Computational Visual Media
https://doi.org/10.1007/s41095-021-0255-3

Research Article

NPRportrait 1.0: A three-level benchmark for non-photorealistic rendering of portraits

Paul L. Rosin[1] (✉), Yu-Kun Lai[1], David Mould[2], Ran Yi[3], Itamar Berger[4], Lars Doyle[2], Seungyong Lee[5], Chuan Li[6], Yong-Jin Liu[3], Amir Semmo[7], Ariel Shamir[4], Minjung Son[8], and Holger Winnemöller[9]

Abstract Recently, there has been an upsurge of activity in image-based non-photorealistic rendering (NPR), and in particular portrait image stylisation, due to the advent of neural style transfer (NST). However, the state of performance evaluation in this field is poor, especially compared to the norms in the computer vision and machine learning communities. Unfortunately, the task of evaluating image stylisation is thus far not well defined, since it involves subjective, perceptual, and aesthetic aspects. To make progress towards a solution, this paper proposes a new structured, three-level, benchmark dataset for the evaluation of stylised portrait images. Rigorous criteria were used for its construction, and its consistency was validated by user studies. Moreover, a new methodology has been developed for evaluating portrait stylisation algorithms, which makes use of the different benchmark levels as well as annotations provided by user studies regarding the characteristics of the faces. We perform evaluation for a wide variety of image stylisation methods (both portrait-specific and general purpose, and also both traditional NPR approaches and NST) using the new benchmark dataset.

Keywords non-photorealistic rendering (NPR); image stylization; style transfer; portrait; evaluation; benchmark

1 School of Computer Science and Informatics, Cardiff University, Cardiff, UK. E-mail: P. L. Rosin, RosinPL@cardiff.ac.uk (✉); Y.-K. Lai, Yukun.Lai@cs.cardiff.ac.uk.

2 School of Computer Science, Carleton University, Ottawa, Canada. E-mail: D. Mould, mould@scs.carleton.ca; L. Doyle, larsdoyle@cmail.carleton.ca.

3 Department of Computer Science and Technology, Tsinghua University, Beijing, China. E-mail: R. Yi, yr16@mails.tsinghua.edu.cn; Y.-J. Liu, liuyongjin@tsinghua.edu.cn.

4 Reichman University (the Interdisciplinary Center), Herzliya, Israel. E-mail: I. Berger, berger.itamar@gmail.com; A. Shamir, arik@idc.ac.il.

5 Department of Computer Science and Engineering, Pohang University of Science and Technology, Pohang, Republic of Korea. E-mail: leesy@postech.ac.kr.

6 Lambda Labs, Inc., San Francisco, USA. E-mail: c@lambdal.com.

7 Hasso Plattner Institute, University of Potsdam, Potsdam, Germany. E-mail: Amir.Semmo@hpi.de.

8 Multimedia Processing Laboratory, Samsung Advanced Institute of Technology, Suwon, Republic of Korea. E-mail: minjungs.son@samsung.com.

9 Adobe Systems, Inc., San Jose, USA. E-mail: hwinnemo@adobe.com.

Manuscript received: 2021-07-17; accepted: 2021-09-16

1 Introduction

Image-based non-photorealistic rendering (NPR) lies at the intersection of computer graphics and computer vision. It has the aim of synthesising new stylised images based on existing images [1]. This paper focusses on portrait image stylisation. A comprehensive historical overview of 30 years of image-based NPR is provided by Kyprianidis et al. [1], while an overview of the state of the art in 2013 is given by Rosin and Collomosse [2]. Shortly after this date the course of NPR dramatically changed with the advent of deep learning and the huge popularity of neural style transfer (NST), initiated by Gatys et al.'s landmark paper [3]. Jing et al. [4] provided a recent review of NST. However, NST methods still have limitations, as discussed by Semmo et al. [5].

Despite substantial activity in NPR and image stylisation, evaluation of reported results is limited, and falls far below the norms in the computer vision

and machine learning communities. Some of these issues were identified by David Salesin in 2002, as recounted by Gooch et al. [6]. Kyprianidis et al. [1] stated that few papers presented structured methodologies for evaluation, with subjective side-by-side visual comparison being more typical. Some authors even argue against performing NPR benchmarking at all! [7] The problem is that evaluation of NPR results is less straightforward than for computer vision or machine learning for a variety of reasons, including:

- Image stylisation tasks lack ground truth. We do not have any pairs of inputs with ideal stylised outputs, and such pairs are not possible even in principle. Moreover, NPR algorithms are often designed to produce novel styles, with no prior examples available.
- Unlike tasks such as classification where all algorithms aim to return the same correct result, stylisation tasks do not have a unique output. Many aspects of stylisations can vary, independently of rendering quality, on dimensions such as medium (oil paint, crayon) or artistic school (impressionist, cubist). The vast range of potential stylisations makes it impossible to generate all possible correct outputs, even for a single test image.
- Even if ground truth were available, it is not clear how to quantify the similarity of a rendered image to ground truth. Some partial solutions appear in the literature; their shortcomings will be described later.

While automated scoring is not possible, evaluation can be made more systematic by using standard image sets. At present, few standard benchmark datasets are available. Mould and Rosin [8] proposed a general set for NPR, and Rosin et al. [9] presented a set for portraits. This paper extends the latter with a refined and extended image set for evaluating portrait stylisation.

There is a huge amount of portrait photography, from formal portraiture to selfies; social media has fueled demand for personalised portraits. The surge in portrait stylisation methods brings a need for more portrait benchmarking resources. This paper seeks to improve evaluation methodology for portrait stylisation, and makes the following three specific contributions:

- We present a new structured, three-level, benchmark dataset for the evaluation of stylised portrait images. Compared to previous benchmarks, more rigorous criteria were used for its construction. User studies supplied annotations.
- We give a new methodology for evaluating portrait stylisation algorithms, making use of the different benchmark levels and annotations.
- We evaluate a wide variety of NPR methods (both portrait-specific and general) using the new benchmark dataset.

This paper follows on from the previous conference version by Rosin et al. [9], and makes substantial changes to the earlier *NPRportrait 0.1*. Overall, less than a third of *NPRportrait 1.0* consists of images from *NPRportrait 0.1*[①]. The major differences are that:

- More rigorous criteria for image selection were used compared to version 0.1. This particularly affected level 1, which was entirely replaced by a better-controlled image set.
- Images are now more rigorously checked against the design matrix requirements by running user studies for validation.
- We extended the benchmark to a third level to provide more challenging test images.
- The set of NPR algorithms that have been systematically evaluated has been expanded to include another six styles from the literature, ensuring that they cover: (i) both portrait-specific and general purpose methods, (ii) both traditional NPR and NST methods, (iii) stylisation of both texture and geometry, and (iv) colour as well as black and white stylisations.
- A new set of experimental procedures is defined, and used to quantitatively evaluate the NPR algorithms, whereas the previous conference version only carried out informal evaluation. Specifically, (i) the correctness of perceived facial characteristics is tested for stylisations (making use of the benchmark annotations), and (ii) the quality of the NPR algorithms' outputs are checked for trends across the benchmark levels.

The benchmark data (images and annotations) are available to the research community at `https://users.cs.cf.ac.uk/Paul.Rosin/NPRportraitV1/`, and provide a framework for others to use and to extend.

① The benchmark released in Ref. [9] was presented at the time as a basic "version 0.1", with the intention of performing user studies and extending the number of levels.

2 Related work

Two critical elements in benchmarking are the datasets and the evaluation of the results.

2.1 Benchmark datasets

CVonline [10] lists 1170 unique computer vision datasets, that collectively (i) incorporate both data and annotations (e.g., class labels, segmentations), (ii) cover many areas (e.g., medicine, agriculture), and (iii) range from high-level applications (e.g., detection of various medical conditions), to specific low-level tasks (e.g., image registration). Over time these benchmark datasets have become increasingly large, especially recently so as to facilitate machine learning.

The situation in NPR is very different. Until recently, there were no benchmark datasets; Kyprianidis et al.'s [1] comprehensive overview does not mention benchmarks, suggesting that they were not part of the prevailing mindset. Although various images were occasionally reused as test cases in the community, these were few in number, and were typically limited to specific styles (e.g., the Peperomia plant for stippling). Mould and Rosin [8] created *NPRgeneral*, designed to provide images for general NPR. It contains 20 images selected to include a variety of attributes and content, such as irregular texture, vivid or muted colors, and long gradients. Images were selected manually, although some low-level image measures (e.g., colourfulness and sharpness) provided guidance. The authors identified that some specific images were generally challenging, suggesting a suitable direction for future research. Other groups of images were found to be very difficult for certain categories of algorithm, but not others, indicating how existing methods can be best deployed according to the expected nature of the input.

Kumar et al. [11] recently produced an NPR benchmark that closely follows the principles of *NPRgeneral*. Its 32 images were intended to augment *NPRgeneral* with more variety and complexity.

The more specialized benchmark dataset *NPRportrait 0.1* was released by Rosin et al. [9]. It contains portrait images, split into two levels of difficulty, each consisting of 20 images. Its first level contained highly constrained portraits: closely-cropped frontal views of faces with simple backgrounds. Its second level relaxed the constraints on pose, lighting, and background, while introducing complications of facial hair and varied expressions. Six NPR algorithms (both portrait-specific and general) were applied to the benchmark dataset. All methods worked reasonably well on level one, and the domain knowledge used by the face-specific methods enabled them to improve the quality of their stylisations, e.g., preserving elements such as eyes. At level two, the performance of the portrait-specific algorithms declined for some images with more complex contents; however, the general-purpose algorithms were equally effective across both levels. *NPRportrait 0.1* took a systematic approach to selecting images, using a design matrix; the new dataset *NPRportrait 1.0* follows that methodology, as described in Section 3.

Following a design matrix ensures that a balanced dataset is created. The issue of data bias has come to the fore in recent years, particularly for race and gender [12]. Although the focus is normally on training data, so as to avoid biased models, here we are more interested in test data, so that any biases in NPR methods can be detected, whether the method uses machine learning or not.

To date, these benchmark datasets have been used in a variety of ways: to include some stylisation results from examples taken from the benchmark [4, 13–16]; to systematically test the performance of stylisation algorithms [17–19]; to provide appropriate test data as part of the optimisation of preset parameters for post-processing filters in BeCasso, an interactive mobile iOS app for image stylisation [20]; and to provide a competitive and common set of test images for a research course on image processing for mobile applications [21].

2.2 Image quality assessment

Evaluating NPR outputs involves the aesthetic qualities of pictures, which is subjective and hard to quantify. Even were a ground-truth stylised image to be available, neither low-level image comparison measures such as MSE, PSNR, or SSIM [22] nor more recent deep learning approaches such as LPIPS [23] suffice. Low-level methods fail to capture important perceptual and aesthetic aspects; while deep learning does better, such methods do not always follow human judgements [24], are prone to overfitting, lack robustness [25], and have not been trained on stylised images.

Making matters worse, ground truth images are

likely to be unavailable. Blind image quality assessments exist (e.g., DIIVINE [26], BRISQUE [27]), and more recently, "opinion-unaware" methods (e.g., IL-NIQE [28]) have appeared, avoiding the need for human subjective scores. However, they are neither developed for nor applicable to evaluating stylisation.

To cope with the lack of ground-truth before-and-after stylisation images, NST researchers have used the Fréchet inception distance (FID) [29], which compares the distributions of two unpaired sets of images (stylised and unstylised). FID has some limitations; it assumes that features have Gaussian distributions, the estimator of FID has a strong bias even for up to 10,000 samples [30], and is also not trained on stylised images. Moreover, it requires a set of ideal images in the target style, which may not be available.

2.3 Alternative approaches to NPR evaluation

The difficulties of performance evaluation in NPR have been identified and discussed thoroughly in the NPR community [7, 31]. A common practice in NPR is to employ proxy metrics [32] in place of directly evaluating the aesthetics of the stylised image: easily quantifiable measures, such as performance on a memory task, could be collected. Unfortunately, the proxy measure may not be related to the quality or aesthetics of the image stylisation.

Mould [33] noted that many tasks in NPR have neither clearly defined success criteria nor ground truth, hampering both automated evaluation and evaluation with user studies. He suggested that the author should identify important characteristics of interest, and use these to inform a transparent and structured visual analysis of the results. This evaluation strategy does not scale up well, but it can be considered as a fallback position.

User studies are a popular alternative means of evaluation, and have the strong advantage that they have the potential to capture all aspects of human perception including semantics, aesthetics, and art history. They are a popular tool in the NST community. However, the traditional NPR community has reservations about their effectiveness [7, 31–33]. Issues abound: participants may guess the hypothesis, and provide biased responses; participants may be careless or may insufficiently understand the task, formulating suitable questions or tasks is difficult, it is infeasible to verify a user study's results other than by re-running the study, and finally, it is difficult to compare results from separate user studies. For example, participants may assess renderings based less on aesthetics and style elements than on source image content. Indeed, "aesthetics" are not defined, with different participants using different criteria.

2.4 Portraiture in NPR

Since the early days of NPR there has been particular interest in generating portraits, from simple line drawings [34] a quarter of a century ago, to modern methods that combine deep learning with a dataset of artists' portraits to enable stylisation of both geometry and texture [35]. However, portraits are one of the most challenging tasks for stylisation algorithms. Kyprianidis et al.'s [1] assessment from 2013 still holds true: "Portraits are an example of subject matter currently rendered poorly by general purpose stylization algorithms since they are particularly sensitive to distortion or detail loss in facial regions." We refer the reader to Zhao and Zhu's work [36] for an overview of portrait-specific NPR methods prior to deep learning, and to Yaniv et al.'s paper [35] for references to more recent methods. In this section we very briefly outline the 11 NPR algorithms which are evaluated in Section 4. More detailed overviews of these algorithms are provided in the Electronic Supplementary Material (ESM).

- Li and Wand's method [37] combines convolutional neural networks and Markov random fields.
- Berger et al. [38] mimic the style of specific artists' line-drawings in a data-driven manner, drawing strokes following the drawing statistics from an artist's stroke database.
- APDrawingGAN by Yi et al. [39] uses a hierarchical system of generative adversarial networks (GANs) along with a line-promoting distance transform loss.
- Rosin and Lai's algorithm [40] stylises the image with abstracted regions of flat colours plus black and white lines, adding skin shading and enhancing facial parts. A modified version of this pipeline renders a more abstract version inspired by the artist Julian Opie.
- Winnemöller et al.'s XDoG filter [41] can be conceptualised as the weighted sum of a blurred source image and a scaled difference-of-Gaussians

(DoG) response of the same image, effectively applying unsharp masking to the DoG response.
- Rosin and Lai [42] create an engraving style rendering using a dither matrix, which is a spatially-varying threshold.
- Son et al. [43] propose a novel method in which dots and hatching lines with varying sizes are regularly spaced along local feature orientations.
- Semmo et al.'s [44] oil paint filter is based on non-linear image smoothing. The method uses Gaussian-based filter kernels aligned to the main feature contours of an image for structure-adaptive filtering.
- Doyle et al.'s [45] pebble mosaic stylisation obtains a superpixel segmentation of the image, and then converts each superpixel into a 2.5D pebble.
- Rosin and Lai [15] apply filtering to generate a watercolour stylisation, incorporating steps such as morphological opening and closing, and local histogram equalisation.

3 Methodology

3.1 Basis

We have constructed a benchmark with three levels of increasing difficulty. The first level contains straightforward images: unoccluded faces with neutral expressions and simple backgrounds. These restrictions are common to many existing portrait stylisation methods. The second level increases the challenge by introducing complications such as facial hair and non-neutral expressions. The third level increases the difficulty even further by relaxing restrictions on lighting complexity, gaze direction, and distractions such as tattoos. All images have been annotated with a number of characteristics which were obtained through a user study.

Overall, our benchmark construction principles follow those of *NPRportrait 0.1*. The key considerations are as follows.

Challenging images: The benchmark needs to include images that are likely to be challenging for stylisation methods. Revealing weaknesses in the state of the art helps drive research progress.

Range of difficulty: The benchmark should include images covering a range of levels of difficulty, so as to better assess the performance of NPR algorithms, i.e., showing when they work, and when they fail. A benchmark that is too difficult will discourage users, limiting community uptake. Furthermore, some algorithms are designed for certain types of input (e.g., frontal faces, uncluttered background). The first level should be attainable by the majority of existing methods.

Small number of images: The portrait benchmark should be as small as possible while still having enough variety to be representative. Three main factors explain the need for a small dataset. First, the benchmark is intended for the image stylisation community, where manual evaluation (e.g., via user studies) is commonplace; evaluating a larger benchmark would require more manual effort. Second, a small benchmark makes comparisons easier, with a common set of images stylised by different algorithms; this is already done informally with images such as Lena, but can be formalised with a standard benchmark. Third, a small benchmark encourages authors to play fair by showing the entire benchmark, rather than presenting images for which their algorithm performs particularly well; this feature would be undercut by a larger benchmark. We elaborate below.

Large benchmark datasets are used elsewhere in computer vision, where ground-truth measurements can summarize performance in a single number. Conversely, in stylisation tasks, it is both common and useful to examine individual results. Evaluations are often manual, whether less structured (just showing results) or more structured (careful discussion of individual images, or conducting a user study). An evaluation protocol with humans in the loop is time-consuming, and the fewer evaluations required, the better.

If the dataset is too large, then researchers will select subsets. Since different researchers would make different selections, the results across different papers would not be comparable, destroying part of the benefit of a common benchmark. Not only that, but it becomes possible for researchers to select non-representative results. These dangers can be reduced by creating a dataset sufficiently small that it can be treated in its entirety.

The above seems to argue that the benchmark should be very small, perhaps only a single image. However, the benchmark should also cover the target domain thoroughly, sampling widely over

potential input images. The tension between these two considerations led to the choice of 20 images per level for both *NPRgeneral* and *NPRportrait 0.1*, balancing the desire for a small benchmark with the need to show varied content. We also use 20 images for *NPRportrait 1.0*.

Representative: Ideally the benchmark should be representative of the population, i.e., balanced in terms of perceptually significant facial characteristics such as gender and ethnicity. This will make the benchmark more useful, as it will ensure that algorithms have been tested on the likely characteristics of the input data when the algorithms are deployed.

Facial characteristics: Each level of the benchmark is built according to a design matrix, where the set of images was chosen to ensure diversity among several high-level dimensions describing possible faces. The dimensions vary per difficulty level; for example, while we enforce neutral facial expressions at level 1, level 2 includes variations in facial expression. Level 1 is intended to show a broad spectrum of different faces with tight constraints on pose, lighting, and visibility so as to make the images straightforward to stylise. Conversely, levels beyond 1 include complications that make stylisations more difficult.

At level 1, we use the characteristics of *gender*, *age*, *attractiveness*, and *ethnicity*. Each characteristic was quantized into discrete categories. Some of these characteristics have the drawback that the categories may not have precise boundaries. Furthermore, participants in user studies will be influenced by their cultural backgrounds, as well as having other biases, in assessing the characteristics. Nevertheless, we perceive benefits in using high-level sociological characteristics over alternative low-level features (e.g., smoothness, angularity). Humans have specialised mechanisms for the visual processing of faces [46], and from infancy develop mechanisms for judging high-level properties such as gender, ethnicity, and attractiveness. In addition, low-level characteristics are not independently distributed across faces; rather, they may be bundled in ways that correspond to high-level groupings.

Gender, age, and ethnicity influence how we perceive and remember faces. Studies of efforts to describe unfamiliar individuals [47, 48] (in the context of eyewitness reports) found a high prevalence of gender, age, and ethnicity as descriptive terms; other frequently-appearing characteristics such as height and build are not readily discernible from portraits. Magnetoencephalography (MEG) studies [49] on neural responses to faces indicate that gender, age and identity are determined within a fraction of a second, and that gender and age information emerge even before identity information.

Gender, age, and ethnicity can be considered basic features to describe faces. To this set of features, we added *attractiveness* to ensure a wider range of potential faces; many available photographs portray models or celebrities, which are not representative of the population more generally.

The set of facial characteristics for each level is used to direct the construction of a design matrix governing the distribution of characteristics within that level: see Section 3.2. A design matrix provides a formal mechanism for ensuring both diversity and balance among the selected characteristics.

Since the image sources will typically not provide all the above facial characteristics, and moreover some are inherently ambiguous or subjective, they will be acquired separately through user studies.

The gap between levels: The difficulty gap between level n and level $n + 1$ should not be too great since we desire fine granularity of what conditions cause algorithms to fail. However, a large number of levels would make the benchmark unwieldy. *NPRportrait 0.1* provided two levels, with the potential for more in the future. *NPRportrait 1.0* extends this to three levels which enables it to include more varied, and therefore more challenging, content. However, there remains scope for further levels which cover both more complicated scenes (e.g., multiple people, full bodies, heavy clutter, extreme poses and expressions) and broader coverage of portrait subjects (e.g., children, the elderly, more ethnicities).

Variety of image sources: The images should come from a wide variety of sources so as to ensure that a variety of cameras, lighting conditions, backgrounds, poses, and varied levels of professionalism of the photographers and the subjects are included. We deliberately decided against creating our own photographs, and instead selected images from existing image collections such as Flickr, making the images diverse in the senses mentioned here.

Image resolution: Most NPR algorithms are

suitable for medium resolution images, and so all images have a fixed height of 1024. This also simplifies running some NPR algorithms as they may have scale parameters that can therefore be held constant across the dataset[①].

Copyright clearance: Since (manual) visual evaluation of results remains important, the benchmark images should not have any copyright restrictions that would prevent them from being published along with the derived results. We drew our images from a variety of sources and they have a variety of copyright terms, with the majority having some type of Creative Commons license (`https://creativecommons.org/licenses/`) that allows reuse and modification. Some images are in the public domain. Details about the terms for specific images can be found in the ESM.

3.2 Design matrix

For each benchmark level, a set of desired characteristics is defined that all the images should have (e.g., be a frontal view). Another set of desired characteristics should vary (e.g., subjects' gender, ethnicity, expression); each of these is constrained to a set of categories (e.g., {young adult, middle-aged adult}). With 20 images in a level, it is not possible to cover all combinations of characteristics. Therefore, rather than a full factorial design, a subset of the possible combinations is sampled in order to create a reasonably balanced design. We have used the optFederov function from the R package AlgDesign [50] with the common, default criterion of D-optimality [51], which seeks to maximize the determinant of the information matrix $|X^{\mathrm{T}}X|$ of the design X, and as a result maximizes the information from the designed experiments. Starting from a random selection of 20 combinations of desired characteristics from the factorial design, the Federov algorithm [52] is applied to iteratively exchange selected and unselected combinations for optimization. Five random initializations are attempted so as to find a solution closer to the global optimum.

3.3 Level 1

Level 1 is intended to be straightforward to stylise, and thus we impose many restrictions. Each image should contain a frontal, approximately upright, unoccluded view of a single face with a forward gaze. The images must contain minimal background objects or clutter, providing a clear separation of the face from the background. The backgrounds should be homogeneous, but natural—not manually masked out. The face should dominate the image, filling most of it, and be cropped approximately at the neck. Other body parts such as the hands are to be excluded. The subject in the portrait should not have facial hair or long hair that partly covers the face, and should omit jewellery or other accessories such as a pipe, glasses, or hat. Harsh or complex lighting is to be avoided; only soft lighting is permitted. All subjects should have approximately neutral expressions.

NPRportrait 0.1 included *face shape* as a variable characteristic, identified using the descriptors {round, square, oval, heart, long}. At the time it was noted that these were not strictly defined, with some attributions of face shape only approximate. Here, image characteristics are validated by user studies, and preliminary tests suggested that face shape could not be reliably determined; hence, we excluded face shape from the current benchmark.

Another change from the previous benchmark is that ethnicity has been expanded from three to four categories, with Asian being split into East Asian (e.g., Chinese) and South Asian (e.g., Indian). Of course, with the large number of ethnic groups in the world [53], and the small size of the dataset, coverage is necessarily incomplete.

The remaining variable characteristics are unchanged from version 0.1: gender, age, and attractiveness. There are two categories for gender, {male and female}, and for age, {young adult, middle-aged adult}. Finally, we specified three levels of attractiveness: {below average, average, above average}. It is important to control attractiveness since there is a tendency in the NPR literature to use aesthetically pleasing images with attractive and/or interesting faces. However, stylisation should also be effective for unattractive or ordinary faces.

3.4 Level 2

The criteria and design matrix for level 2 are unchanged from *NPRportrait 0.1*. Level 2 retains many of the restrictions enforced in level 1: each image contains a frontal, approximately upright, unoccluded view of a single face that fills most of

① However, future NPR benchmarks should revisit the issue of image resolution. Many commercial stylisation applicationss need to operate on images of arbitrary size. Moreover, they typically provide a lower-resolution preview (e.g., when changing interactive settings). Thus stylisation algorithms should ideally be resolution-independent.

the image. The background should be relatively plain, but we relax this requirement slightly: some unobtrusive background content can be present. The requirement for unadorned faces is also relaxed, and so some jewellery is allowed. Likewise, level 1's requirement for moderate lighting is relaxed. Gaze direction should be mostly forwards, but need not be exclusively so. Ages are again restricted to adult, but age is not used as a control variable.

As in level 1, an equal distribution of gender is maintained. Facial expressions are distributed among the categories {negative, neutral, positive}, though extreme facial expressions were avoided in order to maintain reliability of fitting face models (used by some face-specific NPR algorithms). Level 2 also includes facial hair; we used the categories {none, moustache, beard, goatee, stubble}, and assumed that females had no facial hair.

In level 2, the design matrix does not control for age, attractiveness, or ethnicity. This was for practical reasons: with more control factors, it is difficult to source images that satisfy all constraints. However, we endeavoured to maintain a reasonable spread of these characteristics.

3.5 Level 3

Level 3 roughly maintains the previous criteria, but is not as strict. The cropping can be less tight, the pose can be less frontal, and background clutter can be more prominent. Several other factors are modified systematically in the design matrix. A variety of lighting effects are used, categorised as {simple, complex}: "simple" indicates soft frontal lighting as used in the previous two levels, while "complex" encompasses anything else such as side lighting, back lighting, strong lighting, strong shadows, or unusual lighting effects. We employ four categories of expression: {regular, extreme, odd, eyes}: "extreme" indicates an exaggerated expression, "odd" indicates an unusual expression such as pouting, grimacing, an open mouth, etc., and "eyes" indicates that eyes are not open and forward facing as before. The final variations concern either additions to or occlusions of the face: skin markings include scars, strong makeup, or strong specularities, while occlusions are caused by objects such as jewellery, glasses, or hands.

Even though level 3 criteria are similar to those of level 2, level 3 does provide a greater challenge for stylisation algorithms as demonstrated by the outcome of Experiment 2. As shown in Fig. 4, most of the worst rated results (28 out of 33) came from level 3.

3.6 Image selection

The design matrices for levels 1, 2, and 3 are shown in Tables 1, 2, and 3, respectively. The next step was to acquire images that satisfy these design matrices, and are also consistent with our goals of using a variety of image sources and of course have copyright clearance and sufficient image resolution. As noted by Rosin et al. [9], the majority of photographs available online were taken under uncontrolled conditions, with complicated backgrounds, non-frontal view, occlusions, or other factors that make them unsuitable. Moreover, many do not provide sufficient or explicit copyright clearance.

3.6.1 Level 1

Since level 1 requires the most tightly controlled images, it required the most work. We gathered 540 photos from sources such as Wikimedia Commons, Flickr, and Unsplash, as well as photographs from the authors' own collections. A user study was carried out to collect the main characteristics of the faces that

Table 1 Design matrix for level 1. Note that an additional column for attractiveness was generated (and is used in the experiments), but for the benefit of the photographed subjects is not displayed in the paper

Gender	Age	Ethnicity
female	middle	black
female	young	black
male	middle	black
female	young	black
male	middle	black
male	young	black
male	middle	South Asian
male	young	South Asian
female	young	South Asian
female	middle	South Asian
female	middle	East Asian
male	middle	East Asian
female	middle	East Asian
male	young	East Asian
female	young	East Asian
male	young	East Asian
female	young	white
male	young	white
male	middle	white
female	middle	white

Table 2 Design matrix for level 2

Gender	Expression	Facial hair
male	negative	none
male	neutral	none
female	neutral	—
female	positive	—
male	negative	moustache
female	neutral	—
male	positive	moustache
female	positive	—
male	negative	beard
female	negative	—
male	neutral	beard
female	positive	—
female	negative	—
male	neutral	goatee
female	neutral	—
male	positive	goatee
female	negative	—
male	neutral	stubble
female	neutral	—
male	positive	stubble

Table 3 Design matrix for level 3

Gender	Lighting	Expression/eyes	Skin/occlusion
male	complex	extreme	skin marking
female	complex	extreme	skin marking
female	complex	regular	skin marking
male	simple	regular	skin marking
male	complex	odd	skin marking
male	simple	eyes	skin marking
female	simple	eyes	skin marking
female	simple	extreme	occlusion
male	complex	extreme	occlusion
female	complex	regular	occlusion
male	simple	odd	occlusion
female	simple	odd	occlusion
male	complex	eyes	occlusion
female	complex	eyes	occlusion
male	simple	extreme	regular
female	simple	regular	regular
male	complex	regular	regular
male	complex	odd	regular
female	complex	odd	regular
female	complex	eyes	regular

appear in the design matrix: age, attractiveness, and ethnicity. We recruited 260 participants (105 females and 155 males, age 16–78, $\mu = 33.96$, $\sigma = 13.70$) who were self-selected volunteers with diverse backgrounds who responded to a call circulated to the authors' contacts in various places around the world. They used a web-based application on their own platform to participate in the user study. All the subsequent user studies were also conducted with versions of this web-based application. Each of our 260 participants was shown a subset of the images (in practice 49), allowing them to complete the task without undue burden. They were presented each image in turn and asked to rate it for: age on a four-point scale, attractiveness on a three-point scale, and ethnicity from a list of five categories.

Users were given a choice of four categories for the question about age, even though we only use portraits from two age groups. These groups were bracketed above and below by the categories *child* and *old* so that we could reject unsuitable images. We also excluded images that lacked a clear consensus label. Even though age is in principle well-defined, we wanted to assess all characteristics from the appearances in the images.

The most contentious characteristic is attractiveness: the perception of attractiveness is very subjective, depending on many factors such as age and gender [54], ethnicity, cultural background, rural versus urban living [55], and even recent experience [56]. We assigned an attractiveness score to each face, calculated as the mean user judgement, where the user judgements of {below average, average, above average} are scored as {+1, 0, –1}.

To select faces for the three categories of {below average, average, above average} we approximate the distribution of attractiveness scores as normal with zero mean, and extracted the subsets corresponding to images that appear in the distribution in the ranges $[-\infty, -\alpha\sigma]$ and $[\alpha\sigma, \infty]$ as having below and above average attractiveness respectively, while images in the range $[-\beta\sigma, \beta\sigma]$ are treated as having average attractiveness. The categories can be made distinctive by setting large (or respectively, small) values for α (or respectively, β), and therefore discarding a large number of images in the intermediate ranges $[-\alpha\sigma, -\beta\sigma]$ and $[\beta\sigma, \alpha\sigma]$. However, choosing a large separation in α/β values needs to be moderated by the need to have sufficient images to fully populate the design matrix. To balance these needs, we set $\alpha = 1$ and $\beta = 1/2$.

For ethnicity, in addition to the four categories listed in Section 3, namely {South Asian, East Asian,

white, black}, users were allowed the additional category "other", which could be used to exclude portraits that do not clearly fit into the above four categories.

We did not include gender in this study as apparent gender can more reliably be assessed, and omitting it reduced demands on the users. Assessment of gender was done by the authors.

Those images that had less than 50% consistency in responses for age and ethnicity in the user study were excluded. This left a pool of 459 images from which we drew to fulfil the design matrix requirements.

3.6.2 Level 2

Since the design matrix for level 2 did not change, the images previously used in *NPRportrait 0.1* could be potentially retained. However, the characteristics of expressions are subtle, and so a second user study was carried out to determine if the perceived facial expressions were correct. Initial tests showed problems with some images, and so the full user study eventually included the 20 images from *NPRportrait 0.1* plus another 13 images. All 22 participants saw all 33 images. We replaced four of the original images with new images that more clearly display the appropriate expression (i.e., negative, neutral, or positive), as determined by the user study. In addition, one image was moved from a negative to a neutral expression.

3.6.3 Level 3

Since the characteristics of this level that need annotations are straightforward, we did not run a user study for such characteristics.

3.6.4 Full three-level benchmark

The full set of 60 images selected for the three levels of the *NPRportrait 1.0* benchmark is shown in Fig. 1. A further user study in which 56 participants (20 females and 36 males, age 21–79, $\mu = 33.45$, σ=12.53) were shown all 60 images was carried out to check the four characteristics of gender, age, attractiveness, and ethnicity. Not only did this confirm that the image labels were assigned correctly, but it gave us user responses to be used later in experimental evaluation of NPR stylisations.

3.7 Evaluation of stylisations

3.7.1 Experiments

Our benchmark allows researchers to use carefully chosen images to test their NPR algorithms, but as discussed in Section 2.2, carrying out the next step of evaluation is not straightforward. In the context of an application, stylisation may have some precise goal (e.g., mimicking an existing artist, or enabling a viewer to identify the rendered object quickly), which allows for a task-performance metric (e.g., the "deception score" is used to measure the fraction of stylised images classified by a VGG network as being artworks of the artist for which the stylization was produced [57]). However, in this paper we do not assume that such a goal is known (or even exists). To avoid the difficulty of directly comparing outputs of one algorithm against another algorithm, we formulate several experiments which are either based on the aesthetics from single stylisation algorithms, or else operate indirectly on the aesthetic aspects, using the four facial characteristics with which the benchmark dataset is annotated: gender, age, attractiveness, and ethnicity. It is better to ask users to make decisions about such characteristics rather than asking them to score the quality of a stylisation. Asking about stylisation quality involves making aesthetic judgements; not only is this difficult for users and subjective, but the task is often ill-defined given the multiple and interacting factors of content, style, and level of abstraction. In contrast, the four facial characteristics we use are familiar to all participants of the user studies, and they can make judgments with ease.

In summary, we propose two core experiments involving the full set of images in *NPRportrait 1.0* that can be used by researchers. Their general methodology is described below, while details and outcomes of running these experiments for the specific 11 stylisation algorithms covered in this paper are covered in Section 4.

Experiment 1: Correctness of facial characteristics. This experiment evaluates an NPR algorithm by measuring how the stylisation affects the four facial characteristics (gender, age, attractiveness, and ethnicity) captured in the user studies.

The estimates from the source images can be taken as a good approximation of ground truth, and it is expected that in most cases good stylisations would preserve these characteristics, although this may not hold for highly abstracted styles, or stylisations that aim to change characteristics (e.g., beautification). Since the responses in the user studies vary, a

Fig. 1 Images comprising levels 1, 2, and 3 of the *NPRportrait 1.0* benchmark.

distribution should be captured for each question, and an appropriate difference measure applied. One possibility is to compare characteristics with the earth mover's distance (EMD) for ordinal scales (gender, age, attractiveness) and L_1 distance for ethnicity. In addition to the traditional unsigned EMD, we use a signed version, computed by modifying Cha and Srihari's [58] Algorithm 1 to accumulate the signed prefix sum rather than the absolute prefix sum.

Experiment 2: Quality of stylisation across levels. This experiment checks the robustness of an NPR algorithm by directly looking at the quality (as determined by users' ratings or rankings) of its stylisations across the three benchmark levels.

One possibility would be to perform a user study involving a grouping task on the stylised photographs, but if the user studies were to be carried out remotely, then the benchmark contains too many images (60) to view simultaneously on a screen. Instead, ask users to view a triple of stylised images (all from the same NPR algorithm) and rank them according to the quality of the stylisations. The triples are generated

randomly, with one image from level 1, another from level 2, and one from level 3 (although the users are unaware of the three benchmark levels). Robustness is then measured by the correlations between the set of user rankings and the benchmark levels. Since the data is ordinal, it is appropriate to use Kendall's τ correlation. Restricting the elements of the triples to be drawn from different levels implies that their stylisations should be more distinct, and this has a double benefit. First, it makes the user's task easier, as trying to choose between similar quality stylisations is difficult and frustrating. Second, it makes the user study more efficient as the user can answer the questions more quickly and more reliably.

3.7.2 Validating facial characteristics across all levels

Experiment 1 involves analysing facial characteristics across all three levels in the benchmark. Therefore, since not all facial characteristics were carefully controlled at all levels in the benchmark, we need to check the consistency of the participants. For each image, the standard deviation of the user responses was calculated, and averaged over the 20 images in each level. This was done for gender, age, and attractiveness, which can be treated as numerical values, with each possible value in the user study mapped to $\mathbb{N}$. For example, attractiveness values {below average, average, above average} are mapped to {1, 2, 3}. For ethnicity, which is a nominal value, the index of dispersion was used instead[①]. Table 4 shows that gender and age have standard deviations below 0.5; that is, a clear majority of responses fall into the same category. The standard deviations for attractiveness are a little higher, unsurprisingly. The index of dispersion values range from zero (all ratings fall into the same category) to one (all ratings are equally divided between all the categories). Consider two examples. The image with highest ethnicity dispersion is the eighth image in level 3, with user assessments as follows: South Asian: 13, East Asian: 4, white: 3, black: 22, other: 14. The resulting dispersion score is 0.9, which reflects that the mode response (39%) was below an absolute majority. The image is challenging (as befits level 3): the figure in the portrait has closed eyes, exhibits a strong expression, and the lighting level is low.

Since level 1 is controlled for ethnicity, images with significant ambiguity of this characteristic should have been avoided. Indeed, level 1's largest dispersion score is 0.6, from the first image. Still, a majority of users agreed; the responses were: South Asian: 15, East Asian: 2, white: 0, black: 37, other: 2. Overall, as shown in Table 4, for all four face characteristics, in all but one case the variations increase with level. This reflects the increasing variation in the images (e.g., lighting, pose, occlusion, etc.) as the levels increase, and thus the difficulty gap between levels applies to human observers as well as stylisation algorithms.

4 Demonstration: Evaluating 11 NPR algorithms

This section demonstrates the use of *NPRportrait 1.0* to evaluate 11 NPR algorithms which cover a wide range of styles and methods: neural style transfer [37], XDoG [41], oil painting [44], pebble mosaic [45], artistic sketch method [38], APDrawingGAN [39], puppet style [40], engraving [42], hedcut [43], Julian Opie style [40], watercolour [15]. In addition, the results from analysing these stylisations allow us to confirm the requirement (detailed in Section 3) that the benchmark provides a clear range of difficulty across the three levels.

4.1 Experiment 1: Correctness of facial characteristics

We conducted Experiment 1 as described in Section 3.7; with 11 NPR algorithms and the full benchmark of 60 images, there are 660 stylised

Table 4 Variability of user judgements of face characteristics from source images in the *NPRportrait 1.0* benchmark; standard deviations for gender, age, and attractiveness, and the index of dispersion for ethnicity

Characteristic	Gender			Age			Attractiveness			Ethnicity		
Level	1	2	3	1	2	3	1	2	3	1	2	3
Variablity	0.070	0.069	0.087	0.459	0.464	0.486	0.563	0.580	0.603	0.188	0.220	0.301

① A version of the index of dispersion can be applied to nominal values, computed as $D = k(N^2 - \sum_c f_c^2)/[N^2(k-1)]$ where k = number of categories, N = number of samples, and f_c = frequency of category c.

photos. Note that none of these stylisation algorithms explicitly aims to modify the tested facial characteristics. Some stylised results for each algorithm are shown in Fig. 4; the full set of stylisations is shown in the ESM. The 225 participants (79 females and 146 males, aged 17–75, $\mu = 32.75$, σ=11.42) viewed randomly-selected subsets of 30 stylised images, so that each image was seen, on average, by 10 participants. For each image they selected the choice that best matched the image from the list of possible values for each of the face characteristics (gender, age, attractiveness, and ethnicity).

Tables 5–7 list the discrepancies in reported face characteristics for the stylised images, summed across the 20 images in each level, compared to the original portraits. Examples of discrepancy values for individual images are shown in Fig. 2. To aid interpretation, large values (i.e., above a threshold) are marked in red, and in Tables 5 and 7 the threshold is calculated for each facial characteristic as $\mu + \sigma$, where μ and σ are the mean and standard deviation of the 33 values in the table across the 3 levels and 11 NPR algorithms, respectively. For Table 6, μ and σ are computed from absolute versions of the signed distances, and the thresholds are also applied to the absolute versions of the signed distances so that both positive and negative discrepancies are treated equally.

Note that since not all images had the same number of user responses, the histograms are normalised to

Table 5 Evaluation of facial characteristics of 11 NPR algorithms. Discrepancies for gender, age, attractiveness are *signed* EMD distances; for age and attractiveness positive value indicates an increase in judged value after stylisation, while for gender it indicates increased likelihood of assignment as female rather than male. Larger absolute discrepancies are marked in red: gender $\geqslant 1.24$, age $\geqslant 5.48$, attractiveness $\geqslant 4.51$. Yellow highlights indicate significant differences between levels for an NPR method (ANOVA at 0.05 level)

Characteristic	Gender			Age			Attractiveness		
Level	1	2	3	1	2	3	1	2	3
neural style transfer [37]	0.55	1.15	2.36	7.01	9.19	10.10	−6.32	−8.71	−8.11
artistic sketch method [38]	0.17	−0.33	−1.32	0.04	5.05	7.96	−2.42	−3.62	−2.65
APDrawingGAN [39]	−0.45	0.16	0.02	0.79	3.85	7.12	−0.19	−1.49	−2.64
puppet style [40]	0.19	−0.24	0.55	−0.61	3.45	2.06	0.32	−1.29	−0.08
XDoG [41]	−0.29	−0.40	−0.51	2.44	2.42	−0.03	2.09	0.21	3.85
engraving [42]	−0.25	−0.05	0.34	−2.20	0.02	−0.88	1.37	−0.36	3.76
hedcut [43]	0.45	−0.41	1.27	0.24	1.59	2.50	−0.80	−1.58	0.88
oil painting [44]	−0.38	−0.34	0.52	−1.42	0.55	−0.79	4.25	2.06	2.86
Julian Opie style [40]	−1.68	−0.94	−2.76	−3.53	−2.79	−3.74	−2.90	−3.06	−0.44
pebble mosaic [45]	0.03	−0.77	0.73	0.26	2.45	−0.69	2.42	1.44	1.06
watercolour [15]	0.03	−0.24	0.31	−3.16	−2.91	−0.61	2.72	0.51	3.90

Table 6 Evaluation of facial characteristics of 11 NPR algorithms. Discrepancies for gender, age, and attractiveness are *unsigned* EMD distances. Larger discrepancies are marked in red: gender $\geqslant 2.37$, age $\geqslant 8.37$, attractiveness $\geqslant 7.08$. Yellow highlights in two cells in a row indicate significant differences between two levels for an NPR method (ANOVA at 0.05 level). Where significant differences occur between two pairs of levels for one characteristic, these are coloured as pairs of yellow and blue, with the overlap coloured as green

Characteristic	Gender			Age			Attractiveness		
Level	1	2	3	1	2	3	1	2	3
neural style transfer [37]	1.49	2.02	3.53	8.90	11.03	11.23	8.49	10.18	8.58
artistic sketch method [38]	2.21	2.00	4.82	7.57	8.70	11.72	6.94	6.01	6.29
APDrawingGAN [39]	0.97	0.55	2.06	5.50	6.13	9.07	3.81	4.90	7.10
puppet style [40]	0.59	0.73	1.13	6.19	4.74	7.51	5.33	5.06	4.69
XDoG [41]	0.90	0.76	0.99	5.11	5.05	5.02	5.39	4.45	6.25
engraving [42]	0.63	0.61	0.74	4.17	4.34	4.27	4.98	5.29	5.32
hedcut [43]	1.03	1.24	1.45	5.63	4.19	6.31	4.37	4.78	4.79
oil painting [44]	0.65	0.52	0.96	4.23	3.95	3.05	5.32	4.48	4.37
Julian Opie style [40]	1.97	1.09	3.91	6.37	5.88	7.84	6.64	6.16	7.43
pebble mosaic [45]	0.51	1.07	1.81	5.19	4.75	5.87	4.42	5.25	5.98
watercolour [15]	0.49	0.66	0.86	5.49	3.40	4.74	4.62	4.75	5.70

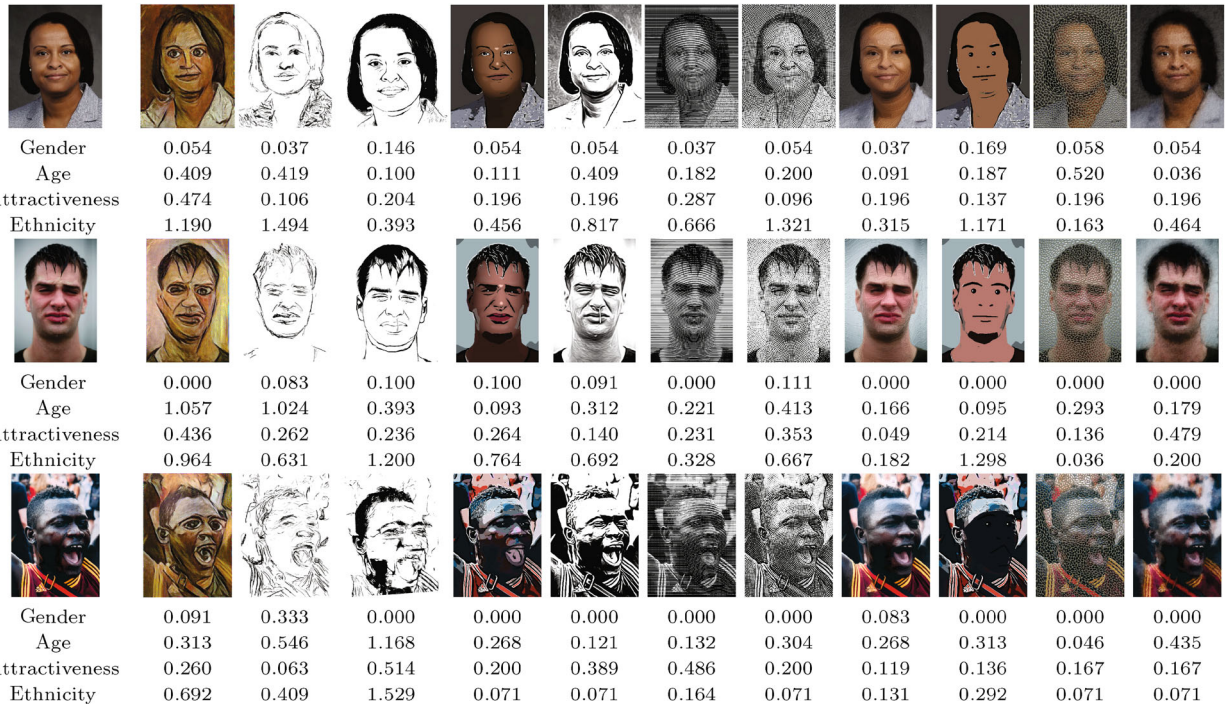


Fig. 2 For the first image at each of the three levels, discrepancies (unsigned distances) for the perceived facial characteristics obtained from the user study are shown for each of the stylisations.

Table 7 Evaluation of facial characteristic of 11 NPR algorithms: ethnicity. Discrepancies are measured using the L_1 distance; larger discrepancies (ethnicity $\geqslant 16.25$) are marked in red. Yellow highlights in two cells in a row indicate significant differences between two levels for an NPR method (ANOVA at 0.05 level). Where significant differences occur between two pairs of levels in a row, these are coloured as pairs of yellow and blue, with the overlap coloured as green

Characteristic	Ethnicity		
Level	1	2	3
neural style transfer [37]	20.93	16.10	17.15
artistic sketch method [38]	18.14	15.51	18.75
APDrawingGAN [39]	11.08	12.64	17.50
puppet style [40]	10.83	14.84	17.81
XDoG [41]	8.90	8.02	9.75
engraving [42]	6.37	6.54	7.88
hedcut [43]	9.20	8.48	9.58
oil painting [44]	4.74	4.51	7.58
Julian Opie style [40]	15.82	17.03	16.79
pebble mosaic [45]	6.10	5.91	12.66
watercolour [15]	5.77	6.22	5.66

unit area before computing distances. The signed EMD distances are useful in showing trends in the signs of differences. For instance, the neural style transfer [37] stylisation has a slight trend to make people look more feminine[①], older, and less attractive. On the other hand, the Julian Opie style [40] tends to make people look more masculine and a little younger.

Under the signed EMD distance, opposite sign movements (differences) cancel out, so it is useful to look at the unsigned EMD distances to check the overall discrepancy. Table 6 shows that both neural style transfer [37] and the artistic sketch method [38] produce renderings that differ substantially from the ground truth on all face characteristics. This is due to their highly stylised output, which incorporates geometric abstraction and distortion. Of course, this distortion is deliberate to match the target style. APDrawingGAN [39] is seen to be sensitive to the complexity of the input; its discrepancies are reasonably low for level 1, but double at level 3 for some characteristics. Table 7 shows that ethnicity is poorly recognised on outputs from the puppet style [40], which is due to low lighting levels causing the shading effect to make the faces dark. For instance, at level 3 the main discrepancies came from five such images which were unambiguously classified as white from the source portraits, but between 44% and 88% users classified the puppet style faces as black. Significant discrepancies were also made in determining ethnicity from the Julian Opie style [40], unsurprising given the high level of abstraction.

① The value of 2.36 for shift in gender at level 3 is mostly accounted for by five of the images that had movements of between one and three quarters of their distribution from male to female. Three of these images had a change in the majority gender compared to the ground truth.

We applied ANOVA tests to the signed and unsigned distances to check for significant differences between levels for each characteristic and stylisation. This allowed us to check the effects of increasing complexity of source images on the NPR algorithms. Both the artistic sketch method [38] and APDrawingGAN [39] show significant increases in the perceived age of the portraits as image complexity increases. This is probably due to the increased difficulty in generating clean renderings, and the increased number of spurious lines that appear in the renderings. Table 6 indicates that the perceived attractiveness of images stylised by APDrawingGAN [39] exhibits a consistently increasing divergence from the original photos across levels, and that this is statistically significant. Although the pebble mosaic stylisation [45] generally produces lower discrepancies for ethnicity than most of the other stylisations, we see a statistically significant increase in these discrepancies as the image complexity increases. This may be due to the constant colour mosaic boundaries, which effectively dilute skin tone, potentially causing confusion under challenging lighting conditions.

Table 8 summarises the findings in Tables 5–7. The "discrepancies" column indicates if an algorithm provoked discrepancies in judgement from the experimental subjects in Experiment 1 for any facial characteristic; the "levels" column indicates whether the discrepancies for an algorithm showed significant differences between the benchmark levels. These entries do not necessarily indicate that a stylisation method is "good" or "bad" since the table only captures a limited aspect of stylisation; rather, the table is beneficial in focusing on how a stylisation method can be further developed. It is interesting to note that a method such as the engraving style can be considered to be effective both in terms of accuracy of depicting facial characteristics, and its stability across the levels. Although engravings have a distinct style, they are capable of capturing tone and spatial detail if the engraving has sufficient resolution. However, this does not necessarily make it a preferred algorithm, as it also has aesthetic limitations.

Table 8 Summary of the performance of 11 NPR algorithms detailed in Tables 5–7. Column "discrepancies" indicates if an algorithm produced large discrepancies in Experiment 1 for any facial characteristic; "levels" indicates whether the discrepancies for an algorithm showed statistically significant differences between the benchmark levels

Method	Discrepancies	Levels
neural style transfer [37]	yes	
artistic sketch method [38]	yes	yes
APDrawingGAN [39]	yes	yes
puppet style [40]	yes	
XDoG [41]		yes
engraving [42]		
hedcut [43]	yes	yes
oil painting [44]		
Julian Opie style [40]	yes	
pebble mosaic [45]		yes
watercolour [15]		yes

Table 9 Correlation coefficients between triplet rankings and benchmark levels

Method	Kendall
neural style transfer [37]	0.363
artistic sketch method [38]	0.306
APDrawingGAN [39]	0.346
puppet style [40]	0.284
XDoG [41]	0.130
engraving [42]	0.154
hedcut [43]	0.202
oil painting [44]	−0.017
Julian Opie style [40]	0.266
pebble mosaic [45]	0.207
watercolour [15]	0.113

4.2 Experiment 2: Quality of stylisation across levels

Experiment 2 described in Section 3.7 was run on the same 11 NPR algorithms as Experiment 1. There were therefore $11 \times 20 \times 20 \times 20 = 88{,}000$ possible stylised triplets. The user study had 213 participants (113 females and 100 males, age 17–60, $\mu = 24.85$, $\sigma = 9.18$) who saw 30 triples of images, randomly generated with replacement, leading to 6390 triples, of which 6171 triplets were unique. Kendall's τ correlation is shown in Table 9; the low correlation values confirm that general-purpose filtering approaches such as XDoG [41] and watercolour [15] are least affected by the increasing complexity across the benchmark levels. Although they are face-specific, watercolour [15] and engraving [42] are also fairly robust since their renderings are not highly dependent on the face model, and their results are reasonable despite inaccurate face detection. The techniques with highest correlation to the levels are neural style transfer [37], which has a tendency to create more spurious facial features (e.g., misplaced eyes) as the images become

more cluttered; and both the line drawing methods (artistic sketch method [38] and APDrawingGAN [39]) which often produce fragmented or spurious lines when there are variations in lighting.

This user study can be further used to analyse both the benchmark and the NPR algorithms. The triplets were converted to a global ranking using Wauthier et al.'s [59] balanced rank estimation method, applied both (i) separately for each NPR algorithm, and also (ii) across all NPR algorithms, by aggregating the local scores for each benchmark image across the stylisations. Ranking the images in this way enables us to see which aspects of images lead to good stylisations either for a specific algorithm, or more generally across a range of algorithms. Figure 3 reveals that the top-ranked images, those more amenable to current stylisation algorithms, tend to be portraits with frontal views, fairly neutral expressions, good lighting, and plain backgrounds. Conversely, bottom-ranked images have one or more of the following characteristics: non-frontal views, strong expressions, patterns on the face, strong lighting effects, and cluttered backgrounds.

The top and bottom three ranked results for each of the 11 NPR methods are shown in Fig. 4. The lowest-ranked results contain a variety of artifacts, including messy rendering, segmentation errors, and rendering that does not clearly delineate facial components and structure. We note that the top and bottom ranked images in Fig. 3 appear in many of the top and bottom three rankings in Fig. 4 (7 and 6 out of 11, respectively). However, it is possible that, despite instructions to rate stylisation quality, the users' responses in Experiment 2 were biased by other factors. The images that were top and bottom ranked in Fig. 3 according to the overall quality of their stylisations also have the most and least attractiveness ratings for source images in the *NPRportrait 1.0* dataset. There is a moderate degree of correlation (0.4666) between the overall stylisation rankings and the source image attractiveness rating.

5 Conclusions and future work

Image stylisation is hampered by a lack of benchmark datasets and objective measures; most papers provide limited and rudimentary performance evaluation. This paper has presented a benchmark dataset for portrait stylisation, structured into three levels to provide clearly specified degrees of difficulty. The criteria for selecting images for each level are clearly specified, and were used to construct a design matrix. User studies were used to validate the suitability of each image with respect to the design matrix.

Alongside the new dataset a new methodology has been proposed for evaluating portrait stylisation algorithms. Rather than relying on aesthetic judgments, a challenging and ill-defined task, the user studies also incorporate more straightforward judgments, such as identification of gender or age.

The new benchmark and methodology enabled us to evaluate 11 NPR algorithms, both portrait-specific and general-purpose, and quantitatively compare them. We identified the most problematic images for each NPR algorithm; typical defects are inappropriate rendering of facial features, messy rendering, segmentation errors, and rendering that does not clearly delineate facial components and structure. Some image types are problematic for many state-of-the-art algorithms. Typically they

Fig. 3 *NPRportrait 1.0* benchmark ranked according to Experiment 2 aggregated over all 11 NPR styles.

Fig. 4 Images from *NPRportrait 1.0* benchmark stylised by the 11 NPR algorithms; the rows show (in order): neural style transfer [37], artistic sketch method [38], APDrawingGAN [39], puppet style [40], XDoG [41], engraving [42], hedcut [43], oil painting [44], Julian Opie style [40], pebble mosaic [45], watercolour [15]. The stylisations are ranked according to the outcomes of Experiment 2; for each method we show the top three results on the left and the bottom three on the right.

contained non-frontal views, strong expressions, patterns on the face, strong lighting effects, and cluttered backgrounds.

Identifying challenging cases will help direct future research. Further, there is scope for increasing the benchmark with additional levels, covering more complicated scenes as well as a broader range of portrait subjects. Possible complications include images with multiple people, full bodies, substantial occlusion, heavily cluttered background, extreme poses and expressions. Additional portrait subjects could include children, the elderly, and more ethnicities. In addition, more NPR benchmarks should be developed for different kinds of content. For example, landscapes, cityscapes, and animal portraiture have different requirements, and have evolved traditionally distinctive depiction styles. Finally, whereas curating images is relatively tractable, capturing the perceptual and artistic aspects of stylisations in an evaluation measure is challenging. Further exploration of evaluation methods is a key area for future work in NPR.

Declaration of competing interest

The authors have no competing interests to declare that are relevant to the content of this article.

Electronic Supplementary Material

Supplementary materials including

- fuller descriptions of the evaluated NPR algorithms,
- the complete set of results of the application of the 11 NPR algorithms to NPRportrait 1.0,
- examples of some additional stylisations of NPRportrait 1.0 images, and
- photo credits and copyright information

are available in the online version of this article at `https://doi.org/10.1007/s41095-021-0255-3`.

References

[1] Kyprianidis, J. E.; Collomosse, J.; Wang, T. H.; Isenberg, T. State of the "art": A taxonomy of artistic stylization techniques for images and video. *IEEE Transactions on Visualization and Computer Graphics* Vol. 19, No. 5, 866–885, 2013.

[2] Rosin, P.; Collomosse, J. *Image and Video-Based Artistic Stylisation*. London: Springer London, 2013.

[3] Gatys, L. A.; Ecker, A. S.; Bethge, M. Image style transfer using convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2414–2423, 2016.

[4] Jing, Y. C.; Yang, Y. Z.; Feng, Z. L.; Ye, J. W.; Yu, Y. Z.; Song, M. L. Neural style transfer: A review. *IEEE Transactions on Visualization and Computer Graphics* Vol. 26, No. 11, 3365–3385, 2020.

[5] Semmo, A.; Isenberg, T.; Döllner, J. Neural style transfer: A paradigm shift for image-based artistic rendering? In: Proceedings of the Symposium on Non-Photorealistic Animation and Rendering, Article No. 5, 2017.

[6] Gooch, A. A.; Long, J.; Ji, L.; Estey, A.; Gooch, B. S. Viewing progress in non-photorealistic rendering through Heinlein's lens. In: Proceedings of the 8th International Symposium on Non-Photorealistic Animation and Rendering, 165–171, 2010.

[7] Hall, P.; Lehmann, A.-S. Don't measure—Appreciate! NPR seen through the prism of art history. In: *Image and Video-Based Artistic Stylisation. Computational Imaging and Vision, Vol. 42*. Rosin, P.; Collomosse, J. Eds. Springe London, 333–351, 2013.

[8] Mould, D.; Rosin, P. L. Developing and applying a benchmark for evaluating image stylization. *Computers & Graphics* Vol. 67, 58–76, 2017.

[9] Rosin, P. L.; Mould, D.; Berger, I.; Collomosse, J.; Lai, Y.; Li, C.; Li, H.; Shamir, A.; Wand, M.; Wang, T.; et al. Benchmarking non-photorealistic rendering of portraits. In: Proceedings of the Symposium on Non-Photorealistic Animation and Rendering, Article No. 11, 2017.

[10] Fisher, R. B. CVonline. Available at http://homepages.inf.ed.ac.uk/rbf/CVonline.

[11] Kumar, M. P. P.; Poornima, B.; Nagendraswamy, H. S.; Manjunath, C. A comprehensive survey on non-photorealistic rendering and benchmark developments for image abstraction and stylization. *Iran Journal of Computer Science* Vol. 2, No. 3, 131–165, 2019.

[12] Buolamwini, J.; Gebru, T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In: Proceedings of the Conference on Fairness, Accountability and Transparency, 77–91, 2018.

[13] Azami, R.; Mould, D. Detail and color enhancement in photo stylization. In: Proceedings of the Symposium on Computational Aesthetics, Article No. 5, 2017.

[14] Du, L. How much deep learning does neural style transfer really need? An ablation study. In: Proceedings of the IEEE Winter Conference on Applications of Computer Vision, 3139–3148, 2020.

[15] Rosin, P. L.; Lai, Y. K. Watercolour rendering of portraits. In: *Image and Video Technology. Lecture Notes in Computer Science, Vol. 10799*. Satoh, S. Ed. Springer Cham, 268–282, 2018.

[16] Wu, T.; Chen, X.; Lu, L. Q. Field coupling-based image filter for sand painting stylization. *Mathematical Problems in Engineering* Vol. 2018, 3670498, 2018.

[17] Low, P. E.; Wong, L. K.; See, J.; Ng, R. Pic2PolyArt: Transforming a photograph into polygon-based geometric art. *Signal Processing*: *Image Communication* Vol. 91, 116090, 2021.

[18] Meier, P.; Lohweg, V. Content representation for neural style transfer algorithms based on structural similarity. In: Proceedings of the Computational Intelligence Workshop, 2019.

[19] Shen, Q.; Zou, L.; Wang, F. J.; Huang, Z. J. A scale-adaptive color preservation neural style transfer method. In: Proceedings of the 5th International Conference on Mathematics and Artificial Intelligence, 5–9, 2020.

[20] Klingbeil, M.; Pasewaldt, S.; Semmo, A.; Döllner, J. Challenges in user experience design of image filtering apps. In: Proceedings of the SIGGRAPH Asia Mobile Graphics & Interactive Applications, Article No. 22, 2017.

[21] Trapp, M.; Pasewaldt, S.; Dürschmid, T.; Semmo, A.; Döllner, J. Teaching image-processing programming for mobile devices: A software development perspective. In: Proceedings of the Annual European Association for Computer Graphics Conference: Education Papers, 17–24, 2018.

[22] Wang, Z.; Bovik, A. C.; Sheikh, H. R.; Simoncelli, E. P. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing* Vol. 13, No. 4, 600–612, 2004.

[23] Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; Wang, O. The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 586–595, 2018.

[24] Zamir, S. W.; Vazquez-Corral, J.; Bertalmío, M. Vision models for wide color gamut imaging in cinema. *IEEE Transactions on Pattern Analysis and Machine Intelligence* Vol. 43, No. 5, 1777–1790, 2019.

[25] Kettunen, M.; Härkönen, E.; Lehtinen, J. E-LPIPS: Robust perceptual image similarity via random transformation ensembles. *arXiv preprint* arXiv:1906.03973, 2019.

[26] Moorthy, A. K.; Bovik, A. C. Blind image quality assessment: From natural scene statistics to perceptual quality. *IEEE Transactions on Image Processing* Vol. 20, No. 12, 3350–3364, 2011.

[27] Mittal, A.; Moorthy, A. K.; Bovik, A. C. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing* Vol. 21, No. 12, 4695–4708, 2012.

[28] Zhang, L.; Zhang, L.; Bovik, A. C. A feature-enriched completely blind image quality evaluator. *IEEE Transactions on Image Processing* Vol. 24, No. 8, 2579–2591, 2015.

[29] Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler,

B.; Hochreiter, S. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, 6629–6640, 2017.

[30] Bińkowski, M.; Sutherland, D. J.; Arbel, M.; Gretton, A. Demystifying MMD GANs. In: Proceedings of the 6th International Conference on Learning Representations, 2018.

[31] Isenberg, T. Evaluating and validating non-photorealistic and illustrative rendering. In: *Image and Video-Based Artistic Stylisation. Computational Imaging and Vision, Vol. 42*. Rosin, P.; Collomosse, J. Eds. Springer London, 311–331, 2013.

[32] Hertzmann, A. Non-Photorealistic Rendering and the science of art. In: Proceedings of the 8th International Symposium on Non-Photorealistic Animation and Rendering, 147–157, 2010.

[33] Mould, D. Authorial subjective evaluation of non-photorealistic images. In: Proceedings of the Workshop on Non-Photorealistic Animation and Rendering, 49–56, 2014.

[34] Li, Y. Z.; Kobatake, H. Extraction of facial sketch images and expression transformation based on FACS. In: Proceedings of the International Conference on Image Processing, 520–523, 1995.

[35] Yaniv, J.; Newman, Y.; Shamir, A. The face of art: Landmark detection and geometric style in portraits. *ACM Transactions on Graphics* Vol. 38, No. 4, Article No. 60, 2019.

[36] Zhao, M.; Zhu, S.-C. Artistic rendering of portraits. In: *Image and Video-Based Artistic Stylisation. Computational Imaging and Vision, Vol. 42*. Rosin, P.; Collomosse, J. Eds. Springer London, 237–253, 2013.

[37] Li, C.; Wand, M. Combining Markov random fields and convolutional neural networks for image synthesis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2479–2486, 2016.

[38] Berger, I.; Shamir, A.; Mahler, M.; Carter, E.; Hodgins, J. Style and abstraction in portrait sketching. *ACM Transactions on Graphics* Vol. 32, No. 4, Article No. 55, 2013.

[39] Yi, R.; Liu, Y. J.; Lai, Y. K.; Rosin, P. L. APDrawingGAN: Generating artistic portrait drawings from face photos with hierarchical GANs. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 10735–10744, 2019.

[40] Rosin, P. L.; Lai, Y.-K. Non-photorealistic rendering of portraits. In: Proceedings of the workshop on Computational Aesthetics, 159–170, 2015.

[41] Winnemöller, H.; Kyprianidis, J. E.; Olsen, S. C. XDoG: An eXtended difference-of-Gaussians compendium including advanced image stylization. *Computers & Graphics* Vol. 36, No. 6, 740–753, 2012.

[42] Rosin, P. L.; Lai, Y. K. Image-based portrait engraving. *arXiv preprint* arXiv:2008.05336, 2020.

[43] Son, M.; Lee, Y. J.; Kang, H.; Lee, S. Structure grid for directional stippling. *Graphical Models* Vol. 73, No. 3, 74–87, 2011.

[44] Semmo, A.; Limberger, D.; Kyprianidis, J. E.; Döllner, J. Image stylization by interactive oil paint filtering. *Computers & Graphics* Vol. 55, 157–171, 2016.

[45] Doyle, L.; Anderson, F.; Choy, E.; Mould, D. Automated pebble mosaic stylization of images. *Computational Visual Media* Vol. 5, No. 1, 33–44, 2019.

[46] Bruce, V.; Young, A. *Face Perception*. Psychology Press, 2013.

[47] Van Koppen, P. J.; Lochun, S. K. Portraying perpetrators: The validity of offender descriptions by witnesses. *Law and Human Behavior* Vol. 21, No. 6, 661–685, 1997.

[48] Fahsing, I. A.; Ask, K.; Granhag, P. A. The man behind the mask: Accuracy and predictors of eyewitness offender descriptions. *Journal of Applied Psychology* Vol. 89, No. 4, 722–729, 2004.

[49] Dobs, K.; Isik, L.; Pantazis, D.; Kanwisher, N. How face perception unfolds over time. *Nature Communications* Vol. 10, No. 1, 1258, 2019.

[50] Wheeler, B. AlgDesign: Algorithmic experimental design. R package version 1.1-7. 2014. Available at https://cran.rproject.org/web/packages/AlgDesign/.

[51] Atkinson, A.; Donev, A.; Tobias, R. *Optimum Experimental Designs, with SAS, Volume 34*. Oxford University Press, 2007.

[52] Fedorov, V. *Theory of Optimal Experiments*. Academic Press, 1972.

[53] Doyle, R. Ethnic groups in the world. *Scientific American* Vol. 279, No. 3, 30, 1998.

[54] McLellan, B.; McKelvie, S. J. Effects of age and gender on perceived facial attractiveness. *Canadian Journal of Behavioural Science/Revue Canadienne des Sciences du Comportement* Vol. 25, No. 1, 135–142, 1993.

[55] Batres, C.; Kannan, M.; Perrett, D. I. Familiarity with own population's appearance influences facial preferences. *Human Nature* Vol. 28, No. 3, 344–354, 2017.

[56] Cooper, P. A.; Maurer, D. The influence of recent experience on perceptions of attractiveness. *Perception* Vol. 37, No. 8, 1216–1226, 2008.

[57] Sanakoyeu, A.; Kotovenko, D.; Lang, S.; Ommer, B. A style-aware content loss for real-time HD style transfer. In: *Computer Vision – ECCV 2018. Lecture Notes in Computer Science, Vol. 11212.* Ferrari, V.; Hebert M.; Sminchisescu C.; Weiss, Y. Eds. Springer Cham, 698–714, 2018.

[58] Cha, S. H.; Srihari, S. N. On measuring the distance between histograms. *Pattern Recognition* Vol. 35, No. 6, 1355–1370, 2002.

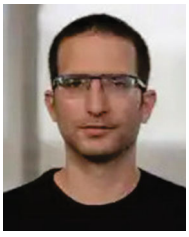
[59] Wauthier, F. L.; Jordan, M. I.; Jojic, N. Efficient ranking from pairwise comparisons. In: Proceedings of the 30th International Conference on Machine Learning, Vol. 28, III-109–III-117, 2013.

Paul L. Rosin is a professor in the School of Computer Science and Informatics, Cardiff University, UK. He received his Ph.D. degree from City University, London, in 1988. Previous posts were at Brunel University, UK; the Institute for Remote Sensing Applications, Joint Research Centre, Italy; and Curtin University of Technology, Australia. His research interests include low-level image processing, performance evaluation, shape analysis, facial analysis, medical image analysis, 3D mesh processing, cellular automata, non-photorealistic rendering, and cultural heritage.

Yu-Kun Lai is a professor in the School of Computer Science and Informatics, Cardiff University. He received his B.S. and Ph.D. degrees in computer science from Tsinghua University, China, in 2003 and 2008 respectively. His research interests include computer graphics, computer vision, geometric modeling, and image processing.

David Mould received his Ph.D. degree from the University of Toronto in 2002. Following a faculty appointment at the University of Saskatchewan, he became a professor at Carleton University, where he founded the Graphics, Imaging, and Games Lab. He is broadly interested in algorithmic creation of aesthetic objects, including images, music, 3D models, and computer-mediated experiences. His research centres on computer graphics and interactive systems, with particular emphasis on image stylisation, computer games, and procedural modeling.

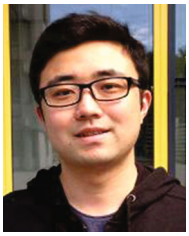
Ran Yi is a Ph.D. student in the Department of Computer Science and Technology, Tsinghua University, where she received her B.Eng. degree, in 2016. Her research interests include computational geometry, computer vision, and computer graphics.



Itamar Berger received his M.Sc. degree in computer science in 2012 from the Efi Arazi School of Computer Science at the Interdisciplinary Center in Israel, specializing in computer graphics, deep learning, and augmented reality.

Lars Doyle is a Ph.D. student in the School of Computer Science at Carleton University where he works in the Graphics, Imaging, and Games Lab. His research interests focus on image processing, image stylization, and super-resolution. He received his master and bachelor degrees in computer science from Carleton University. Previously, he worked as a graphic designer.

Seungyong Lee is a professor of computer science and engineering at Pohang University of Science and Technology (POSTECH), Republic of Korea. He received his Ph.D. degree in computer science from Korea Advanced Institute of Science and Technology (KAIST) in 1995. His current research interests include image and video processing, deep learning based computational photography, and 3D scene reconstruction.



Chuan Li is a research scientist at Lambda Labs. His work focuses specifically on the convergence of computer graphics, computer vision, and machine learning. He completed his Ph.D. degree in image-based modeling at the University of Bath. Before joining Lambda Labs, he was a postdoctoral researcher at the Max Planck Institute of Informatics and a research associate at Utrecht University and at Mainz University. His research in visual data analysis and synthesis has been published at CVPR, ICCV, ECCV, NIPS, and SIGGRAPH.

Yong-Jin Liu is a tenured full professor in the Department of Computer Science and Technology, Tsinghua University. He received his B.Eng. degree from Tianjin University, China, in 1998, and his Ph.D. degree from Hong Kong University of Science and Technology, China, in 2004. His research interests include cognition computation, computational geometry, computer graphics, and computer vision.

Amir Semmo is a post-doctoral researcher with the Visual Computing & Visual Analytics group of the Hasso Plattner Institute, Germany, and is the head of R&D at Digital Masterpieces. In 2016, he received his doctoral degree on non-photorealistic rendering for 3D geospatial data. His main research topics include image and video processing, computer vision, and GPU computing. He is particularly interested in expressive rendering on mobile devices, image stylisation, and the processing of multi-dimensional video data.

Ariel Shamir is the dean of the Efi Arazi School of Computer Science at the Interdisciplinary Center in Israel. He received his Ph.D. degree in computer science in 2000 from the Hebrew University in Jerusalem, and spent two years as a postodctor at the University of Texas in Austin. He is currently an associate editor for ACM TOG and CVM. He was named one of the most highly cited researchers on the Thomson Reuters list in 2015. He has a broad commercial experience consulting for various companies. He specializes in geometric modeling, computer graphics, image processing, and machine learning.

Minjung Son received her B.S., M.S., and Ph.D. degrees from Pohang University of Science and Technology (POSTECH) in 2005, 2007, and 2014, respectively, all in computer science and engineering. Since 2014, she has been with the Samsung Advanced Institute of Technology, Suwon, Republic of Korea, as a senior/principal researcher.

Holger Winnemöller received his B.Sc., B.Sc. (Hons), and M.Sc. degrees in computer science from Rhodes University, South Africa, between 1998 and 2002. He then moved to the US, where in 2006 he received his Ph.D. degree from Northwestern University. Since 2007, he has been with Adobe Research in Seattle, Washington, where he is currently a principal scientist. His research domains include nonphotorealistic rendering and novel digital media, while his current research focuses on creative tools for aspiring (nonprofessional) artists and casual creativity.