# Phần I

# Content-based image retrieval

## 0.1 Problem Definition: The Semantic Gap in CBIR

The primary goal of a Content-Based Image Retrieval (CBIR) system is to retrieve images from a vast database that are visually and conceptually similar to a given query image. Unlike traditional metadata-based search engines that rely on text tags or keywords, a CBIR system must process the raw pixel data to identify relevant objects, textures, and architectures.

### 0.1.1 The Semantic Gap

The fundamental challenge in CBIR is bridging the **Semantic Gap**. In computer vision, pixels are merely numerical values representing intensity or color; however, a functional retrieval system must interpret these numbers as high-level concepts such as "landmark," "baroque architecture," or "specific object identity".

### 0.1.2 Objectives

To build an effective system, the following objectives must be met:

- **Feature Extraction:** The model must extract robust, invariant features from images using deep convolutional neural networks (CNNs).
- **Embedding Space:** The system must map these features into a high-dimensional vector space (embedding space) where similarity is measured by distance.
- **Scalability:** The system must handle large-scale datasets, such as Oxford5k and Paris6k, while maintaining high precision and retrieval speed.

### 0.1.3 Applications

This technology is essential for modern digital asset management, e-commerce visual searches, and mobile visual search engines like Google Lens.

## 0.2 Methodology: Deep Metric Learning

To solve the image retrieval problem, this project utilizes **Deep Metric Learning**. Unlike standard classification, which assigns an image to a fixed category, metric learning aims to learn a mapping from the input image space to a high-dimensional embedding space. In this space, the Euclidean distance between two vectors directly corresponds to the visual similarity between the two respective images.

### 0.2.1 Siamese Network with Triplet Architecture

The system is built upon a Siamese Network using a **Triplet Architecture**. During each training iteration, the network is presented with three distinct images:

- **Anchor ($x_a$):** The reference image from the dataset.
- **Positive ($x_p$):** An image representing the same object or landmark as the

anchor.

- **Negative ($x_n$):** An image representing a different object or landmark.

The network shares the same weights across these three inputs to ensure that all images are projected into the same embedding space using the same feature extraction logic.

#### 0.2.2 Objective Function: Triplet Margin Loss

The model is optimized using the **Triplet Margin Loss** function. The mathematical objective is to minimize the distance between the Anchor and the Positive ($d(a, p)$) while maximizing the distance between the Anchor and the Negative ($d(a, n)$) beyond a specific margin ($\alpha$).

The loss function $L$ is defined as follows:

$$L(a, p, n) = \max(d(a_i, p_i) - d(a_i, n_i) + \alpha, 0)$$

This ensures that the positive pair is closer to the anchor than the negative pair by at least the margin $\alpha$. As learning progresses, the network refines its internal weights to "pull" similar images together and "push" dissimilar images apart in the 2048-dimensional vector space.

### 0.3 Experimental Setup and Configuration

To evaluate the effectiveness of deep metric learning for image retrieval, the models were trained and tested using standardized benchmark datasets and a controlled set of hyperparameters. This ensures the comparability of results between different neural network backbones.

#### 0.3.1 Benchmark Datasets

The system was evaluated on two widely recognized landmark datasets, which provide the structural variety needed to test the "Semantic Gap":

- **Oxford5k:** This dataset contains 5,063 images of 11 distinct Oxford landmarks. For each landmark, images are categorized as "good," "ok," "junk," or "query" to facilitate precise evaluation.
- **Paris6k:** A similar dataset consisting of approximately 6,000 images of Paris landmarks, used to test the model's generalization across different architectural styles.

### 0.3.2 Training Configuration and Hyperparameters

The training process was governed by a specific configuration to allow for convergence over 80 epochs. Both the ResNet-50 and VGG-19 models were trained using the following parameters derived from the system configuration:

- **Optimizer:** The Adam optimizer was used to manage the learning process.
- **Learning Rate:** A consistent rate of $1 \times 10^{-5}$ was maintained.
- **Batch Size:** Set to 32 for efficient GPU utilization.
- **Embedding Dimension:** Each image is projected into a 2048-dimensional vector space.
- **Hardware Acceleration:** The training was executed on CUDA-enabled devices to handle the computational intensity of deep metric learning.

### 0.3.3 Evaluation Metrics

The retrieval performance is quantified using several standard information retrieval metrics:

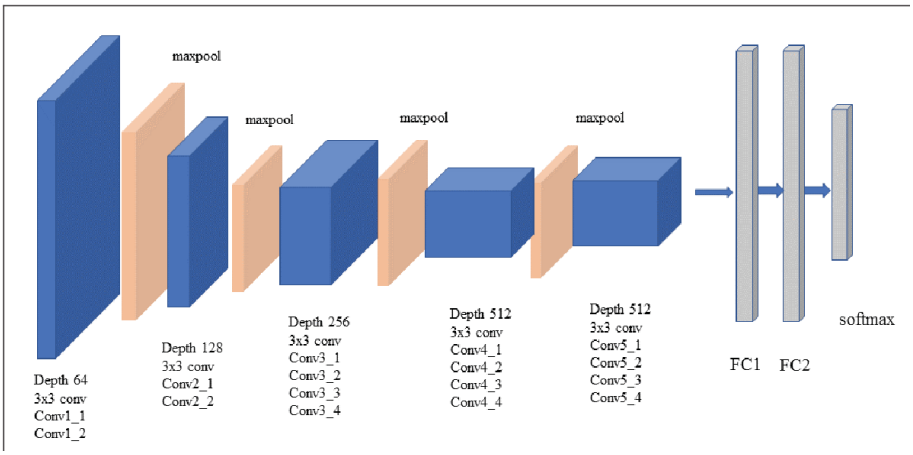
- **mean Average Precision (mAP):** Measures the overall accuracy of the ranked results.
- **Mean Reciprocal Rank (MRR):** Focuses on the position of the first relevant result.
- **NDCG@10 & NDCG@20:** Normalized Discounted Cumulative Gain at specific ranks to evaluate the quality of the top-ranked retrievals.

## 0.4 Architectural Comparison: VGG-19 vs. ResNet-50

The backbone of the CBIR system is responsible for transforming raw pixel input into a compact 2048-dimensional feature vector. In this project, we compare a traditional deep linear network (VGG-19) with a more modern residual-based architecture (ResNet-50).

### 0.4.1 VGG-19 Architecture

VGG-19 is characterized by its simplicity and depth, utilizing a sequence of $3 \times 3$ convolutional filters with increasing channel depths (from 64 up to 512).

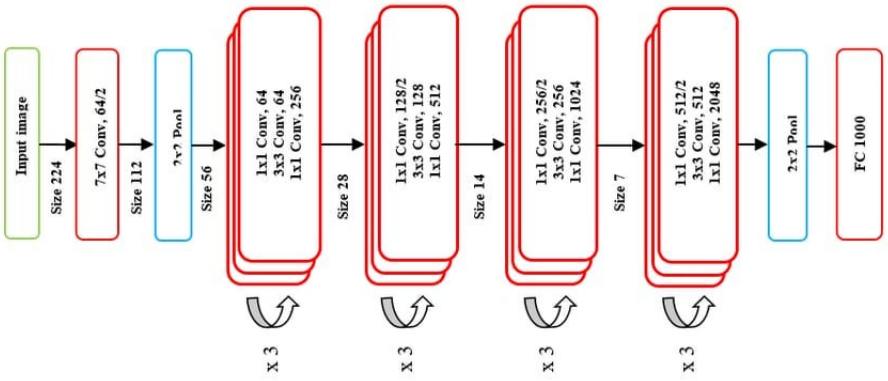



**Hình 0.1:** VGG-19 Architecture

- **Structure:** Consists of 16 convolutional layers followed by 3 fully connected layers.
- **Feature Extraction:** The model relies on stacked convolutions to increase the receptive field, allowing it to learn hierarchical features.
- **Limitation:** As a deep linear network, it is more susceptible to the vanishing gradient problem compared to residual networks.

### 0.4.2 ResNet-50 Architecture

ResNet-50 (Residual Network) introduces "shortcut connections" or "skip connections" that allow gradients to flow more easily through the network during backpropagation.




**Hình 0.2:** Resnet-50 Architecture

- **Structure:** Composed of 50 layers organized into bottleneck blocks.
- **Residual Learning:** Instead of learning a direct mapping, the network learns the residual mapping ($F(x) = H(x) - x$), which makes it easier to optimize very deep architectures.
- **Efficiency:** Despite having more layers than VGG-19, the use of bottleneck

layers and global average pooling makes it computationally efficient and highly effective at capturing architectural nuances.

### 0.4.3 Adaptation for Metric Learning

For both models, the final classification layer (Softmax) was replaced with a fully connected linear layer. This layer projects the extracted features into the common 2048-dimensional embedding space required for the Triplet Margin Loss calculation.

## 0.5 Quantitative Results and Performance Analysis

The evaluation phase involves testing the trained ResNet-50 and VGG-19 models on a combined test set of 6,960 images. To ensure a robust assessment of retrieval quality, several metrics were employed, including mean Average Precision (mAP), Mean Reciprocal Rank (MRR), and Normalized Discounted Cumulative Gain (NDCG) at ranks 10 and 20.

### 0.5.1 Performance Metrics Comparison

The following table summarizes the quantitative performance of both architectures. While both models were trained under identical conditions (80 epochs, $1 \times 10^{-5}$ learning rate), the results reveal a clear performance gap.

| Metric | ResNet-50 | VGG-19 |
|---|---|---|
| **Index Size (MB)** | 54.90 | 54.90 |
| **Vector Dimension** | 2048 | 2048 |
| **mAP** | **0.8640** | 0.8002 |
| **MRR** | **0.9864** | 0.9301 |
| **NDCG@10** | **0.9879** | 0.9277 |
| **NDCG@20** | **0.9851** | 0.9314 |

**Bảng 1:** Quantitative Results Comparison on Benchmark Datasets

### 0.5.2 Analysis of Results

The data indicates that **ResNet-50** significantly outperforms VGG-19 across every accuracy-based metric:

- **mAP Advantage:** ResNet-50 achieved an mAP of 0.8640, approximately 8% higher than the 0.8002 recorded by VGG-19. This suggests that the residual connections allow for superior feature representation in the embedding space.
- **High-Rank Precision:** The MRR of 0.9864 for ResNet-50 implies that, in nearly every query, the first relevant image is returned at the very top of the list.
- **Structural Efficiency:** Despite the difference in retrieval quality, both models

maintain an identical index size of 54.90 MB and a vector dimension of 2048, proving that ResNet-50 provides higher semantic density within the same memory footprint.

### 0.5.3 Comparative Analysis: Why the Performance Differs

The quantitative results demonstrate a clear superiority of the ResNet-50 architecture over VGG-19 for the task of landmark retrieval. This disparity can be attributed to several key technical factors:

- **Residual Learning vs. Plain Stacking:** The defining factor is ResNet's use of skip connections. In VGG-19, information must pass through every layer linearly, which can lead to the "vanishing gradient" problem and a degradation of fine-grained architectural features. ResNet-50's residual blocks allow the model to learn identity mappings, preserving essential low-level structural details (like the specific curves of a dome) while simultaneously learning high-level semantic identities.
- **Feature Representational Power:** Although both models project images into a 2048-dimensional space, the quality of these embeddings differs. ResNet-50 captures more "discriminative" features. This is evidenced by the mAP of 0.8640 compared to VGG-19's 0.8002, suggesting that ResNet is more effective at grouping similar landmarks together even under varying angles and lighting conditions.
- **Generalization Across Datasets:** The high MRR and NDCG values for ResNet-50 (0.98+) across both Oxford5k and Paris6k indicate that its deep features are highly robust to geographical and architectural variations. While VGG-19 performs well, its slightly lower MRR (0.9301) suggests it occasionally ranks "distractor" images or similar-looking but incorrect buildings higher than the true positive.
- **Embedding Density:** Both models produce the same index size (54.90 MB) for the 6,960 images. However, ResNet-50 achieves higher retrieval accuracy within this identical memory footprint. This indicates that ResNet-50 achieves a higher "semantic density"—packing more relevant information into the same 2048-dimensional vector than VGG-19.

## 0.6 Qualitative Assessment: Visual Retrieval Results

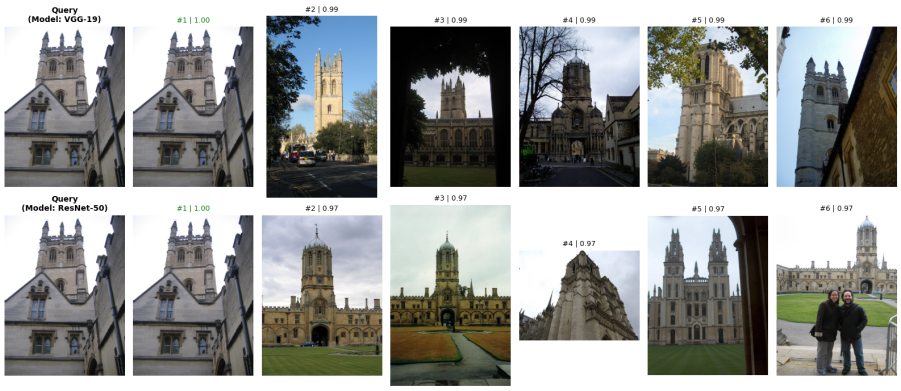
A qualitative assessment allows us to verify the model's ability to handle visual variations such as lighting, perspective, and occlusions. The following figures demonstrate the top-6 retrieved images for various landmark queries using both VGG-19 and

ResNet-50 backbones.

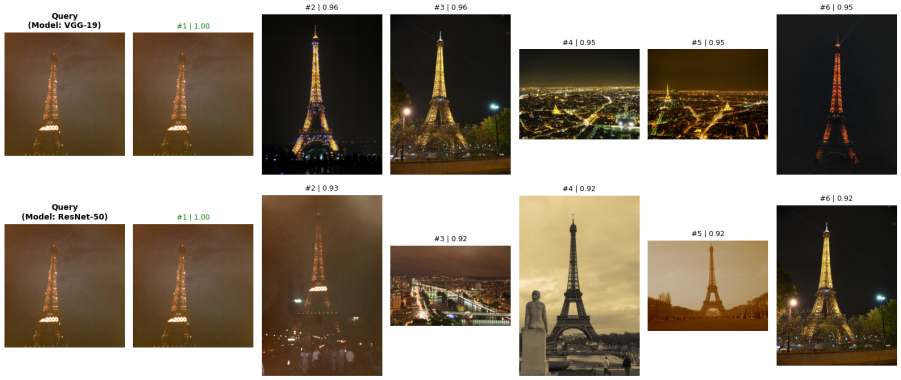
### 0.6.1 Visual Comparison: Dome and Cathedral Landmarks

In queries involving complex architectural details like domes (e.g., the Invalides or Sacré-Cœur), both models demonstrate high precision. However, the similarity scores returned by ResNet-50 are consistently more stable across the top ranks.










### 0.6.2 Handling Complex Backgrounds and Night Scenes

Queries performed on the Eiffel Tower at night demonstrate the robustness of the deep features. Despite the challenging "night-to-night" or "night-to-day" matching, the ResNet-50 backbone effectively ignores the complex dark backgrounds to focus on the structural identity of the tower.

## 0.7 Conclusion

This report explored two distinct domains of computer vision: classical morphological counting and modern deep metric learning for image retrieval.

1. **Rice Grain Counting:** The classical pipeline proved highly effective for controlled

agricultural environments. By combining Top-Hat transforms, directional kernels ($1 \times 80$), and the Watershed algorithm, the system achieved accurate counts (94–98 grains) despite noise and complex backgrounds.

2. **Image Retrieval (CBIR):** The deep learning approach successfully bridged the "Semantic Gap". The comparative study showed that **ResNet-50** is the superior backbone for this task, achieving an mAP of 0.8640. This confirms that residual architectures are better suited for learning high-density semantic embeddings than traditional linear stacks like VGG-19.

The integration of these techniques provides a robust foundation for automated visual analysis systems, ranging from precision agriculture to large-scale digital asset management.