# Developing and Applying A Benchmark for Evaluating Image Stylization

David Mould[a], Paul L. Rosin[b]

[a] *Carleton University*
[b] *Cardiff University*

**Abstract**

The non-photorealistic rendering community has had difficulty evaluating its research results. Other areas of computer graphics, and related disciplines such as computer vision, have made progress by comparing algorithms' performance on common datasets, or *benchmarks*. We argue for the benefits of establishing a benchmark image set to which image stylization methods can be applied, simplifying the comparison of methods, and broadening the testing to which a given method is subjected. We propose a set of benchmark images, representing a range of possible subject matter and image features of interest to researchers, and we describe the policies, tradeoffs, and reasoning that led us to the particular images in the set. Then, we apply six previously existing stylization algorithms to the benchmark images; we discuss observations arising from the interactions between the algorithms and the benchmark images. Inasmuch as the benchmark images were able to thoroughly exercise the algorithms and produce new material for discussion, we can conclude that the benchmark will be effective for its stated aim.

*Keywords:* non-photorealistic rendering, image processing, abstraction, stylization, evaluation, benchmark

*Email addresses:* mould@scs.carleton.ca (David Mould),
Paul.Rosin@cs.cf.ac.uk (Paul L. Rosin)

## 1. Introduction

During the early days of a research topic, there is more focus on trailblazing than on formal analysis and evaluation. However, as the research area matures, many competing approaches are developed, and it becomes more difficult to distinguish between their relative benefits. In NPR, just as in other disciplines, a systematic and objective approach to comparative evaluation is necessary [1, 2, 3].

An ideal method for evaluation should be general purpose, applicable to a wide variety of algorithms. The standard approach used in computer vision is to define a ground truth result against which an algorithm's results are compared. Unfortunately, for NPR no ground truth is available. Not only are many different stylizations possible, often radically different in appearance, but an individual stylization (etching, say) can come in many varieties. In computer vision, some "no-reference" image measures have been developed, which do not need ground truth images, and are generally based on low-level features extracted from the image. However, while this has proven popular for image quality assessment [4], it is not easy to find "no-reference" measures for other assessment tasks. In addition, "no-reference" measures tend to lack discriminatory power compared to measures that have access to ground truth. While proxy measures [1] are fairly general and have been applied to NPR, they are at best loosely connected to the quantities of interest, such as the aesthetic appeal of the image.

Hall and Lehmann [5] agree with Hertzmann [1] in arguing that NPR cannot be assessed by human-subject experiments. Inspired by practices in Art History, they suggest that stylized images should be assessed by comparison to other existing (e.g. art) works, as well as existing criteria ("norms") used implicitly by people in the field, such as automation, algorithmic elegance, novelty, or "wow factor". This paper concentrates on facilitating comparison: the relative strengths and weaknesses of different algorithms can be revealed by applying them to a common dataset.

We use the term *benchmark* to refer to a standard set of data that algorithms use as input so as to produce comparable output. Usually, the evaluation is numerically scored, but that is not presently feasible in NPR. Nevertheless, an NPR benchmark can still provide a useful resource. At the most basic level, it facilitates comparison of NPR algorithms by providing a common set of images. Comparisons on common images already occur informally and sporadically, as images from some published papers are occasionally reused by later authors. Our intent is to encourage more systematic comparisons through use of a common dataset.

We propose an NPR benchmark, named *NPRgeneral*, in which the images collectively exhibit a wide range of possible features of interest, such as texture, contrast, complex edges, and semantically meaningful structures such as human faces. Details are given in Section 3. The benchmark can be used to compare algorithms, by inspecting the results of different algorithms on independently chosen input, and it can be used directly to help evaluate a single algorithm, showing the results over a variety of input images. Many of the images are quite challenging, and we do not expect every algorithm to succeed with every input. The failure cases are potentially of even more interest to the research community than the successes, since they embody unsolved problems and hence illuminate directions for future work. This benchmark is not specific to any particular style or subject matter, and is intended for use by algorithms that can take arbitrary image input, hence the name "NPRgeneral".

This paper is an extended version of the conference paper that initially introduced the benchmark set [6]. In the current paper, we recapitulate the discussion of the need for an NPR benchmark and the reasoning behind the specific benchmark images chosen; our new contribution is to apply six existing stylization algorithms to the full benchmark set and assess the results. The discussion of the stylized benchmark images serves both as an example of how we imagine others using the benchmark in their own future papers, and as a demonstration of the effectiveness of the benchmark set: the benchmark contains sufficient variety of content that we can gain some insight into the behavior of stylization algorithms by examining the stylized benchmark images.

Note that our goal in this paper is to present and assess the benchmark for its ability to exercise image stylization methods. We use some existing methods as examples to demonstrate the breadth of content in the benchmark image set. Turning a critical eye to the filtered images, we will point out particular aspects that strike us as noteworthy, as arising from interactions between the stylization algorithms and the contents of the benchmark images. We are not, per se, making a general assessment of the effectiveness of the existing methods, nor making direct comparisons between the methods we discuss. The reader can consult the original papers to see the objectives of the original authors and their evaluation of the methods' effectiveness.

With dozens or perhaps hundreds of image stylization algorithms available in the literature, we must be selective in this paper. We chose six stylization meth-

ods, grouped into two broad categories: *abstraction*, in the sense of stylization through detail removal, and *reduced-palette rendering*, where the image is communicated without color and using only a limited tonal range. Both categories represent overall objectives shared by numerous methods in NPR. Within each group, we chose three recent algorithms to apply. Discussion of the benchmark images themselves, and our observations on the interaction between the chosen stylization methods and the benchmark images, make up the bulk of this paper.

## 2. Previous Work

Evaluation within the NPR discipline has been limited, both in terms of the amount of evaluation that has been carried out, and also regarding the variety of approaches taken to the evaluation [7, 2]. Proxy metrics and variously formal and informal user studies are common. Mould [3] argues for a principled form of subjective evaluation from proponents of stylization methods, to augment objective metrics and instead of user studies.

When the rendering style is tightly controlled, and moreover corresponds to a traditional artistic style, it is possible to obtain artists' drawings that can stand in for ground truth data. The similarity between artist and algorithmically generated images can then be compared by performing a user study. For example, Isenberg et al. [8] compared a variety of pen-and-ink line drawing styles generated by human artists and algorithms. Images were shown to participants who were asked to sort the images into piles according to style, realism, aesthetics, or other considerations they thought helpful. While the participants could distinguish between the artist-generated and computer-generated drawings, the latter were still highly rated. The even more restricted task of drawing a single pencil line was explored by AlMeraj et al. [9]. Subjects were given the two-alternative forced choice task of deciding whether an image showed a line that was hand-drawn or computer-generated. Their tests indicated that the computer-generated line drawings were often perceived as hand-drawn.

An example of the proxy measure approach referred to by Hertzmann [1] is the memory game, used by Winnemöller et al. [10] to evaluate their NPR algorithm [10]. Participants were shown a $3 \times 6$ grid of cards with back side up; every time the player clicked a pair of cards they were revealed for a short time. If the cards uncovered by two consecutive clicks match, then both cards were removed; otherwise, they were turned back over to hide their contents. The time to complete the game and also the number of cards turned during the game were used to measure the performance of the player. When the memory game contained stylised images, the players' performances improved. From this, it was argued that the stylisation produced distinctive imagery. Other authors [11, 12, 13] have also used matching tasks for evaluation of the authors' NPR algorithms, even though the purpose of the stylizations was not always or only to create memorable images.

Proper evaluation of image stylization methods requires comparisons between multiple approaches. Ideally, the algorithms would be run on common data so that meaningful conclusions could be drawn from the output; researchers should therefore coordinate on a common dataset. In computer graphics, informal reuse of well-known models is common, with models such as the bunny, Buddha, and armadillo seen in many papers, and of course the ubiquitous teapot. Similarly, images such as Lena have seen informal and widespread usage in image processing papers. Stronger coordination becomes possible when researchers agree on a suitable benchmark dataset.

In recent years, image benchmarks have proliferated. There are now literally hundreds of publicly available benchmarks suitable for a wide range of topics, including analysis of faces, gestures, biometrics, object retrieval, pedestrian and vehicle tracking, medical images, character recognition, image segmentation, stereo, saliency, and more. For facial analysis alone, many such benchmark databases exist [14]. Early efforts reused existing collections of photographs, such as Brodatz's *Photographic Album for Artists and Designers* [15], which became popular for testing texture analysis algorithms. A later trend was to create bespoke image benchmarks, so as to enable careful control of the content. For example, the CMU PIE Database [16] captured 41,368 face images of 68 people in 13 poses, with controlled lighting and facial expressions. Recently some extremely large benchmarks have been been created. For instance, the SUN Database [17] collected 130,519 images containing 99 categories from the Internet using online search queries for each scene category term, while the Large Scale Visual Recognition Challenge 2015 [18] used 150,000 images which had been collected from Flickr and other search engines, and then hand-labeled with the presence or absence of 1000 object categories. Recently an image benchmark containing 3.4 million annotated images across 70 classes containing regions of interest was released for Plankton Classification [19]. The largest image dataset of which we are aware is the YFCC100M dataset, containing one hundred million multimedia objects, 99M of which are

photographs [20].

Thomee et al. [20] discuss some of the issues around image databases. While many image datasets have been proposed, most contain content with restrictive licenses, whether because the copyright owner must give permission for use, because the benchmark creator requires a license agreement as a condition of access, or because the benchmark is intended for use in a specific competition and access is restricted to competitors. The YFCC100M dataset contains only content with some sort of Creative Commons license. While Thomee et al. suggest that the massive size of YFCC100M is a key strength, its vastness poses problems as well. In NPR, where researchers labor over the evaluation of individual images, a small benchmark set is needed. Thomee et al. suggest mechanisms for communicating subset selection logic; in this paper, we directly propose a dataset of twenty images, small enough that researchers will find it easy to apply their methods to all of them.

### *2.1. Demonstration*

Once we have created our benchmark dataset, we will present a demonstration of one possible manner in which it can be used: take a candidate stylization technique, apply it to the entire dataset, and observe the effectiveness of the technique in different contexts.

The community has developed a great many methods for image stylization. We chose two categories of method, "reduced-palette" and "abstraction"; within each of those categories, we identified three algorithms for use in this paper. We tried to choose methods that were somewhat different from one another. Pragmatically, we also chose algorithms where implementations were available.

#### *2.1.1. Reduced-Palette*

The area of *reduced-palette* rendering seeks to represent an image without color and using only a few greylevels [21]. In an extreme case, an input color photograph might be reproduced as a binary black-and-white image. More commonly, multiple greylevels are permitted, even if the output image is predominantly black and white. Reduced-palette algorithms might use an arbitrary raster output or might represent the image with small primitives such as lines or dots.

*Stippling* is an artistic style where an image is composed of many small dots. It has a long history in NPR, beginning with the work of Deussen et al. [22] and continuing to the present day. The "structure-preserving stippling" (SPS) method of Li and Mould [23] adapts a halftoning method to stippling. Built on contrast-aware halftoning [24], it traverses the input image's pixels, thresholding each pixel and pushing the error to the surrounding pixels. The key idea is to match the error distribution to the local pixel trend: darker pixels receive more negative error but less positive error, and lighter pixels attract more positive error and less negative error. Pixels are processed in a priority order, the most extremal-valued pixels first. Because there should be fewer stipples than black pixels, the algorithm also includes an error manipulation step, where positive error (lightening the image) is amplified and negative error (darkening the image) is reduced.

Winnemöller et al. [25] presented the eXtended Difference-of-Gaussians method (XDoG), a versatile stylization algorithm capable of producing images in black and white or resembling pencil, charcoal, and other artistic media. The method depends on the Edge Tangent Flow (ETF) field proposed by Kang et al. [26] to generate a direction field, then smooths the field perpendicular to the edge directions. By varying the smoothing locally, and thresholding the resulting smoothed image, XDoG can produce high-quality stylizations over a wide range of inputs.

Rosin and Lai [13] presented a family of algorithms for stylization; we use their central algorithm, where ETF-based lines and region-based tonal blocks are combined to create a two-tone image. The tonal blocks were extracted using a series of image processing operations: multilevel thresholding, morphological cleaning, and grab-cut refinement. The algorithm is referred to as "minimal rendering", where "minimal" is in the sense both of using few primitives to communicate the image and using few colors.

In the present paper, the original method was modified with the intent of improving performance over a wide range of images while not demanding image-specific parameter tuning. The modifications are as follows.

First, two versions of each image were processed: the original, and a version with the intensity channel preprocessed by contrast-limited histogram equalization with a slope limit of 4. The processed results from each version of the image were averaged, creating a three-tone output. This modification helped the method handle the wide range of intensities seen in the benchmark set.

Second, the algorithm was applied at two scales: the original and a downscaled version (one-quarter area). The processed version of the downscaled image was resized to the original scale and the processed results were combined using a logical AND. Finally, to ensure that the white lines from the high-resolution version were re-

tained, these lines were reinserted. The multiscale processing permitted the method to deal with the range of structure sizes seen in the benchmark set without any need to change parameters.

Third, the raster image resulting was processed with Potrace [27], producing a smooth vector image.

#### *2.1.2. Abstraction*

Abstraction has been a general concern of non-photorealism since the inception of the field. Most work in NPR involves abstraction to some degree, often explicitly so. Of many available abstraction algorithms, we selected three for discussion in this paper: the texture-removing abstraction of Papari et al. [28]; the texture-preserving abstraction of Mould [29]; and the directional oil paint of Semmo et al. [30]. Although the method of Semmo et al. somewhat approximates the appearance of oil paint, with the deliberate inclusion of apparent brush strokes, all three methods can be considered a somewhat generic abstraction, lacking strong connection to any particular historical style.

Papari et al. present a variant of the Kuwahara filter. At each pixel, the nearby pixels are allocated to sectors; the average color of each sector is determined, and the pixel's output color is determined by the distribution of colors over the sectors. The method is able to obtain good edge and corner preservation, especially when the edges and sector boundaries align; details small enough to fit within a sector, or high-frequency details distributed over multiple sectors, are suppressed. The outcome is an abstracted image that preserves and enhances edges but removes textures and details below the filter scale.

Mould's "cumulative range geodesic filter" uses a box filter over a mask customized to every pixel. The mask contains the nearest $n$ pixels to the origin, where "nearest" is with respect to shortest paths whose incremental distance includes the color distance between the new pixel and the original pixel. Such masks were shown to have excellent feature-preserving properties at scales near and above the mask size, yielding an appealing abstraction.

Semmo et al. create an oil painting effect with a multi-stage pipeline. Their algorithm has two main elements: first, color quantization, where a palette of dominant colors is derived from the image; second, a direction field, obtained from an eigenanalysis of the smoothed structure tensor. Line integral convolution along the direction field yields pixel colors, and stroke texture is combined with these colors to produce a final image resembling a painting.

## 3. Principles of Image Dataset

Before presenting our set of selected images, we will discuss the policies that led us to that particular set. The set of images is by no means unique in satisfying our constraints; indeed, we expect the benchmark set to evolve with input from the community. However, the set does serve to demonstrate the possibility of a plausible compromise among complex and sometimes opposing considerations.

**Challenging images**: The benchmark needs to include challenging images that are likely to be problematic for NPR algorithms. As different types of algorithms typically have different weaknesses, the images should be challenging in different ways. This requirement will help uncover weak spots in current algorithms, and identify limitations which can be addressed in future work. Thus it helps enable a realistic appraisal of the state of the art (undercutting over-selling algorithms) and will help push algorithmic development.

**Range of difficulty**: The benchmark should include images covering a range of levels of difficulty, so that the overall results of applying an NPR algorithm to the benchmark should not be a binary pass or fail. Indeed, a given image should not be a binary pass or fail, but some more complex measure of how effective the algorithm is, possibly qualitative. Including only challenging images would likely discourage many potential users, especially if they were developing more experimental methods.

**Small number of images**: Whereas the trend for benchmark datasets such as those used in computer vision is to contain thousands or millions of images, our requirement is the opposite. Image analysis methods that produce results such as classifications can easily compute assessments using automatic scoring. Conversely, most evaluation in NPR will be done manually. A small dataset is essential for manual evaluation to be manageable; given a large dataset, we expect that users would only use small selections, and since different users would make different selections, the results across different papers would not be comparable, defeating the original purpose of using a common benchmark. A sufficiently small dataset can be treated in its entirety.

Notice that that the criteria of using a small dataset and providing a broad coverage are in conflict. Since we cannot sacrifice the compactness of the dataset, its coverage is necessarily limited. In particular, semantic variation of the photographic subjects somewhat suffers. However, low-level details are still extremely varied within the set we chose.

**Photographic images**: The images in the dataset should be conventional photographs, as we think that stylizing captured real-world scenes offers the most difficult and widely relevant problems. There may be some utility in stylizing hand-drawn or other artistic images, but work (in sketch-based modeling, for example) based on handmade images generally uses rich data including a history of marks and information about the primitives involved. We are not aware of work that concentrates on stylizing general handmade images using only the images themselves. Computer-generated synthetic images are another possibility; even more than handmade images, computer-generated images would typically contain much more than simply color information, with additional channels available such as depth, normal, object ID, and surface texture coordinates. These additional channels can potentially be exploited by stylization algorithms to good effect. We recommend creating an entirely separate dataset of 3D models and scenes for benchmarking evaluation of such methods.

**Still images**: We have deliberately excluded video from the present benchmark, not because we think it is unimportant, but rather because we think it is important enough to do a good job with it and it is distinct enough from images that its considerations will need to be addressed separately. Video has the added complications of time and motion. Complex motions and apparent motions owing to changes in camera parameters (focus, zoom, orientation) and the movement of the camera and of objects in the scene need to be considered carefully. Even basic questions like the appropriate duration of a shot do not have obviously correct answers.

**Standard painting types**: Many captured images follow standard topics. For instance, in the AVA dataset [31] landscape, still life, animals, and portraits are all popular tags. NPR has been influenced by historical artistic practices, and standard painting genres such as landscapes, portraits, and still lifes should be represented in the benchmark.

**Aesthetics**: Many stylization algorithms are designed with the intention of generating aesthetically pleasing results, researchers in this area tend to use source images that in their original state are also aesthetically pleasing.

**Metadata**: Including metadata such as numerical ratings of image characteristics is a useful adjunct, as these can then help characterise the performance of NPR algorithms. Correspondences between scores in the metadata and measures of image quality can be enlightening. For example, metadata could reveal that a specific algorithm has problems with images that are low contrast and contain large amounts of fine detail. The metadata can be provided by subjective human annotations and by objective measures using automatic image processing. We provided measurements of some characteristics of interest for the images in our proposed dataset.

**Copyright clearance**: Since NPR relies on manual evaluation (rather than listing numerical scores), it is essential that all the benchmark images have copyright clearance so that they can be published along with the derived results. We took images from Flickr, selecting only those whose license permits distribution of modified versions.

**Image size**: We wanted images for which large sizes – at least 2048 linear pixels – were available. We will make at least two sizes available for benchmark users: a large size and a smaller size, standardized at 1024 pixels width. Aspect ratios vary slightly; by chance, all our images had a landscape or square aspect ratio, but we did not particularly use aspect ratio as a selection criterion.

### *3.1. Image characteristics*

The following is a list of image properties we sought to include. We selected images so that each property can be found in several images in the benchmark set. While not all properties are equally important, each property is doubtless of interest to some subset of stylization algorithms. For example, an algorithm may have an inherent scale parameter, and it is worthwhile to test its effect on images where the elements vary in size. Many stroke-based algorithms have difficulty conveying fine-scale detail or high-frequency texture. Conversely, while filter-based algorithms with local thresholding can handle texture and fine detail well, long gradients may prove problematic. We do not intend to claim that the list is exhaustive; we welcome suggestions for additions that can guide the future development of the benchmark image set.

- **Variation in scale** of the elements in the image.
- **Fine detail**: high-frequency structure, whether fine-scale texture or semantically important elements that are quite small.
- **Variation in texture**, usually arising from multiple types and scales of texture within a single image.
- **Regular structure**, encompassing both regular patterns and clean shapes such as straight lines, 90-degree angles, and circular arcs.
- **Irregular texture** such as foliage or unkempt hair.
- **Visual clutter**: prominent visual elements that are irrelevant to communicating the main content of the image.
- **Vivid and varied colors** over the image.
- **Muted colors**, such that the image contains unsaturated colors and the color contrast is low.

• **Low contrast**: some important image elements have low dynamic range.

• **Mixed contrast**: different image regions have different dynamic ranges, or use similar dynamic ranges with different average intensities.

• **Complex edges**: some of the silhouettes or other important edges are long and complicated; the silhouette of a tree would be an example.

• **Thin features** such as wires or tree branches are present in the image.

• **Indistinct edges** where the semantics of the scene indicate an edge to a human observer, but the pixels exhibit only a small change in intensity or color.

• **Long gradients** of intensity or color in the image plane, perhaps due to curved surfaces or lighting changes.

• **Human faces** are of particular interest to human artists and audiences; we count only images where a face makes up a significant portion of the image, as in a portrait.

• **High key** (or generally light images) and **low key** (dark images) are included to confirm the robustness of the methods against more extreme inputs (which are nevertheless often generated for artistic effect).

### *3.2. Limitations*

The principles articulated above provide guidelines for selecting images. However, these guidelines are not necessarily complete, and they leave considerable room for judgement in deciding precisely which images should be included. We do propose a specific set of images, discussed in the next section; we consider this to be "version 1.0" of the benchmark. It is ready for use. Over time, we may release further versions, pending additional suggestions from the NPR community arising from experience using the presented image set.

The principles constrain the benchmark content, sometimes with a negative impact on the applicability of the benchmark. By restricting our benchmark to a small set, we necessarily sacrifice detailed coverage of image variations. For machine learning applications, a much larger dataset is required. For specialized methods such as portraiture, many of our images are irrelevant and the benchmark is insufficient by itself.

The most salient constraint arises from our deliberate decision to exclude video from the current version of the benchmark. Image stylization methods can be applied to video straightforwardly, if not always effectively, by stylizing each video frame separately; a video benchmark set would help standardize evaluation of video stylization. As discussed above, though, video has many considerations that images lack. We concentrated on images in this paper, but intend to extend the benchmark to include video as well.

The benchmark also excludes items such as 3D scenes and models. While this to some extent reflects the research interests of the present paper's authors, we also think that the need for a benchmark set is not as crucial there, given de facto benchmarking in using common models such as the Stanford bunny.

This paper presents a basic version of the benchmark. Expanded benchmarks are possible. One vision of an expanded benchmark would be a hierarchical dataset, where a core subset would be considered mandatory, and then preselected sections of the full dataset could be used according to the requirements of the method. The risk of a larger dataset, even with a defined core, is that authors might be tempted to pick and choose subsets, undermining the usefulness of the common benchmark. Nonetheless, special-purpose image sets can be useful for more focussed methods. In parallel work, Rosin et al. have devised a dedicated set of face images meant to help evaluate portrait stylization texhniques [32]. Other special-purpose modules can be added if there is sufficient interest.

## 4. Proposed Benchmark Set

This section discusses our tentative benchmark set. All 20 images can be seen in Figure 1. Top row: dark woods; mountains; cabbage; Mac. Second: angel; barn; toque. Third: Oparara; arch; headlight. Fourth: Yemeni; daisy; snow. Fifth: athletes; desert; tomatoes. Last row: city; rim lighting; cat; berries.

Table 1 summarizes the list of image properties and shows which of our images possess them. The decision about whether or not to identify a given property with a given image is of necessity subjective, although our choices are informed by numerical measurements of related traits, summarized in Table 2. We measured *colorfulness, complexity, contrast, sharpness, lineness, noise,* and the mean and standard deviation of intensity. These low-level features vary widely over our image set, giving us some confidence that the benchmark provides a broad spectrum of test cases. In addition, they will enable ratings derived from user studies of NPR results to be correlated against image measures, so that relationships (e.g. a certain algorithm may perform poorly on noisy or low contrast images) can be easily identified. Details of the measurements are given next. Default parameters from the relevant papers are used unless otherwise stated. **Image colourfulness:** computed following Hasler and Süsstrunk [33]. They use a simple mea-

| | angel | arch | ath | barn | berr | cabb | cat | city | daisy | dark wood | desert | head light | mac | mtns | opa | rim | snow | toma | toque | yem |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| varied scale | ✓ | ✓ | | ✓ | | ✓ | | ✓ | ✓ | | | | | | | | | ✓ | | |
| fine detail | | ✓ | ✓ | ✓ | | | ✓ | | | | ✓ | | | ✓ | ✓ | | | | ✓ | ✓ |
| varied texture | | ✓ | | ✓ | ✓ | | | | | ✓ | ✓ | ✓ | | ✓ | ✓ | ✓ | | | ✓ | ✓ |
| regular | | | | ✓ | | | | ✓ | | | | ✓ | | | | | | | ✓ | |
| irregular | | ✓ | | ✓ | ✓ | | ✓ | | | ✓ | ✓ | | | ✓ | ✓ | | ✓ | | | |
| clutter | | | ✓ | | ✓ | | ✓ | | | | | | | | ✓ | | ✓ | ✓ | | |
| color | | ✓ | ✓ | ✓ | ✓ | | | | | | | | | | | | | ✓ | | |
| muted | ✓ | | | | | ✓ | | ✓ | | ✓ | ✓ | ✓ | | ✓ | | | | | | |
| low contrast | ✓ | | | | | | | | ✓ | | ✓ | | | ✓ | | | ✓ | | | |
| mixed contrast | ✓ | | | | | ✓ | | ✓ | | ✓ | ✓ | ✓ | | ✓ | ✓ | | ✓ | | | ✓ |
| thin features | | | | ✓ | | ✓ | ✓ | ✓ | | ✓ | ✓ | ✓ | ✓ | | ✓ | | ✓ | | ✓ | ✓ |
| complex edges | | | ✓ | | | ✓ | ✓ | | | ✓ | ✓ | | | ✓ | | ✓ | ✓ | ✓ | | ✓ |
| gradients | | ✓ | | | | | | | ✓ | | ✓ | ✓ | ✓ | | | | ✓ | ✓ | | |
| indistinct | ✓ | | | | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | | | ✓ | ✓ | | |
| low key | ✓ | | | | | | | | | ✓ | | | | | ✓ | ✓ | | | | |
| high key | | | | | | | | | ✓ | | | | ✓ | ✓ | | | ✓ | | | |
| portrait | | | | | | | | | | | | | ✓ | | | ✓ | | | ✓ | ✓ |

Table 1: Our assessment of which images possess which properties of interest.

sure involving the means and standard deviations of the image pixels in the red-green and yellow-blue channels of opponent colour space, with the weighting of these terms determined by a perceptual experiment.

**Image complexity:** computed following Machado and Cardoso [34], who encode the image using JPEG compression at a fixed quality factor quality; we used 50. For more complex images, compression will incur a high error, and also yield a low file size compression ratio. Therefore the ratio of these two terms is an estimate of the complexity of the original image.

**Image contrast:** computed following Matković et al. [35]. A non-linear mapping is applied to the image intensities to match them better to human perception. Each pixel's local contrast is measured as the mean absolute difference with respect to its neighbouring four pixels, and contrast is summed over the image. The process is repeated at multiple (specifically 9) image resolutions, and a weighted sum of these contrasts provides the final measure.

**Image sharpness:** computed following Bahrami and Kot [36], who compute for each pixel the maximum difference with respect to its 8-neighborhood, termed the maximum local variation MLV. The distribution of MLV in an image is modelled by a Generalized Gaussian Distribution (GGD) with a weighting to increase sensitivity to large MLV values. The sharpness measure is taken as the standard deviation of the GGD.

**Lineness:** since a standard approach was not available, we developed a new measure. In a similar manner in which the summed edge strength over the image is used to measure sharpness [37], we have used the absolute value of the Laplacian of Gaussian (LoG) at two scales ($\sigma = \{2, 4\}$) to measure the response to dark or bright lines. However, the LoG also produces responses adjacent to edges, and so these have been suppressed following the approach taken by Rosin and Lai [13].

**Image noise estimation:** computed following Immerkaer [38]. His approach assumes that the estimation should be insensitive to edges, and so the image is convolved with a Laplacian, which should give no response at edges. Assuming normally distributed noise, the noise level is derived from the Laplacian response. None of our images is particularly noisy; for our purposes the measurement is better thought of as an estimate of the image's high-frequency content.

Of course, the image measures are not mutually independent. Below, we give the covariance matrix of the values in Table 2. Since the numerical ranges of the measures are not standardised we have first normalised each measure to have unit standard deviation.

| *color* | *complex* | *contrast* | *sharp* | *line* | *mean* | *stdev* | *noise* |
|---|---|---|---|---|---|---|---|
| 1 | .315 | .372 | .039 | .174 | .197 | .427 | .314 |
| .315 | 1 | .623 | .559 | .787 | −.108 | .107 | .994 |
| .372 | .623 | 1 | .541 | .607 | −.378 | .605 | .615 |
| .039 | .559 | .541 | 1 | .698 | −.446 | .066 | .550 |
| .174 | .787 | .607 | .698 | 1 | .005 | .276 | .738 |
| .197 | −.108 | −.378 | −.446 | .005 | 1 | .177 | −.128 |
| .427 | .107 | .605 | .066 | .276 | .177 | 1 | .103 |
| .314 | .994 | .615 | .550 | .738 | −.128 | .103 | 1 |

There is a strong correlation between complexity and

| | colourfulness | complexity | contrast | sharpness | lineness | mean | standard deviation | noise |
|---|---|---|---|---|---|---|---|---|
| angel | 12.97 | 0.48 | 4.84 | 0.19 | 3.32 | 67.77 | **29.96** | 3.48 |
| arch | 71.48 | **2.98** | **10.68** | 0.21 | **8.20** | 113.00 | 62.14 | **24.05** |
| athletes | 45.68 | 0.21 | 5.50 | 0.16 | 2.43 | 136.37 | 47.76 | 1.54 |
| barn | 73.39 | 0.82 | 6.08 | 0.18 | 4.59 | 142.15 | 63.10 | 6.08 |
| berries | **101.74** | 0.71 | 9.38 | 0.20 | 4.46 | 105.74 | 66.24 | 5.30 |
| cabbage | 26.18 | 0.39 | 5.40 | 0.16 | 4.26 | 108.64 | 38.30 | 2.55 |
| cat | 38.39 | 0.93 | 7.18 | **0.22** | 4.37 | 112.73 | 55.49 | 8.86 |
| city | 47.86 | 0.48 | 6.76 | 0.18 | 6.14 | 148.45 | 72.47 | 2.80 |
| daisy | 32.04 | 0.08 | **1.99** | **0.08** | 1.48 | **208.87** | 32.47 | **0.59** |
| darkwoods | 34.81 | 1.32 | 6.69 | 0.20 | 4.69 | 55.17 | 39.78 | 10.36 |
| desert | 46.59 | 0.52 | 4.29 | 0.13 | 2.57 | 114.06 | 39.54 | 4.88 |
| headlight | 24.49 | 0.15 | 6.28 | 0.14 | 4.02 | 93.86 | 59.71 | 0.88 |
| mac | 57.52 | **0.07** | 2.58 | 0.12 | **1.09** | 159.93 | 46.03 | 1.00 |
| mountains | 41.64 | 0.13 | 5.09 | 0.10 | 1.38 | 145.03 | 64.74 | 1.25 |
| oparara | 27.65 | 0.56 | 6.88 | 0.18 | 3.45 | 44.71 | 42.54 | 4.09 |
| rim | **11.98** | 0.21 | 5.62 | 0.21 | 2.26 | **23.72** | 42.34 | 2.51 |
| snow | 20.73 | 0.79 | 4.76 | 0.22 | 6.47 | 183.55 | 58.11 | 5.86 |
| tomato | 60.68 | 0.21 | 6.63 | 0.15 | 2.05 | 90.31 | 69.47 | 2.41 |
| toque | 23.10 | 0.36 | 9.14 | 0.14 | 3.11 | 121.71 | **81.45** | 3.00 |
| yemeni | 57.72 | 0.35 | 6.26 | 0.16 | 2.94 | 90.01 | 63.70 | 3.17 |

Table 2: Numerical measurements of image properties for each of the images in the benchmark; minimum and maximum values are highlighted for each measure.

the noise measure. The lineness measure also has high correlations with complexity, noise, and sharpness. A set of lower, but still reasonably high, correlations exist between contrast and complexity, noise, lineness, and the standard deviation of intensity. Note that no significant correlations exist between any image measures and colourfulness or mean intensity. Thus we see that, despite some correlations, the image measures still capture a reasonable range of image characteristics.

In the remainder of this section, we discuss the individual images in our benchmark. Each image has a combination of low-level and higher-level features of interest. Not all features are of equal importance, nor equally widespread throughout the benchmark set; a simple count of features does not give a very good estimate of the value of a particular image. In addition to seeking variety of content and image features in the set, we tried to make all the individual images reasonably appealing.

**Angel.** The stone of this image is fairly dark overall, but high intensities along the angel's arm and torso produce areas of high contrast. Lower contrast makes some important image elements difficult to see, such as the angel's nose and wing and the lower faces; overall, we assess the contrast as mixed. Elements exist over multiple scales, from the largest structures such as the arm and wings, to smaller structures such as facial features, feathers, and the leaves of the wreath. Colors are muted, and some edges are indiscernible owing to the lighting and low color contrast. Texture details in the stone surface add further visual interest.

**Arch.** This image depicts the Liberty Bell Arch in Nevada. It has strong and moderately interesting silhouettes, and it was included in the benchmark because of its irregular rock textures. There are high-frequency textures throughout the image, but the image-space scale of the rocks varies across the image, from the larger objects on the leftmost part to the smaller structures in the lower middle and right. The color range of the rocks is limited. The sky contributes a long vertical gradient. Communicating the sometimes indistinct structure of the plants and features of the rock will be a challenge for many stylization methods.

**Athletes.** Unlike the other images in the benchmark, we see the full human figure in this action scene. The high contrasts and bright colors make the image superficially straightforward for many methods, but there is potential for distraction from the irregular albeit blurry background, and some edges, such as the hair and the cleats, will be complex if the structure is preserved faithfully. Researchers will often want to preserve fine details of the facial expressions of the athletes; we do

not label this image as a portrait, though, since the faces occupy so little of the image plane.

**Barn.** This colorful image contains objects over a wide range of scales, from the largest objects such as the barn and silo, through intermediate-sized objects such as trees and the component parts of the buildings, down to very small structures such as tree branches, boards on the barn's front, and the ladder leading up the silo. Many features are thin, including tree branches and the struts and rafters visible on the nearest part of the barn. Texture is varied, with irregular texture in the vegetation and more regular texture on the silo and the face of the barn.

**Berries.** This is the most colorful image in our collection, as judged both subjectively and by the automatic "colourfulness" measure. It contains objects of somewhat different sizes – the strawberries, blackberries, blueberries, and spoon – and one could consider the image to be cluttered; not only is the pattern on the plate a potential distraction, unusually, the image is an example of foreground clutter, where not all details in the foreground necessarily need to be retained in order to communicate the image content. There is a mix of edge strengths. The overall image might be considered a variable texture, and the textures on the strawberries and blackberries differ.

**Cabbage.** This image has little color range but a wide range of intensities. The leaf boundaries are convoluted and sometimes difficult to detect; in places, they can be confused by the interior edges of the leaf veins. The veins themselves are thin features and occur at different scales, being larger on the outer leaves than the inner ones. The lighting is varied over the image. We anticipate the cabbage being a moderately challenging image for stylization methods.

**Cat.** The complex patterns and detail in the fur of the cat provide most of the visual interest of this image. The blurry but varied background may be challenging, with an indistinct boundary separating it from the furry foreground. The cat's whiskers are thin but definite features. Edge shapes in this image (e.g., the fur of the cat's ears) will be complex even when well defined.

**City.** The masterly composition of this image provides a high level of visual interest throughout. Colors are generally muted, but the contrast is usually high; the dark clothing of the human figures provides a focal point. In the city itself, windows and building silhouettes are regular structures, while more distant buildings have reduced contrast and ultimately vanish. Wiring in the interior and architectural elements on building exteriors are thin features. The perspective yields structures over a wide range of scales.

**Daisy.** A high-key image with some sharp and some blurry edges. The petals vary in size considerably; gradients across the largest petals, caused both by soft shadows and by curvature, offer a mild complication to algorithms. The central texture is quite regular; anisotropic textures along several petals provide a different regular texture. Most of the image has little contrast, as the dynamic range is low in the first place. The image would be more challenging if it were less abstract, but nonetheless provides a way to weakly test methods on a large number of possible image features.

**Dark Woods.** A generally low-key image, the majority of the content of this photograph is the complex, irregular textures from the tree trunks and foliage. The trees themselves supply thin features to test algorithms. Contrast is variable, with low contrast in some of the more shadowed areas and stronger contrast between the darker trees and sunlit leaves behind them.

**Desert.** A composition with mixed texture, structure, and smooth gradients along the sand dunes. The colors are muted but there is a variety of intensity edges, including simple edges such as the lighter sand against the shadows and the darker region behind, and more complex and indistinct edge shapes such as the low-contrast texture edges in the uppermost region and the tree branch silhouettes. The mix of content plus generally low contrast makes this a challenging image.

**Headlight.** An image with regular patterns of variable contrast. Long gradients across the curving metal offer challenges to segmentation methods and threshold-based techniques. Reflections on the paint, as well as the grille, contain indistinct edges. Although it contains a recognizable object, the regular geometry makes the image seem a little abstract as well.

**Mac.** A portrait of a Mac user, with generally light tones. The man's features are partially occluded by the Mac, slightly complicating stylization; the presence of glasses and facial hair may also pose a problem for some dedicated portraiture methods. Though the glasses are very clear, they are thin features; there is some small-scale texture across the man's forehead. Some edges are blurred owing to the shallow depth of field, and the Mac itself supplies large-scale gradients.

**Mountains.** An overall light image owing to the mix of snow and cloud. Snow on the mountaintops provides irregular texture. The contrast is overall low. Some edges, such as those within the clouds, or the blue mountain against blue sky, are indistinct, but the strong silhouette of the trees provides a definite and complex edge shape for stylization algorithms to work on.

**Oparara.** This depicts a limestone arch over the Oparara River in New Zealand. It is unusually dark for

a photograph, but its dark areas contain some variation and texture. The textures in the image are highly varied, including multiple scales and types of rock surface, ripples on the river, and foliage seen through the arch. We considered the image slightly cluttered, as the details of the rainforest visible through the arch are probably unimportant, and the silhouette of the arch is obscured by hanging vegetation. This image is likely to prove a challenge to many stylisation algorithms.

**Rim Lighting.** A portrait with a clean background and strong rim lighting. The darkest image in the benchmark, this image can be used to test algorithms for failures on near-uniform backgrounds. The high contrast along the rim may mask weaker but important contrasts on the man's facial features and clothing. In general, though, we do not expect this image to be especially challenging; it is a basic sanity check.

**Snow.** This is a largely high-contrast image that nonetheless may be challenging because of its overall light tone and the weak contrasts of some snow-covered branches in the midground. Dense arrangements of branches form irregular textures, while more prominent branches are thin features. There is also some muted texture on the barn. The silhouettes of the treetops and the branches against the barn are complex edges. While this image might be difficult to convey thoroughly in a stylization, we expect that its straightforward semantics may make it reasonably forgiving.

**Tomato.** The still-life composition of a bowl of tomatoes contains many features of interest. It has good contrast and strong colors as well as fine details (the hairs on the turnip root, the texture on the table and the curtain). The image contains structure across multiple scales – fine-scale texture and small structures such as stems and the tiny flower, medium-scale tomatoes, and the bowl and curtain at the largest scale. The curtain might be considered clutter. Still, the clarity of the composition and the overall clean edges will probably make this one of the simpler images to treat with image stylization techniques.

**Toque.** This is a largely straightforward portrait image, whose subject shows well-defined facial features. The regular knitted textures of the toque and scarf offer some interest; the relatively fine texture of the toque, combined with the lighting gradient, is especially noteworthy. Smaller gradients across the jacket, showing its shape, may or may not be preserved through stylization. The background, although very blurred, has high contrast. Finally, some regions of the silhouette are fairly complex, such as the fuzzy detail of the toque, the hairs on the image left, and the fur on the lower right.

**Yemeni.** A portrait of a man from Yemen, his strong features providing some inherent interest while including complications such as deep lines and a variable beard. The texture and coloration of his headgear afford additional opportunities for stylization. The shadows and lighting provide a challenge; some strong intensity edges, such as those on the tip of the man's nose, are unimportant, while weaker edges such as those on the right half of the man's face are critical.

## 5. On Adoption of the Benchmark

The benchmark is only of any benefit if the NPR community actually uses it. We envision two main use cases. First, researchers can include selected benchmark results in the pages of their published papers. Since the full benchmark has been kept sufficiently compact to be displayed in a single page, it is also feasible for them to include full results within their papers. Second, researchers can provide a more extensive set of benchmark results on a project page, augmenting the publication and helping future researchers by making comparisons easy.

We believe that widespread adoption of the benchmark will benefit the science of non-photorealistic rendering. It will encourage a more systematic approach to evaluation and a thorough disclosure of algorithms' behaviour so that weaknesses can be known and addressed by followup work. It will also help researchers in other ways, by providing sensible defaults for testing image stylization algorithms and improving access to past results for purposes of comparison.

At the same time, we recognize that benchmarks have drawbacks. A benchmark that is not representative of real data will lead to conclusions of dubious validity; we have tried to make our dataset as broad as possible, and anyway do not expect that researchers will concentrate single-mindedly on the benchmark to the exclusion of other images. The related issue of overfitting is a serious potential problem, where algorithms are finely tuned to the benchmark data and do not attain equally good performance on other data. In fact, the problem is worse in the absence of a benchmark, since researchers are free to choose inputs where their algorithms perform well; an independently chosen dataset eliminates the suspicion that the inputs were chosen excessively selectively. Lastly, saturation is a potential long-term issue, in which a benchmark was at first challenging but the discipline later advances to the point that its images are simple. In the absence of quantitative evaluation, and given the diversity of possible objectives for stylization algorithms, we do not think that saturation will be a problem in the non-photorealistic rendering field.

We encourage researchers to apply their algorithms to the entire dataset, as we have in this paper. Some benefits, such as the transparency of using a dataset chosen independently rather than by the researchers, only become possible when the entire dataset is used. In cases where the algorithm is only meant to apply to a certain image type – e.g., specialized methods for portrait rendering – the appropriate subset of the benchmark can be extracted. Where algorithms are intended for more general use, however, the entire benchmark set should be shown; even if only selected images will fit into the paper, the benchmark results can be reported as supplementary material.

The dataset in this paper should be considered "version 1.0." Researchers should use the benchmark images in their evaluation of new methods. We welcome further feedback from the community and we may release refined versions of the benchmark in the years to come.

## 6. Evaluation

In this section, we give our observations about the stylized benchmark images. We are not trying to compare the algorithms to one another or to conclude that one method is better than another. Rather, we mean to concentrate on the interaction between the algorithms and the features in the benchmark. Where the interactions provoke discussion, we can conclude that the benchmark has provided some help in understanding the behaviour of the algorithm. Where we discuss deficiencies in the stylized images, we are seeking to characterize areas of interest for future work on stylization algorithms.

We show excerpts from the stylized images in Figures 2 through 7. We found that showing the full stylized images at a small scale made it impossible to discern small details at print resolution; also, we use the excerpts to draw attention to particular elements within the full images. We generally use the same excerpts over all stylizations. Occasionally, though, we make a different choice to show particular details for a specific method; for example, in the stippled version of the barn image, we focussed on the sky to emphasize the problem with light clouds on a darker background. A summary of the common excerpts is given in Table 3. The complete stylized images are available as supplementary material and we encourage the interested reader to consult them while reading about our observations.

| image | detail |
|---|---|
| angel | wreath and wing |
| arch | sky and rock texture |
| athletes | expression and bokeh |
| barn | tree branches, sky |
| berries | different types |
| cabbage | leaf venation |
| cat | muzzle and whiskers |
| city | businessmen, distant buildings |
| daisy | detail of centre |
| desert | dunes, desert plant |
| headlight | light, grille |
| Mac | glasses, monitor |
| mountains | treeline, clouds |
| rim lighting | portion of portrait |
| snow | foreground and background trees |
| tomato | individual tomatoes, turnip root |
| toque | facial features, hat |
| Yemeni | face, beard |

Table 3: Image excerpt contents.

### *6.1. Reduced-Palette Rendering*

We applied the structure-preserving stippling technique of Li and Mould, the eXtended Difference-of-Gaussians method of Winnemöller et al., and the minimal rendering method of Rosin and Lai. The methods are fairly different technically and produce visually distinct results. They differ in the degree of abstraction, with SPS doing the least and minimal rendering the most. The greater abstraction of XDoG and minimal rendering allows them to emphasize salient objects and features; see, for example, the headlight image, where the higher contrast makes a more striking image than the relatively faithful greytones produced by SPS. However, stronger abstraction comes with the danger that inappropriate features are selected for emphasis or important features are missed; this risk is particularly acute for the minimal rendering method, with the mountains image a notable failure case.

Some issues are common across all three methods. An obvious observation is that **color** is eliminated when images are re-rendered in monochrome. Accordingly, aspects of the benchmark relating to color content cannot be evaluated using these styles. Conversely, without the distraction of color, other aspects can be assessed more keenly.

**High contrast** is the best case for all methods. Thresholding is most effective when a clear separation is available, thus making the output fairly insensitive to the choice of threshold. Similarly, stippling can be quite

effective at showing sharp boundaries. Early work on stippling sometimes used images with deliberately exaggerated contrast, or employed thresholding to enforce sharp stipple boundaries. All three methods were generally quite successful at conveying high-contrast image elements.

**Thin features** were often an area of difficulty. The scale of the ETF was often too large to capture very fine features such as tree branches in the barn and desert images, or the roots of the turnip in the tomato image. Somewhat thicker features like the glasses frames in the Mac image are thick enough to be captured by all reduced-palette methods; the glasses also have high intensity contrast, making them even easier. In principle, it is straightforward to use stipples to present linear features: the stipples can be aggregated along the feature, creating a discernible object. However, in many images, fine features were not well captured by stippling either.

**Long gradients** are a problem for reduced-palette rendering methods, since thresholding will eliminate them and potentially introduce spurious edges. XDoG can bypass the problem by declining to threshold, as in the headlight image. Rosin and Lai's minimal rendering algorithm deliberately uses few tones, so does not have this option; long gradients can be suggested using the three tones, as in the angel image, or omitted, as in the headlight image. Contrariwise, long gradients are generally well communicated through stippling: stippling is an effective halftoning effect for low-frequency structures. Thus, long, smooth gradients are treated well by SPS; for example, the shape of the headlight is nicely conveyed.

### *6.1.1. XDoG*

XDoG is a versatile method. With suitable parameter changes, it can produce different effects; the gallery of Figure 3 shows a range of outcomes, with parameters adjusted on a per-image basis. The method's flexibility has allowed it to provide high-quality results for most of the images in the benchmark set, including challenging **low-contrast** images such as the daisy and mountains.

When the scale of the XDoG matches the image detail scale, results are excellent. The fur of the cat produces a textured effect that, although exaggerated, is particularly appealing. Details at multiple scales, however, are pushed towards a single scale. In the arches image, for example, the varied details are condensed and become difficult to interpret. The snow image contains low contrast and more distant background trees along with two superimposed foreground trees which are visually quite distinct from the background trees. Although the XDoG treatment is reasonably attractive, it produces a uniform appearance over all the trees.

XDoG relies on the edge tangent field, and the smoothed direction field poorly represents very-high-frequency irregular texture. The darkwoods image, with its varied and irregular textures, has become nearly indecipherable. The city image also demonstrates a mismatch in scale between the filter and some image structures (e.g., the faces, the windows). The ETF also sometimes introduces spurious texture, visible in the barn, desert, and Mac images, among others. Further, the smoothed ETF will often simplify complex edges: the treeline in the mountains image is greatly simplified, and the cleats in the athletes image have been stripped away.

Like many methods, XDoG has difficulty handling clutter. The background in the athletes image is particularly peculiar. In principle, it should be possible to deal with clutter through human intervention, by specifying different parameter settings in different regions of an image.

**Gradients** are preserved when parameters are set to avoid thresholding. The headlight image, for example, has a very clean gradient. With similar settings, the face of the Yemeni presents an embossed look. The stylization of the tomato image conveys a three-dimensional impression of the scene. Using different settings, thresholding can produce a striking black and white effect in some images, such as the cat; in others, such as the rim lighting image, the effect is harsh. We imagined that the high contrasts of the rim lighting image would make it an easy case for limited-tone rendering, but both XDoG and the minimal rendering algorithm struggle to produce attractive output.

### *6.1.2. Structure-preserving Stippling*

The structure-preserving stippling method thrives on image contrast. **Complex edges** are crisply preserved where the contrast is sufficient. Similarly, given enough contrast, **fine details** and **texture** are conveyed well: consider the weave pattern in the toque image, and the rock texture in Oparara. Quite small details can be shown, such as the text on the button on the rim lighting image (not quite legible).

Stipples can show fine details and can change density to show gradients, and hence are effective at representing varying levels of detail in the images. The angel and arches image show this nicely, where the texture on the rocks as well as larger structures in the image are both communicated effectively. As mentioned above, even quite thin features such as the cat's whiskers or the glasses frames in the Mac image can be shown well; the

whiskers, though, appear as negative space surrounded by stipples, and it might be worthwhile to consider intensity inversion for such structures in future research.

**Low key** images represent an open problem for stippling: stippling algorithms notoriously have difficulty portraying very dark areas. An extremely high density of stipples is needed in order to accurately reproduce the darkness, and sometimes irregular holes appear between stipples, giving dark regions a mottled look. Conventional stippling, with dark stipples on a light background, can only convey light objects by omitting stipples, leaving a blank area. For example, the rim lighting image is mostly full of stipples, with the rim itself shown using the white paper behind. In this particular case, the problem can readily be solved by reversing the stipple and paper colors, with light stipples on a dark background. However, many images contain both light-on-dark and dark-on-light; consider the barn image, where the foreground barn is nicely shown by dark stipples, but the white clouds are not as neatly communicated. The mountains image poses a similar problem, but here the issue is even more pronounced because of the low contrast between the sky, snow, and clouds. It is not clear how best to use stipples to convey variable image details such as these.

While **high key** images are less troublesome than low key images, they can still pose a problem: very light areas may receive few stipples, so that detail is lost. The daisy image is overall very light and its content is difficult to discern in the stippled version. This is partially due to the low contrast in the image and partly because of a misguided effort to preserve the original greylevel: bright regions receive a low stipple count, so that details cannot be seen. In the snow image, for example, there are too few stipples to show the tree branches. The daisy image is almost unrecognizable except for the darker details at the flower's centre. As noted above, low intensity can also be a problem, as the stipples crowd together and their pattern fails to reveal the underlying image structure. The solution would seem to be to preprocess the image to ensure a good distribution of greylevels; although histogram equalization or manual retouching of the inputs are common, there does not appear to be a systematic treatment of preprocessing techniques for stippling algorithms.

A faithful reproduction of details means that **clutter** is not treated well either. In the athletes image, for example, the bokeh appears as a distracting background pattern. The background objects in the cat image remain to clutter the image. Reproduction of greylevels with an irregular stipple distribution also means that flat image regions (e.g., the sky in the arches image) which were formerly uncluttered now potentially exhibit spurious texture; this is particularly objectionable in faces, such as the toque image.

*6.1.3. Lines and Tonal Blocks*

Rosin and Lai's minimal rendering algorithm has two main components: tonal blocks, where graph cuts provide a coarse separation into foreground and background, and line drawing, where ETF-based lines are superimposed on the blocks. The blocks provide larger-scale structure and the lines give detail.

The tonal blocks are effective in images where a clear subject can be automatically deduced. The cat, rim lighting, and especially the toque image show good separation. Even in cases where the cut does not distinguish foreground and background, it can help to guide the viewer's attention in the image: in the cabbage image, for example, some leaf boundaries are enhanced by the cut. The daisy image is similar, where the cut helps to emphasize image features.

The lines are used to show local and small-scale details. In cases with high contrast and clear edges, such as the grille in the headlight image, the method is effective. Conversely, where the edges are less clear, such as the mountains image, the lines do not fully convey the image content. Similarly, the stylization of the tomato image flattens the image, removing much sense of its three-dimensional quality. In other cases, the tonal block helps to provide contrast; for example, the snow image is made more effective because the background trees are rendered in a sketchy manner, and the gray tonal block separates them from the foreground.

**Complex edges** are smoothed: the treeline in the mountains image is somewhat present, but simplified. Similarly, in regions of **irregular texture**, the smoothed direction field produces a confused sense of the image; the arch and darkwoods images provide examples.

**Thin features** are shown using the ETF lines. Line inversion helps to ensure that details remain visible: for example, the trees of the snow image are shown clearly. The ETF is less reliable for indistinct and chaotic edges, such as those around the turnip root in the tomato image or the clouds in the mountains and barn images.

Because of its reliance on thresholding and segmentation, the algorithm struggles with **long gradients**. The headlight image, though overall successful, has had the gradient removed. Gradients in the skies of the barn, mountains, and arch images have been obliterated. The thresholding can also be a problem in cluttered images, such as the berries, or when a large-scale intensity difference is present but need not be emphasized, such as in the barn image.

The modified algorithm merges results from two distinct input sizes, helping it to manage issues of **scale**. Both large and small details are apparent in images with variable scale such as the angel and arches images. The method might benefit from additional scales, allowing it to preserve yet finer details in these and other images, although presenting such details would move the method further away from its goal of "minimal rendering". In general, we think that issues of scale have been underserved by automatic image stylization methods.

**Faces** receive special treatment by the algorithm, with a dedicated face detector helping to shape the tonal blocks around faces. The rim lighting image has a good outcome, with the third tonal level helping to illustrate the interior shading in the face region. The toque image conveys the facial features clearly. The XDoG algorithm also works effectively on this example, whereas the result from stippling has less distinct facial features, and the face is not well separated from the background. However, some of the faces in the benchmark set are very challenging for face detectors: the Mac image, with its partly-occluded face, and the Yemeni image, with its severe lighting and other complications, proved especially problematic.

Probably the most effective results are obtained when the images contain strong geometric structures, such as in the headlight image. The method is less successful when strong structures are missing and the image becomes difficult to segment. Scenes containing mainly texture, such as the darkwoods image, are the most challenging for this method, and probably for other segmentation-based methods as well.

### *6.2. Abstraction*

We first discuss characteristics common to all three abstraction algorithms.

All three methods treat **fine structure** and small details in a similar way. An apparent **characteristic scale** is present in each method: details below a certain size become muted or are eliminated. Therefore, for images such as the headlight which contain little fine detail, the degree of stylisation produced by the three methods is not substantial.

The characteristic scale is particularly pronounced in the case of Papari et al.'s method, where new features of a particular size are sometimes created; these can be seen, for example, on the wreath in the angel image and in the background of the athletes image. In the range geodesic method, the scale is present as a region size, but regions can have arbitrary shapes, so that the scale is not as immediately visually identifiable as in the other two methods. The results of Papari et al.'s and Mould's methods are similar for some images, such as athletes, daisy, and tomato; however, they differ in their treatment of fine detail, which is particularly evident in images such as darkwoods and snow.

All three methods produce full-color output images. None depend on edge detection or thresholds per se, so neither high key nor low key images provide particular challenges. The methods have difficulty adequately abstracting textures, fine details, and thin features. We now turn to discussing the results from each individual algorithm.

#### *6.2.1. Texture removal*

This method's most prominent characteristic is its effect on texture and small details. **Texture** is removed, a deliberate choice on the part of the authors. However, it is not replaced with gradients or otherwise smooth structures. Rather, irregular convex blobs are created in textured areas: this behaviour is particularly noticeable in the arches and dark woods images. In general, **fine details** are removed when they are smaller than the method's scale, and even **gradients** are sometimes replaced with semi-quantized blotches, as seen in the berries and tomato images. Similarly, **variation in scale** is not entirely captured by this method: at scales above the filter size, structures are preserved, and smaller structures are simplified or removed. The arches image provides a good demonstration.

**Regular structure** at a small scale, such as outside the window of the city image, is replaced by irregular structure. The same effect is observed in **faces**: irregular blotches at the method's characteristic scale appear. The Yemeni and Mac images show this very strongly, but it is present to a degree in the toque image and on the faces of the athletes.

**Thin features** are handled well when they are distinct, such as the frames of the glasses in the Mac image. Less distinct thin features, such as the tree branches in the snow image or the cat whiskers, become blurred.

Simple edges with high contrast are preserved. **Complex edges** are altered, being generally simplified to remove details of very high frequency; consider the tree silhouettes in the mountains image, which remain sharp while having their shape smoothed. Because of the detail removal imposed by this method, strong edges can even be clarified; the rock silhouette against the sky of the arches image provides an example. Less distinct complex edges, such as the Yemeni's beard or the silhouette of the cat's fur, are blurred. Nonetheless, **low contrast** images such as the daisy and mountain images are handled nicely by this method. The petals of the daisy are quite distinct in the processed image.

*6.2.2. Geodesic filtering*

This method, like that of Papari et al., shows a characteristic scale: with our parameter setting of $n = 240$, the scale is approximately a radius of 9 pixels. The characteristic scale can be seen in the Yemeni's head covering: sufficiently large markings are preserved, while smaller ones fade. As this example shows, even high-contrast features can be removed from the image when they are too small. Where the image lacks small details and textures, the image is hardly altered at all, perhaps making the viewer question whether any stylization effect has been deployed.

**Textures** are not preserved; the intent was to suggest them, and while this objective is partially met in some cases (such as the fur of the cat), very high-frequency textures with high contrasts stand out in an unappealing way after processing. The problem arises in regions where tiny features are located near a reservoir of pixels of similar color: such features retain their original color while the surroundings are smoothed, yielding spurious shapes with high visibility. The effect can be seen in the leaves of the dark woods and in the rock texture in the arch image. Where no such reservoirs exist, such as in the middle of the toque, high-frequency elements are dimmed.

Similarly, but with better visual effect, **complex edges** are preserved: for example, the tree silhouettes in the mountain image retain their complexity and sharpness. Even very small details, such as the fuzz of the toque or the edge of the hat in the rim lighting image, can be preserved. Again, though, the contrast must be sufficiently high; the complex fur silhouette in the cat image is not very visible in the output.

**Smooth gradients** remain clean. If very small, they can be slightly flattened (e.g., the highlights on the tomato image) but larger-scale gradients are preserved.

**Thin features** are only preserved when the contrast is high and the feature is large enough. The glasses in the Mac image remain, as do the details of the cabbage leaves, but many tree branches in the barn and snow images are blurred until they become nearly unrecognizable. Other fine but low-contrast features – the roots of the turnip in the tomato image, and the branches of the bush in the desert – are equally poorly preserved.

**Color** is preserved where there are regions of solid color. Small colored regions can blend together, or become muddied by their surroundings; the bright greens in the dark woods image are darkened, for example. The effect is sometimes subtle, but it always pushes in the direction of making the image's color duller.

Detail in areas of **low contrast** is sometimes preserved, such as in the daisy image, when there is no competition from stronger contrast nearby. In general, though, there is room for improvement in this method's ability to handle low contrast.

*6.2.3. Oil painting*

**Contrast** is generally handled well, owing to the color quantization. The highlights in the angel are strong. The lighting variation in the Oparara image is quite striking. However, contrast enhancement through quantization does not always produce desirable outcomes: the face in the toque image is now unnaturally pale, even ghostly. Low contrast, conversely, is eliminated. Much of the structure of the daisy has disappeared. In the mountain image, the clouds and mountain peak merge confusingly. In the tomato image, the strokes generally provide a good impression of the tomato shapes, but weak contrast has caused the leftmost tomato to merge with the partially visible tomato behind it, distorting the shape peculiarly.

**Gradients** suffer from spurious flow lines; the headlight image and the desert image show this well. The directionality of the flow field is helpful in communicating the image content in regions where a clear direction is present, such as the fur of the cat or the beard of the Yemeni. However, in regions of **clutter** or high detail, the direction field becomes confused: for example, spurious flow lines appear prominently in the background of the athletes image. Similarly, the direction field can become confused in flat areas. An example can be seen in the snow image, where the strokes above the trees have been extended in a uniform vertical pattern, which is distracting.

**Fine detail** can also confuse the direction field: the field has a characteristic scale, and features smaller than this scale cause trouble. The branches of the tree in the desert image provide an example. Two aspects of the algorithm act to remove fine detail: first, the directional blurring, and second, the addition of stroke and canvas texture, which overwhelm any detail that remains. Similarly, complex edges such as the leaf margin in the cabbage image are simplified by this approach.

**Colors** in the image are well-preserved, and even enhanced by the palettization process. The yellows in the berries image, for example, are warmer than in the original. The palettization is less effective in areas of **low key**, however, darkening the image too aggressively. The dark woods image, for example, has become much more difficult to appreciate after stylization. Similarly, the face in the rim image does not benefit from color reduction.

**Faces** are in general not treated well by this approach. Fine facial details are not necessarily preserved, and spurious texture can be distracting. For example, the face of the Yemeni has been covered by stroke texture.

## 7. Effectiveness of the benchmark

In the preceding section, we described the outcome of an exercise where we applied several stylization algorithms to the benchmark images and then inspected those images and reported our observations. Our hope is that later researchers will be inspired to follow suit when presenting their new stylization algorithms. In this section, we reflect on the outcome of the exercise and discuss the strengths and weaknesses of the benchmark.

Overall, we found the exercise to be useful. The variety of scenarios in the benchmark set provided ample raw material for revealing varied behaviours in the stylization algorithms. We found the athletes, cat, mountains, and Yemeni images generally challenging for the algorithms and hence these images provided greater insight and potential grounds for future work in stylization. Conversely, all methods gave good results on the headlight image. The tomato image, while not particularly problematic, does contain a range of features to which different methods responded differently. Although not all images were informative for all algorithms, different images proved informative for different algorithms and for different reasons.

In the main, the algorithms worked quite well on this dataset, generating appealing stylized output. The basic role of the benchmark as a sanity check is thus successful. Beyond this, though, the benchmark allowed us to test the algorithms thoroughly. The variety of textures – stone, leaves, fur, fabric, clouds, and cityscape, among others – gave us a nuanced picture of the behaviour of the algorithms over different shapes and scales of details in the input. The broad range of tone and color in the input images was very informative for evaluating the reduced-palette algorithms, and somewhat unexpectedly for Semmo et al.'s oil painting stylization as well. Also, to our knowledge, no one has previously identified the challenge of using stipples to portray a light foreground against a dark background.

Through the exercise, we noticed some shortcomings in the benchmark. Clutter, text, corners, and faces are underrepresented in the current dataset. Complex edges mostly coincide with thin features; rare exceptions include the treeline in the mountains image and the cleats in the athletes image. It would therefore be beneficial to have additional examples of high-contrast complex boundaries.

The representation of clutter is a possible weak point in the benchmark. We should add some more definitely cluttered photos in order to better reflect everyday photos; amateur photographs are often taken with cell phones whose optics produce a very broad depth of field, with many extraneous elements in focus.

The images generally lack precise symbolic content, such as text, characters, or digits, which could be strictly evaluated for post-stylization readability. Very thin high-contrast elements are lacking. Also, sharp corners are rare in this image set. This set of limitations could potentially be solved simultaneously, with the addition of urban scenes containing street signs and similar elements.

We did include faces in several images: the Mac, rim lighting, toque, and Yemeni images are portraits, and the angel, city, and athletes images contain faces at smaller scales. Still, a greater range of facial types, scales, and facial expressions would be welcome. In part, this should be resolved with a dedicated portrait database. Even given a portrait database, though, the general database would benefit from at least one image containing several faces at different scales.

## 8. Conclusion and Future Work

We presented NPRgeneral, a set of benchmark images to test image stylization algorithms; more importantly, we articulated a set of considerations that can guide the development of future benchmark sets. The image set should be large enough to include all the features of interest, but not so large that it becomes unwieldy for manual assessment. Features of interest include low-level features such as variable contrast and high-frequency structure, as well as high-level features such as human faces and clutter. Pragmatically, the images should be of adequate resolution and must be free of copyright encumbrances that would prevent distribution of modified images.

As the benchmark will only be beneficial when researchers use it, this paper is also a call to action to the community to take up the benchmark and report the results of new and old algorithms. Researchers can showcase benchmark results in their papers as well as hosting the results of the full benchmark on a project page.

The set of images presented in this paper are "version 1.0" of the benchmark. The benchmark images themselves and some additional example stylizations will be publicly available from `expressive.graphics/`

benchmark as well as from gigl.scs.carleton.ca/benchmark.

Future versions of the benchmark may be extended to include video, or other image types such as depth images or plenoptic images. Often, stylization methods build on standard methods; we can facilitate comparisons by providing salience maps and pre-segmented images, for example, and perhaps other forms of standard preprocessing would be helpful. Additional metadata in the form of manual annotations of the images – e.g., manual foreground/background segmentation, or labelings of regions of interest – could be included in later versions of the benchmark.

**Acknowledgements**

All original images came from Flickr, provided by these photographers: Eole Wind (angel), Nathan Congleton (athletes), MrClean1982 (barn), HelmutZen (berries), Leonard Chien (cabbage), Theen Moy (cat), Rob Schneider (city), mgaloseau (daisy), JB Banks (dark woods), Charles Roffey (desert), Photos By Clark (headlight), Martin Kenney (Mac), Jenny Pansing (mountains), trevorklatko (Oparara), James Marvin Phelps (arch), Paul Stevenson (rim lighting), John Anes (snow), Greg Myers (tomatoes), sicknotepix (toque), Richard Messenger (Yemeni).

Thanks go out as well to Hua Li, Holger Winnemöller, and Amir Semmo for providing us with the results for the stippling, XDoG, and oil painting algorithms, and further thanks to Papari et al. for providing code to run their method.

Figure 1: The set of 20 benchmark images.

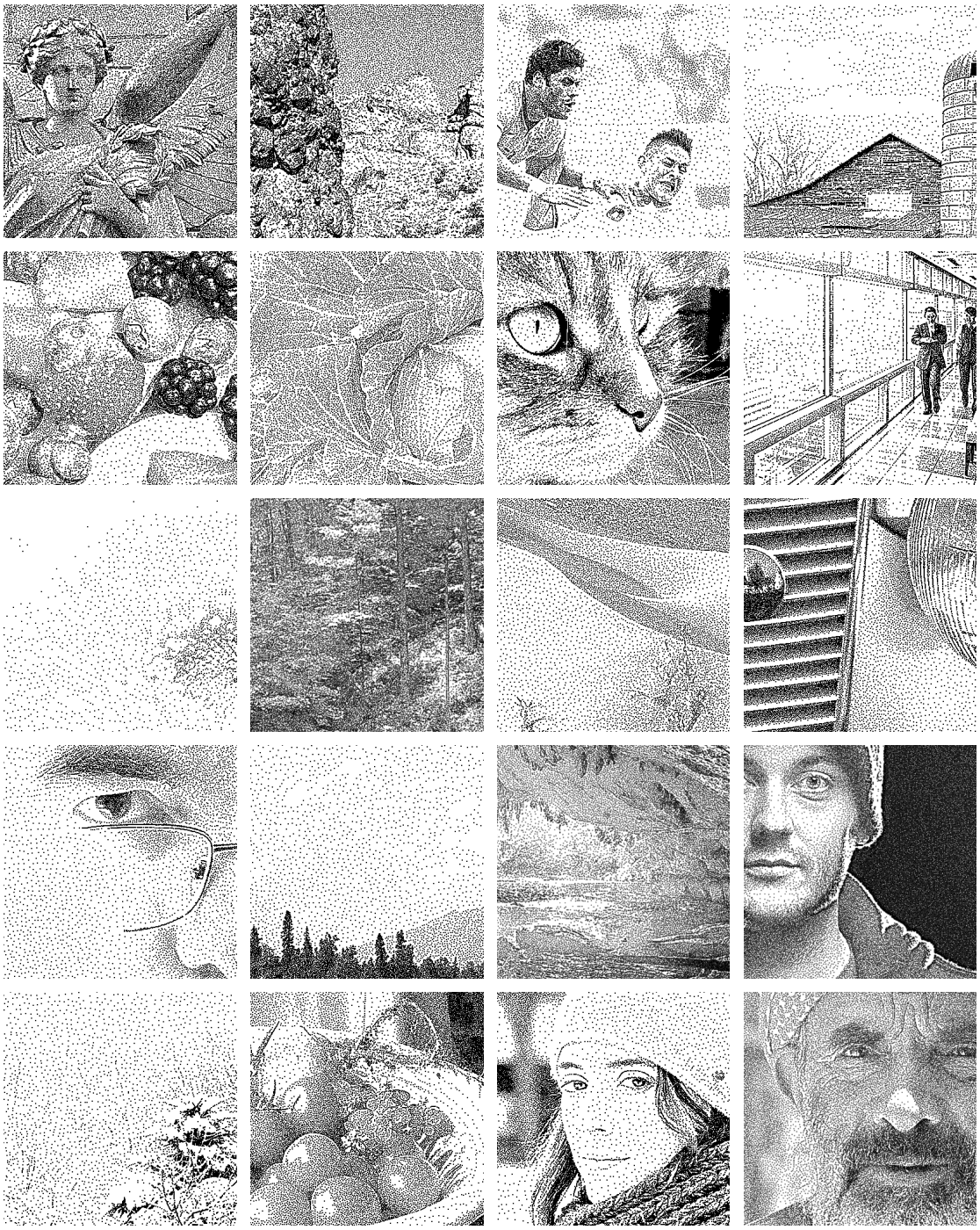

Figure 2: Details of benchmark images stylized with structure-preserving stippling [23].

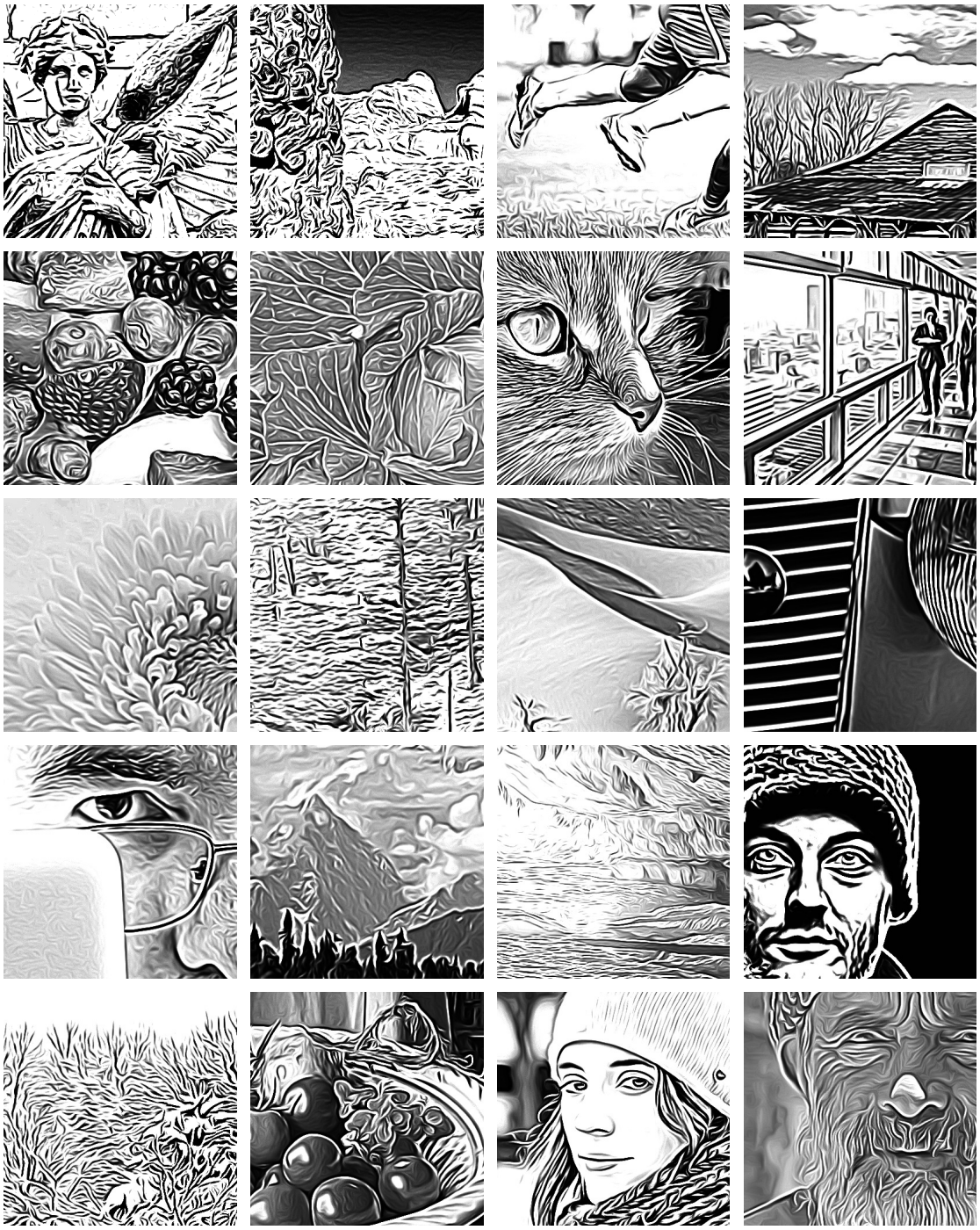

Figure 3: Details of benchmark images stylized with XDoG [25].

Figure 4: Details of benchmark images stylized with minimal rendering [13].

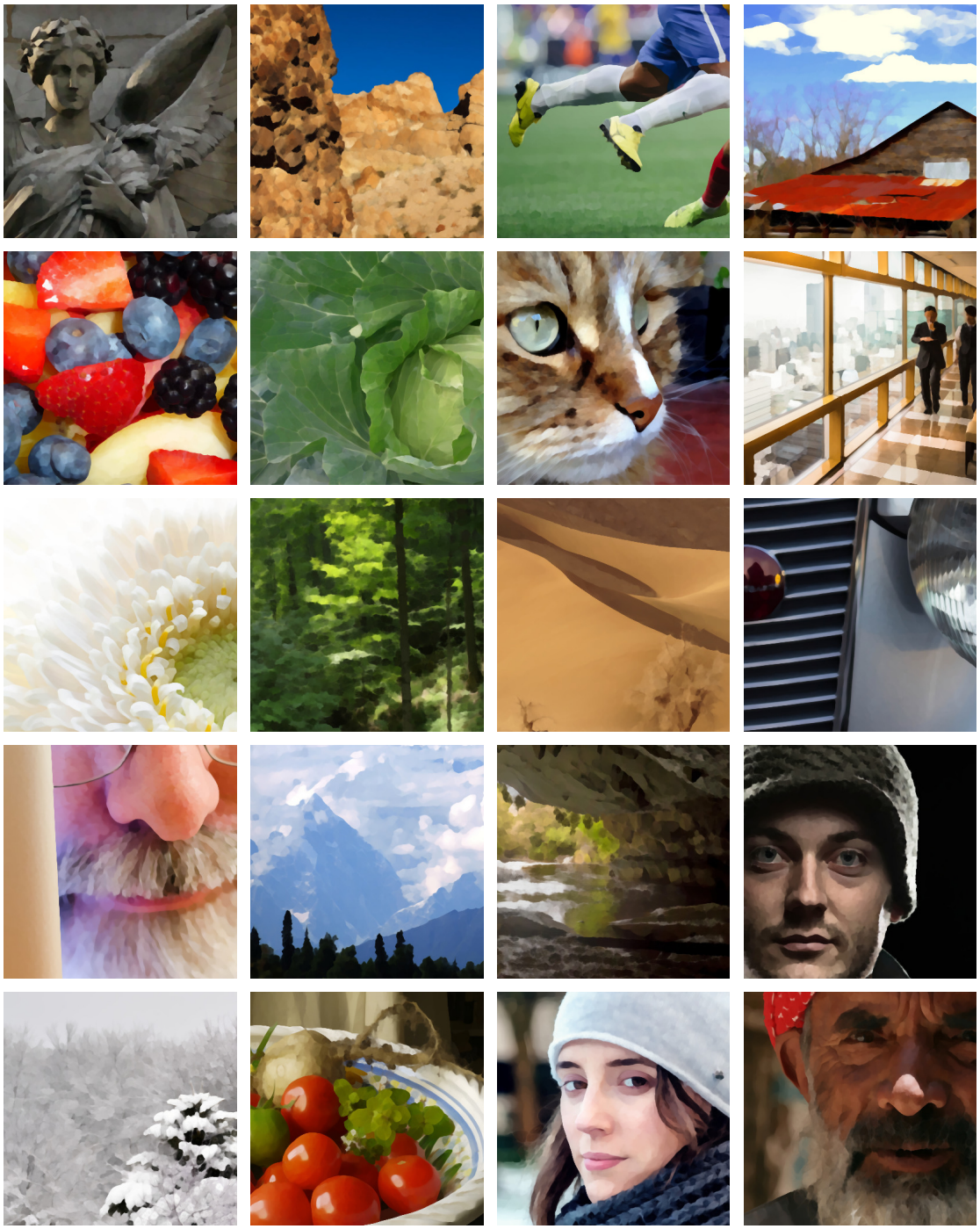

Figure 5: Details of benchmark images stylized with Papari et al.'s texture-removing abstraction [28].

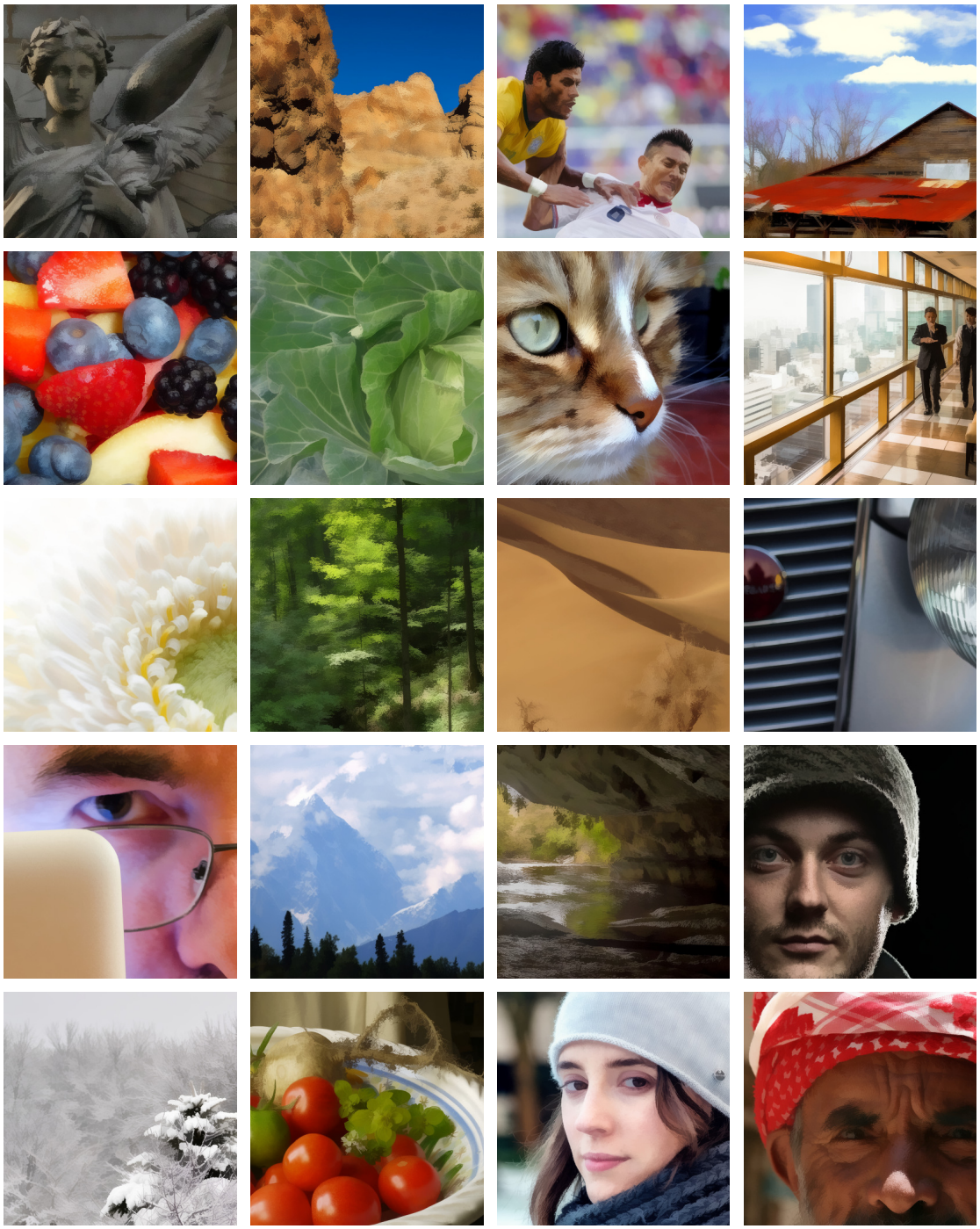

Figure 6: Details of benchmark images stylized with geodesic abstraction [29].

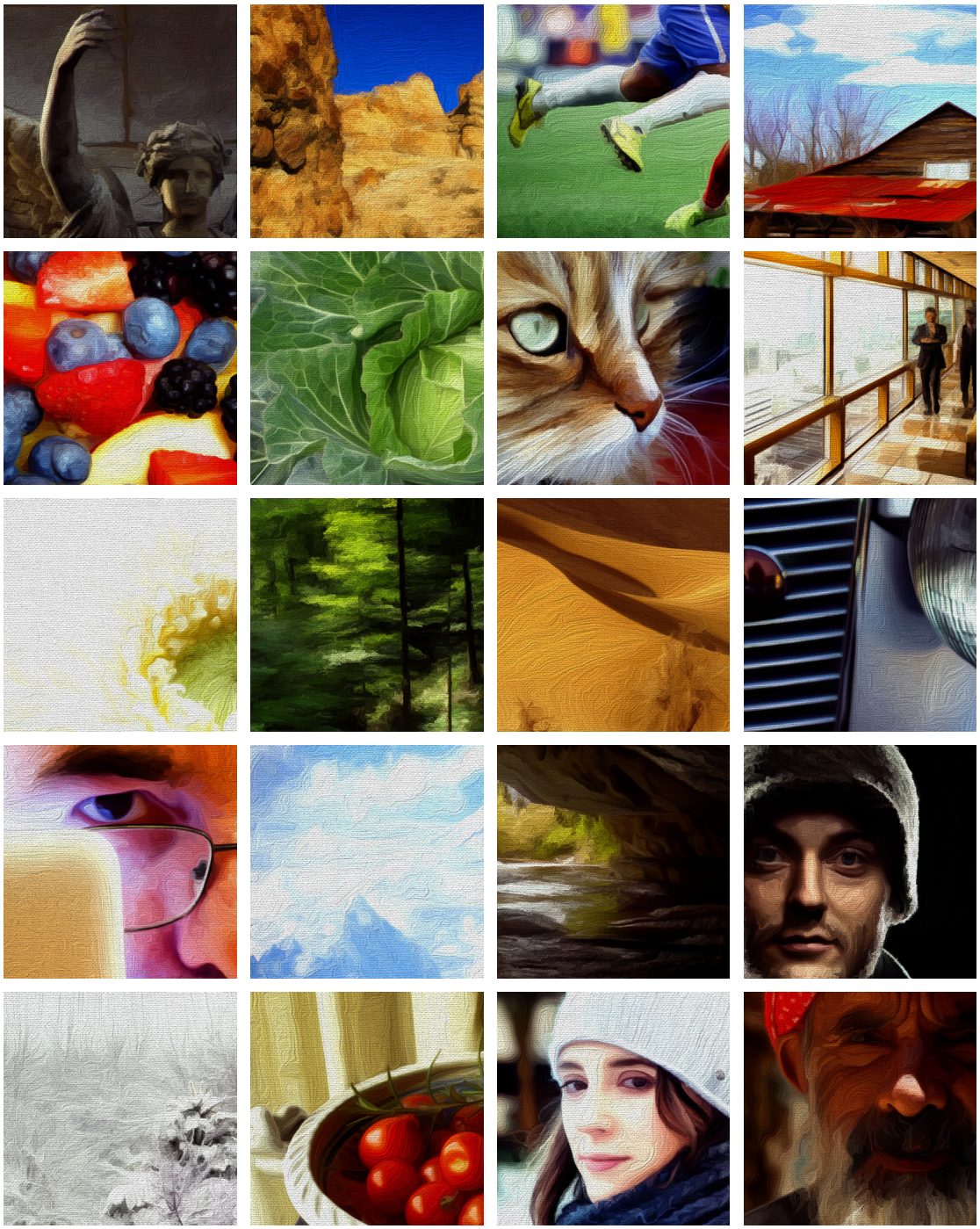

Figure 7: Details of benchmark images stylized with oil painting [30].

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

[2] Isenberg, T.. Evaluating and validating non-photorealistic and illustrative rendering. In: Rosin, P.L., Collomosse, J.P., editors. Image and Video-Based Artistic Stylisation. Springer; 2013, p. 311–331.

[3] Mould, D.. Authorial subjective evaluation of non-photorealistic images. In: Proceedings of the Workshop on Non-Photorealistic Animation and Rendering. NPAR '14; New York, NY, USA: ACM. ISBN 978-1-4503-3020-6; 2014, p. 49–56.

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

[5] Hall, P., Lehmann, A.S.. Don't measure – appreciate! NPR seen through the prism of art history. In: Rosin, P.L., Collomosse, J.P., editors. Image and Video-Based Artistic Stylisation. Springer; 2013, p. 333–351.

[6] Mould, D., Rosin, P.L.. A benchmark image set for evaluating stylization. In: Proceedings of the Joint Symposium on Computational Aesthetics and Sketch Based Interfaces and Modeling and Non-Photorealistic Animation and Rendering. Expresive '16; Aire-la-Ville, Switzerland, Switzerland: Eurographics Association; 2016, p. 11–20.

[7] Gatzidis, C., Papakonstantinou, S., Brujic-Okretic, V., Baker, S.. Recent advances in the user evaluation methods and studies of non-photorealisitic visualisation and rendering techniques. In: Proc. Info. Vis. 2008, p. 475–480.

[8] Isenberg, T., Neumann, P., Carpendale, S., Sousa, M.C., Jorge, J.A.. Non-photorealistic rendering in context: an observational study. In: ACM Symp. NPAR. 2006, p. 115–126.

[9] AlMeraj, Z., Wyvill, B., Isenberg, T., Gooch, A.A., Guy, R.. Automatically mimicking unique hand-drawn pencil lines. Computers & Graphics 2009;33(4):496–508.

[10] Winnemöller, H., Olsen, S., Gooch, B.. Real-time video abstraction. ACM Trans Graphics 2006;25(3):1221–1226.

[11] Gooch, B., Reinhard, E., Gooch, A.. Human facial illustrations: Creation and psychophysical evaluation. ACM Trans Graph 2004;23(1):27–44.

[12] Zhao, M., Zhu, S.C.. Abstract painting with interactive control of perceptual entropy. ACM Transactions on Applied Perception (TAP) 2013;10(1):5.

[13] Rosin, P.L., Lai, Y.K.. Artistic minimal rendering with lines and blocks. Graphical Models 2013;75(4):208–229.

[14] Bousmalis, K., Mehu, M., Pantic, M.. Towards the automatic detection of spontaneous agreement and disagreement based on nonverbal behaviour: A survey of related cues, databases, and tools. Image Vision Computing 2013;31(2):203–221.

[15] Brodatz, P.. Textures : a photographic album for artists and designers. New York: Dover Publications; 1966. ISBN 0-486-21669-1.

[16] Sim, T., Baker, S., Bsat, M.. The CMU pose, illumination, and expression database. IEEE Trans Pattern Anal Mach Intell 2003;25(12):1615–1618.

[17] Xiao, J., Hays, J., Ehinger, K.A., Oliva, A., Torralba, A.. SUN database: Large-scale scene recognition from abbey to zoo. In: Conference on Computer Vision and Pattern Recognition. 2010, p. 3485–3492.

[18] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., et al. ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision 2015;115(3):211–252.

[19] Orenstein, E.C., Beijbom, O., Peacock, E.E., Sosik, H.M.. WHOI-Plankton – A large scale fine grained visual recognition benchmark dataset for plankton classification. CoRR 2015;abs/1510.00745.

[20] Thomee, B., Shamma, D.A., Friedland, G., Elizalde, B., Ni, K., Poland, D., et al. YFCC100M: The new data in multimedia research. Commun ACM 2016;59(2):64–73.

[21] Lai, Y.K., Rosin, P.L.. Image and Video-Based Artistic Stylisation; chap. Non-photorealistic Rendering with Reduced Colour Palettes. Springer; 2013, p. 211–236.

[22] Deussen, O., Hiller, S., Van Overveld, C., Strothotte, T.. Floating points: A method for computing stipple drawings. Computer Graphics Forum 2000;19(3):41–50.

[23] Li, H., Mould, D.. Structure-preserving stippling by priority-based error diffusion. In: Proceedings of Graphics Interface 2011. GI '11; School of Computer Science, University of Waterloo, Waterloo, Ontario, Canada: Canadian Human-Computer Communications Society. ISBN 978-1-4503-0693-5; 2011, p. 127–134.

[24] Li, H., Mould, D.. Contrast-aware Halftoning. Computer Graphics Forum 2010;.

[25] Winnemöller, H., Kyprianidis, J.E., Olsen, S.C.. Xdog: An extended difference-of-gaussians compendium including advanced image stylization. Computers & Graphics 2012;36(6):740–753.

[26] Kang, H., Lee, S., Chui, C.. Coherent line drawing. In: ACM Symp. Non-photorealistic Animation and Rendering. 2007, p. 43–50.

[27] Selinger, P.. Potrace: a polygon-based tracing algorithm; 2003. URL `http://potrace.sourceforge.net/`.

[28] Papari, G., Petkov, N., Campisi, P.. Artistic edge and corner enhancing smoothing. IEEE Transactions on Image Processing 2007;16(10):2449–2662.

[29] Mould, D.. Image and video abstraction using cumulative range geodesic filtering. Computers & Graphics 2013;37(5):413–430.

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

[31] Murray, N., Marchesotti, L., Perronnin, F.. AVA: A large-scale database for aesthetic visual analysis. In: Conf. Computer Vision and Pattern Recognition. 2012, p. 2408–2415.

[32] Rosin, P.L., Mould, D., Berger, I., Collomosse, J.h., Lai, Y.K., Li, C., et al. Benchmarking non-photorealistic rendering of portraits. In: Proceedings of the Joint Symposium on Computational Aesthetics and Sketch Based Interfaces and Modeling and Non-Photorealistic Animation and Rendering. Expresive '17; ACM; 2017,.

[33] Hasler, D., Süsstrunk, S.. Measuring colourfulness in natural images. In: Proc. SPIE Human Vision and Electronic Imaging. 2003, p. 87–95.

[34] Machado, P., Cardoso, A.. Computing aesthetics. In: Brazilian Symposium on Artificial Intelligence. 1998, p. 219–228.

[35] Matković, K., Neumann, L., Neumann, A., Psik, T., Purgathofer, W.. Global contrast factor – a new approach to image contrast. In: Computational Aesthetics. 2005, p. 159–167.

[36] Bahrami, K., Kot, A.C.. A fast approach for no-reference image sharpness assessment based on maximum local variation. IEEE Signal Processing Letters 2014;21(6):751–755.

[37] Redi, M., Rasiwasia, N., Aggarwal, G., Jaimes, A.. The beauty of capturing faces: Rating the quality of digital portraits. In: Conf. on Automatic Face and Gesture Recognition. 2015, p. 1–8.

[38] Immerkaer, J.. Fast noise variance estimation. Computer Vision and Image Understanding 1996;64:300–302.