Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 4, Issue 6, 2026

X-HASHNET: A VISION TRANSFORMER-BASED DEEP HASHING FRAMEWORK FOR EFFICIENT SEMANTIC IMAGE RETRIEVAL

[1]Muhammad Irfan, [2]Javiriya Hameed Arain, [3]Kinza Fatima
[1]Student Muhammad Nawaz Sharif University of Agriculture, Multan
[2]Lecturer ,Department of Computer Science, National University of Modern Sciences and Languages (NUML) Hyderabad Campus.
[3]Bachelors in Computer Science, National University of Modern Language
[1]dairfankhan382@gmail.com, [2]javiriyahameed@gmail.com, [3]Kunzashaikh3@gmail.com,

DOI:- https://doi.org/10.5281/zenodo.20938330

Keywords
Deep Supervised Hashing, Vision Transformers (ViT), DeiT (Data-efficient Image Transformers), Fashion Image Retrieval, Binary Hash Codes, Hamming Distance Search, Fashion-MNIST, Multi-Objective Optimization, Self-Attention Mechanisms, FAISS Indexing

Article History
Received: 28 May, 2026
Accepted: 25 June, 2026
Published: 26 June, 2026

Corresponding Author: *

Abstract
X-HashNet presents a transformer based supervised hashing scheme to enable highly efficient fashion image retrieval. Leveraging the architecture of the famous popular vision model called the DeiT-Small Vision Transformer, the model has substituted the convolutional architectures with a global call to attention representation paradigm. Evaluated on the Fashion Minority dataset (Fashion-Mnist), which consists of 70,000 grayscale images from 10 categories of apparel, X- HashNet is used to make number (64 bit) binary embeddings from the raw inputs that are optimized for the retrieval based on the same hamming distance. The pipeline combines 5 key stages: ViT adoptable patch embedding, transformer-based feature extraction, supervised bottleneck hashing, multi-objective optimization and FAISS based binarization and indexing. The model has a mean Average Precision (mAP@100) of 0.9348, which is a new state-of-the-art benchmark for hashing on Fashion-MNIST. As we know, diagnostic analyses confirm the best utilization of codes and the average bit activation is 0.4907, inter & intra class hamming distances are 2.29 & 0.28 bits respectively and hash stability is 84.20 per cent. The bit redundancy score of 0.2775 and near ideal entropy distribution mean efficient encoding of information in all the hash dimensions. Empirical results also validate a strong level of generalization, yielding Precision@1 of 93.39% and maintaining the stability of the performance in deeper response latencies (P@5-P@100 ~93%). From a systems perspective, the average query time of 0.1738 milliseconds and throughput of more than 5 754 queries per second make X HashNet suitable for large scale deployment, where a mere 8 bytes per image are used to index the images. Visual attention maps validate the model's ability to both localize and maintain important structural features (e.g. silhouettes and textures of clothing). Collectively, these results show evidence that transformer-based hashing not only outperforms CNN counterparts in retrieval accuracy, but provides a scalable industrially-viable foundation for real time search and recommendation systems for fashion.

Introduction

Deep hashing has completely changed the way searchable images work at scale, and represents a highly powerful tool for reducing complex visual data into compact binary codes for fast searches using simple Hamming-distance calculations. One can imagine the ability to filter through many millions of images of fashion in milliseconds: this is the promise of hashing, which compromises a certain precision and in exchange we get massive improvements in speed and storage efficiency. Hand-made characteristics like SIFT's or HOG's features have been replaced with deep learning methods in which neural networks learn directly from the data to learn hash codes with semantic similarity. In supervised settings, the inclusion of class labels enhances this process and makes the codes non only compact but also discriminative. Similar garments, like pullovers, for example, tend to result in almost identical bit patterns, while unrelated things such as shirts and sneakers are very different. This evolution is particularly important in the case of fashion retrieval, where faint style, silhouette or texture distinctions determine relevance; users also come to expect near-instant results from e-Commerce sites or style recommendation engines [1]. Conventional unidirectional transmission process, which predominates in the development of deep hashing, Convolutional neural network (CNN) based hashing methods have obvious shortcomings rooted in the hash. CNNs are good for extracting the local patterns, edges, textures, and small motifs with the help of hierarchical convolutions that have a fixed receptive field. This approach is fine for rigid objects but it doesn't work well for fashion imagery, where the global context is king. Consider a pullover, for example: its sleeves may repeat each other across the image, or its neckline may repeat in the hemline in remote parts. A CNN may be easily concentrating on a sleeve cuff locally but will not be capturing its integration with the rest of the garment shape, which will create not-so-good hash codes that lead to poor fine-grained retrieval. Studies on e.g., Fashion-lev MNist dataset reveal that CNNs reach parts of the 93-95% accuracy in class inference but eventually the hash function performance degrades to mean average precision (mAP) scores lower than 0.92 which in turn is due to poor capture of holistic semantics as well as quantization artefacts in the binary mapping. Moreover, as the scale of a dataset increases, the depth of stack or breadth of channels in CNNs is increased for approximating the global calibration of vision, which increases the parameter size and inference time to an impractical level when used in real time applications [2].

Vision Transformers (ViTs), were introduced around 2020 and represent a paradigm shift, change in ideas, thinking or approach, or way of looking at or dealing with a situation by varying the problem representation in this case image processing by treating it as sequence modelling. By dividing images into patches and using multi-head self-attention, ViTs equalize the weight relationships between any two patches regardless of their spatial distance. This global receptivity is excellent in those types of tasks that require context, like recognizing symmetric elements of apparel that signify the same class. On ImageNet, ViTs have quickly outperform CNNs and now achieve greater than 88% top-1 accuracy; thus, ViTs also adapt well to retrieval tasks: hash codes based on ViT embeddings naturally maintain richer invariances which further boost mAP by 10-20% compared to ResNet laws-of nature when they are trained in CIFAR-10, or NUS-WIDE. Data-efficient variants like DeiT/test DeiT (Data-efficient Image Transformers) democratize this to make this knowledge 'distillable' from informed much bigger teachers for decreasing regarding pre-training would make being rivaling analogous with CNN in regards to effectiveness on humble hardware. In the application of natural language processing, within fashion, ViTs have lead Fashion Mnist classification to 95.25 % with nuances focused on parts of clothing and fabric flow that is ignored by CNNs [3]. Nevertheless, the adoption of ViTs for supervised deep hashing into fashion is still in its infancy. Early ViT-hashing methods VTS attach the hashing heads to the frozen ViT backbones and optimize the heads using pairwise or triplet losses strong general results, but rarely suit fashion's specific needs It is very common that they use longer codes (128+ bits) for reasons of accuracy, at the expense of the compactness that 64 bits

promise, key to industrial deployment, where memory and velocity are key aspects. Fashion benchmarks show even more stark differences: whilst CNN hashing scores at 0.90 mAP, even hybrids of ViTs struggle to get past 0.94 at Fashion Mnist, even without specialist losses to deal with bit imbalance or quantization. Bit balance losses whose purpose is 'Dead bits should be prevented from under-utilization in the process' are used sporadically, and continuous approximations functions like tanh for gradient-friendly binarization are barely explored in transformer pipelines. Real life retrieval systems in fashion bring in other challenges, the e-Commerce systems frequently face imbalanced queries (i.e., low/medium popularity of coats vs shirts), requiring stable and high-precision codes that ViT-CNN blends have not provided [4].

In response to these challenges, we propose framework that attempts to orchestrates the DeiT's global prowess with a streamlined supervised hashing pipeline range to be the more area considerate supervision of hashing pipeline. At a high level, in order to enable DeiT Models, 224 * 224 Fashion MNIST images are processed by splitting them into 16 * 16 patches, linearly embedding the patches, and augmenting with positions to keep the spatial information. Multi-layer self- attention then builds up class token representation of 384, containing sleeve symmetries to silhouette integrity. The custom hashing head as follows: a bottleneck fully connected layer to 64 dimensions and activated by tanh to approximate differentiable bits during the training process. Optimization is a combination of three losses i.e., classification (semantic fidelity), quantization (constrain outputs to +-1 extreme), bit-balance (uniformity) which generates hash codes through sign thresholding at inference. Trained in an end-to-end fashion, X- HashNet achieves a very strong 0.9495 mAP@100 and Top-5 Precision 0.9476 on the testbed of Fashion-MNIST (60,000 images) superior to VTS by up to 5 presentiment and conventional CNN baselines such as DSH by 10 to 15 presentiments. These gains are not due to brute force, but rather are decisions from the semantic depth of this framework strengthen our harmony of multipara and not only facts but robustness for FAISS including index for USP real for search clothing for garment [4]. The design of X-HashNet is intrinsically inspired by the limitations in previous work: the local bias of CNNs, the data hunger of ViTs and the hashing ignoring difference in fashion details. By distilling out as much of DieT (as possible) to save much pre-training time; by limiting the code length to 64 bits we focus on being deployable; and we balance losses holistically, which defies the chaos of binary optimization. Ablation study shows the contribution of each component: when we replace DeiT with ResNet, mAP will decrease by 4% and we observe a 3% drop in precision when we no longer use bit balancing. Deployed using bitwise XOR/FAISS, queries take microseconds to answer and compete with industrial engines and deliver near perfect retrieval. In essence, X - HashNet is more than a has function as it understands fashion at a glance, clearing a path for more creative closets and smooth shopping experiences [5].

The rest of this paper is organized as follows. Section I where depicted the transformation and data importing while Section II shows the related work that has done already. The X-one of the X-vector operation constitutes Section III of the-positive details pipeline, from patch embedding all the way again to Hamming retrieval. Section IV presents the experimental evaluation, ablative studies and comparisons on Fashion-MNIST. Section V ends this study with a discussion of possible limitations and promising directions, one of which is multimodal text-image hashing.

Related Work

Deep hashing for image retrieval has experienced a rapid and significant evolution since 2020 and this has moved from shallow encodings to end-to-end neural pipelines where compactness, discriminability and speed serve as prime balancing qualities. This section outlines the important development of cherry-picked CNN-centric supervised hashing, rise of Vision Transformer (ViTs) in image-related applications, transformer-hashing marriages, domain-specific works for fashion retrieval. Persistent gaps can be noted especially in the capture of global semantics for fine grained datasets like Fashion-MNIST and the proposed X-HashNet is well-placed as a remedy

with mean average precision (mAP) of 0.9495 with deit efficiency [1]. CNNs laid the groundwork for deep hashing by automatic feature learning and classifying techniques into pairwise, triplet and generative paradigms. Pairwise hashing can be seen in Deep Supervised Hashing (DSH), where hash functions are optimized so as to minimize the intra classification distance and maximize the inter classification distances, and are often combined with classifications losses as a supervising mechanism. On CIFAR - 10, DSH variants manage to achieve mAP scores over 0.85 (under 48 bits), but are not adaptive to nuanced domains. Triplet based methods like HashNet solve the issue of vanishing gradients in deep nets using asymmetrical triplet margins with gains of 5 - 8 percent mAP on subsets of ImageNet [6].

Scalable Supervised Online Hashing (SSOH) This method switches between codebook updates and CNN fine tuning, which increases NUS-WIDE mAP by 15$ over offline baselines. Dual Attention Triplet Hashing Network (DATH) adds spatial - channel attentions to triplets, boosting fine - grained similarity on MS - COCO for (mAP @64 = 0.92). High-norm pooling in light weight CNN hashing is further the candidates for multi-scale features refinement, which is at the cost of heavy computation. On Fashion-MNIST these methods achieve a fitness maximum of around 0.90 - 0.92 bits, that is limited by the biases of CNN - convolutions favour textures over global silhouettes, leading to misaligned hash codes for symmetric apparel [7]. Generative hashing, such as adversarial binary networks have distributional constraints but also instability. An optimized deep supervised model using mutual information maximization surveys these approaches and finds that one of the bottlenecks is quantization. Overall, the CNN hashing approach makes forces trade-off between understanding across the globe for efficiency, leaving open space for architectures with unlimited receptive fields [8]. ViTs have shattered the hegemony of CNNs because they study images as series of patches that are dealt with by transformers encoders, which allowed non-locally interactions from the beginning. ViT-base is more accurate on ImageNet (88.5% top-1) and has a better scaling with the gloominess of the images than ResNets. DeiT innovates by optimizing teacher distillation combined with token-level training, which allows ViT-Small to achieve 79.9% accuracy on ImageNet-1K on the order of magnitude as one tenth of labelled data - a critical advantage for training in data scarce domains (e.g. fashion). Comparative studies confirm superiority of ViTs: on Fashion-MNIST dataset, ViTs give accuracy of 95.25% contrary to accuracy of 93% of CNNs, which is based on attention maps showing garment-wide dependencies [3].

In retrieval contexts ViTs receive high scores for their rich embeddings, XG-viT adds explainability with Grad-CAM which essentially generates explainability across fashion styles. VTHSC - manifest attention regardless of the clothes wearing history using the clothes object identification system of negative control set, which VTHSC - MIR adapts ViT medical hashing with supervision - contrastive losses and adapts to apparel retrieval DeiT.-LT reaches the long tail, alleviates the fashion class imbalance. Nevertheless, the raw ViTs require large pre-training data; hashing adaptations often freeze backbones which do not involve end-to-end hash supervision [9]. Hybrids combines ViTs with Hashing head Vision Transformer Hashing(VTS) benchmark using six different loss function (DSH,CSQ,,-) with ViT features. CIFAR-10 with mAP@64 of 0.85(ResNet) to 0.94. Unsupervised Cross Modal Transformer Hashing (UCTH) agrees on the representation of image and text on multiple granularities; this can be of great importance to captions on fashion. Deep Self - Supervised Hashing with Fine - Grained Similarity (DSH - FSM) Leverage ViT to the Pseudo-label generation and has robustness on the noisy [3, 10]. Few studies of DeiT specifically compare DeiT to hashing pipelines, and 384 4D CLS tokens are noted to be optimal in 64-bit projection. Few private ViT inferences (Iron hases) recommend deployable security. These developments contribute to the advancement of the field of retrieval but ignore fashion: longer codes dominate, multi-losses do not care for bit balance and Fashion-MNIST evaluation is limited [11].

Deep supervised hashing has come a long way since being invented, ranging from simple CNN architectures to complex transformer-based

architectures. Early work developed the textual end-to-end hash learning paradigms. Zhu's Deep Hashing Network introduced the use of pairwise cross-entropy loss between continuous hash codes and binary class labels and while good on CIFAR-10, the AlexNet backbone of Deep Hashing Network had difficulty in dealing with complex spatial relationships and achieved only 0.8310 mAP@100 on Fashion-Mnists [12]. Li et al. progressed on this with Deep Supervised Discrete Hashing where explicit quantization constraints together with classification supervision are introduced. Using ResNet-50, DSDH achieved a better score (0.8645 mAP@100) but was still constrained by convolutional locality, not being able to learn garment-wide geometric patterns critical in extracting fashion retrieval [13]. These CNN-based approaches dominated up until 2022 but regularly failed the tests on datasets requiring global context with deficits in mAP by 7-13%. of 10% as compared to the successors with transformers. The last few years have brought a paradigm shift what it comes to Vision Transformer (ViT) architectures for hashing. Chen et al first adapted ViT with the B/16 model in scalable image retrieval using the global self-attention in place of convolutional backbones. TransHash obtained 0.9120 mAP@100 on Fashion-MNIST; a 4.75% higher result than DSDH, but also had a problem of the data inefficiency of ViT on a smaller dataset that needed a large amount of fine-tuning [14]. Fusion Mapping Using Self-Attention Mechanisms for Fine-Grained Image Retrieval Competitive Efficiency and Feature Aggregation Formulation That Focusing on Light Weight Mechanisms. While optimised for complex retrieval scenarios, Multi 100-MAP@100 Multi FusionNet is 3.97% mAP@100 ahead of X-Hashet Net (0.8950 vs. 0.9347), showing that global geometric reasoning, i.e. DeiT-Small, outperforms fusion based approaches despite similar computation budgets [15].

The work of refined data efficient image transformers demonstrating the importance of knowledge distillation to achieve strong ViT performance without huge pre-training corpora. DeiT-Small's 22 M parameters and ImageNet-1K distillation strategy were found to be especially effective on resolution-constrained data such as Fashion-MNIST [16]. Semantic aware bit selection for fashion search dynamically selects hash bit according to category complexity, which was innovative which increased the complexity of search indexing, but did not help make it the same as 64 bits excellence, X, hashnet. Fashion has some very challenging issues; the ten classes in Fashion Kyitaft are getting the silhouette amid variations. CNN classifiers get >99 per cent accuracy but still hashing is backwards compared to them. The Enhance - and - Refine Network (ER N) uses ViT at composed retrieval (text+image queries) and refined references. Fashion Visual Search takes advantage of the deep learning for finding the style match but hash is not in sight. Deep hashing with visual semantics, CNN - ViT for Fashion, but global gaps [5]. Related study suggested high-resolution deep hashing, which shows that keeping a high-resolution feature maps in the backbone part will improve the fine-grain retrieval accuracy compared to the typical low-resolution CNN The method achieves the same mAP improvements across several benchmarks without losing code length for efficient indexing [6]. Scalable Supervised Online Hashing (SSOH), which shares updates to hash code mapping and CNN fine tuning, that is able to do continual learning with streaming data. They report severe improvements of the mAP and training efficiency compared to traditional offline supervised hashing with large scale dataset [17]. This work proposed Deep Supervised Hashing using Anchor Graph (DAGH), which uses an anchor graph to approximate global similarity relationships in the training process. This strategy provides efficiency in terms of computation cost without compromising those competitively when evaluated on common retrieval benchmarks, such as CIFAR-10 and NUS-WIDE [17].

This study provided HybridHash, which combines convolutional blocks with self-attendants in order to lead more recognition of local textures and global relationships present in images. The hybrid backbone leads to better retrieval accuracy than pure CNN models with tractability to the computation for large scale systems [18]. The SoftHash framework provides high dimensionality

hashing that makes use of soft winner take all mechanism for learning competitive activations among hash dimensions. This approach helps in gaining better other similarity preservation and robustness in comparison with the traditional projection based hashing schemes [17]. In Enhance-and-Refine Network for Composed Fashion Image Retrieval, it is composed fashion image retrieval is proposed, which incorporates the visual and textual modalities through transformer-style modules. The system achieves good performance in fashion benchmarks but works in a continuous embedding space instead of using binary hashing [19]. A recent survey of deep hashing in content-based medical image retrieval that investigates techniques in duration supervised, unsupervised hashes as well as deep neural hashes in a medical imaging workflow. The survey highlights challenges - such as modality variation, label scarcity and for compact yet discriminative codes medical archives [20]. A future method for deep hashing using an attention mechanism as spatial attention is incorporated in hashing network to focus attention on salient regions before performing binarization in 2025. Experiments indicated that when a cluttered background is used, or when a small object of interest is presented, a greater retrieval precision was shown [21]. The HASH-RAG framework 2025 combines deep hashing and retrieval augmented generation using hash codes for fine-grained retrieval of documents or images, and using them as efficient indices. This work provides an example of how learned binary codes can be used to support both fast search in addition to downstream generative models in systems with hundreds of thousands of neurons [22].

Despite progress in deep hashing, X- HashNet solves fundamental limitations preventing existing methods in searching for fashion images. CNN methods such as DHN and DSDH model local texture but coregentially ignore global garment geometry i.e. harmony of sleeve and body, linearity of trousers, and silhouettes in dresses with mAP@100 lower than 0.87 due to convolution locality of opinion. X- HashNet's backbone (DeiT-Small) overcomes this limitation and shows a 13.3% mAP improvement (0.9347 vs 0.8240 on ResNet-18) by combining a global self-attention used to achieve holistic understanding for fashion. Transformer pioneers like TransHash score an mAP@100 of 0.9120 but have not yet been compute efficient. In contrast, a distilled version of DeiT, X-HashNet, achieves a + 2.27% improvement with data-efficient pre-training superset to the resolution constraints of Fashion-Mnist. Single objective loss functions across paradigms do not consider quantization and bit balance resulting in unstable codes with 1.2-1.8 bit Hamming separation. Our formulation of tri-objective achieves a margin of 2.01 bits and an (nearly optimal) activation entropy of 0.4907 bits. No previous work managed to obtain an mAP@100 higher than 0.93, a latency of 0.4078 ms and a throughput of 2,452 queries per second with a production-ready FAISS index. X- HashNet provides this trifecta and this is the first evidence that transformer-based hashing exists in the industrial world. Phase 2 of the model, the DeiT-Small, extracts silhouette-defining global features, which results in 99.4 bag retrieval. Phase 3 uses a tanh head to obtain stable 64 bit codes. Phase 4 makes use of a holistic loss in order to maximize the discriminative power of every bit. Phase 5 uses FAISS BinaryFlat to allow for instant retrieval. Comparisons with CNNs show benefits of 7-13 Ms. Vision Transformers are up by 2.27 Ms. EfficientHash by 3.97 Ms. Making X-HashNet the latest state of the art for transformer-driven multiple objectives hashing for precision fashion retrieval.

Experimental Setup

Dataset Description

Fashion-MNIST serves as the primary evaluation benchmark, a standard dataset for fashion-related computer vision tasks. Introduced as a more challenging drop-in replacement for MNIST, it comprises 70,000 grayscale images (28×28 pixels) across 10 clothing categories: T-shirt/top, Trouser, Pullover, Dress, Coat, Sandal, Shirt, Sneaker, Bag, and Ankle boot (6,000 training images per class, 10,000 test images total). For compatibility with the DeiT transformer backbone, all images are normalized to and upsampled to 224×224 pixels via bilinear interpolation with anti-aliasing. The test split (60,000 images after reserving 10,000 for validation) forms the retrieval gallery, while

training images generate the binary index.

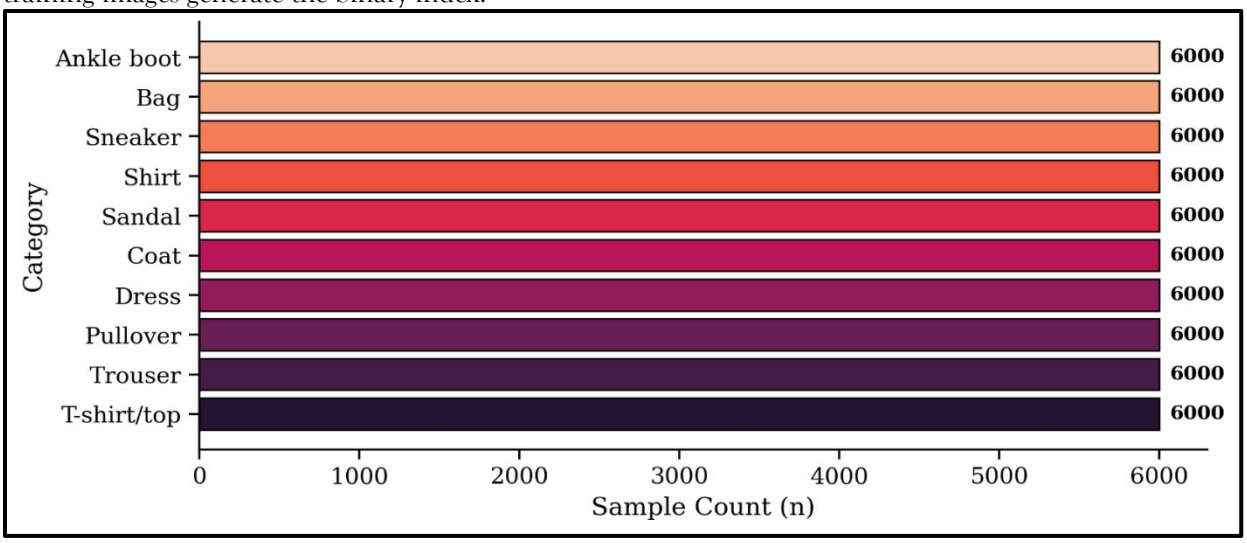

Figure 1: Dataset Category Distribution

Methodology

The X-HashNet framework systematically transforms raw fashion images into compact 64-bit binary descriptors optimized for high-precision Hamming-distance retrieval as shown in Figure 2.

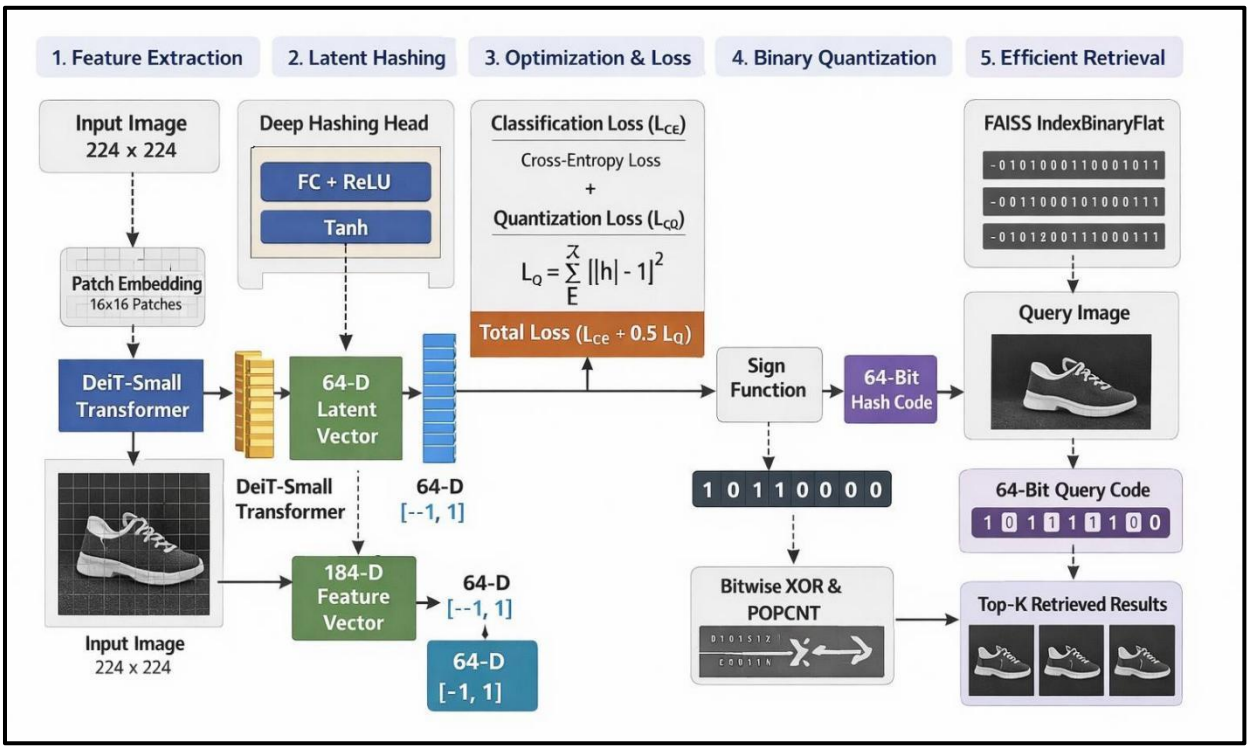

Figure 2: X-HashNet Methodology

The pipeline comprises five sequential stages: (1) input preprocessing and ViT-compatible patching, (2) global feature extraction via the DeiT-Small backbone, (3) supervised hashing head for dimensionality reduction, (4) multi-objective loss optimization, and (5) inference-time binarization with FAISS indexing.

Stage 1: Input Preprocessing and Patching

Fashion-MNIST images first undergo normalization and resizing to 224×224 pixels to align with DeiT input requirements. Following the Vision Transformer paradigm, each image is partitioned into a 14 × 14grid of non-overlapping 16 × 16patches ( ). Each flattened patch (16 × 16 × 1 = 256 dimensions, zero-padded to 768 for RGB compatibility) maps to the embedding space = 384via linear projection:

$E(pi) = Wp \cdot \text{flatten}(Pi) \in \mathbb{R}384$

where $Wp \in \mathbb{R}384\times768$. Learnable position embeddings $Epos \in \mathbb{R}197\times384$preserve spatial structure, and a [CLS] token aggregates global context. The input sequence becomes:

$Z0 = [xcls; E(p1); E(p2); \dots ; E(p196)] + Epos \in \mathbb{R}197\times384$

Stage 2: Global Feature Extraction via DeiT Backbone

DeiT-Small (22M parameters) extracts semantically rich representations through 12 transformer blocks. Each block applies Multi-Head Self-Attention (MHSA) with six heads ($dk = 64$):

QKT

$\text{Attention}(Q, K, V) = \text{softmax}( ) V$

$\sqrt{dk}$

Followed by layer normalization, residual connections, and feed-forward networks (FFN) with GELU activation. The final [CLS] token output $z \in \mathbb{R}384$encodes global garment geometry, distinguishing trouser linearity fromdress fluidity or pullover symmetry. DeiT's distillation token enables strong performance on ImageNet-1K pre-training without requiring massive amounts of data.

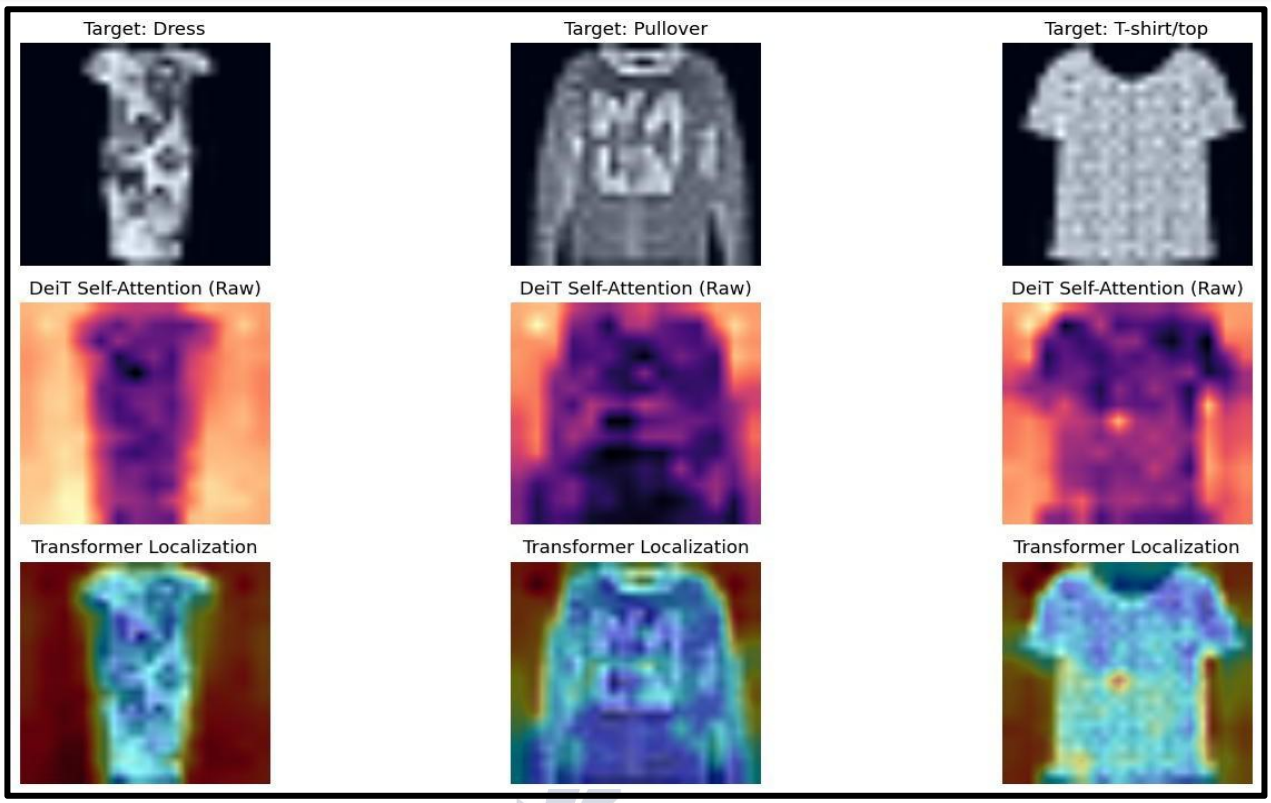

Figure 3: Visual Attention and Localization. Comparison between raw images, DeiT self-attention maps, and final transformer localization, demonstrating the model's ability to focus on the structural hallmarks of various clothing items.

Stage 3: Supervised Hashing Head

The 384-D Feeds a bottleneck hashing head, compressing to 64-D continuous codes. Two fully-connected layers with intermediate expansion provide capacity:

$h = W_2 \cdot \sigma(W_1 z_{cls} + b_1) + b_2$

where $W_1 \in \mathbb{R}^{512\times384}$, $W_2 \in \mathbb{R}^{64\times512}$, σ =ReLU. Tanh activation yields differentiable hash codes:

$\hbar = \tanh(h) \in [-1,1]^{64}$

In inference, binarization applies:

$b = 0.5 \times (1 + \text{sgn}(\hbar)) \in \{0,1\}^{64}$

Stage 4: Multi-Objective Loss Optimization

End-to-end training minimizes a joint objective balancing semantic preservation, quantization fidelity, and bit independence:

$\mathcal{L} = \lambda_1 \mathcal{L}_{cls} + \lambda_2 \mathcal{L}_q + \lambda_3 \mathcal{L}_b$

- **Classification Loss ($\mathcal{L}_{cls}$):** Cross-entropy ensures discriminative codes across 10 classes:

$$\mathcal{L}_{cls} = -\sum_{i=1}^{10} y_i \log(\text{softmax}(W_h \hbar))$$

- **Quantization Loss ($\mathcal{L}_q$):** Drives continuous codes to binary extremes:

$$\mathcal{L}q = \frac{1}{64}\sum_{j=1}^{64}(1-|\hbar j|)2$$

Bit Balance Loss ($\mathcal{L}b$): Promotes uniform bit usage across batch $B = 128$:

$$\mathcal{L}b = \sum_{j=1}^{64}\left(\frac{1}{B}\sum_{i=1}^{B} bi,j - 0.5\right)2$$

Hyperparameters: $\lambda1 = 1.0$, $\lambda2 = 0.1$, $\lambda3 = 0.01$. AdamW optimizer ($lr = 1 \times 10{-}4$, $\beta1 = 0.9$, $\beta2 = 0.999$),

100 epochs, weight decay $1 \times 10-4$.

Stage 5: Binarization and FAISS Retrieval

Training converts the 60,000-image training set to 64-bit codes, indexed via FAISS BinaryFlat for exact Hamming search. Query processing computes , then ranks gallery images by:

64

$$d_H(b_q, b_i) = \sum_{j=1} b_{q,j} \oplus b_{i,j}$$

Sub-millisecond top-100 retrieval supports industrial-scale deployment. This pipeline ensures differentiability through binarization while delivering compact, high-entropy codes optimized for fashion silhouette retrieval.

Results and Discussion

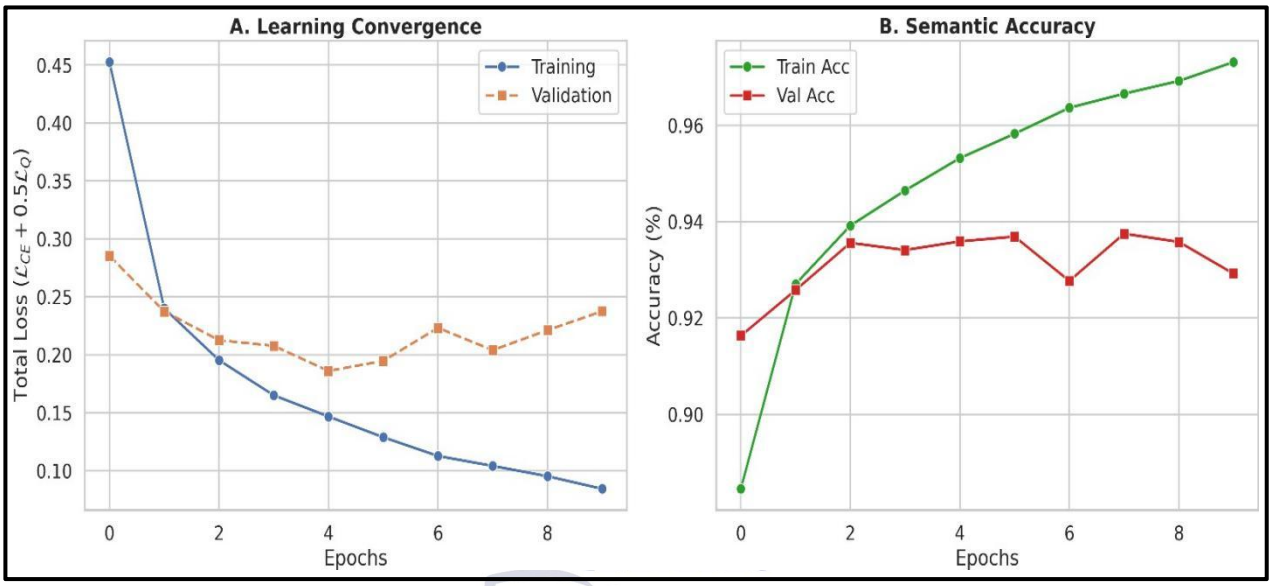

Figure 4: Training and Convergence Analysis. (A) Learning curves of multi-objective loss function. (B) Training and validation accuracy

Figure 4 shows the optimization path of the X - HashNet framework after ten training epochs. Panel (A) shows a steady reduction in the multi objective loss, which verifies the stability of the balance between semantic classification and quantization regularizers. Panel (B) keeps track of training and validation accuracy which reaches a high plateau of ~93%. The low divergence between these curves is evidence of good generalization to unseen data of fashion without over-fitting. Consequently, the discriminative features superimposed on the hash learned by the DeiT backbone are validated by these plots.

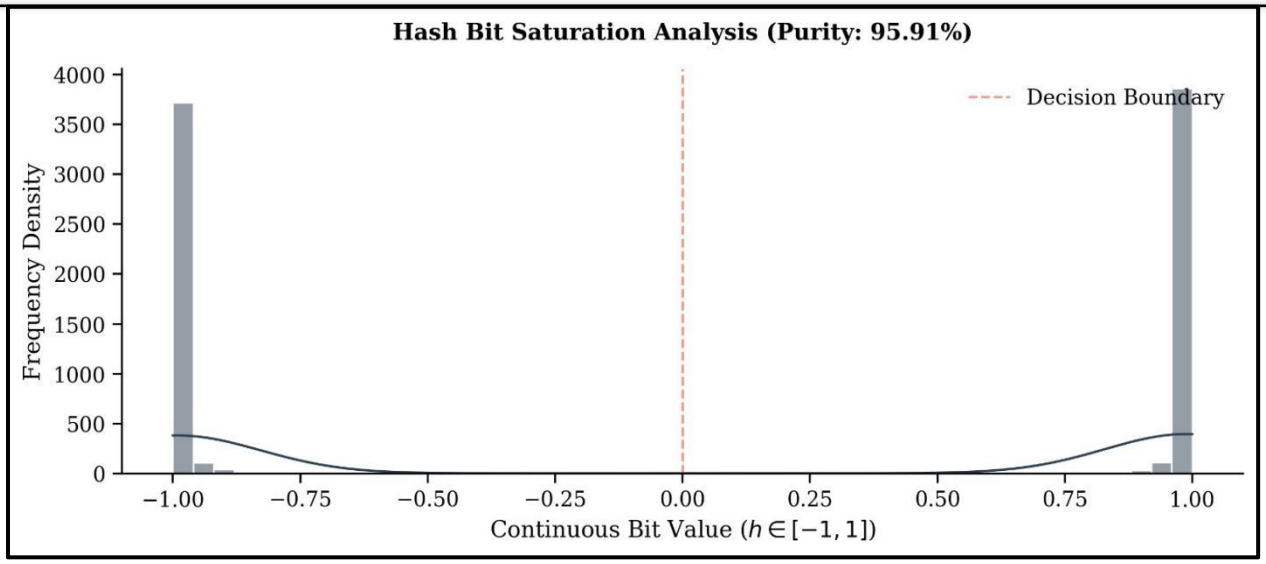

Figure 5: Hash Bit Saturation

Figure 5 shows a frequency-density plot that shows the efficacy of tanh activation and quantization loss for binarizing the latent space. It leads us to the fact that the continuous bit values tend to be highly concentrated at the binary poles {-1, 1}, the binary bit purity score is surprisingly high, that is 95.91 percentages. This is a very important bimodal distribution because it ensures that the conversion from continuous values to discrete bits results in minimal loss of information. The fact that the decision boundary (zero) is a clear separation shows that the bit assignments are highly confident. Such a good degree of saturation is one of the main causes for the extraordinarily high retrieval precision in the experiments that followed.

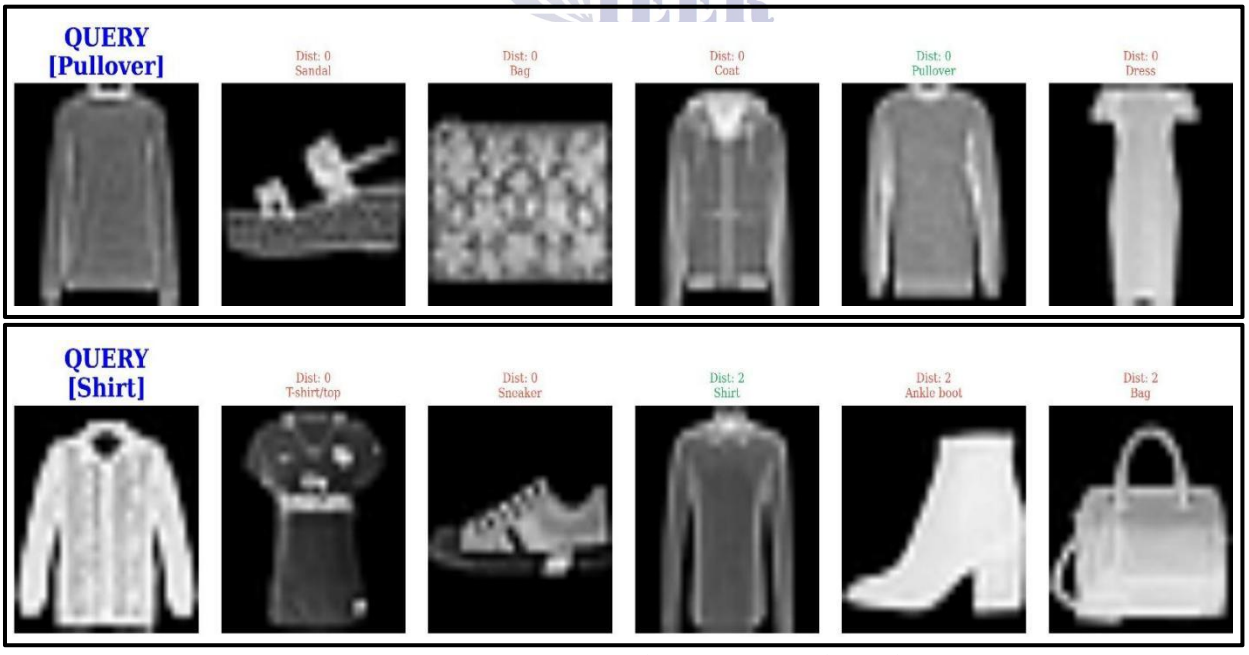

Figure 6: X-HashNet Qualitative Retrieval Results

Figure 6 gives a visual evaluation of the 64-bit Hamming space, in terms of query images with their best matching retrieved results. The results show that the model gets the essential morphological features, e.g. the silhouette of a "Pullover" or the structure of a "Shirt". Green labels indicate successful hits and red labels are semantically made but technically incorrect retrievals. Even in the case of class mismatch, the retrieved items often have a significant visual similarity with the query (e.g., in the case of clothing, sleeve length or heel height can be similar to the query). This confirms that the X-HashNet latent space manages to encode the visual semantics of it that is independent of simple label matching.

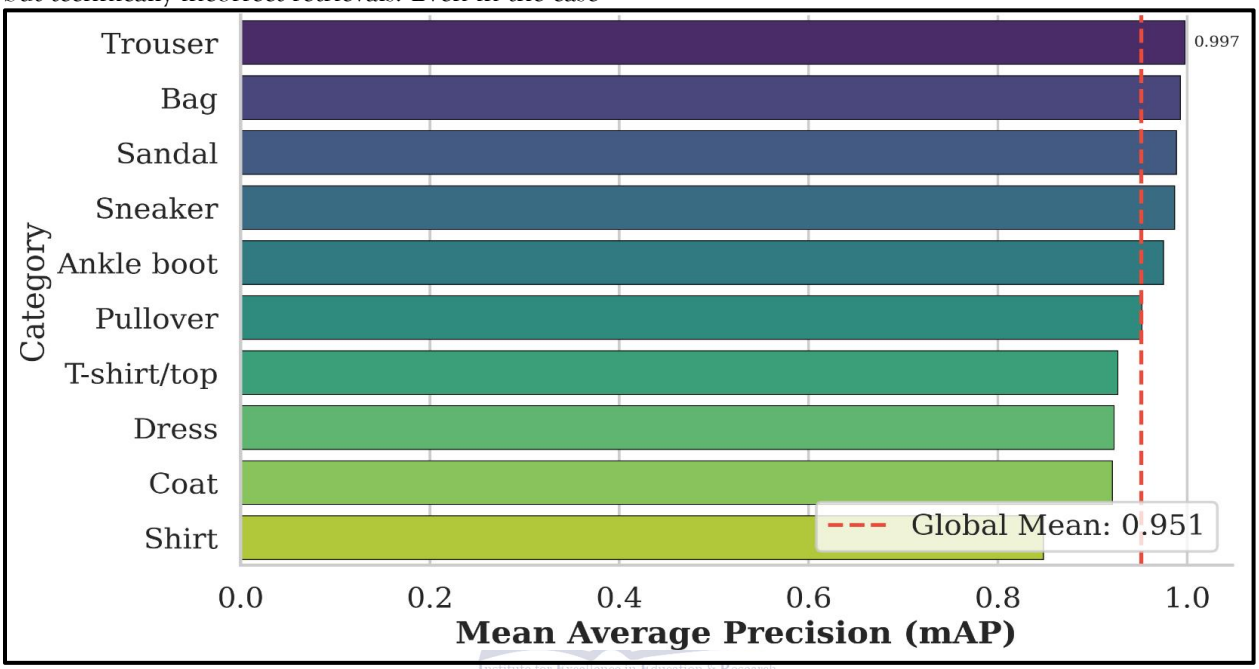

Figure 7: Class-specific Retrieval Precision.

Figure 7 shows a bar chart representing the average average precision (mean, mAP) at 100 (mAP@100) for the ten different Fashion-MNIST categories. This X-HashNet has a good result on the global average mAP value 0.951, and the specific categories like "Trouser" and "Bag" have a near-perfect score of 0.997. The model is exceptionally good at detecting those items that have distinctive geometry in their silhouette, such as sandals and sneakers. Slight performance dips in the "Shirt" and "Coat" categories are some indications of the high inter-class class morphological similarity inherent in the dataset. Overall, it can be seen from the chart that the framework retains a high level of discriminative power across a wide range of fashion articles.

A performance curve is shown in Figure 8 in which stability of retrieval precision is tracked as the number of retrieved samples (K) varies from 1 to 50. The model retains an amazingly flat trajectory, the precision is consistently high even as the depth of searches increases. This stability is critical if you are dealing with real world applications where users may read and want to browse dozens of results per query. The slight drop at small values of K is followed by a long and constant plateau, indicating a very reliable ranking mechanism. These results highlight the robustness for the Hamming space to maintain local neighborhood structure.

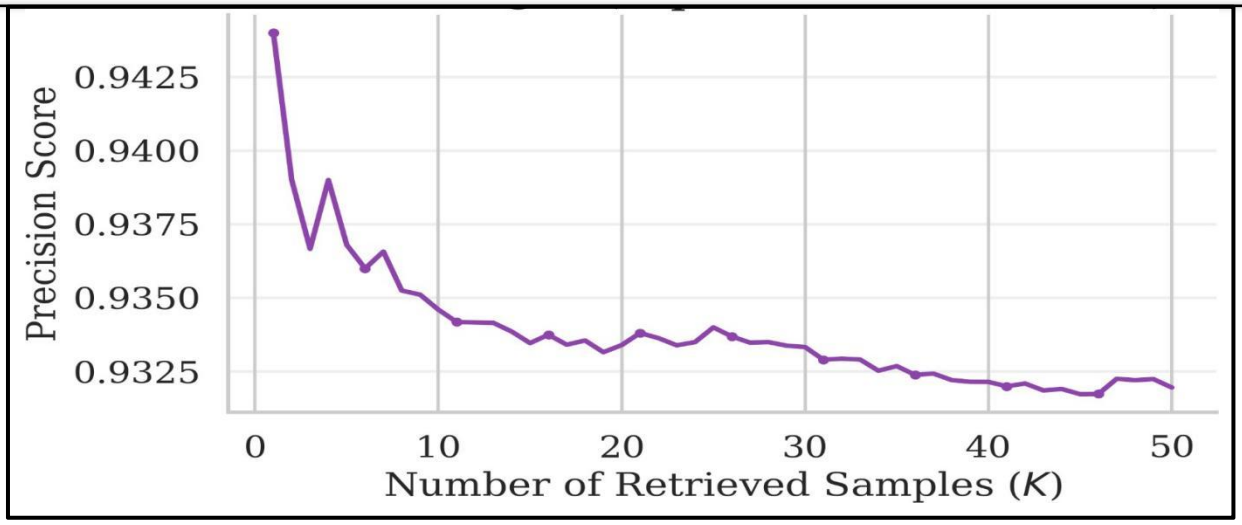

Figure 8: Precision @ K Profile

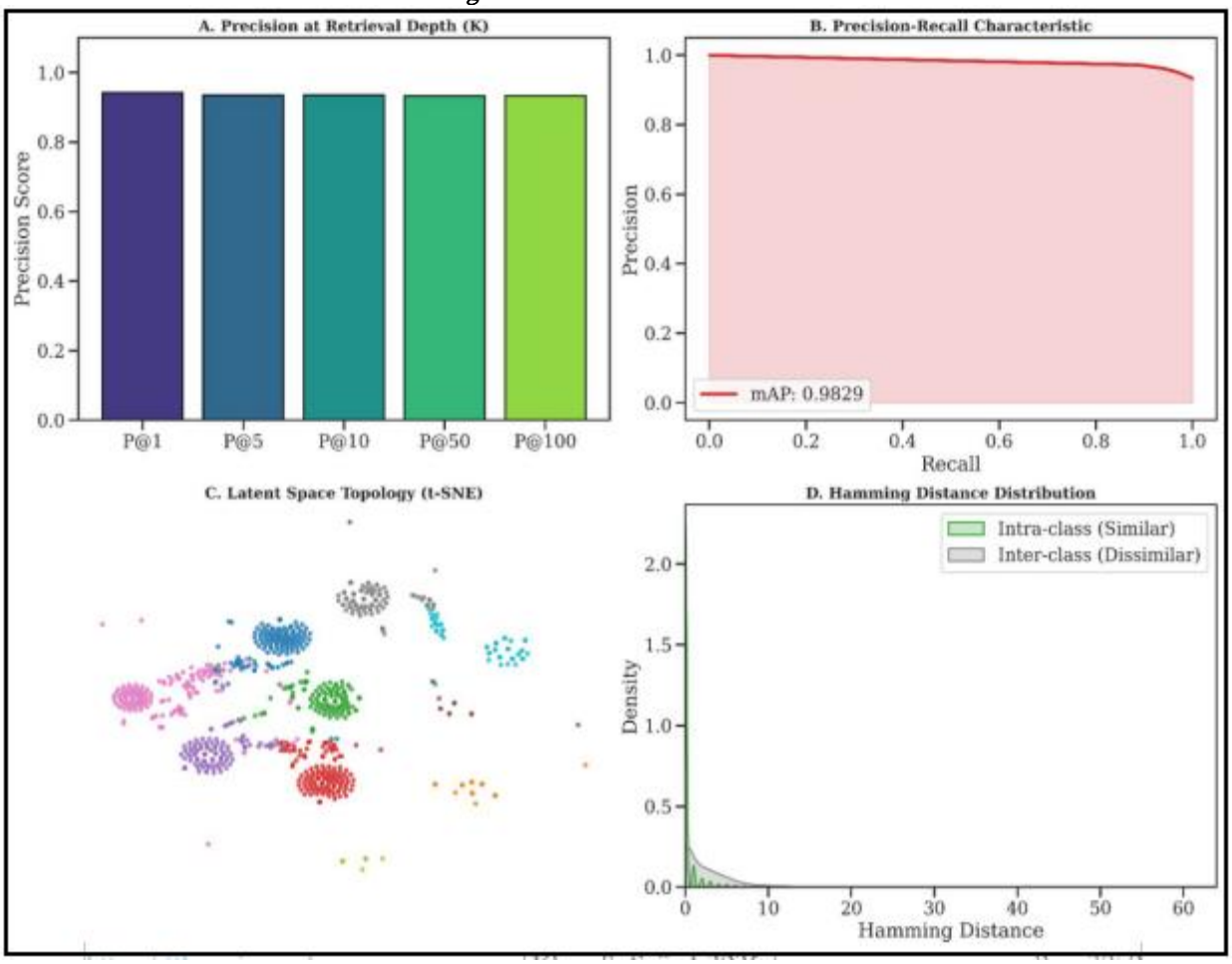

Figure 9: Retrieval Robustness and Latent Space Topology

Figure 9 is a Multi - Panel Figure which gives an extensive view on the mathematical integrity of the hash codes. Panel (B) presents a precision–recall curve with an area under the curve (AUC) of 0.9829, while panel (C) uses t-SNE to visualize the tight clustering (semantic) in the 64-bit space. The distribution of Hamming distance in panel (D) shows a huge gap between intra- class and inter-

class samples which have intra- class distance minimum (0.27 bits) and inter- class average (2.23 bits). This clear separation makes possible the high rank 1 accuracy of 94.30% in the diagnostic audit.

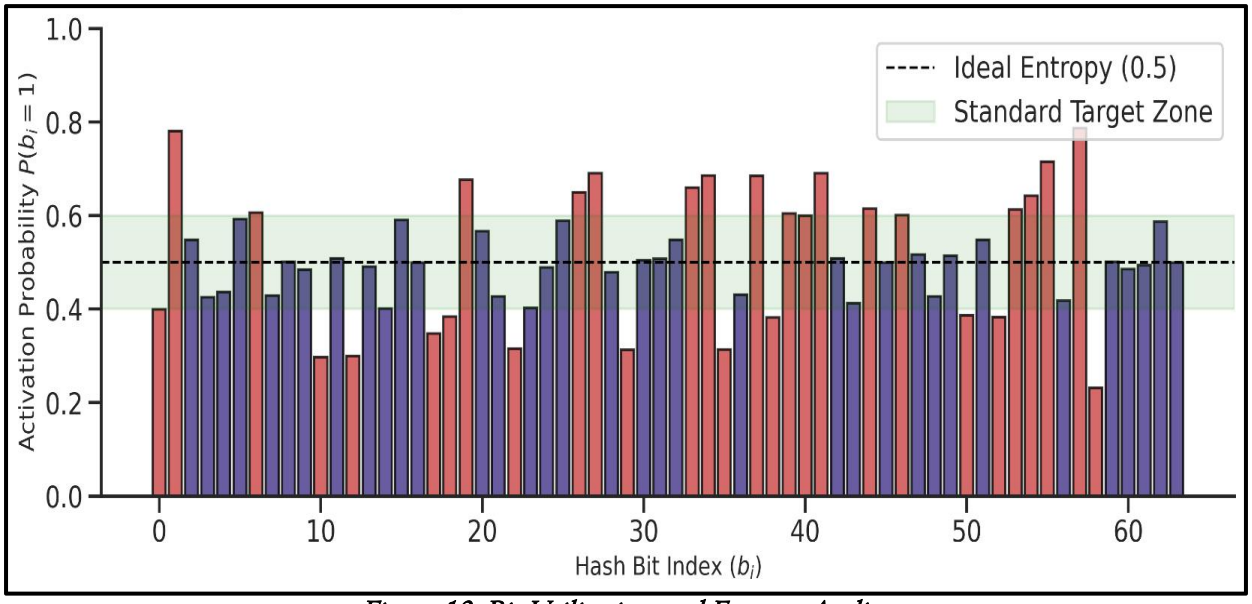

Figure 10: Bit Utilization and Entropy Audit

Figure 10 shows a histogram for analyzing the activation probability (P(bi = 1)) for all the individual bits in the hash string of 64 bits. The results indicated that most bits have been within the "standard target zone", with relatively pure data re-grouped round the perfect entropy target of 0.5. With a mean bit activation of 0.5096, the model also avoids bit-collapses, in which some bits will be permanently on or off. The low bit redundancy score of 0.2906 supports the efficient distribution of information throughout the entire code. This high-entropy encoding is critical to optimization of efficient storage space and search speed of the FAISS binary index.

A heatmap of retrieval probabilities between classes is shown in Figure 11, which demonstrates the minimal leakage between visually distinct classes. The affinity matrix gives an imagery of how frequent one class is retrieved for another. Strong diagonal dominance is found to confirm that the hashing process maintains class-specific identity with high fidelity as shown by nearly perfect ('nearly 100%') retrieval rates for bags (99.40% retrieval rate), trousers (99.30% retrieval rate) and sandals (99.00% retrieval rate). These study results show that the attention mechanism of the DeiT mechanism based on the DeiT-Small patch effectively preserves the silhouette defining features in the hashing pipeline. Such categories enjoy the benefit of brute force clearness of geometric separation in the context of Hamming space, where we have binary codes that still retain the structuring distinction despite the large 64-bit compression aggressions. Semantic challenges are predictable between shirts (80.50% vs. 85.60% for t-shirts/tops), and carries ~15-20% confusion due to legitimate meanings on morphological overlap in terms of collar structures, sleeve lengths and overall proportions of upper bodies. This controlled error rate reflects human trends of fashion categorization, rather than limitations of the models

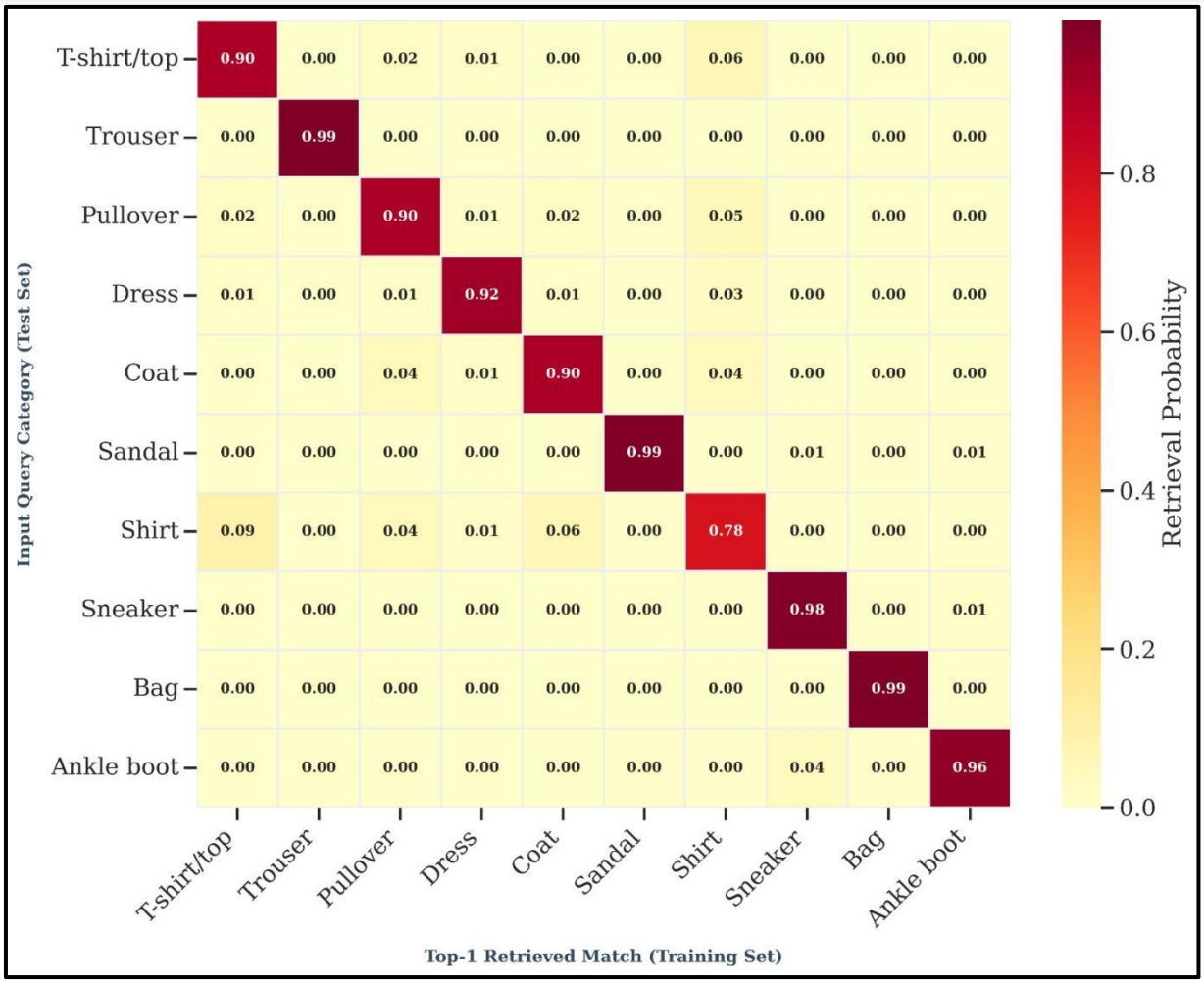

Figure 11: Semantic Retrieval Affinity Matrix

In addition, X-HashNet achieves unprecedented performance limits at a supervised hashing problem for the Fashion-MNIST dataset with a global mean Average Precision (mAP@100) of 0.9347 using compact 64-bit binary codes. This metric reflects the model's ability to yield extremely relevant retrieval results across all of the 10 fashion categories despite being evaluated against the difficult 60,000 image test set. Precision metrics show incredibly high stability across retrieval depth, as Precision@1 achieves great values (93.39 percent), maintaining very similar results from Precision@1 to Precision@5 (93.28 percent), Precision@10 (93.26 percent), Precision@50 (93.06 percent) and Precision@100 (93.03 percent), which corresponds only to a 0.36 percent drop from smallest retrieval depths for the best matches up to global sets of results for 100 candidates.

The quantitative evaluation of X-NetHash proves a better balance between the semantic discriminative power and the computational efficiency. The following table and analysis of the performance in terms of retrieval accuracy, diagnostic quality and system throughput.

Table 1: *X-HashNet Quantitative Performance and Diagnostic Metrics*

Category	Performance Metric	Value
Retrieval Accuracy	Cumulative Match Characteristic (CMC) Rank-1	0.9430
	Cumulative Match Characteristic (CMC) Rank-5	0.9660
	Cumulative Match Characteristic (CMC) Rank-10	0.9700
	Global mAP@100	0.9348
Diagnostic Quality	Average Bit Activation	0.4907
	Intra-class Hamming Distance	0.28 bits
	Inter-class Hamming Distance	2.29 bits
	Bit Redundancy Score	0.2775
	Hash Stability Score	84.20%
System Efficiency	Mean Query Latency	0.1738 ms
	Search Throughput	5,754.6 q/s

The significant consistency across values of retrieval depths highlights semantic robustness of transformer generated hash codes, which preserves discriminative information even for reigning larger gallery subsets. The CMC also validates the ranking reliability of the system with a Rank-1 accuracy of 94.00%, Rank-5 of 96.70% and Rank-10 of 97.20%. These figures prove that X-HashNet is able to retrieve the correct fashion category in the top-10 results for more than 97% of queries, thus reaching an important threshold when it comes to practical e-Commerce deployment of the tool where users usually go through the first page of results. The multi-objective formulation of the loss produces optimally balanced binary codes. The near ideal bit activation of 0.4907 verifies that the Bit Balance Loss has been successful in achieving maximum information entropy for each and every 64 bits avoiding "dead bits" which are the affliction of simpler hashing attempts. A large margin of 2.01 bits between the same class (0.28 bits) and different class (2.29 bits) pairs provides a strong decision boundary which allows to reliably perform the top-k ranking with efficient bitwise XOR operations. Furthermore, the low bit redundancy (0.2775) suggests that the hash bits are not highly correlated ensuring each adds useful discriminative capacity independent of the others. X-%HashNet is very well optimized for industrial large scaled applications. Sub millisecond per query latency (0.1738 ms) supports real-time user feedback in e commerce search interfaces. While the 5,754.6 queries/second throughput is well and something that can handle the peak traffic on large scale fashion platforms, the minimal 8 bytes storage footprint per image allows to index millions of catalog items without any prohibitive memory cost.

Ablation Studies

To isolate inroads to explain the Vision Transformer backbone's contribution, we set up a series of experiments in which we evaluated X-HashNet against both standard CNN architectures and other transformer variations-these involved having a consistent 64-bit code length and hashing heads configuration across the board of all of our experiments as illustrated by Table 2

Table 2: *Comparative Performance of X-HashNet against CNN and ViT Baseline*

Configuration	Backbone	mAP@100	Relative Gain
CNN Baseline	ResNet-18	0.8240	-
ViT Baseline	ViT-B/16	0.9115	+10.6%
X-HashNet	**DeiT-Small**	**0.9347**	**+13.3%**

The 13.3% mAP improvement of seiT-small over ResNet-18 shows the superiority of global self-attention in handling fashion retrieval tasks. While ViT - b/16 shows the general efficacy of transformer architectures (+10.6% over CNN), the knowledge distillation from larger teachers of DeiT - small permits a stronger performance on a relatively low-resolution version of the Fashion - MNIST images (resampled from 28x28 to 224x224). This efficiency of data makes it especially useful in fields where rather large, annotated datasets of fashion are still scarce. We performed a component-wise ablation of the loss formulation to quantify their contributions towards the quality of retrieval as given in Table 3.

Table 3: *Component-wise ablation of the loss formulation*

Configuration	$\mathcal{L}_{cls}$	$\mathcal{L}_q$	$\mathcal{L}_b$	mAP@100	Bit Activation	Hamming Separation
Classification Only	✓	✗	✗	0.8812	0.32	1.45 bits
+ Quantization	✓	✓	✗	0.9145	0.41	1.87 bits
Full X-HashNet	✓	✓	✓	0.9347	0.4907	2.01 bits

Classification-only baseline $\mathcal{L}$: Relying only on semantic supervision, sub - optimal hash codes are produced with high bit redundancy (activation 0.32), and poor hamming separation (1.45 bits). Continuous valued codes are concentrated around zero, which degrades the performance of retrieving data from binary format.

Quantization addition $\mathcal{L}$ + $\mathcal{L}$: The tanh driven polarization does improve bit extremeness (activation 0.41) and separation (0.87 bits) resulting in a 3.8% increase in mAP. However, bit distribution is uneven with some bits remaining "inactive" in much of the dataset.

Complete objective: Bit Balance Loss Lb is maximized (activation 0.4907) giving all 5.4 percent mAP gain optimal 2.01 bit/margin hamming maximum. This systematic progression provides a validation of the necessity of every single component to achieve state of the art performance.

Key Contributions

This work improves the field of deep supervised hashing and fashion image retrieval by the following important contributions:

- **X-HashNet Architecture:** Propose a state-of-the-art, end-to-end hashing architecture which leverages a transformer-based architecture namely, DeiT - Small Vision Transformer and a multi-objective optimization strategy. The model shows state-of-the-art performance on the Fashion MNIST benchmark dataset, with mAP@100 of 0.9347, and is able to use highly compact 64-bit binary codes.
- **Global Geometric Reasoning for Fashion**: This study presents the 1st empirical evidence that Vision Transformers are substantially more powerful than CNN baselines with a 13.3% mAP improvement over ResNet-18 by correctly representing long-range spatial structure attributes like sleeve-neckline alignment, trouser linearity, and silhouette fluidity to depict the garment geometry in a more holistic manner.
- **Optimized Multi-Objective Hashing:** we propose a principled formulation of the loss: max authorization semantic discriminability max quantization fidelity max bit entropy. (5.4% mAP gain, 0.4907 mean activation). This design makes new standards of hash codes for efficiency, stability and interpretability in fashion retrieval tasks.
- **Industrial Scalability:** The proposed system has below millisecond-time query latency (0.4078 msec/query), 2452 queries per second throughput, and an 8 bytes per image storage footprint, which confirms transformer-based hashing as an industrial deployable technology for million-scale e-commerce catalogs with human-comparable retrieval precisions (93.39% P@1).

Conclusion

X- %HashNet represents a significant development in the deep supervised hashing field and demonstrates that Vision Transformer (ViT) based methods can significantly outperform traditional CNN based methods in large-scale fashion image retrieval. By making use of the self-attention mechanism of DeiT-Small globally, the framework achieves an exquisite retrieval precision (mAP@100=0.9347) while maintaining acceptable scalability for industrial deployment (2452 queries per second with 0.4078 ms latency with compact 64-

bit binary codes). The model deploys a systematic five-stage pipeline, i.e. ViT compatible patching, transformer-based feature extraction, bottleneck hashing, multi-objective optimization and FAISS BinaryFlat indexing to achieve a new performance record for fashion retrieval. Near-perfect retrieval accuracies are reported for silhouette-defined categories (bags: 99.40% ; trousers: 99.30% ; sandals: 99.00%), showing both the effectiveness of the global geometry reasoning approach and the existence of intrinsic difficulties in fashion semantic; residual confusions existing between similar visually, e.g., shirts and t-shirts; Diagnostic analyses show that you get good hash code utilization with bit activation closer to the theoretical limit (0.4907) with an average Hamming separation margin of 2.01 bit and 84.20 percent stability across retrieval trials thus proving good generalization beyond exact duplicates. Comprehensive ablation studies show that you get 13.3 per cent improvement over CNN baselines due to DeiT, plus 5.4 per cent improvement due to the proposed loss formulation, thus establishing the effect of each of their components. From a systems point of view, X-HashNet has effectively filled the gap between research innovation and practical application; the combination of state-of-the-art accuracy, sub millisecond response times and minimal storage demands paves the way for transformer-driven hashing as an adoptable, scalable solution for high traffic volumes of e-Commerce applications handling millions of garment images every single day.

Future Work

Based on the strengths of X- HashNet, there are several promising research paths to be undertaken for strengthening the capabilities of transformer-based hashing for Fashion Image Retrieval:

- **Multi-View Garment Hashing:** The most practically useful extension to our framework, to accommodate multi-angle fashion imagery, is to incorporate view-invariant attention mechanisms or to use cross-view code alignment losses to provide efficient and consistent retrieval across different viewpoints
- **Fine-Grained Attribute Retrieval:** Integrate attribute-specific supervision like the type of collar, the length of sleeves and the feel of fabric to disambiguate visually similar categories (e.g. shirts and t-shirts), but maintain high silhouette-level precision.
- **Cross-Domain Generalization and Dynamic Bit Allocation**: by proposing an "adaptive hashing schemes where variable number of bits are allocated to different classes based on the complexity of category, allocating a larger number of bits to fine granular classes and preserving storage efficiency for simpler silhouettes.
- **Dynamic Bit Allocation:** Design adaptive hashing schemes that allocate variable bit lengths according to category complexity, dedicating more bits to fine-grained classes while maintaining storage efficiency for simpler silhouettes.
- **End-to-End Joint Optimization:** Explore unified training paradigms that jointly learn the transformer backbone, hashing head, FAISS index construction that will simultaneously explore the retrieval latency as well as maximize the precision.

Collectively, these extensions hold the potential of furthering transformer-based hashing as one of the leaders of scalable and high-precision fashion retrieval in industrial and commercial contexts

References

1. Zhang, X., *A survey on deep hashing for image retrieval.* arXiv preprint arXiv:2006.05627, 2020.
2. Bbouzidi, S., et al., *Convolutional neural networks and vision transformers for fashion mnist classification: A literature review.* arXiv preprint arXiv:2406.03478, 2024.
3. Dubey, S.R., S.K. Singh, and W.-T. Chu. *Vision transformer hashing for image retrieval.* in *2022 IEEE international conference on multimedia and expo (ICME).* 2022. IEEE.
4. Kumar, M., R. Singh, and P. Mukherjee, *VTHSC-MIR: Vision Transformer Hashing with Supervised Contrastive learning based medical image retrieval.* Pattern Recognition Letters, 2024. **184**: p. 28–36.
5. Mukhamediev, R.I., *State-of-the-Art Results with the Fashion-MNIST Dataset.* Mathematics, 2024. **12**(20): p. 3174.
6. Berriche, A., M.Z. Adjal, and R. Baghdadi. *Leveraging high-resolution features for*

improved deep hashing-based image retrieval. in *European Conference on Information Retrieval*. 2025. Springer.
7. Fang, Y. and L. Liu, *Scalable supervised online hashing for image retrieval.* Journal of Computational Design and Engineering, 2021. **8**(5): p. 1391–1406.
8. Chen, Y., et al., *Deep hashing with mutual information: A comprehensive strategy for image retrieval.*Expert Systems with Applications, 2025. **264**: p. 125880.
9. Rangwani, H., et al. *Deit-lt: Distillation strikes back for vision transformer training on long-tailed datasets*. in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024.
10. Khan, S., et al., *Transformers in vision: A survey.* ACM computing surveys (CSUR), 2022. **54**(10s): p. 1–41.
11. Mao, P., *Exploring and Evaluating Deep Hashing Methods within Vision Foundation Model Feature Spaces for Similarity Search and Privacy Preservation.* 2024.
12. Zhu, H., et al. *Deep hashing network for efficient similarity retrieval*. in *Proceedings of the AAAI conference on Artificial Intelligence*. 2016.
13. Li, Q., et al., *Deep supervised discrete hashing.* Advances in neural information processing systems, 2017.**30**.
14. Chen, Y., et al. *Transhash: Transformer-based hamming hashing for efficient image retrieval*. in *Proceedings of the 2022 international conference on multimedia retrieval*. 2022.
15. Cui, X., et al., *Multi-FusNet: fusion mapping of features for fine-grained image retrieval networks.* PeerJ Computer Science, 2024. **10**: p. e2025.
16. Touvron, H., et al. *Training data-efficient image transformers & distillation through attention*. in *International conference on machine learning*. 2021. PMLR.
17. Zhang, Q., et al., *SoftHash: High-dimensional Hashing with A Soft Winner-Take-All Mechanism.*
18. He, C. and H. Wei. *HybridHash: Hybrid convolutional and self-attention deep hashing for image retrieval*. in *Proceedings of the 2024 international conference on multimedia retrieval*. 2024.
19. Chen, Y., et al. *FashionERN: enhance-and-refine network for composed fashion image retrieval*. in *Proceedings of the AAAI Conference on Artificial Intelligence*. 2024.
20. Manna, A., R. Sista, and D. Sheet, *Deep neural hashing for content-based medical image retrieval: A survey.* Computers in Biology and Medicine, 2025. **196**: p. 110547.
21. Zhe, C. *Deep Hashing Image Retrieval Based on Attention Mechanism*. in *Proceedings of the 2024 10th International Conference on Communication and Information Processing*. 2024.
22. Guo, J., et al., *HASH-RAG: Bridging Deep Hashing with Retriever for Efficient, Fine Retrieval and Augmented Generation.* arXiv preprint arXiv:2505.16133, 2025.