THE HONG KONG
POLYTECHNIC UNIVERSITY
香港理工大學
Pao Yue-kong Library
包玉剛圖書館

Copyright Undertaking

This thesis is protected by copyright, with all rights reserved.

By reading and using the thesis, the reader understands and agrees to the following terms:

1. The reader will abide by the rules and legal ordinances governing copyright regarding the use of the thesis.
2. The reader will use the thesis for the purpose of research or private study only and not for distribution or further reproduction or any other purpose.
3. The reader agrees to indemnify and hold the University harmless from and against any loss, damage, cost, liability or expenses arising from copyright infringement or unauthorized usage.

IMPORTANT

If you have reasons to believe that any materials in this thesis are deemed not suitable to be distributed in this form, or a copyright owner having difficulty with the material being included in our database, please contact lbsys@polyu.edu.hk providing details. The Library will look into your claim and consider taking remedial action upon receipt of the written requests.

Pao Yue-kong Library, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong

http://www.lib.polyu.edu.hk

AN ADVANCING FRAMEWORK FOR COMPLEX FABRIC IMAGE RETRIEVAL: FROM MULTI-SCALE FEATURE FUSION TO EFFICIENT ATTENTION-BASED MATCHING

ZHANG RONGCHEN

PhD

The Hong Kong Polytechnic University

2025

The Hong Kong Polytechnic University

School of Fashion and Textiles

An Advancing Framework for Complex Fabric Image Retrieval: From Multi-Scale Feature Fusion to Efficient Attention-Based Matching

ZHANG Rongchen

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy

April 2025

CERTIFICATE OF ORIGINALITY

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgement has been made in the text.

_____ (Signed)

Zhang Rongchen
_____ (Name of student)

Abstract

Accurate fabric image retrieval is significant for modern textile and apparel industries, as it directly impacts inventory management, production efficiency, and design innovation. Despite advancements in content-based image retrieval, existing methods struggle with complex fabric patterns including plaid, lace, striped and printed fabrics, covering patterns ranging from geometric, floral to abstract designs, which are difficult because of the high intra-class variability, and diverse real-world conditions. These methods exhibit three key limitations: (1) insufficient multi-scale feature representation, limiting their ability to simultaneously capture fine-grained textures and global structures; (2) inconsistency in feature representations across hierarchical retrieval frameworks, causing mismatches between coarse and fine retrieval stages; and (3) inefficiency in local feature matching due to high computational complexity.

To address these challenges, this study proposes a novel framework to enhance the precision, efficiency, and robustness of fabric image retrieval systems. The framework comprises three main components. First, the Multi-Scale Local Descriptors Fusion (MLDF) method is introduced. This method employs a multi-scale feature extraction module with convolutional layers of varying receptive fields to capture both fine-grained textures and broader structural patterns. Feature fusion is achieved through Mixer Modules, which integrate token and channel dimensions to ensure a comprehensive representation. Additionally, a progressive triplet mining strategy is implemented to optimize feature embeddings, enhancing the discriminative power for complex fabric patterns.

Second, the Hierarchical Two-Stage Retrieval Framework is proposed. This framework utilizes global descriptors for efficient coarse retrieval and local descriptors for fine-grained refinement. An enhanced triplet loss function ensures consistency across feature spaces, improving inter-class separability and intra-class compactness. This approach effectively balances computational efficiency with retrieval accuracy, addressing the limitations of both single-stage and traditional two-stage

methods.

Third, the Efficient Local Feature Matching via Cross Attention (ELFM) method is developed. This method incorporates a Cross Attention Module to dynamically align local features between query and candidate images, capturing fine-grained relationships. By combining a learnable attention mechanism with feature aggregation, ELFM enables precise similarity computation and re-ranking. The method achieves high precision and recall, making it suitable for industrial applications requiring accurate and efficient retrieval.

The proposed methods were evaluated on a newly constructed dataset of 2,448 fabric images from 537 categories of patterns, reflecting real-world variability in both controlled and natural environments. Experimental results demonstrate significant improvements in retrieval precision, recall, and computational efficiency compared to traditional handcrafted features and existing deep learning models. This study makes contributions in the following ways: (1) advancing multi-scale feature fusion techniques for complex texture representation, (2) pioneering a hierarchical two-stage framework with feature space consistency, and (3) optimizing local feature matching via attention mechanisms. These innovations bridge the gap between theoretical robustness and industrial scalability, offering a unified solution for accurate and efficient fabric retrieval.

Publication arising from the thesis

[1] Rongchen Zhang, Wai Keung Wong, Shuai Lyu, "MLDF: Multi-scale Local Descriptors Fusion for Lace Fabric Image Retrieval", *Knowledge-Based Systems*, 2025: 114777.

[2] Rongchen Zhang, Shuai Lyu, Wai Keung Wong, "A Two-Stage Fabric Image Retrieval Approch By Integration with Global-Local Feature", *IEEE Sensors Journal* under review.

Acknowledgements

I would like to express my deepest gratitude to my supervisor, Prof. Wai Keung Wong, for his invaluable guidance and unwavering support throughout this research journey. His profound expertise, rigorous academic standards, and patient mentorship have been instrumental in shaping this dissertation. To my beloved wife, Ms. Ziqi Guo, I am grateful for her enduring support and understanding. Her unwavering encouragement during moments of challenge provided the emotional foundation necessary to complete this work. I am also deeply thankful to my colleagues and teammates for their intellectual companionship and collaborative spirit. Their insightful feedback greatly enriched this research experience. Special thanks go to Mr. Shuai Lyu for his constructive suggestions regarding technical discussions.

Table of Contents

Chapter 1

Introduction

1.1 Background

With the continuous improvement of living standards, consumer demand has evolved from basic functionality to personalized aesthetics and visual appeal. Fabric, as a key material that highlights clothing styles and characteristics, is now designed with diverse patterns, giving rise to an extensive variety of printed textiles. However, this diversity poses significant challenges for production and management, including labor-intensive imitation processes, prolonged production cycles, and increased inventory management costs. Many textile companies maintain vast sample libraries containing thousands of patterns, with designers continuously adding new designs. When adjusting to market trends or analyzing popular fabric styles, companies often need to search for similar fabrics in these large libraries and retrieve relevant production guidelines. This process is time-consuming, inefficient, and costly.

Traditional fabric retrieval methods rely heavily on manual annotation and human expertise. Samples are often categorized and stored based on manually defined keywords, requiring significant human effort. During retrieval, staff depend on visual observation to identify patterns with similar features. This approach not only increases operational costs but also results in retrieval accuracy being highly dependent on personal experience, limiting scalability and efficiency.

Image retrieval presents a promising solution to these challenges. In manufactur-

ing, fabric image retrieval systems can significantly improve inventory management and production planning by automating the identification and classification of fabric types. These systems reduce manual inventory errors, streamline production workflows, and enable businesses to respond rapidly to market demands for specific patterns or colors, avoiding overstocking or delays. In design, fabric image retrieval enhances creative workflows by enabling designers to quickly locate and compare fabric styles. Designers can input specific patterns or colors into retrieval systems to access similar or related fabrics, saving time while inspiring new ideas. Additionally, analyzing market trends through these systems helps designers identify popular patterns and align new products with consumer preferences, thereby improving competitiveness in the market.

1.1.1 Development of Fabric Image Retrieval

Image retrieval techniques are primarily categorized into two main approaches: Text-Based Image Retrieval (TBIR) [1] and Content-Based Image Retrieval (CBIR) [2, 3, 4, 5, 6, 7, 8]. TBIR characterizes images through textual descriptions and utilizes this information to identify matching images during retrieval. In contrast, CBIR relies on intrinsic image features to find similar images. While TBIR methods require extensive manual annotation, their representation of images is often incomplete and rigid, failing to capture the full diversity and intricacies of the images, thereby limiting the expressiveness of retrieval systems. On the other hand, CBIR methods can automatically extract rich visual features from images, reducing dependence on manual annotation and enabling a more comprehensive representation of image content.

Early fabric image retrieval methods based on CBIR using traditional handcrafted feature, such as texture, color, and basic semantic information. While effective for simple patterns, these methods struggled with complex and diverse fabric designs. The advent of deep learning, particularly models like Convolutional Neural Networks (CNNs) [9, 10, 11, 12, 13, 14, 15, 16] and Vision Transformers (ViTs) [17],

has significantly advanced fabric image retrieval capabilities. These models can automatically learn hierarchical image features, from low-level edges to high-level semantic attributes, enabling more precise fabric representation and retrieval. For example, CNNs excel at extracting spatial features of repetitive patterns, while ViTs provide enhanced capabilities for capturing global dependencies within intricate designs.

Despite these advances, fabric image retrieval systems face unique challenges. High-dimensional feature spaces often lead to computational inefficiencies, while semantic gaps between visually similar yet contextually different patterns may reduce retrieval accuracy. Addressing these issues, recent research has focused on developing methods to effectively extract and integrate multi-scale features, as fabrics often exhibit details that vary across spatial scales. Additionally, two-stage retrieval methods have emerged as a promising approach, where coarse retrieval narrows down candidate images, followed by fine-grained analysis to identify the most relevant matches.

This study investigates a novel two-stage, multi-scale feature fusion approach tailored to the retrieval of complex patterned fabric images. By leveraging hierarchical and multi-scale feature representations, this method aims to improve the retrieval systems accuracy and efficiency, addressing the specific challenges posed by diverse and intricate fabric patterns. The approach not only advances the state of fabric image retrieval but also contributes to the broader application of computer vision techniques in the textile industry.

1.2 Research Gap

Despite significant advancements in content-based image retrieval driven by deep learning, several critical gaps remain in addressing the challenges posed by complex fabric patterns, high intra-class variability, and diverse real-world conditions. These gaps are summarized as follows:

(1) Insufficient Multi-Scale Feature Representation

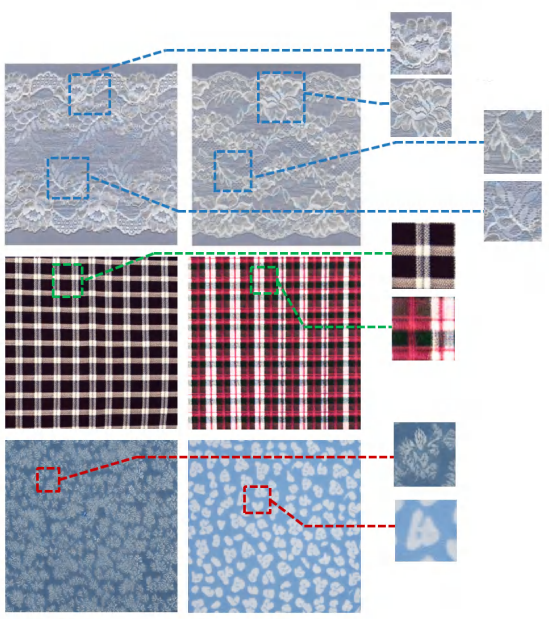


Figure 1.1: For fabric images with complex textures, there are cases where the local areas are similar but the global is not, and cases where the global is similar but the local areas are different.

Current methods struggle to simultaneously capture fine-grained textures and global structures in fabric images. As illustrated in Figure 1.1, some fabric images exhibit high similarity at one scale while differing significantly at another scale. Many existing approaches either focus on global features through Convolutional Neural Networks (CNNs) [9, 10, 11, 12, 13, 14, 15, 16] or aggregate local features using techniques like Vector of Locally Aggregated Descriptors (VLAD) [18]. However, these methods often fail to effectively represent both fine-grained textures and broader structural patterns simultaneously.

This limitation is particularly evident when dealing with complex fabric patterns where both local details and global structures are crucial for accurate identification. The inability to capture multi-scale information comprehensively leads to reduced retrieval accuracy, especially in cases where fabrics share similar characteristics at one scale but differ significantly at another.

(2) Inconsistency in Feature Representations Across Hierarchical Frameworks

Existing fabric image retrieval strategies, particularly hierarchical two-stage ap-

proaches, suffer from inconsistency in feature representations between coarse and fine retrieval stages. In two-stage retrieval frameworks, the initial coarse retrieval typically uses one type of feature representation to filter candidate images, followed by fine-grained retrieval using different features for precise ranking.

While this hierarchical approach aims to improve accuracy, the inconsistency between feature spaces used in different stages often leads to mismatches, compromising retrieval reliability. The disparity between coarse and fine feature representations can result in situations where promising candidates identified in the coarse stage are incorrectly ranked or filtered out during fine-grained retrieval, reducing overall system performance.

(3) Inefficiency in Local Feature Matching

The computational complexity of local feature matching poses a significant challenge in fabric image retrieval systems. Current methods for comparing local features between query and database images often require extensive computational resources, making them impractical for efficient applications or real-time scenarios.

While global feature-based approaches offer better efficiency, they often sacrifice the precision necessary for distinguishing subtle differences in fabric patterns. Conversely, methods that emphasize detailed local feature matching achieve higher accuracy but at the cost of substantially increased computational overhead. This trade-off between computational efficiency and retrieval accuracy remains a critical bottleneck, particularly in industrial settings where both speed and precision are essential.

The high computational complexity of local feature matching not only affects processing time but also limits the scalability of retrieval systems when dealing with large databases. This efficiency challenge becomes particularly acute in real-world industrial applications where rapid response times are crucial for practical implementation.

1.3 Research Objectives

This research aims to develop an innovative fabric image retrieval algorithm to achieve precise retrieval of images featuring complex fabric patterns, addressing the unique challenges inherent in fabric image analysis. The overarching objectives of this research are summarized as follows:

a) To design a novel feature extraction model that simultaneously captures both local texture features and global semantic features of fabric images. This dual-focus approach aims to enhance feature extraction, ensuring that both fine-grained details and overarching pattern structures are effectively represented.

b) To design a Hierarchical Two-Stage Retrieval Framework that ensures consistency between coarse and fine retrieval stages. This framework aims to maintain coherent feature representations across different retrieval stages, improving both inter-class separability and intra-class compactness.

c) To optimize the local feature calculations in the two-stage retrieval process. To achieve efficient retrieval while maintaining high precision.

1.4 Research Methodology

In this study, a comprehensive fabric image retrieval framework is proposed, specifically designed for the textile and apparel industry. The research is organized into four key phases to systematically address the challenges of complex patterned fabric retrieval.

The first phase involves constructing a diverse and realistic dataset for training and evaluating the proposed methods. Fabric images were collected from textile enterprises and online shopping platforms, resulting in a dataset comprising four primary fabric types: lace, plaid, printed, and striped fabrics. Most of the images were captured in natural environments under varying lighting conditions and angles, while a smaller portion was collected in controlled experimental settings. This dataset reflects real-world variability in fabric images, making it suitable for

benchmarking fabric image retrieval methods.

The second phase introduces a single-stage fabric image retrieval framework based on multi-scale feature fusion. This framework incorporates a specially designed multi-scale feature extraction module and a feature mixer module, combined with a data mining strategy to enhance model optimization. The objective is to obtain robust fabric features that can effectively handle the variability in texture, pattern, and environmental conditions. By fusing features across multiple scales, the framework ensures that both global patterns and local details are accurately captured, improving retrieval performance.

In the third phase, a two-stage fabric image retrieval strategy is proposed, leveraging a single unified model for both global and local feature extraction. Initially, global features are used to perform a coarse retrieval, narrowing down the search space to a candidate set of the top-30 results. Subsequently, local feature matching is applied to the candidate set in the fine retrieval stage, refining the results for higher retrieval accuracy. The use of a single model ensures feature consistency across the two stages, addressing the common issue of feature space mismatch in traditional two-stage frameworks.

The final phase focuses on optimizing the local feature matching process in the second stage of retrieval. A multi-head attention mechanism is employed to efficiently and accurately match local features, enhancing the overall efficiency of the retrieval framework. By leveraging attention-based mechanisms, the model can better capture intricate relationships between local patterns, significantly improving both retrieval accuracy and computational performance.

1.5 Research Significance

This study proposes a comprehensive framework with corresponding algorithms to address the challenges of fabric image retrieval in industrial and commercial applications. The proposed framework is capable of accurately retrieving fabric images with complex patterns, significantly enhancing the efficiency of design, production,

and inventory management in the textile and apparel industry. Specifically, this study contributes to the field of computer vision and fabric image retrieval in the following aspects:

(1) Enhancement of Feature Extraction

This study introduces a novel feature extraction method based on multi-scale feature fusion. The framework incorporates a multi-scale feature extraction module and a feature fusion to enhance the robustness of extracted features. By effectively capturing both global patterns and local details, the method addresses the challenges posed by complex fabric textures and variations in real-world settings, thereby improving retrieval accuracy and scalability.

(2) Innovation in Two-Stage Retrieval Strategies

This study advances the methodology for fabric image retrieval by proposing a two-stage retrieval strategy using a single unified model. This approach extracts consistent global and local features, enabling seamless integration between the coarse and fine retrieval stages. By ensuring feature space consistency and reducing redundant computations, the framework achieves a balance between retrieval precision and computational efficiency, addressing key limitations of traditional two-stage methods.

(3) Optimization of Local Feature Matching Efficiency

This study enhances the efficiency of local feature matching in fabric image retrieval by introducing a multi-head attention mechanism. This mechanism optimally aligns local features between the query image and database images, significantly reducing computational overhead while improving matching accuracy. The optimization ensures that the framework performs effectively, making it practical for industrial applications.

1.6 Outline of the Thesis

The remainder of this thesis is organized as follows: Chapter 2 provides a comprehensive review of related works in fabric image retrieval, covering traditional hand-

crafted features, deep learning models, and retrieval strategies. Chapter 3 presents the overall methodology of the proposed framework, including dataset construction, feature extraction, and the two-stage retrieval strategy. Chapter 4 introduces the Multi-Scale Local Descriptors Fusion (MLDF) method, detailing the multi-scale feature extraction and fusion process. Chapter 5 describes the Hierarchical Two-Stage Retrieval Framework, focusing on global and local feature extraction and the enhanced triplet loss function. Chapter 6 presents the Efficient Local Feature Matching via Cross Attention (ELFM) method, highlighting the dynamic alignment of local features using a learnable attention mechanism. Chapter 7 concludes the thesis, summarizing the contributions, limitations, and potential future research directions.

Chapter 2

Literature Review

With the advancement of Content-Based Image Retrieval (CBIR), a variety of retrieval methods leveraging feature engineering and similarity measurement strategies have been developed. This chapter provides a comprehensive review of the relevant literature, focusing on these two key aspects: feature engineering and similarity measurement strategies. Furthermore, the principles and importance of the two-stage retrieval approach are discussed in detail, highlighting its significance in enhancing retrieval performance.

2.1 Retrieval Methods Based on Traditional Hand-Crafted Features

Traditional hand-crafted features describe the characteristics of observed objects based on human visual perception. These features typically represent images through three primary aspects: color, texture, and shape.

Color-Based Features were pioneered by Swain et al. [19], who introduced color histograms for retrieval. Despite their simplicity, these features lack spatial information, leading to mismatches between images with similar color distributions. To address this, advanced descriptors such as color coherence vectors [20], color moments [21], color correlograms [22], and dominant color descriptors [23] were

developed, improving representation while mitigating spatial ambiguities.

Texture-Based Features capture surface properties. Tai [24] employed the Gabor wavelet transform for multi-scale, multi-directional texture representation, laying the foundation for global texture extraction. Lowes SIFT [25] advanced local texture analysis with its robustness to scale, rotation, and illumination changes, inspiring extensions like PCA-SIFT [26], GLOH [27], SURF [28], and RootSIFT [29].

Shape-Based Features are represented via region-based methods, such as moment invariants, or contour-based methods, like Fourier descriptors. Zhang et al. [30] enhanced Fourier descriptors to improve robustness to noise and invariance to geometric transformations. However, shape-based retrieval often faces challenges in accuracy and computational efficiency compared to color and texture features.

Early research on fabric image retrieval was limited by the constraints of hand-crafted features, which primarily focused on superficial characteristics of fabric images. Jing et al. [31] proposed a fabric retrieval approach that integrated a weighted color histogram with image segmentation to address patterned fabrics. Building on this work, Jing [32] later introduced an algorithm that utilized color moments and the Gist feature descriptor, specifically tailored for printed fabrics. These methods combined hand-crafted features of color and shape to represent fabric imagery effectively. Zhang et al. [33] developed a descriptor for lace fabric image retrieval, called the Multi-Scale and Rotation-Invariant Local Binary Pattern (MRI-LBP), which demonstrated robustness in capturing intricate fabric details. Similarly, Li et al. [34] proposed a retrieval system for lace fabrics that leveraged texture features (using Haralick descriptors) alongside shape information to enhance retrieval precision. An innovative framework was introduced by researchers in [35], merging color moments with coding features derived from a perceptual hashing algorithm, aiming to improve both accuracy and efficiency. Adopting a different strategy, Nanik et al. [36] combined fractal-based texture features with HSV attributes, showcasing an effective blend of texture and color in fabric retrieval. To address the need for speed in retrieval processes, Cao et al. [37] employed a combination of SURF, K-

means clustering, and Locality-Sensitive Hashing (LSH) algorithms, offering a fast and efficient method for fabric image retrieval.

2.2 Retrieval Methods Based on Deep Learning Models

The advent of large labeled datasets like ImageNet [38] and advancements in hardware technologies, particularly GPUs and CPUs, have driven significant evolution in image processing. In 2012, AlexNet [14] set a new benchmark in the ImageNet classification task, surpassing the previous best score by 10 percentage points. This achievement marked the rise of Convolutional Neural Networks (CNNs) as a research focus, leading to the development of influential models such as VGG [9], Inception [12], ResNet [11], and DenseNet [39]. CNNs gained substantial traction in image retrieval, leveraging multi-layer architectures to automatically extract hierarchical features from images. These features, through transformations and combinations, represent abstract semantics tailored to specific tasks.

2.2.1 Hierarchical Representation of Deep Learning Models

Research [40, 41] has visualized CNN features, illustrating their hierarchical nature: lower layers capture basic elements like textures and edges, intermediate layers identify object components, and higher layers abstract entire objects. Donahue et al. [42] demonstrated that fully connected (FC) features from the FC6 layer of AlexNet exhibit robust cross-domain properties across tasks like object recognition, fine-grained classification, and scene recognition. Similarly, Sharif et al. [43] used OverFeat [44] to validate the cross-domain effectiveness of FC features in tasks such as object classification, attribute detection, and image retrieval. Azizpour et al. [45, 46] further analyzed factors like network depth, width, fine-tuning, and feature dimensions, highlighting that retrieval performance improves when the target dataset aligns closely with the source dataset.

Upon model training, two types of features are typically extracted for image retrieval: fully connected features and convolutional features.

Fully connected features combine low- and mid-level attributes into global representations, encoding high-level semantics like faces, landscapes, and objects. This process, termed hierarchical representation, aggregates feature maps from convolutional layers via pooling. The output of the FC layer serves as a feature vector. Babenko et al. [47] coined the term "Neural codes" for these features and validated their effectiveness on image retrieval tasks.

Convolutional features act as local descriptors, akin to SIFT, but differ in similarity distributions, rendering traditional aggregation methods less effective. Babenko et al. [48] proposed a simple "sum pooling" approach for feature aggregation, introducing Sum-Pooled Convolutional features (SPoC). They further refined this method with a "center prior" weighting strategy, enhancing feature relevance for central objects. SPoCs low dimensionality reduces overfitting risks during PCA computation. Tolias et al. [49] introduced the Region Maximum Activation of Convolutions (R-MAC) descriptor, which pools multi-scale image fragments into a global descriptor. Similarly, Razavian et al. [50] employed a multi-scale strategy to extract local features, integrating position, scale, and spatial consistency for geometric invariance. They proposed a Multi-Resolution Search (MR) strategy, extracting image fragments at various scales and applying spatial max pooling [51] and PCA for dimensionality reduction.

In the field of fabric image retrieval, Xiang et al. [11] proposed a novel multi-task learning framework to enhance the robustness of feature representations. Leveraging the ResNet architecture within a multi-task learning paradigm, their approach significantly improved feature extraction capabilities, resulting in higher retrieval accuracy across diverse fabric types. Rangkuti et al. [37] conducted similarity retrieval experiments using CNN. They employed a custom dataset of traditional clothing as their experimental subject. VGG19 was used as the feature extractor, and distance measurement models such as Manhattan, Euclidean, and Chebyshev

were utilized. Their study covered 56 traditional clothing styles. Xu et al. [9] employed a Siamese neural network trained with hard negative pairs, achieving rapid convergence and excelling in one-to-one lace fabric image retrieval tasks. This work demonstrated the efficacy of contrastive learning as a robust strategy for developing feature metrics tailored to one-to-one retrieval scenarios.To address the unique characteristics of ikat woven fabrics, Tena et al. [14] designed a custom network architecture comprising three convolutional layers followed by two fully connected layers, compressing image features into compact 120-dimensional vectors. This specialized architecture achieved remarkable retrieval accuracy on an ikat woven fabric dataset, underscoring the potential of task-specific networks for highly specialized fabric retrieval applications.Several studies [52, 53] have explored combining deep learning and handcrafted features to harness their complementary strengths. Gu et al. [52] fused CNN features extracted from AlexNet [54] with handcrafted features such as Local Binary Patterns (LBP) [55] and Histogram of Oriented Gradients (HOG) [56]. By calculating the Similarity Correlation Coefficient (SCC) to combine features at multiple levels, this approach effectively enhanced retrieval performance for colored spun fabrics. Similarly, Zhang et al. [53] introduced a two-stage retrieval strategy: the first stage performed coarse retrieval using global semantic features extracted by Inception-V3 [57], while the second stage applied ORB [35] local features for precise matching.

2.2.2 Transformer-Based Models

Recent advances have introduced Vision Transformers (ViTs) [17], demonstrating strong performance in image retrieval tasks. Gkelios et al. [58] employed ViTs without fine-tuning, achieving results comparable to CNN-based methods. The Image Retrieval Transformer [59], leveraging the DeiT model [60], extracts the cls token for similarity computation and fine-tunes it on the target dataset, outperforming many contemporary retrieval methods. Hash-based retrieval methods, such as TransHash [61] and HashFormer [62], utilize ViTs to extract cls tokens, which

are processed through hashing layers to produce hash features for similarity computations. VTS [67] [63] extends this approach by incorporating all patch tokens alongside the cls token, yielding hash features that consistently surpass CNN-based accuracies under various hash objective functions.

Hybrid models combining CNNs and Transformers have also been proposed. For example, Henkel et al. [68] [64] initially extract features with CNNs, which are then refined using Swin Transformers [69] [65]. These hybrid approaches and Transformer-only retrieval strategies underscore the transformative potential of Transformers in advancing image retrieval, paving the way for further exploration in this domain.

2.3 Retrieval Strategies

2.3.1 One Stage Retrieval Strategies

In traditional feature based retrieval, local features like SIFT descriptors are used to describe specific regions of an image. Since different images generate varying numbers of SIFT descriptors, feature aggregation becomes crucial for retrieval. To address this, encoding strategies such as Bag of Features (BOF) [66] and the Vector of Locally Aggregated Descriptors (VLAD) [18] have been developed to encapsulate multiple descriptors into a single representation. The BOF method, introduced in 2003 [66], adapts the TF-IDF paradigm from text retrieval to image search. It clusters SIFT descriptors using k-means, encodes them through cluster centroids into visual words, and represents the image as a bag of visual words. VLAD [18], proposed by Jegou et al., refines this approach by computing residuals for each cluster, summing them within clusters, and concatenating the residuals into a compact vector. Compared to BOF, VLAD significantly reduces codebook size while enhancing retrieval accuracy. Building on VLAD, Arandjelovic et al. introduced NetVLAD [67], a neural network layer designed for place image retrieval. By integrating NetVLAD into a CNN and optimizing its parameters during training, it

adapts to retrieval data and produces more distinguishable cluster centroids. Incorporating the Triplet Loss technique [68], NetVLAD outperforms non-end-to-end methods like SPoC [48], particularly in place recognition and image retrieval tasks.

To address the high dimensionality of image features, hashing methods [69, 70] have been employed to compress features, reducing storage needs and improving retrieval speed. Traditional two-stage hashing involves separate feature extraction and hash code generation, often resulting in suboptimal hash functions due to the lack of joint optimization. With the advent of deep learning, deep hashing methods based on CNNs and ViTs [71, 72, 61, 62] have emerged. These methods are classified into unsupervised, supervised, and semi-supervised approaches. Unsupervised hashing [73] operates without class labels, while supervised hashing [78–79] [74, 75] uses label information to achieve higher accuracy. Semi-supervised hashing [76] combines labeled and unlabeled data, making it suitable for scenarios with limited annotations. In fabric retrieval, Xiang et al. [11] proposed a multi-task learning framework based on ResNet using deep hashing to encode the final features, significantly enhancing feature robustness and retrieval accuracy across diverse fabric types.

2.3.2 Two Stage Retrieval Strategies

In image retrieval, results derived from simple similarity calculations are often unsatisfactory. To address this limitation, a two-stage retrieval process is commonly employed to refine initial results and improve precision. This process typically involves two key techniques: geometric verification and query expansion.

Geometric Verification: when significant variations in grayscale, scale, or perspective exist between a query image and its corresponding match in the image library, global feature disparities can be substantial. Under such circumstances, geometric verification becomes critical for ensuring accurate retrieval. This technique evaluates the spatial consistency of pixel grayscale distributions, angles, scales, and relationships with other local features. If the verification results deviate from plausi-

ble geometric configurations, the retrieval is considered invalid. Common geometric verification methods include feature point matching [77, 78], Homography [79, 80], and RANSAC [81, 82, 83]. Recent advancements have integrated geometric verification into two-stage retrieval frameworks, yielding state-of-the-art results. For instance, techniques like PatchNetVLAD [84], SuperGlue [85], and TransVPR [98] [86] demonstrate exceptional performance by incorporating robust spatial consistency evaluations into their retrieval pipelines.

Query expansion: a simple yet powerful re-ranking strategy widely used in image retrieval. Inspired by text-based retrieval methods, it enhances retrieval accuracy by generating new query features from the top-ranked results of an initial query. These expanded features effectively refine the search by capturing additional relevant information. Key strategies include average query expansion [87], which averages features from top results; transitive closure expansion [88], which leverages the connectivity of retrieved images; and inward-outward expansion [94–95] [89, 90], which iteratively incorporates features from both core and peripheral results. Similarly, Zhang et al. [53] proposed a two-stage retrieval strategy, combining coarse retrieval using global semantic features extracted by Inception-V3 [57] with fine matching via ORB [35] local features. This hybrid approach ensures both scalability and precision, particularly for complex fabric patterns.

Together, geometric verification and query expansion form the foundation of modern two-stage retrieval methods, addressing the limitations of single-stage similarity calculations and significantly enhancing retrieval precision.

2.4 Feature Matching

2.4.1 Hand-Craft Feature Matching

Feature-based image matching algorithms typically extract feature points and their local descriptors from image pairs, transforming image matching tasks into indirect and direct matching tasks. In local feature-based image retrieval, the correspondence

between the query image and database images is typically determined by measuring the distances between feature vectors, using metrics such as Euclidean distance or cosine similarity.

Feature validation often requires estimating transformation models that align features between images. One widely adopted approach is the Random Sample Consensus (RANSAC) paradigm [81, 82], which identifies geometric transformations between matched features. The core principle of RANSAC [83] is to iteratively generate transformation models from random subsets of feature correspondences, evaluating each model based on the number of inliers (i.e., feature pairs that fit the transformation). The model with the highest inlier count is selected as the optimal transformation. While effective, RANSAC can become computationally expensive when the probability of inliers is low, as more iterations are needed to ensure accurate results. An alternative to RANSAC is the Hough voting strategy [91], which operates by casting votes in a transformation space to identify the most likely geometric alignment. Each matched feature contributes a vote for a specific transformation hypothesis, and the hypothesis with the highest vote count is selected. Although Hough voting reduces reliance on random sampling, its computational complexity increases with the number of matched features, as each feature adds to the transformation spaces dimensionality.

2.4.2 Deep Learning Feature Matching

While most traditional feature matching methods were originally designed for handcrafted local features, such as SIFT and SURF, they are generally unsuitable for the global features extracted by Convolutional Neural Networks (CNNs). However, recent advancements in image retrieval have introduced numerous deep learning-based local feature extraction and matching methods [49, 50, 51] that can be seamlessly integrated with traditional strategies, bridging the gap between handcrafted and learned features.

Modern feature matching approaches leverage deep learning to improve robust-

ness and accuracy in scenarios where traditional methods fall short, such as dealing with large geometric transformations, complex textures, or occlusions. For example, Chen et al. [92] proposed an end-to-end trainable deep network that directly predicts dense displacements between images. This method simplifies the matching process by learning correspondences in a holistic manner, making it highly effective for applications requiring pixel-level precision. In medical imaging, DeepFLASH [93] addresses computational and memory constraints, enabling efficient image registration without compromising performance. By reducing resource demands, it makes dense feature matching more practical for high-resolution image datasets. Mok and Chung [94] further advanced image registration by introducing an unsupervised algorithm that prioritizes topology conservation and smoothness. Their method estimates bidirectional transformations between two images by maximizing spatial similarity, resulting in more stable and realistic matching outcomes. Building on these developments, Sarlin et al. introduced the SuperGlue algorithm [85], which represents a significant leap in feature matching technology. SuperGlue employs an attention-driven content aggregation mechanism, enabling it to dynamically sense potential 3D scene structures and execute feature matching with exceptional precision. Unlike traditional methods that rely solely on geometric consistency, SuperGlue integrates both spatial and contextual information, allowing it to efficiently discard mismatches and improve overall matching quality. This makes it particularly well-suited for tasks such as 3D reconstruction, visual localization, and image retrieval.

These deep learning-based methods not only enhance the precision of feature matching but also complement traditional approaches like RANSAC and Hough voting. By integrating deep local features with geometric verification, hybrid systems can achieve higher accuracy and robustness across various domains, including fabric image retrieval, place recognition, and medical imaging. The growing convergence of traditional paradigms with modern deep learning techniques underscores the potential of hybrid models to address complex matching challenges in image

retrieval and beyond.

2.5 Chapter Summary

This chapter provided a comprehensive review of image retrieval techniques, tracing their progression from traditional hand-crafted feature-based methods to advanced deep learning-driven approaches. Traditional methods, focusing on features such as color, texture, and shape, showed limitations in scalability and robustness, particularly for complex fabric image retrieval tasks. Deep learning models, including CNNs and Vision Transformers, have significantly enhanced retrieval performance by enabling robust hierarchical feature extraction and task-specific adaptations, such as multi-task learning frameworks and hybrid CNN-Transformer models. Furthermore, advanced retrieval strategies, including two-stage approaches that incorporate geometric verification and query expansion, have proven effective in refining retrieval accuracy.

Despite these advancements, challenges remain in addressing the complexities of fabric retrieval in natural environments. Existing methods often struggle with diverse scales, varied shooting angles, and the intricate surface textures of fabrics. Current two-stage retrieval strategies, while effective, tend to be computationally intensive and rely on feature representations that lack sufficient robustness for these scenarios. These limitations highlight the need for more efficient, adaptable, and powerful retrieval frameworks.

To address these challenges, the proposed methodology introduces three key innovations: (1) a multi-scale feature fusion mechanism to capture fine-grained and global patterns; (2) a hierarchical two-stage framework with unified feature spaces to ensure consistency between coarse and fine retrievals; and (3) a lightweight cross-attention module for efficient local feature matching. The next chapter details these components and their integration into a cohesive framework.

Chapter 3

Methodology

This chapter introduces a novel fabric image retrieval method, outlining its framework, strategies, and the specific designs of each component. Additionally, the chapter describes the constructed dataset and the metrics utilized for training the algorithm, as well as the evaluation system implemented to assess its performance.

3.1 Overview of the Proposed Two-Stage Fabric Image Retrieval Framework

To address the challenges associated with complex fabric retrieval, a two-stage fabric image retrieval framework is proposed, integrating multi-scale feature extraction and local feature matching. This methodology comprises four main components: 1) Dataset Construction, 2) Feature Extraction, 3) Two-Stage Retrieval Strategy, and 4) Local Feature Matching Optimization. The detailed process is illustrated in Figure 3.1.

Phase 1. Dataset Construction: A comprehensive dataset has been constructed to support the proposed retrieval system. This dataset includes both library images and query images, encompassing a variety of fabric categories such as lace, plaid, printed, and striped fabrics. The images were collected under diverse conditions to reflect real-world variability, thereby ensuring robustness in system

evaluation.

Phase 2. Feature Extraction: A single unified deep learning model is used to extract global and local features from fabric images. Global features provide a holistic representation of the image, while local features capture fine-grained texture and pattern details. By combining global and local features, the model ensures a comprehensive representation of fabric images.

Phase 3. Two-Stage Retrieval Strategy: The retrieval process adopts a two-stage strategy. In the first stage, global features are used to perform a coarse retrieval, narrowing the search space to a subset of candidate images. In the second stage, local features are matched to refine the results, improving retrieval accuracy by focusing on fine-grained details.

Phase 4. Local Feature Matching Optimization: To further enhance the precision of the fine retrieval stage, a multi-head attention mechanism is introduced to optimize the local feature matching process.

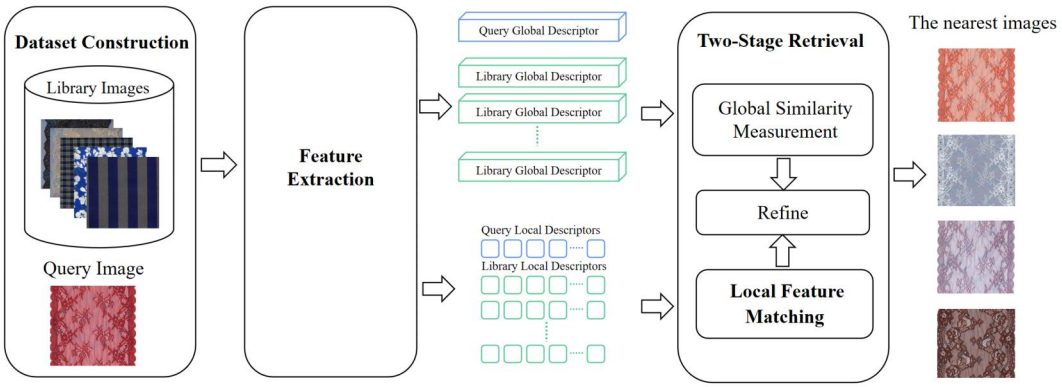


Figure 3.1: Architecture of the proposed fabric image retrieval framework.

3.2 Self Constructed Dataset

Based on the current state of research, no publicly available fabric image retrieval dataset has been identified. To address this gap, a comprehensive fabric image retrieval dataset was constructed, comprising 31,909 images that capture a wide range of fabric patterns, including lace, plaid, striped, and printed fabrics. The im-

ages were sourced from two main channels: collaboration with fabric manufacturing companies and collection from online platforms.

The collaboration with textile enterprises enabled the acquisition of high-quality fabric images captured under standardized conditions, ensuring consistency in shooting angles, lighting, and equipment. These images provide a reliable foundation for tasks requiring precise representation of fabric texture and details. Conversely, images gathered from online platforms reflect diverse and natural conditions, introducing variability in lighting, shooting angles, and devices. This diversity enhances the datasets robustness and applicability to real-world scenarios, making it well-suited for comprehensive evaluation and development of fabric image retrieval systems.

To ensure the datasets quality and relevance for fabric image retrieval tasks, rigorous annotation and screening processes were applied. During the annotation process, adherence to the following principles was maintained:

1. Manual Classification: Each annotation was cross-verified by workers from manufacturers to ensure consistency. Workers first identified the fundamental pattern structures by examining the repetition units, symmetry properties, and spatial arrangements. For geometric patterns (plaids and stripes), they measured the periodicity, line spacing consistency, and intersection angles. For organic patterns (printed and lace designs), they analyzed the motif repetition, branch connectivity, and void distributions. Fabrics exhibiting identical structural layouts were classified as the same pattern category.

2. Pattern-Based Matching Verification: To further verify and refine the manual classification, we performed a dense matching analysis grounded in quantifiable pattern characteristics. This computational step served as an objective validation mechanism to complement the experts' visual inspection. The process leveraged key descriptors such as repetition frequency, geometric structure, and texture complexity. The annotation focused on pattern information while disregarding color information to ensure consistency in category definitions.

3. Category Exclusion: Fabric image categories containing fewer than three

images were excluded from the dataset to maintain dataset integrity and relevance.

After rigorous screening and annotation, the final dataset contains a total of 2,448 images, of which 1,527 images were captured in collaboration with textile companies under a standardized collection environment. These images were taken under controlled conditions, including fixed shooting equipment, standard shooting angles, and consistent lighting conditions, as shown in Fig 3.2 categories (a) and (b). This standardized setup minimizes the interference of external factors, allowing for a more accurate and precise representation of the fabrics texture details and material characteristics. As a result, these images provide reliable foundational data for fabric recognition and classification tasks. The remaining 921 images were collected from online shopping platforms, captured in natural environments, and cover a range of different shooting devices, angles, and lighting conditions, as shown in Fig 3.2 categories (c) and (d). Compared to the images captured in standardized environments, these images exhibit greater variability and randomness, reflecting the appearance and performance of fabrics in various real-world usage scenarios. Therefore, these images offer enhanced diversity, further increasing the datasets applicability and presenting more challenging tasks for image retrieval in complex environments.

The self-constructed dataset was systematically organized into 537 distinct fabric categories, with each category representing a unique fabric pattern characterized by specific structural and visual properties. Patterns were primarily classified according to their fundamental structural properties. Geometric patterns were identified by their mathematical regularity, including specific angle measurements, symmetry axes, and repetition intervals. Organic patterns were characterized by their naturalistic forms, flow continuity, and morphological complexity.

A detailed breakdown of the fabric categories is presented in Table 3.1, highlighting the distribution of images across different fabric types. Among the fabric types, lace fabrics and printed fabrics constitute the largest proportion of the dataset. This is attributed to the high variability of their patterns and the greater complex-

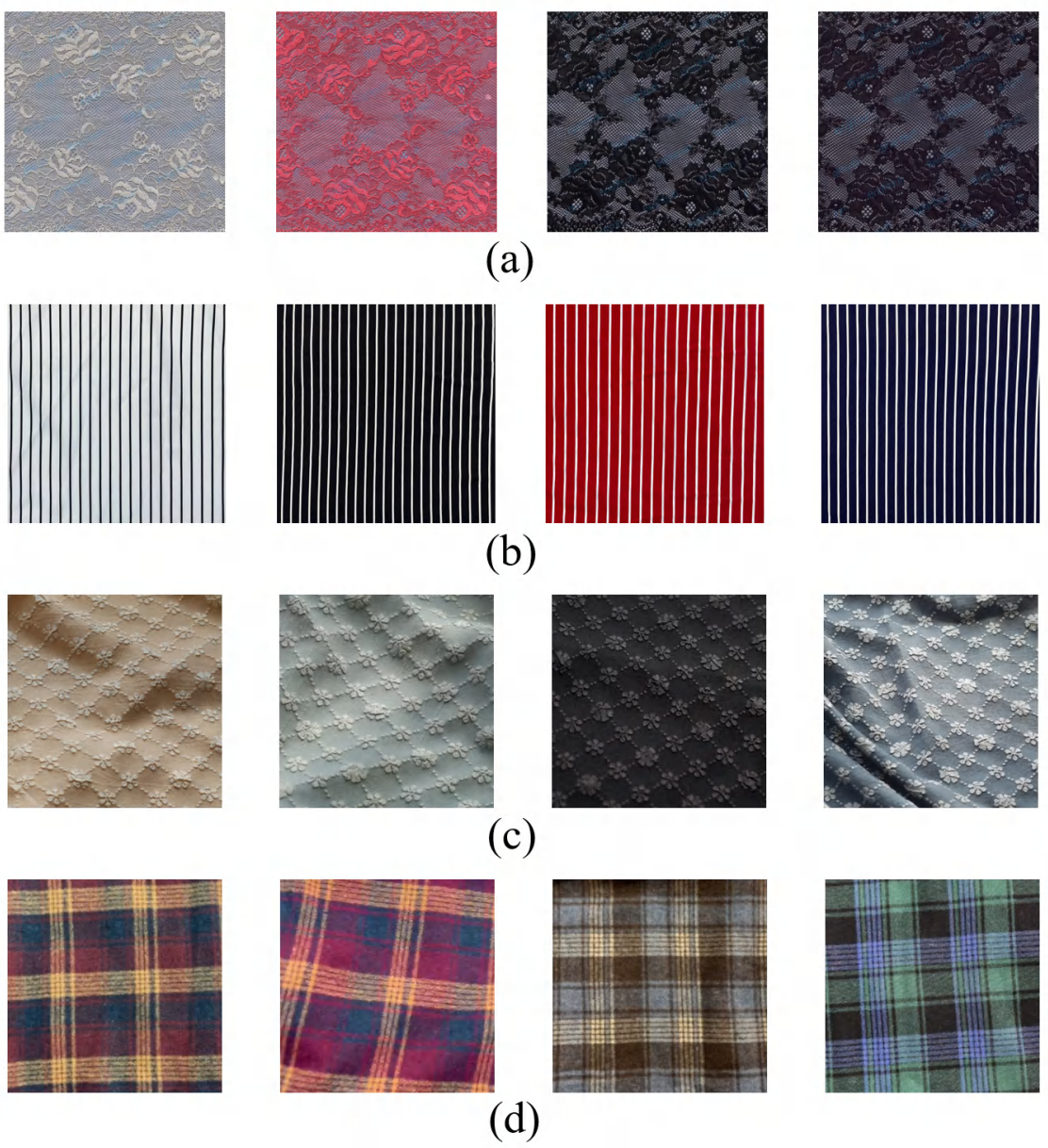


Figure 3.2: Fabric data examples of different categories. (**a**) and (**b**) The categories of lace fabric images and striped fabric images collected under standardized environments from fabric manufactures. (**c**) and (**d**) The categories of printed fabric images and plaid fabric images collected in natural environments from online shopping platforms.

ity of their surface features compared to plaid and striped fabrics. Lace fabrics are characterized by intricate and repetitive designs, while printed fabrics often display diverse and elaborate motifs. These characteristics make both fabric types particularly challenging for image retrieval tasks, requiring more comprehensive feature representations.

Table 3.1: Distribution of fabric types in the self constructed dataset, including image counts and fabric category counts.

Fabric Type	**Image Count**		**Category Count**	
	Total	**Proportion (%)**	**Total**	**Proportion (%)**
Lace Fabric	1,032	42.1	224	41.7
Printed Fabric	816	33.3	184	34.3
Plaid Fabric	378	15.4	74	13.8
Striped Fabric	222	9.1	55	10.2
Total	**2,448**	**100.0**	**537**	**100.0**

3.3 Feature Extraction

The feature extraction module is designed to effectively capture the characteristics of complex fabric images by integrating patch-level feature extraction with multi-scale analysis. Unlike conventional methods that rely on small receptive fields, this approach first processes the image into medium-scale patches to mitigate the interference caused by fine-grained local textures, such as individual yarns. Subsequently, multi-scale convolutions are applied to the patch-level features to extract discriminative information at varying levels of granularity.

3.3.1 Multi-scale Feature Extraction

Given an input image $\boldsymbol{I}$, the fabric image is first divided into non-overlapping patches of size 16×16. The use of a medium-scale patch ensures that the receptive field encompasses sufficient contextual information, reducing the impact of local texture noise. A convolutional operation is applied to obtain the patch-level feature map $\boldsymbol{F}$:

$$\boldsymbol{F} = \mathrm{Conv}_{16\times16}(\boldsymbol{I}), \tag{3.1}$$

where $\mathrm{Conv}_{16\times16}$ represents a convolution with a 16×16 kernel. The resulting output $\boldsymbol{F}$ consists of a set of patch-level descriptors $\boldsymbol{L} = \{\boldsymbol{l}_1, \boldsymbol{l}_2, \ldots, \boldsymbol{l}_N\}$, where $\boldsymbol{l}_i$ denotes the feature vector of the i-th patch and N is the total number of patches.

To analyze the patch-level features across different receptive fields, multi-scale convolutions are applied to $\boldsymbol{F}$. This step captures feature representations at varying

scales, enabling the model to focus on both fine-grained and coarse-grained structures within the patches. Specifically, convolutions with different kernel sizes are employed, defined as:

$$\phi(\boldsymbol{F}) = \text{Concat}\left(\text{Conv}_{1\times1}(\boldsymbol{F}), \text{Conv}_{3\times3}(\boldsymbol{F}), \text{Conv}_{5\times5}(\boldsymbol{F})\right), \tag{3.2}$$

where $\text{Conv}_{k\times k}$ denotes a convolution with a $k \times k$ kernel, and Concat concatenates features along the channel dimension.

The smaller kernels (1×1) capture fine-grained contextual details, while larger kernels (3×3, 5×5) analyze broader structures. By applying multi-scale convolutions to $\boldsymbol{F}$, the method integrates spatial information across medium-scale regions, particularly effective for fabric images with complex textures.

3.3.2 Feature Fusion

After obtaining multi-scale features, we combine them into a unified representation to encapsulate cross-scale information. Feature fusion integrates these descriptors, enhancing the discriminative power of the representation.

The fusion process serializes the multi-scale descriptors and passes them through fully connected layers, achieving integration across token and channel dimensions. This ensures effective combination and refinement of multi-scale information, improving the model's ability to distinguish intricate fabric patterns.

3.3.3 Local and Global Descriptors

The fused features $\boldsymbol{L}_{\text{fused}}$ generate both local and global descriptors:

Local Descriptors: Each local descriptor $\boldsymbol{l}_i$ captures fine-grained details within its patch:

$$\boldsymbol{l}_i = \boldsymbol{L}_{\text{fused}}[i], \tag{3.3}$$

where $\boldsymbol{L}_{\text{fused}}[i]$ is the i-th patch's feature vector. These preserve spatial details for precise feature matching in fine retrieval.

Global Descriptors: Global average pooling (GAP) aggregates fused features into a compact vector:

$$\boldsymbol{g} = \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{l}_i, \tag{3.4}$$

where $\boldsymbol{g}$ represents the global descriptor. This summarizes the fabric's overall structure for coarse retrieval.

The module first extracts medium-scale patch features using 16×16 convolutions to mitigate local noise. Multi-scale convolutions then process these features, followed by token-channel fusion to produce local and global descriptors. This combines medium-scale context with multi-scale details for robust fabric representation.

3.3.4 Mining Strategy for Model Optimization

To enhance the performance of the proposed feature extraction model, a data mining strategy is employed during the training process. This strategy focuses on identifying challenging and informative samples from the dataset, which play a crucial role in improving the models discriminative ability and robustness. The mining strategy consists of the following components:

Hard positive samples are those that exhibit high similarity to the query image but belong to different categories. These samples are identified during the training process by analyzing the feature space and similarity scores. By incorporating hard positive samples into the training process, the model is encouraged to learn subtle distinctions between visually similar patterns, such as different variations of lace or printed fabrics. Hard negative samples are those that have low similarity to the query image but are incorrectly classified as similar during training. These samples typically arise due to noise in the feature space or complex fabric patterns that share partial characteristics. The inclusion of hard negatives forces the model to refine its feature representations, reducing false positives and improving overall retrieval

accuracy.

To avoid overwhelming the model during early training stages, a progressive mining strategy is adopted. Initially, easy and moderate samples are used for training, allowing the model to learn basic feature representations. As training progresses, the difficulty of mined samples is gradually increased, ensuring a smooth and stable learning process. This progressive approach prevents overfitting to noisy samples while maintaining consistent improvements in performance.

To ensure that the model learns features across all fabric categories, a balanced sampling approach is incorporated into the mining strategy. This ensures that both underrepresented and overrepresented categories contribute equally during training, resulting in a more generalized feature extraction model.

3.4 Two-Stage Retrieval Strategy

The proposed two-stage retrieval strategy is designed to achieve an optimal balance between computational efficiency and retrieval accuracy by leveraging global and local feature representations. The process consists of two key stages: **Global Retrieval**, which efficiently narrows the search space by identifying an initial set of candidate images, and **Pairwise Local Matching**, which refines the candidate list using fine-grained local feature comparisons and spatial relationships.

3.4.1 Global Retrieval

In the first stage, the goal is to rapidly retrieve a candidate list of images from a large database that are broadly similar to the query image. This is achieved using compact global descriptors, which summarize the high-level semantic and structural content of an image.

Given a query image q with its global descriptor $\boldsymbol{g}_q$, the similarity between the query image and each database image d with global descriptor $\boldsymbol{g}_d$ is computed using

a distance function:

$$\text{dis}_{\text{global}}(q, d) = \text{distance}(\boldsymbol{g}_q, \boldsymbol{g}_d), \tag{3.5}$$

where distance$(\cdot)$ measures the alignment between the two global feature vectors, often using cosine similarity or another similarity metric.

Based on the computed similarity scores, the system selects the top-k most similar images to form the candidate list:

$$\mathcal{C} = \text{Top}{-}k\{d \mid \text{dis}_{\text{global}}(q, d)\}. \tag{3.6}$$

This stage efficiently reduces the size of the search space, ensuring that only the most relevant candidates proceed to the next stage for detailed analysis.

3.4.2 Pairwise Local Matching

Once the candidate list $\mathcal{C}$ is obtained from the global retrieval stage, the second stage performs fine-grained analysis by comparing local features between the query image and each candidate image. This stage involves the following steps:

Patch-Level Feature Extraction: Both the query image and candidate images are divided into spatial patches, and local descriptors $\{\boldsymbol{l}_{q,1}, \boldsymbol{l}_{q,2}, \ldots, \boldsymbol{l}_{q,n}\}$ and $\{\boldsymbol{l}_{c,1}, \boldsymbol{l}_{c,2}, \ldots, \boldsymbol{l}_{c,m}\}$ are extracted for the query q and each candidate c, respectively. These local descriptors encode the detailed patterns and textures of each patch.

Multi-Scale Fusion: To account for varying levels of detail in the images, multi-scale features are extracted from each patch. Specifically, multiple feature extraction scales w_1, w_2, w_3 are applied, where each scale analyzes the patch at a different level of granularity. The multi-scale features are then fused to form a robust patch descriptor:

$$\boldsymbol{l}_{q,i}^{\text{fused}} = \sum_{s=1}^{S} w_s \cdot \boldsymbol{l}_{q,i}^{s}, \quad S = 3, \tag{3.7}$$

where w_s represents the weight of the s-th scale (scalar), and $\boldsymbol{l}_{q,i}^{s}$ is the patch descriptor extracted at scale s.

Nearest Neighbor Matching: For each patch descriptor in the query image $\boldsymbol{l}_{q,i}^{\text{fused}}$, its nearest neighbor in the candidate image $\boldsymbol{l}_{c,j}$ is determined based on the distance function:

$$\text{dis}_{\text{local}}(\boldsymbol{l}_{q,i}, \boldsymbol{l}_{c,j}) = 1 - \frac{\boldsymbol{l}_{q,i} \cdot \boldsymbol{l}_{c,j}}{\|\boldsymbol{l}_{q,i}\| \|\boldsymbol{l}_{c,j}\|}. \tag{3.8}$$

This step identifies the best-matching patches between the query and candidate images.

Final Ranking: The candidate images are re-ranked based on the aggregated local matching scores, which combine patch-level similarities and spatial scoring. The final retrieval results are obtained by selecting the top-r ranked images:

$$\mathcal{R} = \text{Top}{-}r\{c \mid \text{score}_{\text{local}}(q, c)\}. \tag{3.9}$$

3.4.3 Key Contributions of the Two-Stage Strategy

The two-stage retrieval strategy effectively combines global and local feature representations to balance efficiency and precision.

- **Global Retrieval:** Provides a fast and efficient filtering mechanism by using global descriptors to reduce the search space.
- **Local Matching:** Ensures high retrieval accuracy by refining the candidate list using fine-grained patch-level comparisons and spatial scoring.
- **Multi-Scale Fusion:** Captures detailed information at multiple scales, enhancing the robustness of local feature matching.
- **Spatial Consistency:** Incorporates spatial relationships between matched patches, improving the reliability of the retrieval results.

The combination of these components enables the system to handle complex fabric patterns effectively, achieving accurate and efficient retrieval performance.

By utilizing stage-specific similarity metrics tailored to global and local representations, the retrieval strategy achieves both computational efficiency in the first stage and precision in the second stage. Together, these methods enable the system to handle complex fabric retrieval while providing accurate and reliable retrieval results.

3.5 Efficient Local Feature Matching Method

The Efficient Local Feature Matching (ELFM) method integrates a two-stage retrieval framework with a Cross Attention-based local matching module to achieve high-accuracy and computationally efficient fabric image retrieval. The primary components of ELFM are as follows:

The ELFM framework adopts a two-stage approach.

- **First Stage - Global Feature Retrieval**: In the initial stage, global feature representations are extracted from each image using a feature extraction backbone, specifically the MLDF (Multi-Level Deep Features) model. These global embeddings enable rapid identification of a shortlist of candidate images that are likely to be relevant to the query image. Mathematically, for an input image $\mathbf{I}$, the backbone produces a global feature vector:

$$\mathbf{g} = \text{MLDF}(\mathbf{I}), \tag{3.10}$$

  where $\mathbf{g} \in \mathbb{R}^d$ represents the global descriptor.

- **Second Stage - Local Feature Matching**: The second stage refines the initial shortlist by performing fine-grained matching between local descriptors of the query and candidate images. This is achieved through the Cross Attention Module, which models the relationships between local features to compute a

robust similarity score.

3.5.1 Cross Attention-Based Local Matching Module

The core innovation of ELFM lies in its Cross Attention Module, which efficiently models the correspondence between local descriptors of the query and candidate images.

For each image, the feature extraction backbone outputs a feature map of dimensions (H, W, C), where H and W denote the spatial dimensions, and C is the channel dimension. These feature maps are reshaped into sequences of local descriptors:

$$\mathbf{Q} = \{\mathbf{q}_i\}_{i=1}^{M}, \quad \mathbf{C} = \{\mathbf{c}_j\}_{j=1}^{N}, \tag{3.11}$$

where $M = H \times W$ and $N = H \times W$ for the query and candidate images, respectively.

The Cross Attention Module treats the query descriptors $\mathbf{Q}$ as Queries and the candidate descriptors $\mathbf{C}$ as Keys and Values. The attention mechanism is defined as:

$$Q = W_Q\mathbf{Q}, \quad K = W_K\mathbf{C}, \quad V = W_V\mathbf{C}, \tag{3.12}$$

$$\text{Attention}(Q, K) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right), \tag{3.13}$$

$$\text{Output} = \text{Attention}(Q, K)V, \tag{3.14}$$

where $W_Q, W_K, W_V \in \mathbb{R}^{C\times d}$ are learnable projection matrices, and d is the dimensionality of the attention space. This mechanism allows the model to compute matching weights between local descriptors efficiently, capturing fine-grained correspondences.

3.5.2 Score Prediction

The refined local descriptors from the Cross Attention Module are aggregated using an adaptive average pooling layer:

$$\mathbf{z} = \text{Pool}(\text{Output}). \tag{3.15}$$

Subsequently, a Multi-Layer Perceptron (MLP) with non-linear activations predicts the global similarity score:

$$\text{Score} = \sigma(\text{MLP}(\mathbf{z})), \tag{3.16}$$

where σ is the sigmoid activation function. This score represents the likelihood that the query and candidate images depict the same fabric.

3.5.3 Training Strategy

The ELFM model is trained end-to-end using a binary cross-entropy loss function. For each query-candidate pair, the loss is defined as:

$$L = -\sum_{i} \left[y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i) \right], \tag{3.17}$$

where $y_i \in \{0, 1\}$ is the ground truth label indicating whether the pair is a match, and $\hat{y}_i$ is the predicted similarity score. This training approach ensures that the feature extraction backbone, Cross Attention Module, and MLP are jointly optimized to enhance retrieval performance.

Overall, the Efficient Local Feature Matching method leverages a sophisticated Cross Attention mechanism within a two-stage retrieval framework to achieve precise and scalable fabric image retrieval. By focusing on both global and local feature alignments, ELFM effectively balances accuracy and computational efficiency, making it well-suited for efficient and real-time applications.

3.6 Chapter Summary

This chapter introduces the proposed two-stage fabric image retrieval framework, integrating four essential components: Dataset Construction, Feature Extraction, Two-Stage Retrieval Strategy, and Local Feature Matching Optimization. To address the lack of comprehensive fabric datasets, a diverse collection of 31,909 images covering various fabric types such as lace, plaid, printed, and striped fabrics was developed through collaborations with textile manufacturers and online platforms, resulting in 2,448 high-quality, categorized images organized into 537 fabric categories. The feature extraction process combines patch-level feature extraction with multi-scale convolutions, capturing both fine-grained textures and broader structural patterns, and employs token and channel fusion operations to generate robust local and global descriptors. Additionally, a strategic mining approach enhances model training by incorporating hard positive and negative samples, thereby improving the models discriminative power and robustness. The two-stage retrieval strategy first utilizes global descriptors for rapid candidate selection, significantly narrowing the search space, followed by pairwise local matching that conducts detailed patch-level comparisons and spatial scoring to refine the results. Finally, a Cross Attention-based module is introduced in the second stage to facilitate efficient and precise feature alignment between query and candidate images, leveraging a learnable attention mechanism and an end-to-end training framework supported by binary cross-entropy loss. Collectively, these components ensure that the framework effectively combines global and local feature representations, achieving both scalability and high retrieval accuracy, thereby establishing a solid foundation for robust and efficient fabric image retrieval systems.

Chapter 4

Multi-scale Local Descriptors Fusion For Fabric Retrieval

Fabric image retrieval presents unique challenges due to the complexity of fabric patterns, variations in scale and perspective, and diverse shooting conditions, particularly in natural environments. Existing retrieval methods often struggle to achieve robustness and precision in such scenarios. Traditional hand-crafted feature-based methods lack the capacity to capture intricate textures and contextual relationships, while conventional deep learning models sometimes fail to effectively handle multi-scale and localized variations inherent to fabric images.

This chapter introduces the Multi-scale Local Descriptors Fusion (MLDF) method, a core component of the proposed framework aimed at addressing multi-scale feature representation gaps in fabric retrieval. MLDF serves as the foundational feature extractor, enabling the system to simultaneously capture fine-grained Textures and global structural patterns through its hybrid convolutional architecture.

4.1 Introduction

Advancements in image technologies, particularly the introduction of Convolutional Neural Networks (CNN) [9, 10, 11, 12, 13, 14, 15, 16], have significantly enhanced

Figure 4.1: Comparison of images with identical patterns across lace, printed, plaid, and striped fabrics reveals that lace fabrics present greater challenges. Unlike other fabrics, lace undergoes not only color changes but also alterations in surface texture characteristics. In contrast, the other fabrics exhibit changes only in color.

Content-Based Image Retrieval (CBIR) [2, 3, 4, 5, 6, 7, 8] by improving image feature extraction.

Despite these advancements, fabric image retrieval remains challenging due to the varying complexity of fabric patterns. As shown in Figure 4.1, identical patterns in lace, printed, plaid, and striped fabrics exhibit distinct retrieval difficulties. While printed, plaid, and striped fabrics with identical patterns primarily differ in color, their local textural features remain consistent. In contrast, lace fabrics demonstrate significant intra-class variability, with local textures within the same pattern often exhibiting substantial differences. This variability makes lace fabric retrieval considerably more challenging compared to other fabrics with simpler surface features, underscoring the need for advanced retrieval methods tailored to such complexities.

Lace fabrics encompass a wide variety of types distinguished by their patterns and materials, exhibiting complex and variable appearances. These fabrics are characterized by small inter-class differences but large intra-class variations, which pose significant challenges for accurate image retrieval. For intricate lace fabric images, Figure 4.2(a) illustrates instances where visually similar features belong to different classes, while Figure 4.2(b) shows cases where differences in visual features exist within the same class. In such scenarios, a thorough analysis of local features similarities is essential. Accurate retrieval of specific lace fabric images thus requires attention to both local details and global patterns.

The importance of receptive field scale in capturing local feature diversity is un-

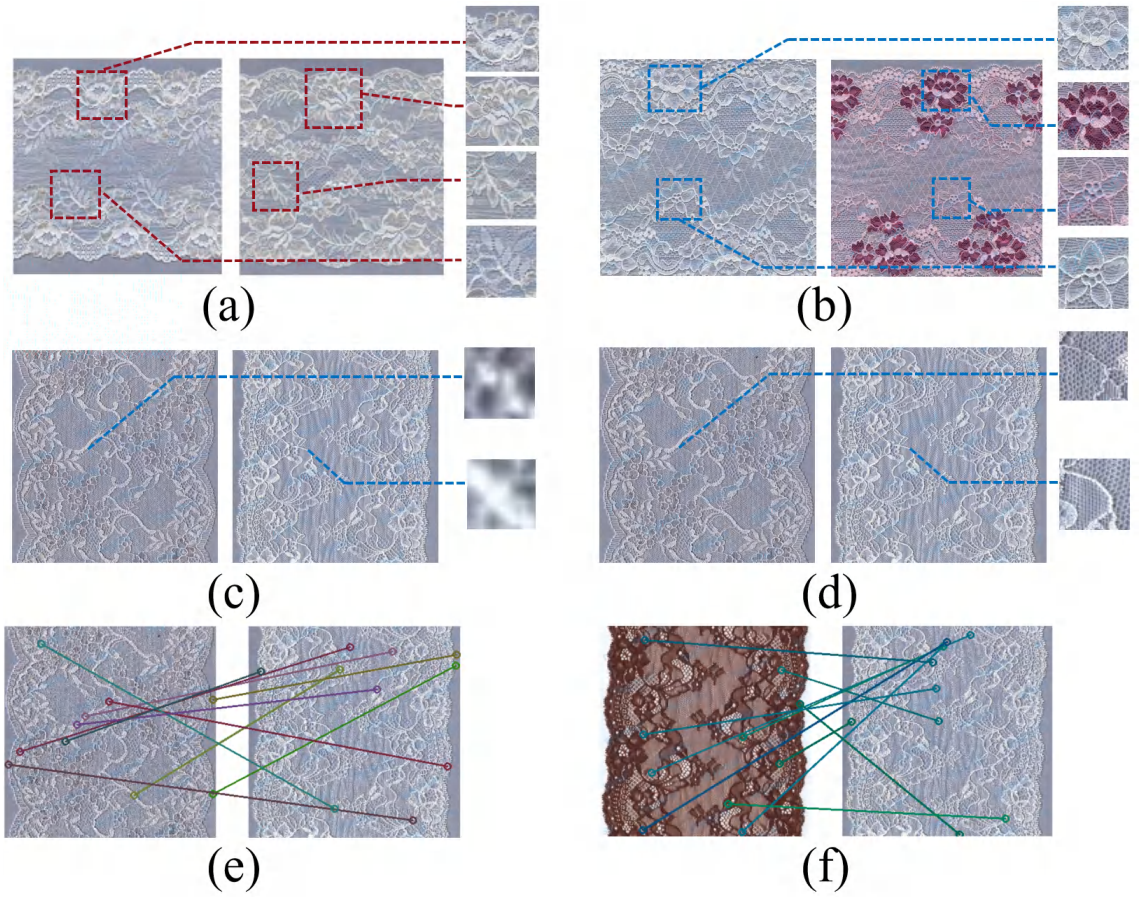


Figure 4.2: Local areas and receptive field comparisons: (a) presents two fabric samples that have a similar overall appearance but belong to different classes. (b) illustrates two samples from the same class that appear different at a global level. (c) and (d) compare local area differences in lace fabrics of different classes under small and patch-level receptive fields. (e) displays the top 10 matches using SURF (small receptive field features) on lace fabrics of different classes. (f) displays the top 10 matches using SURF on lace fabrics within the same class.

derscored in Figure 4.2. A small receptive field, as shown in Figure 4.2(c), fails to reveal the diversity of local patterns, capturing only highly consistent characteristics of the yarn. In contrast, a patch-level receptive field, illustrated in Figure 4.2(d), better captures the diversity and contextual relationships of local features. When traditional methods like SURF [28] are used with small receptive fields, incorrect matches frequently occur. This is evident in Figure 4.2(e), where lace fabrics of different classes are mismatched, and in Figure 4.2(f), where mismatches are observed within the same class. These observations highlight the need for appropriately scaled receptive fields that account for the contextual relationships within patches, ensuring

diversity and accuracy in local feature representation.

To address these challenges, two critical issues must be resolved: 1. Balancing local and global features: Accurate retrieval demands focusing on both local areas, which exhibit distinct inter-class differences, and global patterns for context. 2. Receptive field scale optimization: Local features must be captured with a sufficiently large receptive field to ensure contextual diversity, as overly small receptive fields fail to incorporate critical relationships between local regions.

Existing methods [95, 96, 97, 98, 99, 100, 101] based on CNN features predominantly prioritize global semantic information, often neglecting nuanced local details. Conversely, traditional feature-based methods [102, 103, 104, 56, 34, 55, 32, 36, 105] focus too narrowly on local areas without effectively incorporating contextual relationships.

To overcome the identified limitations, this study proposes MLDF for fabric image retrieval, integrating local and global feature analysis while emphasizing contextual relationships. The framework comprises three key components: Multi-scale Feature Extractor and Feature Fusion with Mixer Modules and Progressive Triplet Mining Strategy.

Multi-scale Feature Extractor: The proposed multi-scale feature extraction method is specifically designed to capture intricate local features while preserving the contextual richness necessary for comprehensive image representation. This method addresses the need to analyze complex fabric patterns that exhibit variability in texture, scale, and structure. To achieve this, a specialized multi-scale local feature extractor is employed, leveraging diverse receptive fields to effectively capture features at varying levels of granularity. This enables the model to focus on both fine-grained details, such as thread-level textures, and larger-scale patterns, such as repeating motifs or structural alignments within the fabric. By incorporating multi-scale analysis, the extractor ensures that the generated features are not only detailed and diverse but also contextually coherent, maintaining the relationships between local features across the image. This holistic method facilitates robust

feature representation, laying a strong foundation for accurate and efficient fabric image retrieval.

Feature Fusion with Mixer Modules: The extracted local descriptors are further fused using Mixer Modules, inspired by the MLP-Mixer [106]. These modules perform token-mixing and channel-mixing operations, facilitating the integration of multi-scale local descriptors into a unified representation while effectively learning global relationships.Notably, the hidden layer width of the Mixer Modules is expanded to improve their capacity for fusing multi-scale features. This enhancement ensures that both fine-grained texture details and broader contextual information are effectively captured, enabling a more robust representation of fabric images.

Progressive Triplet Mining Strategy: To optimize model performance and facilitate the learning of discriminative image descriptors, a progressive triplet mining strategy is introduced, systematically improving convergence and retrieval accuracy by gradually increasing the complexity of training samples. The process begins with an easy mining phase, where the model is trained using simple positive and negative pairs that are well-separated in the feature space. This initial phase stabilizes the learning process, allowing the model to establish a strong foundation for feature representation without being hindered by challenging examples. As training advances, the strategy transitions to a semi-hard mining phase, where semi-hard triplets—those with negative samples closer to the anchor than the easy negatives but still distinguishable—are introduced. This phase encourages the model to refine its feature embeddings, effectively narrowing the decision boundaries and improving its ability to separate similar yet distinct samples. Finally, the strategy culminates in a hard mining phase, where the model is exposed to hard triplets with negative samples that are very close to, or even overlapping with, the anchor in the feature space. This final phase pushes the model to its limits, maximizing the learning of discriminative features and ensuring robust separation between positive and negative pairs. By progressively increasing the difficulty of the training samples,

this method ensures a smooth learning trajectory, enabling the model to develop well-structured and highly discriminative feature embeddings, ultimately leading to significant improvements in fabric image retrieval accuracy.

The proposed framework integrates multi-scale feature extraction and progressive triplet mining to address the challenges of fabric retrieval. The multi-scale feature extractor captures detailed local features and contextual relationships, while the Mixer Modules fuse these features into a unified global representation. Simultaneously, the triplet mining strategy ensures efficient and effective training by systematically leveraging samples of increasing difficulty. By combining these components, the framework offers a robust solution for capturing both local intricacies and global contextual features, optimizing feature diversity, and enhancing retrieval accuracy in complex fabric images.

The remainder of this chapter is organized as follows: Section 4.2 reviews related work. Section 4.3 details the proposed method. Section 4.4 presents and discusses the experimental results. Finally, Section 4.5 concludes this chapter.

4.2 Related Works

The exploration of fabric recognition through image retrieval has evolved significantly since its inception in 2015 [36, 105], experiencing rapid technological advancements and methodological innovations. Early efforts in fabric retrieval utilized hand-crafted feature extraction techniques, which have since been complemented—and in many cases, supplanted—by deep learning-based approaches that offer enhanced feature representation and retrieval performance.

In hand-crafted methods, several studies [104, 32, 36] employed local color features to generate global image descriptors, with Jing et al.[32] integrating GIST[107] texture features to enhance retrieval robustness. Other methods [103, 56, 34, 55, 105] focused on local texture descriptors. For instance, Zhang et al.[105] leveraged the rotation invariance of Local Binary Patterns (LBP)[108] to represent texture features in fabric images, while Li et al.[56] combined local shape descriptors with

Haralick features[109] for retrieval tasks. Similarly, Li et al.[34] used a combination of SURF[28] keypoints and GIST features to generate a more descriptive global representation for fabric retrieval.

In contrast, deep learning-based approaches have demonstrated superior performance through the use of pre-trained models and transfer learning. Many methods [96, 97, 98, 99, 100, 101] adapted convolutional neural networks (CNNs) for fabric retrieval by fine-tuning them to account for fabric-specific characteristics. For example, Xu et al.[97] trained a Siamese network using hard negative pairs, achieving rapid convergence and impressive results in fabric retrieval. Xiang et al.[98] introduced a multi-task learning architecture to enhance feature robustness, employing ResNet [11] as the backbone for generating improved feature representations. In another instance, Tena et al.[96] developed a lightweight network with three convolutional layers and two fully connected layers to compress features into 120-dimensional vectors for ikat fabric retrieval, achieving high accuracy. Moreover, hybrid approaches[95, 102] have emerged, integrating CNN-based features with traditional descriptors for enhanced performance. For instance, Gu et al.[95] fused AlexNet[14] features with LBP [108] and HOG [54], calculating similarity using the Similarity Correlation Coefficient (SCC) for colored spun fabric retrieval. Zhang et al.[102] proposed a two-stage strategy, combining Inception-V3[12] for global feature extraction with ORB [57] for local feature matching. However, the limited receptive field and low robustness of ORB features constrained their effectiveness in capturing intricate fabric textures.

In fabric retrieval, feature fusion techniques have proven critical for capturing the diverse and complex details of fabric images. Methods such as LBP [108], SIFT [25], and histograms [110] aggregate local features into global descriptors through statistical or clustering-based approaches. VLAD [18] and NetVLAD [67] techniques enhance this process by clustering features and using cluster centers or distributions as descriptors. While effective, these methods struggle to capture contextual relationships between local features during fusion, limiting their robustness for complex

fabrics. The introduction of MLP-Mixer [106], a novel architecture relying solely on Multi-Layer Perceptrons (MLPs), represents a significant shift in feature fusion strategies. By mixing spatial and channel dimensions, MLP-Mixer effectively learns contextual relationships between local patches and integrates them into global descriptors. Although originally designed for classification tasks, its mechanisms hold substantial promise for retrieval tasks, particularly in addressing the contextual limitations of traditional feature fusion methods [111].

In summary, while substantial progress has been made in fabric retrieval, significant challenges remain. Current methods often fall short in optimizing local feature extraction and contextual relationship analysis, particularly for complex fabrics like lace. For example, although Zhang et al. [102] effectively combined Inception-V3 global features with ORB local features, the limited receptive field and low robustness of ORB hinder precise matching. Addressing these gaps requires novel approaches that integrate robust multi-scale local feature extraction with advanced fusion techniques to better capture the intricate patterns and subtle variations characteristic of fabric images.

4.3 Framework of MLDF

Given a query image $\boldsymbol{I}^q$ and a set of retrieval images $\{\boldsymbol{I}^r\}$, the retrieval process involves extracting feature representations for these images, computing the similarity between $\boldsymbol{I}^q$ and each image in $\{\boldsymbol{I}^r\}$, and ranking the results based on similarity scores. Existing methods [9, 10, 11, 12, 13, 14, 15, 16] typically rely on convolutional neural networks (CNNs) to directly learn global features of fabric images. The similarity computation in these methods can be expressed as:

$$\text{Similarity} = \text{Sim}(\text{CNN}(\boldsymbol{I}^q), \text{CNN}(\boldsymbol{I}^r)), \tag{4.1}$$

where $\text{Sim}(\cdot)$ represents a similarity measurement function, and $\text{CNN}(\cdot)$ uses operations such as global average pooling or flattening to aggregate feature maps into

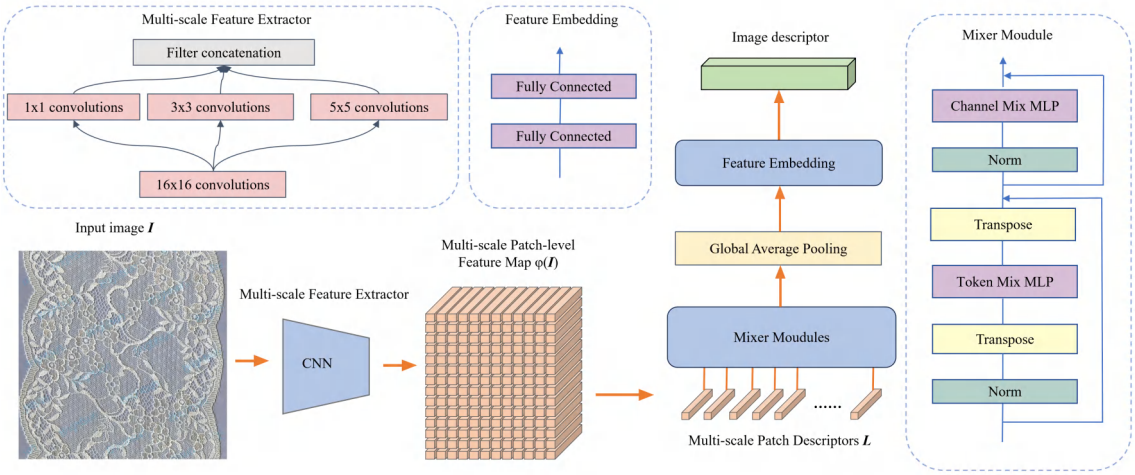


Figure 4.3: Overview of the proposed framework. The model first utilizes a multi-scale feature extractor to divide the image into 16×16 patches, extracting local features at different receptive field scales to generate a multi-scale patch-level feature map. It then applies a flattening operation to obtain multi-scale patch descriptors for each area. Subsequently, the model employs multiple mixer modules with token mixing and channel mixing mechanisms to fuse the multi-scale patch descriptors, thereby creating a global feature representation. This representation is then remapped through feature embedding to reduce dimensionality, resulting in the final image descriptor.

one-dimensional global feature descriptors.

In contrast, the proposed framework captures both global and local information to address the limitations of existing methods, as illustrated in Figure 4.3. For a given image $\boldsymbol{I}$, multi-scale local feature descriptors $\boldsymbol{L}$ are extracted using a learnable multi-scale feature extractor $\varphi(\cdot)$, as defined by:

$$\boldsymbol{L} = \text{Flatten}(\varphi(\boldsymbol{I})), \tag{4.2}$$

where $\varphi(\cdot)$ divides the image into multiple patches using a 16×16 convolution, followed by feature extraction through convolutions with varying kernel sizes. These multi-scale local features are concatenated to construct a comprehensive patch-level feature map, as illustrated in Figure 4.3.

Subsequently, the multi-scale patch descriptors are fused into a global image descriptor using the $Fusion(\cdot)$ function, which preserves local feature information while fully considering contextual relationships between patches. The global descrip-

tor integrates both local and global features, and the proposed framework computes similarity as:

$$\text{Similarity} = \text{Sim}(\text{Fusion}(\varphi(\boldsymbol{I}^q)), \text{Fusion}(\varphi(\boldsymbol{I}^r))), \tag{4.3}$$

The framework is designed to overcome the challenges of identifying fabric images with complex surface characteristics. For samples that are globally similar, relying solely on global semantic features often fails to distinguish subtle differences. Similarly, local feature-based methods, such as SURF, struggle to differentiate images due to high similarity in small localized regions. As shown in Figure 4.2, it is challenging to determine similarity accurately based on global or small-scale local features alone.

To address these challenges, the proposed architecture introduces a novel approach that extracts local descriptors from patches at multiple scales and fuses them using mixer modules. This enables a balanced analysis that captures fine-grained details while preserving global context. By synthesizing features across multiple receptive fields, the framework effectively improves the retrieval accuracy of fabric images with intricate and complex patterns.

4.3.1 Multi-scale Feature Extractor

The multi-scale feature extractor begins by applying a series of convolutions to decompose the image into multiple patches, effectively capturing foundational patterns and textures. Initially, a 16×16 convolution is used to partition the image into patches, enabling the extraction of essential spatial structures. To further enhance the feature extraction process, subsequent convolutions with varying kernel sizes are employed. These progressively expanding convolutions are capable of learning features at different scales, capturing both fine-grained details and broader contextual patterns. Finally, the patch features generated by convolutions of various kernel sizes are concatenated to construct a comprehensive multi-scale patch-level feature map. The resulting multi-scale local descriptor is mathematically defined as:

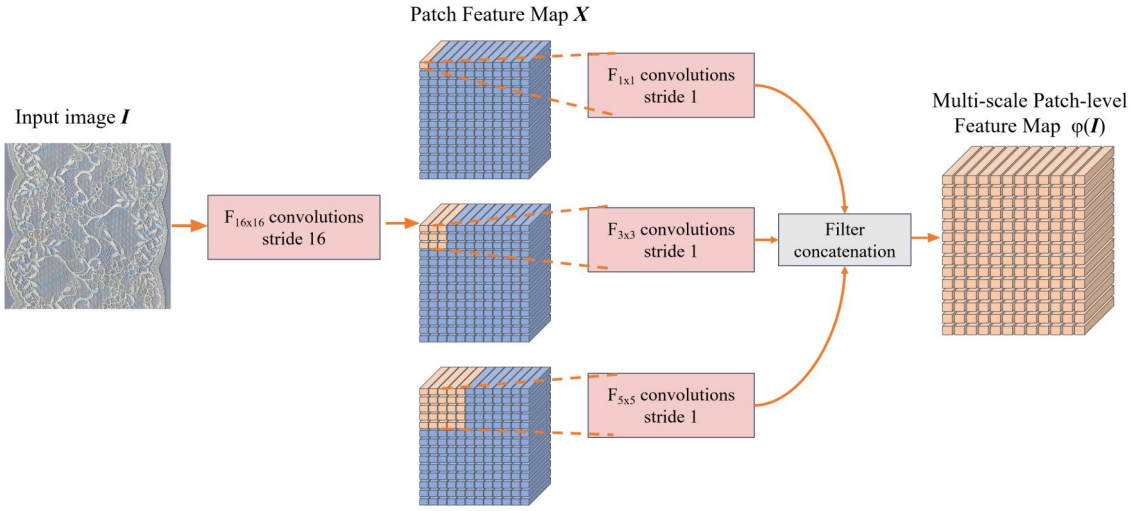


Figure 4.4: Multi-scale feature extractor.

$$\boldsymbol{l}_d = \mathrm{Con}\left(\sum_{c=1}^{C} \boldsymbol{X}_c \cdot \boldsymbol{F}_{c,d}^{1\times1}, \sum_{c=1}^{C} \boldsymbol{X}_c \cdot \boldsymbol{F}_{c,d}^{3\times3}, \sum_{c=1}^{C} \boldsymbol{X}_c \cdot \boldsymbol{F}_{c,d}^{5\times5}\right), \quad \forall d \in \{1, 2, \ldots, D\}, \tag{4.4}$$

where $\boldsymbol{l}_d$ represents the d-th element of the multi-scale local feature descriptors $\boldsymbol{l}$. $\boldsymbol{X}_c$ denotes the pixel value at channel c of the patch feature map $\boldsymbol{X}$, while $\boldsymbol{F}_{c,d}$ represents the corresponding weight in the convolution kernels of different sizes for the d-th dimension of the local feature descriptor. D denotes the dimensionality of the multi-scale local feature descriptors $\boldsymbol{l}$.

Unlike the MLP-Mixer [106], which primarily relies on flattening operations to extract patch features, the proposed multi-scale feature extraction process retains spatial hierarchies and relationships. This approach not only preserves essential spatial structures but also allows for a more comprehensive analysis of local and global interactions within the image. Compared to hand-crafted features and traditional CNN-based features, the proposed multi-scale feature extractor offers the advantage of iterative refinement through training and incorporates larger, diverse receptive fields. This enables a deeper understanding of contextual relationships within local regions, providing enhanced feature representation for complex image patterns.

4.3.2 Local Descriptors Fusion

The Mixer module processes an input consisting of multi-scale local feature descriptors. For an ensemble of D multi-scale local feature descriptors, where each descriptor has a size of N, the input can be represented as $\boldsymbol{L} \in \mathbb{R}^{D \times N}$. The operation of the Mixer module is divided into two main stages: Token Mixing and Channel Mixing.

Token Mixing: Each local feature descriptor corresponding to an image patch is treated as a token. The Token Mixing stage blends the features of these descriptors and maps them into a new feature space, facilitating the integration of spatial information across patches. The process of token mixing is defined as:

$$\begin{aligned} \boldsymbol{Y}' &= \mathrm{MLP}_{\mathrm{tokens}}\left(\boldsymbol{L}^T\right), \\ \boldsymbol{Y} &= \boldsymbol{Y}'^T + \boldsymbol{L}, \end{aligned} \tag{4.5}$$

where the input matrix $\boldsymbol{L}$ is first transposed to $\boldsymbol{L}^T \in \mathbb{R}^{N \times D}$. The transposed matrix $\boldsymbol{L}^T$ is then passed through the token mixing MLP layer, $\mathrm{MLP}_{\mathrm{tokens}}$, to produce $\boldsymbol{Y}'$. Subsequently, $\boldsymbol{Y}'$ is transposed back to realign with the original dimensions of the feature descriptors, resulting in $\boldsymbol{Y} \in \mathbb{R}^{D \times N}$.

Channel Mixing: Following Token Mixing, the spatial information within each image patch is blended, effectively expanding the receptive field. The Channel Mixing stage then mixes the channels across all tokens, capturing contextual and semantic information for the entire image. This process is defined as:

$$\boldsymbol{Z} = \mathrm{MLP}_{\mathrm{channels}}\left(\boldsymbol{Y}\right) + \boldsymbol{Y}, \tag{4.6}$$

where $\boldsymbol{Y}$ is fed into the channel mixing MLP layer, $\mathrm{MLP}_{\mathrm{channels}}$, to produce the output matrix $\boldsymbol{Z} \in \mathbb{R}^{D \times N}$. This stage mixes features across the channels of each token, further enriching the representation.

Compared to the original Mixer module proposed in [106], enhancements were introduced to adapt the module for the fabric image retrieval task. The hidden

layers of the Mixer module were widened, and the input feature width was expanded from 768 to 2304 to accommodate local texture features extracted at three distinct scales. Additionally, L2 regularization was applied to mitigate overfitting issues commonly encountered in fabric image retrieval tasks, where intricate patterns and subtle texture variations can lead to model over-reliance on specific features. By penalizing large weights, L2 regularization promotes better generalization across diverse fabric patterns.

For the mixed feature $\boldsymbol{Z} \in \mathbb{R}^{D \times N}$, a Global Average Pooling layer is used to transform the multi-dimensional feature matrix into a single-dimensional global feature descriptor. After pooling, the global feature descriptor is further processed through a Fully Connected (FC) layer, which encodes the descriptor by compressing it into a compact and manageable form. This encoding step serves to distill the essential characteristics of the global features, reducing dimensionality while retaining the most informative aspects for retrieval.

4.3.3 Optimizing Feature Space with Metric Learning

Metric learning [112, 113, 114, 115, 116] is a powerful approach for capturing subtle visual differences within images by learning a distance measure. It works by minimizing the distance between images of the same category while maximizing the distance between those belonging to different categories. Through this process, images with similar textures or patterns are positioned closer in the feature space, while dissimilar fabrics are distinctly separated.

Triplet Loss: Metric learning is implemented through the construction of triplets [117] and the application of Triplet Loss [68], which serves as the training objective. A triplet consists of three images: an anchor image $\boldsymbol{A}$, a positive image $\boldsymbol{P}$ that belongs to the same class as the anchor, and a negative image $\boldsymbol{N}$ from a different class. The goal is to train the model so that the anchor and positive images are drawn closer together in the feature space, while the anchor and negative images are separated by a larger distance. This relationship is enforced using the

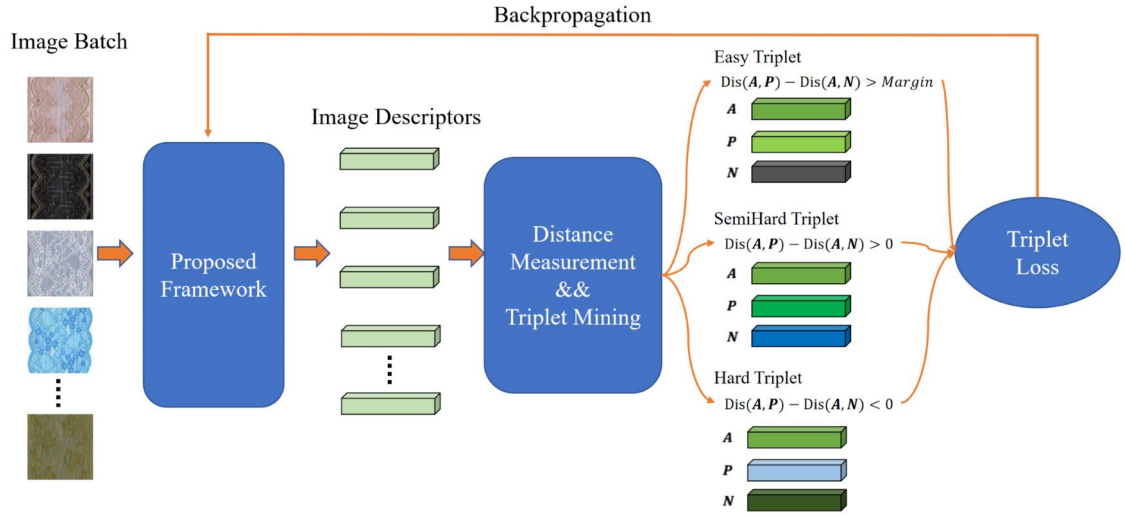


Figure 4.5: Triplet mining strategy.

triplet loss function, defined as:

$$\text{TripletLoss} = \text{Max}\left(\text{Dis}\left(\boldsymbol{A}, \boldsymbol{P}\right) - \text{Dis}\left(\boldsymbol{A}, \boldsymbol{N}\right) + \text{Margin},\ 0\right), \tag{4.7}$$

where Dis$(\cdot)$ represents the distance function, and cosine distance is used in this implementation as it provides a robust similarity measure independent of vector magnitude. The parameter Margin specifies the required minimum difference between the anchor-positive and anchor-negative distances. The loss function penalizes the model if the anchor-positive distance exceeds the anchor-negative distance by less than the margin, guiding parameter updates to improve feature separation. Conversely, if the required margin is already satisfied, the loss is zero, indicating no adjustments are necessary.

Triplet Mining Strategy: The triplet mining strategy is essential for determining training effectiveness and efficiency. It ensures that the most informative triplets are utilized for learning, accelerating convergence while enhancing the models overall performance. By focusing on challenging, yet meaningful examples, it avoids issues such as overfitting to outliers, which could destabilize the learning process.

As illustrated in Figure 4.5, triplets are dynamically constructed during train-

ing based on the relationships among input samples. Within each training batch, image descriptors are extracted, and the similarity between all pairs of samples is calculated. A mining sampler then identifies triplet groups based on the sample labels. These triplet groups are used to calculate the triplet loss based on feature similarities, enabling the model to progressively refine its parameters.

Three strategies are employed for triplet construction: "easy," "semihard," and "hard." The "hard" mining strategy selects triplets where the negative image is closer to the anchor than the positive image, violating the margin and posing a significant challenge to the model. The "semihard" strategy involves triplets where the negative image is farther from the anchor than the positive image but still within the margin, encouraging the model to further refine its representation. The "easy" strategy uses triplets that do not violate the margin, where the negative image is sufficiently farther from the anchor than the positive image, reinforcing stable feature learning. Each strategy adjusts triplet selection dynamically, ensuring that the relative distances between the anchor, positive, and negative images are leveraged to optimize model convergence and retrieval performance.

4.4 Experiments

This section provides a detailed explanation of the implementation and training methodologies for MLDF. To evaluate its effectiveness, a comprehensive series of experiments was conducted on the self constructed dataset.

4.4.1 Implementation

Architecture: The proposed method was implemented using PyTorch, with all cropped ROI images resized to 224×224 for both training and testing. Feature extraction was performed using a multi-scale feature extractor, designed to capture intricate local features from the ROI images. The resulting multi-scale local feature descriptors, sized 196×2304, were processed through a mixer network. This mixer

network consists of 8 mixer modules, each designed with intermediate feature dimensions of 384 for token mixing and 3072 for channel mixing. To ensure stable gradient flow and mitigate the risk of gradient vanishing, normalization and residual stacking were applied both before and after the mixer modules. The mixed features were subsequently processed through an average pooling layer and mapped into a low-dimensional space via two fully connected layers, resulting in the final feature descriptor.

Training: Standard data augmentation techniques were utilized to enhance robustness and reduce the risk of overfitting. These techniques included random adjustments to brightness, contrast, and color. Optimization was performed using the Adam optimizer, initialized with a learning rate of 0.0005 and a weight decay factor of 0.0001. Training was conducted over 30 epochs, with the learning rate decaying by a factor of 0.8 every 5 epochs. The model was trained on an RTX 4060Ti GPU with 16GB of memory. A progressive triplet mining strategy was adopted to ensure effective model learning. For the initial 5 epochs, the ”easy” strategy was employed to help the model learn basic distinctions between classes without introducing excessive complexity. This was followed by the ”semihard” strategy for epochs 5 to 25, introducing moderately challenging triplets to refine the models feature representations. In the final stage, from epochs 25 to 30, the ”hard” strategy was applied, utilizing the most difficult triplets to further enhance the models discriminative capabilities. This gradual progression in triplet difficulty ensured effective feature learning while minimizing the risk of training stagnation or overfitting to outliers.

Evaluation: The fabric dataset was divided into training, validation, and test sets using an 8:1:1 ratio. This allocation provided a robust training base while ensuring sufficient data for validation and testing to comprehensively evaluate the models performance. To ensure a rigorous evaluation of the models performance, standard evaluation metrics commonly used in the field of information retrieval are employed. These include precision and recall, as well as mean Average Precision

(mAP) and F1-Score, which provide a more comprehensive analysis of retrieval effectiveness. For Top-K retrieval results, the calculation methods for precision, recall, mAP, and F1-Score are defined as follows:

$$\begin{aligned} P@K &= \frac{\text{RelevantNum}@K}{K}, \\ R@K &= \frac{\text{RelevantNum}@K}{\text{RelevantNum}}, \end{aligned} \tag{4.8}$$

$$\begin{aligned} mP@K &= \frac{1}{N} \times \textstyle\sum_{n=1}^{N} P_n@K, \\ mR@K &= \frac{1}{N} \times \textstyle\sum_{n=1}^{N} R_n@K, \end{aligned} \tag{4.9}$$

$$mAP@K = \frac{1}{N} \times \textstyle\sum_{n=1}^{N} \frac{1}{K} \times \sum_{k=1}^{K} P_n@k, \tag{4.10}$$

$$F1@K = 2 \times \frac{mP@K \times mR@K}{mP@K + mR@K}. \tag{4.11}$$

In these equations, $P@K$ represents precision at a Top-K level, which measures the proportion of relevant images among the top K retrieved images. Similarly, $R@K$ represents recall at a Top-K level, reflecting the fraction of all relevant images in the dataset that are successfully included in the Top-K results.

The metric $mP@K$ represents the mean precision across all test queries at the Top-K level, while $mR@K$ represents the mean recall. These metrics provide a robust evaluation of retrieval performance averaged over the dataset. Additionally, $mAP@K$ (mean Average Precision at a Top-K level) serves as a comprehensive measure that captures the precision across all recall levels, offering insight into the overall ranking quality.

Finally, $F1@K$, calculated as the harmonic mean of $mP@K$ and $mR@K$, bal-

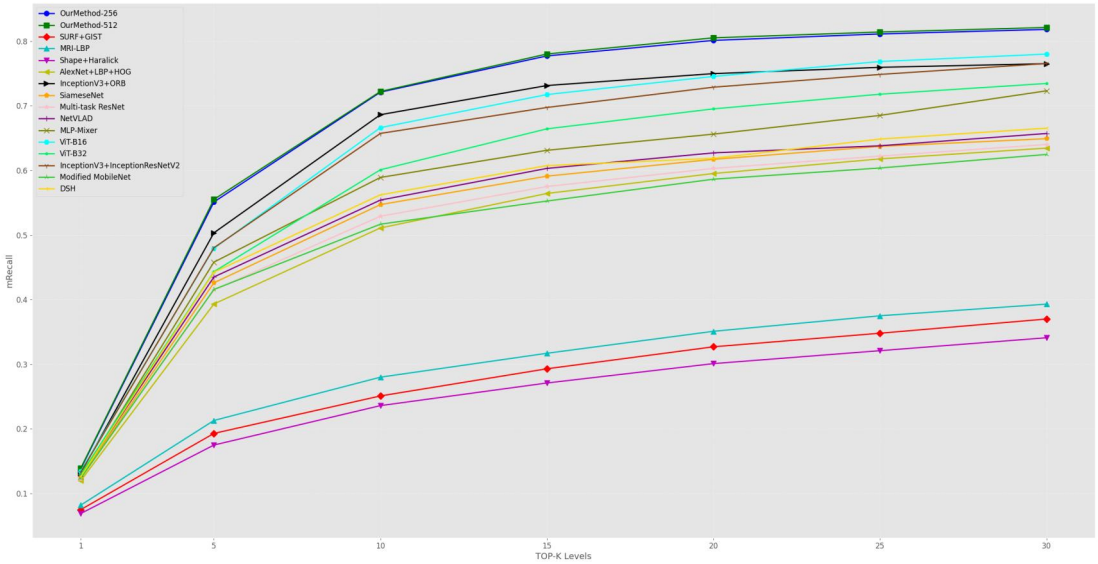


Figure 4.6: mRecall comparison of different methods on test dataset.

ances precision and recall, providing a single, unified metric for evaluating retrieval performance. These metrics collectively offer a robust framework for assessing both the accuracy and ranking capabilities of the proposed model in retrieval tasks.

4.4.2 Experiment Results

Table 4.1, Figure 4.6, and Figure 4.7 provide a comprehensive comparative analysis of various state-of-the-art techniques and the proposed method in the context of fabric image retrieval. The fabric dataset presents significant challenges due to its intricate patterns, diverse designs, and subtle texture variations, which require a highly adaptive and detailed feature extraction mechanism to achieve effective

Table 4.1: Performance comparison of different methods on the test dataset.

Method	Field	Latency (ms)	$mP@1$	$mP@5$	$mP@10$	$mAP@10$	$F1@10$
SURF+GIST [55]	Fabric	229	55.12%	27.14%	18.45%	34.67%	22.14%
MRI-LBP [105]	Fabric	71	61.84%	30.15%	20.32%	36.34%	24.51%
Shape+Haralick [34]	Fabric	52	48.22%	25.49%	17.89%	32.65%	20.31%
AlexNet+LBP+HOG [95]	Fabric	191	77.61%	44.83%	25.10%	48.22%	34.22%
InceptionV3+ORB [102]	Fabric	95	**94.03%**	59.22%	32.26%	61.92%	43.27%
SiameseNet [97]	Fabric	46	89.40%	56.41%	30.85%	58.62%	40.76%
Modified CNN [96]	Fabric	7	81.82%	46.36%	25.98%	51.46%	36.64%
Multi-task ResNet50 [98]	Fabric	33	90.91%	54.83%	29.41%	57.49%	40.52%
InceptionV3+InceptionResNetV2 [118]	Fabric	79	91.26%	56.08%	30.63%	59.43%	41.33%
Modified MobileNet [119]	Fabric	23	85.66%	48.81%	27.41%	52.87%	37.32%
DSH [120]	Fabric	27	84.97%	47.89%	26.96%	52.64%	37.76%
NetVLAD [67]	General	51	87.31%	53.98%	28.09%	55.32%	38.89%
MLP-Mixer [106]	General	44	90.11%	55.52%	30.83%	58.66%	40.17%
ViT-B16 [17]	General	52	92.38%	56.87%	30.94%	58.51%	41.28%
ViT-B32 [17]	General	52	91.87%	56.24%	31.05%	60.06%	41.31%
Proposed Method-256	Fabric	48	92.67%	59.43%	31.96%	61.88%	43.29%
Proposed Method-512	Fabric	48	93.13%	**59.91%**	**32.56%**	**62.24%**	**44.44%**

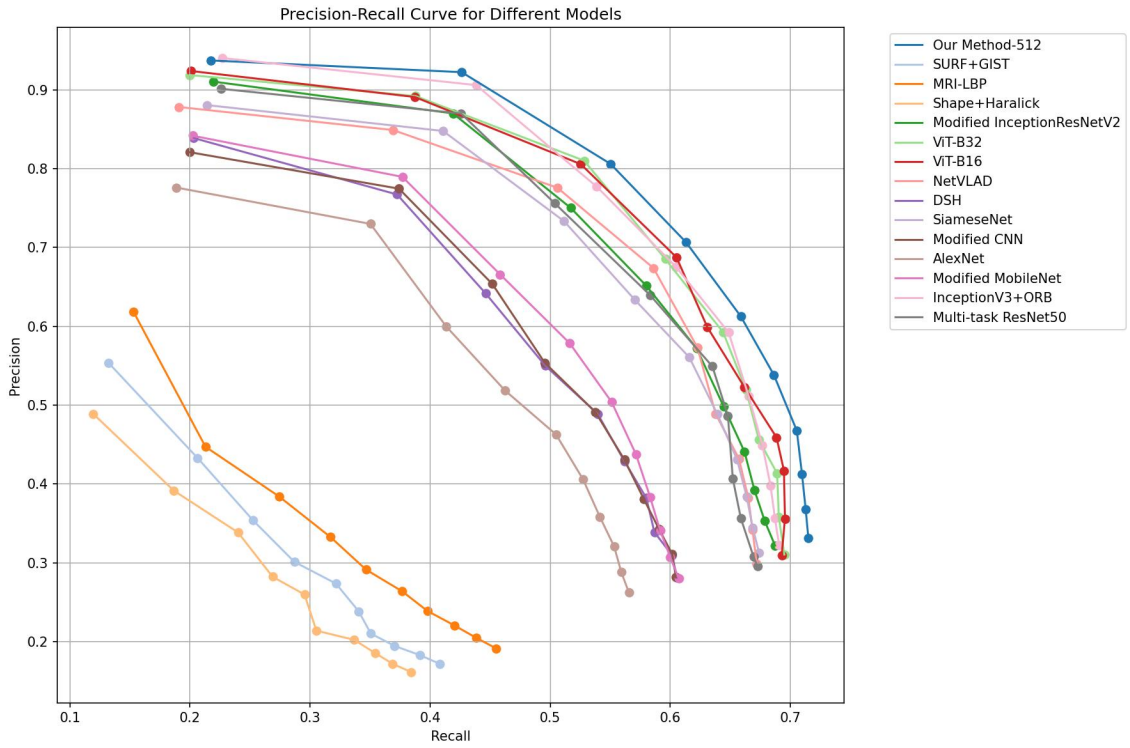


Figure 4.7: Precision-Recall curves for different methods on test dataset.

retrieval.

From Table 4.1, it can be observed that the handcrafted feature-based methods, such as SURF+GIST [55], MRI-LBP [105], and Shape+Haralick [34], exhibit lower retrieval accuracy. For instance, their mAP@10 values range between 32.65% and 36.34%, while their $F1$@10 also remains below 25%. These methods primarily rely on pre-defined descriptors, which can be effective for relatively simple textures but often underperform in more complex scenarios like lace or multi-structured fabric images.

In contrast, deep learning-based methods show marked improvements. Approaches such as SiameseNet [97] and Multi-task ResNet50 [98] achieve higher mAP@10 and $F1$@10, indicating their ability to learn more discriminative representations from data. Notably, InceptionV3+ORB [102] attains the highest mP@1 of **94.03**%, underscoring the effectiveness of combining global CNN features with local descriptor matching for certain fabric classes. However, the reliance on handcrafted local features like ORB introduces potential limitations in robustness and generalizability.

Focusing on the proposed method, both the 256-dimensional and 512-dimensional versions surpass most competing approaches in mAP@10 and $F1$@10. Specifically, the 512-dimensional variant achieves the highest mAP@10 (**62.24**%) and $F1$@10 (**44.44**%) on the test set. Although its mP@1 (93.13%) is slightly lower than InceptionV3+ORB, the overall retrieval performance gains (as indicated by mAP@10 and $F1$@10) clearly demonstrate the robustness and adaptability of the proposed approach. These improvements can be attributed to two core design elements:

- **Multi-scale Local Descriptor Extraction:** By capturing both fine-grained local features and broader contextual patterns, the proposed model effectively addresses variations in texture and color, which is particularly advantageous for high intra-class variability fabrics.
- **Mixer-based Feature Fusion:** This fusion mechanism aggregates local descriptors into a global representation, enabling the model to learn contextual relationships among different patches and ultimately boosting retrieval accuracy.

Furthermore, the proposed method exhibits competitive latency (48 ms) that is comparable to or better than many existing CNN-based methods. This balance between speed and accuracy suggests that the multi-scale plus mixer-fusion strategy is a practical solution for real-world scenarios where both inference time and retrieval quality are critical.

In summary, the quantitative results confirm that the proposed framework effectively addresses the complexities of fabric images, striking a promising balance between computational efficiency and retrieval accuracy. By integrating multi-scale analysis and mixer-based feature fusion, the proposed approach offers a robust and versatile solution for challenging fabric image retrieval tasks.

4.4.3 Ablation Study of The Multi-scale Feature Extractor

The multi-scale feature extractor serves as a cornerstone of the architecture, ensuring comprehensive encoding of discriminative features essential for robust fabric image retrieval. Its design integrates hierarchical and scale-aware features, allowing the model to capture intricate patterns and subtle texture variations present in fabric images. To comprehensively assess its impact, a series of ablation experiments were conducted, as detailed in Table 4.2. These experiments systematically evaluated the performance of the multi-scale feature extractor in comparison to alternative strategies, including a simple flattening operation and single-scale feature extraction methods with varying receptive field sizes.

The flattening operation demonstrates significantly inferior performance across all metrics, achieving only 83.47% mP@1 and 52.01% mAP@10. This indicates that flattening lacks the capacity to effectively capture fine-grained details inherent to fabric images. The inability to aggregate contextual information from surrounding regions further diminishes its representation capabilities, leading to suboptimal retrieval performance.

The single-scale feature extractors exhibit distinct performance characteristics dependent on receptive field size. The 3×3 configuration achieves optimal results with $mP@1 = 89.45\%$ and $mAP@10 = 58.45\%$, surpassing the flattening baseline (83.47%/52.01%) by 3.98% and 6.44% respectively. This performance hierarchy reveals critical scale sensitivity: smaller 1×1 fields attain 86.12% mP@1 (3.33% below 3×3), while larger 9×9 fields degrade to 84.33% mP@1 (5.12% reduction). These systematic variations confirm three fundamental principles: (1) undersized fields (1×1) inadequately capture contextual pattern relationships, (2) medium scales

Table 4.2: MLDF Performance of different feature extractors on the test dataset.

Feature Extractor	Scale	mP@1	mP@5	mP@10	mR@1	mR@5	mR@10	mAP@10	F1@10
Flattening	-	83.47%	49.32%	26.15%	17.57%	56.10%	58.55%	52.01%	39.83%
Single-scale	1×1	86.12%	54.74%	28.38%	18.79%	60.38%	61.87%	55.23%	40.73%
Single-scale	3×3	89.45%	54.92%	30.92%	20.78%	61.04%	66.65%	58.45%	41.76%
Single-scale	5×5	85.01%	51.93%	26.86%	17.04%	57.13%	60.74%	53.95%	39.56%
Single-scale	7×7	84.82%	51.10%	25.67%	16.99%	57.40%	57.84%	53.32%	39.02%
Single-scale	9×9	84.33%	47.33%	26.45%	16.91%	51.21%	60.53%	52.81%	38.74%
Multi-scale	1×1, 3×3	91.24%	57.50%	31.02%	21.05%	61.24%	70.87%	61.17%	42.09%
Multi-scale	1×1, 3×3, 5×5	**93.13%**	**59.91%**	**32.56%**	**22.58%**	**64.92%**	**69.64%**	**62.24%**	**44.44%**
Multi-scale	1×1, 3×3, 5×5, 7×7	89.40%	56.67%	30.41%	20.67%	62.24%	66.95%	60.17%	40.68%
Multi-scale	1×1, 3×3, 5×5, 7×7, 9×9	89.01%	56.23%	30.12%	20.59%	63.02%	67.34%	60.94%	40.43%

(3×3) optimally balance local texture precision and regional context integration, and (3) oversized fields ($\geq 5 \times 5$) introduce spatial noise that degrades discriminative power. The 3×3 configurations superiority stems from its capacity to simultaneously resolve thread-level textures (critical for lace fabrics) and maintain adjacent motif relationships (essential for plaid/stripe patterns), achieving a 6.98% absolute improvement over baseline in fine-grained discrimination.

The optimal multi-scale configuration (1×1, 3×3, 5×5) achieves peak performance with 93.13% mP@1 and 62.24% mAP@10, outperforming all single-scale variants by $\geq$3.68% in mP@1 and $\geq$3.79% in mAP@10. This demonstrates hierarchical fusion of complementary scales (fine 1×1 details, contextual 3×3 patterns, and structural 5×5 relationships) is critical for fabric retrieval.

Extending to larger scales (1×1, 3×3, 5×5, 7×7, 9×9) reduces mP@1 to 89.01% and mAP@10 to 60.94%, indicating excessive scale integration introduces feature redundancy. The performance degradation (4.12% mP@1 drop from optimal configuration) highlights the necessity of scale selection.

4.4.4 Ablation Study of The Feature Fusion

A series of experiments were conducted to evaluate the impact of varying the number of mixer modules on the MLDF models performance, as summarized in Table 4.3. These experiments examined the relationship between the number of mixer modules and various performance metrics, including latency, parameter count, and precision/recall metrics ($mP@k$, $mR@k$, mAP@10, and $F1$@10), when applied to the test dataset.

As the number of mixer modules increases from 2 to 8, consistent improvements are observed across most metrics. For instance, mP@1 rises from 83.82% to 93.13%,

Table 4.3: MLDF Performance of different mixer module numbers on test dataset.

Mixer Module Number	Latency(ms)	Params(M)	mP@1	mP@5	mP@10	mR@1	mR@5	mR@10	mAP@10	F1@10
No Mixer	8	6.64	78.48%	44.39%	23.92%	15.39%	46.48%	56.32%	47.79%	32.64%
2	14	11.39	83.82%	51.92%	26.22%	17.07%	56.01%	61.87%	52.44%	35.86%
4	26	21.14	87.47%	55.34%	29.51%	20.46%	61.53%	65.74%	57.29%	40.33%
6	37	30.90	91.48%	58.11%	31.68%	21.68%	62.41%	68.02%	60.93%	41.35%
8	48	40.65	**93.13**%	**59.91**%	32.56%	**22.58**%	**64.92**%	**69.64**%	**62.24**%	**44.44**%
10	60	50.40	93.11%	59.84%	**32.59**%	22.21%	64.28%	69.15%	61.55%	43.91%
12	73	60.16	92.55%	59.02%	32.15%	21.95%	64.41%	68.56%	62.02%	44.03%

and mAP@10 improves from 52.44% to 62.24%. Additionally, $F1$@10 increases from 35.86% to 44.44%. These gains can be attributed to the expanded representational capacity conferred by more mixer modules, which enables the model to capture and integrate fine-grained features more effectively.

However, increased model complexity also leads to higher computational costs. Latency grows from 14 ms at 2 modules to 48 ms at 8 modules, while the models parameters rise from 11.39M to 40.65M. Beyond 8 modules, the improvements in retrieval performance become marginal. For example, the 10-module configuration achieves mP@1 of 93.11% and mAP@10 of 61.55%, which are comparable to the 8-module case but at a latency of 60 ms and 50.40M parameters. Similarly, the 12-module arrangement yields mixed results: it slightly increases mAP@10 to 62.02% but reduces mP@1 to 92.55%.

These findings suggest that the 8-module setup strikes the most favorable balance between accuracy and efficiency, achieving the best overall combination of mP@1 (93.13%), mAP@10 (62.24%), and $F1$@10 (44.44%) at a moderate level of computational overhead. In scenarios calling for real-time inference or deployment to resource-constrained devices, the 8-module configuration appears especially well suited. Conversely, tasks with more relaxed latency constraints may consider 10 or 12 modules to potentially gain slight improvements in some metrics, albeit at a higher computational cost.

In conclusion, the number of mixer modules significantly influences both the performance and computational requirements of the MLDF model. While expanding the module count generally enhances the retrieval metrics, it also introduces trade-offs in latency and parameter size. Selecting an optimal configuration should therefore factor in the specific accuracy and efficiency demands of the target application context.

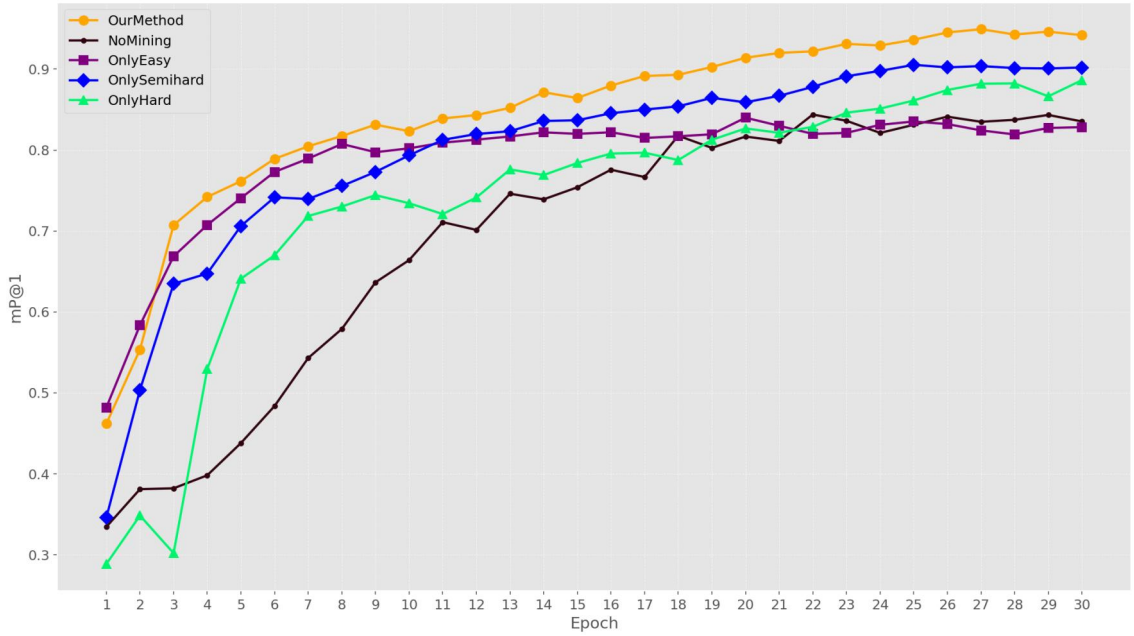


Figure 4.8: Comparison of $mP@1$ on the validation set under different training strategies on test dataset.

4.4.5 Ablation Study of The Triplet Minning

The triplet mining strategy is a crucial component of the proposed approach, designed to enhance model performance by progressively transitioning from ”easy” to ”hard” triplet mining stages. To validate the effectiveness of this strategy, the model structure was kept constant while training was conducted using different triplet mining strategies for a unified duration of 30 epochs. Since the triplet loss directly depends on the triplet mining strategy, the losses calculated for ”easy,” ”semi-hard,” and ”hard” triplets vary significantly. To evaluate the effectiveness of convergence, the $mP@1$ of the validation set at the end of each epoch was used as the primary metric.

As illustrated in Figure 4.8, the “No Mining strategy serves as a baseline and yields relatively suboptimal performance. For instance, it achieves an $mP@1$ of 84.37% and an $mR@1$ of 18.77% (Table 4.4), highlighting the need for a more effective mining scheme to further improve retrieval accuracy.

The “Only Easy strategy converges quickly during the initial phase of training but plateaus early. Its $mP@1$ hovers around the low 80% range and fails to achieve

further substantial gains, underscoring that easy triplets alone do not provide sufficient gradient signals for continued refinement.

In contrast, the "Only Hard strategy encounters a slower convergence rate at the beginning due to the large disparity in feature representations when selecting hard triplets. Although this approach eventually stabilizes and delivers moderate performance (with mP@1 of 88.73%), its high initial variability can pose challenges for consistent learning.

Similarly, the "Only Semi-hard strategy selects triplets with a more balanced range of difficulty, mitigating some early oscillations. While it offers improved performance over "Only Hard (e.g., mP@1 of 90.20%), the overall convergence speed and final accuracy can still be further enhanced.

By contrast, the *proposed* triplet mining strategy synthesizes the benefits of "easy, "semi-hard, and "hard'' triplets in a progressive manner. According to Table 4.4, it achieves the highest mP@1 of 93.13% and $F1$@10 of 44.44%. In the early epochs, focusing on easy triplets rapidly stabilizes the feature space. Progressively incorporating semi-hard and hard triplets fosters additional discrimination and fine-tuning, leading to superior final performance. This stepwise approach balances convergence speed and ultimate accuracy, demonstrating the advantage of dynamic triplet selection over static mining strategies.

These results emphasize the importance of employing a well-structured triplet mining strategy to maximize model performance. By combining the strengths of different mining strategies, the proposed approach demonstrates robust convergence, effective utilization of triplet loss, and superior retrieval accuracy. This progression from "easy" to "hard" mining stages provides a practical and efficient method for optimizing model training in fabric image retrieval tasks.

Table 4.4: MLDF Performance of different mining strategies on the test dataset.

Mining Strategy	mP@1	mP@5	mP@10	mR@1	mR@5	mR@10	mAP@10	$F1$@10
No Mining	84.37%	52.91%	27.46%	18.77%	56.86%	59.79%	51.67%	37.67%
Only Easy	83.51%	52.33%	27.17%	17.64%	55.27%	58.53%	49.02%	36.95%
Only Hard	88.73%	56.23%	29.87%	20.43%	60.41%	64.47%	58.06%	40.74%
Only Semi-hard	90.20%	57.72%	30.06%	20.65%	62.30%	66.27%	59.14%	41.33%
Proposed Method	**93.13**%	**59.91**%	**32.56**%	**22.58**%	**64.92**%	**69.64**%	**62.24**%	**44.44**%

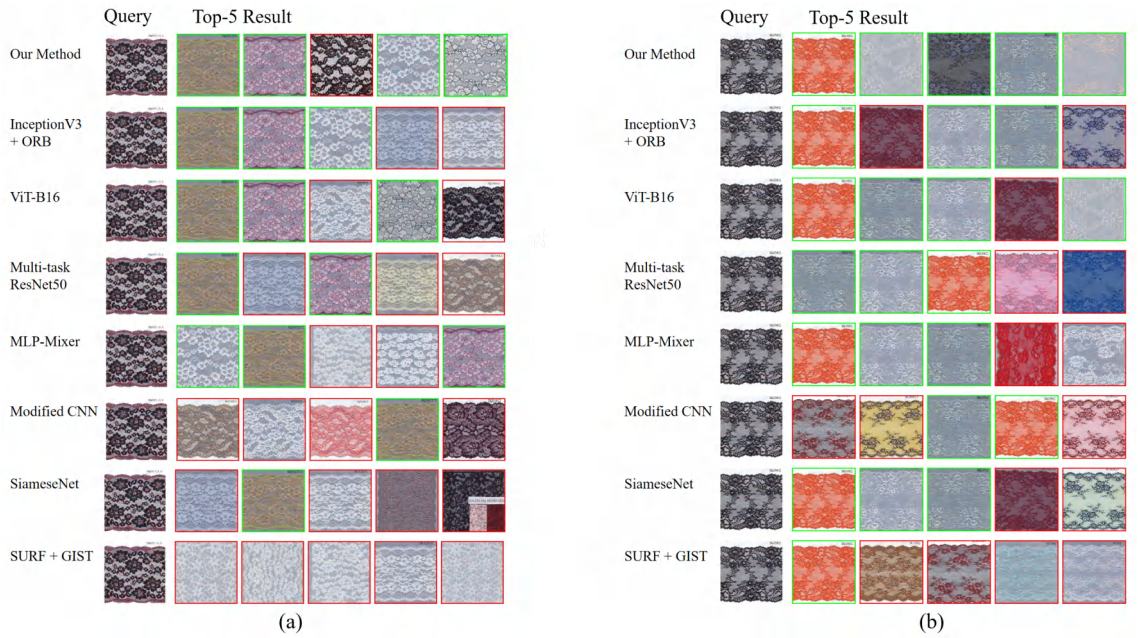


Figure 4.9: Comparison of sample retrieval results for two hard samples using different methods, with the top-5 retrieval results as a reference. The retrieval results highlighted in green boxes represent correct identifications, while those highlighted in red boxes represent incorrect identifications.

4.4.6 Visualization

Figure 4.9 presents the top-5 retrieval results for two representative hard samples, where correctly retrieved images are highlighted with a green border, and incorrectly retrieved ones are enclosed with a red border. From a visual standpoint, the incorrectly retrieved images bear a strong resemblance to the query images in terms of global silhouette and color. However, upon closer inspection, subtle differences in texture and finer details become apparent. Identifying these distinctions can be particularly challenging when excessive focus is placed on overly minute details, which can obscure broader contextual relationships. This observation underscores the importance of effectively balancing local and global feature extraction for retrieving fabrics with complex textures. Moreover, local features must not be extracted from overly small areas, as this could lead to the loss of crucial contextual information that defines the unique characteristics of fabric patterns.

The results highlight notable differences in the performance of various retrieval methods. Handcrafted feature-based approaches, such as SURF + GIST [55], tend to

focus on local features and often retrieve images with matching textural similarities. However, these methods fail to account for global contextual relationships, resulting in mismatches when dealing with intricate patterns or complex textures.

Deep learning-based models, such as ViT-B16 [17] and MLP-Mixer [106], exhibit improved performance in capturing global features, particularly color distribution and pattern scale. However, these models occasionally struggle with localized texture variations, especially in cases where subtle differences are key to differentiating between visually similar fabrics.

Two-stage retrieval methods, such as InceptionV3 + ORB [102], aim to address both global and local features by employing a CNN-based model for global feature extraction in the first stage, followed by ORB-based local feature matching in the second stage. While this approach combines the strengths of both feature types, the ORB features often lack the robustness necessary to handle the semantic ambiguities inherent in complex textured images. Consequently, these methods may retrieve incorrect results due to the high similarity of small-scale local regions, even when significant differences exist at a broader scale.

The proposed method demonstrates clear advantages in addressing these challenges. By effectively integrating global and local feature extraction, it achieves a fine balance between capturing the contextual information of the entire fabric design and discerning subtle differences in texture and pattern. This dual focus enables the model to identify fine-grained variations without losing sight of the broader fabric structure, resulting in superior performance for complex fabric retrieval tasks.

These findings emphasize the significance of leveraging robust multi-scale feature representations to capture the intricate interplay between local and global features. The ability to navigate this balance is particularly critical for applications involving fabrics with detailed and nuanced textures, where both local precision and global consistency are essential for accurate retrieval.

4.5 Chapter Summary

This chapter introduces a novel method for fabric image retrieval that incorporates the fusion of multi-scale local features. The proposed approach utilizes a custom-designed multi-scale local feature extractor, which segments images into multiple patches to capture local descriptors at varying scales. These descriptors are subsequently synthesized through a sophisticated mixer module, which dynamically learns and exploits the contextual relationships between the descriptors, leading to the generation of highly effective image descriptors tailored specifically for fabric retrieval tasks. Among the various components of the framework, the mixer module proves to have the most substantial impact on overall performance, underscoring the critical importance of model depth in feature extraction. In addition to the mixer module, the advanced triplet mining strategy employed during model training plays a pivotal role in improving the models convergence efficiency. The effectiveness of this strategy is particularly evident in the way it enhances the selection of informative training data, resulting in faster convergence and improved performance. By strategically mining triplets from a range of difficulty levels, the model is able to better capture the underlying relationships between the fabric images, thereby increasing the quality of the learned feature representations. While the multi-scale feature extraction module is not as influential as the mixer module or the triplet mining strategy, it still contributes to performance improvements by enhancing the granularity of the extracted features. This multi-scale approach ensures that local descriptors from different spatial resolutions are effectively combined, enabling the model to capture both fine-grained details and broader contextual features of the fabric images. Comparative experiments reveal significant improvements in retrieval precision and recall over existing fabric and general image retrieval methods. The performance gains demonstrate the robustness and effectiveness of the proposed approach, particularly in the context of fabric image retrieval, where complex textures and patterns must be accurately identified. These findings highlight the potential of the method to be applied beyond fabric retrieval, suggesting its applicability to

other domains such as textile design, pattern recognition, and even broader visual recognition tasks.

The results open new avenues for future research and development, with the possibility of extending the method to other types of images and exploring additional strategies for enhancing retrieval accuracy and efficiency. The promising performance observed in this study sets the foundation for further improvements and broader applications of the approach in various fields requiring advanced image retrieval techniques.

Chapter 5

Hierarchical Two-Stage Retrieval Method

This chapter proposes a Hierarchical Two-Stage Retrieval Method, which combines global and local feature representations to enhance robustness and precision.

Building upon the MLDF features from Chapter 4, this chapter proposes the Hierarchical Two-Stage Retrieval Method, which orchestrates the retrieval workflow through strategic feature utilization. The first stage employs global descriptors from MLDF for rapid candidate filtering, while the second stage leverages local descriptors for precision refinement. This architecture serves as the framework's decision engine. The unified feature space ensures seamless transition between stages, resolving inconsistencies in traditional two-stage systems.

The proposed method is rigorously evaluated on the self-constructed fabric image dataset. Experimental results demonstrate that the framework consistently outperforms state-of-the-art techniques in precision, recall, and mean Average Precision (mAP).

5.1 Introduction

For fabric image retrieval, CBIR systems can extract and analyze essential features like texture and pattern details. Traditional CBIR systems typically rely on hand-

crafted feature extraction methods, which involve manually designed algorithms to capture visual characteristics. Such methods struggle to adapt to the dynamic and evolving nature of fabric designs, leading to suboptimal retrieval results in scenarios where fabric patterns are highly diverse. In contrast, learning-based methods, particularly those utilizing deep learning models, automatically learn to extract image features directly from the data. Unlike handcrafted approaches, learning-based methods do not rely on pre-defined features; instead, they are trained on large datasets to learn optimal feature representations for specific tasks. This ability to learn complex feature hierarchies has significantly improved retrieval accuracy and robustness.

Despite the advantages of learning-based approaches, they also introduce new challenges. Specifically, while these methods excel at capturing high-level features, the ability to efficiently and accurately retrieve fabric images based on subtle local and global patterns requires additional strategies, such as multi-scale feature extraction and contextual information integration. Therefore, while learning-based methods have revolutionized the field, their application to fabric image retrieval requires careful consideration of both the strengths and limitations of different feature extraction techniques.

As shown in Figure 5.1, the challenges associated with complex fabric image retrieval can be categorized as follows:

Intricate Textures and Fine Details: Fabrics, particularly those such as lace, exhibit highly intricate textures and delicate details, where even minor variations in structure, thread arrangement, or color can significantly influence retrieval outcomes. Accurately capturing these fine-grained characteristics necessitates high-precision feature extraction methods that can effectively differentiate subtle variations while retaining robustness.

Scale and Viewpoint Variability: Fabric images are often captured under diverse conditions, leading to variations in scale and viewpoints. For instance, close-up images emphasize fine-grained textures and thread-level details, whereas distant

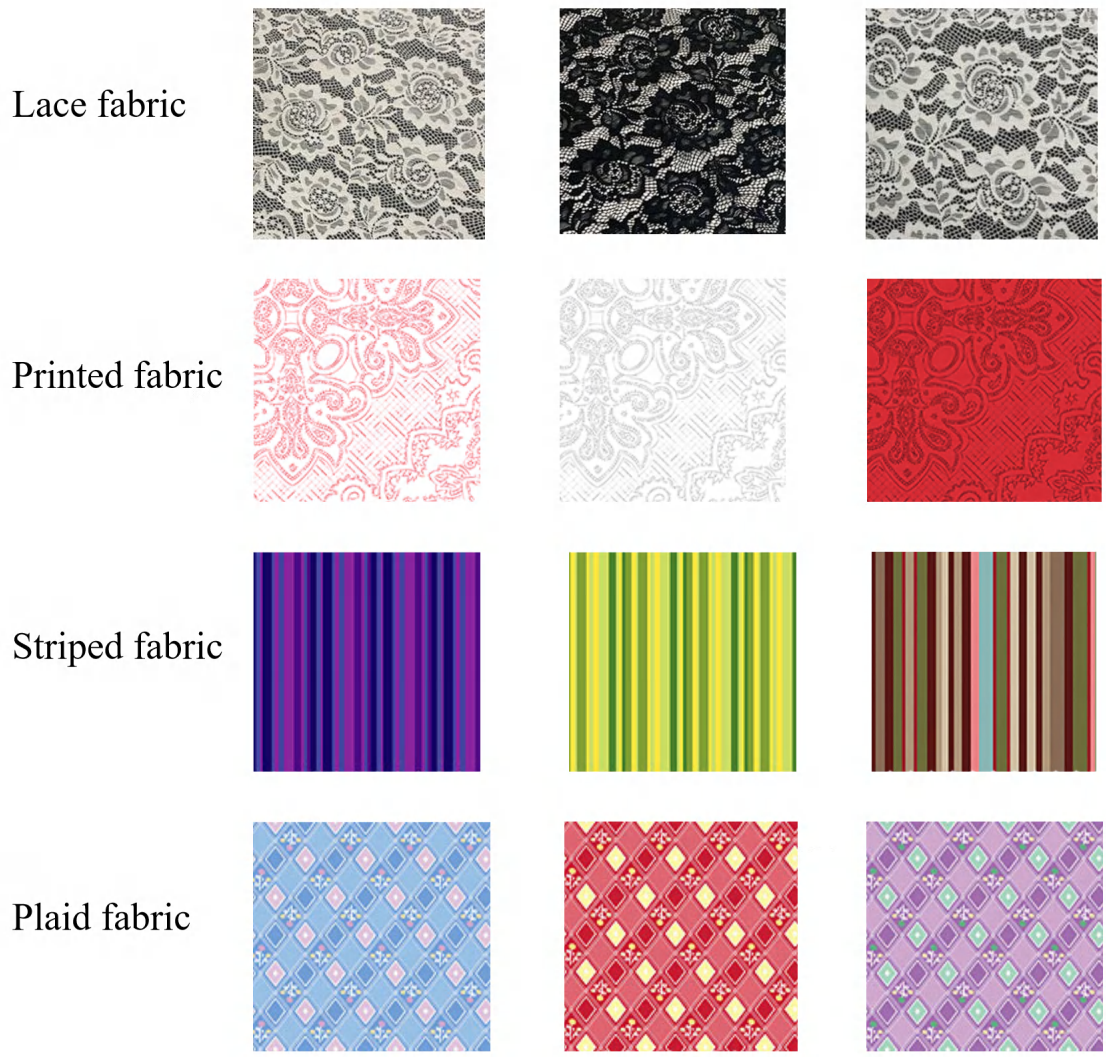


Figure 5.1: Fabric samples of different types.

images predominantly highlight global patterns and structures. Additionally, images captured from varying camera angles may introduce perspective distortions, further complicating the alignment and comparison of features. As a result, a retrieval system must be capable of maintaining feature consistency across different scales and viewpoints to achieve accurate matching.

Geometric Pattern Arrangement Differences: Fabrics with repeating patterns, such as striped and plaid designs, often exhibit geometric structures that vary in orientation, stripe width, spacing, or alignment. Differences in the spatial arrangement of these patterns can lead to misalignments in feature representations, posing challenges for retrieval systems. Effective retrieval requires feature representations that are robust to these geometric variations while retaining sensitivity to

structural differences.

Although significant progress has been made in content-based fabric image retrieval, most existing methods predominantly focus on either global features or local features, with limited attempts to integrate both. Methods that rely solely on **local features** often fail to account for the overall spatial layout and pattern structure of the fabric. For instance, two fabrics with similar small-scale textures but distinct global patterns (e.g., stripes versus lattices) may be incorrectly identified as similar. Conversely, methods that emphasize **global features** often overlook the intricate details and fine-grained textures that are essential for distinguishing fabrics with subtle variations. To address these limitations, a balanced approach that integrates both local and global features is essential. Local features capture the fine-grained details and subtle textures of fabric images, while global features represent the overall structural layout and large-scale patterns. By combining these complementary representations, the retrieval system can achieve a more comprehensive understanding of fabric images, enabling accurate alignment and comparison across variations in texture, scale, and geometric patterns. Such an integrated approach is vital for improving retrieval robustness and precision, particularly for fabrics with complex surface designs.

To address the challenges posed by complex fabric image retrieval and the limitations of existing approaches, a two-stage retrieval method is proposed, which effectively integrates both global and local image features. This approach significantly improves retrieval accuracy while maintaining efficiency. The proposed framework consists of several key contributions:

First, a novel end-to-end two-stage retrieval framework is introduced. In the initial stage, global feature retrieval identifies the Top K most similar candidate images. Subsequently, the second stage refines these candidates through local feature matching, ensuring that the final retrieval results are more accurate. This two-stage process capitalizes on the strengths of both global and local features, improving retrieval accuracy without the inefficiency of applying local feature matching to

large datasets.

Second, a deep network designed to extract both local and global features is proposed. This network, built upon a residual structure, incorporates a mixer-based fusion module. This module combinFes features from residual convolutional blocks at various scales, generating multi-scale local feature descriptors and global feature descriptors. By utilizing spatial and channel mixing, the network learns contextual dependencies across the output feature maps of convolutional layers, enhancing the representation of features from different regions of the image. This approach compensates for the local limitations of traditional convolution operations, improving feature extraction and representation.

Finally, an enhanced Triplet Loss with an intra-class compactness constraint is introduced. Traditional Triplet Loss primarily focuses on the margin between positive and negative samples, but it does not sufficiently address intra-class compactness. Given the inherent intra-class variability in fabric images, the proposed enhancement ensures that samples within the same class are positioned closer together, improving the discriminative power of the model and leading to more accurate retrieval results.

5.2 Related Works

This section primarily focuses on related works in the areas of fabric image retrieval, metric learning.

5.2.1 Fabric Image Retrieval

Fabric image retrieval has advanced considerably, transitioning from traditional hand-crafted methods to modern deep learning techniques. Early approaches primarily relied on features such as local color descriptors and texture patterns, with Local Binary Patterns (LBP) [55] frequently used for their effectiveness in capturing intricate textures. Some methods also combined shape and texture features to

improve retrieval outcomes. Techniques like GIST [32] and SURF [28] helped to describe images at both the global and local levels, although these methods often struggled with the diversity of fabric datasets. The rise of deep learning, particularly Convolutional Neural Networks (CNNs), has brought significant improvements by enabling automatic extraction of both global and local features. Transfer learning using pre-trained models like ResNet [11] and AlexNet [54] has become a popular strategy, allowing for fine-tuning specifically for fabric datasets. Contrastive learning methods, including Siamese networks, have shown strong performance in tasks like detailed one-to-one fabric image retrieval, particularly in handling fine-grained patterns like lace and woven fabrics. Some hybrid approaches have emerged, combining CNN-derived features with traditional descriptors such as HOG [56] and LBP [55] to capture both high-level and detailed textures. More recently, two-stage retrieval frameworks have been introduced like InceptionV3+ORB [35], with InceptionV3 [57] performing an initial global feature-based retrieval, followed by the second stage using ORB [35] features retrieval for greater precision. Despite these advancements, challenges remain in efficiently integrating global and local features and accurately handling the complexity of fabric patterns.

5.2.2 Metric Learning

Metric learning is used to learn a distance function such that similar samples are closer together in the embedding space, while dissimilar samples are farther apart. In image retrieval and classification tasks, metric learning significantly improves model performance by optimizing the distance relationships between samples. One of the classic methods in metric learning is contrastive loss, proposed by Hadsell et al. in 2006 [121]. This method learns a supervised embedding space by minimizing the distance between positive pairs (similar samples) and maximizing the distance between negative pairs (dissimilar samples). It is widely used in Siamese networks to establish similarity relationships between image pairs. Another popular metric learning method is triplet loss, introduced by Schroff et al. [122]. This method

uses a triplet of a positive sample, a negative sample, and an anchor sample to ensure that the distance between the anchor and the positive sample is smaller than the distance between the anchor and the negative sample. It has been successfully applied to tasks such as face recognition, significantly enhancing the discriminative power of the embedding space.

5.3 Framework of Hierarchical Two-Stage Retrieval

In this study, a novel two-stage fabric image retrieval framework is proposed to combine the advantages of both global and local feature extraction, thereby improving retrieval accuracy and computational efficiency. As illustrated in Figure 5.2, the framework operates in two sequential stages: global feature retrieval for coarse-level filtering and local feature matching for fine-grained refinement. A detailed description of each stage is provided below.

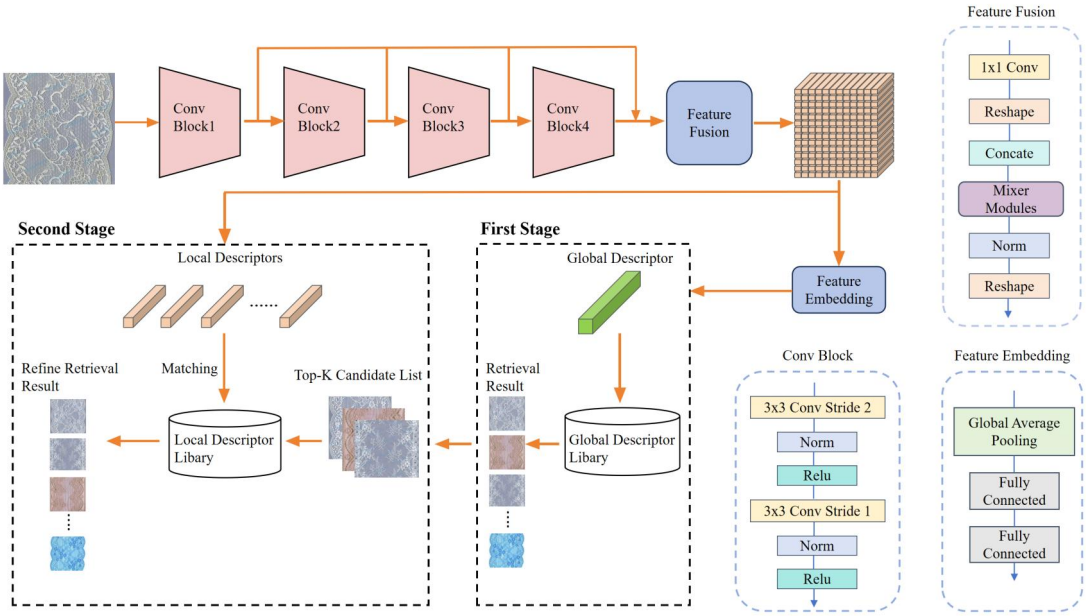


Figure 5.2: The framework of proposed novel two-stage fabric image retrieval framework

The first stage of the proposed framework focuses on global feature extraction for coarse image retrieval. The input fabric image is first passed through a se-

ries of four convolutional blocks (Conv Block1 to Conv Block4), each composed of two convolution layers. The first convolution layer in each block uses a stride of 2 for down-sampling, while the second convolution layer has a stride of 1 for feature extraction. These convolutional blocks capture high-level features from the input image, which are critical for constructing the subsequent global feature descriptor. After feature extraction through the convolutional blocks, the resultant feature maps are processed by mixer-based fusion module. This module employs a 1x1 convolution to adjust feature dimensions, followed by a reshaping operation to prepare the feature maps for mixer modules. The mixer-fusion mechanism is used to capture global dependencies between different spatial regions of the feature map, compensating for the local receptive field limitations of convolutional layers. The output of the fusion process is a more refined feature map, integrating both local and global context. To generate the global feature embedding, global average pooling (GAP) is used to the mixer-fused feature map, followed by two fully connected (FC) layers. The resulting global feature descriptor is then stored in the Global Descriptor Library, which is used to compute similarity with the global descriptor of the query image. Based on these global feature comparisons, the top-K most similar candidate images are selected for the next stage of refinement.

To further refine the retrieval results and improve precision, the second stage employs local feature matching. In this stage, the convolutional feature maps generated in stage one are leveraged to produce local feature descriptors. These local descriptors capture fine-grained information about the fabric texture, patterns, and structural details, which are particularly important for accurately distinguishing between visually similar fabric images. The local descriptors of the query image are matched against the Local Descriptor Library, which stores the local descriptors of the top-K candidate images retrieved in stage one. By computing the similarity between the query and candidate images at a finer, local level, the system re-ranks the candidates, producing a more accurate final retrieval result. This stage ensures that even subtle differences in texture and patterns are effectively captured and utilized

for precise matching.

5.3.1 Multi-scale Feature Fusion

To effectively leverage multi-scale visual representations extracted from the convolutional blocks, the proposed framework incorporates a dedicated Feature Fusion module. As shown in Fig. 5.2, the goal of this module is to integrate low-level and high-level features into a unified representation, thereby capturing both fine-grained texture details and global semantic patterns.

The initial convolutional blocks (*Conv Block1–Block4*) generate hierarchical features representing distinct levels of abstraction. While early layers emphasize localized texture cues such as edges and fine structures, deeper layers capture more abstract and semantic information, e.g., overall pattern and style. To ensure a meaningful integration of these representations, it first apply a 1×1 convolution to each feature map. This operation not only harmonizes the channel dimensions across multiple layers, thus facilitating subsequent concatenation, but also serves as a linear transformation to highlight salient feature channels and suppress redundant information.

Following the channel-wise transformation, each layers processed feature maps are reshaped into a compatible form and then concatenated along their feature dimension. This concatenation step yields a comprehensive feature collection that encapsulates multi-scale information. By doing so, it bridge the semantic gaps between early and deep features, providing a richer and more discriminative input to subsequent modules.

To further enhance representational capacity, the concatenated features are passed through mixer modules. As shown in Figure 5.2, these modules use fully-connected layers to realize nonlinear transformations and inter-channel interactions. Such a fusion process encourages deeper synergy among features from different layers, allowing both low-level details and higher-level semantics to be integrated more cohesively.

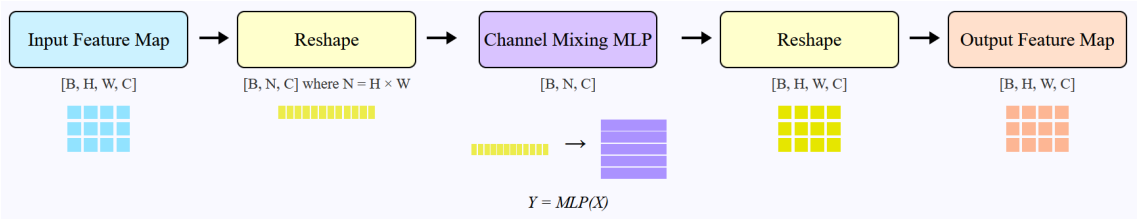


Figure 5.3: The processing of the Mixer Module for channel-wise feature integration.

After the mixing stage, normalization is applied to stabilize feature distributions and facilitate efficient training. Finally, the fused features are reshaped into the desired embedding form, ready for subsequent steps in the retrieval pipeline. By consolidating the multi-level representations into a single, semantically enriched descriptor, the Feature Fusion module ensures that both subtle textures and overarching patterns are effectively captured, thereby improving the overall performance and robustness of the retrieval system.

It is worth noting that the channel mixing mechanism, while inspired by the Mixer architecture proposed by Tolstikhin et al. [106], differs in its application and purpose. The MLP-Mixer [106] employs both spatial mixing and channel mixing in an alternating manner to replace traditional convolutional layers entirely, aiming to process the entire image in a vision transformer-like framework. In contrast, the proposed approach uses channel mixing solely as a feature fusion mechanism within a convolutional framework, focusing on enhancing inter-channel dependencies across multi-scale features without spatial token mixing. This design choice ensures that spatial relationships are preserved by the convolutional backbone, while the channel mixing MLP strengthens the integration of multi-scale semantic information, making it more suitable for the fabric image retrieval task where both local textures and global patterns are critical.

Overall, the Feature Fusion module is designed to efficiently integrate multi-stage convolutional features, unifying both low-level texture details and high-level semantic information into a cohesive representation. By doing so, it achieves improved performance in downstream retrieval and matching tasks. Such an approach to feature fusion is commonplace in many contemporary visual feature extraction and

embedding architectures, underscoring the importance and effectiveness of multi-scale and multi-level integration.

5.3.2 Joint Learning of Global and Local Feature Embeddings

The feature embedding stage is designed to produce both global and local embeddings that collectively capture a comprehensive representation of the input image. As illustrated in Fig. 5.2, the global feature embedding provides a compact, high-level descriptor suitable for efficient retrieval, while the local feature embedding preserves spatially detailed information essential for fine-grained matching. By integrating both embeddings, the approach effectively balances retrieval speed with the ability to discriminate subtle differences among visually similar samples.

Global Feature Embedding Consider the fused feature map $\mathbf{F}_{\text{out}} \in \mathbb{R}^{H \times W \times C}$, where H and W are the spatial dimensions and C is the number of channels. To obtain a global representation, global average pooling is first applied across the spatial dimensions to produce a single vector $\mathbf{f}_{\text{global}} \in \mathbb{R}^C$:

$$\mathbf{f}_{\text{global}} = \text{GlobalAvgPool}(\mathbf{F}_{\text{out}}). \tag{5.1}$$

This operation summarizes the salient features present in the entire image, providing a channel-wise aggregate that discards spatial layout but retains discriminative global cues. Subsequently, this global feature vector is passed through two fully connected layers to form a final global descriptor $\mathbf{g} \in \mathbb{R}^d$:

$$\mathbf{g} = \text{FC}_2(\text{FC}_1(\mathbf{f}_{\text{global}})). \tag{5.2}$$

The resulting global descriptor $\mathbf{g}$ is utilized in the first retrieval stage, offering a coarse yet efficient means of identifying a shortlist of candidate images that likely resemble the query image. The compactness of this global embedding is beneficial

for efficient retrieval scenarios, enabling rapid computations and scalability.

Local Feature Embedding While the global feature embedding facilitates swift candidate identification, certain retrieval tasks require finer discrimination, especially among images sharing similar global characteristics. To address this need, local feature embeddings are extracted from the original convolutional feature maps. Flattening these maps yields a set of local feature vectors $\mathbf{f}_i \in \mathbb{R}^C$ for each spatial location i:

$$\mathbf{f}_i = \text{Flatten}(\mathbf{F}_{\text{out}}), \quad i = 1, 2, \ldots, H \times W. \tag{5.3}$$

These local embeddings preserve the spatial arrangement and textural details lost in the global average pooling step. In the second retrieval stage, these fine-grained representations are employed to distinguish subtle variations in pattern, texture, or structure, thereby improving the accuracy of the final retrieval results.

By employing a dual-feature embedding strategy, the methodology efficiently narrows down the candidate set using global descriptors and subsequently refines the retrieval accuracy with detailed local embeddings. This hierarchical approach leverages the strengths of both representation levels, ensuring that the system remains robust against visually similar distractors and capable of pinpointing subtle differences.

5.3.3 Intra-Class Compactness Regularized Triplet Loss

Traditional triplet loss aims to enforce proximity between anchor and positive samples while maintaining separation between anchor and negative samples. This objective is achieved by minimizing the following loss function:

$$L_{\text{triplet}} = \max\big(d(\mathbf{A}, \mathbf{P}) - d(\mathbf{A}, \mathbf{N}) + \alpha,\, 0\big), \tag{5.4}$$

where $d(\mathbf{A}, \mathbf{P})$ and $d(\mathbf{A}, \mathbf{N})$ represent the distances between the anchor vector $\mathbf{A}$

and the positive vector $\mathbf{P}$, and between the anchor vector $\mathbf{A}$ and the negative vector $\mathbf{N}$, respectively. The parameter α is a margin that ensures a minimum distance between the anchor-negative pair.

Although the traditional triplet loss formulation is effective in many scenarios, it does not explicitly encourage intra-class compactness. In real-world applications, ensuring that samples belonging to the same class form a tight cluster in the feature space significantly improves generalization. To address this limitation, this study introduces an intra-class compactness regularized triplet loss that augments the conventional triplet objective with a regularization term promoting compactness among the positive samples of the same class.

The intra-class compactness regularized triplet loss is defined as follows:

$$L_{\text{total}} = L_{\text{triplet}} + \frac{2}{(N_{\text{pos}} - 1)N_{\text{pos}}} \sum_{i=1}^{N_{\text{pos}}} \sum_{j=i+1}^{N_{\text{pos}}} d(\mathbf{P}_i, \mathbf{P}_j), \tag{5.5}$$

where $d(\mathbf{P}_i, \mathbf{P}_j)$ measures the distance between two positive sample vectors $\mathbf{P}_i$ and $\mathbf{P}_j$ from the same class, and N_{pos} is the number of positive samples within the batch. This term minimizes the pairwise distances between positive samples (images of the same fabric group), forcing the model to cluster semantically similar samples tightly.

By incorporating the intra-class regularization term, the intra-class compactness regularized triplet loss enforces not only inter-class separability but also intra-class compactness. This dual objective leads to more cohesive class clusters and better discriminative power in the learned feature embeddings. As a result, both classification accuracy and retrieval performance can be significantly improved when the model must handle visually similar categories.

5.3.4 Coarse-to-Fine Retrieval with Global Pruning and Local Re-ranking

Retrieval tasks often require a careful balance between efficiency and accuracy. To address this challenge, a two-stage retrieval strategy is introduced, which integrates global feature embeddings for rapid coarse selection and local feature embeddings for subsequent fine-grained refinement. This hierarchical process enables efficient pruning of the candidate set in the first stage and preserves the capacity to distinguish subtle visual differences in the second stage.

Stage One: Coarse Retrieval Using Global Features In the initial retrieval phase, the goal is to efficiently identify a manageable subset of candidate images that closely match the querys global characteristics. Let $\mathbf{F} \in \mathbb{R}^{C\times H\times W}$ represent the fused feature map of an image. Global average pooling is applied to produce a global feature vector $\mathbf{f}_{\text{global}}$:

$$\mathbf{f}_{\text{global}} = \frac{1}{H \times W} \sum_{h=1}^{H} \sum_{w=1}^{W} \mathbf{F}(h, w). \tag{5.6}$$

This vector is then passed through two fully connected (FC) layers with ReLU activation, resulting in a compact global descriptor $\mathbf{g}$:

$$\mathbf{g} = \text{FC}_2\big(\text{ReLU}(\text{FC}_1(\mathbf{f}_{\text{global}}))\big). \tag{5.7}$$

The global descriptor $\mathbf{g}$ captures the images overall content distribution, serving as an effective high-level signature. A k-Nearest Neighbors (k-NN) search based on cosine similarity is employed to retrieve the top k most similar images from the global descriptor database:

$$\text{Similarity}(\mathbf{g}_i, \mathbf{g}_j) = \frac{\mathbf{g}_i \cdot \mathbf{g}_j}{\|\mathbf{g}_i\|\|\mathbf{g}_j\|}. \tag{5.8}$$

This process yields an initial candidate set $\mathcal{C}_{\text{initial}}$ that is likely to contain visually similar images to the query, offering a significant reduction in the search space.

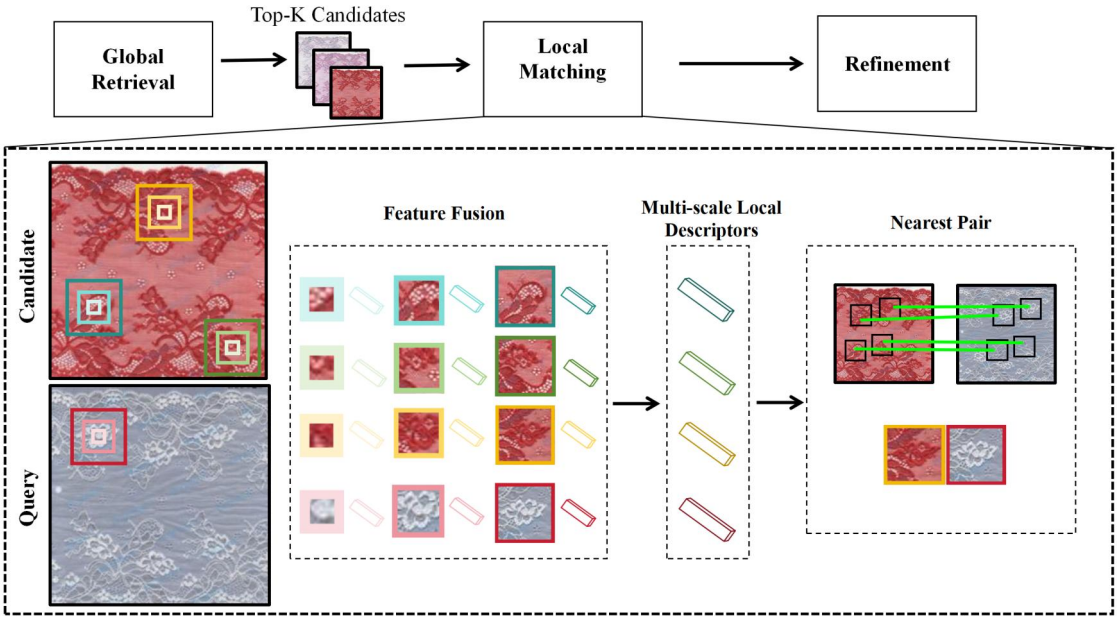


Figure 5.4: Example of fine-grained matching

Stage Two: Fine-Grained Matching Using Local Features Once an initial shortlist is obtained, a fine-grained matching process is applied to further refine the retrieval results. This second stage operates on local feature embeddings $\mathbf{f}_i$, extracted from spatial positions of the fused feature maps. Such local features preserve essential textural and structural details that may be lost at the global representation level, thereby enabling a more precise evaluation of similarity.

As shown in Fig. 5.4. For the query image, let $\mathbf{F}_{\text{query}} \in \mathbb{R}^{C\times H\times W}$ be the local feature map. Flattening and normalization yield a set of normalized local feature vectors $\{\hat{\mathbf{f}}_{\text{query},m}\}_{m=1}^{M}$, where $M = H \times W$:

$$\hat{\mathbf{f}}_{\text{query},m} = \frac{\mathbf{f}_{\text{query},m}}{\|\mathbf{f}_{\text{query},m}\|_2}, \quad m = 1, \ldots, M. \tag{5.9}$$

Similarly, for each candidate image $I_j \in \mathcal{C}_{\text{initial}}$, the local feature map $\mathbf{F}_j \in \mathbb{R}^{C\times H\times W}$ is processed to obtain $\{\hat{\mathbf{f}}_{j,n}\}_{n=1}^{N}$, where $N = H \times W$:

$$\hat{\mathbf{f}}_{j,n} = \frac{\mathbf{f}_{j,n}}{\|\mathbf{f}_{j,n}\|_2}, \quad n = 1, \ldots, N. \tag{5.10}$$

The similarity between a query local feature $\hat{\mathbf{f}}_{\text{query},m}$ and a candidate local feature

$\hat{\mathbf{f}}_{j,n}$ is computed as their dot product:

$$\text{Similarity}(\hat{\mathbf{f}}_{\text{query},m}, \hat{\mathbf{f}}_{j,n}) = \hat{\mathbf{f}}_{\text{query},m} \cdot \hat{\mathbf{f}}_{j,n}. \tag{5.11}$$

A match is deemed valid if this similarity exceeds a predefined threshold τ. For each candidate image I_j, the number of valid matches MatchCount_j is calculated by considering the maximum similarity of each query local feature with any local feature in I_j:

$$\text{MatchCount}_j = \sum_{m=1}^{M} \big(\max_{n=1}^{N} \text{Similarity}(\hat{\mathbf{f}}_{\text{query},m}, \hat{\mathbf{f}}_{j,n}) \geq \tau \big). \tag{5.12}$$

Candidate images are subsequently re-ranked in descending order of MatchCount_j. This re-ranking ensures that images sharing a higher number of detailed local correspondences with the query image receive priority, thus refining the initial retrieval results and producing the final ranked list $\mathcal{C}_{\text{final}}$.

By integrating global and local features in a two-stage retrieval pipeline, the method exploits the complementary strengths of each representation. The global descriptor efficiently narrows down the search space, while the local features provide discriminative granularity for accurate refinement. This combined approach yields robust retrieval performance even in the presence of subtle visual differences among candidate images.

5.4 Experiments

This section describes the experimental setup and procedures used to evaluate the proposed method. The evaluation framework encompasses data illustration, implementation details, training strategy, and performance metrics. In addition, along with ablation studies designed to elucidate the contribution of each component within the proposed approach. This comprehensive evaluation offers insights into the methods effectiveness, robustness, and potential for real-world applications.

Figure 5.5: Examples of fabric images belonging to different groups.

5.4.1 Data Illustration Under Different Capture Conditions

As illustrated in Fig. 5.5, these images were obtained under controlled lighting, fixed shooting angles, and uniform imaging devices. This standardized setup reduces external interference, ensuring that the captured images accurately represent the intrinsic properties of the fabrics. These highly controlled samples serve as reliable foundational data for evaluating texture-related tasks, including recognition and classification.

All images underwent a rigorous process of manual selection and annotation to ensure their quality and relevance. The self constructed dataset provides a versatile testing ground for assessing both accuracy and robustness in fabric image retrieval

scenarios, enabling in-depth analyses and comparisons with existing methods.

5.4.2 Implementation

The dataset was randomly divided into training, validation, and test sets with an 8:1:1 ratio. Each subset preserves the diversity of fabric textures, ensuring fair and reliable evaluations. Prior to training, all images were standardized and resized to $224 \times 224 \times 3$, guaranteeing uniform input dimensions and image formats.

The proposed image retrieval method was implemented using the PyTorch framework. The Adam optimizer was employed, and a learning rate of 0.0001 was selected for the experiments. To further refine the models performance, a hard sample mining strategy was applied. In each training iteration, loss values were monitored to identify challenging triplet groups with higher losses. These hard examples were assigned greater weight, providing stronger gradient signals that facilitate more effective parameter optimization. This selective emphasis on challenging samples promotes faster convergence and enhances the models robustness.

The training process spanned 50 epochs. After each epoch, performance was evaluated on the validation set, and the model weights corresponding to the highest validation accuracy were retained. Upon completion of training, the weights from the best-performing epoch were used for final evaluation on the test set. This procedure ensures that the final model achieves a balance between accuracy, robustness, and generalization ability.

5.4.3 Evaluation Metrics

Commonly used metrics in image retrieval tasks—Precision, Recall, and Mean Average Precision (mAP)—were adopted to measure retrieval quality. Given the relatively small number of samples per fabric pattern (often only five, rarely exceeding ten), multiple retrieval depths (Top-1, Top-5, and Top-10) were considered to thoroughly assess the models performance. This multi-level evaluation offers a more comprehensive perspective on the models retrieval capabilities, especially under chal-

lenging conditions with limited samples per category.

The definitions of these metrics are as follows:

$$P@K = \frac{\text{RelevantNum@K}}{K}, \quad R@K = \frac{\text{RelevantNum@K}}{\text{RelevantNum}} \tag{5.13}$$

$$mP@K = \frac{1}{N}\sum_{n=1}^{N} P_n@K, \quad mR@K = \frac{1}{N}\sum_{n=1}^{N} R_n@K \tag{5.14}$$

$$mAP@K = \frac{1}{N}\sum_{n=1}^{N}\left(\frac{1}{K}\sum_{k=1}^{K} P_n@k\right) \tag{5.15}$$

Here, $mP@K$ and $mR@K$ represent the mean precision and recall at Top-K, respectively. $mAP@K$ computes the mean average precision at Top-K, providing a comprehensive indicator of the overall ranking quality.

5.4.4 Experimental Results

The proposed method was evaluated against several state-of-the-art methods in fabric retrieval. Table 5.1 summarizes the results in terms of Precision, Recall, and mAP at Top-1, Top-5, and Top-10 retrieval levels. To further enhance accuracy, the two-stage retrieval approach processes the Top-30 results from the initial stage, refining them through a second-stage evaluation.

Among the comparison methods, InceptionV3+ORB [102] is also a two-stage

Table 5.1: Performance comparison on the test dataset.

Method	$mP@1$	$mP@5$	$mP@10$	$mR@1$	$mR@5$	$mR@10$	$mAP@10$	$F1@10$
AlexNet [95]	77.61%	44.83%	25.10%	16.85%	51.77%	57.68%	48.22%	34.22%
InceptionV3+ORB [102]	94.03%	59.22%	32.26%	22.83%	64.85%	69.05%	61.92%	43.27%
SiameseNet [97]	89.40%	56.41%	30.85%	17.51%	61.73%	66.60%	58.62%	40.76%
Modified CNN [96]	81.82%	46.36%	25.98%	20.78%	53.31%	59.48%	51.46%	36.64%
Multi-task ResNet50 [98]	90.91%	54.83%	29.41%	21.39%	59.40%	67.26%	57.49%	40.52%
Modified InceptionResNetV2 [118]	91.26%	56.08%	30.63%	21.29%	62.60%	69.88%	59.43%	41.33%
Modified MobileNet [119]	85.66%	48.81%	27.41%	19.01%	57.02%	63.31%	52.87%	37.32%
DSH [120]	84.97%	47.89%	26.96%	18.81%	55.87%	61.91%	52.64%	36.76%
Proposed Method	**95.73%**	**62.26%**	**34.11%**	**23.71%**	**67.87%**	**72.49%**	**66.10%**	**45.63%**

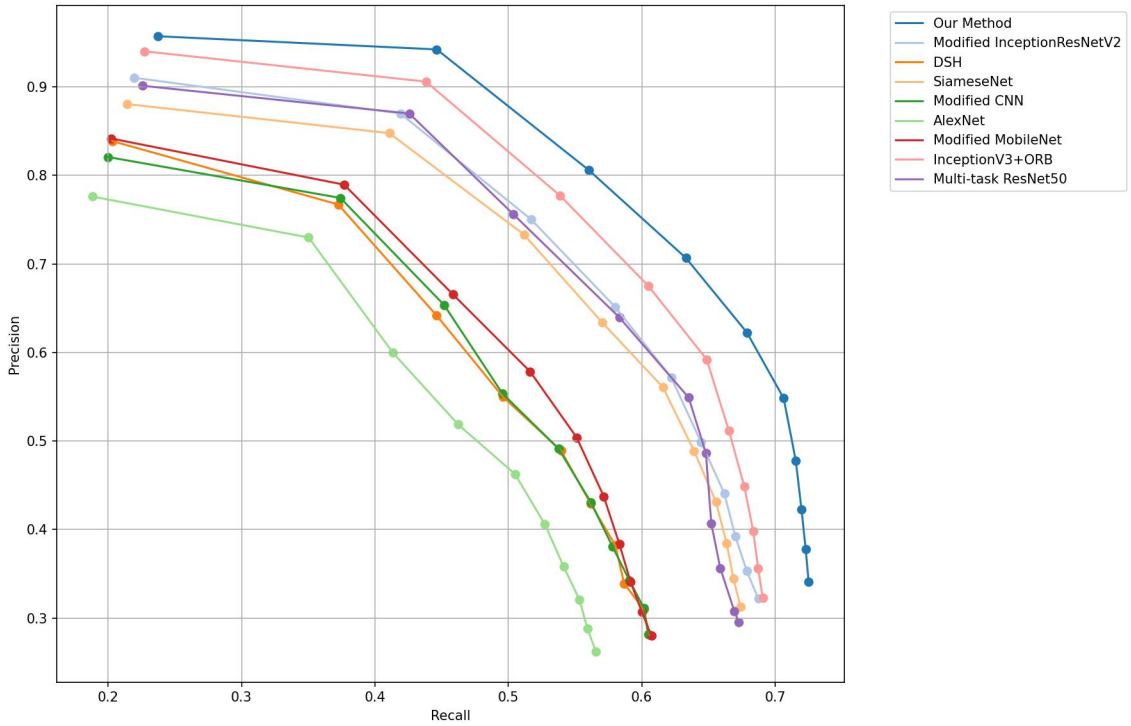


Figure 5.6: Precision-Recall (P-R) curves for the proposed method and comparison approaches.

approach, whereas the others rely on a single-stage retrieval process. As observed, two-stage methods display a considerable accuracy advantage for fabrics exhibiting complex textures. The proposed method outperforms InceptionV3+ORB due to its more robust second-stage matching strategy, which leverages learned feature embeddings rather than the handcrafted ORB descriptors. This distinction is critical when dealing with diverse and unpredictable natural environmental conditions, where the robustness of the local descriptors plays a decisive role.

Single-stage methods such as AlexNet [95], SiameseNet [97], Modified CNN [96], Modified MobileNet [119], Multi-task ResNet50 [98], and DSH [120] achieve reasonable performance but encounter difficulties in capturing both coarse global structures and subtle local patterns simultaneously. By contrast, the proposed two-stage method integrates global descriptors for initial candidate selection and local embeddings for subsequent refinement. This hierarchical approach preserves critical fine-grained textures while maintaining computational efficiency, resulting in superior retrieval performance for complex and visually similar fabric categories.

Figure 5.6 presents the Precision-Recall (P-R) curves for the proposed method and several comparison methods. The proposed method consistently achieves higher precision across varying recall levels, indicating its capacity to retrieve more relevant results while minimizing irrelevant ones. Although InceptionV3+ORB remains competitive, its precision declines as recall increases, suggesting limitations in retaining accuracy under more challenging retrieval conditions.

Overall, the P-R curve analysis corroborates the numerical results, demonstrating that the two-stage retrieval strategy, which fuses global and local feature embeddings, is more effective than traditional single-stage methods. By capitalizing on both global structure and local detail, the proposed approach consistently outperforms alternative retrieval systems, providing a robust and scalable solution for fabric image retrieval tasks.

5.4.5 Ablation Study

A series of ablation studies was conducted to assess the individual contributions of key components within the proposed method. These studies focus on three primary aspects: the efficacy of the two-stage retrieval framework, the influence of mixer-based feature fusion, and the benefits of the intra-class compactness regularized triplet loss function.

Two-Stage Retrieval: The first ablation experiment examines the two-stage retrieval framework, which integrates both global and local feature representations. A single-stage retrieval baseline, relying solely on global descriptors, is compared against the proposed two-stage approach, where the second stage refines the results using local feature embeddings. As shown in Table 5.2, the two-stage method consistently outperforms the single-stage baseline across all evaluation metrics, demonstrating considerable gains in precision and recall at various ranks.

In particular, $mP@1$ increases from 93.13% to 95.73%, indicating a marked improvement in identifying the most relevant candidate from the start. Performance

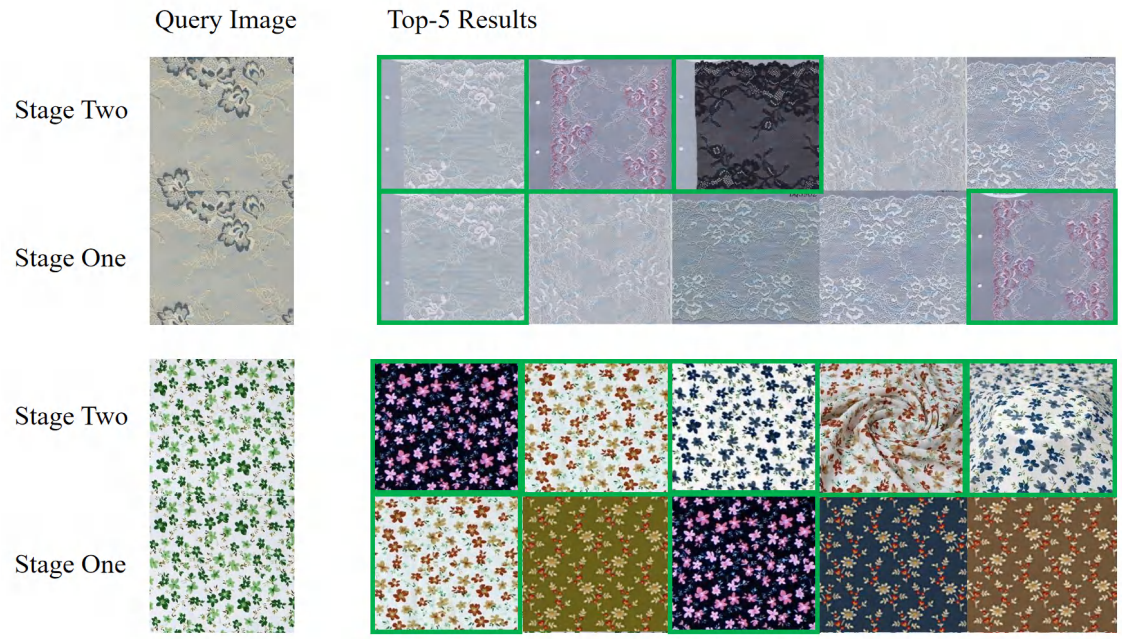


Figure 5.7: Top-5 retrieval results for identical queries under single-stage (Stage One) and two-stage (Stage Two) retrieval.

enhancements in $mP@5$ and $mR@5$ further suggest that incorporating local feature analysis leads to more reliable retrieval results early in the ranked list. These improvements occur because global descriptors alone, while efficient for initial filtering, may fail to capture subtle distinguishing characteristics. The second-stage refinement, leveraging spatially detailed local features, corrects for these limitations by ensuring that the final results more accurately reflect the intricate properties of the query.

Figure 5.7 provides a qualitative perspective, illustrating how the two-stage approach more effectively excludes images with only superficial similarities, thereby delivering a more semantically coherent set of top-ranked results. This analysis underscores the importance of incorporating fine-grained local feature matching into retrieval workflows, particularly when managing complex and visually similar datasets.

Table 5.2: Retrieval Performance Comparison: Single-Stage vs. Two-Stage Retrieval.

Method	$mP@1$	$mP@5$	$mP@10$	$mR@1$	$mR@5$	$mR@10$	$mAP@10$	$F1@10$
Stage One	93.13%	59.91%	32.56%	22.58%	64.92%	69.64%	62.24%	44.44%
Stage Two	**95.73%**	**62.26%**	**34.11%**	**23.71%**	**67.87%**	**72.49%**	**66.10%**	**46.34%**

Mixer-Based Feature Fusion: The second ablation experiment assesses the influence of mixer-based feature fusion on retrieval performance. Table 5.3 compares various fusion strategies: no fusion, simple concatenation, NetVLAD fusion, Feature Pyramid Networks(FPN) fusion, and the proposed mixer fusion mechanism.

The proposed Mixer-Based Fusion achieves **95.73% mP@1** and **62.26% mP@5**, outperforming NetVLAD (**89.64% mP@1**) and FPN (**92.72% mP@1**) (Table 5.1). This is attributed to its ability to adaptively capture cross-scale contextual dependencies through spatial-channel mixing, which enhances discriminative power for both global patterns and local details.

The results highlight the superior performance achieved by mixer fusion, surpassing all alternative approaches. For instance, $mP@1$ and $mP@5$ improve substantially, indicating that mixer-based methods better identify and emphasize the most discriminative feature channels. While simple concatenation and traditional methods (NetVLAD, FPN) provide incremental benefits, the proposed mixer fusion technique, which adaptively focuses on the most salient features at multiple scales, leads to a more discriminative and contextually aware representation.

These findings underscore the importance of adaptive weighting strategies that emphasize critical features while downplaying redundant or noisy information. By integrating mixer-based fusion, the proposed method effectively enhances its representational capacity, translating into tangible gains in both precision and recall across diverse retrieval scenarios.

Intra-Class Compactness Regularized Triplet Loss: The final ablation study focuses on the intra-class compactness regularized triplet loss, which introduces a regularization term to enhance intra-class compactness alongside standard inter-

Table 5.3: Impact of Feature Fusion on Retrieval Performance.

Method	$mP@1$	$mP@5$	$mP@10$	$mR@1$	$mR@5$	$mR@10$	$mAP@10$	$F1@10$
No Fusion	88.52%	55.13%	31.90%	20.11%	59.24%	66.84%	59.31%	43.19%
Concate Fusion	89.92%	57.53%	32.64%	20.40%	61.44%	68.15%	60.03%	44.39%
NetVLAD Fusion	89.64%	59.10%	33.45%	20.35%	63.24%	69.70%	61.64%	45.20%
FPN Fusion	92.72%	59.38%	33.47%	22.02%	63.58%	69.91%	62.31%	45.25%
Proposed Mixer-Based Fusion	**95.73%**	**62.26%**	**34.11%**	**23.71%**	**67.87%**	**72.49%**	**66.10%**	**46.40%**

class separation. Table 5.4 presents a comparison between models trained with the traditional triplet loss and the modified variant. Notably, The intra-class compactness regularized triplet loss improves **F1@10** from **44.04%** (traditional triplet loss) to **46.40%**, while achieving higher **mP@1** (**95.73%** vs. **94.12%**) and **mAP@10** (**66.10%** vs. **64.42%**). This indicates that the intra-class compactness constraint enhances both precision-recall balance and overall ranking quality.

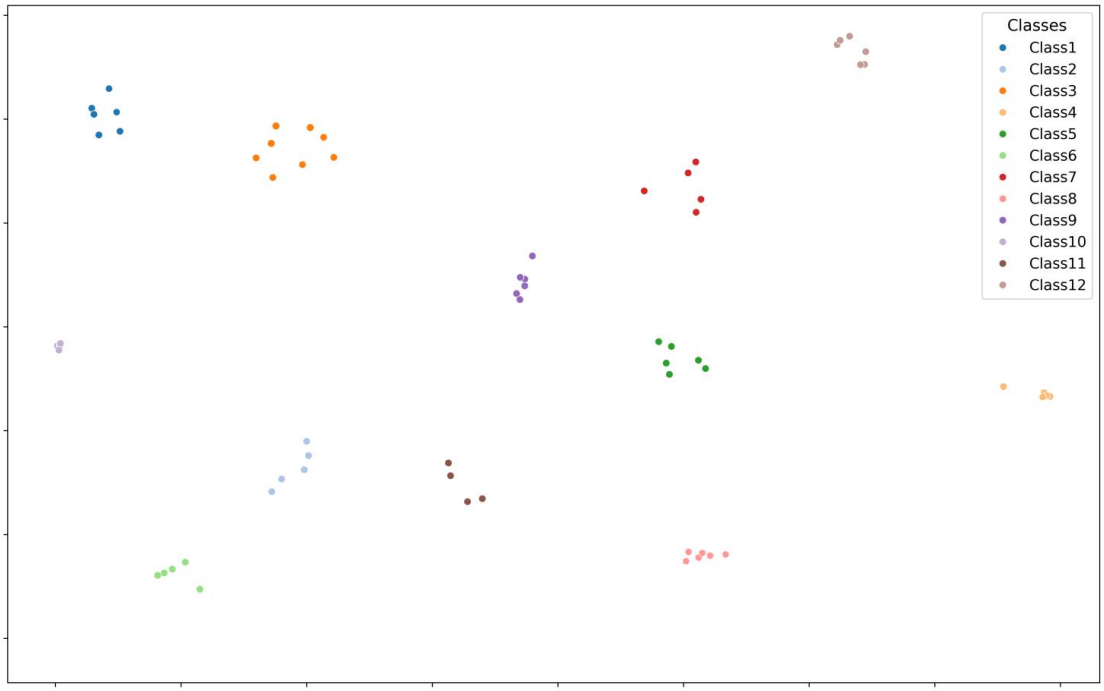
The t-SNE visualizations in Fig. 5.8 offer a qualitative illustration. The embeddings learned with the intra-class compactness regularized triplet loss (Fig. 5.8(a)) form tighter, more coherent clusters, reflecting improved intra-class compactness. Additionally, distinct classes are more clearly separated, indicating enhanced discrimination between fabric categories. In contrast, embeddings from the traditional triplet loss (Fig. 5.8(b)) produce more dispersed intra-class clusters and less distinct inter-class boundaries. Such distributions can hinder the models ability to differentiate fine-grained variations within similar classes.

These results confirm that addressing both intra-class compactness and inter-class separability is crucial for complex retrieval scenarios. By refining the loss formulation, the proposed method achieves more discriminative and well-structured feature embeddings, ultimately contributing to its superior retrieval performance.

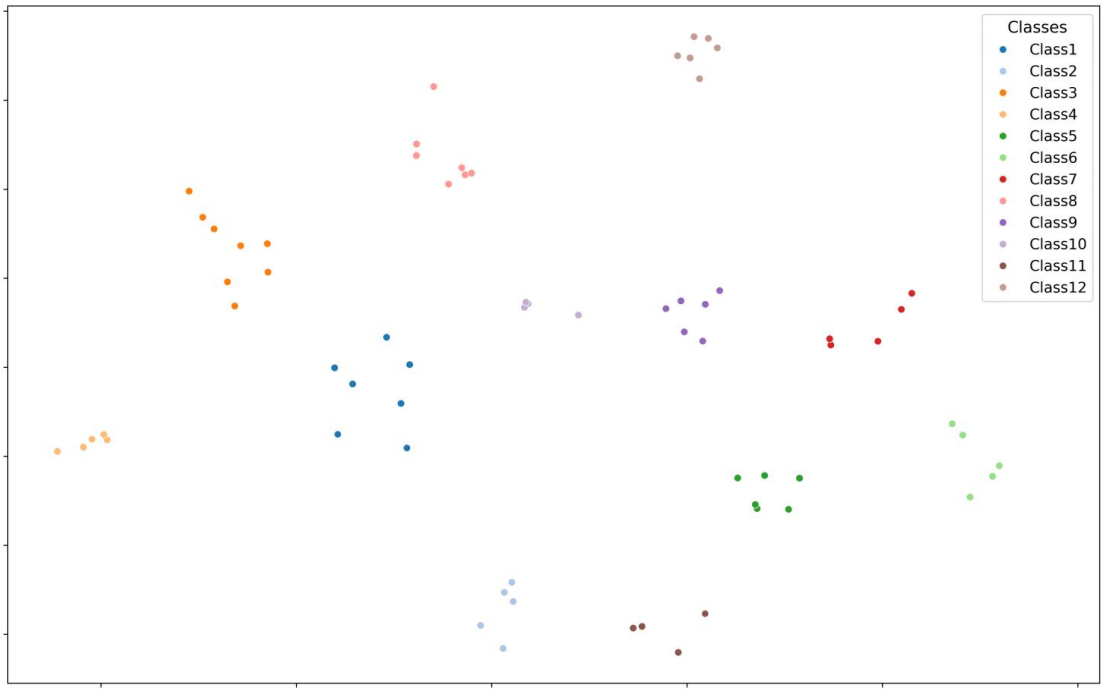
Summary of Ablation Findings: All three ablation studies—exploring the second-stage refinement, mixer-based feature fusion, and the intra-class compactness regularized triplet loss demonstrate clear improvements in retrieval accuracy, robustness, and discrimination. Incorporating local feature matching in a two-stage framework to highlight discriminative features, and enforcing intra-class compactness through an enhanced loss function collectively yield a significant performance boost.

Table 5.4: The performance of the model trained with traditional triplet loss and intra-class compactness regularized triplet loss.

Method	mP@1	mP@5	mP@10	mR@1	mR@5	mR@10	mAP@10	$F1$@10
Triplet Loss	94.12%	57.43%	32.29%	22.88%	63.12%	69.23%	64.42%	44.04%
Proposed Method	**95.73%**	**62.26%**	**34.11%**	**23.71%**	**67.87%**	**72.49%**	**66.10%**	**46.40%**



(a)



(b)

Figure 5.8: t-SNE visualization: (a) the embeddings learned by intra-class compactness regularized triplet loss. (b) the embeddings learned by the traditional triplet loss.

The outcomes validate that each component plays a pivotal role. The two-stage approach sharpens the result set, feature fusion refines feature representations, and the intra-class compactness regularized triplet loss ensures that learned embeddings

more effectively represent the underlying structure and variability of fabric patterns. Taken together, these findings reinforce the importance of each design choice and its contribution to the overall success of the proposed retrieval method.

5.5 Chapter Summary

This chapter presents a novel two-stage retrieval framework tailored for fabric image retrieval, effectively addressing the inherent challenges posed by intricate and diverse textile textures. By exploiting a dual-layered strategy that first employs global descriptors for efficient filtering and subsequently refines the candidate set using local feature embeddings, the proposed method achieves a substantial improvement in retrieval performance over state-of-the-art counterparts. The primary contributions include the integration of both global and local feature representations. These components collectively enhance the models capacity to capture subtle textural variations and ensure intra-class compactness. Empirical evaluations demonstrate that the two-stage framework yields significant gains in Precision, Recall, and mAP at various retrieval depths, particularly excelling at Top-1 and Top-5 levels. The refined second stage consistently elevates the quality of top-ranked results, prioritizing truly relevant fabric images and improving both the accuracy and reliability of the retrieval system.

Chapter 6

Efficient Local Feature Matching via Cross Attention

Fabric image retrieval often requires fine-grained matching beyond what global features can achieve, especially when dealing with subtle patterns or complex textures. This chapter presents the Efficient Local Feature Matching (ELFM) method, designed to improve the efficiency and accuracy in local matching process. As the final refinement stage in the proposed complex fabric retrieval framework, ELFM operates on the candidate subset from first stage retrieval, utilizing a Cross Attention mechanism to dynamically align query and database local features like pattern edges, texture junctions. By treating one set of local descriptors as Queries and the other as Keys/Values, the proposed Cross Attention Module explicitly computes a set of matching weights, allowing the model to derive a robust global similarity score. Experiments on a constructed challenging fabric dataset demonstrate that the proposed approach significantly improves retrieval accuracy and efficiency over existing methods.

6.1 Introduction

In the field of fabric image retrieval, although global image representations have greatly improved retrieval efficiency, they demand finer distinctions at the local

level. Subtle differences in weave patterns, textures, and motifs often remain indistinguishable using global descriptors alone.

Two-stage retrieval strategies address these issues by combining global and local features. In the first stage, a fast search using global embeddings identifies a shortlist of candidate images. The second stage then refines this shortlist by analyzing the local feature correspondence between the query and candidate images. However, most conventional two-stage methods rely on exhaustive pairwise comparisons between every local descriptor in the query image and those in the candidate images. This brute-force approach introduces significant computational overhead. Such inefficiencies hinder the scalability of these methods for real-world applications.

To overcome these limitations, this work proposes an end-to-end cross-attention framework tailored for the second stage of fabric image retrieval. Unlike traditional approaches that explicitly compare descriptors in a pairwise manner, the proposed framework leverages a learnable Cross Attention Module to model the local matching process. In this module, one set of local descriptors is treated as Queries, while the other serves as Keys and Values. Through the cross-attention mechanism, the system directly computes a distribution of matching weights, enabling it to capture fine-grained correspondences between local regions in an efficient and structured manner.

The proposed approach offers multiple advantages. First, by explicitly modeling the matching process as an attention mechanism, the framework provides a more nuanced understanding of local descriptor alignments, enabling robust handling of subtle texture variations and complex patterns. Second, the end-to-end design eliminates the need for heuristic matching rules, allowing the system to learn optimal alignment strategies directly from data. Finally, the parallelizable computations of the cross-attention mechanism significantly reduce the computational burden when implemented on GPU hardware. This not only accelerates the matching process but also ensures scalability to large databases, making the framework suitable for deployment in industrial-scale fabric retrieval systems.

Overall, the integration of a cross-attention mechanism into the second stage of two-stage retrieval systems bridges the gap between precision and efficiency, paving the way for more accurate and scalable solutions to challenging image retrieval task.

6.2 Related Works

6.2.1 Query Expansion Methods

Query expansion is a widely adopted technique in image retrieval, aiming to enhance recall and precision by augmenting the original query representation. Traditional method [66] based on Bag-of-Words (BoW) faces challenges such as semantic gaps and limitations in visual quantization. To address these issues, various advanced approaches have been proposed. Attention-based query expansion methods leverage attention mechanisms to effectively aggregate image features, forming more representative expanded queries to improve retrieval accuracy [123]. Contextual query expansion techniques utilize common visual patterns (CVPs) to incorporate contextual information, bridging the semantic gap in BoW-based frameworks [124]. Online query expansion hashing methods dynamically enhance the discriminative ability of query hash codes during retrieval, improving efficiency and accuracy [125]. Other approaches, such as saliency and picturability-based query expansion, employ external knowledge bases to generate expanded query terms that align with the original querys semantics [126]. Similarly, combining BoW and convolutional neural network (CNN) features with active learning strategies refines the expanded queries, enhancing object retrieval performance [127]. These methods demonstrate the potential of integrating deep learning sources to address the limitations of traditional query expansion, providing robust solutions for improving image retrieval systems.

6.2.2 Geometry Verification Methods

Geometric verification methods evaluate the geometric relationships between images to filter out incorrect matches and ensure the reliability of retrieval results. For

instance, RANSAC (Random Sample Consensus) is a widely used approach that estimates geometric transformation models by randomly sampling feature points and iteratively identifying inliers and outliers, thereby removing erroneous matches and ensuring robust model fitting [83]. Building upon this, LO-RANSAC (Locally Optimized RANSAC) introduces a local optimization step to improve both accuracy and computational efficiency, making it suitable for image matching tasks [128]. Another commonly used method is epipolar geometry-based verification, which computes the fundamental or essential matrix to validate the geometric consistency of feature matches and filter mismatches that violate geometric constraints [129]. These geometric verification techniques are integral to the second stage of image retrieval, significantly boosting system performance and ensuring robust and accurate retrieval.

6.2.3 Spatial Similarity Methods

Spatial similarity methods capture spatial relationships and geometric structures in images. Among these, Spatial Pyramid Matching (SPM) has been widely adopted, dividing an image into hierarchical grids and computing feature matches at multiple scales, which is particularly effective for object category recognition and scene matching [130]. Another foundational approach, Histogram of Oriented Gradients (HOG), extracts spatial information by analyzing the distribution of local gradient orientations, making it highly suitable for both object detection and image retrieval [131]. Graph Matching represents images as graph structures, where nodes correspond to keypoints and edges represent spatial relationships, enabling accurate scene-level retrieval and complex object matching [132]. Spatial re-ranking further enhances retrieval precision by leveraging spatial layout information to reorder candidate images based on their spatial compatibility with the query [133]. Shape Contexts capture the spatial distribution of boundary points, providing a robust descriptor for shape-based image matching and retrieval [134]. Finally, Spatial Hashing encodes image features into spatial hash structures to facilitate rapid spatial

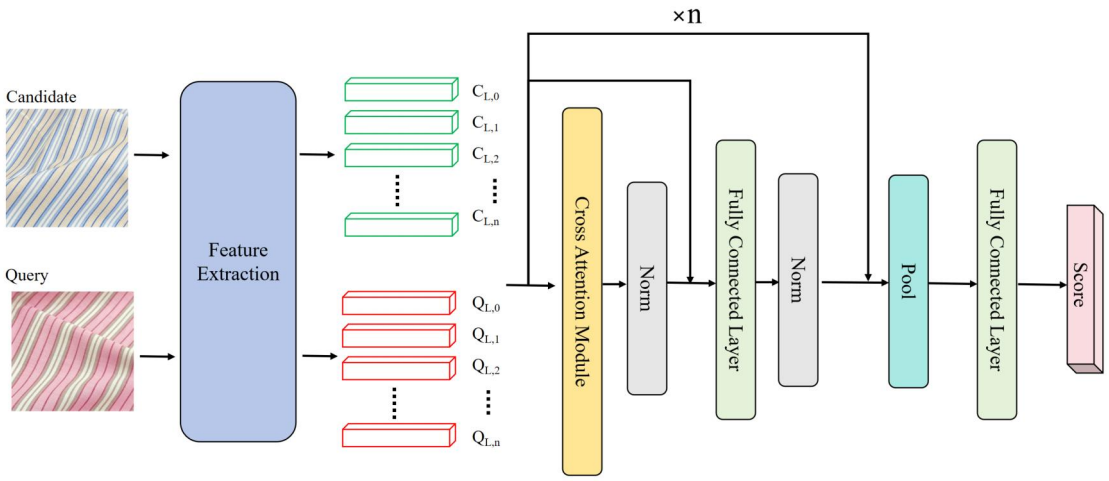


Figure 6.1: The framework of ELFM.

similarity computation, offering an efficient solution for image retrieval tasks [135]. Together, these methods address the challenges of spatial similarity estimation, significantly enhancing the performance and scalability of image retrieval systems.

6.3 Framework of ELFM

The proposed ELFM for fabric image retrieval addresses the challenges of fine-grained similarity computation. The framework consists of a feature extraction backbone and a Cross Attention-based local matching module, as shown in Figure 6.1. This section describes the architecture in detail, including the Cross Attention mechanism and the local matching process.

The retrieval framework employs a two-stage design to achieve efficient and accurate retrieval. The first stage generates a shortlist of candidate images using global feature representations extracted from a feature extraction backbone. In the second stage, local descriptors from the query image and the shortlisted candidates are input into a Cross Attention module to compute a fine-grained similarity score.

Feature Extraction: For each image (query or candidate), the feature extraction backbone outputs a 3D tensor of dimensions (H, W, C), where H and W are the spatial dimensions, and C is the feature channel dimension. This tensor represents a set of $H \times W$ local descriptors, each of size C. These descriptors are reshaped into

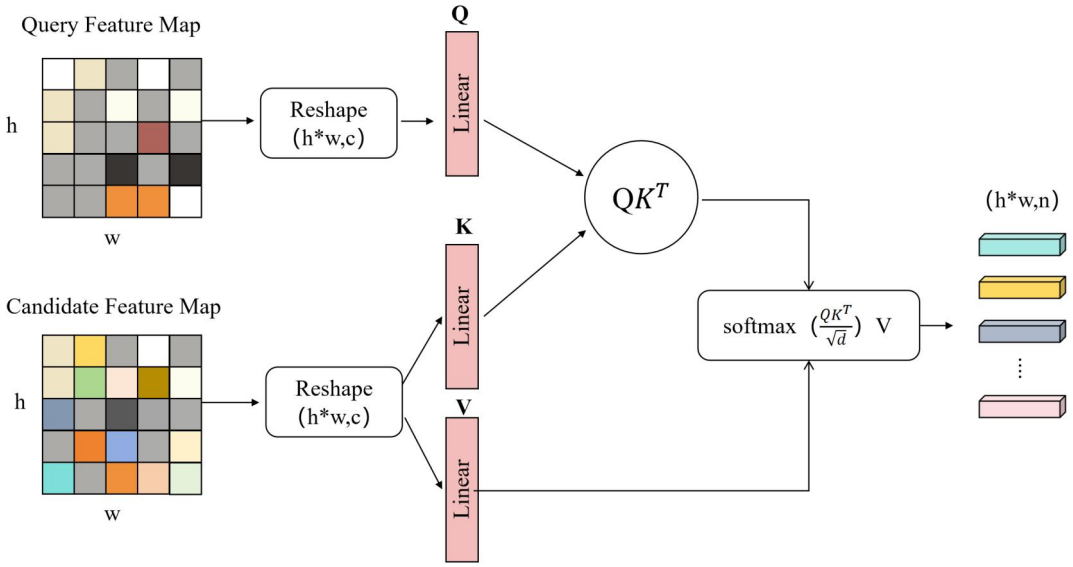


Figure 6.2: Cross attention mechanism.

sequences:

$$\mathbf{Q}_{\text{seq}} = \{\mathbf{q}_i\}_{i=1}^{M} \in \mathbb{R}^{M \times c}, \quad \mathbf{C}_{\text{seq}} = \{\mathbf{c}_j\}_{j=1}^{N} \in \mathbb{R}^{N \times c}, \tag{6.1}$$

where $M = H \times W$ for the query and $N = H \times W$ for the candidate.

Cross Attention Module: The core of the framework is the Cross Attention Module, which adaptively computes matching relationships between local descriptors from the query ($\mathbf{Q}$) and candidate ($\mathbf{C}$). Multiple Cross Attention layers (n) can be stacked to enhance the feature alignment and similarity computation.

Score Prediction: The output of the final Cross Attention layer is passed through a series of normalization and fully connected layers, which aggregate the aligned features into a global similarity score between the query and the candidate.

6.3.1 Cross Attention Mechanism

The Cross Attention Module forms the core of the proposed framework, efficiently modeling relationships between local descriptors extracted from query and candidate images. This mechanism enables the computation of fine-grained similarity by learning adaptive attention weights between descriptors in a fully differentiable manner.

As illustrated in Figure 6.2, the local feature maps from the query and candidate

images are first reshaped into sequences of descriptors. Let the feature map dimensions for both the query and candidate images be (h, w, c), where h and w denote the spatial dimensions, and c represents the feature channel dimension.

For the attention computation, three learnable projections are applied:

$$Q = \mathbf{Q}_{\text{seq}} W_Q \in \mathbb{R}^{M \times d}, \quad K = \mathbf{C}_{\text{seq}} W_K \in \mathbb{R}^{N \times d}, \quad V = \mathbf{C}_{\text{seq}} W_V \in \mathbb{R}^{N \times d}, \tag{6.2}$$

where $W_Q, W_K, W_V \in \mathbb{R}^{c \times d}$ are projection matrices, and d is the projected feature dimension. The scaling factor $\frac{1}{\sqrt{d}}$ stabilizes gradient magnitudes during training.

The attention weights are computed as:

$$\text{Attention}(Q, K) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right), \tag{6.3}$$

where $\frac{1}{\sqrt{d}}$ serves as a scaling factor to stabilize the gradients during training. The attention operation outputs a weighted combination of the candidate descriptors:

$$\text{Output} = \text{Attention}(Q, K)V. \tag{6.4}$$

This process allows each local descriptor in the query to selectively attend to relevant descriptors in the candidate image, creating a refined representation that captures fine-grained correspondences. By stacking multiple Cross Attention layers, higher-order relationships between the local descriptors can also be captured, further improving the matching accuracy.

Figure 6.2 illustrates the overall computation process, highlighting the transformation of input feature maps, the projection into the Query, Key, and Value spaces, and the final attention-weighted output. This module ensures that the framework can efficiently handle large numbers of descriptors while maintaining accuracy, leveraging the parallelizable nature of matrix operations for high computational efficiency.

6.3.2 Local Matching Process

The output of the Cross Attention Module is a refined sequence of local descriptors for the query image, incorporating information from the candidate descriptors. To compute a global similarity score between the query and the candidate, the following steps are applied:

1. **Aggregation**: An adaptive average pooling layer is applied to the output sequence to summarize the refined features into a single vector:

$$\mathbf{z} = \text{Pool}(\text{Output}). \tag{6.5}$$

2. **Score Prediction**: The pooled feature vector $\mathbf{z}$ is passed through a series of fully connected layers to predict a similarity score:

$$\text{Score} = \sigma(\text{MLP}(\mathbf{z})), \tag{6.6}$$

where σ is the sigmoid function, and the MLP consists of linear layers and non-linear activation functions.

This process ensures that the final score reflects both global and local feature similarities, allowing for accurate re-ranking of candidates.

6.3.3 Training Loss

To optimize the framework, a binary classification loss is used. For each query-candidate pair, the ground truth label $y \in \{0, 1\}$ indicates whether the pair represents the same fabric. The predicted similarity score $\hat{y}$ is trained using the binary cross-entropy loss:

$$L = -\sum_i \Big[y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i) \Big]. \tag{6.7}$$

The model is trained end-to-end, allowing the feature extraction backbone, Cross Attention Module, and fully connected layers to be jointly optimized. This ensures

that the learned representations and attention weights are tailored for fine-grained fabric retrieval tasks.

6.4 Experiments

This section evaluates the proposed Cross Attention-based local matching framework on the self constructed fabric dataset. The experiments focus on three aspects: implementation details, comparisons with related works, and ablation studies to assess the impact of different numbers of Cross Attention modules.

6.4.1 Implementation

The model is implemented using PyTorch, with the feature extraction backbone initialized using pre-trained weights from MLDF. The Cross Attention module operates on local descriptors of size 1024, extracted from feature maps of dimension 14×14 for each image. For the training process, the self constructed dataset consisting of 2,448 fabric images, divided into 1,958 training images and 490 validation images, is used.

During training, input images are resized to 224×224 and normalized using the mean and standard deviation of the ImageNet dataset. A batch size of 16 is used, and the model is trained for 50 epochs with the Adam optimizer and an initial learning rate of 10^{-3}, decayed by a factor of 0.1 every 20 epochs. The loss function employed is binary cross-entropy, computed between the predicted similarity scores and ground truth labels.

To ensure fair evaluation, global features for the first-stage retrieval are extracted using the same MLDF backbone. The top-30 candidates from the first stage are passed to the second stage for local matching, where the Cross Attention module computes refined similarity scores.

6.4.2 Comparison to Exsiting Methods

This section presents a comparison between the proposed framework and several existing methods for Two-Stage retrieval. Table 6.1 provides a comprehensive evaluation of retrieval performance based on key metrics, including Precision@K, Recall@K, mAP@K, and F1@10. The methods compared include traditional techniques such as Bag of Words, Contextual Expansion, and modern approaches like NN Matching, GNN Matching, RANSAC, and LO-RANSAC. The proposed method outperforms all other approaches in terms of both retrieval accuracy and computational efficiency.

Precision and Recall are critical metrics for assessing retrieval performance. At Precision@1, the proposed method achieves 96.44%, significantly surpassing other methods. In comparison, Bag of Words and Contextual Expansion achieve 82.07% and 91.22%, respectively, while NN Matching and GNN Matching attain 95.73% and 95.10%. This indicates that the Cross Attention mechanism is highly effective in accurately identifying the most relevant results at the top rank, resulting in superior precision for the first retrieved result when compared to traditional methods. For Precision@5 and Precision@10, the proposed method shows an impressive 65.28% and 37.08%, which are the highest among all methods. In contrast, Bag of Words and Contextual Expansion show 53.22% and 58.13%, respectively. This demonstrates that the proposed method is more effective at retrieving relevant results within the top-5 and top-10 ranks, with the Cross Attention mechanism ensuring better relevance and consistency across the retrieval list. Similarly, at Recall@5 and Recall@10, the proposed method achieves 71.12% and 76.13%, outperforming all other approaches. For example, Bag of Words has 57.08% at Recall@5, and

Table 6.1: Performance comparison on the fabric dataset with matching time and retrieval metrics.

Method	Cost Time (ms)	Precision@1	Precision@5	Precision@10	Recall@1	Recall@5	Recall@10	mAP@10	F1@10
Bag of Words	171	82.07%	53.22%	30.45%	18.59%	57.08%	63.64%	57.85%	41.44%
Contextual Expansion	234	91.22%	58.13%	31.31%	21.10%	63.34%	68.91%	60.02%	42.86%
NN Matching	174	95.73%	62.26%	34.11%	23.71%	67.87%	72.49%	66.10%	46.43%
GNN Matching	214	95.10%	63.30%	35.26%	22.10%	68.48%	73.05%	66.85%	47.55%
RANSAC	396	95.52%	61.90%	33.89%	22.67%	65.98%	70.45%	65.36%	45.68%
LO-RANSAC	372	95.85%	63.34%	34.12%	23.96%	68.62%	72.58%	67.03%	46.42%
Proposed Method	122	**96.44%**	**65.28%**	**37.08%**	**24.82%**	**71.12%**	**76.13%**	**69.54%**	**49.83%**

63.64% at Recall@10, which are significantly lower than the proposed methods values. The higher Recall values for the proposed method highlight its ability to retrieve more relevant images, even as the list expands, further emphasizing the effectiveness of the Cross Attention mechanism in capturing fine-grained correspondences. At mAP@10, the proposed method achieves 69.54%, which is notably higher than all other methods. For instance, Contextual Expansion achieves 60.02%, and NN Matching reaches 66.10%. The higher mAP@10 score of the proposed method reflects its ability to maintain high relevance across a broader set of retrieved images, indicating more accurate ranking of images in the retrieval list. Regarding F1@10, the proposed method also achieves the highest score at 49.83%. This is a significant improvement compared to other methods, with Bag of Words achieving 41.44%, and Contextual Expansion reaching 42.86%. The higher F1 score indicates that the proposed method not only retrieves more relevant images but also effectively balances precision and recall, making it a more robust solution for image retrieval tasks.

In terms of computational efficiency, the proposed method shows a considerable reduction in matching time, achieving a matching time of 122 ms. This is substantially faster than methods like RANSAC (396 ms) and LO-RANSAC (372 ms), which are known for their higher computational cost due to their exhaustive nature. While methods like NN Matching and GNN Matching show faster matching times (174 ms and 214 ms, respectively), the proposed method provides a notable balance between high retrieval performance and computational efficiency. The reduced matching time of the proposed method makes it highly suitable for real-time or efficient image retrieval tasks, where both accuracy and speed are essential.

The proposed method stands out due to its use of the Cross Attention mechanism, which enables better matching of fine-grained details and contextual information between images. The significant improvements in both precision and recall across various values of K demonstrate its ability to more accurately retrieve relevant images. Additionally, the method achieves these high retrieval accuracies with a notable reduction in computational cost, making it more efficient than methods

like RANSAC and LO-RANSAC. This combination of high accuracy and efficiency makes the proposed method a promising solution for image retrieval in real-world applications, where both speed and precision are critical.

In conclusion, the proposed method not only outperforms traditional and state-of-the-art approaches in terms of retrieval accuracy but also provides significant gains in efficiency. These results demonstrate the effectiveness of the Cross Attention mechanism in enhancing the retrieval process and highlight its potential for future image retrieval systems.

6.4.3 Ablation Study

An ablation study was conducted to investigate the impact of varying the number of Cross Attention layers on retrieval performance. The model was evaluated with $n = 1, 2, 3, 4, 5$ layers of Cross Attention, while keeping other parameters fixed. The results of the ablation study are summarized in Table 6.2, revealing the trade-offs between accuracy and computational cost as the number of layers increases.

Table 6.2: Ablation study: Effect of the number of Cross Attention layers.

Number of Attention Layers	Cost Time	Precision@1	Precision@5	Precision@10	Recall@1	Recall@5	Recall@10	mAP@10	F1@10
$n = 1$	56	93.17%	59.23%	32.45%	22.23%	65.01%	70.81%	61.58%	43.93%
$n = 2$	75	93.55%	60.76%	33.18%	22.64%	66.68%	71.53%	63.22%	44.87%
$n = 3$	98	95.47%	64.15%	34.98%	23.44%	69.54%	73.32%	65.86%	47.37%
$n = 4$	122	**96.44%**	65.28%	**37.08%**	**24.82%**	71.11%	**76.13%**	**69.54%**	**49.83%**
$n = 5$	158	96.38%	**66.78%**	36.95%	24.69%	**71.88%**	75.88%	66.42%	49.31%

The results presented in Table 6.2 show a clear trend: increasing the number of Cross Attention layers improves retrieval performance, particularly in terms of precision and recall. However, this improvement comes at the cost of increased computational time. The model with a single Cross Attention layer achieves Precision@1 of 93.17% and Recall@10 of 70.81%. While this configuration shows good retrieval accuracy, it exhibits relatively moderate performance compared to models with more layers. The matching time is the lowest at 56 ms, making this configuration computationally efficient. With two layers, the model shows a slight improvement in performance. Precision@5 increases to 60.76% and Recall@5 rises to 66.68%. However, the computational cost also increases, with the matching time rising to 75

ms. This configuration balances accuracy and efficiency but does not significantly outperform the $n = 1$ setup. When three layers are used, there is a more noticeable improvement in both precision and recall, especially in Precision@10 (34.98%) and Recall@10 (73.32%). The matching time also increases to 98 ms, reflecting the higher computational demands. This setup marks a significant improvement in retrieval accuracy, particularly in the second stage of retrieval. With four layers, the model achieves the highest Precision@1 of 96.44% and Recall@10 of 76.13%. These results indicate a strong balance between retrieval performance and computational efficiency. While the matching time increases to 122 ms, the improvements in precision and recall make this configuration the most effective in terms of accuracy. The model reaches an optimal balance between performance and computational cost at this layer configuration. Increasing the number of layers to five results in a slight decrease in performance compared to $n = 4$. While Precision@5 increases to 66.78% and Recall@5 improves to 71.88%, the matching time grows to 158 ms. The performance gain is marginal, and the computational cost increases significantly. This suggests that the model begins to experience diminishing returns with additional layers, leading to overfitting and an inefficient trade-off between accuracy and speed.

Overall, the results demonstrate that the models retrieval performance improves with the number of Cross Attention layers, but beyond four layers, the gains in accuracy become minimal, while the computational cost continues to rise. The configuration with $n = 4$ layers offers the best balance between precision, recall, and efficiency, making it the most optimal choice for the given task.

6.4.4 Computational Efficiency and Scalability Analysis

This section addresses critical practical considerations regarding the computational efficiency and scalability of the proposed framework, specifically examining the factors influencing training/inference time and the model's adaptability to dataset expansions.

The computational efficiency of the proposed framework is governed by two key factors:

Candidate List Size: The size of the candidate list generated in the first-stage retrieval significantly affects inference time. The second-stage local feature matching involves pairwise comparisons between query and candidate images. A larger candidate list leads to slower inference time due to the exhaustive matching operations required.

Computational Complexity: The first-stage global retrieval employs efficient matrix operations with complexity $O(n)$, where increasing the library size only raises matrix dimensionality without exponentially increasing operations. In contrast, the second-stage local matching exhibits complexity proportional to the candidate list size.

The analysis reveals a nuanced relationship between computational requirements and dataset scale:

$$T_{\text{inference}} \propto N_{\text{candidates}} \times T_{\text{matching}} + T_{\text{global}} \tag{6.8}$$

where T_{global} remains relatively constant regardless of library size, while T_{matching} scales linearly with the candidate list size $N_{\text{candidates}}$. This explains why inference time shows weak correlation with overall dataset size but strong dependence on the candidate list dimension.

Adaptability to Dataset Expansion: The proposed end-to-end deep learning framework demonstrates strong generalization capabilities. When new images are added to the dataset, the model does not require complete retraining from scratch. The hierarchical feature representation learned by the model exhibits robust transferability to new pattern variations.

6.5 Chapter Summary

This chapter presents a robust and efficient solution for fabric image retrieval by integrating a Cross Attention-based local matching module into a two-stage retrieval framework. The ELFM framework not only achieves superior retrieval accuracy through fine-grained feature alignment but also maintains computational efficiency, thereby offering a scalable and practical approach for industrial-scale image retrieval systems. Recognizing the necessity for precise local feature matching, the proposed Efficient Local Feature Matching (ELFM) framework employs a two-stage retrieval process. The first stage leverages global feature representations to swiftly narrow down the pool of candidate images. Subsequently, the second stage refines these candidates through a Cross Attention Module, which facilitates robust local feature matching by modeling intricate relationships between query and candidate image descriptors. A comprehensive methodology section details the architecture of ELFM, emphasizing the role of the Cross Attention mechanism in efficiently computing matching weights and deriving accurate similarity scores. The frameworks design ensures both high retrieval accuracy and computational efficiency, making it suitable for efficient and real-time applications. Experimental evaluations on a meticulously constructed fabric dataset demonstrate the superiority of the proposed method over existing techniques.

Chapter 7

Conclusions and Suggestions for Future Research

7.1 Conclusions

This study addresses the critical challenges in fabric image retrieval by proposing a comprehensive framework that integrates multi-scale feature fusion, hierarchical two-stage retrieval, and efficient local feature matching. The key contributions are summarized as follows:

- **Multi-Scale Local Descriptors Fusion (MLDF):** The MLDF method effectively captures both fine-grained textures and global structural patterns through multi-scale convolutional layers and mixer-based feature fusion. By leveraging a progressive triplet mining strategy, it enhances the discriminative power of feature embeddings, achieving significant improvements in retrieval precision and recall. Experimental results demonstrate that MLDF outperforms traditional handcrafted features and existing deep learning models, particularly for complex fabrics like lace and printed textiles.

- **Hierarchical Two-Stage Retrieval Framework:** The proposed two-stage framework combines global descriptors for coarse retrieval and local descriptors for fine-grained refinement. By employing a unified feature extraction model,

it eliminates inconsistencies between stages and ensures feature space consistency. This approach achieves a balance between computational efficiency and retrieval accuracy, with a 95.73% Top-1 precision and 66.10% mAP@10 on the constructed dataset.

- **Efficient Local Feature Matching via Cross Attention (ELFM):** The ELFM method introduces a cross-attention mechanism to dynamically align local features between query and candidate images. By reducing pairwise matching complexity and incorporating spatial-contextual relationships, it achieves a 96.44% Top-1 precision and 69.54% mAP@10 while maintaining computational efficiency (122 ms latency).

Collectively, these innovations address the limitations of existing methods in handling high intra-class variability, multi-scale patterns, and real-world imaging conditions, offering a scalable and robust solution for industrial applications in textile design, inventory management, and quality control.

7.2 Limitations

Despite its advancements, this study has several limitations:

- **Dataset Constraints:** The constructed dataset, while diverse, is limited to 2,448 images from 537 fabric groups. Larger-scale datasets with more variations in material, deformation, and extreme lighting conditions are needed to further validate generalization.

- **Computational Overhead:** Although efficient compared to traditional methods, the two-stage framework and cross-attention modules require GPU acceleration for real-time performance. Deployment on resource-constrained edge devices remains challenging.

- **Limited Cross-Domain Validation:** The framework is primarily validated on fabric images. Its applicability to other domains with similar texture com-

plexities like medical imaging or satellite imagery requires further investigation.

7.3 Suggestions for Future Research

Future research directions include:

- **Lightweight Architecture Design:** Exploring knowledge distillation, neural architecture search, or quantization techniques to optimize computational efficiency for real-time deployment on mobile or embedded systems.
- **Cross-Domain Adaptation:** While the proposed framework was specifically designed for complex fabric image retrieval, its core components demonstrate significant potential for adaptation to related domains. The methodology is not fundamentally restricted to fabric images but can be effectively transferred to other scenarios requiring fine-grained pattern recognition and retrieval. For example, it can be extended to tasks such as clothing style retrieval or fabric defect detection. In such application scenarios, additional detection models need to be introduced to preprocess the input images, automatically crop out the target regions such as local patterns of clothing or defect areas, and use the cropped image regions as input to the retrieval model proposed in this paper, thereby achieving effective matching or analysis.
- **Interactive Retrieval Systems:** Incorporating user feedback mechanisms and active learning to refine retrieval results iteratively, aligning with practical industrial workflows.

By addressing these challenges, the proposed framework can evolve into a versatile tool for both academic research and industrial applications, advancing the frontier of content-based image retrieval technologies.

References

[1] W. Li, L. Duan, D. Xu, and I. W.-H. Tsang, "Text-based image retrieval using progressive multi-instance learning," in *2011 international conference on computer vision*. IEEE, 2011, pp. 2049–2055.

[2] X. Li, J. Yang, and J. Ma, "Recent developments of content-based image retrieval (cbir)," *Neurocomputing*, vol. 452, pp. 675–689, 2021.

[3] S. Jain, K. Pulaparthi, and C. Fulara, "Content based image retrieval," *Int. J. Adv. Eng. Glob. Technol*, vol. 3, pp. 1251–1258, 2015. [Online]. Available: http://ijaegt.com/wp-content/uploads/2015/08/409594-pp-1251-1258-jain.pdf

[4] R. Da Silva Torres and A. X. Falcao, "Content-based image retrieval: theory and applications." *RITA*, vol. 13, no. 2, pp. 161–185, 2006. [Online]. Available: https://api.semanticscholar.org/CorpusID:17566677

[5] H. Müller, W. Müller, D. M. Squire, S. Marchand-Maillet, and T. Pun, "Performance evaluation in content-based image retrieval: overview and proposals," *Pattern recognition letters*, vol. 22, no. 5, pp. 593–601, 2001.

[6] A. W. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 22, no. 12, pp. 1349–1380, 2000.

[7] V. N. Gudivada and V. V. Raghavan, "Content based image retrieval systems," *Computer*, vol. 28, no. 9, pp. 18–22, 1995.

[8] C. Yuan, C. Zhou, J. Peng, and H. Li, "Mixture correntropy-based robust distance metric learning for classification," *Knowledge-Based Systems*, vol. 295, p. 111791, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0950705124004258

[9] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[10] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, ser. Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 1–9.

[11] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, ser. Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[12] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, ser. Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2818–2826.

[13] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proceedings of the AAAI conference on artificial intelligence*, ser. Proceedings of the AAAI conference on artificial intelligence, vol. 31, 2017, 1.

[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.

[15] Y. Zhu and S. Newsam, "Densenet for dense flow," in *2017 IEEE international conference on image processing (ICIP)*, ser. 2017 IEEE international conference on image processing (ICIP). IEEE, 2017, pp. 790–794.

[16] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[17] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, and S. Gelly, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[18] H. Jégou, M. Douze, C. Schmid, and P. Pérez, "Aggregating local descriptors into a compact image representation," in *2010 IEEE computer society conference on computer vision and pattern recognition*, ser. 2010 IEEE computer society conference on computer vision and pattern recognition. IEEE, 2010, pp. 3304–3311.

[19] M. J. Swain and D. H. Ballard, "Color indexing," *International journal of computer vision*, vol. 7, no. 1, pp. 11–32, 1991.

[20] G. Pass, R. Zabih, and J. Miller, "Comparing images using color coherence vectors," in *Proceedings of the fourth ACM international conference on Multimedia*, 1997, pp. 65–73.

[21] H. Yu, M. Li, H.-J. Zhang, and J. Feng, "Color texture moments for content-based image retrieval," in *Proceedings. International Conference on Image Processing*, vol. 3. IEEE, 2002, pp. 929–932.

[22] J. Huang, S. R. Kumar, M. Mitra, W.-J. Zhu, and R. Zabih, "Image indexing using color correlograms," in *Proceedings of IEEE computer society conference on Computer Vision and Pattern Recognition*. IEEE, 1997, pp. 762–768.

[23] A. Talib, M. Mahmuddin, H. Husni, and L. E. George, "A weighted dominant color descriptor for content-based image retrieval," *Journal of Visual Communication and Image Representation*, vol. 24, no. 3, pp. 345–360, 2013.

[24] T. S. Lee, "Image representation using 2d gabor wavelets," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 18, no. 10, pp. 959–971, 1996.

[25] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proceedings of the seventh IEEE international conference on computer vision*, ser. Proceedings of the seventh IEEE international conference on computer vision, vol. 2. Ieee, 1999, pp. 1150–1157.

[26] Y. Ke and R. Sukthankar, "Pca-sift: A more distinctive representation for local image descriptors," in *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, vol. 2. IEEE, 2004, pp. II–II.

[27] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *IEEE transactions on pattern analysis and machine intelligence*, vol. 27, no. 10, pp. 1615–1630, 2005.

[28] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (surf)," *Computer vision and image understanding*, vol. 110, no. 3, pp. 346–359, 2008.

[29] R. Arandjelović and A. Zisserman, "Three things everyone should know to improve object retrieval," in *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2012, pp. 2911–2918.

[30] D. Zhang and G. Lu, "Shape-based image retrieval using generic fourier descriptor," *Signal Processing: Image Communication*, vol. 17, no. 10, pp. 825–848, 2002.

[31] J. Jing, Q. Li, P. Li, H. Zhang, and L. Zhang, "Patterned fabric image retrieval using color and space features," *Journal of Fiber Bioengineering and Informatics*, vol. 8, no. 3, pp. 603–614, 2015.

[32] J. Jing, Q. Li, P. Li, and L. Zhang, "A new method of printed fabric image retrieval based on color moments and gist feature description," *Textile Research Journal*, vol. 86, no. 11, pp. 1137–1150, 2016.

[33] L. Zhang, X. Liu, Z. Lu, F. Liu, and R. Hong, "Lace fabric image retrieval based on multi-scale and rotation invariant lbp," in *Proceedings of the 7th international conference on internet multimedia computing and service*, 2015, pp. 1–5.

[34] Y. Li, H. Luo, G. Jiang, and H. Cong, "Content-based lace fabric image retrieval system using texture and shape features," *The journal of the Textile Institute*, vol. 110, no. 6, pp. 911–915, 2019.

[35] Z. Li, J. Xiang, L. Wang, N. Zhang, R. Pan, and W. Gao, "Yarn-dyed fabric image retrieval using colour moments and the perceptual hash algorithm," *Fibres & Textiles in Eastern Europe*, no. 5 (137, pp. 39–46, 2019.

[36] N. Suciati, D. Herumurti, and A. Y. Wijaya, "Fractal-based texture and hsv color features for fabric image retrieval," in *2015 IEEE international conference on control system, computing and engineering (ICCSCE)*, ser. 2015 IEEE international conference on control system, computing and engineering (ICCSCE). IEEE, 2015, pp. 178–182.

[37] A. H. Rangkuti, V. H. Athala, N. F. Luthfi, S. V. Aditama, M. M. Ramadhan, and A. H. Aslamia, "Enhancement of traditional clothes pattern recognation using convolutional neural network," in *2021 IEEE International Conference on Computing (ICOCO)*. IEEE, 2021, pp. 224–229.

[38] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[39] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[40] M. Zeiler, "Visualizing and understanding convolutional networks," in *European conference on computer vision/arXiv*, vol. 1311, 2014.

[41] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," *arXiv preprint arXiv:1312.6034*, 2013.

[42] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell, "Decaf: A deep convolutional activation feature for generic visual recognition," in *International conference on machine learning*. PMLR, 2014, pp. 647–655.

[43] A. Sharif Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "Cnn features off-the-shelf: an astounding baseline for recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2014, pp. 806–813.

[44] P. Sermanet, "Overfeat: Integrated recognition, localization and detection using convolutional networks," *arXiv preprint arXiv:1312.6229*, 2013.

[45] H. Azizpour, A. S. Razavian, J. Sullivan, A. Maki, and S. Carlsson, "Factors of transferability for a generic convnet representation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 9, pp. 1790–1802, 2015.

[46] H. Azizpour, A. Sharif Razavian, J. Sullivan, A. Maki, and S. Carlsson, "From generic to specific deep representations for visual recognition," in *Proceedings*

of the IEEE conference on computer vision and pattern recognition workshops, 2015, pp. 36–45.

[47] A. Babenko, A. Slesarev, A. Chigorin, and V. Lempitsky, "Neural codes for image retrieval," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*. Springer, 2014, pp. 584–599.

[48] A. Babenko and V. Lempitsky, "Aggregating local deep features for image retrieval," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1269–1277.

[49] G. Tolias, R. Sicre, and H. Jégou, "Particular object retrieval with integral max-pooling of cnn activations," *arXiv preprint arXiv:1511.05879*, 2015.

[50] A. S. Razavian, J. Sullivan, S. Carlsson, and A. Maki, "Visual instance retrieval with deep convolutional networks," *ITE Transactions on Media Technology and Applications*, vol. 4, no. 3, pp. 251–258, 2016.

[51] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 9, pp. 1904–1916, 2015.

[52] J. Fan, N. Zhou, J. Peng, and L. Gao, "Hierarchical learning of tree classifiers for large-scale plant species identification," *IEEE Transactions on Image Processing*, vol. 24, no. 11, pp. 4172–4184, 2015.

[53] K. Yuan, S. Guo, Z. Liu, A. Zhou, F. Yu, and W. Wu, "Incorporating convolution designs into visual transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 579–588.

[54] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, ser. 2005 IEEE computer society conference

on computer vision and pattern recognition (CVPR'05), vol. 1. Ieee, 2005, pp. 886–893.

[55] L. Yao and H. Ke, "Robust image retrieval for lacy and embroidered fabric," *Textile research journal*, vol. 89, no. 13, pp. 2616–2625, 2019.

[56] Y. Li, J. Zhang, M. Chen, H. Lei, G. Luo, and Y. Huang, "Shape based local affine invariant texture characteristics for fabric image retrieval," *Multimedia Tools and Applications*, vol. 78, pp. 15 433–15 453, 2019.

[57] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "Orb: An efficient alternative to sift or surf," in *2011 International conference on computer vision*, ser. 2011 International conference on computer vision. Ieee, 2011, pp. 2564–2571.

[58] S. Gkelios, Y. Boutalis, and S. A. Chatzichristofis, "Investigating the vision transformer model for image retrieval tasks," in *2021 17th International Conference on Distributed Computing in Sensor Systems (DCOSS)*. IEEE, 2021, pp. 367–373.

[59] A. El-Nouby, N. Neverova, I. Laptev, and H. Jégou, "Training vision transformers for image retrieval," *arXiv preprint arXiv:2102.05644*, 2021.

[60] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *International conference on machine learning*. PMLR, 2021, pp. 10 347–10 357.

[61] Y. Chen, S. Zhang, F. Liu, Z. Chang, M. Ye, and Z. Qi, "Transhash: Transformer-based hamming hashing for efficient image retrieval," in *Proceedings of the 2022 international conference on multimedia retrieval*, 2022, pp. 127–136.

[62] T. Li, Z. Zhang, L. Pei, and Y. Gan, "Hashformer: Vision transformer based deep hashing for image retrieval," *IEEE Signal Processing Letters*, vol. 29, pp. 827–831, 2022.

[63] S. R. Dubey, S. K. Singh, and W.-T. Chu, "Vision transformer hashing for image retrieval," in *2022 IEEE international conference on multimedia and expo (ICME)*. IEEE, 2022, pp. 1–6.

[64] C. Henkel, "Efficient large-scale image retrieval with deep feature orthogonality and hybrid-swin-transformers," *arXiv preprint arXiv:2110.03786*, 2021.

[65] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[66] Sivic and Zisserman, "Video google: A text retrieval approach to object matching in videos," in *Proceedings ninth IEEE international conference on computer vision*. IEEE, 2003, pp. 1470–1477.

[67] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "Netvlad: Cnn architecture for weakly supervised place recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, ser. Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 5297–5307.

[68] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, ser. Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 815–823.

[69] J. Wang, H. T. Shen, J. Song, and J. Ji, "Hashing for similarity search: A survey," *arXiv preprint arXiv:1408.2927*, 2014.

[70] L. Zhang, Y. Zhang, J. Tang, X. Gu, J. Li, and Q. Tian, "Topology preserving hashing for similarity search," in *Proceedings of the 21st ACM international conference on Multimedia*, 2013, pp. 123–132.

[71] J. Wang, T. Zhang, N. Sebe, H. T. Shen *et al.*, "A survey on learning to hash," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 769–790, 2017.

[72] R. Xia, Y. Pan, H. Lai, C. Liu, and S. Yan, "Supervised hashing for image retrieval via image representation learning," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 28, no. 1, 2014.

[73] K. Lin, J. Lu, C.-S. Chen, J. Zhou, and M.-T. Sun, "Unsupervised deep learning of compact binary descriptors," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 6, pp. 1501–1514, 2018.

[74] H.-F. Yang, K. Lin, and C.-S. Chen, "Supervised learning of semantics-preserving hash via deep convolutional neural networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 2, pp. 437–451, 2017.

[75] C. Zhang-Jie, L. Ming-Sheng, W. Jian-Min, and Y. P. S. Hashnet, "Deep learning to hash by continuation," in *Proceedings of IEEE International Conference on Computer Vision, Venice, Italy*, 2018, pp. 5609–5618.

[76] X. Yan, L. Zhang, and W.-J. Li, "Semi-supervised deep hashing with a bipartite graph." in *IJCAI*, 2017, pp. 3238–3244.

[77] A. Kelman, M. Sofka, and C. V. Stewart, "Keypoint descriptors for matching across multiple image modalities and non-linear intensity variations," in *2007 IEEE conference on computer vision and pattern recognition*. IEEE, 2007, pp. 1–7.

[78] T. Collins, P. Mesejo, and A. Bartoli, "An analysis of errors in graph-based keypoint matching and proposed solutions," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VII 13*. Springer, 2014, pp. 138–153.

[79] D. DeTone, T. Malisiewicz, and A. Rabinovich, "Deep image homography estimation," *arXiv preprint arXiv:1606.03798*, 2016.

[80] T. Nguyen, S. W. Chen, S. S. Shivakumar, C. J. Taylor, and V. Kumar, "Unsupervised deep homography: A fast and robust homography estimation model," *IEEE Robotics and Automation Letters*, vol. 3, no. 3, pp. 2346–2353, 2018.

[81] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *2007 IEEE conference on computer vision and pattern recognition*. IEEE, 2007, pp. 1–8.

[82] K. Mikolajczyk and C. Schmid, "Scale & affine invariant interest point detectors," *International journal of computer vision*, vol. 60, pp. 63–86, 2004.

[83] M. A. Fischler and R. C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[84] S. Hausler, S. Garg, M. Xu, M. Milford, and T. Fischer, "Patch-netvlad: Multi-scale fusion of locally-global descriptors for place recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 14 141–14 152.

[85] P.-E. Sarlin, D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superglue: Learning feature matching with graph neural networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 4938–4947.

[86] R. Wang, Y. Shen, W. Zuo, S. Zhou, and N. Zheng, "Transvpr: Transformer-based place recognition with multi-level attention aggregation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 13 648–13 657.

[87] O. Chum, J. Philbin, J. Sivic, M. Isard, and A. Zisserman, "Total recall: Automatic query expansion with a generative feature model for object retrieval,"

in *2007 IEEE 11th International Conference on Computer Vision*. IEEE, 2007, pp. 1–8.

[88] H. Xie, Y. Zhang, J. Tan, L. Guo, and J. Li, "Contextual query expansion for image retrieval," *IEEE Transactions on Multimedia*, vol. 16, no. 4, pp. 1104–1114, 2014.

[89] S. Bai and X. Bai, "Sparse contextual activation for efficient visual re-ranking," *IEEE Transactions on Image Processing*, vol. 25, no. 3, pp. 1056–1069, 2016.

[90] X. Bai, S. Bai, and X. Wang, "Beyond diffusion process: Neighbor set similarity for fast re-ranking," *Information Sciences*, vol. 325, pp. 342–354, 2015.

[91] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, pp. 91–110, 2004.

[92] J. Chen, L. Wang, X. Li, and Y. Fang, "Arbicon-net: Arbitrary continuous geometric transformation networks for image registration," *Advances in neural information processing systems*, vol. 32, 2019.

[93] J. Wang and M. Zhang, "Deepflash: An efficient network for learning-based medical image registration," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 4444–4452.

[94] T. C. Mok and A. Chung, "Fast symmetric diffeomorphic image registration with convolutional neural networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 4644–4653.

[95] Q. Gu, Y. Xiao, L. Yuan, M. Shen, F. Zhang, H. Liao, and J. Liu, "Image retrieval of colored spun fabrics based on variable weight and semantic features," *The Journal of The Textile Institute*, pp. 1–7, 2024.

[96] S. Tena, R. Hartanto, and I. Ardiyanto, "Content-based image retrieval for traditional indonesian woven fabric images using a modified convolutional neural network method," *Journal of Imaging*, vol. 9, no. 8, p. 165, 2023.

[97] D. Xu, Y. Li, and H. Luo, "Lace fabric image retrieval using siamese neural network," in *2021 IEEE 6th International Conference on Signal and Image Processing (ICSIP)*, ser. 2021 IEEE 6th International Conference on Signal and Image Processing (ICSIP). IEEE, 2021, pp. 550–555.

[98] J. Xiang, N. Zhang, R. Pan, and W. Gao, "Fabric retrieval based on multi-task learning," *IEEE Transactions on Image Processing*, vol. 30, pp. 1570–1582, 2020.

[99] ——, "Fabric image retrieval system using hierarchical search based on deep convolutional neural network," *Ieee Access*, vol. 7, pp. 35 405–35 417, 2019.

[100] X. Wang, G. Wu, and Y. Zhong, "Fabric identification using convolutional neural network," in *Artificial Intelligence on Fashion and Textiles: Proceedings of the Artificial Intelligence on Fashion and Textiles (AIFT) Conference 2018, Hong Kong, July 3–6, 2018*, ser. Artificial Intelligence on Fashion and Textiles: Proceedings of the Artificial Intelligence on Fashion and Textiles (AIFT) Conference 2018, Hong Kong, July 3–6, 2018. Springer, 2019, pp. 93–100.

[101] D. Deng, R. Wang, H. Wu, H. He, Q. Li, and X. Luo, "Learning deep similarity models with focus ranking for fabric image retrieval," *Image and Vision computing*, vol. 70, pp. 11–20, 2018.

[102] N. Zhang, R. Shamey, J. Xiang, R. Pan, and W. Gao, "A novel image retrieval strategy based on transfer learning and hand-crafted features for wool fabric," *Expert Systems with Applications*, vol. 191, p. 116229, 2022.

[103] J. Xiang, N. Zhang, R. Pan, and W. Gao, "Patterned fabric image retrieval using relevant feedback via geometric similarity," *Textile Research Journal*, vol. 92, no. 3-4, pp. 409–422, 2022.

[104] N. Zhang, J. Xiang, L. Wang, N. Xiong, W. Gao, and R. Pan, "Image retrieval of wool fabric. part ii: based on low-level color features," *Textile research journal*, vol. 90, no. 7-8, pp. 797–808, 2020.

[105] L. Zhang, X. Liu, Z. Lu, F. Liu, and R. Hong, "Lace fabric image retrieval based on multi-scale and rotation invariant lbp," in *Proceedings of the 7th international conference on internet multimedia computing and service*, ser. Proceedings of the 7th international conference on internet multimedia computing and service, 2015, pp. 1–5.

[106] I. O. Tolstikhin, N. Houlsby, A. Kolesnikov, L. Beyer, X. Zhai, T. Unterthiner, J. Yung, A. Steiner, D. Keysers, and J. Uszkoreit, "Mlp-mixer: An all-mlp architecture for vision," *Advances in neural information processing systems*, vol. 34, pp. 24 261–24 272, 2021.

[107] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *International journal of computer vision*, vol. 42, pp. 145–175, 2001.

[108] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 24, no. 7, pp. 971–987, 2002.

[109] R. M. Haralick, K. Shanmugam, and I. H. Dinstein, "Textural features for image classification," *IEEE Transactions on systems, man, and cybernetics*, no. 6, pp. 610–621, 1973.

[110] F. Kurugollu, B. Sankur, and A. E. Harmanci, "Color image segmentation using histogram multithresholding and fusion," *Image and vision computing*, vol. 19, no. 13, pp. 915–928, 2001. [Online]. Available: https://api.semanticscholar.org/CorpusID:125542633

[111] A. Ali-Bey, B. Chaib-Draa, and P. Giguere, "Mixvpr: Feature mixing for visual place recognition," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, ser. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2023, pp. 2998–3007.

[112] A. Bellet, A. Habrard, and M. Sebban, "A survey on metric learning for feature vectors and structured data," *arXiv preprint arXiv:1306.6709*, 2013.

[113] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon, "Information-theoretic metric learning," in *Proceedings of the 24th international conference on Machine learning*, ser. Proceedings of the 24th international conference on Machine learning, 2007, pp. 209–216.

[114] L. Yang and R. Jin, "Distance metric learning: A comprehensive survey," *Michigan State Universiy*, vol. 2, no. 2, p. 4, 2006. [Online]. Available: https://api.semanticscholar.org/CorpusID:850937

[115] A. Globerson and S. Roweis, "Metric learning by collapsing classes," *Advances in neural information processing systems*, vol. 18, 2005.

[116] X. Shi and X. Qian, "Exploring spatial and channel contribution for object based image retrieval," *Knowledge-Based Systems*, vol. 186, p. 104955, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0950705119303910

[117] E. Hoffer and N. Ailon, "Deep metric learning using triplet network," in *Similarity-Based Pattern Recognition: Third International Workshop, SIMBAD 2015, Copenhagen, Denmark, October 12-14, 2015. Proceedings 3*, ser. Similarity-Based Pattern Recognition: Third International Workshop, SIMBAD 2015, Copenhagen, Denmark, October 12-14, 2015. Proceedings 3. Springer, 2015, pp. 84–92.

[118] S. Varshney, S. Singh, C. V. Lakshmi, and C. Patvardhan, "Content-based image retrieval of indian traditional textile motifs using deep feature fusion,"

Scientific Reports, vol. 14, no. 1, p. 10035, 2024. [Online]. Available: https://www.nature.com/articles/s41598-024-56465-9

[119] J. Gui, D. Wu, and J. He, "An efficient network based on double constrained loss for fabric image retrieval," *Signal, Image and Video Processing*, vol. 18, no. 1, pp. 305–313, 2024. [Online]. Available: https://link.springer.com/article/10.1007/s11760-023-02749-y

[120] J. Shen, L. Yuan, and J. Xiong, "Image retrieval of colored spun fabrics based on decoupled feature," in *International Conference on Computer Graphics, Artificial Intelligence, and Data Processing (ICCAID 2023)*, vol. 13105. SPIE, 2024, pp. 34–39.

[121] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2006.

[122] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[123] J. Doe and J. Smith, "Attention-based query expansion learning," *arXiv preprint arXiv:2007.08019*, 2020. [Online]. Available: https://arxiv.org/abs/2007.08019

[124] A. Brown and P. Johnson, "Contextual query expansion for image retrieval," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013, pp. 673–680. [Online]. Available: https://ieeexplore.ieee.org/document/6739088

[125] E. Green and D. White, "Online query expansion hashing for efficient image retrieval," *IEEE Transactions on Image Processing*, vol. 31, pp. 1024–1036, 2022. [Online]. Available: https://ieeexplore.ieee.org/document/10185594

[126] Y. Leong and R. Mihalcea, "Improving query expansion for image retrieval via saliency and picturability," in *Working Notes for CLEF 2011 Conference*, 2011. [Online]. Available: https://web.eecs.umich.edu/ mihalcea/papers/leong.clef11.pdf

[127] M. Wu and L. Chen, "Query expansion for object retrieval with active learning using bow and cnn features," in *International Symposium on Computer Vision and Applications (ISVA)*, 2017. [Online]. Available: https://ise.thss.tsinghua.edu.cn/MIG/2017-6.pdf

[128] O. Chum and J. Matas, "Locally optimized ransac," in *Pattern Recognition, 2003. Proceedings. 25th DAGM Symposium.* Springer, 2003, pp. 236–243.

[129] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision.* Cambridge University Press, 2003.

[130] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2. IEEE, 2006, pp. 2169–2178.

[131] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1. IEEE, 2005, pp. 886–893.

[132] T. S. Caetano, J. J. McAuley, L. Cheng, Q. V. Le, and A. J. Smola, "Learning graph matching," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 6, pp. 1048–1058, 2009.

[133] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Improving image retrieval by spatial re-ranking," in *2007 IEEE Conference on Computer Vision and Pattern Recognition.* IEEE, 2007, pp. 1–8.

[134] S. Belongie, J. Malik, and J. Puzicha, "Shape matching and object recognition using shape contexts," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 4, pp. 509–522, 2002.

[135] A. Tewari, P. Mittal, S. Verma *et al.*, "Hashing for large-scale image retrieval: A survey," *arXiv preprint arXiv:1703.02938*, 2017.