Matching Faces and Attributes Between the Artistic and the Real Domain: the PersonArt Approach

MARCELLA CORNIA, University of Modena and Reggio Emilia, Italy
MATTEO TOMEI, University of Modena and Reggio Emilia, Italy
LORENZO BARALDI, University of Modena and Reggio Emilia, Italy
RITA CUCCHIARA, University of Modena and Reggio Emilia, Italy

In this paper, we present an approach for retrieving similar faces between the artistic and the real domain. The application we refer to is an interactive exhibition inside a museum, in which a visitor can take a photo of himself and search for a lookalike in the collection of paintings. The task requires not only to identify faces but also to extract discriminative features from artistic and photo-realistic images, tackling a significant domain shift. Our method integrates feature extraction networks which account for the aesthetic similarity of two faces and their correspondences in terms of semantic attributes. Also, it addresses the domain shift between realistic images and paintings by translating photo-realistic images into the artistic domain. Noticeably, by exploiting the same technique, our model does not need to rely on annotated data in the artistic domain. Experimental results are conducted on different paired datasets to show the effectiveness of the proposed solution in terms of identity and attribute preservation. The approach is also evaluated on unpaired settings and in combination with an interactive relevance feedback strategy. Finally, we show how the proposed algorithm has been implemented in a real showcase at the Gallerie Estensi museum in Italy, with the participation of more than 1,100 visitors in just three days.

CCS Concepts: • **Computing methodologies** → **Visual content-based indexing and retrieval**; **Matching**; • **Applied computing** → *Fine arts*.

Additional Key Words and Phrases: face similarity, face retrieval, face recognition, cultural heritage.

ACM Reference Format:
Marcella Cornia, Matteo Tomei, Lorenzo Baraldi, and Rita Cucchiara. 2022. Matching Faces and Attributes Between the Artistic and the Real Domain: the PersonArt Approach. *ACM Trans. Multimedia Comput. Commun. Appl.* 18, 3, Article 77 (2022),

1 INTRODUCTION

Even though there isn't an artwork depicting yourself, there are probably some which contain faces that look just like you, as testified by the number of people who found their "Doppelgänger" in a painting while visiting a museum[1] [16]. Taking inspiration from these accidental discoveries, we develop an interactive multimedia solution which can retrieve similar faces in a collection of paintings given a query photo from a visitor. Once the visitor has arrived at the museum, we imagine a situation where he can take a photo of himself, and get back the name and the location of the painting where his Doppelgänger lies, as a possible starting point for his visit.

This setting requires to detect faces in both paintings and selfies taken by the visitors, and also to retrieve faces that look similar between the real domain (that of photographs) and the artistic one

[1]https://www.thoughtco.com/art-museum-doppelgangers-4154789

Authors' addresses: Marcella Cornia, marcella.cornia@unimore.it, University of Modena and Reggio Emilia, Modena, Italy; Matteo Tomei, matteo.tomei@unimore.it, University of Modena and Reggio Emilia, Modena, Italy; Lorenzo Baraldi, lorenzo.baraldi@unimore.it, University of Modena and Reggio Emilia, Modena, Italy; Rita Cucchiara, rita.cucchiara@unimore.it, University of Modena and Reggio Emilia, Modena, Italy.

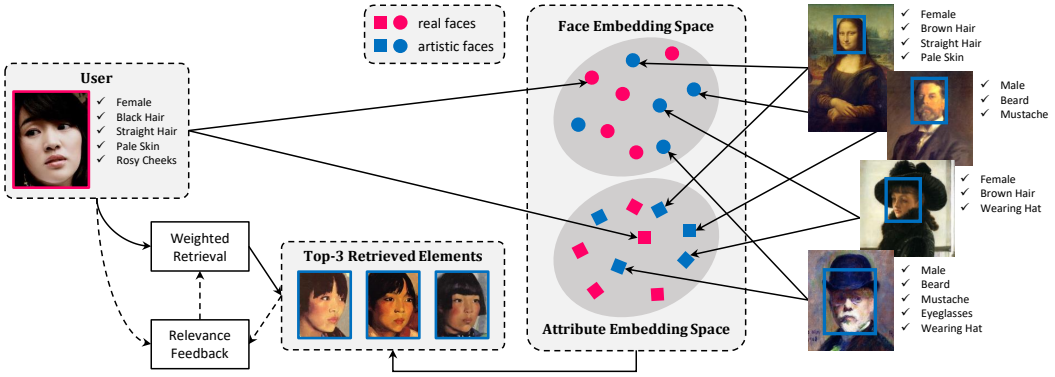


Fig. 1. Overview of our approach: given a real face as query, we retrieve similar faces from paintings, jointly taking into account the aesthetic similarity and the correspondence between semantic attributes. The user can further refine the retrieved set by providing feedbacks and imposing constraints on attributes.

(that of paintings). As images from these two domains can look very different in terms of low-level statistics, due to the presence of strokes and the peculiarities of the artistic style, the task demands for a domain adaptation stage, in which features extracted from the two domains can be merged and compared to get the final retrieval result. Additionally, as often happens when dealing with artistic data, there are no annotated datasets for the task, since no dataset contains photos of real people annotated with the most similar paintings in a given collection. To tackle these issues, we define a deep learning-based domain translation strategy, which lets us recover artistic proxies of real faces while still exploiting annotations coming from datasets with real faces. This allows us to exploit the supervision given by datasets originally designed for tasks similar to the one we are addressing (*e.g.* face recognition and attribute prediction) while working in a domain where no supervision is given. In this regard, this is the first work to propose a face retrieval system, trained with a shared embedding space and with no explicit supervision, for the artistic domain.

Beyond developing a retrieval algorithm, we also design of a multimedia system for face retrieval in the artistic domain with user interaction. In our system, we let the user customize the final result by interacting with the application. By a first mean, the user can provide a feedback on retrieved results by telling the application which faces look similar to him, and which are not satisfactory. This, in combination with a relevance feedback strategy, allows us to recover better and more subjective results at each iteration. Secondly, we also let the user impose constraints on the attributes of retrieved faces, *e.g.* requiring that retrieved results should be smiling, wearing a hat, having a big nose, and so on. This in turn permits a better exploration of the space of artworks. An overview of our approach is depicted in Figure 1.

The proposal is evaluated on both synthetic and realistic settings, on a variety of different datasets. In particular, we employ a style-transferred version of Celeb-A, which contains annotations for face and attribute recognition. We then evaluate the results on WebCaricature and IIIT-D Sketch, which respectively contain caricatures and sketches of different subjects, and collect two additional test sets, one from DeviantArt, and a second one containing artworks. We will publicly release both datasets to foster research in the field. Experimental evaluations will show the effectiveness of the presented architecture when employing different face embeddings, and the role of both face recognition and attribute detection in retrieval performance. We will also investigate the potential of the interaction with the user and the performances of the proposed relevance feedback strategy.

The proposed algorithm has also been tested in a real setting, by implementing a showcase at the Gallerie Estensi museum in Italy, which has seen the participation of more than 1,100 visitors in three days during the world-wide known Philosophy Festival. The remainder of this paper is organized as follows: in Sec. 2 we review the existing literature on the application of multimedia and computer vision techniques to arts with specific reference to face analysis; in Sec. 3 we present our architecture for weighted retrieval and its combination with relevance feedback. Further, in Sec. 4, 5 and 6 we describe our experimental setting and validate our proposal on different datasets. Finally, in Sec. 7 we present the showcase and the interaction with real users.

2 RELATED WORK

In this section, we review the literature related to the application of multimedia and computer vision techniques to the cultural heritage domain, mainly focusing on generative architectures. Moreover, we provide a brief overview of the most important advancements of face recognition methods and facial attribute classification solutions applied to both real and artistic images.

2.1 Computer vision for artistic content

In the last years, several efforts have been done to apply computer vision techniques to the cultural heritage domain, resulting in different works and applications ranging from generative models to classification and retrieval solutions. On the generative and synthesis side, up-and-coming results have been obtained by style transfer models, which aim to transfer the style of a painting to a real photo. Gatys *et al.* [24, 39] first proposed to encode style and content through pre-trained CNNs to generate a new image with a target content and a target texture. After this work, several methods have been proposed to improve different aspects of style transfer, including the reduction of the computational overhead [39, 46, 86], the improvement of the generation quality [25, 38, 66], and diversity [47, 87]. Other works have instead focused on the combination of different styles [13] and on the generalization to previously unseen styles and paintings [26, 48, 70]. Although all these methods have demonstrated to be effective in transferring artistic styles, they are usually not suitable for being inverted, *i.e.* for creating a realistic representation of a given painting. The latter task has been addressed in [45, 83–85, 101], by proposing generative architectures which can translate from the artistic to the real domain.

On the analysis and feature extraction side, several datasets containing artistic images have been collected and annotated to foster researches on style and genre recognition [40, 56, 62, 93], multi-task learning [76], zero-shot artwork instance recognition [19], and object and people detection [17, 92]. In this context, different traditional [4, 27] and deep learning-based [18, 28, 71, 92] solutions have been introduced to guarantee the application and generalization of visual correspondence and object detection methods to the artistic domain. Similar research efforts have also been made in the context of vision-and-language related tasks, with a particular focus on cross-modal retrieval, in which both supervised [7, 23, 75] and semi-supervised [5, 11, 15] approaches have been presented.

2.2 Generative architectures for content generation

Generative adversarial networks have been applied to several conditional image generation problems, ranging from image inpainting [61, 96] and super-resolution [44] to video prediction [58, 88] and text to image synthesis [63, 64]. In this context, a line of work on image-to-image translation has emerged, in both paired [37] and unpaired settings [49, 101].

Taigman *et al.* [81] tackled the image-to-image translation task by proposing a Domain Transfer Network based on a multiclass GAN loss, an f-constancy component, and a regularizing component. Zhu *et al.* [101] proposed the Cycle-GAN framework, which learns a translation between domains by exploiting a cycle-consistent constraint that guarantees the consistency of generated images

with respect to original ones. On a related line, Liu *et al.* [49] used a combination of generative adversarial networks and variational auto-encoders, while Ma *et al.* [54] focused on instance-level image translation by incorporating the attention mechanism into generative adversarial networks and allowing the incorporation of constraints at both instance and set-level.

A different line of work is multi-domain image-to-image translation [3, 14, 95]: here, the same model can be used for translating images according to multiple attributes (*i.e.* hair color, gender or age). For instance, Choi *et al.* [14] create a unified architecture that allows simultaneous image-to-image translation between multiple datasets and different domains. Other methods, instead, focus on diverse image-to-image translation, in which an image can be translated in multiple ways by encoding different style properties of the target distribution [34, 45, 102].

2.3 Face recognition

After the advent of deep learning, research on face recognition has mainly focused on the collection of training data to learn face representations [10, 30, 68, 82, 89] and on the design of specific architectures and training strategies to effectively address the task [50, 60, 68, 82, 94].

In this context, a considerable effort has been made to develop new loss functions for learning better and more powerful face representations. While first deep learning-based methods for face recognition adopted cross-entropy based softmax losses [79, 82], subsequent works started to explore discriminative loss functions to enhance generalization capabilities. Among them, contrastive and triplet loss functions became one of the leading choices to train face recognition networks. While contrastive loss requires face image pairs to minimize the distance of positive pairs and maximize that of negative pairs [77, 78], triplet loss considers face triplets and tries to minimize the distance between an anchor and a positive sample (*i.e.* same subject) while maximizing the distance between the anchor and a negative sample (*i.e.* different identity) [21, 60, 67, 68]. Other works proposed to use angular or cosine margin-based loss functions to make learned features potentially separable with a larger angular or cosine distance thus obtaining highly discriminative face features [20, 50, 51, 90].

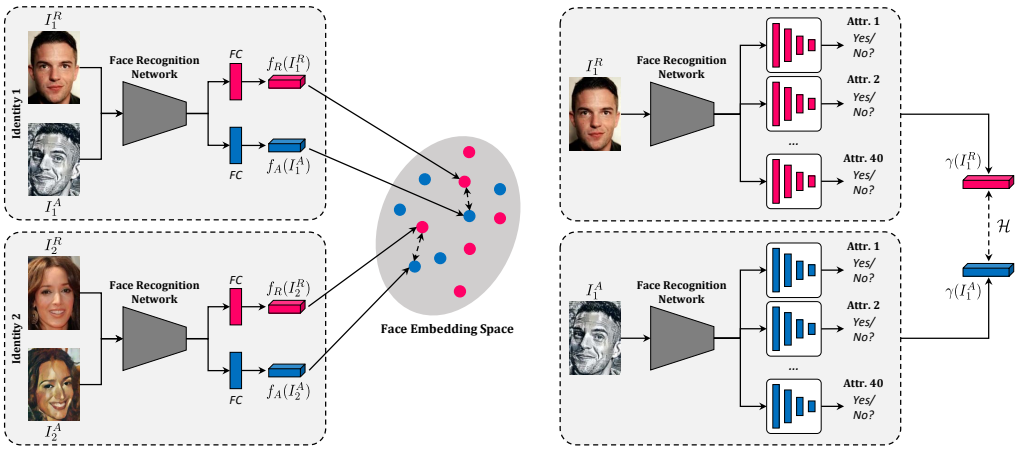
For a complete review of face recognition literature, we refer to recent surveys on this topic [57, 91] focusing on both networks and training strategies specifically designed for the task, along with the datasets used to train and evaluate face recognition architectures.

2.4 Facial attribute classification

Another important field of research on face analysis is focused on automatically recognizing facial attributes. Along with the introduction of different datasets for the task [52], several CNN-based architectures have been proposed [98]. One of the key components is the extraction of powerful and discriminative features to represent faces, either by leveraging on pre-trained networks [99] or by designing customized architectures [53]. While earlier methods combined facial representations with traditional classifiers [9, 43], current literature is focused on the design of deep attribute classifiers trained with specific loss functions [31, 65]. In this context, the most widely used loss function is the sigmoid cross-entropy loss, leading to a binary classification for each attribute [31]. Other works have instead investigated the use of the euclidean loss to train the network, thus considering the facial attribute classification as a regression problem [65]. As recently demonstrated, similar performance can be achieved by using both solutions with minimal changes in the architectural design [29].

2.5 Faces in artworks

While there is a vast corpus of work on face-related tasks on real images, only a few works have studied the extraction of facial features from artistic images. Crowley *et al.* [16] proposed to retrieve paintings using hand-crafted and pre-trained features either with a discriminative



(a) Learning from face identities (b) Learning from semantic attributes

Fig. 2. Overview of face recognition and attribute detection networks.

dimensionality reduction technique or with Exemplar SVMs [55]. Other works have focused on caricatures [1, 35, 36, 42] and cartoon recognition [59]. Among them, Hou *et al.* [35, 36] introduced a novel benchmark for caricature recognition and an approach based on landmark localization and representation. In this context, our work focuses on the domain shift between artistic and photo-realistic images and exploits the integration of face recognition and face attribute detection.

3 LEARNING CROSS-DOMAIN FACE RETRIEVAL

As it is rarely feasible to find and annotate a variety of Doppelgängers from paintings, almost no data exist for training a retrieval algorithm to find similar faces between the artistic and the real domain. While sources of supervision are available on photo-realistic images, our method needs to be aware of the domain shift between artistic and real images, which demands different feature extractors and a domain adaptation strategy.

To address these challenges, we define a translation approach that generates *artistic proxies* of real faces and lets us exploit the annotations of datasets containing real faces. Secondly, we address the lack of training data by exploiting proxy tasks for which training data is available and which share some of the underlying characteristics of our task. We employ face recognition (Figure 2a) and face attribute detection (Figure 2b), as they are closely related to the goal of finding similar faces and define a learned retrieval strategy which lets us recover faces across the two domains. Finally, we also discuss how our architecture can be combined with relevant feedback strategies to consider suggestions from the user during the retrieval phase. As an additional result, we provide explanation maps to highlight which regions of the retrieved results contribute most to the final similarity.

3.1 Addressing the domain shift

To be robust to the domain shift between artistic and real images, we define a way of converting real images to artistic proxies, which we will later employ as a data translation strategy to train the retrieval network. Here, we take inspiration from style transfer techniques, firstly proposed in [24]: given a realistic input face I^R and an artistic face I^A, we build an artistic version of I^R, called I^G, which preserves the content of I^R and the artistic style of I^A. The generated image can be built by

back-propagating directly on its pixel values while minimizing a loss function L that takes into account both the artistic style of I^A and the content of I^R. Formally, pixel values of the input real image are updated as

$$I^G_{i,j} \leftarrow I^G_{i,j} - \eta \frac{\partial L}{\partial I^G_{i,j}}, \tag{1}$$

where (i, j) represents the coordinates of a pixel on the input image, and η is the step size of Stochastic Gradient Descent. To encode the artistic style of I^A, we focus on textures and employ the Gram matrix obtained from the activations of a pre-trained CNN when applied on I^A. Given a layer l from the CNN, we define an ℓ_2 loss function between the Gram matrix of the generated image and that of the artistic image, as follows:

$$E_s = \frac{1}{4} \sum_{i,j} \left(\mathcal{G}^l(I^A)_{i,j} - \mathcal{G}^l(I^G)_{i,j} \right)^2, \tag{2}$$

where $\mathcal{G}^l$ denotes the Gram matrix obtained from activations at layer l, and $\mathcal{G}^l(\cdot)_{i,j}$ denotes its element in position (i, j). To encode the content of I^R, and to make sure that the same content is preserved in I^G, we combine the previously defined loss with a regularization term built on top of CNN activations. Given a layer l, we define an ℓ_2 loss function between raw activations as follows:

$$E_c = \frac{1}{2} \sum_{i,j} \left(\mathcal{F}^l(I^R)_{i,j} - \mathcal{F}^l(I^G)_{i,j} \right)^2, \tag{3}$$

where $\mathcal{F}^l$ denotes the activations at layer l. By combining the style loss with the reconstruction regularizer into the final loss (*i.e.* $L = E_s + \alpha E_c$), and exploiting appropriate pre-trained layers to encode style and content, we obtain a *style-transferred* version of I^R which matches the artistic style of I^A. To encode multi-level and multi-scale information, we also combine the activations extracted at multiple layers and define a separate loss function for each layer. We refer the reader to Figure 4 for a visualization of some qualitative results of the aforementioned translation approach and to Sec. 5 for implementation details.

3.2 Learning from face identities

Having obtained artistic proxies of real faces, we can train a similarity function which is aware of the domain gap between artistic and real faces. Firstly, we employ the preservation of face identities across domains as a proxy task. We employ a face recognition network $\phi(\cdot)$ to extract face-level features and a common embedding space, which is trained to recognize faces of the same identity across the two domains. An overview of the face recognition branch is shown in Figure 2a.

To project the representations of artistic and real images in a common semantic space, we perform a non-symmetric linear projection, followed by a ℓ_2-normalization step, so that the embedding space lies on the ℓ_2 unit ball. When projecting, we also employ a residual connection which was observed to slightly improve residual performance during our preliminary experiments. Formally:

$$f_a(I^A, \mathbf{w}_a, \mathbf{w}_\phi) = \ell_{2,norm}(\mathbf{w}_a^\intercal (1 + \phi(I^A, \mathbf{w}_\phi))) \tag{4}$$

$$f_r(I^R, \mathbf{w}_r, \mathbf{w}_\phi) = \ell_{2,norm}(\mathbf{w}_r^\intercal (1 + \phi(I^R, \mathbf{w}_\phi))), \tag{5}$$

where $\ell_{2,norm}$ is the ℓ_2 normalization function. Being D_ψ the output dimensionality of ϕ and D the dimensionality of the joint embedding space, $\mathbf{w}_r$ and $\mathbf{w}_a$ are $D_\psi \times D$ matrices which are in charge of storing different weights for the artistic and real projection branches. $\mathbf{w}_\phi$ indicates the weights of the face recognition network ϕ.

Artistic and real faces can be compared in the joint embedding space by computing the dot product (*i.e.* the cosine similarity) between their projections, so that the similarity between a (real)

query face I^R and an artistic face I^A becomes

$$s_i(I^R, I^A) = f_a(I^A) \cdot f_r(I^R), \tag{6}$$

where we drop the dependency on weights for brevity. The utility of the joint embedding space is maximized when it exhibits suitable cross-domain matching properties, *i.e.* when distances in the embedding space correspond to meaningful distances across the artistic and the real domain, and when corresponding pairs are matched in the embedding space. When this is verified to some extent, the embedding space acts as a bridge between the two domains and makes it possible to retrieve artistic faces given real faces as queries.

In order to learn an embedding space with such properties, we leverage face recognition datasets partially translated into the artistic domain using the strategy outlined in Sec. 3.1. We train the parameters of the model according to a Hinge triplet ranking loss with maximum violation [22] and margin α, defined as

$$\begin{aligned} \ell(I^R, I^A) = &\max_{\hat{I^A}} \left[\alpha - s(I^R, I^A) + s(I^R, \hat{I^A})\right]_+ + \\ &+ \max_{\hat{I^R}} \left[\alpha - s(I^R, I^A) + s(\hat{I^R}, I^A)\right]_+, \end{aligned} \tag{7}$$

where $[x]_+ = \max(0, x)$. In the equation above, (I^R, I^A) is a matching artistic-real pair of faces (such that the identity of the person in I^R is the same of that in I^A), while $\hat{I^R}$ is a negative real face with respect to I^A (such that the person in $\hat{I^R}$ is different from that of I^A), and $\hat{I^A}$ is a negative artistic face with respect to I^R (such that the person in $\hat{I^A}$ is different from that of I^R). The two terms contained in the loss require that the difference in similarity between the matching and the non-matching pair is higher than a margin α: in the first term, this is done by considering a real anchor and matching or non-matching artistic images; in the latter, instead, an artistic image is used as anchor.

3.3 Learning from semantic attributes

While preserving people identities across domains is a good proxy objective for retrieving similar faces, the network described so far might be unable to maintain face attributes (*i.e.* hair style, presence of glasses, age) in retrieved elements, as it focuses on face identification features rather than considering properties of the face that might change over time. For this reason, we complement the common embedding with attribute detection capabilities, to ensure that the correct attributes are maintained in retrieved faces.

To this aim, we employ two attribute detection networks, one for the artistic and one for the real domain. The structure of the attribute detection network we employ is shown in Figure 2b: we start from a face recognition model and then feed the activations of its last layer to n identical branches, each in charge of predicting the presence of an attribute. In our implementation, each branch is a composition of four fully connected layers, the last having an output size of two. Each branch is then trained, via a categorical cross-entropy loss, to predict the presence of the i-th attribute[2].

At test time, given predictions from the two attribute networks, we binarize them by tresholding, so to obtain a binary vector with length n for each image, and compute a similarity function by means of their Hamming distance as follows:

$$s_a(I^R, I^A) = 1 - \frac{\mathcal{H}(\gamma(I^R), \gamma(I^A))}{n}, \tag{8}$$

[2]We refer the reader to Sec. 5 for implementation details.

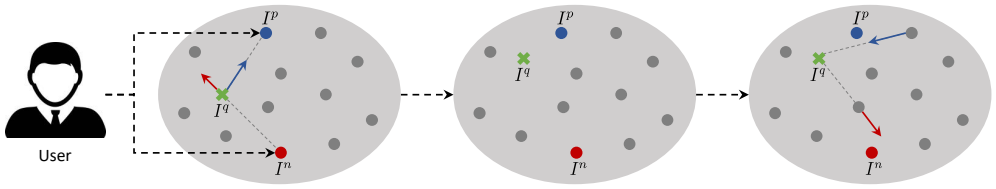


Fig. 3. Schema of the relevance feedback strategy. Given a query I^q, an item labeled by the user as positive I^p and one labeled as negative I^n, we firstly change the position of the query according to the user feedback. Then, we also change the position of the database items (gray dots) employing Feature Space Warping.

where $\gamma(\cdot)$ indicates the binarized prediction from the attribute prediction network, $\mathcal{H}$ the Hamming distance and n the number of attributes.

The final retrieval function is a weighted combination of the similarity computed through cross-domain face recognition and that computed by extracting semantic attributes:

$$s(I^R, I^A) = \frac{1 + s_i(I^R, I^A)}{2} + \lambda s_a(I^R, I^A), \tag{9}$$

where $s_i(I^R, I^A)$, resulting from a cosine similarity, is projected back to the same range of $s_a(I^R, I^A)$, *i.e.* $[0, 1]$. For a given λ, the resulting similarity score jointly takes into account the similarity of face traits and the similarity in attributes.

3.4 Interacting with user feedback

To foster the interactivity of the application and considering the subjectivity of the task, we also let the user give a feedback on retrieved results. This is, in turn, used to iteratively improve the quality of the retrieval following the preferences expressed by the user. In particular, the user is presented with the first k retrieved artworks and can refine the results by labelling a retrieved item either as positive (when it is more satisfactory than the others) or negative (if completely unsatisfactory).

As our retrieval depends on a face recognition and a face attribute branch, and given that the face attribute branch is easily controllable from the exterior (*e.g.* by asking the user which attributes he prefers to see in retrieved results), we here focus on the face recognition branch of our architecture. We employ two relevance feedback strategies to alter the embeddings of the query and database items. Given an item I^p labeled by the user as positive and an item I^n labelled as negative, we firstly change the position of the query I^q in the embedding space according to the user's feedback, as follows:

$$f_r(I^q) \leftarrow \alpha f_r(I^q) + \beta f_a(I^p) - \gamma f_a(I^n). \tag{10}$$

Then, we also change the position of the database items employing Feature Space Warping [8, 12]. This consists in moving all data points towards or away the new query vector $f_r(I^q)$ according to their similarities with the user's feedbacks. Formally,

$$\begin{aligned} f_a(I) \leftarrow f_a(I) &+ \eta e^{-c|f_a(I) - f_a(I^p)|}(f_r(I^q) - f_a(I)) \\ &- \eta e^{-c|f_a(I) - f_a(I^n)|}(f_r(I^q) - f_a(I)), \end{aligned} \tag{11}$$

where c is a bandwith parameter. The procedure above might be repeated at multiple feedback iterations, until the user is satisfied with at least one of the k retrieved results. A graphical overview of the approach is depicted in Fig. 3.

In practice, as it will be reported in Sec. 6, in most cases the initial retrieved set already contains sufficiently similar images and few feedback iterations are enough to retrieve a significant set of

similar faces. Nevertheless the adoption of a relevance feedback strategy is useful for corner cases in which retrieval fails and to accommodate user's preferences.

3.5 Providing explanations

Even in the case that retrieved results are satisfactory before the user provides a feedback or imposes any constraint on attributes, it is important to explain why a particular artistic face was selected and presented to the user. This is particularly significant considering that the user might struggle to perceive his own face, and thus the retrieved results, in an objective way.

Following recent works on Explainable AI [2] and techniques on prediction attribution [69, 72, 74, 80], for each retrieved item I^A we provide the user with a saliency map that visually indicates which regions of I^R have mostly contributed to the final similarity score $s(I^R, I^A)$, as an explanation of why the particular image was selected and presented. Ideally, a pixel in the saliency (or explanation) map $\mathcal{E}$ should be high if the corresponding pixel in I^R has contributed to increase the final similarity, lower otherwise.

To compute an explanation map $\mathcal{E}$, we follow the Integrated Gradients approach [80] by considering the straight-line path from a black baseline image to I^R and computing the gradients of the similarity score at all points along the path. Integrated gradients are obtained by accumulating these gradients. Specifically, a pixel (i, j) in $\mathcal{E}$ is computed as the path integral of the gradients along the straight-line path from 0 to $I^R(i, j)$, *i.e.*

$$\mathcal{E}_{i,j} = I^R_{i,j} \cdot \int_{\alpha=0}^{1} \frac{\partial s(\alpha I^R, I^A)}{\partial I^R_{i,j}} d\alpha. \tag{12}$$

Given the structure of our retrieval network, the computation of the partial derivative in Eq. 12 requires to back-propagate on both the face identity branch of the architecture and the face attribute branch of the network, thus providing an explanation which takes into account the dual nature of the similarity function.

4 DATASETS

To train and evaluate our model, we use – and in some cases extend – different datasets which contain real photos and portraits of people of known identities. To further validate our proposal, we also collect a large set of paintings containing people faces that can be used in real scenarios in which the corresponding portrait of a person is not known in advance.

Celeb-A [52]. This dataset contains more than 200, 000 face images belonging to 10, 177 different celebrities. Each image is annotated with 40 facial attributes ranging from the gender and hair style of the person to the presence of specific accessories such as eyeglasses, hat, or earrings. Since its size is sufficiently large to train a neural network, we employ this dataset to address the domain shift between artistic and real faces and to train the proposed architecture. We use the predefined training, validation, and testing splits, where identities do not overlap. To address the domain shift, we create a copy of this dataset where each image is converted into an artistic representation of itself using the style transfer strategy described in Sec. 3.1. For each image, we apply the style of a randomly selected painting from WikiArt[3] (*i.e.* different images of the same identity are style-transferred with a different painting) and use the annotations of identity and attributes to train the two branches of our architecture. Sample results of the style transfer performed on Celeb-A are shown in Figure 4.

WebCaricature [35, 36]. This dataset contains real photos and caricatures of 252 different subjects. For each person, the number of caricatures ranges from 1 to 114, while the number of real photos

[3]https://www.wikiart.org/

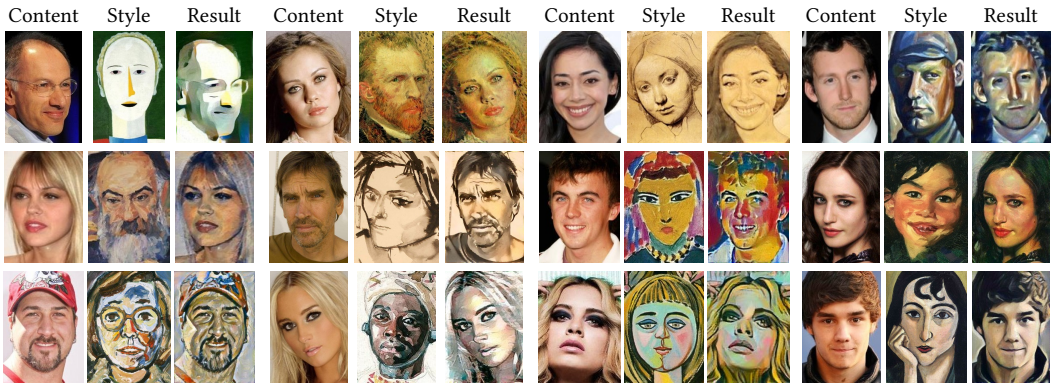


Fig. 4. Sample results of the style transfer strategy for addressing the domain shift between real and artistic faces. Real photos and paintings come from Celeb-A and WikiArt, respectively.

from 7 to 59. Overall, the dataset contains 6,042 caricatures and 5,974 photos. In our experiments, we use the entire set of caricatures as test set and the real photos as queries to test the effectiveness of our approach in the case of caricatures.

IIIT-D Sketch [6]. It comprises real photos and sketches of 238 different subjects. For each person, a single sketch, drawn by a professional sketch artist, is available. Also for this dataset, we use all 238 sketches as test set and the corresponding real photos as queries, thus testing our solution in the case of sketches.

DeviantArt Faces. This is a dataset that we collect and release as part of this work. It features portraits from the DeviantArt website[4], which contains art images produced by the users, divided in different categories (*e.g.* photography, traditional art, cartoons). On this website, many portraits of well known figures, especially celebrities, are available and can be used for this task. In general, images vary from oil or acrylic painted portraits to sketches, and lack of a photo-realistic quality.

To create the dataset, we start from the set of portraits described in [16] containing 1,088 images, one for each person, collected from DeviantArt. Since only the original image URLs are provided, we discard all identities for which the artistic image is no longer available obtaining a starting set of 617 portraits. We then extend this set of images by collecting new portraits of different people manually downloading them from this website. Overall, we obtain a set of 1,215 different identities for which a single portrait is collected. For the real counterpart, we download 5 different photos from Google Images for each identity that we manually validate to ensure the identity preservation and avoid errors. Also for this dataset, we use the entire set of portraits as test set and the real photos as queries.

Even though this dataset contains artworks instead of caricatures or sketches, it still cannot be used to fully test our target task, as it deals with a paired setting in which given a real image artistic portraits of the same person can be retrieved. Our task, instead, deals with an unpaired setting in which given a real face it is only possible to retrieve similar artistic faces.

WikiArt Faces. To qualitatively validate our solution in a real and unpaired scenario, we also collect a large set of paintings from WikiArt in which at least one face is present. To extract face bounding boxes, we use a deep learning-based face detector [97] that experimentally works well also on less realistic images. Then, we manually validate all detected faces, removing any false

[4]https://www.deviantart.com/

positive detection. Overall, we collect 15,116 artistic faces from paintings of different styles and periods. As previously mentioned, we release both DeviantArt Faces and WikiArt Faces.

5 IMPLEMENTATION DETAILS

Before presenting the experimental evaluations and their corresponding results, we here provide implementation and training details of our face retrieval approach.

Face detection. The first stage of our pipeline consists of the detection of the user face. Similarly, all database images are firstly offline fed through a face detector [97] that extracts a bounding box for each detected face. We apply the same face detector for both artistic and real images as we found face detections to be quite reliable on both domains. In our experiments, we keep all detected faces with a lower edge size of at least 100 pixels. To include the whole head, we extend face bounding boxes by a factor of 0.4.

Style-transferred art proxies. To generate art proxies from real faces, we employ a pretrained VGG-19 [73] and extract features from layers `conv1_1`, `conv2_1`, `conv3_1`, `conv4_1`, and `conv5_1` to encode the artistic style, and from layer `conv4_2` to encode the content of the real image. The relative weight α of content and style loss is set to 0.02. The generated face I^G is initialized with the realistic face I^R. During the update of I^G, all CNN parameters are kept fixed, and we backpropagate the loss only with respect to the pixels in I^G, using a L-BFGS optimizer [100] and clamping I^G between 0 and 1 at each iteration. We manually check the generated images to ensure that they have sufficient visual quality and repeat the generation with a different artistic image when needed. Noticing that bad visual quality is often associated with an high loss during the generation, we discard a sample and repeat the random choice until the loss is always lower than a threshold. In our experiments, we set this threshold to 40.0 as it generally corresponds to good generated results.

Face embeddings. To extract face feature vectors, we use and compare different face recognition networks, namely: SphereFace [50], VGG-Face [60], LightCNN [94], and VGG-Face-2 [10]. The feature embedding sizes respectively are of 512, 4096, 256, and 2048. For the Light-CNN network, we use its variants composed of 9 and 29 layers. For the VGG-Face-2, instead, we use two different versions of the model: one based on ResNet-50 [32], while the other based on SE-ResNet-50 with squeeze-and-excitation blocks [33]. We employ the implementations provided by the authors, when available.

Attribute prediction. We test the usage of two separate attribute prediction models, one for artistic images and one for real photos, and that of a shared one. In our experiments, we train the two separate models by using real photos and style-transferred images from Celeb-A, each of them annotated with binary attributes. When employing a shared prediction model, instead, we train it on both real and style-transferred images. The network structure consists of $n = 40$ identical branches, one for each facial attribute. Starting from the feature vector coming from a face recognition network, each branch is composed of four fully connected layers having output sizes of 1024, 512, 64, and 2. We use dropout after the first and second layer with a probability of 0.2 and 0.4, respectively. Also for this model, we extract face feature vectors from different face recognition models.

Training details. To train all our models, we employ the Adam optimizer [41]. For face retrieval, we use an initial learning rate of 10^{-5} decreased by a factor of 0.1 when recall metrics on validation set stop improving. When finetuning the whole face embedding model, we instead use a learning rate of 10^{-6}. The margin of the Hinge ranking loss is set to 0.1. For attribute prediction, we employ an initial learning rate of 10^{-4} decreased by a factor of 0.8 every 3 epochs. For all our experiments, we use a batch size of 32.

Table 1. Face recognition branch performance (Recall@k) on Celeb-A, WebCaricature, IIIT-D Sketch, and DeviantArt Faces.

			Celeb-A			WebCaricature			IIIT-D Sketch			DeviantArt Faces		
Face Embedding	**Emb. Size**	**Finetuning**	**R@1**	**R@5**	**R@10**	**R@1**	**R@5**	**R@10**	**R@1**	**R@5**	**R@10**	**R@1**	**R@5**	**R@10**
SphereFace	512	✗	30.2	44.0	50.4	22.9	40.9	49.3	29.9	57.3	67.5	3.0	7.1	10.4
		✓	**33.0**	**46.9**	**52.8**	**26.8**	**45.7**	**54.2**	**31.6**	**62.0**	**69.7**	**3.8**	**9.5**	**13.2**
LightCNN-9L	256	✗	69.2	80.7	83.8	69.0	84.7	88.8	**57.7**	**81.6**	89.7	19.3	31.9	38.7
		✓	**69.6**	**80.8**	**84.2**	**69.3**	**85.0**	**89.0**	**57.7**	81.2	**90.6**	**19.6**	**32.5**	**39.6**
VGG-Face	4096	✗	70.1	82.7	86.2	70.9	87.2	91.6	61.1	86.8	**91.5**	20.5	37.8	46.4
		✓	**77.7**	**88.2**	**90.9**	**75.2**	**89.7**	**93.0**	**61.5**	**87.2**	90.2	**26.5**	**45.3**	**54.3**
LightCNN-29L	256	✗	87.8	92.4	93.6	82.2	91.6	94.1	52.6	75.6	83.8	35.1	51.9	59.4
		✓	**88.6**	**93.2**	**94.1**	**82.5**	**92.5**	**94.8**	**55.1**	**78.6**	**86.3**	**36.3**	**55.1**	**62.9**
VGG-Face-2 (ResNet)	2048	✗	90.5	94.4	95.2	**90.1**	**96.4**	97.6	**66.7**	**88.5**	92.3	45.7	65.4	73.1
		✓	**90.9**	**94.6**	**95.3**	89.6	96.1	**97.7**	63.7	**88.5**	**93.2**	**45.9**	**65.6**	**73.4**
VGG-Face-2 (SE-ResNet)	2048	✗	91.1	94.6	95.3	**91.6**	**97.4**	**98.5**	**60.7**	**87.6**	92.7	47.6	67.0	**74.8**
		✓	**91.5**	**94.8**	**95.5**	91.3	97.2	98.2	58.1	86.3	**93.6**	**47.9**	**67.1**	**74.8**

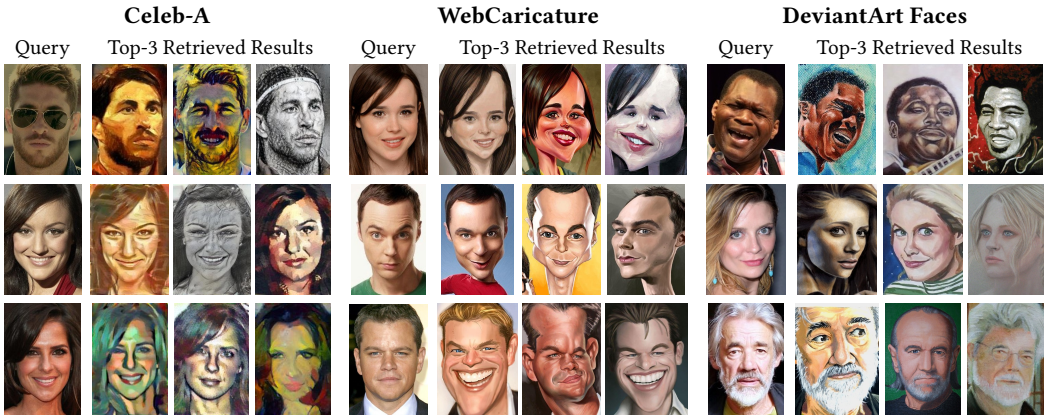


Fig. 5. Top retrieved results from Celeb-A, WebCaricature, and DeviantArt Faces from the face recognition branch ($\lambda = 0$).

Relevance feedback. The parameters of the relevance feedback algorithms are respectively set to: $\alpha = 0.8$, $\beta = 0.1$, $\gamma = 0.1$, $\eta = 0.5$, $c = 0.8$.

6 EXPERIMENTAL RESULTS

In the following, we first evaluate the performance of the two branches of our architecture in presence of the artistic domain gap, and then assess the quality of the retrieved results in the final application scenario, also in combination with the user's feedback.

6.1 Face retrieval

We evaluate the performance of the embedding space trained on the recognition of face identities, by ablating our approach and removing the face attribute networks. We use real faces as query and artistic or style-transferred images as database items. Table 1 shows the performances in terms of Recall@k (k=1, 5, 10) when finetuning the whole feature face recognition network and when training only the non-symmetric linear projections with residuals. For completeness, results are

Table 2. Facial attribute prediction accuracy on Celeb-A.

Model	Training Set	Accuracy (%)	
		Real photos	Style transfer
VGG-Face	Real	**88.9**	85.2
VGG-Face-2 (ResNet)		88.5	**86.3**
VGG-Face-2 (SE-ResNet)		87.7	85.8
VGG-Face	Style Transfer	**88.7**	**87.2**
VGG-Face-2 (ResNet)		87.9	87.0
VGG-Face-2 (SE-ResNet)		87.1	86.3
VGG-Face	Joint	**88.8**	**87.0**
VGG-Face-2 (ResNet)		88.4	87.0
VGG-Face-2 (SE-ResNet)		87.4	86.4

reported for all the face recognition backbones we employ and on Celeb-A, WebCaricature, IIIT-D Sketch, and DeviantArt Faces. For fairness, in the case of Celeb-A we remove the style-transferred version of the query from the set of retrievable elements.

We observe that our strategy for addressing the domain shift can produce significantly effective results on face recognition (with R@1 greater than 90% on some backbones), and further notice that fine-tuning the backbone is generally beneficial. Analyzing the different backbones, VGG-Face-2 provides the best performance, followed by LightCNN. SphereFace, instead, provides the worse performance on our setting, with a R@1 barely over 30%. As it can be seen, the impact of finetuning the backbone is particularly evident in the case of backbones which have low or medium effectiveness. When comparing the performance of the different datasets, instead, it is noticeable that recalls on IIIT-D Sketch and DeviantArt Faces are generally lower than those on Celeb-A and WebCaricature: this is explained by the smaller average number of relevant items for each query (*i.e.* only one for IIIT-D Sketch and DeviantArt Faces, around 20 for Celeb-A and WebCaricature).

Figure 5 qualitatively evaluates the performance of the branch by showing the top-3 retrieved results when using queries from Celeb-A, WebCaricature and DeviantArt Faces. As it can be seen, the branch is capable of preserving the identity of the query most of the times, although as expected it sometimes fails to preserve the facial attributes of the query.

6.2 Face retrieval using semantic attributes

In Table 2, we report the accuracy of the attribute prediction branch of our architecture on Celeb-A. We evaluate the performance of the two attribute prediction networks (one trained on real faces, one on style-transferred versions) on both real and artistic images, and test also the usage of a shared attribute prediction network for both real and artistic images. While VGG-Face tends to perform better in almost all settings, we notice that the obtained results are positive in terms of overall accuracy and always greater than 85% for both real and artistic settings and with all considered backbones. Further, we also notice that using a shared backbone does not decrease performances significantly.

Further, in Table 3 we employ our full architecture and evaluate the role of the two branches by varying the relative weight of the face recognition and the attribute prediction branches. Also in this case we test on Celeb-A, WebCaricature, IIIT-D Sketch, and DeviantArt Faces, and report Recall@1 for face recognition, as well as the attribute accuracy of the first retrieved element with respect to the query. As it can be seen, increasing λ up to 0.2 generally leads to boosting the attribute accuracy of the retrieved elements, without significantly loosing in terms of face recognition, thus

Table 3. Weighted retrieval results on Celeb-A, WebCaricature, IIIT-D Sketch, and DeviantArt Faces. Results are reported for different λ values in terms of recall and attribute preservation accuracy.

	$\lambda = 0.0$		$\lambda = 0.1$		$\lambda = 0.2$		$\lambda = 0.5$		$\lambda = 1.0$		$\lambda = 5.0$		$\lambda = 10.0$	
Face Embedding	**R@1**	**Acc.**	**R@1**	**Acc.**	**R@1**	**Acc.**	**R@1**	**Acc.**	**R@1**	**Acc.**	**R@1**	**Acc.**	**R@1**	**Acc.**
	Celeb-A													
SphereFace	33.0	85.4	34.1	87.7	32.8	88.9	26.0	90.5	17.3	91.3	6.2	91.7	6.1	91.7
LightCNN-9L	69.6	86.5	70.0	87.4	69.5	88.1	63.6	89.3	52.3	90.4	13.5	91.8	9.1	91.9
VGG-Face	77.7	86.8	77.4	87.8	75.7	88.5	68.0	89.6	54.0	90.6	13.6	91.8	9.9	91.8
LightCNN-29	88.6	86.8	88.6	87.6	88.1	88.1	84.2	89.1	74.4	90.0	21.1	91.8	10.8	91.8
VGG-Face-2 (ResNet)	90.9	86.9	90.7	87.6	90.1	88.2	85.5	89.2	73.7	90.2	17.9	91.8	10.6	91.8
VGG-Face-2 (SE-ResNet)	91.5	86.7	91.2	87.6	90.5	88.2	85.7	89.3	73.8	90.3	17.7	91.9	10.7	91.9
	WebCaricature													
SphereFace	26.8	85.0	30.2	89.3	30.6	91.5	26.4	94.6	19.0	96.4	10.9	97.4	10.9	97.4
LightCNN-9L	69.3	86.8	69.9	88.5	69.9	89.7	64.8	92.1	53.0	94.3	17.7	97.4	15.1	97.4
VGG-Face	75.2	87.6	75.5	89.3	73.9	90.5	67.3	92.8	54.3	94.7	19.6	97.3	16.7	97.4
LightCNN-29	82.5	87.1	82.7	88.5	82.6	89.5	77.5	91.7	65.0	93.8	22.5	97.3	16.9	97.4
VGG-Face-2 (ResNet)	89.6	87.5	89.3	88.8	88.2	89.8	81.2	92.0	67.2	94.2	20.6	97.4	17.5	97.4
VGG-Face-2 (SE-ResNet)	91.3	87.5	90.7	88.9	89.8	89.9	83.7	92.1	69.0	94.2	21.5	97.3	17.8	97.4
	IIIT-D Sketch													
SphereFace	31.6	84.0	35.9	84.6	37.6	88.2	31.6	80.6	24.4	92.1	8.5	93.3	8.5	93.3
LightCNN-9L	57.7	84.0	62.0	85.8	62.4	87.0	56.0	88.7	38.9	90.8	11.5	93.2	9.0	93.3
VGG-Face	61.5	85.3	62.0	86.3	62.4	87.1	56.8	88.6	43.2	90.4	9.8	93.2	7.7	93.3
LightCNN-29	55.1	84.5	54.7	85.8	56.4	87.1	50.4	88.7	37.6	90.6	11.5	93.2	8.5	93.3
VGG-Face-2 (ResNet)	63.7	85.0	64.1	86.0	63.2	87.0	52.6	89.2	35.0	91.4	9.4	93.3	8.1	93.3
VGG-Face-2 (SE-ResNet)	58.1	84.9	61.5	86.2	59.8	87.0	51.7	89.0	36.3	91.3	9.8	93.3	8.5	93.3
	DeviantArt Faces													
SphereFace	3.8	84.1	4.8	88.6	5.0	91.0	4.0	94.3	2.7	95.8	1.6	96.5	1.6	96.5
LightCNN-9L	19.6	85.1	20.3	87.6	20.2	89.0	17.4	91.9	11.6	94.2	2.9	96.5	2.4	96.5
VGG-Face	26.5	86.5	26.3	88.2	25.3	89.6	20.2	92.2	13.4	94.3	3.4	96.5	2.7	96.5
LightCNN-29	36.3	85.7	36.6	87.3	35.8	88.5	30.2	91.2	20.0	93.6	4.2	96.5	2.8	96.5
VGG-Face-2 (ResNet)	45.9	86.0	45.6	87.5	43.6	88.8	34.8	91.5	21.5	94.2	3.7	96.5	2.9	96.5
VGG-Face-2 (SE-ResNet)	47.9	86.0	47.3	87.5	45.1	88.7	35.2	91.4	21.8	94.0	3.9	96.5	3.0	96.5

obtaining better quality results in terms of the final similarity and validating the appropriateness of combining face recognition and attribute prediction.

To validate the weighted retrieval approach on our target setting (*i.e.* that of a real museum), in Figure 6 we use queries from the test set of Celeb-A and retrieve artistic faces coming from WikiArt. As it can be noticed, $\lambda = 0.2$ and $\lambda = 0.5$ offer the best result in terms of attribute preservation without significant loss in terms of the overall face similarity. Retrieved paintings generally look similar to the corresponding real queries and tend to preserve most of the facial attributes, in addition to face appearance.

Additionally, we quantitatively evaluate how attributes are preserved in retrieved results. In Table 4, we retrieve elements from WikiArt Faces using our weighted retrieval and queries from the test set of Celeb-A, and measure the attribute accuracy between the query and the top-1 retrieved element. As it can be seen, increasing λ effectively increases the preservation of attributes also in the case of real paintings.

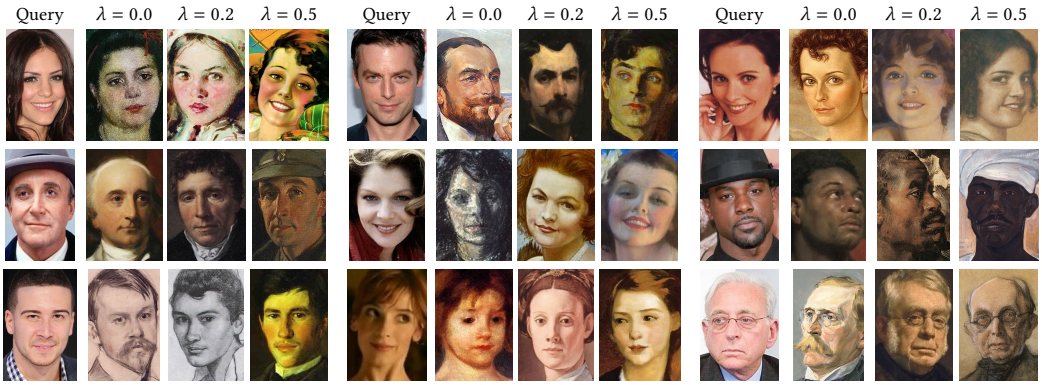


Fig. 6. Top-1 retrieved element from WikiArt Faces using weighted retrieval and different values of λ.

Table 4. Top-1 attribute accuracy on WikiArt Faces using different values of λ.

	Accuracy for different λ values						
Face Embedding	**0.0**	**0.1**	**0.2**	**0.5**	**1.0**	**5.0**	**10.0**
SphereFace	83.3	87.3	89.9	94.0	95.9	96.7	96.7
LightCNN-9	83.1	85.8	87.8	91.5	94.3	96.7	96.7
VGG-Face	83.7	86.4	88.4	92.1	94.5	96.6	96.7
LightCNN-29	83.1	85.4	87.3	91.2	94.0	96.6	96.7
VGG-Face-2 (ResNet)	83.1	85.5	87.6	92.0	94.7	96.7	96.7
VGG-Face-2 (SE-ResNet)	83.1	85.3	87.2	91.0	93.7	96.7	96.7

6.3 Relevance feedback and user interaction

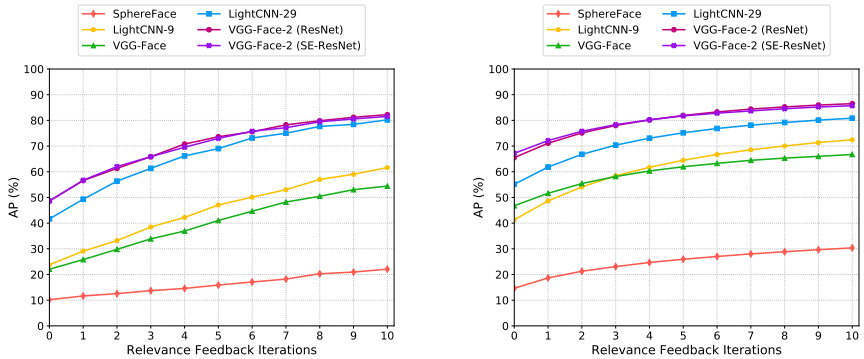
Experimental results provided so far have shown that, when a good face recognition backbone is chosen, the first retrieved element is generally similar enough to the query. However, as the user is presented with the top-k retrieved elements in our application scenario, we also need to go beyond the evaluation of the top-1 and assess the quality of the complete ranking. For this reason, we evaluate the Average Precision (AP) of the predicted ranking, and how it changes when the user interacts with the relevance feedback algorithm.

In the following experiments, we use a subset of the Celeb-A test set composed of a single randomly selected image for each identity (*i.e.* 989 query images) and all style-transferred images of the test set as retrievable items (*i.e.* around 18,000 images). For WebCaricature, we instead perform the experiments on the whole dataset. To simulate user interaction, at each iteration we provide the relevance feedback algorithm with one positive item (randomly selected among images which share the same identity with the query) and one negative item (randomly selected among images which do not share the same identity with the query).

Firstly, we evaluate the proposed relevance feedback model through an ablation study. In Table 5 we report the Average Precision (AP) of the predicted ranking after 1, 5 and 10 relevance feedback iterations, when employing three different variants of the proposed strategy: (1) Query Point Movement (QPM), where at each iteration a new query is reformulated as the mean of positive feedbacks, that corresponds to $\alpha = 0$, $\beta = 1.0$, $\gamma = 0$, $\eta = 0$; (2) Feature Space Warping (FSW) without changing the query point, corresponding to $\alpha = 1.0$, $\beta = 0$, $\gamma = 0$; (3) the full proposal, with both Query Point Movement and Feature Space Warping. The experiments are conducted on both

Table 5. Ablation study on relevance feedback hyper-parameters.

	Parameters					VGG-Face-2 (ResNet)			VGG-Face-2 (SE-ResNet)		
	α	β	γ	η	c	1	5	10	1	5	10
Celeb-A											
QPM	0.0	1.0	0.0	0.0	0.0	56.41	56.20	55.98	59.68	59.64	60.15
FSW	1.0	0.0	0.0	0.5	0.8	56.23	62.74	67.24	**60.02**	64.23	68.32
Ours	0.8	0.1	0.1	0.5	0.8	**56.56**	**73.63**	**82.23**	56.68	**73.00**	**81.53**
WebCaricature											
QPM	0.0	1.0	0.0	0.0	0.0	47.07	46.60	46.79	50.17	50.36	50.72
FSW	1.0	0.0	0.0	0.5	0.8	58.94	62.15	65.67	61.49	64.34	67.43
Ours	0.8	0.1	0.1	0.5	0.8	**71.10**	**81.93**	**86.51**	**72.13**	**81.73**	**85.71**



(a) Celeb-A (b) WebCaricature

Fig. 7. Relevance feedback evaluation with different backbones on Celeb-A and WebCaricature ($\lambda = 0.2$).

Celeb-A and WebCaricature datasets and by using the two versions of VGG-Face-2 (*i.e.* ResNet and SE-ResNet). As it can be seen, our complete strategy leads to the best performance according to both datasets and both considered backbones, with an overall AP over 80% after 10 relevance feedback iterations. On the contrary, when employing only one of the relevance feedback components, the final AP is generally lower than 60% and 70% for the QPM and FSW models respectively, confirming the effectiveness of our proposal.

Then, in Figure 7, we show the AP of different backbones on Celeb-A and WebCaricature, after a variable number of relevance feedback iterations. As it can be seen, the two versions of VGG-Face-2 have similar performances and obtain the best results on both datasets, with an initial AP of 48.5% (ResNet) and 48.6% (SE-ResNet) on Celeb-A, and 65.5% (ResNet) and 67.2% (SE-ResNet) on WebCaricature. In both cases, the initial AP is further increased by 15 points after four relevance feedback iterations (*i.e.* after the user has given four positive and four negative feedbacks), reaching more than 70% and 80% on Celeb-A and WebCaricature respectively.

Additionally, in Figure 8 we show some qualitative examples of using attributes as a further constraint to retrieval. In this case, the user is presented with the list of all possible attributes and can constraint the retrieval so that the retrieved set contains samples with a given attribute.

Fig. 8. Weighted retrieval results with attribute constraints ($\lambda = 0.2$).

7 PERSONART

The proposed retrieval algorithm has been used to build an interactive application that has been placed in the Gallerie Estensi museum in Modena, Italy. Gallerie Estensi is an art gallery centered on the collection of the Este family and houses a range of artworks executed by both notable and local artists, mainly from the 17th century.

The interactive exhibition has been designed as the reproduction of a photo booth, in which the visitor could enter to take a photo of himself and search for his lookalike through an interactive application. To foster a natural interaction, we wall-mounted a small camera and a touchscreen inside the photo booth and designed an application to guide the visitor in the process of capturing the photo and navigating retrieved results. Once a visitor found a satisfactory painting, he could print the result together with a map with guidance to reach the painting. For this purpose, a printer was placed inside a hidden compartment of the booth and made accessible from the exterior through a slot in the plasterboard, so that the visitor could pick up the printed sheet after leaving the booth. Figure 9 shows pictures of the photo booth during a day with high turnout.

The application has been designed following a web-based approach and is run on a browser on the client side. On the server side, instead, a Django server is responsible for generating views and interacting with the retrieval algorithm, implemented in PyTorch. Since the retrieval phase requires an asyncronous interaction between client and server, we also exploited Celery as task queue manager. The flow of the application is visually summarized in Figure 10: after entering personal details and accepting the terms of service, the user can visualize a preview of the video stream coming from the camera and take multiple photos of himself, until reaching a satisfactory result. The retrieval algorithm is then called asynchronously and candidate results are returned. After eventually interacting to refine the result, the user is finally presented with a description of the selected painting and the paper with indications is sent to the printer.

From the collection hosted at Gallerie Estensi, containing 95 pictures, we extracted 226 faces after a manual verification step to remove false positives and occluded faces. Figure 11 presents sample retrieval results obtained during the interactive exhibition on the Gallerie Estensi collection. As can be seen, although the set of retrievable artistic faces was limited in size, the retrieval algorithm was able to find similar faces across the artistic and the real domain.

As mentioned, we also endowed the retrieval algorithm with an explanation strategy to provide the user with saliency maps that can help to justify retrieved results. Figure 12 reports six samples

Fig. 9. Images of PersonArt, an interactive exhibition at the Gallerie Estensi of Modena, Italy.

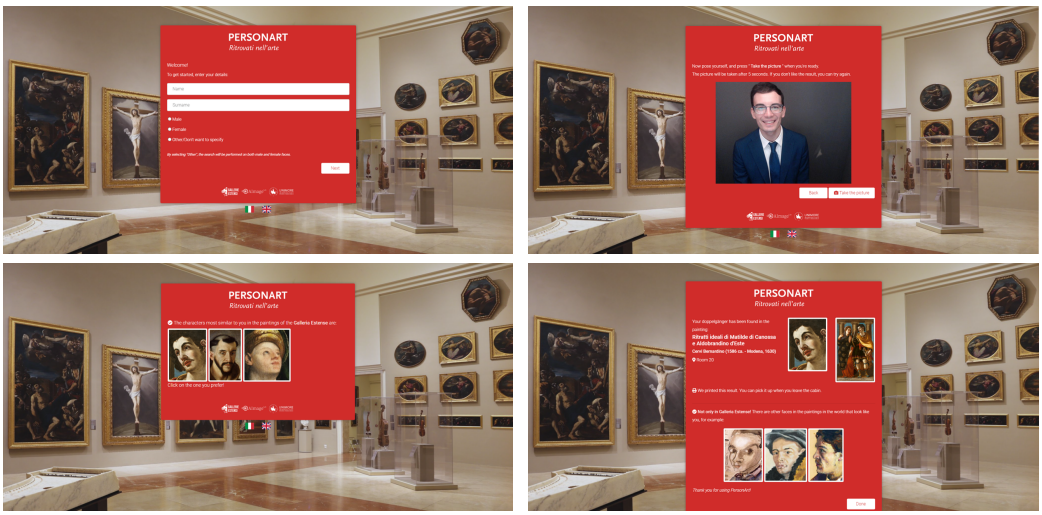


Fig. 10. Screenshots from the flow of the web-based application behind PersonArt.

of explanation maps. As it can be observed by comparing the predicted maps with the two matching faces, the strategy can provide a guide to understand which regions of the user's face contributed to the selection of a particular painting. For example, the middle face on the top row was paired with the reported painting because of the eyebrows, the nose shape, and the location of the corners of the mouth. Similarly, the painting on the right of the bottom row was selected because of similarities with the eyebrows and mustaches of the user, rather than for the beard, which is less bushy than that of the painting.

8 CONCLUSION

In this paper, we have presented a learned weighted retrieval strategy to find faces between the artistic and the real domain, which was also applied as an interactive demo in a museum environment. Our proposed architecture combines face identification and semantic attributes, in a weighted retrieval strategy. Further, the user can interact with the application by providing feedbacks on retrieved results, and by imposing constraints on the attribute space. Experiments,

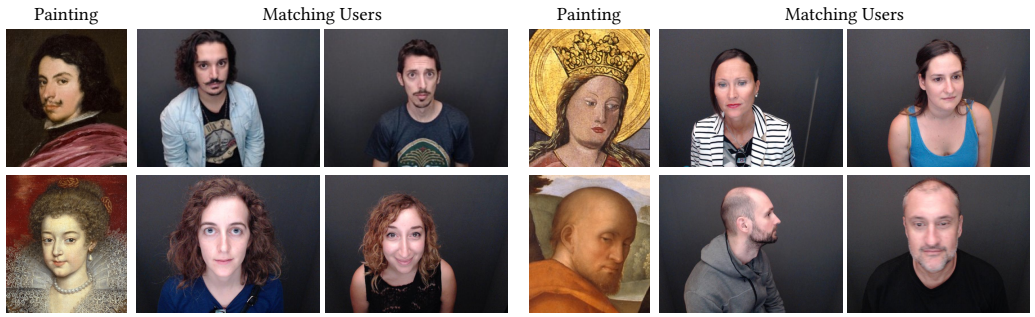


Fig. 11. Retrieval results from the Gallerie Estensi collection.

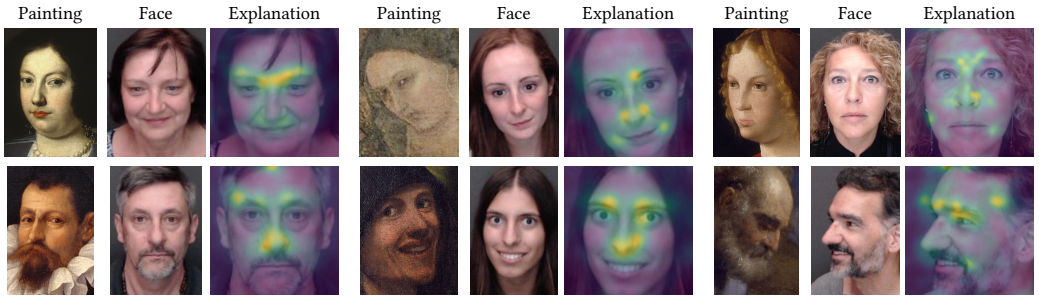


Fig. 12. Explanation examples using the Integrated Gradients approach on the Gallerie Estensi colleciton.

conducted on an augmented version of Celeb-A, WebCaricature, IIIT-D Sketch, and on two novel datasets we have collected, have demonstrated the effectiveness of the proposal in both synthetic and real settings.

ACKNOWLEDGMENTS

This work has been supported by "Fondazione di Modena" under the project "AI for Digital Humanities" and by the national project "IDEHA: Innovation for Data Elaboration in Heritage Areas" (PON ARS01_00421), cofunded by the Italian Ministry of University and Research. We also gratefully acknowledge "Ago – Modena Fabbriche Culturali", the Municipality of Modena, and Gallerie Estensi, in particular Dir. Martina Bagnoli and Arch. Silvia Gaiba for the design and creation of the PersonArt photo booth and for supporting and promoting the initiative.

REFERENCES

[1] Bahri Abaci and Tayfun Akgul. 2015. Matching caricatures to photographs. *Signal, Image and Video Processing* 9, 1 (2015), 295–303.

[2] Sule Anjomshoae, Amro Najjar, Davide Calvaresi, and Kary Främling. 2019. Explainable agents and robots: Results from a systematic literature review. In *Proceedings of the International Conference on Autonomous Agents and MultiAgent Systems*.

[3] Asha Anoosheh, Eirikur Agustsson, Radu Timofte, and Luc Van Gool. 2018. ComboGAN: Unrestrained scalability for image domain translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*.

[4] Mathieu Aubry, Bryan C Russell, and Josef Sivic. 2014. Painting-to-3D model alignment via discriminative visual elements. *ACM Transactions on Graphics* 33, 2 (2014), 14.

[5] Lorenzo Baraldi, Marcella Cornia, Costantino Grana, and Rita Cucchiara. 2018. Aligning text and document illustrations: towards visually explainable digital humanities. In *Proceedings of the International Conference on Pattern Recognition*.

[6] Himanshu S Bhatt, Samarth Bharadwaj, Richa Singh, and Mayank Vatsa. 2012. Memetically optimized MCWLD for matching sketches with digital face images. *IEEE Trans. on Information Forensics and Security* 7, 5 (2012), 1522–1535.

[7] Pietro Bongini, Federico Becattini, Andrew D Bagdanov, and Alberto Del Bimbo. 2020. Visual Question Answering for Cultural Heritage. 949, 1 (2020), 012074.

[8] Daniele Borghesani, Costantino Grana, and Rita Cucchiara. 2014. Miniature illustrations retrieval and innovative interaction for digital illuminated manuscripts. *Multimedia Systems* 20, 1 (2014), 65–79.

[9] Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. 2011. Describing people: A poselet-based approach to attribute classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

[10] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. 2018. VGGFace2: A Dataset for Recognising Faces Across Pose and Age. In *Proceedings of the IEEE International Conference on Automatic Face & Gesture Recognition*.

[11] Angelo Carraggi, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. 2018. Visual-semantic alignment across domains using a semi-supervised approach. In *Proceedings of the European Conference on Computer Vision Workshops*.

[12] Yao-Jen Chang, Keisuke Kamataki, and Tsuhan Chen. 2009. Mean shift feature space warping for relevance feedback. In *Proceedings of the IEEE International Conference on Image Processing*.

[13] Dongdong Chen, Lu Yuan, Jing Liao, Nenghai Yu, and Gang Hua. 2017. Stylebank: An explicit representation for neural image style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[14] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. 2018. StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[15] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, Massimiliano Corsini, and Rita Cucchiara. 2020. Explaining digital humanities by aligning images and textual descriptions. *Pattern Recognition Letters* 129 (2020), 166–172.

[16] Elliot J Crowley, Omkar M Parkhi, and Andrew Zisserman. 2015. Face Painting: Querying Art with Photos. In *Proceedings of the British Machine Vision Conference*.

[17] Elliot J Crowley and Andrew Zisserman. 2014. The State of the Art: Object Retrieval in Paintings using Discriminative Regions. In *Proceedings of the British Machine Vision Conference*.

[18] Elliot J Crowley and Andrew Zisserman. 2016. The art of detection. In *Proceedings of the European Conference on Computer Vision*.

[19] Riccardo Del Chiaro, Andrew D Bagdanov, and Alberto Del Bimbo. 2019. Webly-supervised zero-shot learning for artwork instance recognition. *Pattern Recognition Letters* 128 (2019), 420–426.

[20] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. 2019. ArcFace: Additive Angular Margin Loss for Deep Face Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[21] Changxing Ding and Dacheng Tao. 2015. Robust face recognition via multimodal deep face representation. *IEEE Transactions on Multimedia* 17, 11 (2015), 2049–2058.

[22] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. 2018. VSE++: Improving visual-semantic embeddings with hard negatives. In *Proceedings of the British Machine Vision Conference*.

[23] Noa Garcia and George Vogiatzis. 2018. How to Read Paintings: Semantic Art Understanding with Multi-Modal Retrieval. In *Proceedings of the European Conference on Computer Vision Workshops*.

[24] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. 2016. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[25] Leon A Gatys, Alexander S Ecker, Matthias Bethge, Aaron Hertzmann, and Eli Shechtman. 2017. Controlling perceptual factors in neural style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[26] Golnaz Ghiasi, Honglak Lee, Manjunath Kudlur, Vincent Dumoulin, and Jonathon Shlens. 2017. Exploring the structure of a real-time, arbitrary neural artistic stylization network. In *Proceedings of the British Machine Vision Conference*.

[27] Shiry Ginosar, Daniel Haas, Timothy Brown, and Jitendra Malik. 2014. Detecting people in cubist art. In *Proceedings of the European Conference on Computer Vision Workshops*.

[28] Nicolas Gonthier, Yann Gousseau, Said Ladjal, and Olivier Bonfait. 2018. Weakly Supervised Object Detection in Artworks. In *Proceedings of the European Conference on Computer Vision Workshops*.

[29] Manuel Günther, Andras Rozsa, and Terranee E Boult. 2017. AFFACT: Alignment-free facial attribute classification technique. In *Proceeding of the International Joint Conference on Biometrics*.

[30] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. 2016. MS-Celeb-1M: A dataset and benchmark for large-scale face recognition. In *Proceedings of the European Conference on Computer Vision*.

[31] Emily M Hand and Rama Chellappa. 2017. Attributes for improved attributes: A multi-task network utilizing implicit and explicit relationships for facial attribute classification. In *Proceedings of the AAAI Conference on Artificial*

Intelligence.

[32] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[33] Jie Hu, Li Shen, and Gang Sun. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[34] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. 2018. Multimodal Unsupervised Image-to-image Translation. In *Proceedings of the European Conference on Computer Vision*.

[35] Jing Huo, Yang Gao, Yinghuan Shi, and Hujun Yin. 2017. Variation Robust Cross-Modal Metric Learning for Caricature Recognition. In *Proceedings of the ACM International Conference on Multimedia Workshops*.

[36] Jing Huo, Wenbin Li, Yinghuan Shi, Yang Gao, and Hujun Yin. 2018. WebCaricature: a benchmark for caricature face recognition. In *Proceedings of the British Machine Vision Conference*.

[37] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. 2017. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[38] Yongcheng Jing, Yang Liu, Yezhou Yang, Zunlei Feng, Yizhou Yu, Dacheng Tao, and Mingli Song. 2018. Stroke controllable fast style transfer with adaptive receptive fields. In *Proceedings of the European Conference on Computer Vision*.

[39] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. 2016. Perceptual losses for real-time style transfer and super-resolution. In *Proceedings of the European Conference on Computer Vision*.

[40] Sergey Karayev, Matthew Trentacoste, Helen Han, Aseem Agarwala, Trevor Darrell, Aaron Hertzmann, and Holger Winnemoeller. 2014. Recognizing image style. In *Proceedings of the British Machine Vision Conference*.

[41] Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations*.

[42] Brendan F Klare, Serhat S Bucak, Anil K Jain, and Tayfun Akgul. 2012. Towards automated caricature recognition. In *Proceedings of the International Conference on Biometrics*.

[43] Neeraj Kumar, Alexander C Berg, Peter N Belhumeur, and Shree K Nayar. 2009. Attribute and simile classifiers for face verification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

[44] Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. 2017. Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[45] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Kumar Singh, and Ming-Hsuan Yang. 2018. Diverse Image-to-Image Translation via Disentangled Representations. In *Proceedings of the European Conference on Computer Vision*.

[46] Chuan Li and Michael Wand. 2016. Precomputed real-time texture synthesis with markovian generative adversarial networks. In *Proceedings of the European Conference on Computer Vision*.

[47] Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. 2017. Diversified texture synthesis with feed-forward networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[48] Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. 2017. Universal style transfer via feature transforms. In *Advances in Neural Information Processing Systems*.

[49] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. 2017. Unsupervised image-to-image translation networks. In *Advances in Neural Information Processing Systems*.

[50] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. 2017. Sphereface: Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[51] Weiyang Liu, Yandong Wen, Zhiding Yu, and Meng Yang. 2016. Large-margin softmax loss for convolutional neural networks.. In *Proceedings of the International Conference on Machine Learning*.

[52] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2015. Deep Learning Face Attributes in the Wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

[53] Yongxi Lu, Abhishek Kumar, Shuangfei Zhai, Yu Cheng, Tara Javidi, and Rogerio Feris. 2017. Fully-adaptive feature sharing in multi-task networks with applications in person attribute classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[54] Shuang Ma, Jianlong Fu, Chang Wen Chen, and Tao Mei. 2018. DA-GAN: Instance-level Image Translation by Deep Attention Generative Adversarial Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[55] Tomasz Malisiewicz, Abhinav Gupta, and Alexei A Efros. 2011. Ensemble of exemplar-SVMs for object detection and beyond. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

[56] Hui Mao, Ming Cheung, and James She. 2017. Deepart: Learning joint representations of visual arts. In *Proceedings of the ACM International Conference on Multimedia*.

[57] Iacopo Masi, Yue Wu, Tal Hassner, and Prem Natarajan. 2018. Deep Face Recognition: A Survey. In *Proceedings of the Conference on Graphics, Patterns and Images*.

[58] Michael Mathieu, Camille Couprie, and Yann LeCun. 2016. Deep multi-scale video prediction beyond mean square error. In *Proceedings of the International Conference on Learning Representations*.

[59] Ashutosh Mishra, Shyam Nandan Rai, Anand Mishra, and CV Jawahar. 2016. IIIT-CFW: a benchmark database of cartoon faces in the wild. In *Proceedings of the European Conference on Computer Vision Workshops*.

[60] Omkar M Parkhi, Andrea Vedaldi, and Andrew Zisserman. 2015. Deep Face Recognition. In *Proceedings of the British Machine Vision Conference*.

[61] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. 2016. Context encoders: Feature learning by inpainting. In *Proceedings of the European Conference on Computer Vision*.

[62] David Picard, Philippe-Henri Gosselin, and Marie-Claude Gaspard. 2015. Challenges in content-based image indexing of cultural heritage collections. *IEEE Signal Processing Magazine* 32, 4 (2015), 95–102.

[63] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. 2016. Generative adversarial text to image synthesis. In *Proceedings of the International Conference on Machine Learning*.

[64] Scott E Reed, Zeynep Akata, Santosh Mohan, Samuel Tenka, Bernt Schiele, and Honglak Lee. 2016. Learning what and where to draw. In *Advances in Neural Information Processing Systems*.

[65] Ethan M Rudd, Manuel Günther, and Terrance E Boult. 2016. MOON: A mixed objective optimization network for the recognition of facial attributes. In *Proceedings of the European Conference on Computer Vision*.

[66] Artsiom Sanakoyeu, Dmytro Kotovenko, Sabine Lang, and Björn Ommer. 2018. A Style-Aware Content Loss for Real-time HD Style Transfer. In *Proceedings of the European Conference on Computer Vision*.

[67] Swami Sankaranarayanan, Azadeh Alavi, Carlos D Castillo, and Rama Chellappa. 2016. Triplet probabilistic embedding for face verification and clustering. In *Proceedings of the International Conference on Biometrics Theory, Applications and Systems*.

[68] Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[69] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

[70] Falong Shen, Shuicheng Yan, and Gang Zeng. 2018. Neural Style Transfer via Meta Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[71] Xi Shen, Alexei A Efros, and Mathieu Aubry. 2019. Discovering visual patterns in art collections with spatially-consistent feature learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[72] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *Proceedings of the International Conference on Learning Representations Workshops*.

[73] Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In *Proceedings of the International Conference on Learning Representations*.

[74] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. 2015. Striving for simplicity: The all convolutional net. In *Proceedings of the International Conference on Learning Representations Workshops*.

[75] Matteo Stefanini, Marcella Cornia, Lorenzo Baraldi, Massimiliano Corsini, and Rita Cucchiara. 2019. Artpedia: A new visual-semantic dataset with visual and contextual sentences in the artistic domain. In *Proceedings of the International Conference on Image Analysis and Processing*.

[76] Gjorgji Strezoski and Marcel Worring. 2017. OmniArt: Multi-task Deep Learning for Artistic Data Analysis. *ACM Transactions on Multimedia Computing, Communications, and Applications* 14, 4 (2017), 88:1–88:21.

[77] Yi Sun, Yuheng Chen, Xiaogang Wang, and Xiaoou Tang. 2014. Deep learning face representation by joint identification-verification. In *Advances in Neural Information Processing Systems*.

[78] Yi Sun, Ding Liang, Xiaogang Wang, and Xiaoou Tang. 2015. DeepID3: Face recognition with very deep neural networks. *arXiv preprint arXiv:1502.00873* (2015).

[79] Yi Sun, Xiaogang Wang, and Xiaoou Tang. 2014. Deep learning face representation from predicting 10,000 classes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[80] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the International Conference on Machine Learning*.

[81] Yaniv Taigman, Adam Polyak, and Lior Wolf. 2017. Unsupervised cross-domain image generation. *Proceedings of the International Conference on Learning Representations*.

[82] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. 2014. DeepFace: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[83] Matteo Tomei, Lorenzo Baraldi, Marcella Cornia, and Rita Cucchiara. 2018. What was Monet seeing while painting? Translating artworks to photo-realistic images. In *Proceedings of the European Conference on Computer Vision Workshops*.

[84] Matteo Tomei, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. 2019. Art2Real: Unfolding the Reality of Artworks via Semantically-Aware Image-to-Image Translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[85] Matteo Tomei, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. 2019. Image-to-Image Translation to Unfold the Reality of Artworks: an Empirical Analysis. In *Proceedings of the International Conference on Image Analysis and Processing*.

[86] Dmitry Ulyanov, Vadim Lebedev, Andrea Vedaldi, and Victor S Lempitsky. 2016. Texture Networks: Feed-forward Synthesis of Textures and Stylized Images. In *Proceedings of the International Conference on Machine Learning*.

[87] Dmitry Ulyanov, Andrea Vedaldi, and Victor S Lempitsky. 2017. Improved Texture Networks: Maximizing Quality and Diversity in Feed-forward Stylization and Texture Synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[88] Jacob Walker, Kenneth Marino, Abhinav Gupta, and Martial Hebert. 2017. The pose knows: Video forecasting by generating pose futures. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

[89] Fei Wang, Liren Chen, Cheng Li, Shiyao Huang, Yanjie Chen, Chen Qian, and Chen Change Loy. 2018. The devil of face recognition is in the noise. In *Proceedings of the European Conference on Computer Vision*.

[90] Feng Wang, Jian Cheng, Weiyang Liu, and Haijun Liu. 2018. Additive margin softmax for face verification. *IEEE Signal Processing Letters* 25, 7 (2018), 926–930.

[91] Mei Wang and Weihong Deng. 2018. Deep Face Recognition: A Survey. *arXiv preprint arXiv:1804.06655* (2018).

[92] Nicholas Westlake, Hongping Cai, and Peter Hall. 2016. Detecting people in artwork with CNNs. In *Proceedings of the European Conference on Computer Vision Workshops*.

[93] Michael J Wilber, Chen Fang, Hailin Jin, Aaron Hertzmann, John Collomosse, and Serge Belongie. 2017. Bam! the behance artistic media dataset for recognition beyond photography. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

[94] Xiang Wu, Ran He, Zhenan Sun, and Tieniu Tan. 2018. A Light CNN for Deep Face Representation with Noisy Labels. *IEEE Transactions on Information Forensics and Security* 13, 11 (2018), 2884–2896.

[95] Xuewen Yang, Dongliang Xie, and Xin Wang. 2018. Crossing-Domain Generative Adversarial Networks for Unsupervised Multi-Domain Image-to-Image Translation. In *Proceedings of the ACM International Conference on Multimedia*.

[96] Raymond A Yeh, Chen Chen, Teck-Yian Lim, Alexander G Schwing, Mark Hasegawa-Johnson, and Minh N Do. 2017. Semantic Image Inpainting with Deep Generative Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[97] Shifeng Zhang, Xiangyu Zhu, Zhen Lei, Hailin Shi, Xiaobo Wang, and Stan Z Li. 2017. S3fd: Single shot scale-invariant face detector. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

[98] Xin Zheng, Yanqing Guo, Huaibo Huang, Yi Li, and Ran He. 2020. A Survey of Deep Facial Attribute Analysis. *International Journal of Computer Vision* (2020), 1–33.

[99] Yang Zhong, Josephine Sullivan, and Haibo Li. 2016. Face attribute prediction using off-the-shelf CNN features. In *Proceeding of the International Conference on Biometrics*.

[100] Ciyou Zhu, Richard H Byrd, Peihuang Lu, and Jorge Nocedal. 1997. Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization. *ACM Trans. Math. Software* 23, 4 (1997), 550–560.

[101] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

[102] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. 2017. Toward multimodal image-to-image translation. In *Advances in Neural Information Processing Systems*.