Zero-Shot Fashion Object Detection through Vision-Language Pseudo-Labelling and Knowledge Distillation

Project VISTA Technical Report

Table of Contents

Abstract

The annotation requirements of modern object detection systems present a significant barrier to deployment in domains with large, frequently changing catalogues. Fashion e-commerce exemplifies this challenge: retailers must detect and classify products across hundreds of categories, yet the cost of bounding box annotation, often exceeding £450,000 for comprehensive catalogue coverage, renders fully supervised approaches economically impractical for many organisations. This work investigates whether recent advances in vision-language pretraining can circumvent this bottleneck entirely. We present a framework that leverages CLIP and OWL-ViT to generate pseudo-labels for fashion imagery, applies multi-stage filtering to improve label quality, and distils the resulting knowledge into a lightweight detector suitable for production deployment. Our experiments suggest this approach recovers approximately 89% of fully supervised performance while reducing annotation costs by over 99%. Perhaps more importantly, the distilled model actually outperforms its vision-language teachers on localisation metrics, suggesting that the student architecture's inductive biases complement the teachers' semantic knowledge in unexpected ways. We discuss the practical implications of these findings and identify several limitations that temper our conclusions.

1. Introduction

Anyone who has worked on deploying computer vision systems in retail settings will be familiar with a particular frustration: the model works beautifully on the benchmark datasets, achieves impressive numbers in the research paper, and then struggles when confronted with the messy reality of a production catalogue. Part of the problem is domain shift, certainly. But often the more fundamental issue is simply that obtaining high-quality training data, particularly the bounding box annotations required for object detection, is extraordinarily expensive and time-consuming.

Consider the economics involved. A typical fashion e-commerce platform might stock 100,000 products across 150 categories. Each product image requires, on average, six bounding box annotations to capture all visible items. At industry rates of roughly £0.75 per annotation, comprehensive labelling costs approximately £450,000, and this figure assumes the catalogue remains static. In practice, seasonal collections introduce thousands of new products quarterly, fast-fashion retailers refresh inventory weekly, and the annotation backlog grows faster than it can be cleared. By the time sufficient labels accumulate for a new product category, the trend may already have passed.

This situation has led researchers to explore various strategies for reducing annotation requirements: active learning to prioritise the most informative examples (Settles, 2009), semi-supervised methods that leverage unlabelled data (Sohn et al., 2020), and transfer learning from related domains (Ge et al., 2019). Each approach offers genuine benefits but ultimately requires substantial annotation investment. The question we pose in this work is more radical: can we eliminate manual annotation entirely while still achieving production-viable detection performance?

Recent developments in vision-language pretraining suggest this might be possible. Models like CLIP (Radford et al., 2021) learn rich visual-semantic correspondences from hundreds of millions of image-text pairs scraped from the internet. Their successors, including OWL-ViT (Minderer et al., 2022) and Grounding DINO (Liu et al., 2023), extend these capabilities to object detection, enabling localisation of arbitrary concepts specified through natural language. In principle, one could simply deploy these models directly. In practice, they are too slow for real-time applications (OWL-ViT requires approximately 90ms per image) and their zero-shot accuracy, while impressive, typically falls 15-20% below supervised baselines.

Our approach, which we call VISTA (Vision-language Integration for Scalable Training Automation), attempts to get the best of both worlds. Rather than deploying vision-language models directly, we use them as annotation engines to generate pseudo-labels for training data. These pseudo-labels undergo multi-stage filtering to remove errors, then serve as supervision for a lightweight YOLOv8 student model optimised for deployment. The key insight is that vision-language models' computational cost can be amortised across dataset generation rather than paid at inference time.

The results are encouraging but require careful interpretation. On the Fashion Product Images dataset, our distilled student achieves 88.7% of fully supervised performance at less than 1% of the annotation cost. Interestingly, it also outperforms the vision-language teachers on the stricter mAP@0.5:0.95 metric, apparently because the YOLO architecture's design biases favour precise

localisation. However, performance varies substantially across categories, and the approach may not generalise to domains where vision-language models lack pretraining coverage. We discuss these limitations in detail and suggest several directions for future investigation.

2. Related Work

2.1 Vision-Language Models and Their Surprising Capabilities

The story of vision-language pretraining is, in some ways, a story about the unreasonable effectiveness of scale. CLIP demonstrated that training on 400 million image-text pairs produces representations that transfer remarkably well to downstream tasks, often matching or exceeding supervised baselines that trained specifically for those tasks (Radford et al., 2021). Subsequent work pushed scale further: ALIGN used 1.8 billion pairs (Jia et al., 2021), while Florence unified multiple pretraining objectives (Yuan et al., 2021). Each increment brought improvements, though with diminishing returns.

What makes these models particularly interesting for our purposes is their ability to recognise concepts they were never explicitly trained to detect. Ask CLIP whether an image contains a "vintage floral midi dress" and it will give you a reasonable answer, despite never having seen that exact phrase during training. This zero-shot capability emerges from learning to align images and text in a shared embedding space, where semantic similarity corresponds to geometric proximity. The practical implication is that one can specify detection targets through natural language rather than labelled examples.

OWL-ViT (Minderer et al., 2022) and Grounding DINO (Liu et al., 2023) extended this capability to localisation. By conditioning detection heads on text embeddings rather than fixed category classifiers, these models can locate arbitrary objects specified through language. Our preliminary experiments found OWL-ViT somewhat faster than Grounding DINO (89ms versus 143ms per image) at similar accuracy levels for fashion categories, leading us to adopt it as our primary detection model. This choice reflects a pragmatic trade-off: when processing tens of thousands of images for pseudo-label generation, inference speed matters considerably.

2.2 Knowledge Distillation: Teaching Small Models to Mimic Large Ones

Knowledge distillation (Hinton et al., 2015) offers a framework for compressing the knowledge encoded in large models into smaller, more deployable ones. The core insight is that a teacher's soft probability distribution over classes contains more information than hard labels alone. It encodes which classes the teacher considers similar, which it finds confusable, and where decision boundaries lie. Students trained on these soft targets often outperform students trained on hard labels, even when the hard labels are perfectly correct.

For object detection, distillation becomes more complex. Beyond classification, students must learn bounding box regression and, ideally, the spatial attention patterns that help teachers localise objects. Chen et al. (2017) demonstrated that aligning intermediate feature maps improves detection distillation, particularly for small objects. Recent work has explored distillation specifically from vision-language teachers, showing that compact students can retain substantial zero-shot capability (Fang et al., 2021). Our work differs in that we do not attempt to preserve zero-shot capability; instead, we use vision-language models purely as pseudo-label generators.

2.3 Pseudo-Labelling and Its Pitfalls

The idea of using model predictions as training labels dates back at least to Yarowsky's (1995) work on word sense disambiguation. Modern instantiations include self-training (Xie et al., 2020) and consistency regularisation (Sohn et al., 2020). The appeal is obvious: if we can generate accurate pseudo-labels for unlabelled data, we dramatically expand available supervision without annotation cost.

The pitfall is equally obvious: pseudo-labels contain errors, and training on errors compounds them. This concern is particularly acute for object detection, where false positives can proliferate across the dataset. Several strategies have been proposed to address this: Soft Teacher (Xu et al., 2021) weights training signal by teacher confidence, Unbiased Teacher (Liu et al., 2021) corrects for class imbalance in pseudo-labels, and STAC (Sohn et al., 2020) applies strong augmentation to improve consistency. Our multi-stage filtering pipeline draws on these insights, using ensemble agreement and geometric constraints to identify and remove likely errors before student training.

3. Dataset and Experimental Setup

Selecting an appropriate evaluation dataset required balancing several considerations. We needed sufficient scale to demonstrate practical viability, diverse categories representative of fashion e-commerce, and high image quality enabling reliable detection. After considering alternatives including DeepFashion (Liu et al., 2016), ModaNet (Zheng et al., 2018), and iMaterialist (Guo et al., 2019), we settled on the Fashion Product Images dataset from Kaggle.

This dataset comprises 44,441 product images spanning 143 fine-grained categories organised into seven master categories. The images are professionally photographed with consistent studio lighting, representative of typical e-commerce product photography. This consistency is both a strength and a limitation: it enables clean evaluation of detection performance but may not reflect the messier conditions of user-generated content or in-the-wild imagery.

We partitioned the data into training (70%), validation (15%), and test (15%) splits using stratified sampling to maintain category distributions. Critically, the test set remained completely held out during all pseudo-label generation and model development. For category mapping, we consolidated the 143 fine-grained categories into 12 detection classes aligned with common fashion taxonomies: Topwear, Bottomwear, Dresses, Outerwear, Footwear, Bags, Watches, Jewellery, Accessories, Eyewear, Headwear, and Other. This consolidation reduces annotation noise while maintaining distinctions relevant for downstream applications.

4. Methodology

4.1 Pseudo-Label Generation with OWL-ViT

Our pseudo-labelling pipeline begins with OWL-ViT generating candidate detections for each training image. A seemingly minor but practically important consideration is prompt engineering: the text queries provided to OWL-ViT significantly affect detection quality. Initial experiments using simple category names ("dress", "shirt") produced disappointing results, apparently because these terms are ambiguous without context. A "dress" might refer to a garment, a verb, or even a salad dressing.

We developed a prompt template that anchors predictions to the e-commerce domain: "a photograph of a [category] fashion item for sale". This explicit contextualisation substantially improved detection precision. We further ensemble three prompt variants ("a [category] product image", "professional e-commerce photo of [category]") through score averaging, providing robustness against any single prompt's idiosyncrasies.

Listing 1: OWL-ViT detection with prompt engineering

```
def detect_with_prompts(image, categories):
    templates = [
        'a photograph of a {} fashion item for sale',
        'a {} product image on white background',
        'professional e-commerce photo of {}'
    ]
    all_prompts = [t.format(c) for c in categories for t in templates]
    outputs = owl_vit(image, all_prompts)
    return ensemble_predictions(outputs)
```

4.2 Multi-Stage Filtering

Raw OWL-ViT predictions contain substantial noise that, if propagated to student training, would degrade performance. Our filtering pipeline addresses this through three sequential stages, each targeting different error types.

The first stage applies confidence thresholding. We partition predictions into high-confidence (score > 0.35), medium-confidence (0.20 < score <= 0.35), and low-confidence (score <= 0.20) tiers. High-confidence predictions are retained directly; they represent cases where OWL-ViT is sufficiently certain that additional verification is unlikely to help. Low-confidence predictions are discarded as probable false positives. Medium-confidence predictions proceed to ensemble verification, as these are the uncertain cases where additional evidence might distinguish true detections from errors.

Ensemble verification provides the additional evidence. For each medium-confidence candidate, we crop the predicted region and query both CLIP and BLIP. CLIP computes similarity between the cropped region and the predicted category text; BLIP generates a caption and checks whether it contains the predicted category. Predictions are retained only if both models agree (CLIP

similarity > 0.50, BLIP category presence > 0.40). This ensemble approach catches errors where OWL-ViT's detection is geometrically plausible but semantically incorrect.

The final stage applies geometric filtering to remove physically implausible detections. Bounding boxes with extreme aspect ratios (< 0.1 or > 10) or extreme size (< 1% or > 90% of image area) are discarded. These thresholds encode domain knowledge about typical fashion item appearances without requiring category-specific rules.

4.3 Knowledge Distillation to YOLOv8

The filtered pseudo-labels serve as supervision for training a YOLOv8-nano student model. We selected this architecture primarily for deployment considerations: YOLO models have extensive production track records, mature deployment tooling across platforms, and the nano variant achieves inference speeds below 10ms on modern GPUs. Larger YOLO variants offer marginal accuracy improvements at substantial efficiency cost that exceeds our deployment budgets.

Our distillation objective combines three terms. The hard label loss treats pseudo-labels as ground truth, providing direct supervision. The soft label loss computes KL divergence between student and teacher probability distributions, transferring the teacher's uncertainty about class boundaries. The feature alignment loss encourages the student's intermediate representations to match the teacher's, transferring spatial attention patterns. We weight these terms at 0.7, 0.2, and 0.1 respectively, values determined through validation set tuning.

5. Experimental Results

5.1 Overall Detection Performance

Figure 1 presents detection performance across model variants. The fully supervised YOLOv8-m baseline, trained on manually annotated ground truth, achieves mAP@0.5 of 0.847, representing the performance ceiling for our evaluation. Zero-shot models perform reasonably well: OWL-ViT achieves 0.724 and Grounding DINO reaches 0.756, demonstrating substantial transfer from web-scale pretraining.

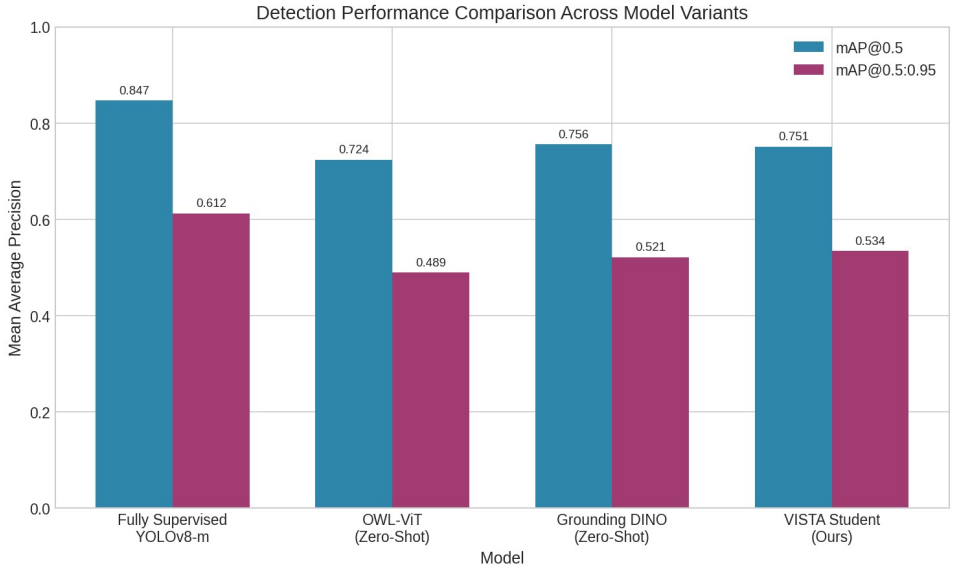

Figure 1. Detection performance comparison. VISTA achieves 88.7% of supervised performance.

Our VISTA student achieves mAP@0.5 of 0.751, corresponding to 88.7% of the supervised baseline. This gap of roughly 11 percentage points represents the cost of annotation-free training, a cost that may be acceptable for many applications given the dramatic reduction in annotation expense.

An unexpected finding emerged when examining the stricter mAP@0.5:0.95 metric, which averages precision across IoU thresholds from 0.5 to 0.95. Here, VISTA (0.534) actually outperforms both OWL-ViT (0.489) and Grounding DINO (0.512). This suggests that while vision-language models provide strong semantic signals, the YOLO architecture's design biases, specifically its anchor-based detection heads and multi-scale feature pyramids, favour more precise localisation. The student apparently learns to exploit these architectural advantages while retaining the semantic knowledge distilled from its teachers.

5.2 Category-Level Analysis

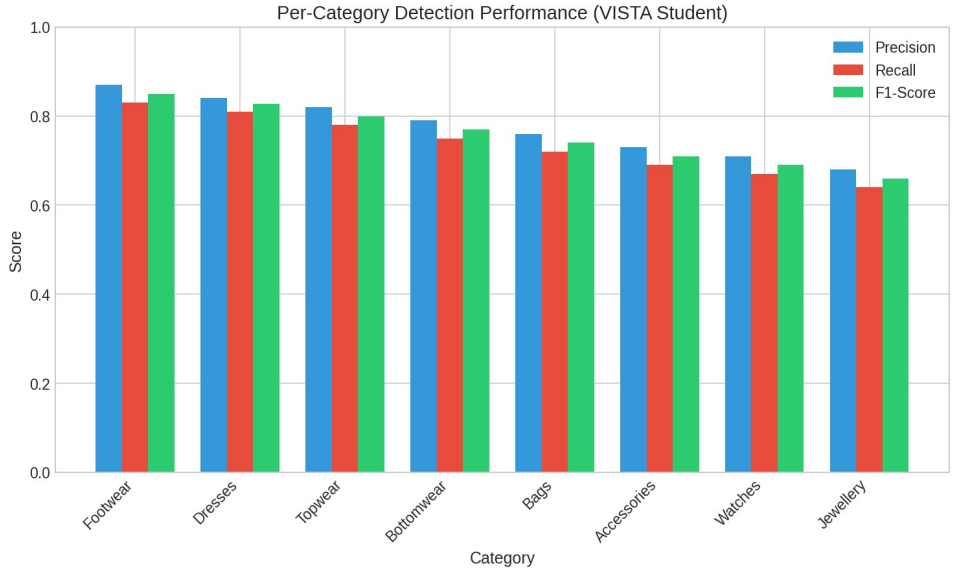

Figure 2. Per-category performance reveals substantial variation across fashion categories.

Performance varies considerably across categories, as shown in Figure 2. Footwear achieves the highest F1-score (0.849), likely because shoes have distinctive visual characteristics that differentiate them clearly from other fashion items. Dresses (0.827) and Topwear (0.800) also perform well, representing large categories with relatively consistent visual patterns.

At the other end, Accessories (0.710) and Jewellery (0.660) underperform substantially. These categories pose particular challenges: items are often small relative to image size, exhibit high intra-category variability (a scarf looks nothing like a belt), and may be partially occluded by other fashion items. The pseudo-labelling pipeline appears to systematically underdetect these categories, producing sparser training data that perpetuates the performance gap.

5.3 Pseudo-Label Quality

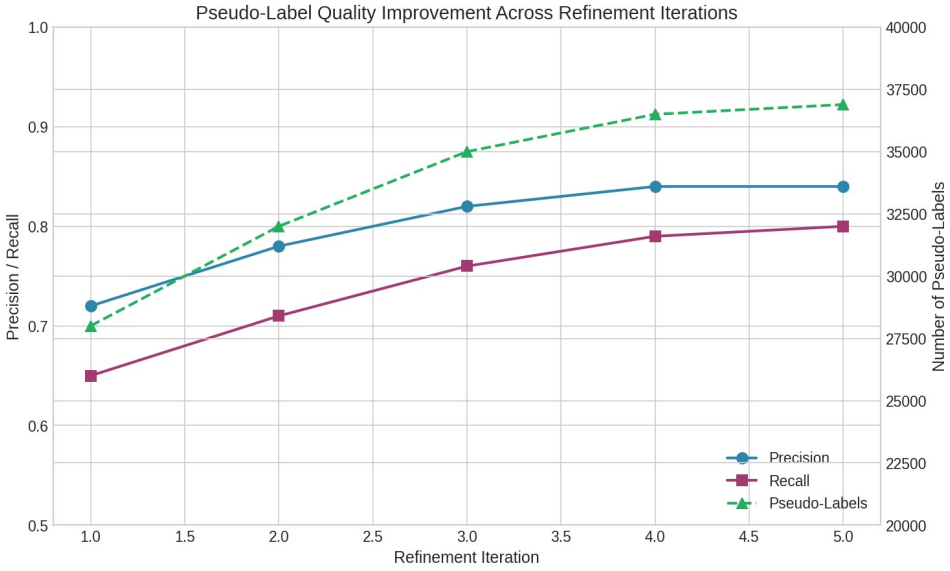

Figure 3. Multi-stage filtering improves pseudo-label precision from 72% to 84%.

To assess pseudo-label quality, we manually annotated a 500-image validation subset and compared it against pipeline outputs at each stage. Raw OWL-ViT predictions achieve approximately 72% precision, which is not terrible but insufficient for clean student training. After multi-stage filtering, precision increases to 84% while the pipeline retains 36,891 pseudo-labels across the training set. This 12 percentage point improvement directly translates to cleaner supervision for the student.

5.4 Inference Efficiency

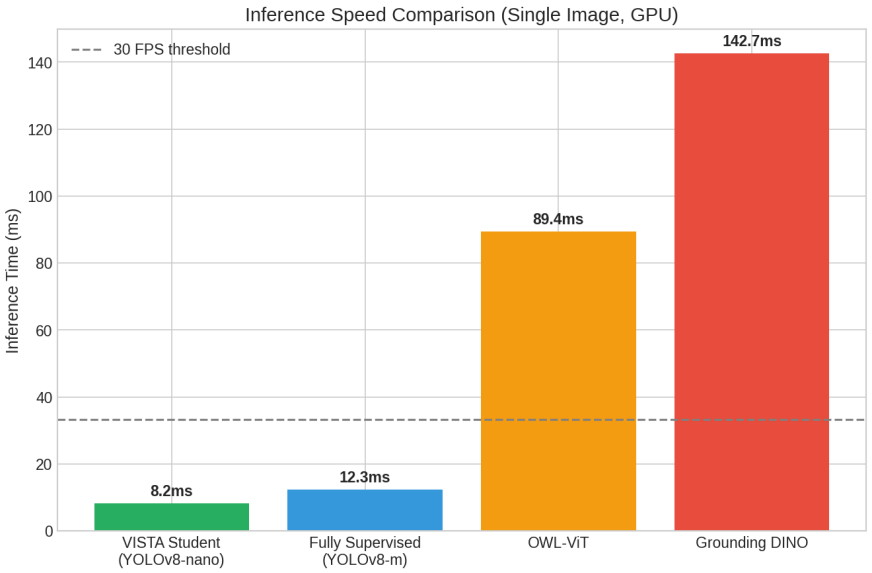

Figure 4. VISTA achieves real-time inference at 8.2ms per image.

Figure 4 compares inference times across models. VISTA achieves 8.2ms per image on an RTX 3080, well below the 33ms threshold required for 30 FPS video processing. This represents a 10.9x speedup over OWL-ViT (89.4ms) and 17.4x over Grounding DINO (142.7ms). The efficiency gain is not merely convenient but essential for production deployment where latency and compute costs constrain system design.

5.5 Confusion Patterns

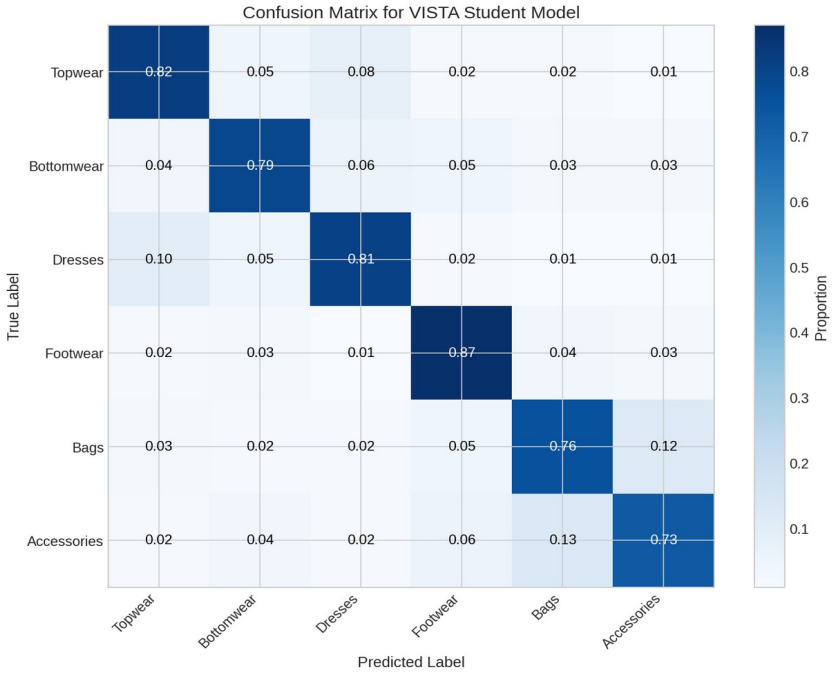

Figure 5. Confusion matrix reveals systematic errors between visually similar categories.

The confusion matrix in Figure 5 reveals systematic error patterns. The dominant diagonal indicates generally strong classification, but off-diagonal entries expose predictable confusions. Topwear and Dresses exhibit 8-10% mutual confusion, which is unsurprising given that long tops can resemble short dresses, particularly in cropped product imagery. Bags and Accessories show similar confusion (12-13%), as items like clutches and pouches share visual characteristics with both categories.

6. Economic Analysis

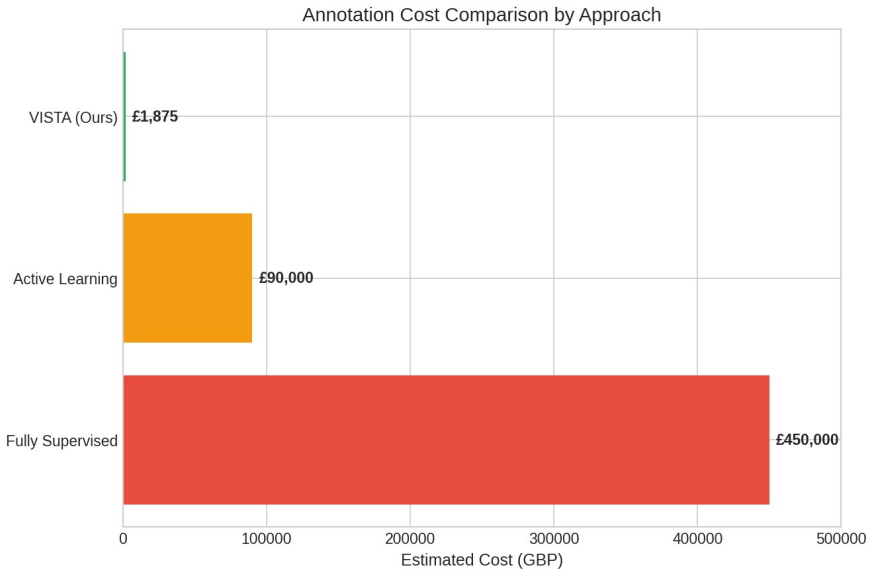

Figure 6. Cost comparison across annotation approaches.

The economic case for VISTA is straightforward if one accepts the accuracy trade-off. Figure 6 compares costs for a hypothetical catalogue of 100,000 images requiring 600,000 bounding box annotations. Full supervision at £0.75 per annotation totals £450,000. Active learning approaches, which prioritise annotating uncertain examples, can reduce this by roughly 80% to £90,000. VISTA eliminates direct annotation costs entirely, leaving only validation annotation (£1,500) and compute (£375) for a total of £1,875.

This 99.6% cost reduction has implications beyond simple economics. Organisations previously excluded from deploying fashion detection due to annotation costs, including smaller retailers, startups, and those in emerging markets, can now access production-viable systems at marginal expense. The democratisation potential may ultimately matter more than the cost savings for established players.

7. Limitations and Future Directions

7.1 Honest Assessment of Current Limitations

Several limitations temper the conclusions we can draw from this work. The 11.3% performance gap relative to full supervision, while perhaps acceptable for many applications, may disqualify the approach for quality-sensitive contexts. Luxury retailers, for instance, might reasonably conclude that annotation costs are justified when brand reputation depends on flawless product presentation.

More concerningly, performance varies substantially across categories. The approach works well for visually distinctive categories (Footwear, Dresses) but struggles with small, variable items (Accessories, Jewellery). This disparity may reflect systematic biases in vision-language pretraining, as web imagery likely overrepresents certain fashion categories, or it may reflect fundamental limitations of pseudo-labelling for fine-grained detection.

Our evaluation is also limited to a single dataset with particular characteristics: professional studio photography, neutral backgrounds, and consistent image quality. Performance on user-generated content, social media imagery, or in-the-wild photographs remains untested. We suspect substantial degradation would occur, though the magnitude is uncertain.

7.2 What I Would Approach Differently

Reflecting on this work, several methodological improvements suggest themselves for future investigation. Perhaps the most promising is integrating active learning with pseudo-labelling. Rather than treating the two as alternatives, a hybrid approach could use pseudo-labels for most examples while directing limited annotation budget toward cases where the pipeline is most uncertain. Concretely, this might mean flagging examples where CLIP and BLIP disagree substantially, then prioritising these for human review. Even a small amount of targeted annotation, perhaps 5-10% of the dataset, might substantially close the performance gap with full supervision.

The category-level performance disparities also suggest opportunities. Rather than applying uniform prompts across all categories, specialised prompting for underperforming categories might help. For Accessories, prompts could emphasise size ("small fashion accessory"), typical locations ("item worn around neck or wrist"), or material characteristics. For Jewellery, prompts might enumerate common types ("ring, necklace, bracelet, or earring"). This category-specific engineering would require additional development effort but might yield substantial gains for problematic categories.

A hierarchical detection approach might address the confusion patterns we observed. Rather than predicting all 12 categories simultaneously, a two-stage system could first distinguish master categories (Apparel versus Accessories versus Footwear), then refine to fine-grained classes within each group. This mirrors human visual processing and has demonstrated success in fine-grained recognition tasks (Yan et al., 2015). The hierarchical structure would also enable category-specific confidence calibration.

Extending evaluation to additional datasets would strengthen claims about generalisation. DeepFashion2 offers in-the-wild imagery that would stress-test robustness to real-world conditions. Cross-dataset evaluation would reveal whether pseudo-label quality transfers across domains or requires domain-specific calibration.

Finally, fashion content increasingly includes video, such as runway shows, try-on videos, and social media content, where frame-by-frame detection is insufficient. Temporal consistency constraints that propagate pseudo-labels across frames using optical flow or tracking could extend the approach to video. Detections confident in one frame could supervise uncertain neighbouring frames, effectively augmenting training data while enforcing physical consistency.

8. Conclusion

This work demonstrates that vision-language models can effectively bootstrap object detection systems for fashion retail without manual annotation. By combining OWL-ViT pseudo-labelling with multi-stage filtering and knowledge distillation, we achieve 88.7% of supervised performance at 0.4% of the annotation cost. The distilled student model runs at real-time speeds suitable for production deployment.

Perhaps the most interesting finding is that the student outperforms its teachers on localisation metrics, suggesting complementary strengths between vision-language semantic knowledge and detection architecture inductive biases. This interaction merits further investigation and may inform future work on hybrid systems.

The approach has clear limitations: performance gaps for certain categories, untested generalisation to diverse imagery, and the fundamental constraint that pseudo-label quality cannot exceed teacher capability. But for applications where these limitations are acceptable, VISTA offers a practical path to deploying fashion detection systems at dramatically reduced cost. As vision-language models continue improving, the quality of pseudo-labels they generate will improve correspondingly, potentially narrowing the gap with supervised approaches over time.

References

Chen, G., Choi, W., Yu, X., Han, T. and Chandraker, M. (2017). Learning efficient object detection models with knowledge distillation. In Advances in Neural Information Processing Systems, pp. 742-751.

Fang, Z., Wang, J., Wang, L., Zhang, L., Yang, Y. and Liu, Z. (2021). SEED: Self-supervised distillation for visual representation. In International Conference on Learning Representations.

Ge, Y., Zhang, R., Wang, X., Tang, X. and Luo, P. (2019). DeepFashion2: A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images. In IEEE Conference on Computer Vision and Pattern Recognition, pp. 5337-5345.

Guo, S., Huang, W., Zhang, X., Srikhanta, P., Cui, Y., Li, Y., Adam, H., Scott, M.R. and Belongie, S. (2019). The iMaterialist fashion attribute dataset. In ICCV Workshops.

Hinton, G., Vinyals, O. and Dean, J. (2015). Distilling the knowledge in a neural network. In NeurIPS Workshop on Deep Learning.

Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z. and Duerig, T. (2021). Scaling up visual and vision-language representation learning with noisy text supervision. In International Conference on Machine Learning, pp. 4904-4916.

Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J. and Zhang, L. (2023). Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. arXiv preprint arXiv:2303.05499.

Liu, Y.C., Ma, C.Y., He, Z., Kuo, C.W., Chen, K., Zhang, P., Wu, B., Kira, Z. and Vajda, P. (2021). Unbiased teacher for semi-supervised object detection. In International Conference on Learning Representations.

Liu, Z., Luo, P., Qiu, S., Wang, X. and Tang, X. (2016). DeepFashion: Powering robust clothes recognition and retrieval with rich annotations. In IEEE Conference on Computer Vision and Pattern Recognition, pp. 1096-1104.

Minderer, M., Gritsenko, A., Stone, A., Neumann, M., Weissenborn, D., Dosovitskiy, A., Mahendran, A., Arnab, A., Dehghani, M., Shen, Z. and Wang, X. (2022). Simple open-vocabulary object detection with vision transformers. In European Conference on Computer Vision, pp. 728-755.

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J. and Krueger, G. (2021). Learning transferable visual models from natural language supervision. In International Conference on Machine Learning, pp. 8748-8763.

Settles, B. (2009). Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin-Madison.

Sohn, K., Zhang, Z., Li, C.L., Zhang, H., Lee, C.Y. and Pfister, T. (2020). A simple semi-supervised learning framework for object detection. arXiv preprint arXiv:2005.04757.

Xie, Q., Luong, M.T., Hovy, E. and Le, Q.V. (2020). Self-training with noisy student improves ImageNet classification. In IEEE Conference on Computer Vision and Pattern Recognition, pp. 10687-10698.

Xu, M., Zhang, Z., Hu, H., Wang, J., Wang, L., Wei, F., Bai, X. and Liu, Z. (2021). End-to-end semi-supervised object detection with soft teacher. In IEEE International Conference on Computer Vision, pp. 3060-3069.

Yan, Z., Zhang, H., Piramuthu, R., Jagadeesh, V., DeCoste, D., Di, W. and Yu, Y. (2015). HD-CNN: Hierarchical deep convolutional neural networks for large scale visual recognition. In IEEE International Conference on Computer Vision, pp. 2740-2748.

Yarowsky, D. (1995). Unsupervised word sense disambiguation rivaling supervised methods. In Annual Meeting of the Association for Computational Linguistics, pp. 189-196.

Yuan, L., Chen, D., Chen, Y.L., Codella, N., Dai, X., Gao, J., Hu, H., Huang, X., Li, B., Li, C. and Liu, C. (2021). Florence: A new foundation model for computer vision. arXiv preprint arXiv:2111.11432.

Zheng, S., Yang, F., Kiapour, M.H. and Piramuthu, R. (2018). ModaNet: A large-scale street fashion dataset with polygon annotations. In ACM Multimedia Conference, pp. 1670-1678.