IJICTR
International Journal of Information & Communication Technology Research

Volume 16- Number 2 – 2024 (34 -44)

Image Retrieval with Missing Regions by Reconstructing and Concatenating Content and Semantic Features

Fatemeh Taheri
Department of Computer Engineering, South Tehran Branch, Islamic Azad University, Tehran, Iran

Kambiz Rahbar*
Department of Computer Engineering, South Tehran Branch, Islamic Azad University, Tehran, Iran
K_rahbar@azad.ac.ir

Ziaeddin Beheshtifard
Department of Computer Engineering, South Tehran Branch, Islamic Azad University, Tehran, Iran

Received: 6 December 2022 – Revised: 30 May 2023 - Accepted: 13 March 2023

Abstract—**In recent years, the performance of deep neural networks in improving the image retrieval process has been remarkable. Utilizing deep neural networks; however, leads to poor results in retrieving images with missing regions. The operators' dysfunctions, who consider the relationship between the image pixels, statistically extract incomplete information from an image, which in turn reduces the number of image features and or leads to features' inaccurate identification. An attempt has been made to eliminate the problem of missing image information through image inpainting techniques; therefore, a content-based image retrieval method is proposed for images with missing regions. In this method, through image inpainting the crucial missing information is reconstructed. The image dataset is being queried to find similar samples. For this purpose, a two-stage inpainting framework based on encoder-decoder is used in the image retrieval system. Also, the features of each image are extracted from the integration and concatenating of content and semantic features. Through using handcraft features such as color and texture image content information is extracted from the Resnet-50 deep neural network. Finally, similar images are retrieved based on the minimum Euclidean distance. The performance of the image retrieval model with missing regions is evaluated with the average precision criterion on the Paris 6K datasets. The best retrieval results are 60.11%, 50.14%, and 42.43% for retrieving the top one, five, and ten samples after reconstructing the image with the most missing regions with a destruction frequency of 6 Hz, respectively.**

Keyword: image inpainting, content features, semantic features, ResNet-50, deep neural network.

Article type: Research Article

Publisher: ICT Research Institute

* Corresponding Author

I. Introduction

Image retrieval, an extension of image classification, is a method of structuring samples of a large dataset of unstructured nature. The classification subject defined in a set with a limited number of predetermined classes. On the other hand, in image retrieval, the limitation of the number of image classes and the need for the image class label is eliminated. Thus, it is easily possible to add new samples to the dataset without considering the mentioned limitations. Therefore, image retrieval drew the attention of researchers in a wide range of application fields. The content-based image retrieval systems attempt to retrieve images with the highest degree of similarities to the user's query image. To meet this end, extracting descriptive features of the query image, and all database images is of prime importance. The extracted feature vector is exploited as a basis for comparison to retrieving similar images. It is necessary to extract the content and semantic features of the image in a way that provides an effective and efficient description of the image. In this regard, the efficiency of deep neural networks in extracting image features has made these networks an effective solution in today's image retrieval systems [1], [2]. However, in the process of querying images with missing regions, the performance of the image retrieval system is severely affected [3]. A mechanism that can effectively reconstruct the images with missing regions leads to improve retrieval results. This is possible with image inpainting techniques.

The content-based image retrieval system is composed of feature extraction and similarity measurement. In the first step, the features of the image are extracted and form the feature vector. In the second step, similarity measurement implemented to retrieve samples similar to the query image. Content features often extracted from the image color, texture, and shape. Color feature extraction using color histogram widely utilized in content-based image retrieval systems due to its simplicity and stability with respect to rotation and scale change. Since the histogram lacks spatial information on color distribution, other improved methods such as the histogram of triangles [4] and spatial color histogram [5] received attention. Also, color's Auto-correlogram [6], and color coherence vector [7] are among other methods of extracting color features to retrieve images. In color's Auto-correlogram technique, correlation and spatial distribution of colors are considered color features. These features are calculated based on the distance of the same colors in the pixels of the image. Texture feature extraction is often accomplished using gray level co-occurrence matrix (GLCM) [8], local binary pattern [9], and Gabor filter [10]. Gray level co-occurrence matrix and local binary pattern (LBP) are texture analysis statistical methods. The distinguishing power and simplicity of calculations are the characteristics of these two approaches. Gabor wavelet transformation is also an effective method for extracting texture features in the form of primary patterns analysis due to the possibility of examining image texture at different scales and angles. Integrating low-level features in the image and the extraction of mid-level features for improving the results of content-based image retrieval have also been studied in this method [11]. However, describing the semantic features of the image with the help of low-level features is considered a shortcoming of these approaches in retrieving similar samples.

The use of convolutional neural networks to extract high-level features or semantic content of the image is recently in use [12][13]. Utilizing the convolution layers for image semantics interpretation is affected by applying the filter to the whole image. There are two categories in implementing the deep neural networks approach to image retrieval: Models based on Re-train neural networks or fine-tuning and pre-trained models. Neural networks performance in network retraining models significantly depends on the training samples. Producing several samples necessary for training is one of the challenges of this approach. Pre-trained neural networks are pre-trained with a large set of data. Image features extracted from different convolutional layers of the network are improved, thus providing a better semantic interpretation of the image than low-level features [13]. Image semantic interpretation with the help of convolution layers is affected by applying the filter on the whole image. Therefore, changes such as luminance, noise, and missing regions in the image can affect the number and accuracy of the extracted features. Some of the pre-trained networks implemented in image retrieval are AlexNet [14], VGG [15], GoogLeNet [16], and ResNet [17]. For example, in [18], image features are extracted using two networks, VGG-19 and GoogleNet. Integrating the features of two networks and then reducing the dimensions of the feature vector is suggested in this method. In [19] first, the high-level features are extracted using the ResNet network, and then the re-ranking and improvement of the initial retrieval results are performed using the content features. Since an image includes concepts and physical features such as color, texture, shape, and semantic concepts, integrating content and semantic features allows to describe the image more precisely. Combining handcraft features with high-level features has also been considered to enhance the desired features in retrieving specific images in [20, 21]. In [20], a combination of high-level features using convolutional neural networks and wavelet transform space is used to extract texture features to retrieve texture images. In the method [21], to reduce the semantic gap between the description of image features and the way humans understand its semantics, the pre-trained AlexNet network features and its integration with handcraft features are utilized. Despite the acceptable performance of convolutional neural networks, extracting semantic features from an image with missing regions is challenging. As mentioned earlier, applying the filter to the invalid space caused by pixels

and missing regions of the image either reduces the number of features or extracts invalid features. As a result, the performance of the image retrieval system decreases [22]. One way to deal with this matter is the imputation approach, which aims to replace missing values with reasonable values [23]. Image inpainting is one of these methods [22].

Image inpainting involves correcting the missing region of an image or video so that it is not visually noticeable. In digital terms, this process is known as inpainting, which reconstruct the missing parts of the image [24]. This technique is used to modify, edit, encode, and transfer the image [25]. Image inpainting techniques are divided into traditional and conventional categories based on how they work. Diffusion-based and path-based methods are traditional techniques [26]. In diffusion-based models, pixel information is spread around the missing region in the image. This process is often accompanied by exemplar-based texture synthesis or exemplar-based structure synthesis [27] around the missing region. In these methods, the statistical information of patterns is used due to the fixed distribution of missing regions and known parts of the image. This method is usually modeled with Markov Random Fields (MRF) [28]. Diffusion-based methods are limited only to the affected regions. The larger the destructed parts, the lesser useful information is generated in its center. This method lacks a high-level semantic understanding of the image. Because of this, the results of image inpainting may lead to apparent inconsistencies in the image context. In patch-based methods, the visible regions of the image are queried to match the destructed part and copied to it [29]. In these methods, the best matching patch is queried for the target region in the image. After finding the most similar part, the pixel values are copied from the source area and overwritten to the target one. The choice of different patch matching criteria, patch size, and the order of filling the missing region influence the results of image inpainting. For this reason, the main challenge of completing missing regions in images is maintaining visual coherence throughout the image.

Modern image inpainting approaches function primarily based on Generative Adversarial Networks [30–33] and convolutional neural networks [34–36]. These methods can inpaint different sizes of missing regions in the image and achieve acceptable results. The visual and semantic features for reconstructing the damaged regions of the image are considered the strength of deep learning-based approaches. For example, in the method [32], considering the challenges of adversarial generator networks in producing high-resolution images and precise texture details, the architecture with multiple generator networks is proposed. The four generative and discriminative networks operate progressively. Primary generators improve the consistency of the overall image structure. The last generators also enhance the image's details, accuracy, and resolution. Also, using a loss function based on the local binary pattern algorithm minimizes the difference between the generated textures and the original sample. The method [30] considers the multi-stage progressive reconstruction of Coarse to fine. A controller filter of changes has been used in each step to avoid the negative effect of image inpainting. The image Saliency region has been suggested as a detail reconstruction controller. Although deep generative models can produce visually acceptable structures and textures, these models cannot produce diverse results for each input. For this reason, in the method [34], a framework based on the convolutional neural network has been introduced to inpaint the image in two consecutive steps in coarse to fine form. At first, the inpainting process is formulated as a regression case. A U-Net convolutional neural network is applied to draw the input to an overall image. Next, pixel matching based on the K-nearest neighbor is utilized for drawing the overall output to the high-quality one with more precise details. By inserting the missing information from different training samples, the second step has made the content of the new inpainted image to be of high quality.

Image retrieval with missing regions has various practical applications in different fields. For instance, in medical imaging, images may contain missing regions due to factors such as patient movement, artifacts, or improper positioning during image labeling. Retrieval of similar images with missing regions can help to provide more complete information about the patient's condition for accurate diagnosis and treatment planning. Additionally, in satellite and aerial image analysis, images may have missing regions due to weather conditions, cloud cover, or other environmental changes. The recovery of missing regions in such images is also effective in tasks such as monitoring land use, natural disasters, and other environmental changes. Furthermore, on social media platforms, images may have missing regions due to subtitles or an inappropriate placement of a media logo or icon. Retrieval of similar images after an image restoration step is effective in organizing visual content and improving user experience.

A content-based image retrieval model for images with missing regions is introduced in this article. Since the missing regions of the image affect image pixels connection and feature extraction, the image reconstruction mechanism before retrieval is considered. The coarse-to-fine two-stage framework based on the encoder-decoder architecture for retrieving the information of the query image is included in the proposed image retrieval model. Also, integrating the image content and semantic features form a feature vector describing that image. The similarity measurement of the query image and the dataset images is also performed based on the Euclidean distance of the feature vector. This article

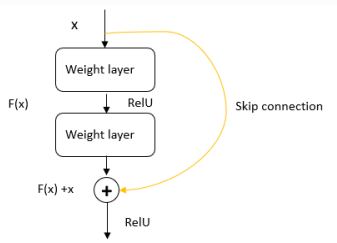

Figure 1. Skip connection block architecture in Resnet-50 network

organizes as follows. In section 2, the proposed method is described. In section 3, the implementation and evaluation results are examined. Section 4 is devoted to the conclusion of the research.

II. IMAGE RECONSTRUCTION AND RETRIEVAL

Fig 1. shows the flowchart of the proposed content-based image retrieval system for the query image with missing regions. First, a two-stage coarse-to-fine architecture is applied to reconstruct the I_g image with missing regions. Then, in the retrieval section of similar images, the semantic features of F_d are extracted using the pre-trained neural network Resnet-50. Content features also include color F_c and texture F_t. The second part includes image features. Concatenating the content and semantic of the final feature vector $F_v = \{F_d, F_c, F_t\}$ forms the query image. In a similar way, the feature vector for all the dataset images is extracted, and a database of image feature vectors is created. The similarity measurement between the query image feature vector and the database images feature vectors is performed based on the Euclidean distance. In this way, images with the highest degree of similarity to the query image are retrieved.

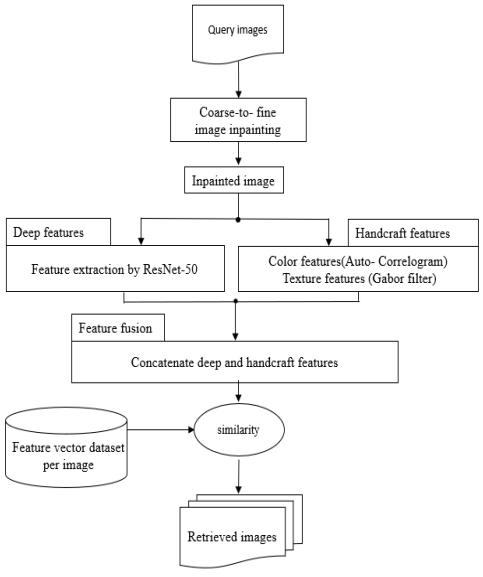

Figure 2. Flowchart of the proposed image retrieval method with missing regions

- **Feature extraction using deep neural network ResNet-50**

ResNet-50 neural network consists of convolution layers, pooling layers, and skip connection. This network introduces image classification [17]. Deep neural networks extract different features from the image in each network layer by applying disparate filters through convolution layers. Therefore, more layers are added to neural networks to extract more complex content and semantic features.

But with the increase of convolution layers, network training is affected by learning parameters and vanishing gradient issues, which consequently reduces network performance. To eliminate these disadvantages, layering process with zero padding or convolution layer method with dimensions 1×1 is utilized to reduce input dimensions. Fig 2 shows a block with residual connections. This shortcut connection crosses one or more layers and connects it to the further one. Therefore, it is possible to add up to 150 more layers to the ResNet-50 network. Because of the extra connection, the x value of the previous layer is also observed. Since the output of the convolution layers has different dimensions along the network, the value of x that adds the shortcut from the previous layers to $f(x)$ possibly creates different dimensions. To eliminate these disadvantages, a layering process with zero padding or a convolution layer method with 1×1 dimension is utilized to reduce input dimensions. The penultimate layer output with 2048 features is used as a feature vector describing the semantic information of the image in the proposed method.

- **Features of color and texture**

Autocorrelation is utilized to extract the image color feature. The Gabor filter is performed to calculate the image texture features. Automatic image correlation is one of the techniques in which spatial information is concatenated with color histograms to extract image color features [37]. Auto-correlogram $\gamma_{c_m}^k(I)$ in image I shows the probability that pixel $p_{c_m}^j$ at distance k from pixel $p_{c_m}^i$ has the same color as c_m. This information, which is a concatenation of color information and spatial information in the image, can be calculated for different distances.

$$\gamma_{c_m}^k(I) = P\left[\left|p_{c_m}^i - p_{c_m}^j\right|\right] = k \qquad p_{c_m}^j \text{and } p_{c_m}^i \in I \tag{1}$$

Gabor wavelet transform is used as a linear filter to analyze the image texture. This filter decomposes the image in different scales and angles. The ability to describe local frequencies and extract texture features based on energy distribution is attributed to this transformation. Each wavelet absorbs energy at a specific frequency and direction. A two-dimensional Gabor wavelet in a modulated Gaussian kernel form integrates with a sinusoidal function. This filter is represented by equation 2, which includes a real and an imaginary part.

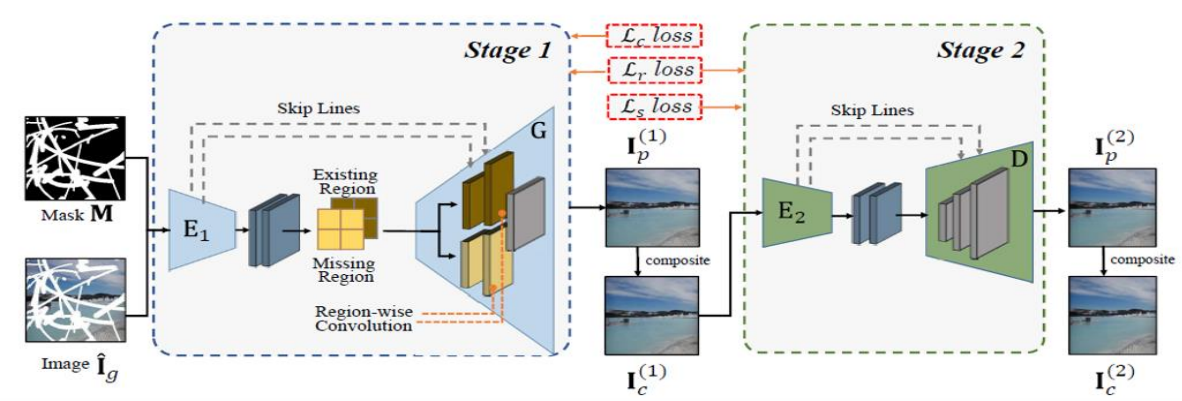

Figure 3. Encoder-decoder architecture for image reconstruction with missing regions [38]

$$\psi_{f,\theta}(x,y) = \exp\left[\left(-\frac{1}{2}\right)\left\{\left(\frac{x'^2}{\sigma_x^2}\right)+\left(\frac{y'^2}{\sigma_y^2}\right)\right\}\right] * \exp(2\pi f\theta_n) \quad (2)$$

$$x' = x\cos(\theta) + y\sin(\theta)$$

$$y' = y\cos(\theta) - x\sin(\theta)$$

$$\theta_n = \left(\frac{\pi}{p}\right) * (n-1);$$

$$n = 1,2,3,..p$$

In this equation, σ_x and σ_y are the standard deviation, f and θ specify the scale and direction, respectively.

- **Image reconstruction with a two-stage coarse-to-fine framework**

The architecture of the two-stage coarse-to-fine framework for reconstructing and inpainting the information on the missing regions of the image includes two stages [38]. In the first step, regional convolution layers performed to identify local features and divisions of different regions. In this way, instead of using the same filter, disparate regional convolution filters are used to deduce the semantic contents of the image from the existing regions, which in turn can help reconstruct and retrieve an overview of missing regions. Then, a non-local operation is employed to model the correlation among different regions globally. The second step is to study the visual compatibility between the missing and the existing regions. Finally, regional convolutions and non-local correlation are integrated with a coarse-to-fine framework so that the images appear semantically and visually real. Fig 3 shows an overview of this network that includes an encoder-decoder.

In this two-stage architecture, the input image I with a binary mask M integrates with the dot multiplication operator ⊙. The result of this operation is the image with missing regions $\hat{I}_g = I \odot M$. Image $\hat{I}_g$ is the input of encoder E_1 to retrieve semantic features using regional convolutions. A decoder, denoted by G, is also included in the first part of the image inpainting architecture to retrieve the semantic concepts of different regions and predict an overall estimate of the image. The estimated image from the first step is named $I_p^{(1)}$. The output of the first step is a composite image presented as equation 3:

$$I_c^{(1)} = \hat{I}_g + I_p^{(1)} + \odot(1-M) \quad (3)$$

The combined image of the first stage is given to the encoder of the second stage E_2. The output of the second stage the decoder, denoted by D, and named $I_p^{(2)}$ is also an estimated image. Finally, the reconstructed output image of the second stage, which is very similar to the original one presented in equation 4, is the result of two stages of the reconstruction process.

$$I_c^{(2)} = \hat{I}_g + I_p^{(2)} + \odot(1-M) \quad (4)$$

Since applying the identical convolution filters to the entire image not only makes retrieving semantic features challenging in different parts but also the result leads to visual artifacts such as variegation and opacity, regional convolutions are exploited. As a result, the decoder can separately retrieve the content of different regions utilizing different convolution filters. Equation 5 defines Regional convolutions in each position.

$$\acute{x} = \begin{cases} W^T x + b, & x \in X \odot M \\ \widehat{W}^T x + \hat{b}, & x \in X \odot (1-M) \end{cases} \quad (5)$$

In this equation, W and $\widehat{W}$ are the weights of convolution filters in terms of region for existing and lost area. b and $\hat{b}$ also refer to bias. x is the convolution window's feature that belongs to the entire X feature map. Therefore, different convolution filters, which are suitable for displaying and describing regional features, defined for the image different parts. The following loss functions are exploited to treat learning in the two-stage encoder-decoder architecture. They are defined in equations 6, 7, and 8, respectively. The reconstruction loss defined in equation 6 is implemented to compare the predicted images in two steps, including the available and missing regions, concerning the original image at the pixel level.

$$l_r = \left\| I_p^{(1)} - I_g \right\|_1 + \left\| I_p^{(2)} - I_g \right\|_1 \quad (6)$$

Correlation loss compares the relationship between local patches in maintaining semantic and visual consistency between missing and available regions. This loss function can help determine the expected non-local operation. For the combined image resulting from the first stage of reconstruction $I_c^{(1)}$ the correlation loss is defined based on $f_{ij}(.)$ as equation 7.

$$l_c = \sigma \sum_{i,j}^{n} \left\| f_{ij}(I_c^{(1)}) - f_{ij}(I_g) \right\|_1 \quad (7)$$

In this equation, σ represents the normalization coefficient based on position. Correlation loss makes the model create images with semantic details very close to the real one.

Although the correlation error retrieves more details, it still cannot prevent visual artifacts in unstable generated models. Therefore, to provide better results and to improve the images more perceptually, a style loss function is also considered. How to calculate the style loss function is given in equation 8. For this purpose, the integrated image feature map $I_c^{(2)}$ extracted from the pth layer of the pre-trained VGG-16 network is utilized. This feature map denoted by $\Phi_p(I_c^{(2)})$ shown in equation 6. Also, δ_p shows the normalization coefficient for the pth layer. Style loss focuses on the relationship between different channels for style transfer for the second-stage combined image.

$$l_s = \sigma \sum_{p} \delta_p \left\| (\Phi_p(I_c^{(2)})^T (\Phi_p(I_c^{(2)}) - (\Phi_p(I_g))^T (\Phi_p(I_g) \right\|_1 \quad (8)$$

Finally, the overall loss function is obtained by summing up the reconstruction loss, correlation and style presented in equation 9.

$$l = l_r + l_c + l_s \quad (9)$$

- **Fusioning of content and semantic features**

Information concatenation at the feature level brings about the integration of multi-source information that can lead to its greater exploitation. Low-level features provide information about the content and visual aspects of an image, such as color details and texture. In contrast, high-level features describe more complex concepts within the image, such as objects and scene understanding. Additionally, these features are robust to changes in lighting and image size. By combining both low-level and high-level features, the feature vector will contain more accurate information and a more complete description of the image. This information is effective in improving the accuracy of retrieving similar samples with the image and concept being searched by the user. The final feature vector which shows with F_v in equation 10 for each image is formed by concatenating semantic and deep features with content features including color and image texture extracted by resnet-50 deep neural network denoted by F_d.

$$F_v = \{ F_d, F_c, F_t \} \quad (10)$$

After concatenating the image semantic and content features and forming the feature vector, the similarity measurement between the query image and the dataset images is carried out by calculating the Euclidean distance. How to calculate the feature vector distance of the query image and dataset images is presented in equation 11. In this equation, p refers to the query image and q refers to an image from the dataset. After calculating the distance for all images in the dataset, the values are sorted in descending order. Finally, images with the highest degree of similarity to the user's query image are retrieved.

$$d(p,q) = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2} \quad (11)$$

III. EVALUATION OF THE RESULT

The proposed image retrieval' model concentrating on image reconstruction with missing regions and retrieving similar images evaluated in this section. Its performance was assessed on Paris 6K dataset [39]. This dataset contains 6412 images with dimensions of 1024x768 in JPEG format and 12 classes. Fig 4 shows sample images of the Paris 6K datasets.

Figure 4. Sample images of Paris 6K dataset

In the proposed method, the Auto-correlogram algorithm has been used to extract color features. These features can be computed in terms of color information and spatial information in the image at different intervals. The extracted features have been calculated at 5 intervals. The Gabor filter has also been used to extract texture features. This filter decomposes the image into different scales and orientations. This algorithm has been adjusted to extract texture features at 4 scales and 8 directions. In total, 384 low-level features have been extracted. High-level features have also been extracted from the second last fully connected layer in the deep neural network. The feature vector dimensions in VGG-16 and VGG-19 networks are both equal toq 4096. The feature vector dimensions for Resnet-50 and MobileNet are 2048 and 1024 respectively.

Binary masks are made using fringe pattern with different frequencies. The Fringe Pattern consists of a pattern of dark and light stripes. This pattern is created

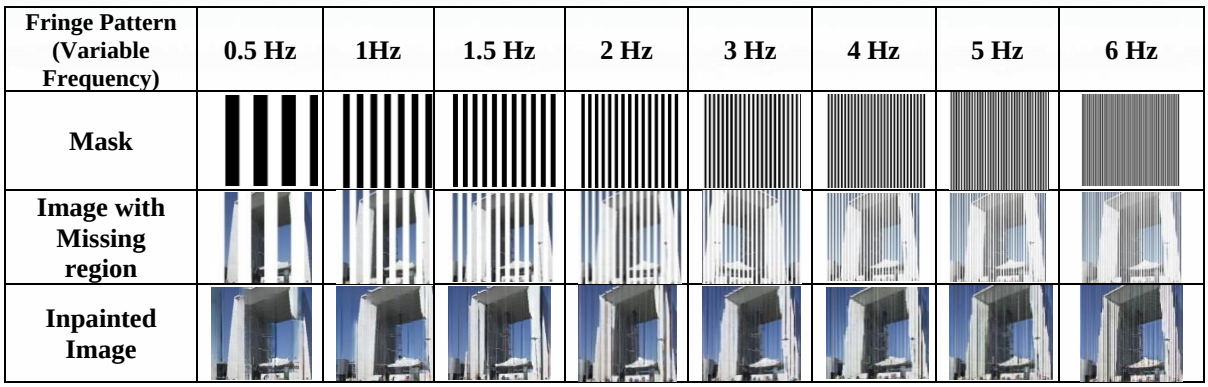

Figure 5. Samples of binary masks with variable frequency fringe pattern and image reconstruction results

when two or more waves interfere with each other. Therefore, this pattern can be used as a controlled mask to determine the degree of image degradation. In this mask, the dark stripes identify the missing areas of the image, while the light stripes identify the undamaged areas. Equation 12 presents the calculation method, where λ is the wavelength and D is the separation between the plane and the slots. d is the distance between the slots. The purpose of using this mask is to establish uniformity of destruction in the entire image using regular patterns of distributed destruction. Increasing the frequency in the fringe mask leads to decreasing the distance between the mask lines and a change in the image destruction model. An example of established masks and the results of image reconstruction with the adversarial generator network for a sample image from the dataset are shown in Fig 5.

$$\beta = (\lambda * D)/d \tag{12}$$

First, a binary mask is applied to an original image to produce a corrupt image for querying in the image retrieval model. Second, queried distorted image reconstructed with the help of an adversarial generative network meanwhile, its performance evaluated with image quality assessment criteria such as structural similarity index measure SSIM [40], Feature similarity image matrix FSIM [41], and peak signal-to-noise ratio PSNR. The SSIM criterion used to measure the similarity of two images based on the three components of Luminance, contrast and structure. The calculation method of the SSIM criterion is provided in equation 13, where x is the query image and y is the retrieved image. μ_x, and μ_y are the average, σ_x and σ_x are the variance, and σ_{xy} is the covariance of x, and y. C_1, C_2, *and* C_3 are numerical constants.

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \tag{13}$$

By calculating the Gradient and phase matching using the Fourier correlation in the image, the FSIM features are another evaluating feature of the two images' similarity. Using this criterion to estimate image quality makes it close to the way of image perception in the human visual system. (PC) emphasizes the characteristics of the image in the frequency domain, and (G) refers to the changes in the direction of image intensity or color. Equation 14 and 15 shows the calculation of (PC) and (G) . Equation 16 shows the calculation method of this criterion. In this equation α and β are used to adjust the importance of $C, and\ G$.

$$S_{pc} = \frac{2PC_xPC_y + T_1}{PC_x^2 + PC_y^2 + T_1} \tag{14}$$

$$S_G = \frac{2G_xG_y + T_2}{G_x^2 + G_y^2 + T_2} \tag{15}$$

$$FSIM(x,y) = [S_{pc}(x,y)]^{\alpha}.[S_G(x,y)]^{\beta} \tag{16}$$

The process of retrieving similar images from the dataset for the reconstructed image is performed by extracting and concatenating the image content and semantic features. The similarity measurement between the query image feature vector and the dataset image is fulfilled by calculating the Euclidean distance.

Table I: RESULTS OF QUALITATIVE EVALUATION OF IMAGE INPAINTING WITH ENCODER-DECODER ARCHITECTURE

Fringe Pattern (Variable Frequency)	0.5 Hz	1 Hz	1.5 Hz	2 Hz	3 Hz	4 Hz	5 Hz	6 Hz
SSIM	0.68±0.04	0.70±0.07	0.70±0.10	0.66±0.11	0.70±0.21	0.65±0.33	0.58±0.20	0.52±0.35
FSIM	0.72±0.04	0.75±0.04	0.78±0.06	0.71±0.21	0.77±0.31	0.71±0.21	0.63±0.47	0.61±0.26
PSNR	24.21±3.41	23.43 ±3.43	24.31 ±4.01	23.30 ±3.82	23.38±4.11	24.51±4.32	24.27±4.41	23.81±4.7

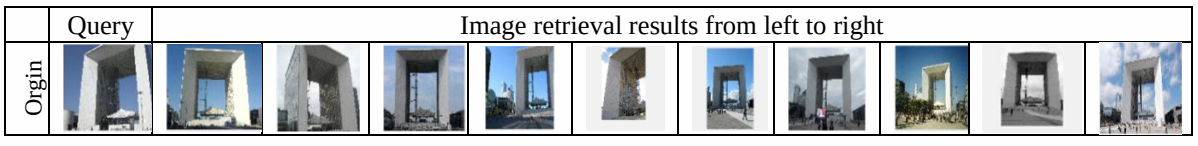

Figure 6. Sample of image retrieval results in the proposed system for the original, destructed, and reconstructed images

Table II: THE AVERAGE PRECISION RESULTS FOR RETRIEVING THE TOP 1, 5, AND 10 SAMPLES

Network Feature vector	Fringe pattern (Variable Frequency)	Average Precision (Top 1)	Average Precision (Top 5)	Average Precision (Top 10)
Handcraft+VGG-16 [15] (3844096)	6 Hz	55.83±1.1	43.49±2.1	36.31±2.6
	5 Hz	60.83±1.4	**56.21±1.9**	42.28±2.3
	4 Hz	**71.16±0.8**	66.47±1.2	60.34±1.9
	3 Hz	71.13±0.6	70.23±2.1	62.31±2.9
	2 Hz	80±1.3	71.11±1.4	65.81±1.8
	1.5 Hz	**76.66±1.6**	72.71±2.2	70.32±3.1
	1 Hz	82.51±1.7	72.32±2.4	68.2±2.7
	0.5 Hz	87.33±1.6	83.21±2.6	78.04±1.9
Handcraft+VGG-19 [15] (3844096)	6 Hz	56.21±1.7	45.32±2.3	34.11±2.5
	5 Hz	60.53±1.4	55.4±1.3	45.78±2.1
	4 Hz	65.23±1.0	64.47±1.5	61.3±2.3
	3 Hz	72.55±1.2	**72.36±2.4**	**64.87±3.1**
	2 Hz	78.6±1.9	**72.39±1.3**	63.01±3.3
	1.5 Hz	70.69±2.1	74.71±2.4	68.45±3.1
	1 Hz	80.43±2.2	75.2±2.0	71.23±2.7
	0.5 Hz	89.41±2.0	82.31±2.6	76.32±2.4
Handcraft+ResNet-50 [17] (384+2048)	6 Hz	**60.11±0.9**	**50.14±1.2**	**42.43±2.4**
	5 Hz	**62.63±1.1**	54.37±1.6	**48.71±2.1**
	4 Hz	70.55±1.3	**72.7±0.9**	**65.83±1.7**
	3 Hz	**75.65±1.0**	67.89±1.3	64.39±1.3
	2 Hz	**82.71±0.8**	69.54±1.4	**68.9±2.6**
	1.5 Hz	72.34±0.4	**79.52±1.7**	**71.44±2.8**
	1 Hz	**83.4±1.9**	**76.58±2.3**	**75.01±2.1**
	0.5 Hz	**92.7±2.3**	**85.7±1.8**	**79.31±3.0**
Handcraft +Mobilenet [20] (384+2048)	6 Hz	50.44±1.2	40.43±2.4	30.07±3.1
	5 Hz	55.89±1.4	52.81±1.7	32.44±3.4
	4 Hz	67.31±2.2	58.34±2.2	46.74±2.7
	3 Hz	64.63±0.7	63.91±2.0	50.37±2.2
	2 Hz	76.71±1.6	61.62±2.6	52.93±2.5
	1.5 Hz	66.32±2.3	69.03±1.7	59.65±2.8
	1 Hz	78.22±0.8	66.33±2.2	64.54±3.1
	0.5 Hz	86.21±1.7	72.65±2.3	68.61±2.6

Table I shows the qualitative results of image reconstruction with missing regions for 300 images from the dataset. The values of the image quality evaluation criteria in Table I represent the quality of the reconstructed image compared to the original one in the form of average along with the standard deviation as ±. The quality of reconstructed images compared to the reference image was evaluated using SSIM, FSIM, and PSNR criteria for various modes of image degradation with Fringe Pattern masks. Degradation was considered in 8 different modes for each image. By increasing the frequency in the Fringe Pattern, the distance between the lines in this mask decreases, and the degree of degradation increases. The results in the table show that the quality of the reconstructed image is affected by the degree of degradation. However, the quality of the reconstructed image for the image with the highest level of degradation (6Hz) is 0.52 and 0.61 according to the SSIM and FSIM criteria, respectively. Furthermore, the best results for image reconstruction according to the SSIM and FSIM criteria belong to the frequency of 3Hz, which are 0.7 and 0.77, respectively.

To study the image retrieval results after reconstructing and inpainting the image with the missing regions, the dataset evaluation processes with the average precision criterion for 1, 5 and 10 top retrieval samples. The calculation precision criterion is the ratio of retrieved correct samples to the sum of correct samples and retrieved incorrect samples in a query. The calculating method of the average precision is provided in equation 17, where TP refers to the number of retrieved correct samples and FP refers to the number of retrieved incorrect samples.

$$Percision = \frac{TP}{TP + FP} \quad (17)$$

Samples of the retrieval results of a query image are shown in Fig 6. At first, the query image, which has been distorted by applying a mask with an average amount, is reconstructed. It is then fed into the image retrieval system to query for similar samples. The retrieved samples for the image with missing regions in the second row of Fig6 show no relevant sample to the query image among the ten retrieved images. This number for the reconstructed image shown in the third row of Fig 6 includes eight similar images out of ten retrieved ones. Table II presents the results of the dataset evaluation with the average precision criterion for the 1, 5 and 10 top retrieval samples.

The feature vector extracted from the four pre-trained networks VGG-16, VGG-19, Resnet-50, MobileNet, and the combination of handcraft features are the basis of comparison. The results show that ResNet-50 deep neural network outperformed compared with other networks. The average precision for retrieving similar images after image reconstruction, with the highest image destructive frequency of 6HZ equals to 60.11%, 50.14%, and 42.43% for the first, the top five, and top 10 retrieved samples, respectively, meanwhile, the retrieval results after image reconstruction with the lowest destructive frequency of 0.5Hz equal to 92.7%, 85.7%, 79.31% for

the first, the top five, and the top ten retrieved samples, respectively.

IV. Conclution

In the present paper, a content-based image retrieval model was introduced for retrieving images with missing regions by the help of reconstructing and concatenating their content and semantic features. In this model, prior to retrieval, the image with the missing regions is reconstructed. Due to the enhancement of the extracted features from the reconstructed image, the retrieval results are improving efficiently. The image reconstruction stage fulfilled using an encoder-decoder framework with a coarse-to-fine architecture. The two-stage reconstructing the image missing regions retrieve images without losing their original information. A content-based image retrieval hybrid model is also exploited to retrieve the relevant images from the dataset. The results of concatenating the image content and semantic features form the feature vector that makes it possible to describe the image more precisely. In the proposed image retrieval model, a pre-trained network performed in both image reconstruction and retrieval stages. The effectiveness of the image retrieval system for images with missing areas is dependent on the quality of the image inpainting technique used to reconstruct the missing regions. If the image inpainting technique is unable to accurately reconstruct the missing areas, the features of the reconstructed image may not be accurate and may lead to incorrect image retrieval results. Additionally, the performance of the image retrieval may be affected by the size and location of the missing regions. If the missing regions are large or located in important areas of the image, the accuracy of the image retrieval may decrease. Overall, advanced image inpainting techniques and accurate feature extraction from the image can be effective in reducing the limitations of retrieving images with missing regions.The proposed model performance was investigated considering the changes in the destructive frequency in the query image, and the reconstructed retrieval results. The results show that retrieving images relevant to query one associated with the highest destructive frequency is 60.11%, 50.14%, and 42.43% for the first, the 5 top, and 10 top samples, respectively. Meanwhile, the image retrieval results after reconstruction with the lowest destructive frequency are 92.7%, 85.7%, and 79.31% for the first, the top five, and the top ten retrieval samples, respectively.

References

[1] H. Hu *et al.*, “Content-based gastric image retrieval using convolutional neural networks,” *Int. J. Imaging Syst. Technol.*, vol. 31, no. 1, pp. 439–449, Mar. 2021, doi: 10.1002/IMA.22470.

[2] O. Tursun, S. Denman, S. Sridharan, E. Goan, and C. Fookes, “An efficient framework for zero-shot sketch-based image retrieval,” *Pattern Recognit.*, vol. 126, p. 108528, Jun. 2022, doi: 10.1016/J.PATCOG.2022.108528.

[3] G. Zhao, M. Zhang, J. Liu, Y. Li, and J. R. Wen, “AP-GAN: Adversarial patch attack on content-based image retrieval systems,” *Geoinformatica*, 2020, doi: 10.1007/s10707-020-00418-7.

[4] N. Ali, K. B. Bajwa, R. Sablatnig, and Z. Mehmood, “Image retrieval by addition of spatial information based on histograms of triangular regions,” *Comput. Electr. Eng.*, vol. 54, pp. 539–550, Aug. 2016, doi: 10.1016/J.COMPELECENG.2016.04.002.

[5] Suryanto, D. H. Kim, H. K. Kim, and S. J. Ko, “Spatial color histogram based center voting method for subsequent object tracking and segmentation,” *Image Vis. Comput.*, vol. 29, no. 12, pp. 850–860, Nov. 2011, doi: 10.1016/J.IMAVIS.2011.09.008.

[6] G. H. Liu and J. Y. Yang, “Content-based image retrieval using color difference histogram,” *Pattern Recognit.*, vol. 46, no. 1, pp. 188–198, Jan. 2013, doi: 10.1016/J.PATCOG.2012.06.001.

[7] M. Salmi and B. Boucheham, “Gradual integration of local color information for image retrieval by content: Application to cell-CCV method,” *Proc. - 2016 Glob. Summit Comput. Inf. Technol. GSCIT 2016*, pp. 54–59, Jul. 2017, doi: 10.1109/GSCIT.2016.23.

[8] D. Srivastava, B. Rajitha, S. Agarwal, and S. Singh, “Pattern-based image retrieval using GLCM,” *Neural Comput. Appl. 2018 3215*, vol. 32, no. 15, pp. 10819–10832, Jul. 2018, doi: 10.1007/S00521-018-3611-1.

[9] M. Garg and G. Dhiman, “A novel content-based image retrieval approach for classification using GLCM features and texture fused LBP variants,” *Neural Comput. Appl. 2020 334*, vol. 33, no. 4, pp. 1311–1328, Jun. 2020, doi: 10.1007/S00521-020-05017-Z.

[10] J. Singh, A. Bajaj, A. Mittal, A. Khanna, and R. Karwayun, “Content Based Image Retrieval using Gabor Filters and Color Coherence Vector,” *Proc. 8th Int. Adv. Comput. Conf. IACC 2018*, pp. 290–295, Jul. 2018, doi: 10.1109/IADCC.2018.8692123.

[11] F. Taheri, K. Rahbar, and P. Salimi, “Effective features in content-based image retrieval from a combination of low-level features and deep Boltzmann machine,” *Multimed. Tools Appl. 2022*, pp. 1–24, Aug. 2022, doi: 10.1007/S11042-022-13670-W.

[12] S. R. Dubey, “A Decade Survey of Content Based Image Retrieval Using Deep Learning,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 5, pp. 2687–2704, May 2022, doi: 10.1109/TCSVT.2021.3080920.

[13] P. Desai, J. Pujari, C. Sujatha, A. Kamble, and A. Kambli, “Hybrid Approach for Content-Based Image Retrieval using VGG16 Layered Architecture and SVM: An Application of Deep Learning,” *SN Comput. Sci. 2021 23*, vol. 2, no. 3, pp. 1–9, Mar. 2021, doi: 10.1007/S42979-021-00529-4.

[14] M. D. Zeiler and R. Fergus, “Visualizing and Understanding Convolutional Networks arXiv:1311.2901v3 [cs.CV] 28 Nov 2013,” *Comput. Vision–ECCV 2014*, vol. 8689, no. PART 1, pp. 818–833, 2014, doi: 10.1007/978-3-319-10590-1_53.

[15]K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc.*, 2015.

[16] C. Szegedy *et al.*, “Going deeper with convolutions,” *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 07-12-June-2015, pp. 1–9, Oct. 2015, doi: 10.1109/CVPR.2015.7298594.

[17] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2016-December, pp. 770–778, Dec. 2016, doi: 10.1109/CVPR.2016.90.

[18] K. T. Ahmed, S. Jaffar, M. G. Hussain, S. Fareed, A. Mehmood, and G. S. Choi, “Maximum Response Deep Learning Using Markov, Retinal Primitive Patch Binding

with GoogLeNet VGG-19 for Large Image Retrieval," *IEEE Access*, vol. 9, pp. 41934–41957, 2021, doi: 10.1109/ACCESS.2021.3063545.
[19] B. Cao, A. Araujo, and J. Sim, "Unifying Deep Local and Global Features for Image Search," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 12365 LNCS, pp. 726–743, 2020, doi: 10.1007/978-3-030-58565-5_43/COVER.
[20] R. Yelchuri, J. K. Dash, P. Singh, A. Mahapatro, and S. Panigrahi, "Exploiting deep and hand-crafted features for texture image retrieval using class membership," *Pattern Recognit. Lett.*, vol. 160, pp. 163–171, Aug. 2022, doi: 10.1016/J.PATREC.2022.06.017.
[21] M. Alrahhal and Supreethi K.P., "Multimedia Image Retrieval System by Combining CNN With Handcraft Features in Three Different Similarity Measures," *Int. J. Comput. Vis. Image Process.*, vol. 10, no. 1, pp. 1–23, Jan. 2020, doi: 10.4018/IJCVIP.2020010101.
[22] M. P. Likowski, M. Smieja, L. Struski, and J. Tabor, "MisConv: Convolutional Neural Networks for Missing Data," *undefined*, pp. 2917–2926, 2022, doi: 10.1109/WACV51458.2022.00297.
[23] H. Khan, X. Wang, and H. Liu, "Handling missing data through deep convolutional neural network," *Inf. Sci. (Ny).*, vol. 595, pp. 278–293, May 2022, doi: 10.1016/J.INS.2022.02.051.
[24] S. A. Chavan and N. M. Choudhari, "Various approaches for video inpainting: A survey," *Proc. - 2019 5th Int. Conf. Comput. Commun. Control Autom. ICCUBEA 2019*, Sep. 2019, doi: 10.1109/ICCUBEA47591.2019.9129266.
[25] O. Elharrouss, N. Almaadeed, S. Al-Maadeed, and Y. Akbari, "Image Inpainting: A Review," *Neural Process. Lett. 2019 512*, vol. 51, no. 2, pp. 2007–2028, Dec. 2019, doi: 10.1007/S11063-019-10163-0.
[26] Y. Zhang, F. Ding, S. Kwong, and G. Zhu, "Feature pyramid network for diffusion-based image inpainting detection," *Inf. Sci. (Ny).*, vol. 572, pp. 29–42, Sep. 2021, doi: 10.1016/J.INS.2021.04.042.
[27] T. Xu, T. Z. Huang, L. J. Deng, X. Le Zhao, and J. F. Hu, "Exemplar-based image inpainting using adaptive two-stage structure-tensor based priority function and nonlocal filtering," *J. Vis. Commun. Image Represent.*, vol. 83, p. 103430, Feb. 2022, doi: 10.1016/J.JVCIR.2021.103430.
[28] B. Ceulemans, S. P. Lu, G. Lafruit, P. Schelkens, and A. Munteanu, "Efficient MRF-based disocclusion inpainting in multiview video," *Proc. - IEEE Int. Conf. Multimed. Expo*, vol. 2016-August, Aug. 2016, doi: 10.1109/ICME.2016.7553000.
[29] M. Shroff and M. S. R. Bombaywala, "A qualitative study of Exemplar based Image Inpainting," *SN Appl. Sci.*, vol. 1, no. 12, pp. 1–8, Dec. 2019, doi: 10.1007/S42452-019-1775-7/TABLES/1.
[30] H. Shao and Y. Wang, "Generative image inpainting with salient prior and relative total variation," *J. Vis. Commun. Image Represent.*, vol. 79, p. 103231, Aug. 2021, doi: 10.1016/J.JVCIR.2021.103231.
[31] X. Zhang *et al.*, "DE-GAN: Domain Embedded GAN for High Quality Face Image Inpainting," *Pattern Recognit.*, vol. 124, p. 108415, Apr. 2022, doi: 10.1016/J.PATCOG.2021.108415.
[32] M. A. Hedjazi and Y. Genc, "Efficient texture-aware multi-GAN for image inpainting," *Knowledge-Based Syst.*, vol. 217, p. 106789, Apr. 2021, doi: 10.1016/J.KNOSYS.2021.106789.
[33] X. Zhang *et al.*, "Face inpainting based on GAN by facial prediction and fusion as guidance information," *Appl. Soft Comput.*, vol. 111, p. 107626, Nov. 2021, doi: 10.1016/J.ASOC.2021.107626.
[34] Y. Zeng, Y. Gong, and J. Zhang, "Feature learning and patch matching for diverse image inpainting," *Pattern Recognit.*, vol. 119, p. 108036, Nov. 2021, doi: 10.1016/J.PATCOG.2021.108036.
[35] N. Farajzadeh and M. Hashemzadeh, "A deep neural network based framework for restoring the damaged persian pottery via digital inpainting," *J. Comput. Sci.*, vol. 56, p. 101486, Nov. 2021, doi: 10.1016/J.JOCS.2021.101486.
[36] L. Liu and Y. Liu, "Load image inpainting: An improved U-Net based load missing data recovery method," *Appl. Energy*, vol. 327, p. 119988, Dec. 2022, doi: 10.1016/J.APENERGY.2022.119988.
[37] H. H. Bu, N. C. Kim, K. W. Park, and S. H. Kim, "Content-based image retrieval using combined texture and color features based on multi-resolution multi-direction filtering and color autocorrelogram," *J. Ambient Intell. Humaniz. Comput. 2019*, pp. 1–9, Oct. 2019, doi: 10.1007/S12652-019-01466-0.
[38] Y. Ma, X. Liu, S. Bai, L. Wang, D. He, and A. Liu, "Coarse-to-fine image inpainting via region-wise convolutions and non-local correlation," *IJCAI Int. Jt. Conf. Artif. Intell.*, vol. 2019-August, pp. 3123–3129, 2019, doi: 10.24963/IJCAI.2019/433.
[39] N. Alajlan, M. S. Kamel, and G. H. Freeman, "Geometry-based image retrieval in binary image databases," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 6, pp. 1003–1013, Jun. 2008, doi: 10.1109/TPAMI.2008.37.
[40] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004, doi: 10.1109/TIP.2003.819861.
[41] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "FSIM: A feature similarity index for image quality assessment," *IEEE Trans. Image Process.*, vol. 20, no. 8, pp. 2378–2386, Aug. 2011, doi: 10.1109/TIP.2011.2109730.
[42] G. Jang, J. woo Lee, J. G. Lee, and Y. Liu, "Distributed fine-tuning of CNNs for image retrieval on multiple mobile devices," *Pervasive Mob. Comput.*, vol. 64, Apr. 2020, doi: 10.1016/J.PMCJ.2020.101134.

Fatemeh Taheri received the M.Sc. degree from the Amirkabir University of Technology, Tehran, Iran. She is currently Ph.D. candidate in Artificial Intelligence from the South Tehran Branch, Islamic Azad University. Her research interests include Image Retrieval, Image Processing and Computer Vision.

Kambiz Rahbar is an expert in the field of computer and artificial intelligence systems. Having earned his Ph.D. in this field, he now serves as an assistant professor at Islamic Azad University, South Tehran Branch where he shares his vast knowledge with students and colleagues alike. Dr. Rahbar's primary research interests lie in computer vision, and
neural networks

Ziaeddin Beheshtifard received his B.Sc., M.Sc. and Ph.D. degrees in computer engineering from University of Tehran, Tarbiat Modares University, and Azad University of Qazvin Branch, Iran, respectively. Currently, he is an assistant professor in computer engineering department of Azad University of South Tehran Branch, Iran. His research interest includes machine learning, computer vision and generative AI.