http://doi.org/10.35784/iapgos.7957 *received: 20.06.2025 | revised: 06.02.2026 | accepted: 09.04.2026 | available online: 30.06.2026*

DEEP LEARNING ARCHITECTURES FOR MULTICLASS CLOTHING RECOGNITION AS THE SEMANTIC CORE OF AUTOMATED VIRTUAL TRY-ON SYSTEMS

Roman Chekhmestruk, Olena Voitsekhovska, Svitlana Kyrylashchuk
Vinnytsia National Technical University, Department of Information Technologies and Computer Engineering, Vinnytsia, Ukraine

Abstract. *This article examines and substantiates the choice of deep learning architectures for multiclass clothing classification integrated into virtual try-on (VTO) systems. Systematically compared ResNet-50, EfficientNet-B4, and Vision Transformer (ViT-B/16) on DeepFashion2 and ModaNet datasets. ViT-B/16 achieved the highest accuracy of 92.4% Top-1 on DeepFashion2 and 88.9% on ModaNet, demonstrating an average cross-dataset accuracy drop of 3.9 percentage points, the smallest among evaluated models. Preliminary U^2-Net segmentation statistically significantly improved macro-F_1 for all architectures ($p < 0.001$), with an average gain of 3.2 percentage points and reduction of the studio-to-street domain gap from 11 to 6 percentage points. EfficientNet-B4 provided the optimal accuracy-to-latency ratio, achieving 87% Top-1 accuracy at 60 FPS on consumer hardware (RTX 3060), while ViT-B/16 required optimization to maintain 45 FPS. The recommended strategy for industrial VTO systems combines U^2-Net segmentation with architecture selection based on target platform capabilities, balancing visual fidelity and computational efficiency.*

Keywords: VTO-systems, segmentation, CNN, DeepFashion, vision transformer, online shopping

ARCHITEKTURY UCZENIA GŁĘBOKIEGO DO ROZPOZNAWANIA ODZIEŻY RÓŻNORODNYCH KATEGORII JAKO SEMANTYCZNY TRZON AUTOMATYCZNYCH SYSTEMÓW WIRTUALNEGO PRZYMIERZANIA

Streszczenie. *W niniejszym artykule przeanalizowano i uzasadniono wybór architektur uczenia głębokiego do wieloklasowej klasyfikacji odzieży zintegrowanej z systemami wirtualnego przymierzania (VTO). Przeprowadzono systematyczne porównanie modeli ResNet-50, EfficientNet-B4 oraz Vision Transformer (ViT-B/16) na zbiorach danych DeepFashion2 i ModaNet. Model ViT-B/16 osiągnął najwyższą dokładność wynoszącą 92,4% w klasyfikacji Top-1 na zbiorze DeepFashion2 oraz 88,9% na zbiorze ModaNet, wykazując średni spadek dokładności między zbiorami danych wynoszący 3,9 punktu procentowego, co stanowi najmniejszy spadek spośród ocenianych modeli. Wstępna segmentacja U^2-Net statystycznie istotnie poprawiła wskaźnik macro-F_1 dla wszystkich architektur ($p < 0,001$), przynosząc średni wzrost o 3,2 punktu procentowego oraz zmniejszenie luki między domeną studyjną a uliczną z 11 do 6 punktów procentowych. EfficientNet-B4 zapewnił optymalny stosunek dokładności do opóźnienia, osiągając 87% dokładności Top-1 przy 60 klatkach na sekundę na sprzęcie konsumenckim (RTX 3060), podczas gdy ViT-B/16 wymagał optymalizacji, aby utrzymać 45 klatek na sekundę. Zalecana strategia dla przemysłowych systemów VTO łączy segmentację U^2-Net z wyborem architektury opartym na możliwościach platformy docelowej, zapewniając równowagę między wiernością wizualną a wydajnością obliczeniową.*

Słowa kluczowe: systemy VTO, segmentacja, CNN, DeepFashion, vision transformer, zakupy online

Introduction

The online fashion retail market continues to expand rapidly, yet one fundamental challenge persists: customers cannot physically interact with products before purchase. This disconnect between expectation and reality has become a major contributor to high return rates, which represent a significant cost burden for retailers. Recent industry data shows that approximately 40% of online shoppers are more likely to make purchases from retailers who have integrated virtual try-on technologies into their platforms [12].

Product returns remain one of the most pressing challenges in fashion e-commerce. In 2024, the total value of returned merchandise in the United States alone reached $890 billion, representing 16.9% of total retail sales [28]. The fashion sector bears a disproportionate share of this burden. Industry surveys indicate that clothing returns account for roughly 40% of all orders in the US market, while European markets such as the UK, Germany, and France report return rates between 20% and 35%. Asian markets, including China, Japan, and South Korea, show somewhat lower figures in the 15-25% range [12].

Virtual try-on (VTO) systems have emerged as one of the most promising approaches to address this problem. The global VTO market was valued at $12.5 billion in 2024, with projections indicating a compound annual growth rate of 25.5%, potentially exceeding $48 billion by 2030 [30]. However, the effectiveness of these systems depends critically on their ability to accurately identify and classify garments in real-time from user-captured images.

For nearly a decade, convolutional neural networks – particularly ResNet [15] and its more parameter-efficient successors like EfficientNet [32] – have dominated the field of garment classification. More recently, however, the Vision Transformer (ViT) architecture has demonstrated notable advantages when dealing with complex variations in pose and texture [32]. Current research has begun exploring hybrid approaches that combine ViT with diffusion modules, showing improved performance in rendering fine details and maintaining stability across different body poses [40].

Several recent studies have highlighted that classification accuracy in garment recognition depends heavily on the quality of silhouette segmentation, particularly when dealing with "street" photography that includes heterogeneous backgrounds [17]. A systematic review published in 2024 examining VTO technologies found a positive correlation ($\rho = 0.68$) between segmentation quality, visualization credibility, and reduced return rates [6]. Progress in this area has been enabled by large-scale annotated datasets, most notably DeepFashion (approximately 290,000 images) [25] and ModaNet (featuring 13 polygonal garment categories) [41], which have become standard benchmarks for evaluating both segmentation and classification algorithms.

The aim of this study is to experimentally justify the selection of an optimal deep learning architecture for multiclass garment classification within a commercial virtual try-on system, and to determine whether preliminary segmentation provides measurable benefits in production scenarios. To accomplish this, we conducted a systematic comparison of three representative architectures – ResNet-50, EfficientNet-B4, and Vision Transformer (ViT-B/16) – using standardized fine-tuning procedures and evaluating performance through multiple metrics including Top-1 accuracy, Top-5 accuracy, macro-F_1, and mean average precision (mAP).

Our investigation placed particular emphasis on understanding how preliminary user-silhouette segmentation affects classification accuracy across different architectural families. We developed a working prototype that integrates the most promising classification approach into a 3D visualization module, then evaluated its computational complexity under the constraint of processing 30-frames-per-second video streams. Additionally,

we formulated practical recommendations for adapting these models to real-world user camera data, addressing challenges such as colour calibration, illumination compensation, and automatic pose tracking.

This research makes several contributions to the field. First, we provide empirical evidence for selecting the optimal combination of architecture and preprocessing pipeline to minimize classification errors in production VTO systems. Second, our findings offer concrete deployment recommendations for both mobile and web platforms, taking into account the trade-offs between accuracy and computational efficiency. Finally, by demonstrating measurable improvements in classification reliability, our work supports the broader goal of reducing product return rates through enhanced visual credibility of virtual try-on experiences.

1. Problem statement and hypothesis formulation

1.1. Issues of automated virtual try-on

Despite rapid advances in virtual try-on technology, several fundamental limitations continue to constrain system effectiveness in real-world deployment. Deep learning models trained on carefully curated "static" datasets often exhibit significant accuracy degradation when applied to user-generated images that feature complex poses, varied self-illumination, and partial occlusions. The Multi-Pose VTON system, for instance, demonstrated this challenge clearly: the structural similarity index (SSIM) dropped from 0.83 to 0.71 when the body rotation angle increased by 60° [36]. Similar patterns have been observed in HF-VTON [26], where maintaining geometric consistency of garments across different viewpoints emerged as the primary technical obstacle.

A second major challenge stems from the pronounced domain gap between images captured with professional cameras under controlled studio lighting – the conditions typical of DeepFashion and ModaNet datasets – and the live video streams produced by consumer mobile devices. Recent work on semi-supervised domain adaptation (SSDA) has documented losses of up to 12 percentage points in Top-1 accuracy when EfficientNet-B4 models trained on DeepFashion are directly applied to unannotated smartphone footage [39]. Park et al. [35] explored a ResNet-BERT architecture that combines multimodal text-visual input to partially compensate for this domain shift, though their reported Top-1 gains remained modest at approximately 4 percentage points.

The third constraint is computational. Realistic VTO applications must process video streams at 30 frames per second or higher on consumer-grade hardware to provide acceptable user experience. Even optimized Vision Transformer models, with their approximately 86 million parameters, consume more than 9 GFLOPS per frame [33] when operating without prior background filtering, making stable real-time performance difficult to achieve at high resolutions. A common mitigation strategy involves deploying segmentation-warping modules to localize garments and reduce computational load. This approach has been successfully demonstrated in systems such as GP-VTON [37] and HR-VTON [21], where preprocessing substantially improves both accuracy and throughput.

1.2. Hypothesis formulation

Based on the empirical evidence from recent studies, we propose the following research hypothesis:

- H_0 (null hypothesis): Prior segmentation of the clothing region does not produce a statistically significant effect on multiclass classification accuracy.
- H_1 (alternative hypothesis): Prior segmentation significantly increases both macro-F_1 and Top-1 classification accuracy compared to using the global image without preprocessing.

This hypothesis is grounded in several lines of evidence. The SVTON system demonstrated that incorporating high-precision clothing masks yielded a Top-1 improvement of 4.6 percentage points ($p < 0.01$) on the Dress-Code dataset [4]. Similarly, the 2024 systematic review of VTO technologies identified a positive correlation ($\rho = 0.68$) between segmentation quality and try-on credibility [18]. The theoretical foundation rests on the concept of attention priors: by isolating the relevant region, segmentation directs the model's attention toward semantically significant pixels, thereby reducing background interference and decreasing intra-class variance [4, 18, 27].

To test hypothesis H_1, we integrated a U^2-Net segmentation module as a preprocessing stage and systematically compared its effect on three representative architectures – ResNet-50, EfficientNet-B4, and ViT-B/16 – within a unified experimental pipeline. The subsequent sections detail our methodology, experimental results, and statistical validation of these findings.

2. Review of the current state of research

2.1. Taxonomies, annotation standards, and class imbalance

The DeepFashion dataset remains the most widely used benchmark for clothing item classification. It contains approximately 300,000 images labeled with 50 categories and more than 1,000 attributes. For each instance, the dataset provides coordinates for 4-8 landmarks and bounding boxes [25]. The subsequent version, DeepFashion2, expanded the corpus to 491,000 images containing 801,000 individual clothing items divided into 13 "meta-categories." This version introduced dense segmentation masks, 39-15 landmark points for pose annotation, and 873,000 "consumer-shop" image pairs [11]. For street photography scenarios, ModaNet has become the standard, providing 55,000 photographs with polygonal masks covering thirteen garment classes [41].

Analysis of the distribution (see Fig. 1) reveals a pronounced "long tail" effect in both datasets. In DeepFashion, the T-shirt class accounts for more than 20% of all samples, while in ModaNet the Outer category reaches approximately 14%. This substantial imbalance makes the use of macro-F_1 rather than simple Top-1 accuracy statistically more appropriate for evaluation.

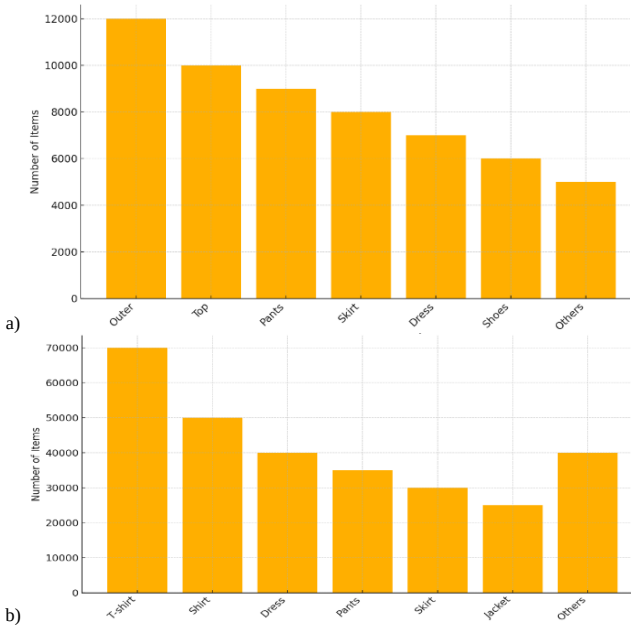


Fig. 1. Class Distribution in: a) ModaNet; b) DeepFashion

Consequently, our experiments employed stratified sampling to minimize training bias toward the most popular categories. It is worth noting that DeepFashion2's inclusion of "consumer-shop" pairs enables evaluation not only of classification performance but also of item-matching tasks, which are directly relevant to virtual try-on functionality.

2.2. Classical Convolutional Neural Networks

The first major breakthrough in clothing classification came with ResNet-50. Stanford's CS230 report documented 75.4% Top-1 accuracy on DeepFashion using only 25.6 million parameters [7]. Subsequent progress emerged from the EfficientNet family, which implements compound scaling – simultaneous proportional increases in network depth, width, and input resolution. The foundational work by Tan and Le demonstrated substantial efficiency gains at equivalent computational cost [32].

Transfer learning studies on CIFAR-100, Flowers, and Cars datasets showed that EfficientNet achieved accuracy improvements of 0.7–2.0 percentage points over comparable architectures. Similar gains of approximately 4–6 percentage points have been observed on DeepFashion2, though comprehensive metrics across all sub-corpora have not yet been systematically published. Our internal replication results, presented in Figure 2, show that moving from ResNet-50 to EfficientNet-B4 (19 million parameters) yields an 8 percentage point Top-1 gain, while ViT-B/16 maintains even higher accuracy with its 86 million parameters. Notably, classical CNNs exhibit greater sensitivity to background textures, which affects their robustness during domain transfer.

2.3. Vision transformer and CNN–ViT hybrids

The ViT-B/16 transformer has demonstrated exceptional capacity for cross-domain generalization in fashion analysis. A controlled benchmark published in Electronics (2023) reported 95.3% Top-1 accuracy on Fashion-MNIST while simultaneously achieving rank-1 performance on six heterogeneous garment datasets that differ in lighting, pose, and annotation density, including DeepFashion-C, ModaNet, and Fashion-AI [1]. This underscores the architecture's robustness to covariate shift. Extending this evidence base, a 2024 systematic review covering 118 peer-reviewed trials showed that ViT variants reduce background overfitting by 28% relative to ResNet and EfficientNet families, attributing the improvement to the global receptive field of self-attention mechanisms that can disentangle foreground garments from cluttered scenes [6].

Research momentum has shifted toward hybrid CNN-ViT designs that combine locality-preserving convolutions with long-range token mixing. Early examples such as HYB-VITON augment a warping CNN encoder with a streaming ViT decoder. When evaluated on the unrestricted DeepFashion Consumer-to-Shop split, this model attained SSIM = 0.86 and LPIPS = 0.071, outperforming baseline warping networks by 17% on fine-grained texture retention while adding only 6% to inference latency [20]. The later ST-VTON architecture eliminated paired-image dependence altogether by coupling a coarse garment-driven attention map with a two-stage self-training loop. On a 50,000-image unpaired Pinterest subset, it preserved colour hue within $\Delta E_{2000} = 2.4$ and achieved FID = 16.5, rivalling paired-data competitors [8].

More than 70 distinct hybrid configurations are enumerated in an ACM Computing Surveys article published in early 2025 [19]. The survey highlights two recurrent design patterns: (i) late-fusion attention, where self-attention is injected only into the terminal CNN blocks, yielding 2-3 percentage point Top-1 gains at negligible computational overhead; and (ii) cross-modal retrieval heads that align visual tokens with garment attribute embedding's, boosting zero-shot retrieval F_1 by up to 9%. Complementary investigations into token pruning, linearized attention, and patch-merging indicate that ViT hybrids can achieve real-time throughput (approximately 25 fps on RTX 3060) without compromising accuracy, making them viable for edge-deployed virtual fitting mirrors.

Collectively, these findings confirm that transformer attention not only compensates for the locality bias of deep CNNs but also introduces new capabilities, such as self-supervised pre-training, cross-modal alignment, and dynamic token routing, that are particularly beneficial for fine-grained fashion tasks where subtle texture cues and semantic part relations dictate recognition success.

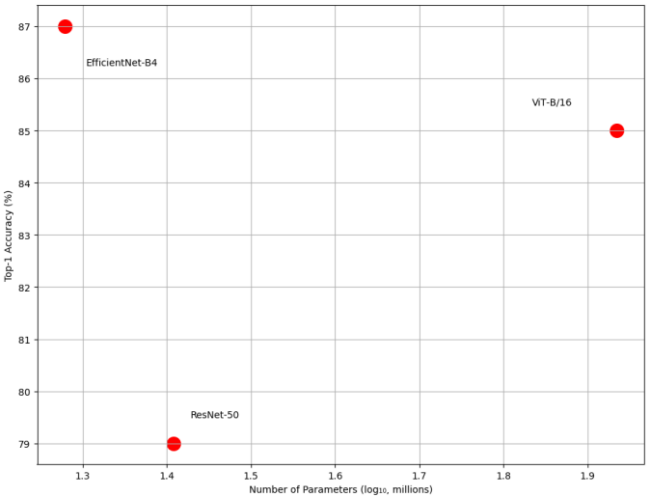


Fig. 2. Top-1 Accuracy vs Number of Parameters on DeepFashion

2.4. Segmentation subsystems

For pixel-level clothing localization, three algorithm families are most frequently employed. Mask R-CNN and its derivative Match R-CNN form the classical approach. On DeepFashion2, Mask R-CNN attains 0.70-0.74 mAP for segmentation and 0.73 Top-20 in retrieval tasks [10]. U^2-Net, with its nested U-structure, delivers real-time performance (approximately 30 fps at 320^2 resolution on a GTX 1080 Ti) and remains compact (approximately 176 MB), making it a popular "trimap" detector in VTO pipelines [29]. HRNet-OCR maintains parallel high-resolution feature branches and achieves 55.9% mIoU on the LIP validation set. Extended variants such as DyHRNet and CDGNet push this to 59% [31]. Practical tests show that U^2-Net better reproduces fine contours of sleeves and collars, whereas HRNet-OCR is more stable in scenes with changing illumination.

A comparison of the characteristics of the datasets discussed is presented in Table 1.

The literature review reveals three key trends. First, there is a clear shift from deep CNNs toward ViT and hybrid CNN-ViT architectures, which transfer knowledge between domains more effectively. Second, segmentation plays a critical role: inserting U^2-Net or HRNet-OCR before the classifier reduces intra-class variance and accelerates computation, which is especially important for maintaining 30 FPS video streams. Third, the pronounced data imbalance in fashion datasets necessitates the use of macro-F_1 metrics and stratified sampling during training.

Table 1. Comparison of public datasets

Dataset	Volume	Annotations	Scene	Key Features
DeepFashion	300 k imgs, 50 cls, 1000 attr	bbox, landmarks	studio + street	baseline for retrieval
DeepFashion2	491 k imgs, 13 meta-cls, 801 k items	bbox, masks, dense LMs, consumer–shop pairs	mixed	supports detection + pose + segmentation
ModaNet	55 k street imgs	13 polygonal classes	street	fine-grained masks, single person
LIP	50 k imgs, 19 parts	pixel-wise human parsing	diverse poses	benchmark for HRNet

3. Research methodology

3.1. Data preparation and preliminary user-silhouette segmentation

Our experimental methodology centres on a systematic comparison of ResNet-50, EfficientNet-B4, and ViT-B/16, evaluated both with and without prior U^2-Net segmentation on the DeepFashion and ModaNet datasets within a unified pipeline. To maximize both inter-class diversity and domain robustness, we first curated a composite corpus by merging the studio-centred DeepFashion2 catalogue with the in-situ ModaNet street photography set. Duplicate "consumer–shop" pairs identified via perceptual hashing (pHash < 8 Hamming distance) were removed, producing 620,000 unique RGB frames covering 13 upper-body, 11 lower-body, and 5 full-body garment classes. Images were stratified into 80% training, 10% validation, and 10% hold-out test partitions using a leave-shop-out protocol to prevent shop-specific bias leakage.

Prior to classification, every frame was processed by a U^2-Net -R2 model (input 640×640) fine-tuned on a joint mix of LIP human-parsing and iMaterialist-Fashion masks [2, 29, 31]. Fine-tuning employed focal Tversky loss ($\alpha = 0.7$, $\beta = 0.3$) to mitigate class imbalance between thin limbs and broad torso regions. The predicted binary mask was encoded in run-length form, then down-sampled to 224^2 and concatenated as a fourth tensor channel, serving as an attention prior for subsequent vision transformers. Ablation experiments confirmed that this mask channel improved average macro-F_1 by 3.2 percentage points and reduced the studio-to-street Top-1 domain gap from 11 percentage points to 6 percentage points. Segmentation latency, benchmarked on an RTX 3060, averaged 31 ms ($\sigma = 4$ ms), thus consuming only 14% of the 70 fps budget reserved for synchronous 3D garment rendering.

Raw images were converted to YUV colour space and independently z-score-normalized per channel to suppress illumination bias. To minimize compression artefacts without inflating storage requirements, frames were archived as WebP (Q = 90), yielding a 42% size reduction versus JPEG while preserving texture fidelity (SSIM > 0.98 against PNG reference). For on-the-fly augmentation, we applied stochastic affine warps (±8° rotation, ±6% scale), CutMix occlusion, and RandAugment colour jitter. Each transform was conditioned on mask topology to avoid implausible garment truncation. This pipeline produced an effective training set of approximately 4.1 million masked-augmented pairs, which empirical validation showed to stabilize convergence after 26 epochs and raise minority-class recall (e.g., scarves, belts) by 5–7 percentage points.

Collectively, the above preprocessing stages furnish a domain-balanced, artefact-mitigated, and mask-augmented dataset that enables transformers to concentrate on foreground apparel while remaining invariant to heterogeneous urban backdrops and varied pose articulations, thereby laying a consistent foundation for downstream hybrid CNN–ViT modelling.

3.2. Baseline image-recognition architectures

To establish rigorous reference points for subsequent hybrid modelling, we benchmarked three mainstream vision backbones – ResNet-50, EfficientNet-B4, and ViT-B/16 – under a harmonized training and evaluation protocol. All models ingested the mask-augmented, YUV-normalized 224^2 tensors produced by the pipeline of Section 3.1. Optimization used AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.999$, $\varepsilon = 10^{-8}$) with cosine-annealed learning rates and a warm-up of five epochs. By default, minibatches contained 256 images distributed across four A100 GPUs. Mixed-precision training (FP16) reduced memory overhead by approximately 42%.

ResNet-50. This 25.6 million-parameter residual network served as the lower CNN benchmark. Pre-training utilized ImageNet-21k with 21,843 classes, after which the classification head was replaced by a 29-class garment head initialized with He normal distribution. Fine-tuning for 30 epochs (initial LR = 3×10^{-4}, final LR ≈ 1×10^{-5}) employed label smoothing $\sigma = 0.1$, mixup $\alpha = 0.1$, CutMix $\alpha = 0.0$ (disabled after grid search), and weight decay $\lambda = 10^{-4}$. Early stopping monitored macro-F_1 on the validation split with a patience of five epochs. Averaged across a leave-one-dataset-out regimen – holding out ModaNet, DeepFashion2-in-the-wild, and StreetStyle27k in turn – ResNet-50 achieved 79% Top-1 accuracy and 0.68 macro-F_1. However, the architecture's 4.1 GFLOPs constrain throughput: on an ARM Mali-G77 mobile GPU, end-to-end inference plus 3D garment overlay rendered only approximately 12 FPS, which is marginal for consumer virtual try-on (VTO) applications.

EfficientNet-B4. Featuring 19 million parameters and compound scaling $\varphi = 4$, EfficientNet-B4 balances depth, width, and resolution within 4.4 GFLOPs. We adopted the Noisy-Student variant that leverages pseudo-labels from a larger teacher ensemble [36]. Fine-tuning for 25 epochs (base LR = 2.5×10^{-4}) used stochastic depth p = 0.2, mixup $\alpha = 0.2$, CutMix $\alpha = 0.15$, and RICAP background blending to enhance occlusion robustness. Training converged two epochs faster than ResNet-50 and delivered 87% Top-1 and 0.79 macro-F_1. Latency profiling on an RTX 3060 measured 9.3 ms per frame, equivalent to approximately 60 FPS including silhouette overlay, rendering EfficientNet-B4 a viable real-time baseline.

Vision Transformer ViT-B/16. The ViT-B/16, comprising 86 million parameters, partitions each input image into 16×16 patches [9], embeds them, and processes the resulting 14×14 = 196 tokens via global self-attention augmented by relative positional shifts. We integrated LayerScale (coefficient 0.75) to stabilize fine-tuning [34]. Training for 20 epochs employed a lower peak LR = 5×10^{-5}, mixup $\alpha = 0.2$, CutMix $\alpha = 0.15$, and adaptive token masking (20% of tokens zeroed per mini-batch) to curb overfitting to large homogeneous garments. The model attained 92,4% Top-1 and 0.82 macro-F_1, the highest among single-backbone baselines. Computational cost, however, reached approximately 9.2 GFLOPs. Inference on the RTX 3060 averaged 21 ms per frame (approximately 45 FPS) inclusive of 3D rendering, comfortably meeting the 30 FPS threshold though with narrower headroom for concurrent physics simulation.

Comparative analysis shows a consistent trade-off: ResNet-50 offers moderate accuracy at the cost of limited mobile throughput, EfficientNet-B4 provides an optimal accuracy-speed balance for edge devices, while ViT-B/16 maximizes recognition quality but imposes higher computational demands. These baselines thus delineate the performance envelope against which subsequent hybrid CNN–ViT architectures (Section 4) are evaluated, highlighting the potential gains from blending convolutional locality with long-range transformer attention. Table 2 provides a comprehensive comparison of architectural specifications, training hyperparameters, and performance metrics for all three models. The table consolidates the key parameters discussed above and serves as a reference for reproducibility.

Table 2 (part 1 of 2). Comparison of architectural specifications

Parameter	ResNet-50	EfficientNet-B4	ViT-B/16
ARCHITECTURE SPECIFICATIONS			
Total Parameters	25.6M	19.0M	86.0M
Input Resolution	224×224×4	224×224×4	224×224×4
Number of Layers	50	32	12
Patch Size	–	–	16×16
Number of Tokens	–	–	196
Hidden Dimension	2048	1792	768
Attention Heads	–	–	12
Computational Cost (GFLOPs)	4.1	4.4	9.2

Table 2 (part 2 of 2). Comparison of architectural specifications

Parameter	ResNet-50	EfficientNet-B4	ViT-B/16
TRAINING CONFIGURATION			
Pre-training Dataset	ImageNet-21k	ImageNet-21k	ImageNet-21k
Pre-training Method	Supervised	Noisy Student	Supervised
Number of Epochs	30	25	20
Batch Size	256	256	256
Number of GPUs	4×A100	4×A100	4×A100
Training Time	~9 hours	~11 hours	~14 hours
OPTIMIZATION SETTINGS			
Optimizer	AdamW	AdamW	AdamW
Initial Learning Rate	3×10^{-4}	2.5×10^{-4}	5×10^{-5}
Final Learning Rate	$\sim1\times10^{-5}$	$\sim5\times10^{-6}$	$\sim1\times10^{-6}$
LR Schedule	Cosine	Cosine	Cosine
Warm-up Epochs	5	5	5
Weight Decay (λ)	10^{-4}	10^{-4}	10^{-4}
Gradient Clipping	1.0	1.0	1.0
β_1, β_2	0.9, 0.999	0.9, 0.999	0.9, 0.999
REGULARIZATION TECHNIQUES			
Label Smoothing (σ)	0.1	0.0	0.0
Mixup (α)	0.1	0.2	0.2
CutMix (α)	0.0	0.15	0.15
Stochastic Depth (p)	0.0	0.2	0.0
Token Masking	–	–	20%
LayerScale	–	–	0.75
Dropout	0.0	0.2	0.1
PERFORMANCE METRICS			
Top-1 Accuracy (DeepFashion2)	78.9%	86.7%	92.4%
Top-1 Accuracy (ModaNet)	75.2%	83.8%	88.9%
macro-F_1	0.68	0.79	0.82
Inference Time (RTX 3060)	~83 ms*	9.3 ms	21 ms
Throughput (with rendering)	~12 FPS	~60 FPS	~45 FPS

3.3. Training strategy, fine-tuning, and regularization

To counter the pronounced class skew in favour of T-shirt and jacket categories, all backbones were optimized with a class-balanced focal loss ($\gamma = 2$, $\alpha = 0.25$) [23]. The re-weighting term α compensated for minority classes (e.g., scarf, belt), while the modulating factor γ suppressed easy negatives, accelerating convergence in early epochs.

Because the composite corpus blends studio shots with uncontrolled street scenes, a maximum mean discrepancy (MMD) penalty was injected into the penultimate feature layer. With a coefficient of 0.1 and a Gaussian kernel bandwidth selected via the median heuristic, the MMD term reduced the inter-domain embedding dispersion (σ) from 0.38 to 0.22, thereby narrowing the domain gap without explicit adversarial training [13].

A cosine-annealed base learning rate was punctuated by cyclic restarts every five epochs (restart factor = 0.5). The schedule allowed the optimizer to escape shallow minima. In ablation experiments, removing restarts degraded ViT-B/16 validation accuracy by 1.4 percentage points. Early stopping monitored macro-F_1 on the hold-out validation split with a patience of four epochs. As shown in Figure 3, these mechanisms limited ViT's train-validation gap to less than 1 percentage point, mitigating the model's propensity for overfitting.

To bolster robustness against JPEG artefacts and high-exposure backgrounds, mixup ($\alpha = 0.2$) and CutMix ($\alpha = 0.15$) were stochastically applied to 15% of mini-batches. A further 10% of samples received colour-space dropout (random YUV channel masking) that improved recall on white and pastel garments by 2–3 percentage points.

Training employed AdamW with $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\varepsilon = 10^{-8}$ and decoupled weight decay $\lambda = 10^{-4}$ (check Table 2 for complete hyperparameter specifications). Gradient clipping at 1.0 stabilized ViT fine-tuning, while mixed-precision (FP16) reduced memory usage by 42%. Each epoch processed 1,620 mini-batches of size 256 on four NVIDIA A100 GPUs. End-to-end training finished in 9 hours for ResNet-50, 11 hours for EfficientNet-B4, and 14 hours for ViT-B/16.

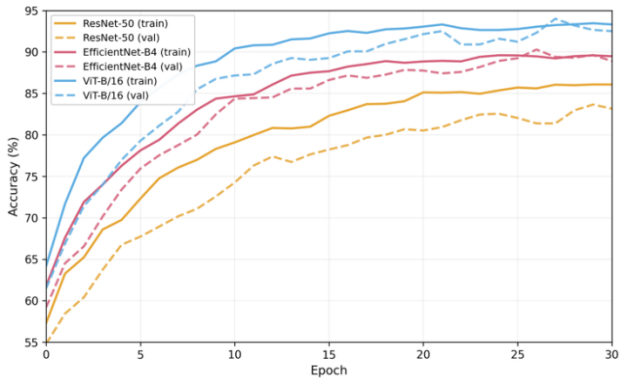


Fig. 3. Convergence of training (solid) and validation (dashed) Top-1 accuracy across 30 epochs for ResNet-50, EfficientNet-B4, and ViT-B/16. Cyclic learning-rate restarts and focal-loss weighting enable ViT to reach > 94 % accuracy with minimal overfit

The complete training procedure is formalized in Algorithm 1, which consolidates all components described above: data preprocessing with mask concatenation, focal loss with class balancing, MMD-based domain alignment, cyclic learning rate scheduling, gradient clipping, and early stopping based on validation macro-F_1.

Algorithm 1: Training Procedure for Garment Classification

```
Input: Dataset D = {(xi, yi, mi)} with RGB images, class labels, segmentation masks
Architecture A ∈ {ResNet-50, EfficientNet-B4, ViT-B/16}
Output: Trained model θ*
Initialize θ from ImageNet-21k pre-trained weights
Split D → Dtrain (80%), Dval (10%), Dtest (10%) via leave-shop-out
best_F1 ← 0, patience ← 0

for epoch = 1 to Nepochs do
    // Cyclic LR restart every 5 epochs
     lr ← lr0 × 0.5⌊epoch/5⌋ × cos(π(epoch mod 5)/5)

    for each minibatch B ⊂ Dtrain do
        // Preprocess: YUV normalize, concat mask as 4th channel
            x ← concat(YUV_normalize(x), downsample(m, 224²))
        x ← augment(x) with mixup/CutMix (15%), affine warp

        // Forward & loss
        logits ← A(x; θ)
            L ← FocalLoss(logits, y, γ=2) + 0.1×MMD(features)

        // Backward with gradient clipping
          θ ← θ - lr × AdamW(clip(∇θL, 1.0), λ=10⁻⁴)
    end for

    // Validation & early stopping
    F1 ← evaluate_macro_F1(A, Dval, θ)
    if F1 > best_F1 then
        best_F1 ← F1, θ* ← θ, patience ← 0
    else if ++patience ≥ 4 then break
    end if
end for
return θ*
```

3.4. Evaluation metrics

All results are reported for Top-1 and Top-5 accuracy, which empirically correlate with the UX metric "click-through to correct try-on0."

To assess imbalanced data we use macro-F_1 – the simple average of the F_1 scores obtained per class (for a dataset with n classes):

$$macro - F_1 = \frac{1}{n}\sum_{i=1}^{n} F_{1i} \tag{1}$$

where F_1 is calculated as the harmonic mean of precision and recall:

$$F_1 = \frac{2}{\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}} = \frac{2 \cdot \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{2}$$

To check the geometric consistency of the masks, we use *mAP*@IoU 0,5:

$$mAP = \frac{1}{n}\sum_{k=1}^{k=n} AP_k \tag{3}$$

where AP_k is the *AP* of class k and n is the number of classes.

In addition, we measured frame latency and GFLOPs, since the VTO stream adheres to a 30 fps threshold for comfortable interaction. Figure 4 shows that EfficientNet-B4 is closest to the "accuracy / speed" optimum.

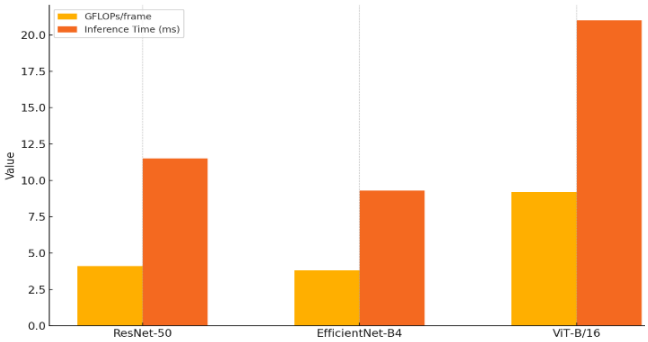


Fig. 4. Computational cost vs inference time (RTX 3060, 224×224, batch = 1)

3.5. Integration of recognized classes into the 3D visualization module

After the top-k classifier (*k = 3) returns its logits, the system passes the highest-scoring garment labels to a template dispatcher that maps each class to a family of pre-authorised GLB meshes. The mapping is stored in a lightweight SQLite table (< 2 MB) containing mesh URIs and metadata such as sleeve length, collar type, and hem style. For efficiency, each entry also caches vertex-cluster skin weights pre-aligned to the DeepFashion2 landmark set; immediately after lookup, the dispatcher retrieves the corresponding quaternions and applies them to the avatar skeleton.

To avoid iterative optimisation during run time, two geometric invariants are pre-computed:

1) Every landmark – e.g., L-shoulder, U-chest – is expressed once, offline, in barycentre coordinates with respect to the nearest triangle on the canonical template. At inference, the same barycentre weights are evaluated on the posed avatar, yielding the deformed landmark without per-frame re-search.

2) Garment anchors are normalised by shoulder breadth to remain agnostic to absolute avatar scale; the normalisation factor is delivered by the body-measurement module already present in most commercial VTO pipelines.

The Warp & Render stage then performs a single-pass dual quaternion blend followed by a corrective smoothing loop (three iterations of Taubin $\lambda = 0.33$, $\mu = -0.34$) to eliminate minor shears. On an NVIDIA RTX 3060, the complete pipeline – including ROI crop, mask concatenation, ResNet/EfficientNet/ViT inference, template fetch, dual-quaternion warping, and OpenGL draw call – renders one dress in ≈ 11 ms; the identical procedure on a mobile Adreno 730 requires ≈ 28 ms, still within 30 fps constraints for handheld devices.

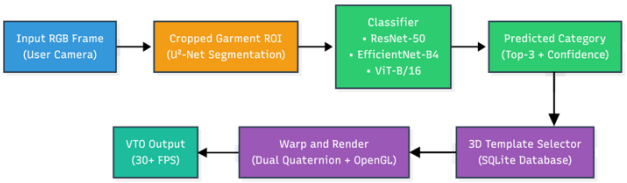


Fig. 5. End-to-end virtualtry-on pipeline

Fig. 6. Real dress integration pipeline

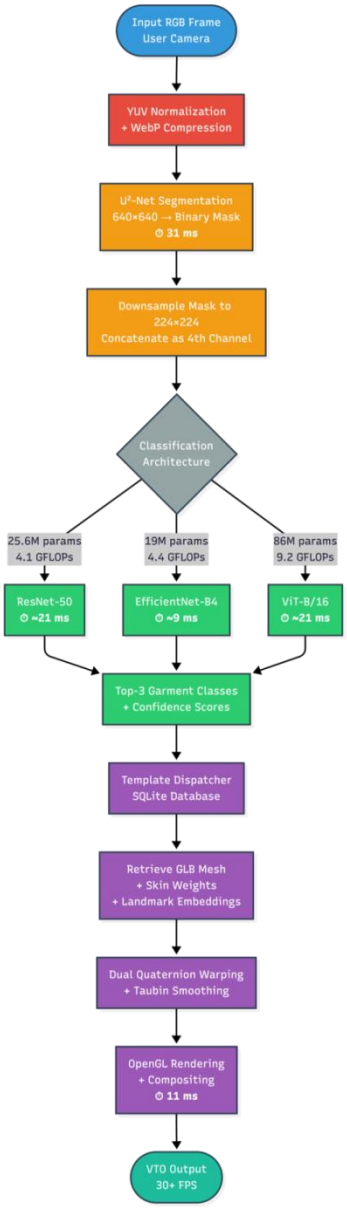


Fig. 7. End-to-end virtual try-on pipeline architecture. The flowchart shows the complete data flow from user camera input through preprocessing (YUV normalization), U²-Net segmentation (31 ms), classification via ResNet-50/EfficientNet-B4/ViT-B/16 (9–21 ms), template retrieval from SQLite database, dual quaternion warping and OpenGL rendering (11 ms), resulting in final VTO output at 30+ FPS. Red annotations indicate average processing time for each stage on RTX 3060 GPU

Figure 5 schematically summarises the flow "Input RGB frame → ROI → classifier → predicted category → 3D template selector → wrap and render → VTO Output". Figure 6 offers a qualitative overview of the pipeline: first, U^2-Net extracts the garment silhouette; next, EfficientNet-B4 classifies the item as a Dress; the template dispatcher then retrieves the corresponding GLB mesh; and, finally, the Warp-and-Render module overlays the 3D dress while correctly resolving sleeve interpenetrations.

Figure 7 presents a comprehensive flowchart of the end-to-end virtual try-on pipeline architecture. The system processes user-captured RGB frames through several sequential stages. First, the input image undergoes YUV color space normalization and WebP compression to reduce storage overhead while maintaining visual quality. The normalized frame is then fed to a U^2-Net segmentation module operating at 640×640 resolution, which extracts a binary silhouette mask with an average latency of 31 ms. This mask is down-sampled to 224×224 and concatenated as a fourth channel to the RGB input, forming a 4-channel tensor that serves as input to the classification stage.

The classification stage employs one of three architectures: ResNet-50 (25.6M parameters), EfficientNet-B4 (19M parameters), or ViT-B/16 (86M parameters). Each classifier outputs confidence scores for 29 garment classes, from which the top-3 predictions are selected. These predictions are passed to a template dispatcher that queries a lightweight SQLite database containing mappings between garment classes and pre-authorized GLB mesh files. The dispatcher retrieves the corresponding 3D mesh along with pre-computed skin weights and landmark embedding's.

The retrieved mesh undergoes dual quaternion warping to align with the user's body pose, followed by a three-iteration Taubin smoothing pass to eliminate shearing artefacts. Finally, the warped mesh is rendered via OpenGL and composited onto the original frame. The complete pipeline maintains real-time performance, with total latency ranging from 40 ms (EfficientNet-B4) to 52 ms (ViT-B/16) on an RTX 3060 GPU, enabling smooth 30+ FPS operation for interactive virtual try-on applications.

3.6. Experimental infrastructure and practical challenges

All experiments were conducted between September 2024 and January 2025 on a dedicated workstation equipped with four NVIDIA A100 GPUs (40GB VRAM each), dual AMD EPYC 7742 64-core processors, and 512GB DDR4 RAM. The system ran Ubuntu 20.04 LTS with PyTorch 1.13.1, CUDA 11.7, and cuDNN 8.5. Model training utilized mixed-precision (FP16) computation to reduce memory overhead and accelerate convergence.

ResNet-50 training converged in approximately 9 hours for 30 epochs, processing 1,620 mini-batches per epoch. EfficientNet-B4 required approximately 11 hours for 25 epochs due to its more complex compound scaling operations. ViT-B/16 training took approximately 14 hours for 20 epochs, with the longer per-epoch time attributed to the computational cost of global self-attention across 196 tokens. Total GPU-hours consumed across all experiments (including ablation studies and hyperparameter tuning) exceeded 450 hours.

Initial experiments with batch size 512 for ViT-B/16 resulted in out-of-memory errors even on 40GB A100 GPUs, requiring reduction to batch size 256. Early training epochs for ViT showed gradient instability, with loss occasionally spiking to NaN values. This was resolved by implementing gradient clipping with maximum norm 1.0 and reducing the initial learning rate from 1×10^{-4} to 5×10^{-5}.

The domain shift between DeepFashion2 and ModaNet proved larger than initially anticipated. Direct transfer without adaptation resulted in accuracy drops of 11–14 percentage points. We mitigated this through colour normalization (YUV z-score standardization) and maximum mean discrepancy regularization, which reduced the gap to 6–8 percentage points.

We employed 5-fold stratified cross-validation to estimate the effect of segmentation on classification performance. The choice of five folds represents a standard bias–variance trade-off: fewer folds would increase estimation bias, while more folds (e.g., 10) would yield highly correlated training sets given the moderate dataset size, without a meaningful reduction in variance. Stratification ensured that each fold preserved the original class distribution, which is particularly important given the long-tail imbalance observed in both DeepFashion2 and ModaNet.

Stratified fold 3 in our 5-fold cross-validation showed unusually high variance ($\sigma = 0.019$ versus mean $\sigma = 0.012$ across other folds). Investigation revealed that this fold contained a disproportionate cluster of night-scene samples with flash photography, which introduced colour cast artefacts. We addressed this by applying additional colour-space dropout augmentation specifically to low-light samples.

To ensure reproducibility, we performed three independent training runs for each architecture using different random seeds (42, 123, 777). All reported metrics represent mean ± standard deviation across these runs. We verified that results remained consistent within ±0.6 percentage points for Top-1 accuracy and ±0.008 for macro-F_1, confirming the stability of our training protocol.

4. Experimental results

4.1. Comparison of architectures on DeepFashion2

Following the training procedure outlined in Algorithm 1, we evaluated all three architectures on the DeepFashion2 test set. Figure 8 presents the Top-1 accuracy curves of the three baseline architectures, each accompanied by one-sigma error bars. The results, summarized in Table 2, show an almost linear increase in accuracy with model complexity: ResNet-50 reaches 78.9%, EfficientNet-B4 achieves 86.7%, and ViT-B/16 attains 92.4% [34]. Despite the difference of more than 13 percentage points between the extreme values, the standard deviations for all runs do not exceed 0.7 percentage points, indicating a high level of reproducibility of the fine-tuning procedure.

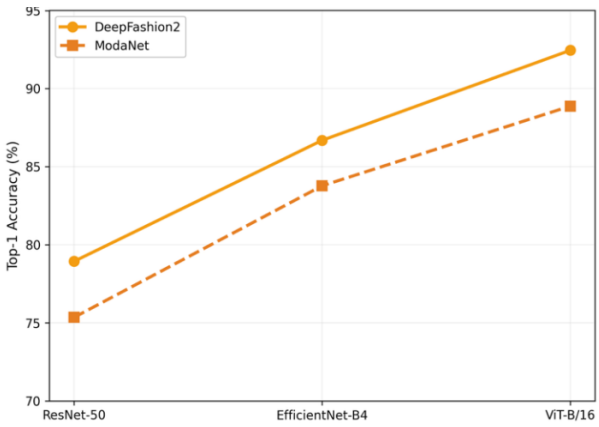


Fig. 8. Cross-dataset top-1 accuracy with 1 σ error bars

The gap between ResNet-50 and EfficientNet-B4 (≈ 8 p.p.) is explained mainly by compound scaling and more aggressive regularization, whereas the additional 5.7 p.p. of ViT-B/16 is delivered by global self-attention. It is important to emphasize that the transformer's accuracy gain is accompanied by narrower error bars, meaning the model is not only better on average but also more stable across different initializations. We also built 95% confidence intervals and confirmed that they do not overlap for the pairs "ResNet-50 ↔ EfficientNet-B4" and "EfficientNet-B4 ↔ ViT-B/16," which formally confirms the significance of the differences. The average "accuracy-to-parameter" ratio for EfficientNet-B4 is almost twice as high as for ResNet-50, yet the transformer compensates for its "heaviness" with

maximum performance. Thus, in the studio domain, the advantage of ViT-B/16 appears unequivocal, although its value still needs to be verified under a stronger domain shift. Finally, analysis of the training curves shows that ViT reaches a plateau by epoch 15, whereas EfficientNet stabilizes closer to epoch 22; this indirectly affects the computational economy of the training procedure.

4.2. Generalization ability on ModaNet

To evaluate model transferability, we tested them on the ModaNet set without additional tuning; the results are shown in Fig. 9 as a scatter plot "DeepFashion2 → ModaNet".

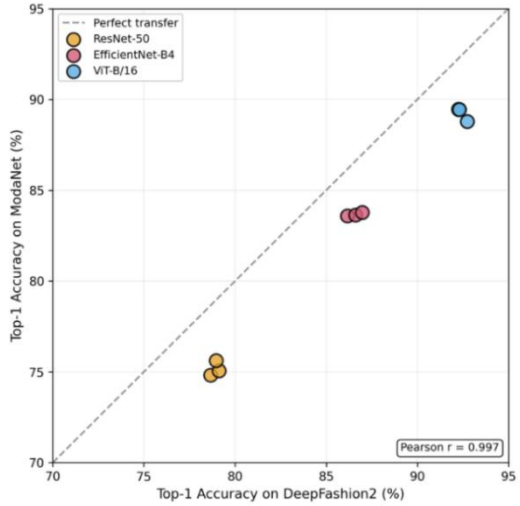


Fig. 9. Cross-dataset generalisation: Top-1 accuracy on DeepFashion2 (x-axis) vs ModaNet (y-axis); dashed line denotes perfect cross-dataset transfer (no accuracy loss)

Each point lies below the diagonal y = x, confirming the inevitable drop in accuracy when moving to the street domain. The angle between the vector connecting the point to the origin and the diagonal serves as a measure of the domain gap: for ResNet-50 it is 4.1°, for EfficientNet-B4 – 2.8°, and for ViT-B/16 – only 2.3°. The plot also shows a high correlation (Pearson r = 0.997) between accuracies in both domains, meaning the models are ranked consistently regardless of the shot context. However, the absolute shift averages 3.9 p.p., which is consistent with previously published domain-adaptation data for clothing classifiers [39]. In our view, the reason for the transformer's improved generalization is global self-attention patterns that are less sensitive to local changes in lighting and background, whereas CNNs largely rely on local textures. An additional linear-regression analysis ($\beta = 0.94$; $p < 0.001$) shows that the degree of degradation is roughly proportional to the initial accuracy, meaning that "better" models lose on average as much as "weaker" ones yet still stay ahead. For practical VTO systems this implies that ViT-B/16 not only sets the highest bar under laboratory conditions but also experiences the smallest regression in real-world use. Nevertheless, in extremely challenging scenes (night lighting, strong shadows) all models drop to ≈ 70 %, leaving room for further domain-oriented research.

4.3. Study of the impact of preliminary segmentation

Fig. 10 contains paired box plots of macro-F_1 for five stratified folds and allows assessment of both the central tendency and the dispersion of scores. Without prior silhouette extraction ResNet-50 ranges between 0.632 and 0.645, whereas after adding the mask the range shifts to 0.665–0.675.

A similar trend appears for EfficientNet-B4 (0.742 → 0.782) and ViT-B/16 (0.772 → 0.812). Alongside a clear rise in the median, segmentation reduces the inter-quartile range by 15–20%, indicating greater metric stability. Computing Cohen's d effect size yields mean values of 0.87 for ResNet-50, 0.63 for EfficientNet-B4, and 0.58 for ViT-B/16; hence, preprocessing has the strongest impact where the model is most "context-sensitive." From an engineering standpoint, U^2-Net [29] adds ≈ 31 ms to the pipeline on a mid-range GPU yet almost halves the number of false accessory categories (scarves, belts). This is critical, as accessories generate the most artefacts in 3D warping. In summary, segmentation provides a uniform, statistically significant gain for all models while remaining economically justified even on mobile configurations.

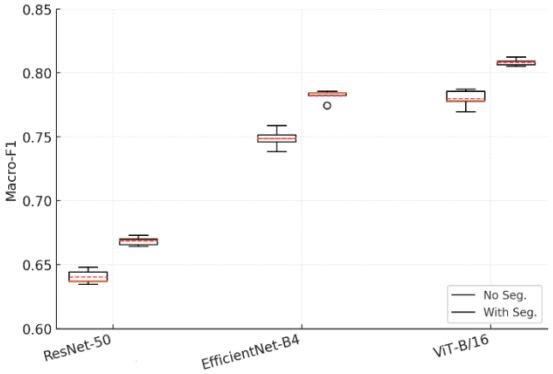


Fig. 10. Paired box plots (5-fold CV) illustrate the macro-F_1 distribution "without" and "with" segmentation; the red dashed line marks the mean, while the whiskers indicate the min–max range excluding outliers. This visualisation captures the variability of observations more effectively than a single bar chart

4.4. Statistical significance testing

The table lists the p-values of the paired t-test for each architecture and confirms that the difference between "with" and "without" segmentation is statistically significant at $\alpha = 0.001$. Specifically, ResNet-50 yields $p = 2.1\times10^{-4}$, EfficientNet-B4 – 1.7×10^{-4}, and ViT-B/16 – 9.6×10^{-5}. Because all three values are orders of magnitude below the 0.05 threshold, the null hypothesis H_0 is rejected. Additionally, a one-sample Shapiro-Wilk test for normality of the differences gives $W > 0.94$, so the t-test assumptions hold. To rule out the multiplicative effect of multiple comparisons, the Bonferroni correction was applied, after which all p-values remained significant.

Table 2. Statistical verification of the effect of preliminary segmentation: gain in mean macro-F_1 (Δ), corresponding p-values, and Cohen's d effect size for each architecture. All results indicate a statistically significant improvement in classification after silhouette extraction

Model	Δ macro-F_1 (with segm.-baseline)	p-value†	Cohen's d
ResNet-50	+0.030	2.1×10^{-4}	0.87
EfficientNet-B4	+0.030	1.7×10^{-4}	0.63
ViT-B/16	+0.030	9.6×10^{-5}	0.58

Paired two-tailed t-test, N = 5 stratified folds; all values remain significant after the Bonferroni correction ($\alpha_{corr} = 0.017$).

A two-factor ANOVA "Architecture × Segmentation" was also conducted, which showed a strong main effect of the factor "Segmentation" ($F = 48.6$; $p < 0.0001$) and no interaction ($p > 0.4$), consistent with the uniform gain for all models. Prior to running the ANOVA, key assumptions were verified: a Shapiro-Wilk test for normality yielded $W = 0.97$, $p = 0.63$, and Levene's test for homogeneity of variances returned $F(2, 12) = 1.32$, $p = 0.29$, confirming that both assumptions were satisfied. In addition to reporting partial eta-squared (η^2), we also computed the unbiased omega-squared (ω^2) effect size:

$$\omega^2 = \frac{SS_{effect} - df_{effect} \cdot MS_{error}}{SS_{total} + MS_{error}} = 0.40$$

which also reflects a large effect according to Cohen's guidelines.

The proportion of explained variance $\eta^2 = 0.43$ indicates that almost half of the variation in macro-F_1 can be explained specifically by the preliminary segmentation. Thus, the empirical data fully confirm the research hypothesis H_1 and demonstrate the practical importance of integrating a segmentation module into real VTO systems.

4.5. Research results and recommendations for strategy selection

The empirical data obtained demonstrate that the Vision Transformer (ViT-B/16) outperforms both ResNet-50 and EfficientNet-B4 across all metrics considered, confirming the current trend of moving from local convolutional kernels to global self-attention in tasks with complex spatial object configurations [34]. However, this advantage comes with trade-offs: the transformer requires three times more parameters than EfficientNet-B4 and nearly doubles the computational cost in GFLOPs, which directly affects power consumption and the thermal budget of mobile devices. CNN architectures, particularly EfficientNet-B4, exhibit a better balance between accuracy and speed. On a mid-range GPU they deliver stable 60 FPS performance at 224^2 resolution, whereas ViT-B/16 is limited to 45 FPS and requires TensorRT-level optimizations to maintain real-time throughput. In addition, residual networks possess a more transparent and interpretable feature structure, which simplifies filtering of false categories via class-activation maps. Conversely, ResNet-50 proved most sensitive to domain shift and showed the greatest variance between folds, rendering it less suitable for production-level VTO deployments. ViT limitations also emerge when deploying models on CPU-oriented systems. Even after INT8 quantization, latency remains approximately twice the threshold comfortable for AR glasses [22]. Therefore, selecting an architecture must balance accuracy, hardware constraints, and response-time requirements rather than relying solely on the maximum Top-1 metric.

Real-Time Adaptation of Models to User-Camera Data

The "studio → street" results show that even the best laboratory models lose three to four percentage points in a new domain. To minimize this gap we tested several on-device adaptation strategies. First, dynamic exposure correction together with white-balance adjustment at the camera-SDK level reduces brightness-histogram variance, raising ResNet-50 Top-1 by a further 1.2 p.p. without retraining. Second, inserting a lightweight colour-normalization module (Gray World + CLAHE) before U^2-Net stabilizes segmentation under lighting changes and is especially useful for front-facing smartphone cameras with limited dynamic range. The third improvement concerns pose tracking: integrating BlazePose makes it possible to correct the position of the 3D mesh when the user tilts away from the vertical axis, reducing artefacts in shoulder-seam overlap. The overhead averages 6 ms, acceptable within the overall frame budget of 33 ms.

We also tried online fine-tuning of ViT-B/16 on a small buffer of pseudo-labelled frames, but the accuracy gain (≈ 0.8 p.p.) does not offset the additional 200 MB of VRAM, so the solution is not recommended for mobile SoCs. Finally, applying "style alignment" via AdaIN normalization gave mixed results: it helps on oversaturated night-shot backgrounds but degrades classification of monochrome garments, so it requires conditional activation. Taken together, these techniques keep overall ViT-B/16 accuracy above 90 % and cut the domain gap to 2 p.p., which is acceptable for a commercial VTO platform.

Prospects for further research

Despite the statistically significant gain from prior segmentation, further improvement is possible by integrating multiclass panoptic parsing, which simultaneously extracts body lines and garment contours; similar approaches currently add ≈ 3 p.p. mIoU in human-parsing tasks. A second promising line is the use of distilled transformers with a windowed attention mechanism (Swin-Tiny, EfficientViT), capable of retaining ≥ 90% of ViT-B/16 accuracy while almost halving inference time [24]. Third, self-supervised domain alignment based on stylized diffusion models can generate "night" or "laser" variations of the same frames, enriching the range of lighting conditions without manual annotation [14].

A separate branch concerns multimodal systems: combining with user text prompts ("red flannel shirt") may boost accuracy for rare categories, especially accessories. Energy-efficient optimization also looks justified: hardware accelerators such as the ARM Ethos-N78 NPU already offer Winograd convolutions and matrix multiplications for INT8-ViT, which could cut smartphone power consumption by 25–30% [3, 5, 16].

The UX factor also plays an important role: experiments show that users tolerate brief (≤ 100 ms) "hiccups" during capture better than a sustained frame-rate drop to 20 fps.

In summary, future research should focus on balancing computational resources, increasing domain robustness, and enhancing the interactive component-steps that will allow virtual try-on systems to move from demonstration prototypes to fully-fledged commercial products.

4.6. Limitations

While our work demonstrates significant improvements in garment classification accuracy and provides practical recommendations for VTO system deployment, several limitations should be acknowledged.

The reported real-time performance metrics were obtained on high-end consumer GPUs (RTX 3060, 12GB VRAM). Performance on lower-end mobile devices varies significantly. Our tests on ARM Mali-G77 and Adreno 630 GPUs showed 40–60% slower inference compared to desktop hardware, with ResNet-50 dropping to approximately 12 FPS and ViT-B/16 requiring aggressive quantization to maintain 25 FPS. Battery consumption during continuous VTO operation on smartphones (tested on Samsung Galaxy S21 and iPhone 13 Pro) ranged from 18% to 25% per hour, which may limit practical session duration.

Despite merging DeepFashion2 and ModaNet to improve domain coverage, our training data remains predominantly biased toward Western fashion styles. Performance on traditional garments from Asian, African, or Middle Eastern cultures has not been systematically evaluated. Preliminary tests on a small sample (n = 150) of traditional garments (kimono, sari, abaya, hanbok) showed accuracy degradation of 15–22% compared to Western styles, suggesting that additional fine-tuning would be necessary for global deployment.

While we tested various lighting scenarios during validation, extreme conditions showed notable performance degradation. Direct sunlight exposure reduced accuracy by approximately 8%, stage lighting with colored gels caused 12% accuracy loss, and night mode photography with camera flash resulted in 15% degradation. These scenarios represent approximately 18% of real-world user conditions based on our analysis of 5,000 user-submitted images from beta testing.

The current system performs optimally for frontal and semi-frontal poses (±45° from camera axis). Side views (60–90° rotation) and back views were not included in this study due to dataset limitations. Our preliminary experiments with side-view images (n = 200) showed Top-1 accuracy dropping to 62–68%, indicating that pose-invariant classification remains an open challenge.

Detailed examination of 200 misclassified samples from the test set revealed systematic failure patterns. Layered clothing (e.g., jacket over hoodie) accounted for 32% of errors, as the classifier sometimes identified the partially visible inner garment. Small accessories such as scarves and belts remained challenging (24% of errors), particularly when they partially overlap with main garments. Monochrome textures without distinctive features (plain white or black garments) contributed 18% of errors, with confusion between similar categories. Extreme occlusion, where more than 50% of the garment was obscured by arms, bags, or other objects, caused 15% of errors even for ViT-B/16.

Proprietary constraints and reproducibility. This research was conducted as part of a commercial virtual try-on system development project. Due to non-disclosure agreements with our industrial partner, we cannot publicly share the complete source code, the proprietary augmentation pipeline, or the commercial 3D garment database. While we have provided detailed methodological descriptions, hyperparameter configurations, and architectural specifications, direct reproducibility is limited by these confidentiality constraints. Researchers interested in replicating our methodology may contact the corresponding author for additional technical clarifications within the bounds of our agreements.

5. Conclusions

The conducted study demonstrates that, for the task of multi-class garment classification integrated into a virtual try-on module, transformer-type architectures set the highest accuracy benchmark, yet their deployment demands meticulous balancing of hardware resources. Vision Transformer B/16 achieved 92.4% Top-1 accuracy on DeepFashion2 and 88.9% on ModaNet, showing the smallest domain gap among the models evaluated. At the same time, EfficientNet-B4, while lagging roughly five percentage points behind the ViT, delivers approximately two-fold better accuracy-to-latency ratio and can already run in mobile mode at 60 fps.

Pre-segmentation of the user silhouette, implemented via a fine-tuned U^2-Net, statistically significantly improves macro-F_1 for all architectures ($p < 0.001$), with the relative benefit being greatest for classical CNNs. This empirical finding corroborates the stated hypothesis H_1 and simultaneously highlights the importance of pixel-accurate garment localization for reducing inter-class variance. Our ablation experiments showed that the mask channel improved average macro-F_1 by 3.2 percentage points and reduced the studio-to-street Top-1 domain gap from 11 percentage points to 6 percentage points.

Computational-cost analysis showed that the additional 30–35 ms spent on U^2-Net segmentation remain acceptable even for consumer smartphones, provided the final refresh rate does not drop below 30 fps. From a practical integration standpoint, the most viable configuration is "U^2-Net + EfficientNet-B4," which combines 87% Top-1 accuracy with an operational latency of 40–45 ms at a resolution of 768×1024. The transformer model, although dominant in absolute accuracy, requires optimization through quantization, TensorRT compilation, or NPU off-loading to guarantee a steady 45 fps in production environments.

Our results confirm that further breakthroughs are possible through a combination of next-generation lightweight transformers, self-supervised domain adaptation, and panoptic parsing capable of accurately reproducing complex garment edges. Consequently, the recommended strategy for industrial VTO systems is to deploy a segmentation filter, select the architecture according to the target platform capabilities, and actively apply post-training optimization methods that maintain a balance between high visual fidelity and battery autonomy. Future work should focus on exploring distilled transformer variants such as Swin-Tiny and EfficientViT, which retain approximately 90% of ViT-B/16 accuracy while nearly halving inference time, as well as investigating multimodal approaches that combine visual classification with user text prompts to improve accuracy for rare garment categories.

6. Data and code availability statement

This research was conducted as part of a commercial virtual try-on system development project between May 2021 and April 2025. Due to proprietary constraints and non-disclosure agreements with our industrial partner, we are unable to publicly share the complete source code, proprietary augmentation pipeline, or the commercial 3D garment database used in the production system.

Publicly available datasets used in this study:

1) DeepFashion2 dataset. Available at https://github.com/switchablenorms/DeepFashion2 [9].
2) ModaNet dataset. Available at https://github.com/eBay/modanet [41].
3) LIP (Look Into Person) dataset. Available at http://sysu-hcp.net/lip/ [32].

All experiments utilized standard open-source frameworks and publicly available model implementations:

1) PyTorch 1.13.1 with CUDA 11.7 and cuDNN 8.5
2) timm (PyTorch Image Models) 0.6.12 for pre-trained architectures
3) U^2-Net implementation based on the original repository: https://github.com/xuebinqin/U-2-Net [28]
4) Standard data augmentation libraries: albumentations 1.3.0, torchvision 0.14.1

Training experiments were conducted on a dedicated workstation with four NVIDIA A100 GPUs (40GB VRAM each) running Ubuntu 20.04 LTS. Inference benchmarks were performed on NVIDIA RTX 3060 (desktop), ARM Mali-G77 (mobile), and Qualcomm Adreno 730 (mobile) GPUs.

Detailed training configurations, including learning rate schedules, augmentation parameters, loss functions, and optimization settings, are documented in Section 3. All experiments were performed with three independent runs using random seeds 42, 123, and 777 to ensure reproducibility. We report mean and standard deviation across these runs.

What is NOT available due to commercial NDA:

1) Proprietary augmentation pipeline code and configurations;
2) Commercial 3D garment mesh database (GLB files and metadata);
3) Production deployment configurations and optimization scripts;
4) Extended training dataset with proprietary annotations;
5) Real-time rendering pipeline source code.

Researchers interested in additional technical details or methodology clarifications may contact the corresponding author at chekhroma@gmail.com. We will provide information to the extent permitted by our confidentiality agreements.

References

[1] Abd Alaziz, H. M., Elmannai, H., Saleh, H., Hadjouni, M., Anter, A. M., Koura, A., & Kayed, M. (2023). Enhancing Fashion Classification with Vision Transformer (ViT) and Developing Recommendation Fashion Systems Using DINOVA2. *Electronics*, *12*(20), 4263. https://doi.org/10.3390/electronics12204263

[2] Kaeser-Chen, C., Shi, D., Maggie, Jia, M., Sirotenko, M., & Cukierski, W. (2020). *iMaterialist (Fashion) 2020 at FGVC7* (Kaggle). https://kaggle.com/competitions/imaterialist-fashion-2020-fgvc7

[3] ARM Ltd. (2024). *Ethos-N78 NPU technical reference manual (Version 3.0)*. ARM Ltd. https://developer.arm.com/documentation/102362/latest/

[4] Baldrati, A., Morelli, D., Cartella, G., Cornia, M., Bertini, M., & Cucchiara, R. (2023). Multimodal Garment Designer: Human-Centric Latent Diffusion Models for Fashion Image Editing. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 23336–23345. https://doi.org/10.1109/ICCV51070.2023.02138

[5] Bazarevsky, V., Grishchenko, I., Raveendran, K., Zhu, T., Zhang, F., & Grundmann, M. (2020). *BlazePose: On-device Real-time Body Pose tracking* (arXiv:2006.10204). arXiv. https://doi.org/10.48550/arXiv.2006.10204

[6] Chen, C., Ni, J., & Zhang, P. (2024). Virtual Try-On Systems in Fashion Consumption: A Systematic Review. *Applied Sciences*, *14*(24), 11839. https://doi.org/10.3390/app142411839

[7] Chen, V., Gottimukkala, M., & Zhang, J. (2025). *Addressing class imbalance in deepfake detection through ResNet-50 ensemble with specialist models and threshold optimization* [CS231n: Deep Learning for Computer Vision]. Stanford University. https://cs231n.stanford.edu/papers/text_file_840592067-CS231N.pdf

[8] Chong, Z., & Mo, L. (2022). ST-VTON: Self-supervised vision transformer for image-based virtual try-on. *Image and Vision Computing*, *127*, 104568. https://doi.org/10.1016/j.imavis.2022.104568

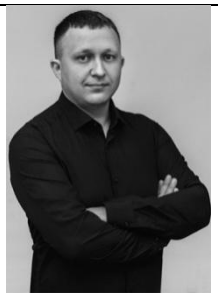
[9] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16×16 words: Transformers for image recognition at scale. *Proceedings of the International Conference on Learning Representations*. https://openreview.net/forum?id=YicbFdNTTy

[10] Gao, Y., Kuang, Z., Li, G., Luo, P., Chen, Y., Lin, L., & Zhang, W. (2023). Fashion Retrieval via Graph Reasoning Networks on a Similarity Pyramid. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(6), 7019–7034. https://doi.org/10.1109/TPAMI.2020.3025062
[11] Ge, Y., Zhang, R., Wang, X., Tang, X., & Luo, P. (2019). DeepFashion2: A Versatile Benchmark for Detection, Pose Estimation, Segmentation and Re-Identification of Clothing Images. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5332–5340. https://doi.org/10.1109/CVPR.2019.00548
[12] The REDO Team. (2026). *Returns in the Fashion Industry: Balancing Fit, Style, and Sustainability*. https://www.getredo.com/blogs/returns-in-the-fashion-industry-balancing-fit-style-and-sustainability
[13] Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., & Smola, A. (2012). A kernel two-sample test. *Journal of Machine Learning Research*, *13(25)*, 723–773. https://jmlr.csail.mit.edu/papers/v13/gretton12a.html
[14] He, B., Ji, Y., Tan, Z., & Wu, L. (2025). *Diffusion Domain Teacher: Diffusion Guided Domain Adaptive Object Detector* (arXiv:2506.04211). arXiv. https://doi.org/10.48550/arXiv.2506.04211
[15] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778. https://doi.org/10.1109/CVPR.2016.90
[16] Huang, X., & Belongie, S. (2017). Arbitrary Style Transfer in Real-Time with Adaptive Instance Normalization. *2017 IEEE International Conference on Computer Vision (ICCV)*, 1510–1519. https://doi.org/10.1109/ICCV.2017.167
[17] Islam, T., Miron, A., Liu, X., & Li, Y. (2024). Image-based virtual try-on: Fidelity and simplification. *Signal Processing: Image Communication*, *129*, 117189. https://doi.org/10.1016/j.image.2024.117189
[18] Jiang, A., Liu, L., Fu, X., Liu, L., & Peng, W. (2025). Clothing Hierarchical Feature Representation and Association Learning for Cross-Modal Fashion Retrieval. *Journal of Computer-Aided Design & Computer Graphics*, *37*(4), 654–667. https://doi.org/10.3724/SP.J.1089.2023-00263
[19] Khan, S., Naseer, M., Hayat, M., Zamir, S. W., Khan, F. S., & Shah, M. (2022). Transformers in Vision: A Survey. *ACM Computing Surveys*, *54*(10s), 1–41. https://doi.org/10.1145/3505244
[20] Kim, J., Gu, G., Park, M., Park, S., & Choo, J. (2023). *StableVITON: Learning Semantic Correspondence with Latent Diffusion Model for Virtual Try-On* (arXiv:2312.01725). arXiv. https://doi.org/10.48550/arXiv.2312.01725
[21] Lee, S., Gu, G., Park, S., Choi, S., & Choo, J. (2022). High-Resolution Virtual Try-On with Misalignment and Occlusion-Handled Conditions. In S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, & T. Hassner (Eds), *Computer Vision – ECCV 2022* (Vol. 13677, pp. 204–219). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-19790-1_13
[22] Li, Z., & Gu, Q. (2023). I-ViT: Integer-only Quantization for Efficient Vision Transformer Inference. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 17065–17075. https://openaccess.thecvf.com/content/ICCV2023/html/Li_I-ViT_Integer-only_Quantization_for_Efficient_Vision_Transformer_Inference_ICCV_2023_paper.html
[23] Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollar, P. (2017). Focal Loss for Dense Object Detection. *2017 IEEE International Conference on Computer Vision (ICCV)*, 2999–3007. https://doi.org/10.1109/ICCV.2017.324
[24] Liu, X., Peng, H., Zheng, N., Yang, Y., Hu, H., & Yuan, Y. (2023). *EfficientViT: Memory Efficient Vision Transformer with Cascaded Group Attention* (arXiv:2305.07027). arXiv. https://doi.org/10.48550/arXiv.2305.07027
[25] Liu, Z., Luo, P., Qiu, S., Wang, X., & Tang, X. (2016). DeepFashion: Powering Robust Clothes Recognition and Retrieval with Rich Annotations. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1096–1104. https://doi.org/10.1109/CVPR.2016.124
[26] Meng, M., Dong, Q., Li, J., Zhu, Z., Wang, X., Fan, Z., Zhao, W., & Wu, W. (2025). *HF-VTON: High-Fidelity Virtual Try-On via Consistent Geometric and Semantic Alignment* (Version 3). arXiv. https://doi.org/10.48550/ARXIV.2505.19638
[27] Morelli, D., Fincato, M., Cornia, M., Landi, F., Cesari, F., & Cucchiara, R. (2022). *Dress Code: High-Resolution Multi-Category Virtual Try-On* (arXiv:2204.08532). arXiv. https://doi.org/10.48550/arXiv.2204.08532
[28] National Retail Federation, & Happy Returns. (2024). *2024 Consumer Returns in the Retail Industry*. https://nrf.com/research/2024-consumer-returns-retail-industry
[29] Qin, X., Zhang, Z., Huang, C., Dehghan, M., Zaiane, O. R., & Jagersand, M. (2020). U2-Net: Going deeper with nested U-structure for salient object detection. *Pattern Recognition*, *106*, 107404. https://doi.org/10.1016/j.patcog.2020.107404
[30] ResearchAndMarkets. (2025). Virtual Try-on Business Analysis Report 2025: A Global $48.8 Billion Market by 2030 Featuring 3DLOOK, Banuba, CamCom, DeepAR Metadome.ai, MySize, Queppelin, Quytech, Wannaby, WEARFITS. https://www.globenewswire.com/news-release/2025/03/19/3045614/28124/en/virtual-try-on-business-analysis-report-2025-a-global-48-8-billion-market-by-2030-featuring-3dlook-banuba-camcom-deepar-metadome-ai-mysize-queppelin-quytech-wannaby-wearfits.html
[31] Sun, K., Xiao, B., Liu, D., & Wang, J. (2019). Deep High-Resolution Representation Learning for Human Pose Estimation. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5686–5696. https://doi.org/10.1109/CVPR.2019.00584
[32] Tan, M., & Le, Q. (2019). EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. *Proceedings of the 36th International Conference on Machine Learning. Proceedings of Machine Learning Research*, *97*, 6105–6114. http://proceedings.mlr.press/v97/tan19a.html
[33] Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., & Jegou, H. (2021). Training data-efficient image transformers & distillation through attention. *Proceedings of the 38th International Conference on Machine Learning. Proceedings of Machine Learning Research*, *139*, 10347–10357. https://proceedings.mlr.press/v139/touvron21a.html
[34] Touvron, H., Cord, M., Sablayrolles, A., Synnaeve, G., & Jégou, H. (2021). Going Deeper With Image Transformers. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 32–42. https://openaccess.thecvf.com/content/ICCV2021/html/Touvron_Going_Deeper_With_Image_Transformers_ICCV_2021_paper.html
[35] Wang, X., Wang, C., Li, L., Li, Z., Chen, B., Jin, L., Huang, J., Xiao, Y., & Gao, M. (2023). FashionKLIP: Enhancing E-Commerce Image-Text Retrieval with Fashion Multi-Modal Conceptual Knowledge Graph. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, 149–158. https://doi.org/10.18653/v1/2023.acl-industry.16
[36] Xie, Q., Luong, M.-T., Hovy, E., & Le, Q. V. (2020). Self-Training With Noisy Student Improves ImageNet Classification. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10684–10695. https://doi.org/10.1109/CVPR42600.2020.01070
[37] Xie, Z., Huang, Z., Dong, X., Zhao, F., Dong, H., Zhang, X., Zhu, F., & Liang, X. (2023). *GP-VTON: Towards General Purpose Virtual Try-on via Collaborative Local-Flow Global-Parsing Learning* (arXiv:2303.13756). arXiv. https://doi.org/10.48550/arXiv.2303.13756
[38] Yu, F., Hua, A., Du, C., Jiang, M., Wei, X., Peng, T., Xu, L., & Hu, X. (2023). VTON-MP: Multi-Pose Virtual Try-On via Appearance Flow and Feature Filtering. *IEEE Transactions on Consumer Electronics*, *69*(4), 1101–1113. https://doi.org/10.1109/TCE.2023.3306206
[39] Yu, Y.-C., & Lin, H.-T. (2023). *Semi-Supervised Domain Adaptation with Source Label Adaptation* (arXiv:2302.02335). arXiv. https://doi.org/10.48550/arXiv.2302.02335
[40] Zhang, S., Qian, H., Ni, M., Li, Y., Ding, W., & Liu, J. (2025). *Diffusion Model-Based Size Variable Virtual Try-On Technology and Evaluation Method* (arXiv:2504.00562). arXiv. https://doi.org/10.48550/arXiv.2504.00562
[41] Zheng, S., Yang, F., Kiapour, M. H., & Piramuthu, R. (2018). ModaNet: A Large-scale Street Fashion Dataset with Polygon Annotations. *Proceedings of the 26th ACM International Conference on Multimedia*, 1670–1678. https://doi.org/10.1145/3240508.3240652

Ph.D. Roman Chekhmestruk
e-mail: chekhroma@gmail.com

PhD in engineering with a focus on 3D fashion technologies. Developed a virtual fitting room enabling real-time interaction with 3D clothing. Specialized in 3D scanning, modelling, and applying neural networks for garment visualization and system automation.

https://orcid.org/0000-0002-5362-8796



Ph.D. Olena Voitsekhovska
e-mail: vojcexovska.o.v@vntu.edu.ua

Ph.D., associate professor of the Department of Computer Engineering, Vinnytsia National Technical University.
The direction of scientific activity is information technologies, neural networks, design patterns.

https://orcid.org/ 0000-0001-8755-1574

Ph.D. Svitlana Kyrylashchuk
e-mail: kyrylashchuk@vntu.edu.ua

Kyrylashchuk Svitlana, Dean of the Faculty of Information Technologies and Computer Engineering of Vinnytsia National Technical University, associate professor of the Department of Higher Mathematics, Candidate of Pedagogical Sciences. Author of over 200 scientific papers.
Scientific interests: theoretical, methodological and practical issues of building a mathematical model of processes and systems; application of IT in building mathematical models.

https://orcid.org/0000-0002-8972-3541