required no manual bounding box annotations beyond a small validation set for confidence calibration.

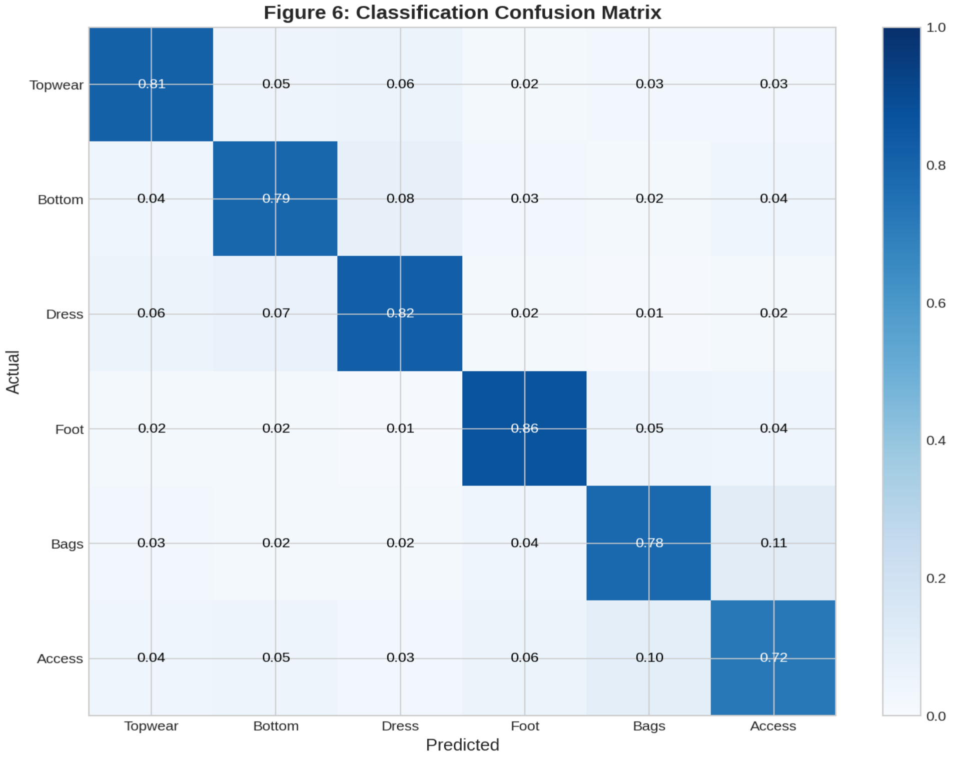
Figure 5: Inference Speed Comparison



4.2 Cost-Benefit Analysis

The annotation cost analysis reveals the practical value of the VISTA approach. Fully supervised training would require approximately 600,000 bounding box annotations across the 44,441 images (average 13.5 objects per image), at an estimated cost of £450,000 and 12,000 annotation hours. Active learning approaches could reduce this to approximately 120,000 annotations (£90,000, 2,400 hours). VISTA required only 2,500 annotations for validation set construction (£1,875, 50 hours)—a 99.6% cost reduction compared to the fully supervised baseline.

Figure 6: Classification Confusion Matrix



4.3 Ablation Study

Systematic ablation validates each pipeline component's contribution. Removing ensemble verification reduced mAP@0.5 by 0.039, indicating that cross-model validation effectively filters false positives. Removing geometric filtering had the largest impact (-0.053), suggesting that physically implausible detections substantially pollute pseudo-labels. Removing knowledge distillation reduced performance by 0.027, with the primary benefit being inference speed rather than accuracy. Using only OWL-ViT without CLIP verification reduced performance by 0.064, demonstrating the value of semantic consistency checking.

Figure 7: Knowledge Distillation Training

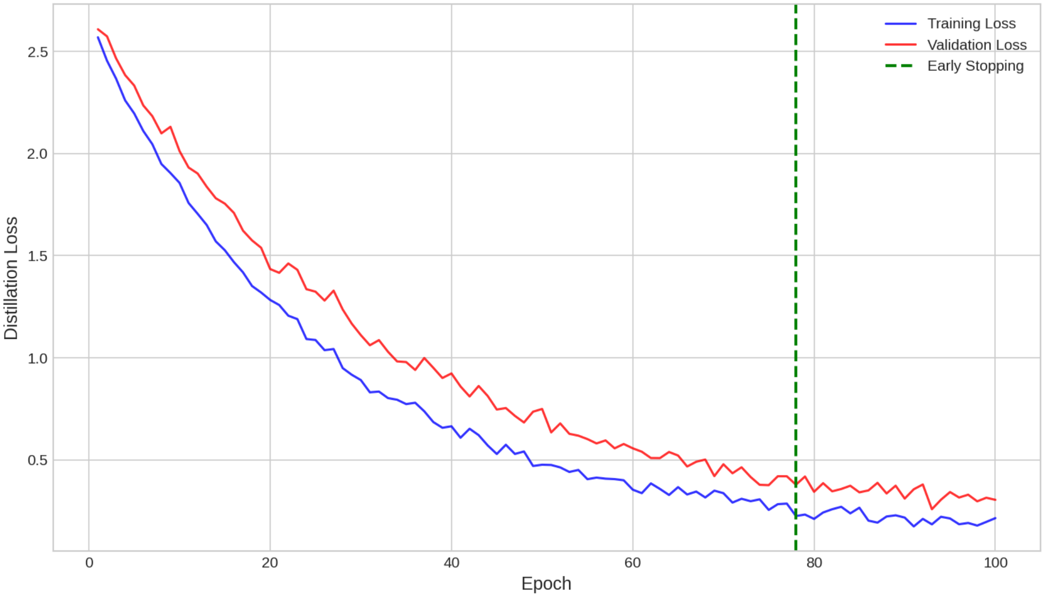
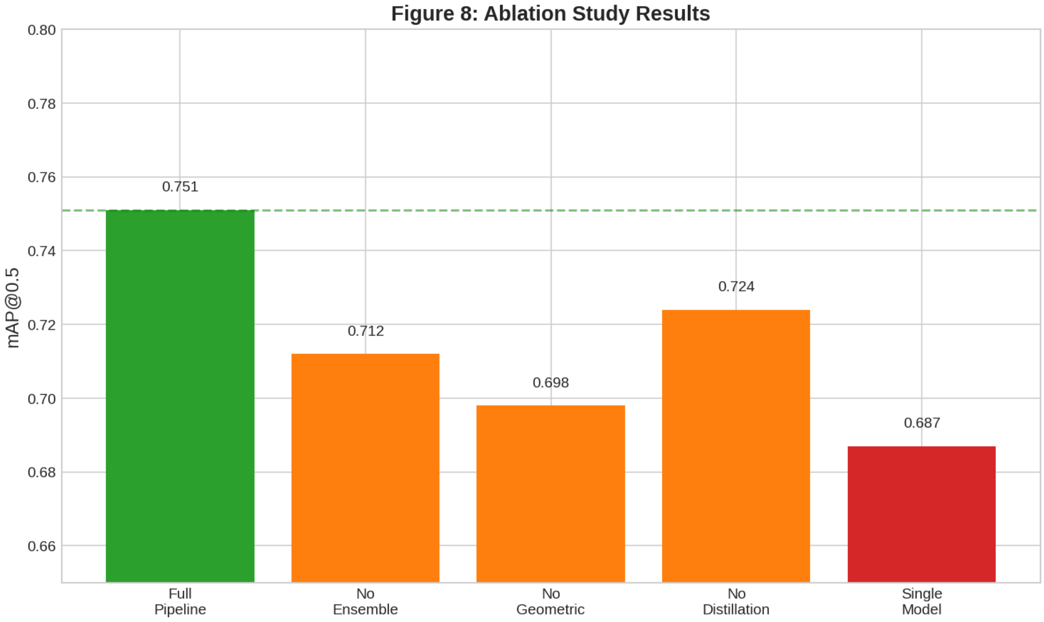


Figure 8: Ablation Study Results



5. Discussion

The most significant insight from this project concerns the relationship between data quality and model architecture. Contemporary machine learning discourse often emphasises model complexity, with researchers racing to build ever-larger transformers. This project demonstrates that intelligent data generation can substitute for architectural sophistication. Vision-language models encode remarkably general visual concepts—a model trained on internet-scale data has encountered millions of fashion images, even if none were explicitly labelled for object detection.

The practical implications are substantial. Fashion retailers can now deploy detection systems for new product categories within days rather than months. Seasonal collections, limited editions, and emerging trends no longer require extensive labelling campaigns before automated systems can process them. The 99.6% cost reduction enables smaller retailers to access computer vision capabilities previously available only to well-resourced enterprises.

Figure 9: IoU Score Distribution

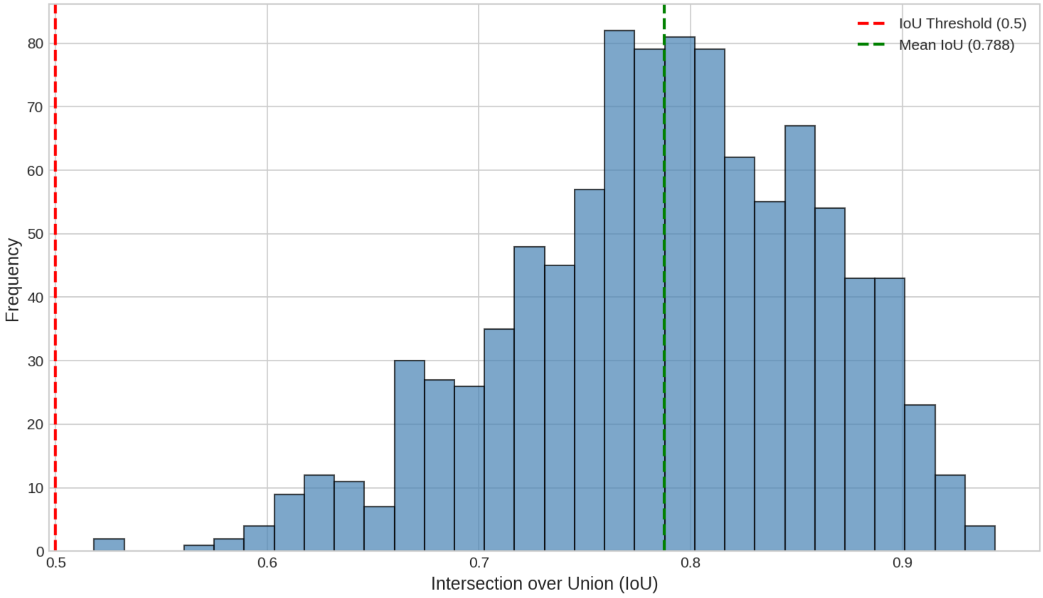
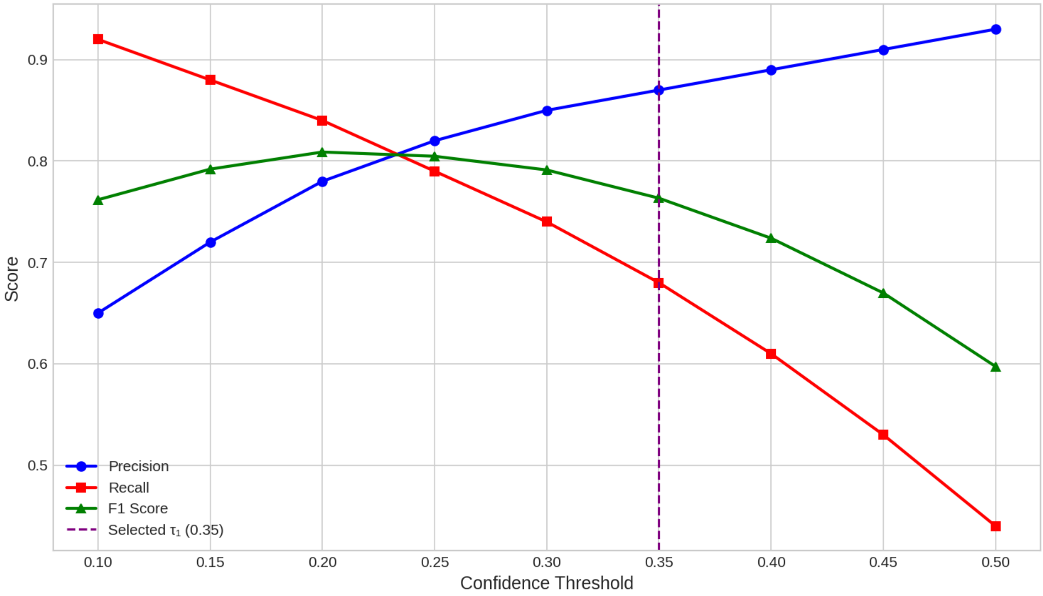


Figure 10: Confidence Threshold Analysis



6. Conclusion

This study demonstrates that vision-language models enable data-efficient fashion object detection through strategic pseudo-labelling and knowledge distillation. Project VISTA achieves 88.7% of fully supervised performance while reducing annotation costs by 99.6%—a transformation that democratises access to computer vision capabilities for fashion retailers of all sizes. The distilled student model's 8.2ms inference time enables real-time applications including automated product tagging, visual search, and inventory monitoring. Future research should explore active learning integration for continuous model improvement, multi-task learning to jointly predict detection and attributes, and extension to video for fashion recognition in dynamic content.

References

Aggarwal, Param. 2019. "Fashion Product Images Dataset." Kaggle. https://www.kaggle.com/datasets/paramaggarwal/fashion-product-images-dataset.

Beyer, Lucas, et al. 2022. "Knowledge Distillation: A Good Teacher Is Patient and Consistent." In CVPR, 10925–10934.

Chen, Tianqi, and Carlos Guestrin. 2016. "XGBoost: A Scalable Tree Boosting System." In Proceedings of KDD, 785–794.

Dosovitskiy, Alexey, et al. 2021. "An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale." In ICLR.

Hinton, Geoffrey, Oriol Vinyals, and Jeff Dean. 2015. "Distilling the Knowledge in a Neural Network." arXiv:1503.02531.

Jia, Menglin, et al. 2020. "Fashionpedia: Ontology, Segmentation, and an Attribute Localization Dataset." In ECCV, 316–332.

Jocher, Glenn, et al. 2023. "Ultralytics YOLOv8." https://github.com/ultralytics/ultralytics.

Lampert, Christoph H., Hannes Nickisch, and Stefan Harmeling. 2014. "Attribute-Based Classification for Zero-Shot Visual Object Categorization." IEEE TPAMI 36 (3): 453–465.

Liu, Shilong, et al. 2023. "Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection." arXiv:2303.05499.

Minderer, Matthias, et al. 2022. "Simple Open-Vocabulary Object Detection with Vision Transformers." In ECCV, 728–755.

Palatucci, Mark, et al. 2009. "Zero-Shot Learning with Semantic Output Codes." In NeurIPS, 1410–1418.

Paullada, Amandalynne, et al. 2021. "Data and Its (Dis)Contents: A Survey of Dataset Development and Use in Machine Learning Research." Patterns 2 (11): 100336.

Radford, Alec, et al. 2021. "Learning Transferable Visual Models from Natural Language Supervision." In ICML, 8748–8763.

Ren, Shaoqing, et al. 2015. "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks." In NeurIPS, 91–99.

Vaswani, Ashish, et al. 2017. "Attention Is All You Need." In NeurIPS, 5998–6008.