Neural Computing and Applications
https://doi.org/10.1007/s00521-023-08966-3

S.I.: VISUAL PATTERN RECOGNITION AND EXTRACTION FOR CULTURAL HERITAGE

Analyzing cultural relationships visual cues through deep learning models in a cross-dataset setting

Lorenzo Stacchio[1] · Alessia Angeli[2] · Giuseppe Lisanti[2] · Gustavo Marfia[3]

Received: 31 December 2022 / Accepted: 15 August 2023

Abstract

To study the evolution of specific cultures and times different kinds of pictures could be adopted. Family album photos may reveal socio-historical insights regarding those specific cultures and times. Along this path, this work addresses the problem of automatically dating an image by resorting to the analysis of an analog family album photo dataset. In particular, the IMAGO collection, which contains Italian photos shot in the 20th century, was considered. Thanks to the IMAGO dataset, it was possible to apply different deep learning-based architectures to date images belonging to photo albums without needing any other sources of information. In addition, we carried out cross-dataset experiments, which also involved models trained on American datasets, observing temporal shifts which may be due to known intercultural influences. We further explore such a possibility by qualitatively analyzing the cross-dataset interpretation of the trained deep-learning models with the Uniform Manifold Approximation and Projection (UMAP) algorithm. In conclusion, deep learning models revealed their potential in terms of possible applications to intercultural research, from different points of view.

Keywords Family album photos · Date estimation · Intercultural influences · Deep learning · Cross-detaset

1 Introduction

Vernacular photography [1], is an umbrella term to indicate pictures made by non-artists capturing everyday life and subjects for a huge range of purposes, including personal and commercial. Among the vernacular photographs, it is possible to define a sub-set considering those contained in family photo albums [2]. Researchers and scholars from different fields, along with public institutions, agree in identifying such collections as capable of capturing salient features regarding the evolution of local communities in space and time. Nevertheless, different contributions in this field often base their findings on the analysis of small corpora of photos [2, 3], since large-scale works are often impeded as they are too many to be processed manually. On one hand, multiple research initiatives have addressed the problem of processing and analyzing digital images. On the other hand, it is challenging to find initiatives focused on analog photos, principally due to the fact that printed images are (a) scattered in numerous public and private collections, (b) of variable quality, and (c) often worn out due to their prolonged use in time. In addition, any analysis employing image processing and computer vision algorithms requires the time-consuming and potentially degrading initial digitization step. Despite all these

✉ Gustavo Marfia
gustavo.marfia@unibo.it

Lorenzo Stacchio
lorenzo.stacchio2@unibo.it

Alessia Angeli
alessia.angeli2@unibo.it

Giuseppe Lisanti
giuseppe.lisanti@unibo.it

[1] Department for Life Quality Studies, University of Bologna, Bologna, Italy

[2] Department of Computer Science and Engineering, University of Bologna, Bologna, Italy

[3] Department for the Arts, University of Bologna, Bologna, Italy

Published online: 09 September 2023

complications, analog photographs continue to represent an unparalleled source of information regarding the recent past [1, 4]. The clothes that people wear, their haircut styles, the natural and urban landscape, etc., and, more in general, the overall environment, may exhibit the culture, and related socio-historical information, of a given time and place. In addition, all of these visual features may amount to important cues to estimate the shooting year of a given image (family album photo in this setting) [5]. Automatically estimating the date of a family photo album has important implications from the analysis of a cultural relationship: as aforementioned, this kind of picture captures the evolution of local communities which is bound by both space and time. Analyzing the time dimensions allows us to search for relationships in human habits among different places and bound possible intercultural influences through time itself. For example, by analyzing changes in fashion, technology, and other visual cues over time, we can gain insights into how cultural practices and social norms have evolved and also identify patterns to connect different communities and how they influenced each other over time. Of particular interest is having an automatic method, based on artificial intelligence, that could learn meaningful visual cues to automatically estimate the picture date could ease such kind of analysis, both from a quantitative and qualitative perspective [6, 7]. This method can be especially valuable when other sources of information, such as written records, could be scarce, hard to find, or unavailable.

This work addresses the problem of dating an image, focusing on the estimation of the shooting year. To do this, the IMAGO collection of family album photos, started in 2004 at the University of Bologna [2] was considered. Such a collection contains digitized versions of analog images with specific characteristics. In particular, each photograph portrays at least one person and has been shot in Italy (mostly in the Emilia-Romagna region) by Italian citizens. In particular, we here perform a dating analysis of the IMAGO collection [8], exploiting different deep learning-based architectures, without using any other source of information.

In [7] we performed an analysis by comparing different Convolutional Neural Network (CNN) architectures for the dating task comprising a multi-input architecture that combines different salient image regions. Moreover, we attempt to verify possible intercultural influences (i.e., the adoption of different customs and habits in different epochs and countries) by analyzing the differences in dating, resulting from a cross-dataset experiment, in which we employ the datasets from [9, 10]. In such work, we extend that contribution by: (i) motivating the importance of such analysis from a cross-cultural perspective; (ii) detailing the procedure we followed to obtain its major contributions in [7], including the cross-dataset experiment accuracies and the error distributions related to the people image crops; (iii) improving our analysis by integrating a qualitative cross-dataset visualization study exploiting the Uniform Manifold Approximation and Projection (UMAP) algorithm [11].

The rest of the paper is organized as follows. In Sect. 2, we review the state-of-the-art works that fall closest to this work. In Sect. 3, we report a description of the considered dataset, along with the pre-processing and splitting steps adopted. In Sects. 4 and 5, we present and validate the deep learning architectures trained on the IMAGO dataset and its human-related crops (IMAGO-FACES and IMAGO-PEOPLE). In Sect. 6, we report and discuss the cross-dataset experiments we performed, focusing on a socio-historical and intercultural influence perspective. Finally, in Sect. 7, we conclude this work with an overall discussion, along with possible future works.

2 Related work

So far only a few works have dealt with the dating of vernacular photographs, also considering analog ones [5, 9, 10, 12, 13]. Most of these works exploited different datasets to train state-of-the-art CNN neural architectures(e.g., ResNet50), which have been successfully applied in several contexts [14–16].

The authors of [9] employed a deep learning approach to date 37, 921 historical frontal-facing American high school yearbook photos taken from 1928 to 2010. They trained a CNN architecture [15, 16] to analyze people's face images and predict their shooting year. In addition, they observed a gender dependency in the performance of the implemented dating models. Again, the authors of [10] presented a dataset containing images of students taken from high school yearbooks, considering 1, 400 photos per year belonging to the 1950 to 2014 time span. They also resorted to CNNs to estimate the date of an image. In addition, they evaluated the quality of color vs. grayscale photos, considering different features: faces, torsos, i.e., people's upper bodies including faces, and random regions of images. They obtained the best performance with the torsos of people, and their results provide cues that human appearance is related to time. Instead, the authors of [13], implemented dating models analyzing images belonging to the years 1930 through 1999. They considered vernacular and landscape photos, including at most 25, 000 pictures per year. In addition, they proposed different baselines relying on CNNs, using both regression and classification approaches. Differently, the authors of [5] formulated the date estimation task as an image-retrieval one where, given a query, the retrieved images are ranked in terms of date

Table 1 Characteristics of existing datasets and IMAGO

Original dataset	Type(s) of photography	Type(s) of camera	Theme	Cardinality	Period
[9, Ginosar et al.]	Portrait	Digital and analog	Frontal face from High school yearbook	168,055	1905 – 2013
[10, Salem et al.]	Portrait	Digital and analog	High school yearbook	ca 600,000	1912 – 2014
[13, Müller et al.]	Vernacular and landscape	Digital and analog	No specific theme	1,029,710	1930 – 1999
IMAGO collection	**Vernacular**	**Analog**	**Family albums**	**ca 80,000**	**1845 – 2009**

The IMAGO collection, the dataset we introduce, is in bold

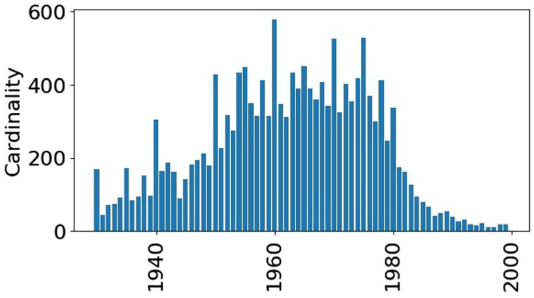


Fig. 1 IMAGO year classes distribution

similarity. In particular, they analyzed the same dataset employed in [13].

On one hand, the presented contributions focused on the dating of vernacular photographs shot in heterogeneous settings, e.g., landscapes and portraits. On the other hand, however, the IMAGO dataset [8] is only composed of family album photos shot during the twentieth century. In Table 1, we report the difference among the considered datasets, considering their main features. Although it may be possible to find scientific contributions which studied datasets comprising historical images, none of the considered ones exposed only family album photos [9, 10, 13]. In addition, to the best of our knowledge, no other works have also considered a cross-dataset and intercultural perspective when approaching the dating task.

3 Dataset, pre-processing and splitting

The IMAGO collection, and the related dataset,[1] were introduced in [8]. The IMAGO project was started in 2004 by socio-historical scholars to study the evolution of Social History through the lenses of family album photos. This produced a digitized collection (namely IMAGO) of analog family album photos gathered year by year and conserved by the Department of the Arts of the University of Bologna[2]. The collection comprises ca 80, 000 photos taken between 1845 and 2009, and belonging to approximately 1500 Italian family albums, offering the opportunity of studying the evolution of Italian society during the twentieth century. Among these, 16, 642 images have been labeled by the bachelor students in the Fashion Cultures and Practices course, under the supervision of the socio-historical faculty. The annotation process followed (and keeps following, as new photos are acquired and annotated every year) a simple but strict protocol [8], and generated two socio-historical metadata per each photo: the shooting year and the socio-historical context [2]. The process of annotation, which is still ongoing as new photos are obtained from incoming bachelor's students in Fashion Cultures and Practices and annotated each year involves several steps: (i) A first lecture is given where the socio-historical background, the IMAGO dataset construction project, and the various classification categories are introduced and explained (ii) A more detailed lecture delves into the annotation problem, with an emphasis on the importance of dependable and authentic sources of socio-historical materials, such as the year of shooting. This entails explaining that the original owner of the photo should be interviewed, whenever feasible. If the original owner is unavailable, such as for very old photos, then a second-hand informed party may be contacted, such as anyone who might be familiar with the context of the photo. Alternatively, if possible, an effort can be made to deduce the socio-historical context and the year of shooting by scrutinizing any written annotations inscribed behind the photo. If none of these options are feasible, then no annotation is added. (iii) The last lecture teaches the students how to label these images from a technical archival point of view. As a result, the data given by the photo's owner serves as the ground truth from a socio-historical perspective since only the owner (or a related individual like a friend or family member) possesses the information that could be exploited to label those images. Nevertheless, in this work we will focus on the image dating task, considering the 16, 642 labeled family album photos shot between 1845 and 2009.

In Fig. 1 is reported the number of labeled images available per year in the 1930 to 1999 time frame,

[1] The IMAGO dataset is available upon request.

[2] http://imago.unibo.it.

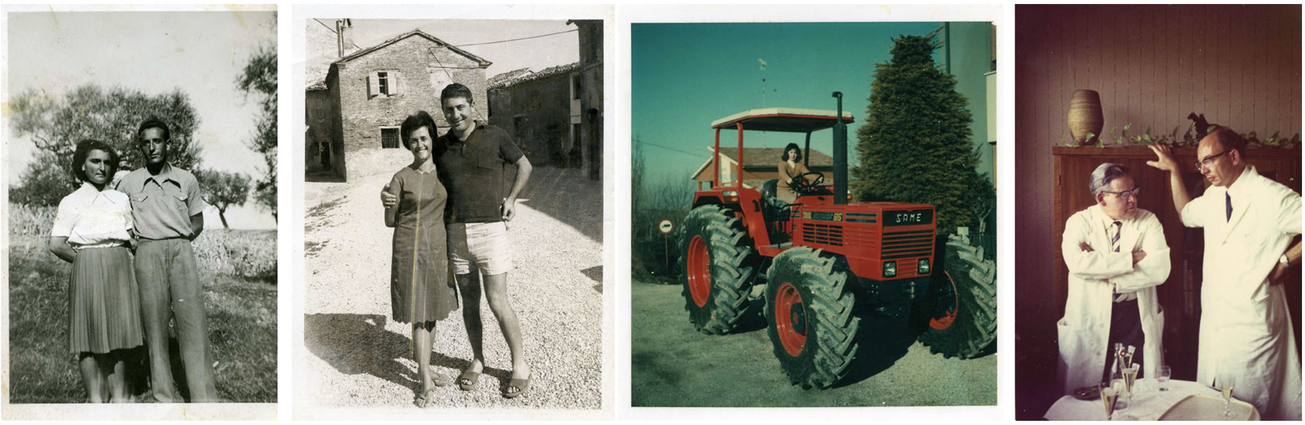


Fig. 2 IMAGO image samples from different epochs

exhibiting the unbalance in terms of the number of photos per year, since most fall between 1950 and 1980. The overall available images in this interval amount to 15,673. Out of such time intervals, the number of available images is too little to be considered. In Fig. 2 are shown, instead, four exemplar images from the IMAGO dataset belonging to different decades. Here, it is possible to appreciate what characterizes each photo (e.g., number of people, clothing, colors, and location), highlighting one of the main ones, i.e., each portrays at least one person.

Through the pre-processing phase, we aimed at isolating the regions of interest, of each image belonging to the IMAGO dataset, which could enhance the performance of the implemented deep learning-based models (details in Sects. 4). Following [9, 10], we extracted all the faces and full figure crops of the people portrayed for each image (referred to as FULL-IMAGES), creating the FACES and PEOPLE images sets, respectively. Importantly, such patches are always present, since all the photos belonging to the IMAGO dataset always include at least one person. For the images of FACES and PEOPLE sets, we processed each image of the IMAGO dataset using, respectively, an open-source implementation of YOLO-FACE [17] and YOLO [18]. The FACES and the PEOPLE images, hence, have been extracted accounting for the number of people portrayed in a photo: adopting a fixed-size bounding box may result in the possible loss of pixels related to the faces or people's full figures; for this reason, we rescaled the provided bounding boxes used to crop a face/people depending on the number of people portrayed in a photo, i.e., the greater the number of people, the smaller the bounding box. In Fig. 3 we reported an IMAGO full-image sample, with the respective crops taken, respectively, from FACES and PEOPLE sets.

It is possible to appreciate that PEOPLE images include details that are not present in FACES ones, such as the clothing of a person.

We then verified the utility of exploiting out-of-the-box image denoising and super-resolution algorithms, as all the images considered in this work derive from scans of the analog prints. For denoising, we tested the neural network model from [19] and the Bilateral Filter [20]. Concerning super-resolution instead, we used an open-source implementation of the ESRGAN model [21] within the Image Restoration Toolbox [22]. The overall improvement obtained from adopting such strategies was revealed to be negligible (less than 1% of overall accuracy with respect to the classical setting), so we hence opted for an analysis based on the original scans of analog photos (to not increase the complexity and the variables of the overall system). The fact that such algorithms didn't perform as expected is reasonable, taking into account that IMAGO pictures were taken with a huge variety of cameras (white &black vs color), scanned with very different devices (e.g., different scanners, and printed on different film paper (Resin Coated vs. Fiber Based)). Since the overall improvement obtained from adopting such strategies was revealed to be negligible, then we opted for an analysis based on the original scans, not considering these operations.

Considering the train, validation, and test set splitting, the FULL-IMAGES dataset (the IMAGO dataset) has been partitioned into three subsets of images. In particular, 80% of images for training, and 20% for testing. In addition, 10% of the training images is used as a validation set. To guarantee the popularity of those subsets, we selected the same partitioning for each year considered in the range provided in the IMAGO dataset (1930-1999). Importantly, for each image in the train, validation, and test sets of IMAGO, the faces and the people there portrayed are extracted and added to the corresponding FACES and PEOPLE sets, respectively. This process guarantees that no faces or people crops from the validation or test sets are observed during the training.

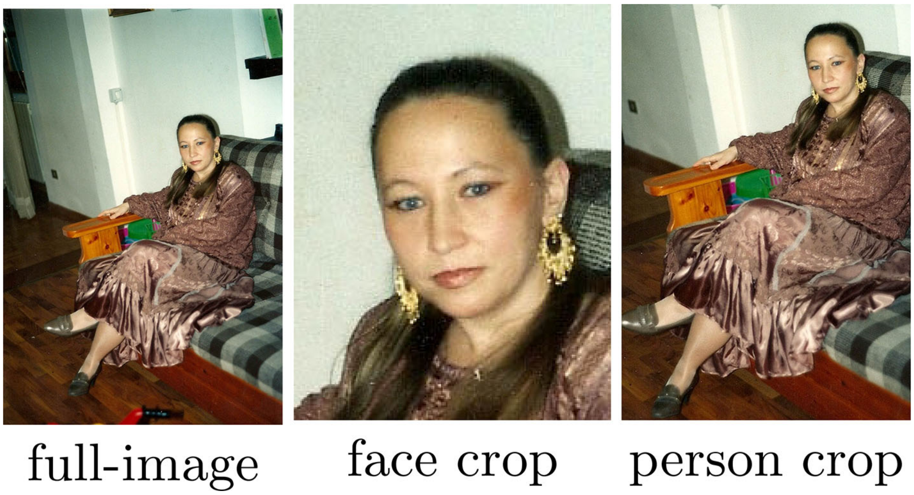


Fig. 3 FULL-IMAGES, FACES, and PEOPLE image samples

4 Model architectures and training settings

Considering the previously introduced IMAGO (and generated image patches) as the target dataset, we exploit single and multi-input deep learning architectures. The former analyzes the FULL-IMAGES and related image patches (FACES and PEOPLE) individually, while the latter combines them. For all our experiments, we employed three well-known CNN architectures: ResNet-50 [23], InceptionV3 [24], and DenseNet121 [25]. In particular, we considered their pre-trained version on ImageNet [26]. For each considered architecture, we replace the last fully connected layer (top-level classifier) with a randomly initialized classification layer, whose structure depends on the network embeddings (input) and the number of output classes (class prediction vector). In addition, the pre-trained convolutional layers were fine-tuned with the given input data.

One single-input classifier for each type of image patch has been trained and named after the considered one: *full-image*, *faces*, and *people*. Concerning FACES and PEOPLE images, instead of evaluating the accuracy for a single face or person, we aggregated the activations for every picture that appeared in the image. This means that if a picture of *n* people was used, the final prediction would be made by providing as input to the softmax function the mean of the activations extrapolated by the model from each face or person. In practice, the average of activation vectors returned by the single-input classifiers for each image was used to compute the most probable class. For the multi-input classifiers, instead, we developed what we defined as the *Merged* model, which merges the single-input classifiers previously mentioned, with the goal of not only exploiting different image patches but also learning how to do so. Specifically, the classification layer was removed from the pre-trained single-input classifiers, retaining the CNN backbone as feature extractors. Adopting such an architecture, the number of faces or persons represented in a picture determines the cardinality of the various extracted feature vectors, and the average of these feature vectors was computed to combine them with the vector derived from the whole image (which is always one feature vector). Multiple FACES and PEOPLE images may originate from a single one in FULL-IMAGES since a photo may feature more than one individual. The three resulting feature vectors (one per image patch) were combined linearly with a weighted sum, whose learnable weights are defined by three different real scalars (i.e, α, β, and γ). The output vector, resulting from the linear combination, is fed to a fully connected layer with a softmax activation, providing the final probability vector (used for classification). A schema of the explained architecture is reported in Fig. 4. In order to teach the newly introduced network how to execute such a combination, a new training session was conducted.

Considering now the training settings, we applied in all our experiments random cropping and horizontal flipping data augmentation. The fine-tuning procedure was carried out by exploiting a weighted cross-entropy loss and an Adam optimizer with a learning rate of $1e-4$ and a weight decay of $5e-4$. For the training of the *full-images* classifier, we fixed the batch size at 32 and for the *faces* and *people* classifiers, at 64 respectively.

5 Experimental results

Since we are here considering the dating task, the performance of the various models are measured in terms of time distance accuracies, as in [9, 10]. The time distance defines the tolerance accepted in predictions concerning the actual year. As an example, if a photo was shot in the year 1945

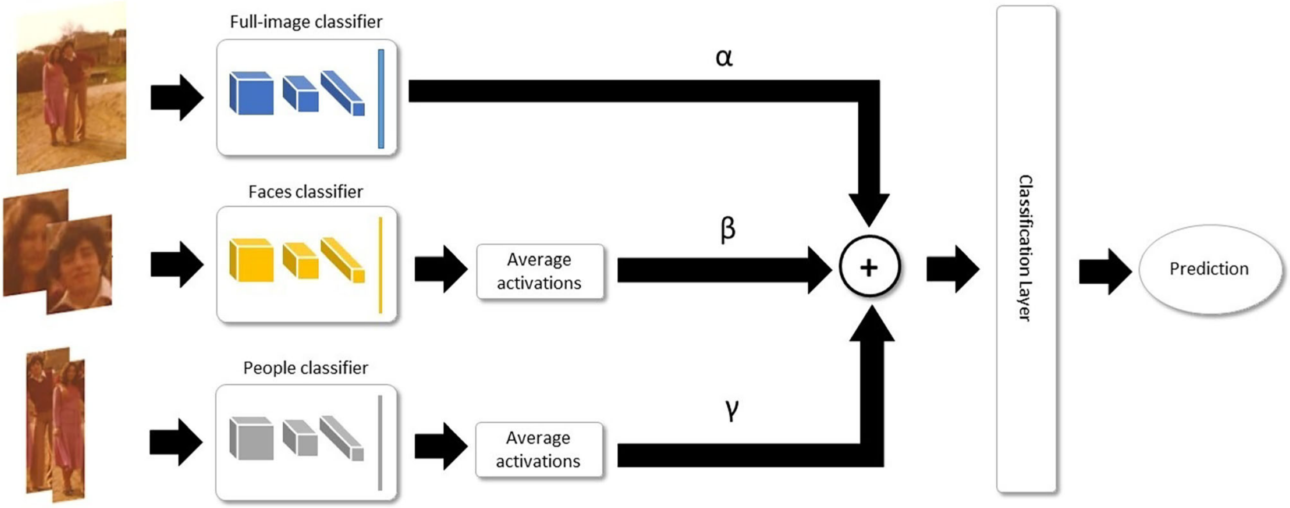


Fig. 4 Merged model architecture, α, β, γ represent the learnable weights

and the model returned 1940 (or even 1950) this would be considered a correct prediction if the time distance is set to be equal or greater than 5, otherwise, it represents an error. When the time distance is set to 0 the performance represents the classical accuracy (we are in a classification context, the years represent the classes). In this work, model accuracies were computed considering temporal distances of 0, 5, and 10 years. The results are reported in Table 2.

It is possible to appreciate that the different considered backbones (i.e., ResNet-50, InceptionV3, DenseNet121) provide similar accuracies for the single-input classifiers considering an intra-dataset perspective (row-wise). Considering instead the same architecture but trained and evaluated on different IMAGO patches (column-wise), Table 2 exhibits different accuracies. In particular, the *faces* and the *people* classifiers slightly outperform the *full-image* one. These results can be first explained by the averaging produced from the ensembling of various image regions, since using more data allows for the control of uncertainty and the reduction of prediction error [27]. However, these results may also be addressed to the fact that each model exploits and focus on different visual cues from people's appearance (e.g., hairstyle, dresses, trousers, earrings). Following such a line of thought, the Merge model improves compared to the single-input classifiers. The Merged model not only combined different visual cues from different image patches (ensembling) but also learn how to do so (feature fusion). The greater accuracy suggests that combining different visual features could effectively improve the year detection.

In the analyses that follow, the ResNet-50 was selected as the reference backbone, since it provided the best tradeoff between accuracy and model dimension [28]. We also took into consideration random patches in order to accurately measure the value in terms of prediction performance of the human-related features (e.g., faces and people) vs. non-human features in image dating. To do so, we created the RANDOM image set, which includes eight

Table 2 Model accuracies for different time distances (d = 0, d = 5, d = 10)

	Single-input classifier		
	ResNet-50	InceptionV3	DenseNet121
Time distance	*Full-image*		
d = 0	**11.31**	10.45	10.68
d = 5	**62.56**	61.38	60.77
d = 10	82.54	**82.82**	82.47
Time distance	*Faces*		
d = *0*	**15.01**	14.60	12.91
d = *5*	**58.09**	56.95	57.81
d = *10*	78.39	78.46	**79.70**
Time distance	*People*		
d = *0*	**15.77**	12.56	13.99
d = *5*	**62.40**	60.04	59.69
d = *10*	**82.47**	81.39	81.42
	Multi-input classifier		
	ResNet-50	InceptionV3	DenseNet121
Time distance	Merged		
d = *0*	**18.71**	17.14	16.22
d = *5*	**67.59**	67.56	66.67
d = *10*	86.17	**86.30**	86.07

Significance of bold is the best performance per given set of parameters

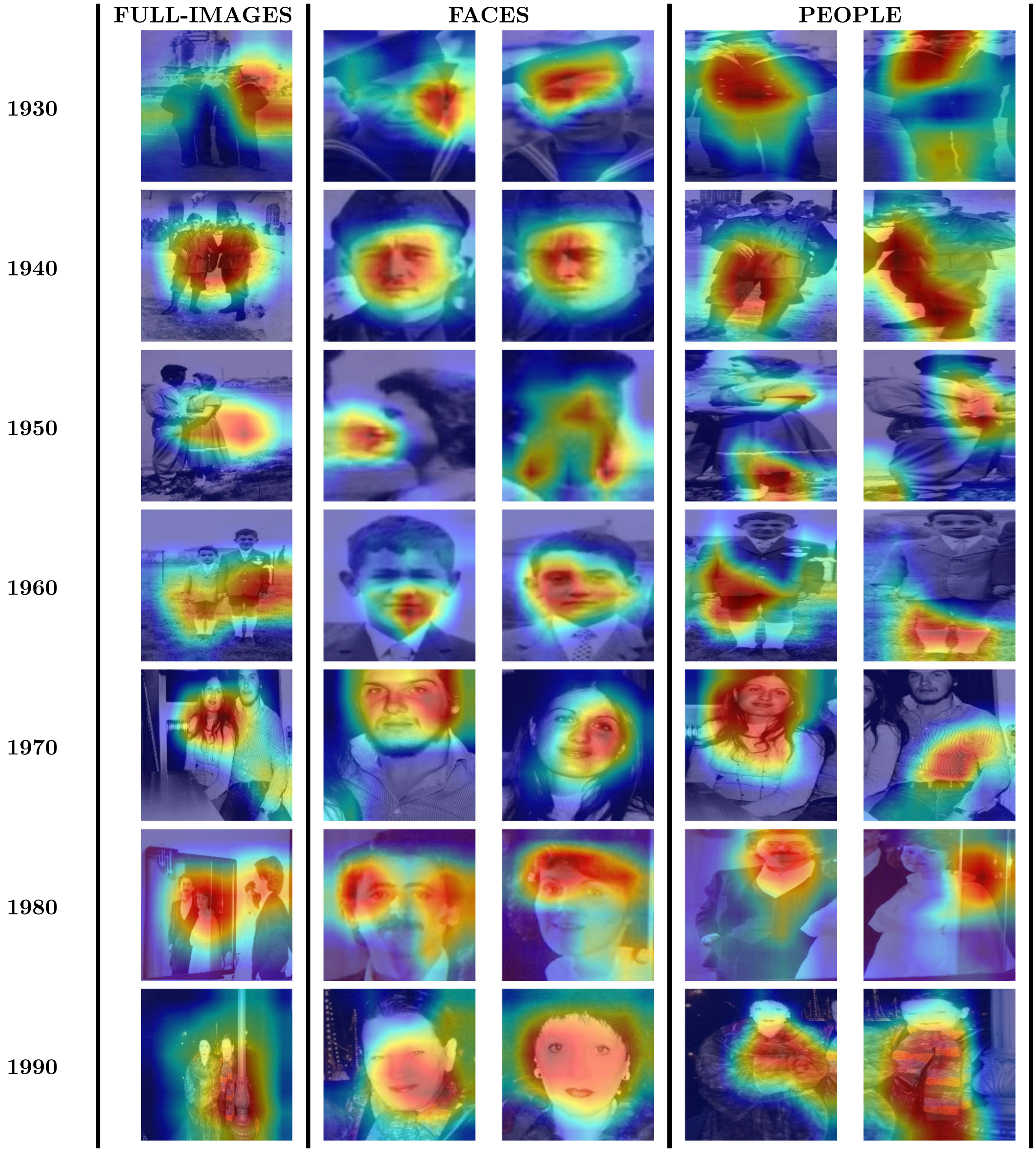


Fig. 5 Grad-CAM image samples spread over the 1930-1990 decades

Table 3 Models settings and accuracies of existing solutions and IMAGO considering the dating task

Original dataset	Architecture	Train cardinality (%)	Test cardinality (%)
Ginosar et al. [9]	VGG16	28,554 (80.0%)	8716 (20.0%)
Salem et al. [10]	AlexNet	72,800 (80.0%)	18,200 (20.0%)
IMAGO collection	ResNet50	11,252 (80.0%)	4,421 (20.0%)

random crop regions, of 128 × 128 pixels, for each image belonging to FULL-IMAGES. Other window sizes were also tested but returned a lower performance. Exploiting this set of images, we fine-tuned the ResNet-50 model to study its performance against the other image patches. The evaluation protocol described for the *faces* and *people* classifiers in Sect. 4 was applied to evaluate the *random* classifier. The obtained accuracies the *random* classifier are **11.64** for time-distance equal to 0 ($\mathbf{d} = \mathbf{0}$), **54.26** for $\mathbf{d} = \mathbf{5}$, and **76.12** for $\mathbf{d} = \mathbf{10}$. As also exhibited by *faces* and

Table 4 Comparison of our faces classifier evaluated on the test set of [9] with the model from [9] evaluated on the IMAGO-FACES test set. We considered the common time slice 1930-1999

	Faces classifier cross-dataset comparison with [9] – range 1930-1999	
Time distance	Our faces classifier tested on [9]	Model from [9] tested on IMAGO-FACES
d = 0	2.50	1.02
d = 5	24.49	12.4
d = 10	41.33	25.68

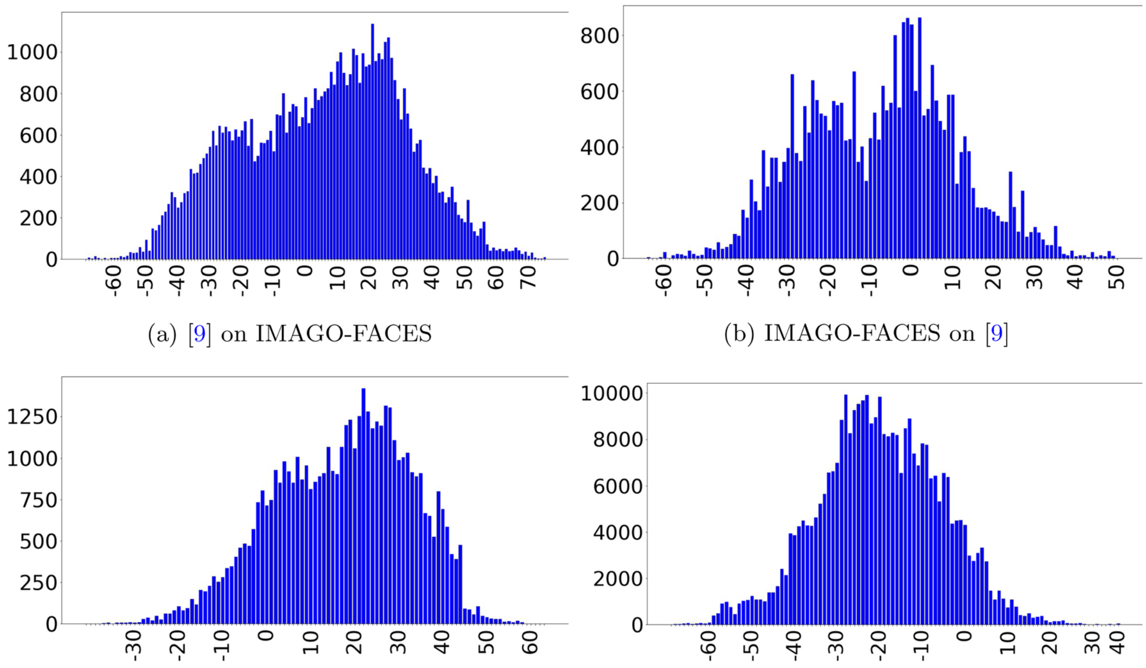
Table 5 Comparison of our faces classifier evaluated on the test set of [10] with the model from [10] evaluated on the IMAGO-FACES test set. We considered the common time slice 1950-1999

	Faces classifier cross-dataset comparison with [10] – range 1950–1999	
Time distance	Our faces classifier tested on [10]	Model from [10] tested on IMAGO-FACES
d = 0	1.45	2.46
d = 5	14.02	25.09
d = 10	26.20	46.13

Table 6 Comparison of our people classifier evaluated on the test set of [10] with the model from [10] evaluated on the IMAGO-PEOPLE test set. We considered the common time slice 1950-1999

	People classifier cross-dataset comparison with [10] – range 1950–1999	
Time distance	Our people classifier tested on [10]	Model from [10] tested on IMAGO-PEOPLE
d = 0	1.49	1.74
d = 5	18.08	18.22
d = 10	35.21	35.43

people classifiers, the *random* one achieved a slightly higher score with respect to the *full-image* classifier when the time distance is set to be equal to 0. However, it exhibited lower accuracies than all the other classifiers considering greater time distances. Even if the averaging effect occurred, the difference in performance between the *random* and the other classifiers could be explained by the different learned visual characteristics which provide useful clues to recognize a given time-slice. From these findings and taking into account that the time distance often



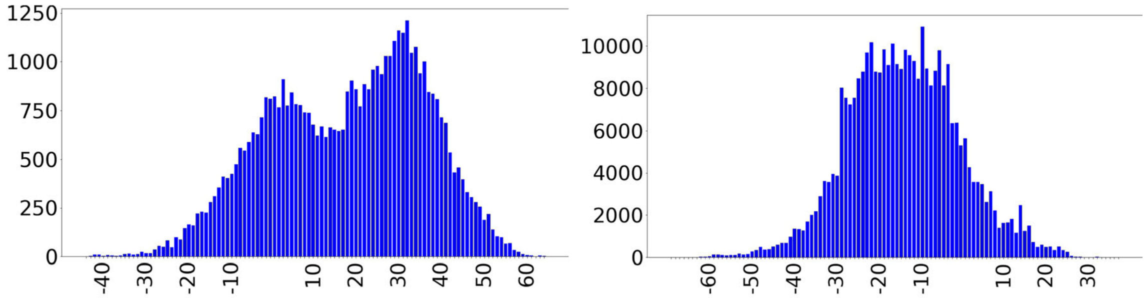
(a) [9] on IMAGO-FACES

(b) IMAGO-FACES on [9]

(c) [10] on IMAGO-FACES

(d) IMAGO-FACES on [10]

Fig. 6 Dating error distributions for faces



(a) [10] on IMAGO-PEOPLE

(b) IMAGO-PEOPLE on [10]

Fig. 7 Dating error distributions for people

used in historical analysis is ±5 years, as described in [2], we did not take into account the RANDOM pictures and the *random* classifier for the experiments that follows in our research.

After evaluating the performance of our models, we decided to investigate which visual cues led the models to determine the year of a family album photo. In this phase, we applied the Grad-CAM algorithm [29] to the single-input classifiers, which produce an overlapping heatmap that highlights the pixel areas exploited by the deep learning models to perform the classification. In Fig. 5 we report some Grad-CAM results for correctly classified images.

A distinct decade is represented by each row, which also contains the Grad-CAM of an IMAGO full-image and the two associated FACES and PEOPLE photos. It is clear that the single-input classifiers concentrated on various visual areas. The enhanced accuracy seen in the multi-input model may be supported by the fact that distinct single-input classifiers take advantage of different visual features. These visual results can be used from a socio-historical perspective to confirm whether the highlighted cues correlate to visual elements that are acknowledged as typical for a certain time period.

6 Cross-dataset experiments: evidence of intercultural influences?

Considering the existence of the USA-Italy cross-cultural influence on visual appearances between individuals, throughout the second half of the 1900 [30, 31] we carried out an analysis to verify whether this effect could be also quantified using deep learning. To achieve such goal, we adopted a cross-dataset approach considering the American-people datasets provided by [9, 10] and IMAGO as Italian counterpart. In particular, among all the relatable datasets [5, 9, 10, 13] no one includes family album photos (each picture contain at least one person). However, [9, 10] share some common traits with IMAGO: they analyzed American datasets comprising people's faces and torsos, where subjects are often in pose and dressed for a specific occasion. This means that it is possible to extract what characterizes all of them: people's faces and torsos. Considering such feature, the cross-dataset experiment will consider along with such datasets the pictures in the IMAGO one that are comparable to them (i.e., IMAGO-FACES and IMAGO-PEOPLE). Finally, all the images within the selected datasets (IMAGO, [9, 10]) were shot during the 20th century.

6.1 Cross-dataset performance evaluation

To perform cross-dataset experiments, the trained models from [9, 10] should be adopted. However, those models weren't available for the framework used in such work to train the IMAGO models (and also for evaluating them on the IMAGO dataset). So, we proceeded by mimicking the training procedure listed in the respective works [9, 10] to define different deep learning-based models that could be adopted to perform the target analysis. To achieve such a goal, we first fine-tuned the VGG16 and AlexNet architectures, respectively used in [9, 10], following the procedures described by the authors. In all the cases, an 80%-20% training-test split was considered. All the information is reported in Table 3. Important to highlight that the dataset introduced in [9] considers only people's faces, while the one introduced in [10] offers both people's faces and torsos. We then evaluated these models on the IMAGO dataset. Vice versa, the *faces* and *people* classifiers, presented in this work, have been evaluated on the corresponding regions offered in the datasets from [9, 10]. For a fair evaluation, the experiments were carried out on the 1930-1999 time-span for the [9] vs. IMAGO comparison, while considering 1950-1999 for the [10] vs. IMAGO one, respectively. The results of such evaluation are reported in Tables 4, 5 and 6. As expected, the final performance is really poor in both directions, i.e., the models fine-tuned on our dataset and evaluated on the test set of the related works and vice versa. This may be due to the domain-shift effect (these datasets have been acquired from multiple locations, using different cameras) [32]. However, another reason for such poor performance could be addressed to the intercultural influence that changes the visual appearance of people in different ages.

To explore such possible influence quantitatively, we collected the error between the predicted and the actual year per each picture. The error distributions are reported in Figs. 6 and 7 for the cross-dataset experiments involving faces and people images. In particular, Figs. 6a and c depict that the date estimation error distributions are shifted towards positive values, while, in Figs. 6b and d towards negative ones. The models built on top of American datasets [9, 10] applied to IMAGO-FACES tend to overestimate the image shooting year while the opposite phenomenon (underestimation) occurs when the model presented in this work is applied to [9] and [10]. The same

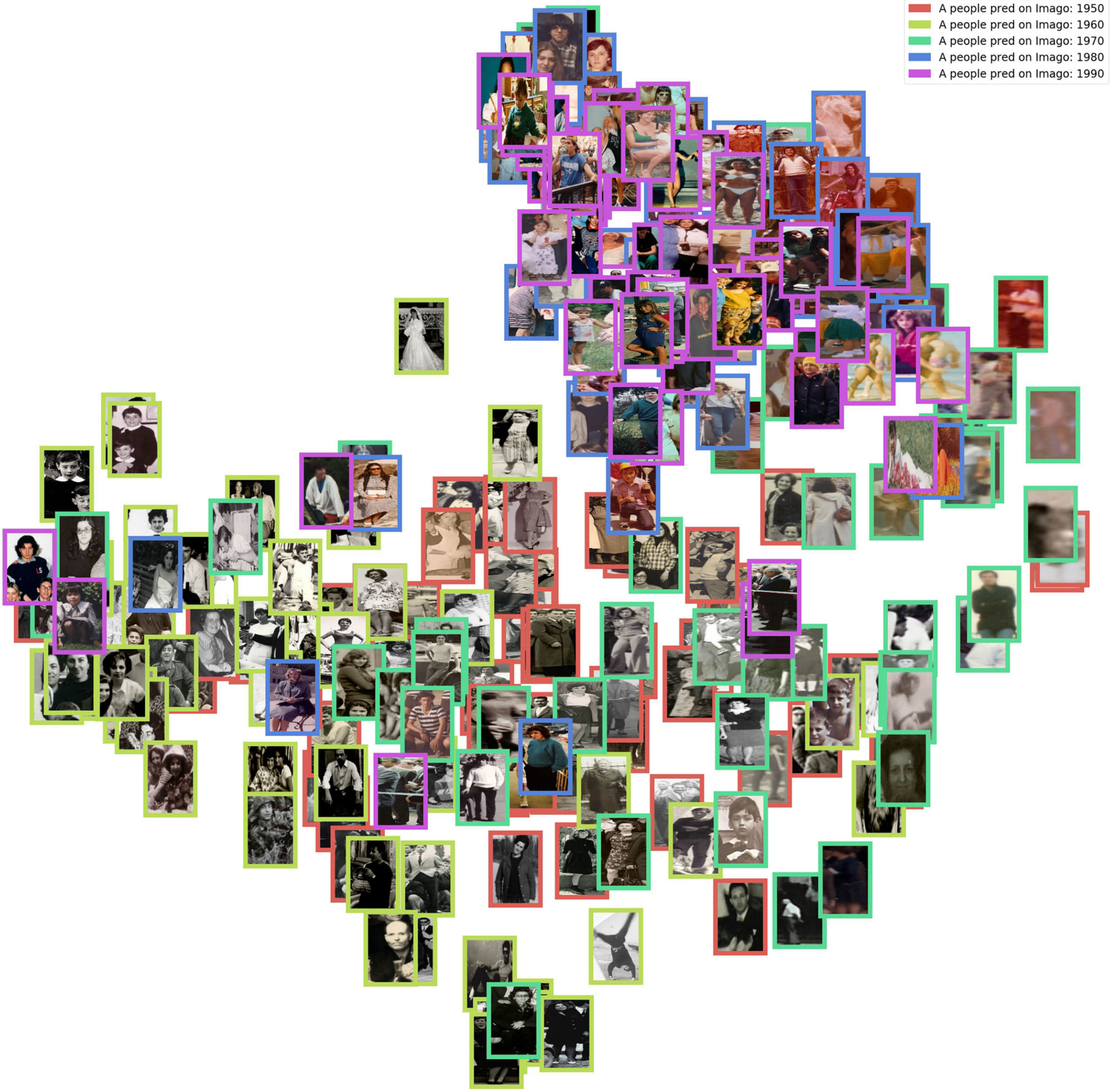

🟥 A people pred on Imago: 1950
🟨 A people pred on Imago: 1960
🟩 A people pred on Imago: 1970
🟦 A people pred on Imago: 1980
🟪 A people pred on Imago: 1990

Fig. 8 UMAP applied to the embeddings of the model trained with [10] (indicated as A) on the IMAGO-PEOPLE dataset. The selected images were correctly predicted by the model within a decade of confidence

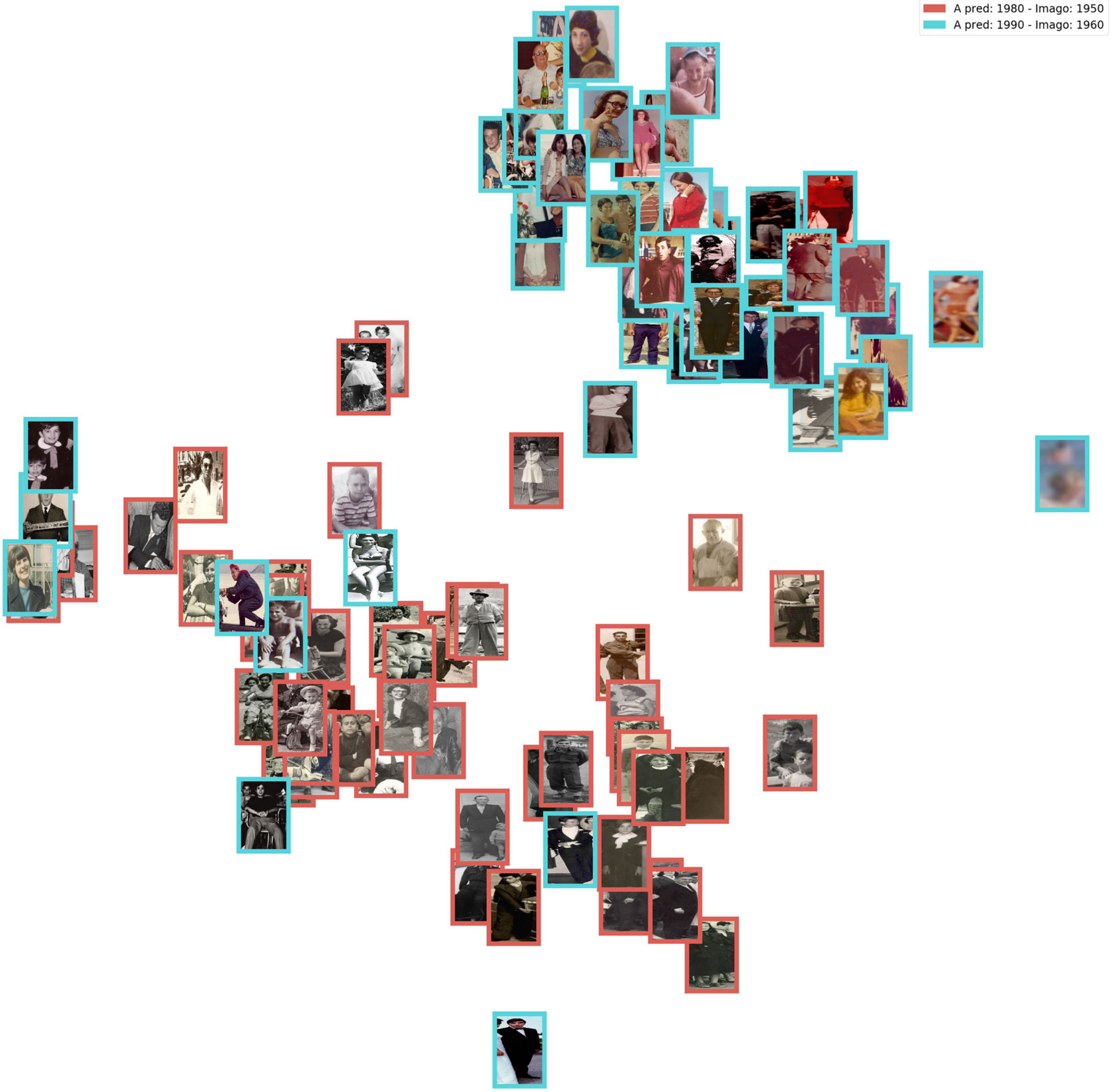


Fig. 9 UMAP applied to the embeddings of the model trained with [10] (indicated as A) on the IMAGO-PEOPLE dataset. The selected images were wrongly predicted to be 30 years forward the real shooting date

phenomenon appeared considering people's torsos. Nevertheless, we were able to analyze such phenomena only for [10] which provides pictures of full-figure instead of only faces. The obtained results are reported in Fig. 7. To further investigate whether the errors were statistically significant, we performed a data analysis process. Firstly, we measured the normality of the error distributions by adopting a normality test that combines skew and kurtosis to produce an omnibus test [33, 34]. The normality test was adopted to discriminate between parametric and non-parametric statistical tests. In our experimental sessions, none of the considered distributions passed the normality test (p-value <0.001, the null hypothesis test that a sample comes from a normal distribution). For this reason, we proceeded by adopting non-parametric tests. In particular, we evaluate whether the difference between the ground truth and model prediction pairs (i.e., error distributions) were statistically significant performing the Wilcoxon signed-rank test. The Wilcoxon signed rank is a non-parametric test where the null hypothesis state: "two

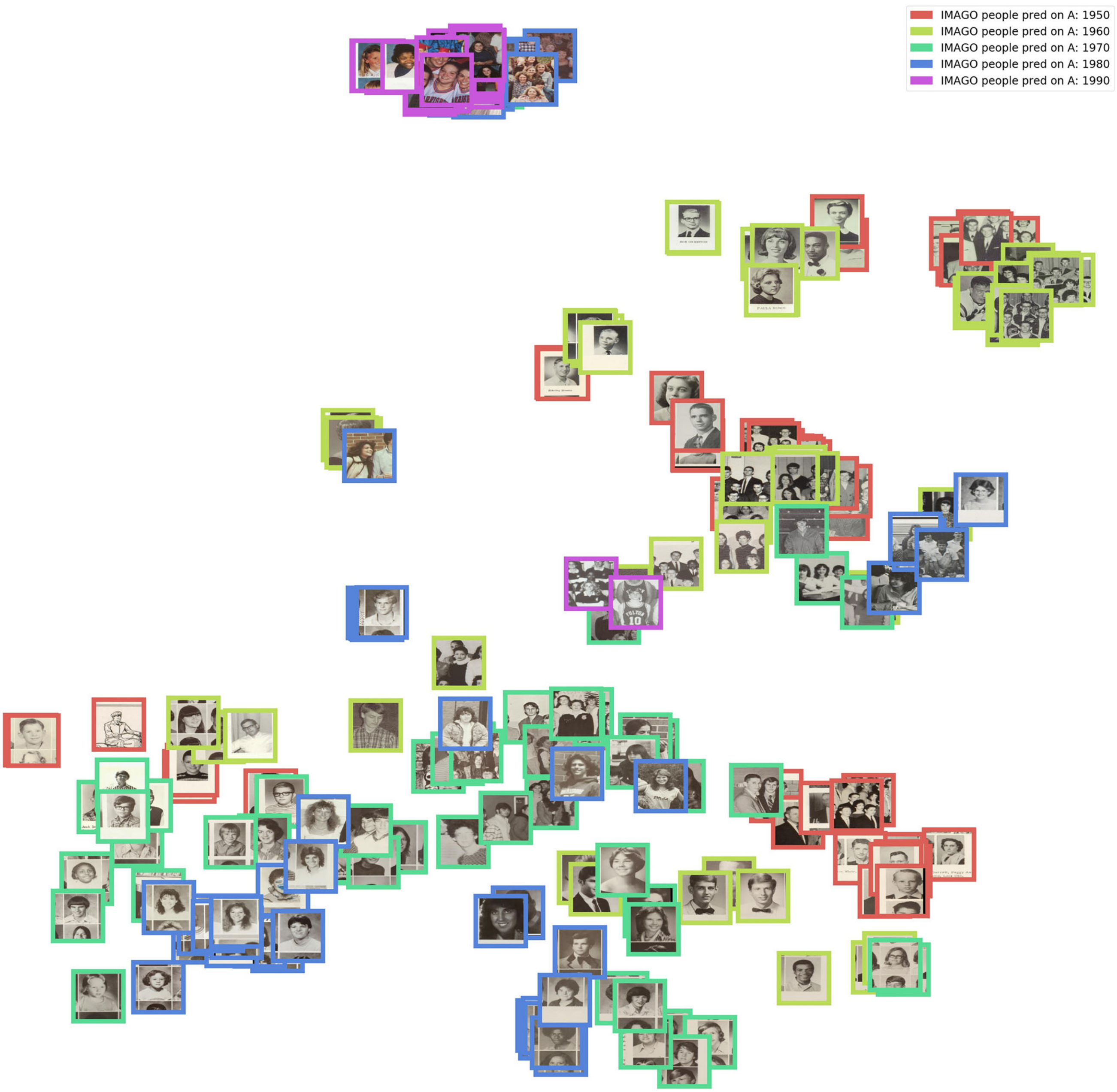


Fig. 10 UMAP applied to the embeddings of the model trained with IMAGO-PEOPLE on [10] (indicated as A). The selected images were correctly predicted by the model within a decade of confidence

related paired samples come from the same distribution". In particular, it tests whether the distribution of the differences is symmetric about zero [35]. Also, in this case, the null hypothesis was rejected for all the conditions (p-value <0.001), indicating that the considered differences exhibit different distributions. Finally, we verified whether the shift between two cross-dataset (e.g.,) settings came effectively from two different distributions with the Mann–Whitney U test [36]. This provides some clues about the significance of the overestimation/underestimation effect. The non-parametric Mann–Whitney U rank test hypothesizes two independent samples and tests the null hypothesis that the distribution underlying the first sample is the same as the distribution underlying the second sample. Even for the Whitney U test the null hypothesis was rejected for all the conditions (p-value <0.001), indicating that the considered cross-shift differences came from different distributions.

These results motivated us to perform an additional visual analysis to qualitatively explore the possible time-shift phenomenon in a cross-dataset setting.

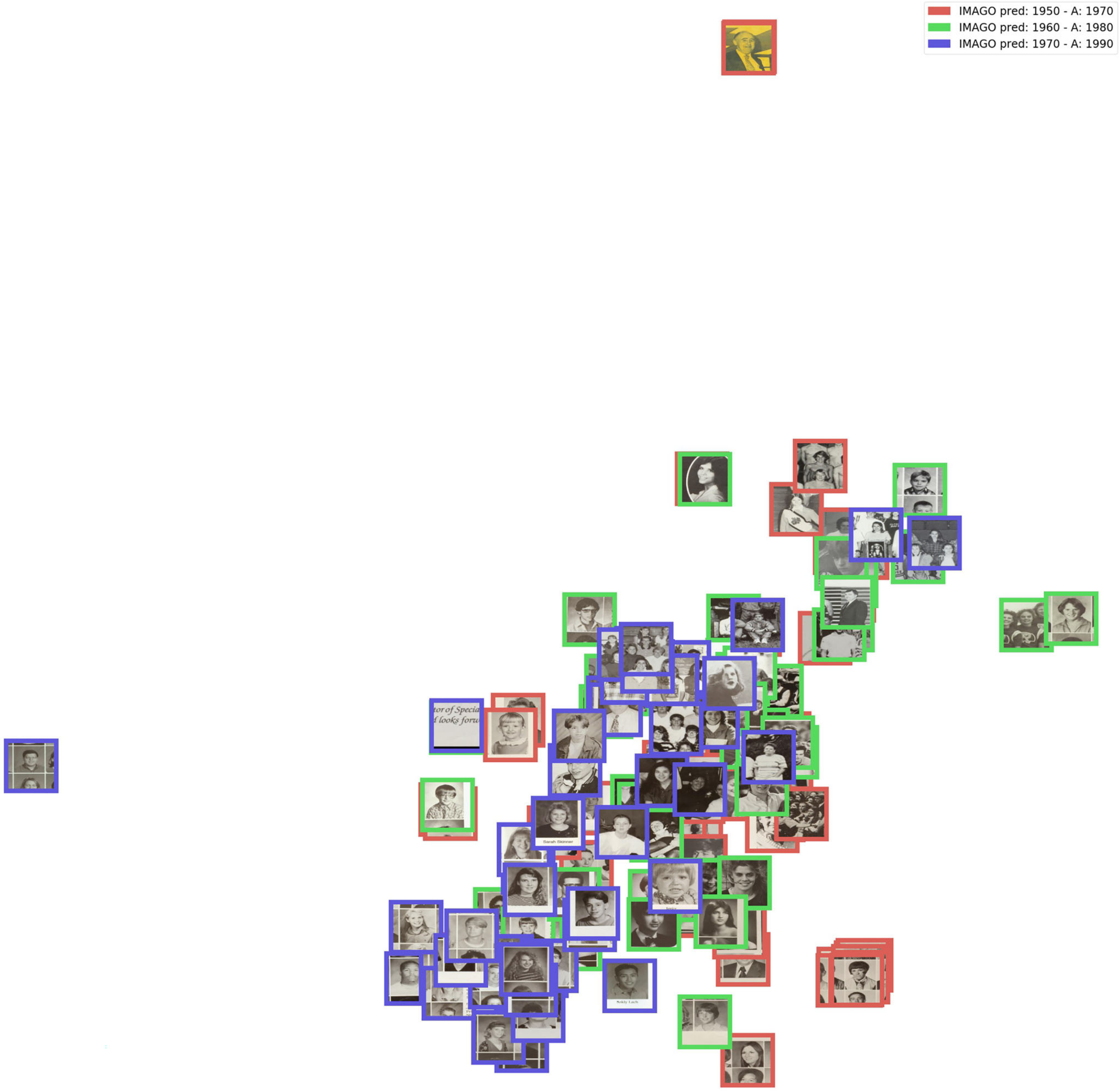


Fig. 11 UMAP applied to the embeddings of the model trained with IMAGO-PEOPLE on [10] (indicated as A). The selected images were wrongly predicted to be −20 years forward the real shooting date

6.2 Evaluate visual intercultural cues with data visualization: a UMAP qualitative analysis

Considering the results reported in Fig. 6, we decided to visually explore the images that were most shifted, from a dating perspective, while evaluating the models described in Sect. 6 on the IMAGO datasets, and the IMAGO models on [9, 10]. In practice, we exploited the CNN extracted feature (embeddings) on the target datasets in a cross-dataset setting. However, for the considered models (ResNet50, VGG, AlexNet), the embeddings lie in a latent space of 2048 or 4096 dimensions. For such reason, we put to good use one of the most used data dimensionality reduction algorithms: UMAP [11]. The aim of dimensionality reduction is to preserve as much of the significant structure of the high-dimensional data as possible in a low-dimensional map (i.e., 2 or 3 dimensions). When the data presents a non-linear structure (as in the case of a CNN latent space), UMAP and the t-distributed stochastic neighbor embedding (t-SNE) represent a valid method to reduce them due to their non-linear nature [11, 37]. However, UMAP is faster and scales better for both dataset dimensionality and cardinality while better preserving the global structure of the data [11]. In particular, t-SNE has been observed to distort distances between clusters in the original high-dimensional space, while UMAP more accurately preserves these distances [38, 39]. In other words, this technique produces high-quality visualizations by reducing the high-dimensional data revealing structures in them also considering large data sets [11]. In our analysis, we employed the official implementation of the UMAP algorithm [40]. To carry out a cross-dataset analysis, we picked as target datasets the one introduced in [10] and IMAGO which includes people's torsos. This choice was mainly driven by the fact that these datasets possess a greater, higher detailed, and more varied number of pictures with respect to the one introduced in [9].

Firstly, we analyzed the clusters extracted by the UMAP algorithm while being applied to the embeddings extracted by inferring date with the model trained with [10] on IMAGO-PEOPLE. In Fig. 8 we reported a sample of images that were correctly predicted for each of the considered decades in the common dataset time-span. In Fig. 9, instead, we report a sample of images that were wrongly predicted with a shift of 30 years, which is the most occurrent shift reported in Fig. 7 (Sect. 6). It is worth noticing that in Fig. 8 the UMAP algorithm was able to highlight clusters for different decades that however possess intersection with clusters of adjacent decades (e.g. some pictures from 1950 are mixed with the ones of 1960). In Fig. 9 instead, it is interesting to note that many samples that were labeled with a 30-year shift are not colored: this could mean that the model exploited other cues apart from the colors to date those images (e.g., the style of men in lower pictures in Fig. 9 possess very similar fashion style).

Secondly, we explored the output of the UMAP algorithm while being evaluated on the embeddings extracted by inferring the date on [10] with the model trained with IMAGO-PEOPLE. In Fig. 10 we report a sample of images that were correctly predicted for each of the considered decades in the common dataset time-span. In Fig. 11 instead, we report a sample of images that were wrongly predicted with a shift of −20 years, which is the majority shift reported in Fig. 7 (Sect. 6). Also, in this case, the UMAP algorithm was able to highlight clusters for different decades that however possess intersection with clusters of adjacent decades (Fig. 10). In Fig. 11, instead, it is interesting to note that the majority of samples that were labeled with a −20 shift are in black-white: this could mean that the model exploited other cues apart from the colors to date those images (e.g. similar female hairstyles are near in the 1990 left-lower cluster in Fig. 11). We want to highlight that these interesting results were obtained in a qualitative analysis setting, and so they cannot be generalized considering also that involved just a subset of the considered dataset [41]. However, the adoption of data visualization algorithms, such as the UMAP, to visualize neighbor images in the latent space ease and speed up the classical approach that would be done in museums or in academia for searching relationships with visual cues. This reduces the time-consuming approach which often subjects this kind of analysis. So, this approach could be a valuable tool for socio-historical researchers, as it allows for a deeper understanding of complex phenomena, such as cross-cultural influences.

7 Discussion, conclusions, and future works

This work analyzed the problem of image dating by exploiting the IMAGO dataset, a collection composed of analog prints belonging to family albums shot during the 20th century considering as the target time-span the 1930-1999 age. We trained and tested single and multi-input deep learning models exploiting different regions (full-image, faces, people) of a given photo to identify its shooting year. Then, we adopted the *faces* and *people* models to search for cues of intercultural influences through cross-dataset experiments. In particular, we applied the models trained on IMAGO-FACES and IMAGO-PEOPLE images and the ones trained on datasets provided by [9, 10], following a cross-dataset configuration. The dating error distributions exhibited an interesting symmetry that motivates us to perform a qualitative UMAP analysis to explore the visual cues that could support this phenomenon.

Despite those interesting results, our cross-cultural visual cues analysis framework has some limitations. We start from the observed domain shift effect, which has led to a high error rate during our cross-dataset experiments [32]. This may be due to a number of reasons. Firstly, we should remind that the three datasets considered in this work are conceptually different. IMAGO mainly contains family album pictures shot in Italy by Italian citizens. The datasets introduced in [9, 10], instead, include pictures extracted from American school yearbooks. Secondly, different digitization devices (e.g., different types of scanners and cameras) could provide changes in textures which CNNs are sensitive to [42, 43]. However, the domain shift effect is partially alleviated considering that the models share similar classification tasks and that the datasets share some common visual features such as people's hairstyles, clothing, and earrings which amount to useful cues to individuate the date of an image [8–10, 42, 43]. To uncover which kind of visual features most influenced dating errors from one domain to another, a Grad-Cam based analysis could be employed in a future contribution [29]. Another interesting aspect that may be further developed amounts to systematically compare the photos of the datasets based on their actual and predicted date. In other words, it could be possible to apply the IMAGO model, for example, to the [9] dataset, collect all the photos misclassified within a decade, and compare those photos to the ones within the IMAGO dataset which have been correctly classified within the same decade. This approach may automate the comparison of different styles across different countries at different times and be supported by the use of well-known visualization tools such as UMAP or t-SNE.

Other approaches could also be employed, which do not solely rely on a comparison of the embeddings extracted from the given datasets. For example, we could use object detectors to identify particular objects in both the pictures that are present in the misclassified images from a dataset and in the correctly classified images of the other dataset (e.g., particular dresses, haircuts, face features, physical objects) [44, 45]. This may lead to the creation of a further layer of knowledge including those objects which most frequently appear in the presence of cross-dataset misclassifications and within-dataset correct classifications. A final but also important aspect concerns improving the performance of the adopted models: modern computer vision architectures such as Vision Transformers could also be adopted [46, 47]. At the same time, we could try advanced restoration deep learning models, such as the one introduced in [48], to reduce noise, picture imperfections, and non-useful cues that could improve the classification performance of the models.

Finally, our work may benefit from the adoption of a multi-modal approach (i.e., image-text) mimicking, even more, the process that is usually carried out by historians in their analyses (i.e., visual analysis along with consultation of textual archival documents). This approach could both support and justify the temporal shift observed in this work.

Funding Open access funding provided by Alma Mater Studiorum - Università di Bologna within the CRUI-CARE Agreement.

Data Availibility Statement The datasets generated during and/or analyzed during the current study are available from the corresponding author upon reasonable request.

Declarations

Conflicts of interest The authors declare that they have no conflict of interest.

References

1. MoMA: Vernacular photography. https://www.moma.org/collection/terms/vernacular-photography (2020)
2. Calanca D (2011) Italians posing between public and private theories and practices of social heritage. Almatourism J Tour Culture Territ Dev 2(3):1–9
3. Sandbye M (2014) Looking at the family photo album: a resumed theoretical discussion of why and how. J Aesthet Culture 6(1):25419
4. Mitman G, Wilder K (2019) Documenting the world: film, photography, and the scientific record. University of Chicago Press, Chicago
5. Molina A, Riba P, Gomez L, Ramos-Terrades O, Lladós J (2021) Date estimation in the wild of scanned historical photos: An image retrieval approach. In: International Conference on Document Analysis and Recognition, pp 306–320. Springer
6. Stacchio L, Angeli A, Lisanti G, Marfia G (2022) Applying deep learning approaches to mixed quantitative-qualitative analyses. In: Proceedings of the 2022 ACM Conference on Information Technology for Social Good, pp 161–166
7. Stacchio L, Angeli A, Lisanti G, Marfia G (2022) Searching for cultural relationships through deep learning models
8. Stacchio L, Angeli A, Lisanti G, Calanca D, Marfia G (2022) Towards a holistic approach to the socio-historical analysis of vernacular photos. ACM Trans Multimed Comput Commun Appl (TOMM) 18(3):1–23
9. Ginosar S, Rakelly K, Sachs S, Yin B, Efros AA (2015) A century of portraits: A visual historical record of american high school yearbooks. In: Proceedings of the IEEE International Conference on Computer Vision Workshops, pp 1–7
10. Salem T, Workman S, Zhai M, Jacobs N (2016) Analyzing human appearance as a cue for dating images. In: 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), pp 1–8. IEEE
11. McInnes L, Healy J, Melville J (2018) Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426
12. Fernando B, Muselet D, Khan R, Tuytelaars T (2014) Color features for dating historical color images. In: 2014 IEEE International Conference on Image Processing (ICIP), pp 2589–2593. IEEE
13. Müller E, Springstein M, Ewerth R (2017) "When was this picture taken?"–image date estimation in the wild. In: European Conference on Information Retrieval, pp. 619–625. Springer
14. Amelio A, Bonifazi G, Corradini E, Di Saverio S, Marchetti M, Ursino D, Virgili L (2022) Defining a deep neural network ensemble for identifying fabric colors. Appl Soft Comput 130:109687
15. Amelio A, Bonifazi G, Corradini E, Ursino D, Virgili L (2023) A multilayer network-based approach to represent, explore and handle convolutional neural networks. Cognit Comput 15(1):61–89
16. Amelio A, Bonifazi G, Cauteruccio F, Corradini E, Marchetti M, Ursino D, Virgili L (2023) Representation and compression of

residual neural networks through a multilayer network based approach. Expert Syst Appl 215:119391

17. Thanh Nguyen: Yolo face implementation. https://github.com/sthanhng/yoloface. Online; accessed 3 August 2020 (2018)
18. Joseph Redmon: YOLO: Real Time Object Detection. https://github.com/pjreddie/darknet/wiki/YOLO:-Real-Time-Object-Detection. Online; accessed 3 August 2020 (2019)
19. Kai Zhang, Wangmeng Zuo, Zhang L (2018) FFDNet: toward a fast and flexible solution for CNN-based image denoising. IEEE Trans Image Process 27(9):4608–4622
20. Paris S, Kornprobst P, Tumblin J, Durand F (2007) A gentle introduction to bilateral filtering and its applications. In: ACM SIGGRAPH 2007 Courses. SIGGRAPH '07, p. 1. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/1281500.1281602
21. Wang X, Yu K, Wu S, Gu J, Liu Y, Dong C, Loy CC, Qiao Y, Tang X (2018) ESRGAN: Enhanced Super-Resolution Generative Adversarial Networks. In: Proceedings of the European conference on computer vision (ECCV) workshops, pp 0-0. 2018
22. Zhang K (2019) Image Restoration Toolbox. https://github.com/cszn/KAIR
23. He K, Zhang X, Ren S, Sun J (2015) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770-778. 2016
24. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z (2015) Rethinking the Inception Architecture for Computer Vision
25. Huang G, Liu Z, van der Maaten L, Weinberger KQ (2018) Densely Connected Convolutional Networks
26. Deng J, Dong W, Socher R, Li L, Kai Li, Li Fei-Fei (2009) Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp 248–255. https://doi.org/10.1109/CVPR.2009.5206848
27. Bishop CM, Nasrabadi NM (2006) Pattern recognition and machine learning. Springer, Cambridge
28. Coleman C, Kang D, Narayanan D, Nardi L, Zhao T, Zhang J, Bailis P, Olukotun K, Re C, Zaharia M (2019) Analysis of DAWNBench, a Time-to-Accuracy Machine Learning Performance Benchmark. ACM SIGOPS Oper Syst Rev 53(1):14–25
29. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D (2019) Grad-cam: visual explanations from deep networks via gradient-based localization. Int J Comput Vis 128(2):336–359. https://doi.org/10.1007/s11263-019-01228-7
30. Gundle S, Guani M (1986) L'americanizzazione del quotidiano televisione. e consumismo nell'italia degli anni cinquanta. Quaderni storici 62:561–594
31. Cannato VJ (2022) How America became Italian. t.ly/fUKb
32. Quinonero-Candela J, Sugiyama M, Schwaighofer A, Lawrence ND (2008) Dataset shift in machine learning. Mit Press, Cambridge
33. DIAgostino R (1971) An omnibus test of normality for moderate and large sample sizes. Biometrika 58(34):1–348
34. D'Agostino R, Pearson ES (1973) Tests for departure from normality empirical results for the distributions of b 2 and b. Biometrika 60(3):613–622
35. Conover WJ (1999) Practical Nonparametric Statistics vol. 350. john wiley & sons, USA
36. Mann HB, Whitney DR (1947) On a test of whether one of two random variables is stochastically larger than the other. Ann Math Stat 18:50–60
37. Van der Maaten L, Hinton G (2008) Visualizing data using t-sne. J Mach Learn Res 9(11):2580
38. Pal K, Sharma M (2020) Performance evaluation of non-linear techniques umap and t-sne for data in higher dimensional topological space. In: 2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC), pp 1106–1110. IEEE
39. Damrich S, Böhm N, Hamprecht FA, Kobak D (2022) From *t*-sne to umap with contrastive learning. In: The Eleventh International Conference on Learning Representations
40. McInnes L, Healy J, Saul N, Grossberger L (2018) Umap: uniform manifold approximation and projection. J Open Source Softw 3(29):861
41. Boeije HR (2009) Analysis in qualitative research. Analysis in qualitative research, 1–240
42. Yosinski J, Clune J, Bengio Y, Lipson H (2014) How transferable are features in deep neural networks? Advances in neural information processing systems. 27
43. Nam H, Lee H, Park J, Yoon W, Yoo D (2021) Reducing domain gap by reducing style bias. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 8690–8699
44. Stacchio L, Angeli A, Hajahmadi S, Marfia G (2021) Revive family photo albums through a collaborative environment exploiting the hololens 2. In: 2021 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct), pp. 378–383. IEEE
45. Zou Z, Chen K, Shi Z, Guo Y, Ye J (2023) Object detection in 20 years: A survey. In: Proceedings of the IEEE
46. Khan S, Naseer M, Hayat M, Zamir SW, Khan FS, Shah M (2022) Transformers in vision: asurvey. ACM Comput Surv (CSUR) 54(10s):1–41
47. Wortsman M, Ilharco G, Gadre SY, Roelofs R, Gontijo-Lopes R, Morcos AS, Namkoong H, Farhadi A, Carmon Y, Kornblith S (2022) Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In: International Conference on Machine Learning, pp. 23965–23998. PMLR
48. Wan Z, Zhang B, Chen D, Zhang P, Chen D, Liao J, Wen F (2020) Bringing old photos back to life. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2747–2757