PTRM: Perceived Terrain Realism Metric

SUREN DEEPAK RAJASEKARAN and HAO KANG, Purdue University, USA
MARTIN ČADÍK, FIT, Brno University of Technology, and FEL, Czech Technical University, Czech Rep.
ERIC GALIN, ERIC GUÉRIN, and ADRIEN PEYTAVIE, Université de Lyon, France
PAVEL SLAVÍK, FEL, Czech Technical University, Czech Rep.
BEDRICH BENES, Purdue University, USA

Terrains are visually prominent and commonly needed objects in many computer graphics applications. While there are many algorithms for synthetic terrain generation, it is rather difficult to assess the realism of a generated output. This article presents a first step toward the direction of perceptual evaluation for terrain models. We gathered and categorized several classes of real terrains, and we generated synthetic terrain models using computer graphics methods. The terrain geometries were rendered by using the same texturing, lighting, and camera position. Two studies on these image sets were conducted, ranking the terrains perceptually, and showing that the synthetic terrains are perceived as lacking realism compared to the real ones. We provide insight into the features that affect the perceived realism by a quantitative evaluation based on localized geomorphology-based landform features (geomorphons) that categorize terrain structures such as valleys, ridges, hollows, and so forth. We show that the presence or absence of certain features has a significant perceptual effect. The importance and presence of the terrain features were confirmed by using a generative deep neural network that transferred the features between the geometric models of the real terrains and the synthetic ones. The feature transfer was followed by another perceptual experiment that further showed their importance and effect on perceived realism. We then introduce *Perceived Terrain Realism Metrics* (PTRM), which estimates human-perceived realism of a terrain represented as a digital elevation map by relating the distribution of terrain features with their perceived realism. This metric can be used on a synthetic terrain, and it will output an estimated level of perceived realism. We validated the proposed metrics on real and synthetic data and compared them to the perceptual studies.

CCS Concepts: • **Computing methodologies** → **Perception**; *Procedural modeling*; Shape analysis;

Additional Key Words and Phrases: Procedural modeling, terrains, visual perception, feature transfer, neural networks

ACM Reference format:
Suren Deepak Rajasekaran, Hao Kang, Martin Čadík, Eric Galin, Eric Guérin, Adrien Peytavie, Pavel Slavík, and Bedrich Benes. 2022. PTRM: Perceived Terrain Realism Metric. *ACM Trans. Appl. Percept.* 19, 2, Article 6 (July 2022), 22 pages.
https://doi.org/10.1145/3514244

This research was funded in part by National Science Foundation grants #10001387, *Functional Proceduralization of 3D Geometric Models*, and project HDW ANR-16-CE33-0001. This work was further supported by project no. LTAIZ19004 *Deep-Learning Approach to Topographical Image Analysis*; by the Ministry of Education, Youth and Sports of the Czech Republic within the activity INTER-EXCELENCE (LT), subactivity INTER-ACTION (LTA), ID: SMSM2019LTAIZ; and by *Research Center for Informatics No. CZ.02.1.01/0.0/0.0/16_019/0000765.*
Authors' addresses: S. D. Rajasekaran, H. Kang, and B. Benes, Department of Computer Science Purdue University, 305 N University St., West Lafayette, IN, 47907-2021, USA; emails: surendeepak.rajasekaran@gmail.com, kayheseri@gmail.com, bbenes@purdue.edu; M. Cadik, FIT, Brno University of Technology, Božetěchova 2, 612 00 Brno, Czech Republic; email: cadik@fit.vutbr.cz; E. Galin, E. Guérin, and A. Peytavie, Laboratoire LIRIS - CNRS, Université Claude Bernard Lyon 1, France; emails: eric.galin@univ-lyon1.fr, eric.guerin@insa-lyon.fr, adrien.peytavie@liris.cnrs.fr; P. Slavík, Praha 2, Karlovo náměstí 13, E-321, Czech Republic; email: slavik@fel.cvut.cz.

1544-3558/2022/07-ART6
https://doi.org/10.1145/3514244

1 INTRODUCTION

Terrains are among the most visually stunning structures, and their modeling has attracted researchers' attention for decades. The patterns in terrains result from eons of complex, interacting geomorphological processes with varying strength at differing spatial and temporal scales, which makes them difficult to simulate. Humans experience terrains throughout their lifetime, and our visual perception system has evolved into a precise tool for judging terrain realism. Humans can quickly detect anomalies [Travers 1984], such as inconsistent rivers, non-realistic shapes of mountains, or incorrectly positioned terrain features. This makes synthetic terrain modeling challenging, and quantifying those inconsistencies remains highly complex.

Although a wide variety of algorithms exist for terrain modeling (see Galin et al. [2019]), existing methods often consider the geomorphological phenomena in separation, and their mutual dependencies are neither well studied nor understood. Previous methods focused on replicating phenomenological processes of terrain formation, but none, to the best of our knowledge, have focused on their perceived realism. Evaluating the results of algorithms simulating natural phenomena has always been a difficult question and is usually addressed by providing a side-by-side comparison of the generated structures or is assumed to be correct if the underlying simulations are physically based. This article attempts to answer two questions: (1) what visually important terrains features make them realistic? and (2) what is the perceived realism of synthetic terrains generated by computer graphics techniques?

Work in geology by Jasiewicz and Stepinski [2013] allows for quantitative classification of terrains according to their geomorphological features such as valleys, ridges, slopes, spurs, hollows, and so forth. It categorizes them into so-called geomorphons and quantifies their presence in the **digital elevation map (DEM)**. We gathered a large amount of DEMs of real and synthetic terrains. We rendered the terrains from an ordinary human perspective (see Figure 1) by using constant texturing and lighting. We performed a user study measuring the perceived realism of real and synthetic terrains, and we related the perceived realism to geomorphons and the geomorphological features. To further validate that some geomorphological features are visually more critical for perceived realism, we used the state-of-the-art deep neural networks, CycleGAN [Zhu et al. 2017] and transferred features (valleys, ridges, etc.) from the DEMs that were ranked high in terms of realism to those ranked low and vice versa. It is important to note that the transfer did not work with the rendered images of terrains but with the DEMs representing terrain geometry. The DEMs were stored as grayscale images. We then performed another user study that showed that the landforms transferred from highly ranked sets to lowly ranked ones improve the perceived realism and that the landforms transferred from low-ranked images to high-ranked ones demote them perceptually. The two user studies combined with the analysis of features show that synthetic terrains do not often include geomorphological features such as depressions, peaks, flats, valleys, ridges, hollows, and spurs, which are important for perceived realism. Eventually, we introduce **Perceived Terrain Realism Metrics (PTRM),** which assigns a normalized value of perceived perception to a terrain represented as a DEM based on the present geomorphons. We validate the PTRM on both real and synthetic terrain models.

An example in Figure 1 shows a real terrain and the distribution of its landform features based on geomorphons as well as a synthetic terrain with its accompanying features. The feature vector of the geomorphons is sorted so that the ones contributing to perceived realism are on the right-hand side. The real terrain was ranked as highly realistic (77%) in our user study, and the synthetic terrain was ranked lower (51%), as can also be seen in the distribution of the geomorphons. We then used deep learning to transfer the features from the procedural terrain to real and vice versa. We show the distribution of the geographic features that indicates that the distributions of the geomorphons changed so that the high ranked worsened and the low ranked improved. This quantitative validation was then confirmed by a perceptual study that showed that the procedural terrain after the style transfer improved its perceived realism to 69%, and the real terrain worsened to 29%. We also show the PTRM that predicts how closely a person would perceive it as realistic (1 = perfect, 0 = poor).

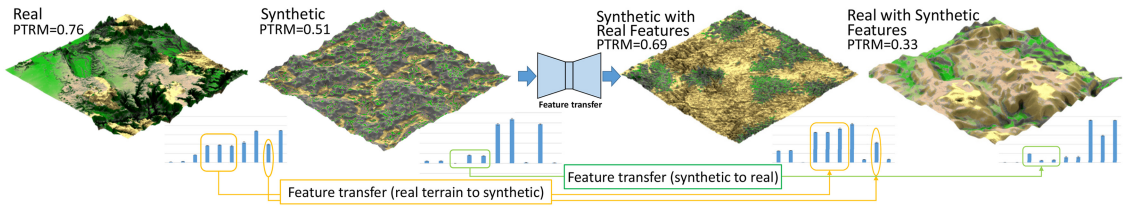

Fig. 1. The real terrain from the state of Arizona, with complex geomorphological patterns, has estimated PTRM = 0.76 (1 = realistic, 0 = poor). It has also been ranked by a perceptual study as the top 78%. The synthetic terrains with patterns generated by thermal erosion have PTRM = 0.51, and it ranked as 49% in the user study. The corresponding feature vector of geomorphons is sorted so that the features that contribute to realism are on the right. It shows the distribution of patterns in each model with a strong presence of valleys, ridges, and hollows landforms in real terrain (circled in the graph) that were not so present in the synthetic model. Using a CycleGAN, we transferred the visually important features to the low-ranked synthetic terrain (orange arrows). We transferred the features in synthetic terrain to the high-ranked real terrain (green arrow), assuming the real terrain should worsen and the synthetic should improve. The second perceptual study showed that the transferred features improved to PTRM = 0.69 (77% ranking in our study), and transferring the visually unattractive features from synthetic terrain to the real one demoted its PTRM = 0.33 (29%). The transferred features are circled in the corresponding graphs of geomorphons.

We claim the following contributions: (1) We introduce Perceived Terrain Realism Metrics, which assigns a normalized value of perceived realism to a terrain represented as a digital elevation model. (2) We conduct user studies that validate and measure the perceived realism of real and synthetic terrain models. (3) We determine geological features that affect the perceived reality of terrains. (4) We provide a publicly available dataset of real and procedural terrains with assigned perceptual evaluation and calculated geomorphons.

2 RELATED WORK

Perception-based computer graphics approaches: The knowledge of human perception has been applied in computer graphics since its beginnings (see, for example, Ferwerda [2003] and Bartz et al. [2008]). A common way is to incorporate it as a computational model of a particular **human visual system (HVS)** feature, e.g., visual attention and saliency [Frintrop et al. 2015; Riche et al. 2013], or to fully replace it by hardware such as an eye tracker [O'Sullivan et al. 2004].

Photorealistic rendering traditionally exploits perception limitations to accelerate costly light transport computations [Weier et al. 2017], and in 3D graphics, HVS models allow removing non-perceptible components [Reddy 2001] and/or predicting popping artifacts [Schwarz and Stamminger 2009].

Perceptual models have been further applied to improving virtual simulations [Ondřej et al. 2016], character animations [Reitsma and Pollard 2003; O'Sullivan et al. 2004], human body modeling [Shi et al. 2017], fluid simulations [Um et al. 2017; Bojrab et al. 2013], and crowd simulations [Wang et al. 2017]. High-dynamic-range imaging and tone mapping benefit from models of human light adaptation [Mantiuk et al. 2006]; color to gray conversions simulate human color sensitivity [Neumann et al. 2007; Smith et al. 2008].

Close to our work is research on procedural textures [Liu et al. 2015] that defines perceptual scales. The perceived quality of a geometry replaced with texture has also been studied by Rushmeier et al. [2000], and a recent work studies perception of tree models [Polasek et al. 2021].

Image quality metrics (IQM) utilize HVS models to predict perceptual image quality. Full-reference IQMs compute perceptual differences between the reference and distorted images [Mantiuk et al. 2011; Wang et al. 2004; Wolski et al. 2018], while no-reference metrics [Herzog et al. 2012; Ye et al. 2014] predict the quality in a reference-less setup. Video quality metrics [Winkler and Mohandas 2008; Aydın et al. 2010] simulate temporal HVS properties to faithfully compare video sequences.

Recent research works study the perceptual quality of 3D models [Lavoué et al. 2016] and meshes [Nader et al. 2016; Guo et al. 2015] including textured models [Guo et al. 2016]. Visual saliency predictors for 3D meshes have been also proposed [Wu et al. 2013].

Unfortunately, we cannot simply take existing metrics and apply them to compare synthetic and real terrain images or models because the terrains have different features. New metrics should be developed to work specifically with terrains and their geometric representations.

Perception of terrains: We are not aware of any computational perception quality metrics that could be applied directly to 3D terrain models. Furthermore, terrain datasets that could be used to evaluate terrain generation methods or train data-driven techniques are missing.

Nevertheless, research on the classification and perception of real-world landscape images has been presented in environmental psychology and geomorphology. In their early works, Bertilone et al. studied spatial statistics of natural-terrain airborne infrared imagery expressed as statistical correlations with perception in the temporal [Bertilone et al. 1997a] and Fourier domain [Bertilone et al. 1997b]. Dragut and Blaschke [2006] proposed a system for landform classification based on profile curvature. Several data layers are extracted from the digital terrain model to feed an image segmentation, classifying the terrain like toe slopes, peaks, shoulders, and so forth. Fractal characteristics of terrains were studied by Hagerhall et al. [2004], who concluded that there is a relationship between preference and the fractal dimension, meaning that the fractal dimension may be part of the basis for choice. Reinhard et al. [2004] studied the perception of images that did not behave as natural images and related the appearance as a second-order image statistic. They showed its applications on fractal terrains and solid textures. Finally, scenic beauty and aesthetics have been addressed by Palmer [2003], Tveit et al. [2012], Tremblet [2016], and Daniel [2001]. These works lay the foundation for landscape perception, but they cannot be directly applied to the quality assessment of synthetic 3D terrain models. Automated tools of measurement and analysis of terrains are sought [Palmer 2003] to advance this area of research. Bergen et al. [1995] studied the validity of computer-generated forest landscapes. They concluded that computer-generated imagery was not a good representation of actual photographs. Lange [2001] examined to what degree the real landscape can be represented utilizing GIS-based 3D models. Approximately 75% of observers assigned the simulated landscape the highest possible degree of realism in the perceptual experiment. Our work deals with terrains and focuses on terrain features without considering other landscape-defining features such as trees and bushes.

Terrains in computer graphics have been studied for decades (see the review [Galin et al. 2019]). Here we list the three significant categories of terrain generation techniques: procedural approaches, erosion simulation, and example based.

The first method to synthesize terrains relied on procedural and fractal approaches. Terrain models were generated so that they exhibit self-similarity either by using subdivisions [Fournier et al. 1982; Miller 1986], faulting [Mandelbrot 1988], or summing weighted noises [Musgrave et al. 1989]. Approaches that control shape [Kelley et al. 1988] or more specific curve-based constructions [Gain et al. 2009] have been introduced. The overall realism of the generated landscape depends on the fine-tuning of control parameters. It requires in-depth knowledge and understanding of the underlying generation process, which restricts those methods to skilled technical artists.

Erosion simulations generate terrain features by approximating the natural phenomena, such as hydraulic [Milne and Sears 1997; St'ava et al. 2008; Benes et al. 2006; Krištof et al. 2009] or thermal erosion [Musgrave et al. 1989; Benes and Forsbach 2002]. They are computationally intensive and only capture a limited set of small-scale structure features [Cordonnier et al. 2018], such as ravines or downstream sediment accretion regions. When combined at a larger scale with uplift [Cordonnier et al. 2016], erosion simulations generate realistic mountain ranges with dendritic ridge networks and their dual drainage network forming rivers.

Another option to obtain realism is by synthesizing new terrains by example, e.g., by stitching terrain patches from existing datasets by using techniques from texture synthesis [Zhou et al. 2007] or using the method of Guérin et al. [2016] that combines sparse modeling with realistic small-scale features. The large-scale plausibility

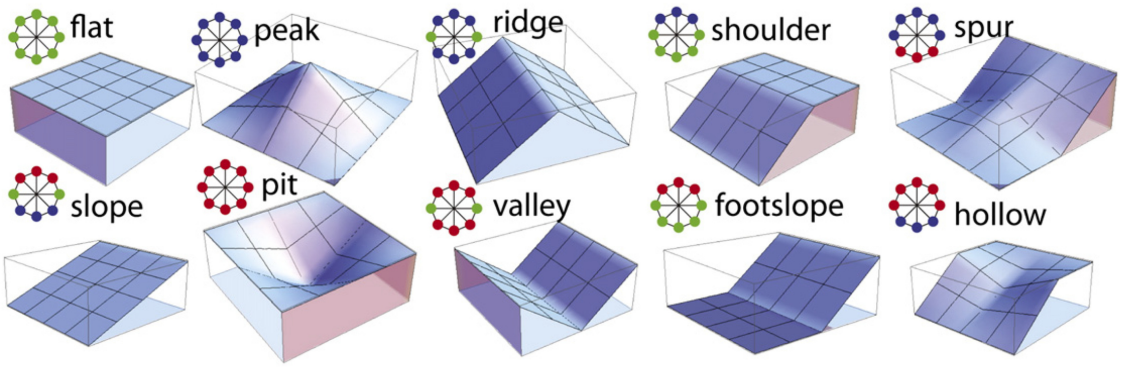

Fig. 2. The 10 most common landform patterns can be uniquely classified by geomorphons from a DEM. Blue disc identifies lower, red higher, and green the same altitude (image from [Jasiewicz and Stepinski 2013]).

remains an open challenge as existing methods, even deep-learning-oriented [Guérin et al. 2017] approaches, rely on user sketching and authoring [Argudo et al. 2019].

Despite the recent advances in terrain modeling, perceptual validation of the geometry of the generated structures remains an outstanding problem and has been addressed only partially. This work is based on the PhD thesis [Rajasekaran 2019].

3 GEOMORPHONS

The fundamental theory behind our method is the recently introduced concept of *geomorphons* [Jasiewicz and Stepinski 2013] that provides an exhaustive classification of terrain features from DEMs. Geomorphons decompose a DEM into local ternary patterns [Liao 2010] based on local curvature that provides an oriented eight-directional feature vector for each location of the DEM, and one value for the Moore neighborhood (see the circles in Figure 2). This gives rise to te10 geomorphons: flat, peak, ridge, shoulder, spur, slope, depression (or pit), valley, footslope, and hollow, as shown in Figure 2 from Jasiewicz and Stepinski [2013]. Geomorphons depend on the resolution of the DEM, and, in our experiments, one pixel of the DEM corresponds to approximately 200m. Therefore, each geomorphon describes an area of about 800×800 m^2. Please note that we follow the definition from Jasiewicz and Stepinski [2013]. Commercial tools such as Grass GIS use a wider variety of features.

We utilize geomorphons to gain insight into the importance of landform features and how they affect the perceived realism of terrains. Later on, we show how they are present or missing in different terrains. The order of the geomorphons in the color coding in Figure 3 is arbitrary, and to compare the wide variety of terrains used in this article, we sort the geomorphons according to their presence in the most realistically perceived terrain category from our user study that is glacial patterns of real terrains (Section 6.1). We use the ascending order of geomorphons as *depression (or pit)* (the least frequently present), *peak, flat, valley, ridge, hollow, spur, shoulder, slope, and footslope* (the most present).

We used an open implementation of geomorphons in the GRASS GIS tool [Neteler and Mitasová 2013] that generates a color-coded image corresponding to the input DEM as shown in the example in Figure 3. The algorithm's output is the normalized coverage of each geomorphon in the input DEM (the values of geomorphons for all datasets from this article are in the supplemental material).

4 METHOD OVERVIEW

The critical problem we intend to address is the perceived realism of terrains. We focus only on terrain geometry and do not consider any additional features such as snow, vegetation, or water bodies. We utilize the concept of

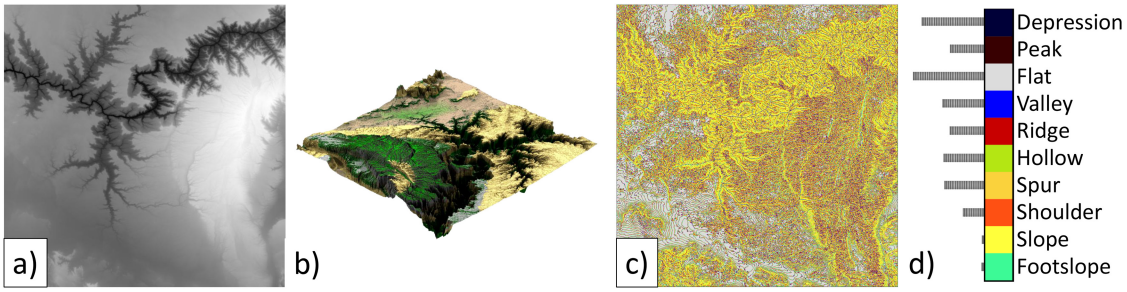

Fig. 3. (a) The input DEM, (b) its rendering, (c) the geomorphons, and (d) the explanation of the color-coding.

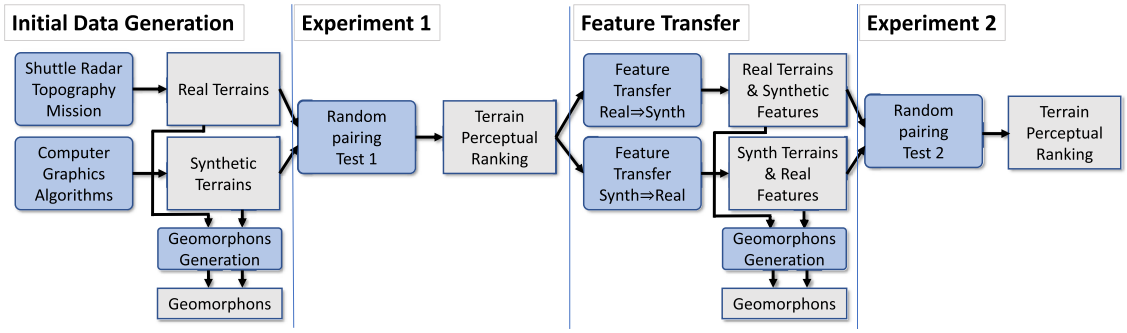

Fig. 4. Overview (rounded boxes—processes, squared boxes—data): The *initial data for Experiment 1* were acquired from two sources: real and synthetic DEMs. They were rendered, and we also generated geomorphons for each DEM that quantitatively describe their landform features. During *Experiment 1* we acquired the perceptual ranking of each image. The *feature transfer* transferred features from highly ranked images (Real⇒Synth) and vice versa (Synth⇒Real), resulting in two new datasets. *Experiment 2* perceptually evaluated the initial data and the newly generated ones, confirming that the transferred features have importance on the perceived realism.

geomorphons, which are the features extracted from DEMs that quantitatively measure the presence of various shapes in terrain (Section 3). We performed two large-scale user studies (Section 6). The first quantifies the perception of real and synthetic datasets, and the second quantifies the effect of the transferred features (see overview in Figure 4).

During the *initial data generation*, we acquired data of real terrains from the Shuttle Radar Topography Mission, and we carefully selected several classes featuring common geological patterns (see Table 1): Aeolian, Coastal, Fluvial, Glacial, and Slope. We proceeded to generate synthetic datasets using terrain generation algorithms used in computer graphics: coastal, thermal, and fluvial erosion; fractional Brownian motion; noise; and ridged noise. Geomorphon values were generated for all the DEMs.

Experiment 1 (E1) was a **two-alternative forced-choice design (2AFC)** by using Mechanical Turk. We have shown pairs of images, and we asked the viewers the question: *"Which terrain looks more realistic (left or right)?"* Each image received multiple rankings, and the votes determined its positioning in the overall test. The experiment provided an initial terrain ranking for each image and category within each group (real or synthetic). The results were used to construct our metric (PTRM) (Section 8) that relates the presence of geomorphons to the perceived realism. To evaluate a new terrain, geomorphons need to be calculated, normalized, and input into the PTRM.

Table 1. Terrain Type (Real/Synthetic/Transferred Features), Categories, Abbreviations, and the Number of Terrain Samples in Each Category

Type	Category	Abbreviation	Number of Sample Terrains
Real (R)	Aeolian	RA	55
	Coastal	RC	19
	Fluvial	RF	64
	Glacial	RG	07
	Slope	RS	05
Synthetic (S)	Coastal	SC	25
	fBm	SM	25
	Fluvial	SF	25
	Noise	SP	25
	Ridged-noise	SR	25
	Thermal	ST	25
Transferred Features (2)	Synth features to real terrains	S2R	25
	Real features to synth terrains	R2S	25

Feature Transfer: We used the CycleGAN [Zhu et al. 2017] to transfer features from the terrains that were ranked high to those ranked low and vice versa (Section 6.2). The motivation for this step is the assumption that certain features have an essential effect on the visual perception of terrains. This step generated two new datasets that we call S2R (synthetic to real) and R2S (real to synthetic). The terminology S2R reads "synthetic to real" (synthetic $\Rightarrow$ real) and indicates that procedural features were transferred to the real terrains from the synthetic ones. Similarly, R2S is the process of transferring features from a real terrain to a synthetic one. Geomorphons were also generated for the new datasets after each transfer.

Experiment 2 (E2) was also 2AFC, and it included the data from E1 and the newly added generated sets S2R and R2S (Section 6.3). The assumption was that the features from the highly ranked terrains would be transferred to the low-ranked terrains, and the resultant terrains that emerge because of the feature transfer would improve their ranking in the perceived realism order. A similar expectation was held for the low-ranked terrains, assuming that the features transferred to high-ranked terrain sets would demote their rank in the perceived realism order. Moreover, we also generated each terrain's geomorphons, and we kept track of which features were transferred (Section 8).

5 TERRAIN DATA

While the procedural generation of terrains is simple, and we could have generated an arbitrary number of DEMs, it is difficult to find good samples of the above-mentioned real patterns. Table 1 shows how many terrain models we had for each category and also establishes nomenclature for each set. Each real image starts with the letter R and synthetic image with S. The second letter indicates a subcategory. We refer to all images from real datasets as R and all synthetic as S. The size of each dataset was the same: $|S| = |R| = 150$.

5.1 Real Terrains

The DEMs for real terrains were retrieved from the **Shuttle Radar Topography Mission (SRTM)** dataset [Farr and Kobrick 2000]. We used three arc-second capture resolution tiles from the dataset that was the highest resolution available for the entire globe. The three arc-second resolution corresponds to $1^\circ \times 1^\circ$ Longitude×Latitude or 100×100 km resolution approximately depending on the DEM's location on Earth. All the DEMs were in the resolution of 512×512, which gives sizes of the land features around 200 meters per pixel.

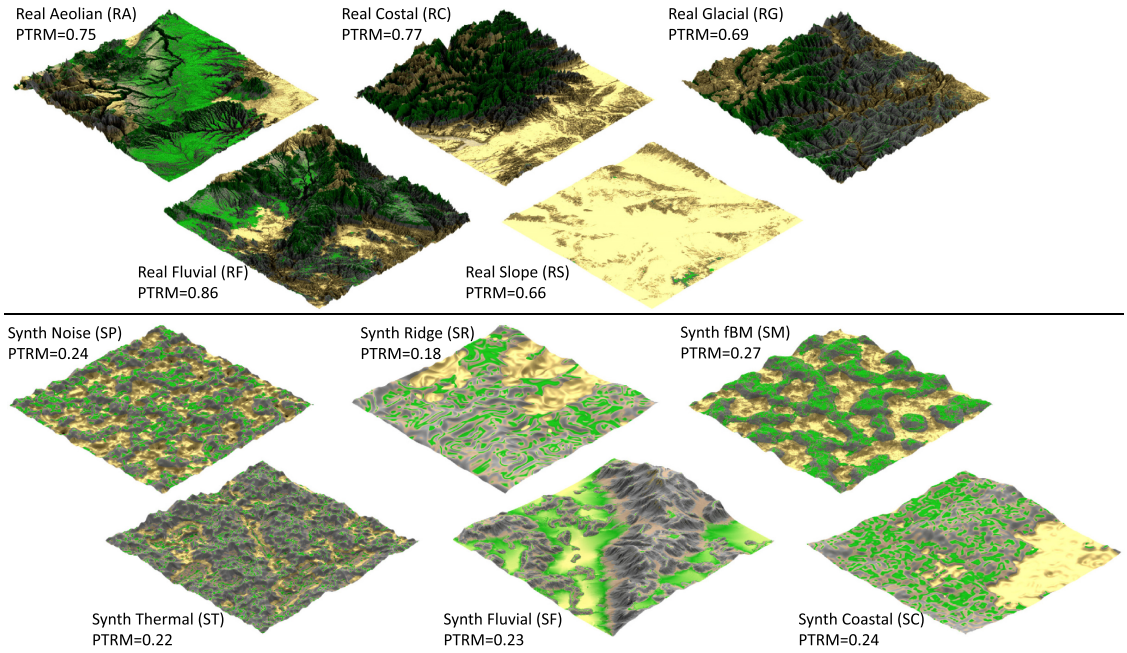

Fig. 5. (Top) Examples of real terrains rendering from our experiment and their PTRM: (RA) aeolian patterns from Moab Arches National Park, Utah; (RC) coastal patterns from Gobi Desert, Mongolia; (RG) glacial erosion patterns from Himachal Pradesh, Western Himalaya, India; (RF) fluvial pattern from Chichiltepec, Mexico, Guerrero; and (RS) slope pattern from Death Valley, California. (Bottom) Examples of synthetic terrains: (SP) noise based, (SR) ridged noise, (SM) fractional Brownian motion surface, (ST) thermal erosion, (SF) fluvial erosion, and (SC) coastal erosion (see supplementary materials for high-resolution images).

After a discussion with a geologist, we used terrains that include patterns that commonly result from aeolian, glacial, coastal, fluvial, and slope processes [Huggett 2016] along with the retrievability of suggested patterns from the SRTM dataset [Farr and Kobrick 2000]. It is important to note that the geoforming processes are not well understood. Most terrains are affected by several processes either at the same time period or in an indeterminable unknown sequence. Instead of discussing processes, we consider terrains that include the specified geomorphological patterns and structures. The two top rows of Figure 5 show examples of several renderings of real terrains, and the supplementary materials include all data.

5.2 Synthetic Terrains

We used terrains generated by noise [Perlin 1985], ridged noise [Galin et al. 2019], **fractional Brownian motion (fBm)** surfaces [Fournier et al. 1982], thermal erosion [Musgrave et al. 1989], fluvial erosion (we used the implementation of Šťava et al. [2008], but Krištof et al. [2009], Anh et al. [2007], and Neidhold et al. [2005] could be also used), and coastal erosion approximated by hydraulic erosion applied only to coastal areas. Eroded terrains were generated from noise-based terrains (Figure 5 two bottom rows). Most procedural methods do not simulate particular geomorphological processes and thus do not include explicit size. For example, fBm can be considered very small and very large because of its fractal (self-similar) nature. Also, synthetic terrains often are not geologically accurate. For example, they do not include outflow, as observed and enforced by a recent work [Scott and Dodgson 2021]. However, we target perceived realism, and geological accuracy is not common knowledge. Please note that the data (the DEMs and the rendered terrains) are available as supplementary material for this article.

5.3 Rendering

All terrains were rendered by using the same settings to minimize bias. We experimented with various settings and lighting, and the following provided the most details and contrast.

The camera position was set to display the terrain from about a 45^o angle, which is a typical viewing distance from the top of a mountain or a low-flying aircraft. This location shows enough details instead of a top view and does not cause common self-occlusions in side views. The camera was positioned above one of the corners, and we assumed viewers were familiar with this viewing angle.

We used sky sphere for illumination with a gradient from 50% of gray near the horizon to full white in zenith. The rendering was performed using global illumination with no additional lights, using 500 reflections and 9× super-sampling for anti-aliasing. We initially experimented with grayscale rendering that did not provide enough details and variation to perceive details. Thus, each terrain was textured by the same color map that changed from low-level and flat areas with yellow color (sand), medium levels that are flat and green (grass) to high and steep slopes gray (stone). We intentionally used non-photorealistic rendering [Gooch et al. 1998] to avoid any bias introduced by the simulation of vegetation and realistic rock, sand, or grass rendering. Moreover, non-photorealistic rendering enhances the shape and structure of the bare elevation of the terrain, which is the focus of this study.

The image resolution used for the perceptual experiment was given by the size of the screen used in Mechanical Turk (1,200 × 720 pixels). The terrain DEM resolution used was 512 × 512. We determined this resolution by scaling down a terrain from 1,201 × 1,201 that was the maximum available resolution for LIDAR scans by 10% down to 128 × 128 and comparing the Peak Signal-to-Noise Ratio of the heightmaps and the rendered images. The error between the maximum resolution and 512 × 512 was only 19.3%. It provided a good compromise in training deep neural networks, rendering, and viewing image pairs without zooming in and out during the ranking process.

6 PERCEPTUAL EXPERIMENTS AND FEATURE TRANSFER

The perceptual study was run on Amazon Mechanical Turk by showing a pair of rendered terrain images without giving any other information about the terrain and asking the subjects: *"Which terrain looks more realistic (left or right)?"* Each image pair was shown only once to each participant, but each image was repeated several times in different pairs. The survey was blinded because the participants only see an image pair with responses restricted to the "Left" or "Right" option. The experiment involved 70 participants with no particular constraints on their education or previous knowledge, and all participants were older than 18 years. However, only qualified "Mechanical Turk Masters," i.e., users who consistently demonstrate accuracy in answers, were allowed to answer the survey. We did not explicitly instruct the participants about what makes a terrain realistic, and we left it to their intuition. The overall expectation is that a large number of participants will arrive at a consensus.

We denote the compared categories by a dash for each image pair, so R-S indicates a pair of images where one is from the real and one from the synthetic sample. The actual position of each image (left or right) was randomized, making this relation symmetrical: R-S is the same as S-R.

6.1 Experiment 1: Real and Synthetic Terrains

We generated random image pairs by one real and one synthetic image, resulting in 150 pairs. This pairing happened five times for each image from R, resulting in $|R - S| = 750$ image pairs. We ensured that the pairing did not miss any image, each image was repeated five times, and pairing always occurred with a different image. The order of the images within each pair was randomized so that the synthetic image could be on the left-hand or right-hand side of the pair with the same probability.

Each image pair was shown to five different participants, resulting in a total of 3,750 image pair observations by 70 subjects with a varying degree of participation (determined based on the unique count of anonymized "workerID" provided by Amazon Mechanical Turk).

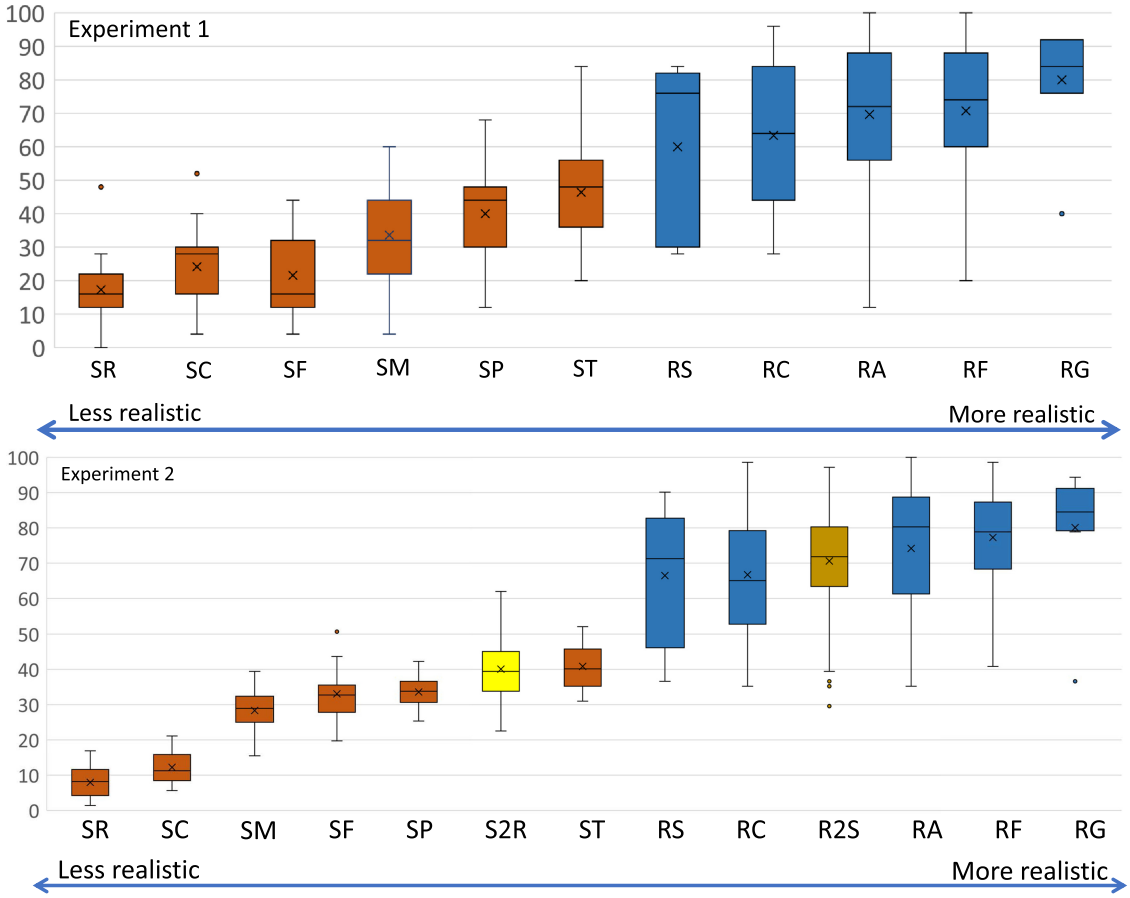

Fig. 6. Perceptual ranking of terrains from E1 (top) and E2 (bottom). The abbreviations are from Table 1, and the terrains are sorted by the average perceived realism from worse (left) to best (right). While the order of the rankings in E2 is very similar to E1, note that the S2R, i.e., synthetic terrains improved with features from real terrains, ranked high. At the same time, real terrain with features transferred from procedural R2S ranked lower. The figure has been plotted based on their average scores. The ×, •, and the − signs represent the mean, outlier points, and median markers.

Each time an image was selected as more realistic, it received a point, and the total number of points determined the overall ranking of each image that was normalized. We also calculated the normalized ranking of each category of real (RA, RC, RG, RF, and RS) and synthetic (SP, SR, SM, ST, SF, and SC) terrains.

Results: Experiment 1 assigned each image a number of how many times it was selected as more realistic in a pairwise choice randomized test. We normalized the counts so that the most realistically perceived image had a score of 1.0. We then calculated the average, standard deviation, mean, and range for each category of R and S from Table 1. The sorted results by the average value are shown in Figure 6 (top). The ranking of terrains from least realistic to the best was SR-SC-SF-SM-SP-ST-RS-RC-RA-RF-RG. All synthetic terrains were perceived as visually less realistic than the real ones. The most realistic synthetic terrains were generated by thermal erosion (ST) (see Table 4). We have also calculated the average and standard deviation of values of the ranking of all images in the sets S and R. An unpaired T-test evaluation suggested that the difference is statistically significant with the two-tailed $p < 0.01$, $DF = 283$, $t = 17.91$, and $\alpha = 0.01$.

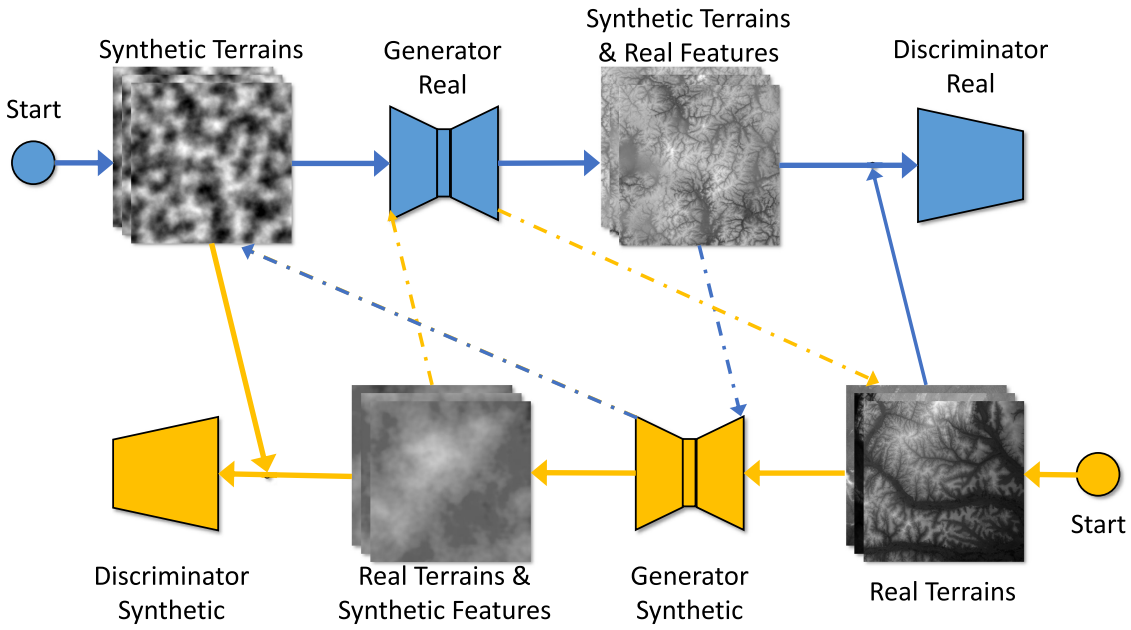

Fig. 7. Feature transfer: The blue arrows indicate the working flow of R2S; the orange arrows indicate the working flow of S2R. The dotted-and-dashed arrows indicate the cycle consistency process.

The perceptual experiment showcases that synthetic terrains in our dataset are perceived as visually significantly less realistic than the real ones.

6.2 Feature Transfer

E1 provided a ranking of each category of real and synthetic terrains. The distribution of geomorphons confirmed (Section 6.4) that they are related to the perceived realism. High-ranking real terrains contained features such as the valley topology in the terrains with fluvial erosion that were almost absent in low-ranking ones.

An important property of DEM is that they can be viewed as pixel images, where each pixel encodes the height of the terrain at the given location. Therefore, we can use deep learning that operates on images and apply it directly to terrain geometry stored as grayscale images such as in the upper left or lower right corner of Figure 7. We assumed that a deep neural network could distinguish the features that make real terrains visually plausible. The features learned from the distinction between the terrain categories can be transferred onto synthetic terrains to make them more visually plausible. Similarly, we hypothesize that the transfer could diminish features if it occurs from synthetic to real terrains, which would further justify the importance of specific features for perceived realism.

We initially experimented with Neural Style Transfer, which did not perform well. The model lacks the capability of transferring consistent global topologies such as long ridges or valleys, as discussed by Gatys et al. [2015]. Because the explicit pairing between the real and synthetic terrains is difficult, we used the unpaired image-to-image translation CycleGan [Zhu et al. 2017] to transfer features from the real domain to the synthetic domain, and vice versa. The micro-structures in the images can be learned and translated. The CycleGan has been shown to translate features such as ice and snow from a winter image to a summer scene. We want to translate the geomorphon features from real terrains to the synthetic to improve the perception and vice versa.

We set up a pair of generators G_R and G_S with a pair of discriminators D_R and D_S (please see details about the CycleGan in Zhu et al. [2017]). Generators are used to create terrain maps—one for creating real terrains G_R

and one for creating the synthetic ones G_S. The created terrains are mixed with the ground-truth terrains. The corresponding discriminators D_R and D_S are used to identify if the terrain maps are from the ground truth or created by the generators. The discriminators learn to identify the created terrains out of the ground truth during the training. In contrast, the generators learn to create terrains similar enough to the ground truth with geomorphons to confuse the discriminators. With such an adversarial process, we can make synthetic terrains include real geomorphons and make real terrains include synthetic noises.

As shown in Figure 7, the blue flow shows how synthetic terrains obtain real features. A synthetic terrain map is an input into the generator G_R to create a new terrain map with more real features. On the one hand, this new terrain map, together with a real terrain map from the ground-truth pool, is passed into the discriminator D_R to let the discriminator identify which one is created. On the other hand, the new terrain map is passed into the other generator G_S to convert (cycle) back to the original synthetic terrain map. CycleGan [Zhu et al. 2017] indicated that if we apply the real generator on a synthetic terrain map to get a new map and then apply the synthetic generator on the new map, we should be able to get the original synthetic terrain map similar to the input. This is defined as the cycle consistency loss, which makes $G_S\left(G_R(s)\right) \approx s$ and $G_R\left(G_S(r)\right) \approx r$. The cycle consistency ensures the high-quality feature transfer. The symmetric yellow flow in Figure 7 shows the other half with real terrain maps.

In other words, the generator G_R translates terrains from the synthetic domain S to the real domain R with real features. The discriminator D_R discriminates between terrains r and $G_R(s)$, where $r \in R$ and $s \in S$. Moreover, G_S translates terrains within the real domain R to the synthetic domain S with synthetic features. Similarly, D_S discriminates between terrains s and $G_S(r)$.

We adopt a nine res-block generator and a 70×70 PatchGAN discriminator [Isola et al. 2016]. The transfer-generated checkerboard patterns caused by fractionally strided convolution and the artifacts decrease if the training epochs increase. We also applied resize-conv with Nearest Neighbor and Bilinear as suggested in Odena et al. [2016].

Our training set contains 9,800 real terrain height maps selected from the SRTM DEMs, excluding the terrains that have been used in E1 and E2. We generated additional synthetic height maps for use in training based on the aforementioned synthetic categorization and the same size as the real terrain training data, which is $9,800$ (see the data collection in Sections 5.1 and 5.2).

We trained the model with 20 epochs and then generated 150 images of real terrains with synthetic features denoted by S2R. The term S2T denotes the *transfer*, meaning "synthetic features were transferred to real terrains." We also generated another 150 images of synthetic terrains with real features denoted by R2S. Figures 1 and 8 show example results of the feature transfer in both directions (from real to synthetic and from synthetic to real). Please note that we provide the training datasets for further experiments and the CycleGan code is freely available.

6.3 Experiment 2: Real, Synthetic, and Terrain Models with Transferred Features

The objective of the second experiment (E2) was to evaluate how the terrains with transferred features score perceptually against real and synthetic terrains. We have reused the 750 R-S image pairs from E1 (Section 6.1) and added another 750 image pairs for each missing combination. Table 2 shows the naming of the image pairs. The first column shows the reused pairs from E1 (R-S). The newly added pairs compare newly created transferred features from synthetic to real R2S combined with all options R2S-R, R2S-S, and S2R-R2S. Also, we added combinations for feature transfer from real to synthetic S2R, i.e., S2R-R and S2R-S. R2S-S2R is already included because it is symmetrical with S2R-R2S.

As in E1, each shuffling was generated five times, resulting in 750 image pairs for each item of Table 2, resulting in a total of 4,500 image pairs. We repeated each test for five independent viewers, resulting in a total of 22,500 views by 128 subjects. All participants were older than 18 years, and we again used only qualified Mechanical

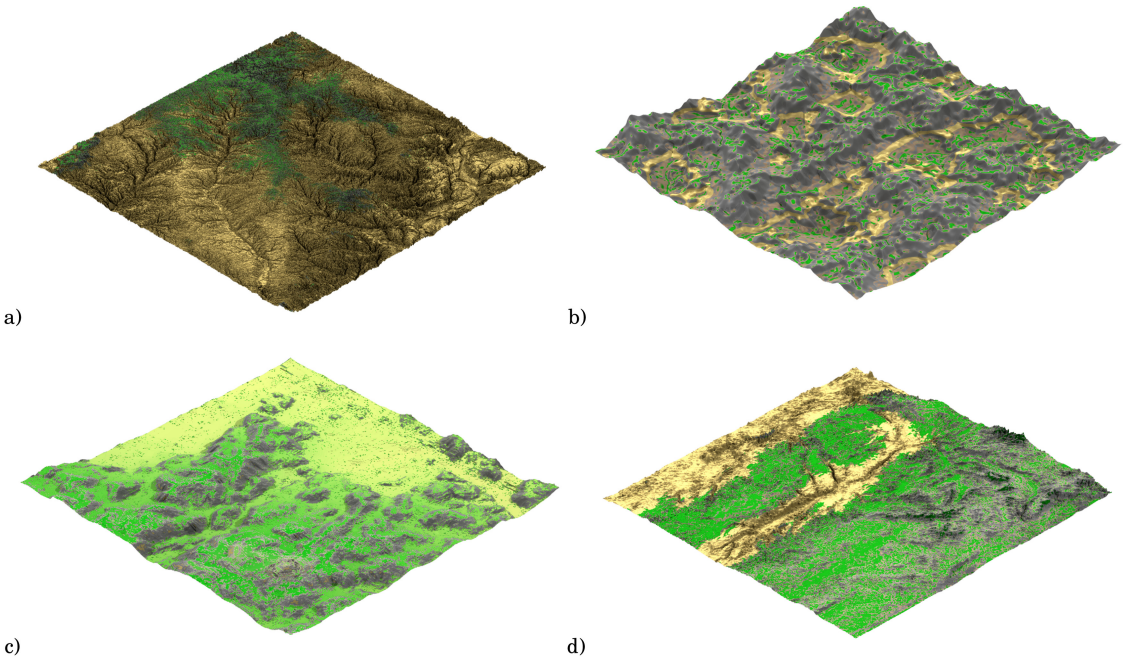

Fig. 8. Example of feature transfer: (a) real terrain with strong fluvial patterns from Colombian Amazonian forest area (S01 W072) (PTRM = 0.67) and (b) synthetic terrain generated by thermal erosion (PTRM = 0.46). (c) Synthetic features transferred to real terrain worsen its perceived visual realism (PTRM = 0.49) and (d) real features transferred to synthetic terrain improve it (PTRM = 0.63).

Table 2. Image Pairing for Experiment 2
(R-S Pairs Are Reused from Experiment 1)

	S	R2S	S2R
R	R-S	R2S-R	S2R-R
S	•	R2S-S	S2R-S
R2S	•	•	S2R-R2S

Turk Masters. Note that because the R-S set from the first experiment was also included in the second one, we have validated the first experiment. The ranking of the R-S results was consistent between E1 and E2, suggesting the data saturation point has been attained.

Results: Experiment 2 repeated E1 with the addition of pairs of images with transferred features (Section 6.2). We assumed that the features transferred from real terrains to synthetic would improve their ranking and that the transfer of features from synthetic to real terrains would do the opposite. The normalized rank of each image and calculated statistics for each category are in Figure 6 and Table 4.

The perceived order of terrain categories is the same as in E1, which confirms the validity of both tests. The synthetic terrains improved with features from real terrains. In the categories with transferred features, R2S came in fourth place, which is understandably higher than all synthetic terrains but also higher than specific real terrain categories (RS and RC). This confirms our hypothesis that feature transfer affects terrain perception. Similarly, the real terrain with transferred procedural features S2R ranked significantly worse than real terrains

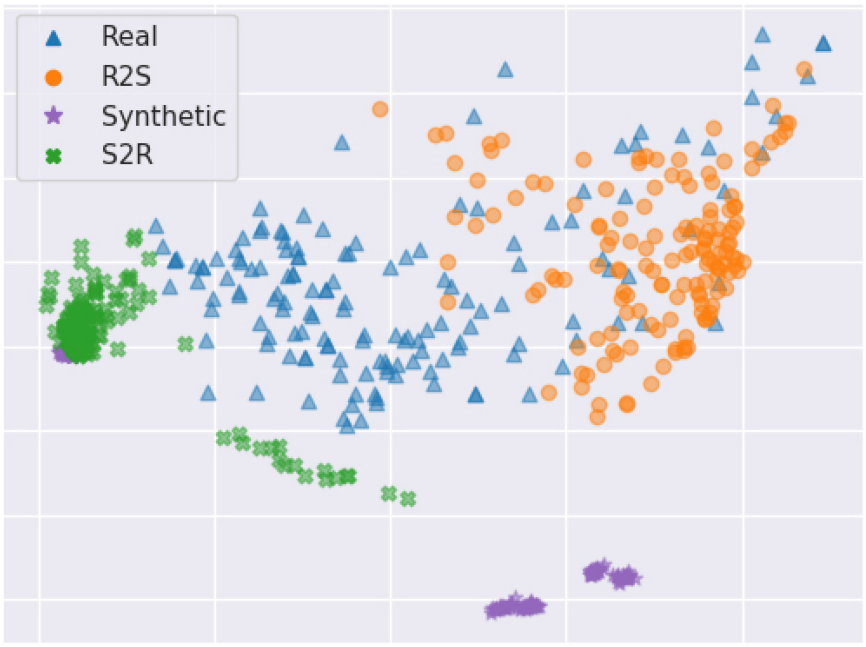

Fig. 9. Projection of geomorphons from all terrains to 2D. Synthetic terrains appear clustered, while real terrains are more scattered. Transfer of real features scatters the terrains and transfer of procedural features clusters the resulting terrains.

and even worse than thermal erosion simulation (ST) in eighth place. This confirmed our hypothesis that features of synthetic terrains do not contribute significantly to terrain realism.

6.4 Geomorphon Transfer

Geomorphons (introduced in Section 3) characterize local terrain features (valleys, ridges, peaks, etc.). A geomorphon is a 10D feature vector that describes a terrain. The spatial distribution of geomorphons brings further insight into the features and the corresponding datasets.

Figure 9 shows the points corresponding to all our datasets (R, S, R2S, and S2R) projected from 10D space to 2D by using the t-Distributed Stochastic Neighbor Embedding algorithm [Maaten and Hinton 2008] that preserves distances among points across the dimensions. Synthetic terrains appear clustered, while features of real terrains are scattered over a wide area. This is further confirmed by the variance of the features, as shown in graphs in Figure 11. When the real features are transferred to synthetic terrains, they tend to scatter the images apart, and when synthetic features are transferred to real terrains, they tend to get close to each other. This seems to indicate that high variability in geomorphological features is beneficial for perceived realism.

Moreover, we visualize domain-wise comparisons among R, R2S, S, and S2R on the distributions of the element-wise geomorphon feature of terrains in Figure 10. The geomorphon features of real terrains (blue curve) tend to distribute generally with a wide span. However, the synthetic features (green curve) show significant differences from the real with multi-modal and low-variability distributions on depression, peak, flat, valley, and ridge (Figure 10, top row). We believe the high-peak distributions of synthetic terrains lead to less attractive perceptions than the real. The process of R2S transfer (orange curve) smooths and normalizes the multi-modal high-peak distributions in the synthetic terrains and improves the perception. It also seems that the lack of geomorphon diversity or variability of individual geomorphons in distribution decreases the perceived terrain realism (Figure 10, radar graphs).

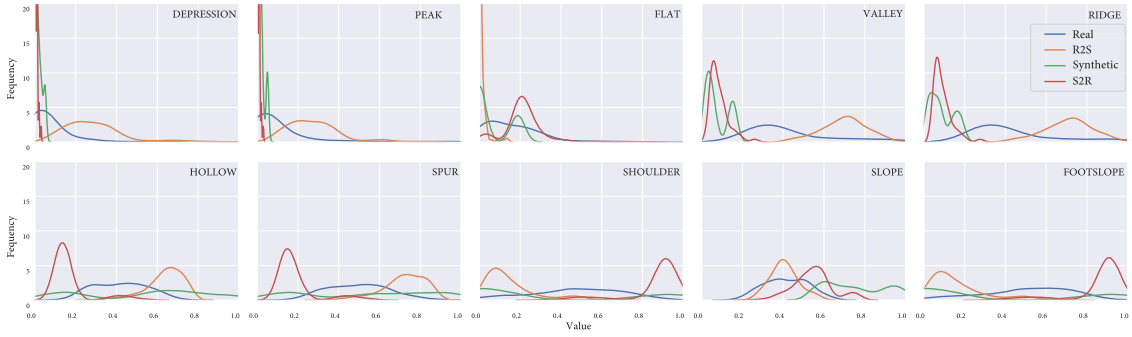

Fig. 10. The geomorphon feature comparisons among Real, R2S, Synthetic, and S2R (x-axis is the normalized value, y-axis the count).

Table 3. Statistical Significance of Terrain Sets from Our Experiments: E1 and E2

	R	S	R2S	S2R
R	•	✓	×	✓
R2S	•	•	✓	✓
S2R	•	•	•	✓
S	•	•	•	•

The ✓ implies that the terrain set in the vertical column is statistically significant over the terrain set in the horizontal row, × marks that the difference is not statistically significant, and • shows that the test is not available or compared already.

7 STATISTICAL TESTS

We performed statistical tests on our normalized perceptual scores to determine any differences in perception of our terrain data groups: R, S, R2S, and S2R. We state the null hypothesis, H_0, for our six statistical tests in E2 as follows: *"There are no significant differences in the visual perception scores between our terrain data groups."*

We used T-test to compare the means and variances of the perception scores, and the results are summarized in Table 4. For testing our candidates in E2, we used the significance level of $\alpha = 0.01$, and get the statistics for, R versus R2S ($p = 0.02$, $DF = 149, t = 2.26$), R versus S2R ($p < 0.01$, $DF = 149, t = 22.10$), R versus S ($p < 0.01$, $DF = 149, t = 22.59$), R2S versus S2R ($p < 0.01$, $DF = 149, t = -23.52$), R2S versus S ($p < 0.01$, $DF = 149, t = 29.12$), and S2R versus S ($p < 0.01$, $DF = 149, t = 10.79$). Table 3 summarizes the perception scores. The scores are statistically different between the terrain groups. The observers perceived the realism of the terrains at different scales except the R vs. R2S. This implies that there are features in real terrains that increase the perceived realism and we can reject our null hypothesis stating that *there is a significant difference in perception of Real Terrains (R), Synthetic Terrains (S), Synthetic Terrains with Real features (R2S), and Real Terrains with Synthetic features (S2R).*

We performed an ANOVA (E1: $calculated\ F = 320.91, critical\ F = 3.87, p < 0.01, df = 298$; E2: $calculated\ F = 465.78, critical\ F = 2.61, p < 0.01, df = 596$) to determine if there are any significant differences in the variances of the scores. After establishing differences in the groups, we proceeded with the T-tests to determine which groups the significant differences lie in. Additionally, a post hoc test, Tukey's **Honestly Significant Difference (HSD)** test, indicated no statistically significant difference in the perception scores between the terrain groups, R, and R2S with a $p = 0.0511$ and standard error of 1.0938. At the same time, there is a statistically significant difference between the rest of the terrain groups with $p < 0.001$ and a standard error of 1.0938 with $\alpha = 0.01$, consistent with T-test results.

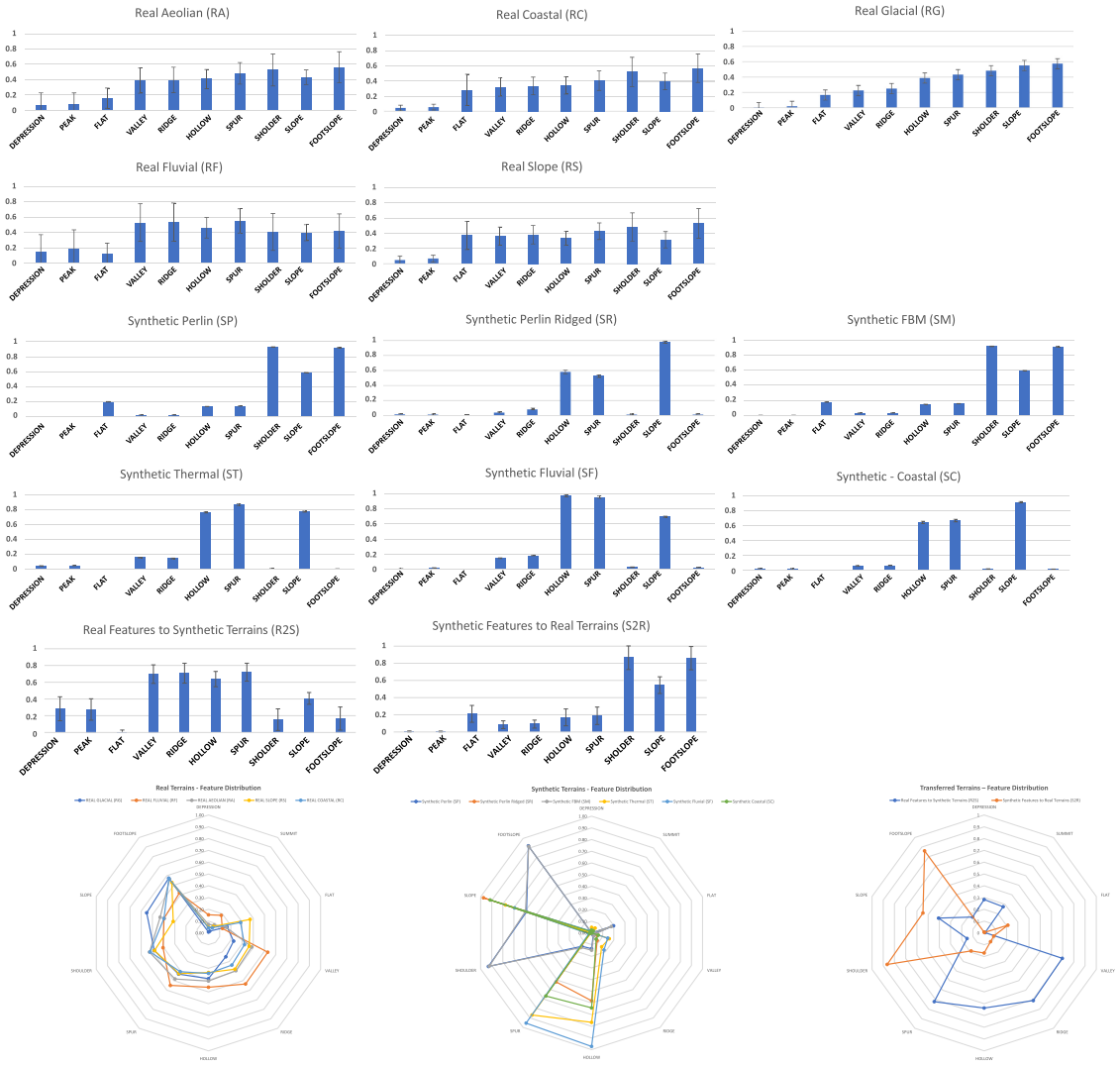

Fig. 11. Distribution of the detected geomorphons in real and synthetic terrains from our dataset.

Additionally, we utilized a scaling mechanism to efficiently capture both ranking and magnitude of differences between these conditions using a pairwise comparison scaling method based on the Thurstone Case V model to measure respondents' attitudes based on probabilities and **Just Objectionable Differences (JODs)** [Perez-Ortiz and Mantiuk 2017]. The standardized scores based on the Thurstonian scaling mechanism are converted to distances for the underlying quality scores to determine JODs. A visualization of these scaling results using Real Terrain (R) as our anchor point (reference condition) is shown in Figure 12. The results are congruous with our previous conclusions.

8 PERCEIVED TERRAIN REALISM METRIC (PTRM)

The Pearson correlation coefficients (Table 5) show a strong correlation between each of the geomorphons (our predictor variables) at various levels (Positive and Negative Correlation) on the Perception Score. The order of

Table 4. Average (AVG), Median (MED), Mode (MODE), RANGE (RNG), Standard Deviation (STDEV), Standard Error (SE), and 95% Confidence Interval (95% CI) of the Normalized Scores for the Terrain Sets: E1 and E2

		E1							E2						
T	**Ab.**	**AVG**	**MED**	**MODE**	**RNG**	**STDEV**	**SE**	**95% C.I.**	**AVG**	**MED**	**MODE**	**RNG**	**STDEV**	**SE**	**95% C.I.**
R	RG	80	84	92	52	19	7	14	80	851	N/A	58	19	7	13
	RF	71	74	88	80	20	3	5	77	79	86	58	13	2	3
	RA	70	72	92	88	22	3	6	74	80	89	65	19	3	5
	RC	63	64	96	68	21	5	9	67	65	63	63	17	4	8
	RS	60	76	N/A	56	28	12	24	67	71	N/A	53	20	8	16
S	ST	46	48	48	64	17	3	7	41	40	32	21	7	1	3
	SP	40	44	48	56	13	3	5	34	34	34	17	5	1	2
	SM	34	32	36	56	14	3	6	28	29	32	24	6	1	2
	SF	22	16	16	40	12	2	5	33	33	35	31	8	2	3
	SC	24	28	28	48	11	2	4	12	11	10	16	5	1	2
	SR	17	16	20	48	10	2	4	8	8	13	16	4	1	2
2	R2S	N/A	N/A	N/A	N/A	N/A	N/A	N/A	71	72	70	68	13	2	2
	S2R	N/A	N/A	N/A	N/A	N/A	N/A	N/A	40	39	33	39	9	1	1

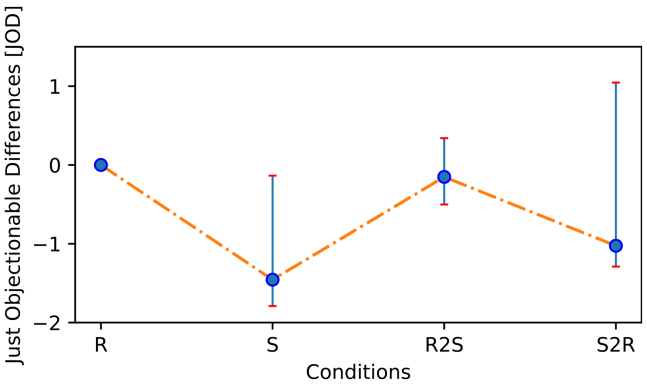

Fig. 12. The graph visualizes the scaling results for experiments E1 and E2 along with their 95% confidence intervals for our dataset. The first condition (R) is our reference condition; therefore, it is always set at 0 and hence, there are no intervals. A difference of one JOD unit indicates that 75% of the participants chose a condition over the other.

the influence on the perception score is given by Valley (0.66), Ridge (0.64), Peak (0.44), Depression (0.42), Spur (0.33), Hollow (0.22), Flat (−0.10), Foot (−0.15), Shoulder (−0.17), and Slope (−0.65). This testifies that the presence of a variety of geomorphons is a good indicator of perceived realism. We devised a PTRM that takes input into a set of normalized geomorphons for a DEM terrain with the spatial resolution of 200m per pixel and returns the estimated perceived realism.

We performed a **multiple linear regression (MLR)** model on our dataset with the hypothesis, H_0: *"There is no linear relationship between the 10 geomorphon landform categories and the perception scores for our terrain data groups."* The regression gave us the following statistics: $DFn = 10, DFd = 588, F = 153.5276, p < 0.01$, with $\alpha = 0.01$. Therefore, we rejected the null hypothesis concluding that the coefficients are statistically significant with $p < 0.01$.

The coefficients from the linear regression model between the 10 geomorphon categories are then used to weight the effect of each geomorphon giving the PTRM ($0.0 \leq PTRM \leq 1.0$, 0 = poor, 1 = realistic):

$$\begin{aligned} PTRM \quad = \quad & (-38.02 + 3.55G_{depression} + 1.75G_{peak} \\ & + 25.12G_{flat} + 9.61G_{valley} + 7.59G_{ridge} \\ & + 6.71G_{hollow} + 9.02G_{spur} + 7.31G_{shoulder} \\ & + 28.95G_{slope} + 7.63G_{footslope})/69.96. \end{aligned} \tag{1}$$

Table 5. Correlations among 10 Geomorphons (Depression, Peak, Flat, Valley, Ridge, Hollow, Spur, Shoulder, Slope, and Footslope) and the Perception Score

CORR	DEPP.	SUMM.	FLAT	VALL.	RIDG.	HOLL.	SPUR.	SHOU.	SLOP.	FOOT	SCORE
DEPR.	1.00	•	•	•	•	•	•	•	•	•	•
SUMM.	0.99	1.00	•	•	•	•	•	•	•	•	•
FLAT	−0.41	−0.41	1.00	•	•	•	•	•	•	•	•
VALL.	0.85	0.87	−0.42	1.00	•	•	•	•	•	•	•
RIDG.	0.86	0.87	−0.42	1.00	1.00	•	•	•	•	•	•
HOLL.	0.41	0.42	−0.77	0.49	0.49	1.00	•	•	•	•	•
SPUR	0.45	0.46	−0.76	0.56	0.57	0.99	1.00	•	•	•	•
SHOU.	−0.50	−0.51	0.71	−0.53	−0.54	−0.94	−0.93	1.00	•	•	•
SLOP.	−0.51	−0.53	−0.32	−0.67	−0.66	0.18	0.08	−0.18	1.00	•	•
FOOT	−0.50	−0.50	0.72	−0.51	−0.52	−0.95	−0.93	1.00	−0.20	1.00	•
SCORE	0.42	0.44	−0.10	0.66	0.64	0.22	0.33	−0.17	−0.65	−0.15	1.00

Table 6. Comparison of Perception Scores Generated Based on Our Introduced Metric and Our Previously Normalized Score from the Study

Type	**Category**	**Measured Perception Score**	**PTRM**
Real (R)	RG	0.61	0.57
	RF	0.78	0.73
	RA	0.75	0.69
	RS	0.73	0.74
	RC	0.69	0.65
Synthetic (S)	ST	0.50	0.53
	SP	0.35	0.36
	SF	0.40	0.42
	SM	0.35	0.36
	SC	0.24	0.24
	SR	0.02	0.02
Transfers (T)	R2S	0.67	0.71
	S2R	0.38	0.41

Table 6 and Figure 13 show the comparison of PTRM with the calculated perception score averages for each category.

The resulting R-squared value for the PTRM is 0.72, signifying that the 72% of variation in the visual realism of terrains (i.e., the perception score) can be explained by the full model with all of our predictor variables, i.e., 10 normalized geomorphon distribution values with a standard error of 0.13. The landform factors are significant predictors of the visual terrain realism perception score.

PTRM Validation: We have collected a large dataset of real and synthetic DEMs (Section 5.1). We validated the PTRM by splitting the data five times randomly into 80:20%, recalculating the PTRM (Equation (2)) on the 80%, and validating on the remaining 20%. The average regression PTRM for the five datasets is

$$\begin{aligned} PTRM \quad = \quad & (-38.44 + 3.61G_{depression} + 1.77G_{peak} \\ & + 25.40G_{flat} + 9.71G_{valley} + 7.65G_{ridge} \\ & + 6.77G_{hollow} + 9.14G_{spur} + 7.40G_{shoulder} \\ & + 29.26G_{slope} + 7.69G_{footslope})/69.22, \end{aligned} \tag{2}$$

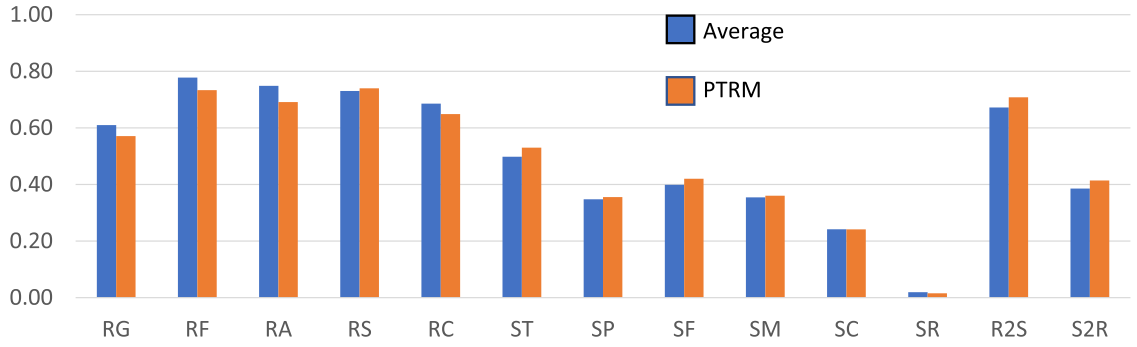

Fig. 13. Average value of measured perception score for each category vs. the PTRM.

which is very close to the PTRM model with the amount of explained variation (72%) and standard error (0.13). Both values remained consistent with the regression model with a 95% confidence interval.

We show various PTRM examples in this article, in particular in Figures 1, 5, and 8. Moreover, all PTRM values for all images as well as perceived scores are in the supplemental material.

9 CONCLUSION

This article presented a first step in evaluating the perceptual quality of procedural models of terrains. We have conducted two large-scale perceptual studies that allowed us to rank synthetic and real terrains. Our results show that the tested synthetic terrains were perceived worse than the tested real terrains with statistical significance. We have performed a quantitative study using geomorphons that indicate that features such as valleys, ridges, peaks, depressions, spurs, and hollows have perceptual importance. We used a deep neural network to transfer the geomorphological features, and the second perceptual study confirmed this observation. Eventually, we designed PTRM, a novel perceptual metric based on geomorphons that allows us to calculate a number for the perceived realism of a generated terrain.

Our study has several **limitations**. Geomorphons are localized to small terrain areas, and they do not reflect the distributions of the large features such as rivers, large valleys, and so forth. Two terrains with the same feature vector may be perceived as different because of the variety of distributions. Our study made several assumptions on terrain size. Changing the scale of the terrains may affect our results because features of different scales would be captured. Another limitation is the assumption about the terrain classification. While we motivated our classification into terrains with different geomorphological patterns, probably every terrain on Earth has been exposed to various phenomena. It is not entirely clear what caused the patterns. Additionally, we assumed the fixed position of the camera, consistent texturing, and illumination. While these aspects were carefully selected and made constant, it would be interesting to see the effect of variation on each of them over the results. Last but not least, learning-based feature transfer with deep neural networks provides limited control on the content to be or not to be transferred. With the metric we offered, the transferred results can be improved in perception with a better control schema of the generative network. We also did not study the spatial correlation between geomorphons and how those interactions may impact perception. There is a concrete correlation between variability in geomorphons and perceived realism in our tests. If combined with our proposed metric, the same correlation can also be used for synthesizing terrains. Additionally, a side effect that should also be noted is that certain real terrains can be perceived as less realistic due to their lack of variety in geomorphons.

There are many possible avenues for **future work**. Perceptual studies can answer longstanding questions about the visual quality of procedural models. Our work is based on the underlying concept of geomorphons that may be difficult to generalize to different domains. A global metric considering large geomorphological structures [Argudo et al. 2019] could also be combined with our perceptual study to create other such

perceptual metrics. We intentionally used non-experts to evaluate terrains. It would be interesting to use professional geologists to provide a perceptual evaluation. Our study attempts to answer the general question for most of the terrains. However, many natural structures may be perceived as unrealistic, yet they are real (Zhangjiajie National Forest Park in China). Our metrics will very likely fail on these cases, and it would be interesting to study these phenomena in future work. We did not have an implementation of example-based methods, such as Zhou et al. [2007]. It would be interesting to see how the terrains composed from various examples of real terrain are perceived. They include geologically correct features, but they may be combined in a way that does not make them realistic. Geomorphons should provide efficient encoding across various scales, as noted by Jasiewicz and Stepinski [2013]. However, the decreasing resolution will lose information that could also affect the perceived realism. The scaling and the effect of geomorphons would be worthy of future investigation.

ACKNOWLEDGMENT

The authors would like to thank Terragen and Vue (e-on Software) for providing student licenses of their software packages. We also want to thank the reviewers for their valuable comments.

REFERENCES

Nguyen Hoang Anh, Alexei Sourin, and Parimal Aswani. 2007. Physically based hydraulic erosion simulation on graphics processing unit. In *Proc. of the Graphite*. ACM, 257–264.

Oscar Argudo, Eric Galin, Adrien Peytavie, Axel Paris, James Gain, and Eric Guérin. 2019. Orometry-based terrain analysis and synthesis. *ACM Trans. on Graph.* 38, 6 (2019), 1–12.

Tunç Ozan Aydın, Martin Čadík, Karol Myszkowski, and Hans-Peter Seidel. 2010. Video quality assessment for computer graphics applications. *ACM Trans. on Graph.* 29, 6, Article 161 (2010), 1–12. DOI: https://doi.org/10.1145/1882261.1866187

Dirk Bartz, Douglas W. Cunningham, Jan Fischer, and Christian Wallraven. 2008. The role of perception for computer graphics. In *Eurographics (State of the Art Reports)*. 59–80.

Bedrich Benes and Rafael Forsbach. 2002. Visual simulation of hydraulic erosion. *J. of WSCG* 10, 1 (2002), 79–86.

Bedrich Benes, Václav Těšínský, Jan Hornyš, and Sanjiv K. Bhatia. 2006. Hydraulic erosion. *Comp. Anim. and Virtual Worlds* 17, 2 (2006), 99–108.

Scott D. Bergen, Christoph Ulbricht, James L. Fridley, and Mark A. Ganter. 1995. The validity of computer-generated graphic images of forest landscape. *J. of Environ. Psychology* 15, 2 (1995), 135–146. https://doi.org/10.1016/0272-4944(95)90021-7

Derek C. Bertilone, Robert S. Caprari, Steven Angeli, and Garry N. Newsam. 1997a. Spatial statistics of natural-terrain imagery. I. Non-Gaussian IR backgrounds and long-range correlations. *Appl. Opt.* 36, 35 (Dec. 1997), 9167–9176. https://doi.org/10.1364/AO.36.009167

Derek C. Bertilone, Robert S. Caprari, Philip B. Chapple, and Steven Angeli. 1997b. Spatial statistics of natural-terrain imagery. II. Oblique visible backgrounds and stochastic simulation. *Appl. Opt.* 36, 35 (Dec. 1997), 9177–9185. https://doi.org/10.1364/AO.36.009177

Micah Bojrab, Michel Abdul-Massih, and Bedrich Benes. 2013. Perceptual importance of lighting phenomena in rendering of animated water. *ACM Trans. on Appl. Percept.* 10, 1, Article 2 (March 2013), 18 pages.

Guillaume Cordonnier, Jean Braun, Marie-Paule Cani, Bedrich Benes, Eric Galin, Adrien Peytavie, and Eric Guérin. 2016. Large scale terrain generation from tectonic uplift and fluvial erosion. In *Comp. Gr. Forum*, Vol. 35. Wiley Online Library, 165–175.

Guillaume Cordonnier, Marie-Paule Cani, Bedrich Benes, Jean Braun, and Eric Galin. 2018. Sculpting mountains: Interactive terrain modeling based on subsurface geology. *IEEE TVCG* 24, 5 (2018), 1756–1769. https://doi.org/10.1109/TVCG.2017.2689022

Terry C. Daniel. 2001. Whither scenic beauty? Visual landscape quality assessment in the 21st century. *Landsc. and Urban Plan.* 54, 1 (2001), 267–281. Our Visual Landscape: analysis, modeling, visualization and protection.

Lucian Draguţ and Thomas Blaschke. 2006. Automated classification of landform elements using object-based image analysis. *Geomorphology* 81 (2006), 330–344. https://doi.org/10.1016/j.geomorph.2006.04.013

Tom G. Farr and Mike Kobrick. 2000. Shuttle radar topography mission produces a wealth of data. *EOS, Trans. Am. Geophys. Union* 81, 48 (2000), 583–585.

James A. Ferwerda. 2003. Three varieties of realism in computer graphics. In *Human Vision and Electronic Imaging VIII*, Vol. 5007. International Society for Optics and Photonics, 290–297.

Alain Fournier, Don Fussell, and Loren Carpenter. 1982. Computer rendering of stochastic models. *Comp. Graph.* 25, 6 (1982), 371–384.

Simone Frintrop, Thomas Werner, and German M. Garca. 2015. Traditional saliency reloaded: A good old model in new shape. In *IEEE Computer Vision and Pattern Recognition (CVPR)*. 82–90. https://doi.org/10.1109/CVPR.2015.7298603

James E. Gain, Patrick Marais, and Wolfgang Strasser. 2009. Terrain sketching. In *Proc. of the Symposium on Interactive 3D Graphics and Games*. ACM, 31–38.

Eric Galin, Eric Guérin, Adrien Peytavie, Guillaume Cordonnier, Marie-Paule Cani, Bedrich Benes, and James Gain. 2019. A review of digital terrain modeling. *Comp. Gr. Forum* 38, 2 (2019), 553–577.

Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. 2015. A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576* (2015).

Amy Gooch, Bruce Gooch, Peter Shirley, and Elaine Cohen. 1998. A non-photorealistic lighting model for automatic technical illustration. In *SIGGRAPH*, Vol. 98. 447–452.

Eric Guérin, Julie Digne, Eric Galin, and Adrien Peytavie. 2016. Sparse representation of terrains for procedural modeling. *Comp. Gr. Forum (Proc. of Eurographics)* 35, 2 (2016), 177–187.

Eric Guérin, Julie Digne, Eric Galin, Adrien Peytavie, Christian Wolf, Bedrich Benes, and Benoit Martinez. 2017. Interactive example-based terrain authoring with conditional generative adversarial networks. *ACM Trans. on Graph.* 36, 6 (2017), 1–13.

Jinjiang Guo, Vincent Vidal, Atilla Baskurt, and Guillaume Lavoué. 2015. Evaluating the local visibility of geometric artifacts. In *Proc. of the ACM SIGGRAPH Symp. on Applied Perception*. ACM, 91–98.

Jinjiang Guo, Vincent Vidal, Irene Cheng, Anup Basu, Atilla Baskurt, and Guillaume Lavoue. 2016. Subjective and objective visual quality assessment of textured 3D meshes. *ACM Trans. Appl. Percept.* 14, 2, Article 11 (2016), 20 pages.

Caroline M. Hagerhall, Terry Purcell, and Richard Taylor. 2004. Fractal dimension of landscape silhouette outlines as a predictor of landscape preference. *J. of Environ. Psychology* 24, 2 (2004), 247–255.

Robert Herzog, Martin Čadík, Tunç O. Aydın, Kwawng In Kim, Karol Myszkowski, and Hans-Peter Seidel. 2012. NoRM: No-reference image quality metric for realistic image synthesis. *Comp. Gr. Forum* 31, 2 (2012), 545–554. https://doi.org/10.1111/j.1467-8659.2012.03055.x

Richard Huggett. 2016. *Fundamentals of Geomorphology*. Routledge.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. 2016. Image-to-image translation with conditional adversarial networks. *arxiv* (2016).

Jaroslaw Jasiewicz and Tomasz F. Stepinski. 2013. Geomorphons - A pattern recognition approach to classification and mapping of landforms. *Geomorphology* 182 (2013), 147–156. https://doi.org/10.1016/j.geomorph.2012.11.005

Alex D. Kelley, Michael C. Malin, and Gregory M. Nielson. 1988. Terrain simulation using a model of stream erosion. *Comp. Graph.* 22, 4 (1988), 263–268.

Peter Krištof, Bedrich Benes, Jaroslav Křivánek, and Ondřej Šťava. 2009. Hydraulic erosion using smoothed particle hydrodynamics. *Comp. Gr. Forum* 28, 2 (2009), 219–228.

Eckart Lange. 2001. The limits of realism: Perceptions of virtual landscapes. *Landsc. and Urban Plan.* 54, 1 (2001), 163–182. https://doi.org/10.1016/S0169-2046(01)00134-7. Our Visual Landscape: analysis, modeling, visualization and protection.

Guillaume Lavoué, Mohamed-Chaker Larabi, and Libor Vása. 2016. On the efficiency of image metrics for evaluating the visual quality of 3D models. *IEEE TVCG* 22, 8 (2016), 1987–1999.

Wen-Hung Liao. 2010. Region description using extended local ternary patterns. In *2010 20th International Conference on Pattern Recognition*. IEEE, 1003–1006.

Jun Liu, Junyu Dong, Xiaoxu Cai, Lin Qi, and Mike Chantler. 2015. Visual perception of procedural textures: Identifying perceptual dimensions and predicting generation models. *PLOS ONE* 10, 6 (2015), 1–22. https://doi.org/10.1371/journal.pone.0130335

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *J. of Mach. Learn. Res.* 9 (2008), 2579–2605.

Benoit B. Mandelbrot. 1988. Fractal landscapes without creases and with rivers. In *The Science of Fractal Images* (1st ed.). Springer, 243–260.

Rafal Mantiuk, Kil Joong Kim, Allan G. Rempel, and Wolfgang Heidrich. 2011. HDR-VDP-2: A calibrated visual metric for visibility and quality predictions in all luminance conditions. *ACM Trans. on Graph.* 30, 4, Article 40 (July 2011), 14 pages.

Rafal Mantiuk, Karol Myszkowski, and Hans-Peter Seidel. 2006. A perceptual framework for contrast processing of high dynamic range images. *ACM Trans. Appl. Percept.* 3, 3 (2006), 286–308.

Gavin Miller. 1986. The definition and rendering of terrain maps. *Comp. Graph.* 20, 4 (1986), 39–48.

J. A. Milne and D. A. Sears. 1997. Modelling river channel topography using GIS. *Int. J. Geograph. Inf. Sci.* 11, 5 (1997), 499–519. https://doi.org/10.1080/136588197242275

Forest Kenton Musgrave, Craig E. Kolb, and Robert S. Mace. 1989. The synthesis and rendering of eroded fractal terrains. *Comp. Graph.* 23, 3 (1989), 41–50.

Georges Nader, Kai Wang, Franck Htroy-Wheeler, and Florent Dupont. 2016. Visual contrast sensitivity and discrimination for 3D meshes and their applications. *Comp. Gr. Forum* 35, 7 (2016), 497–506. https://doi.org/10.1111/cgf.13046

Benjamin Neidhold, Markus Wacker, and Oliver Deussen. 2005. Interactive physically based fluid and erosion simulation. In *Proc. of Eurographics Conference on Natural Phenomena*. 25–33. https://doi.org/10.2312/NPH/NPH05/025-032

Markus Neteler and Helena Mitasová. 2013. *Open Source GIS: A GRASS GIS Approach*, Vol. 689. Springer Science & Business Media.

Laszlo Neumann, Martin Čadík, and Antal Nemcsics. 2007. An efficient perception-based adaptive color to gray transformation. In *Proc. of Computational Aesthetics*. Eurographics Association, 73–80.

Augustus Odena, Vincent Dumoulin, and Chris Olah. 2016. Deconvolution and checkerboard artifacts. *Distill* 1, 10 (2016), e3.

Jan Ondřej, Cathy Ennis, Niamh A. Merriman, and Carol O'Sullivan. 2016. FrankenFolk: Distinctiveness and attractiveness of voice and motion. *ACM Trans. Appl. Percept.* 13, 4, Article 20 (July 2016), 13 pages.

Carol O'Sullivan, Sarah Howlett, Rachel McDonnell, Yann Morvan, and Keith O'Conor. 2004. Perceptually adaptive graphics. In *Eurographics 2004 - STARs*. Eurographics Association. https://doi.org/10.2312/egst.20041029

James Palmer. 2003. Research agenda for landscape perception. In *Trends in Landscape Modelling, Proc. at Anhalt University of Applied Sciences*.
Maria Perez-Ortiz and Rafal K. Mantiuk. 2017. A practical guide and software for analysing pairwise comparison experiments. *arXiv preprint arXiv:1712.03686* (2017).
Ken Perlin. 1985. An image synthesizer. *SIGGRAPH Comp. Graph.* 19, 3 (1985), 287–296.
Tomas Polasek, David Hrusa, Bedrich Benes, and Martin Cadik. 2021. ICTree: Automatic perceptual metrics for tree models. *ACM Trans. on Graph.* 40, 6, Article 230 (Dec. 2021), 15 pages. https://doi.org/10.1145/3478513.3480519
Suren Deepak Rajasekaran. 2019. *Perceptual Evaluation and Metric for Terrain Models*. Ph.D. Dissertation. Purdue University, Purdue University Graduate School.
Martin Reddy. 2001. Perceptually optimized 3D graphics. *IEEE Comp. Graph. Appl.* 21, 5 (Sept. 2001), 68–75. https://doi.org/10.1109/38.946633
Erik Reinhard, Peter Shirley, Michael Ashikhmin, and Tom Troscianko. 2004. Second order image statistics in computer graphics. In *Proceedings of the 1st Symposium on Applied Perception in Graphics and Visualization (APGV'04)*. Association for Computing Machinery, New York, NY, 99–106. https://doi.org/10.1145/1012551.1012568
Paul S. A. Reitsma and Nancy S. Pollard. 2003. Perceptual metrics for character animation: Sensitivity to errors in ballistic motion. *ACM Trans. on Graph.* 22, 3 (July 2003), 537–542.
Nicolas Riche, Matthieu Duvinage, Matei Mancas, Bernard Gosselin, and Thierry Dutoit. 2013. Saliency and human fixations: State-of-the-art and study of comparison metrics. In *IEEE ICCV*. 1153–1160. https://doi.org/10.1109/ICCV.2013.147
Holly Rushmeier, Bernice Rogowitz, and Christine Piatko. 2000. Perceptual issues in substituting texture for geometry. In *SPIE: Human Vision and Electronic Imaging*, Vol. 3959. 372–383. https://doi.org/10.1117/12.387174
Michael Schwarz and Marc Stamminger. 2009. On predicting visual popping in dynamic scenes. In *Proc. of the APGV*. ACM, 93–100. https://doi.org/10.1145/1620993.1621012
Joshua J. Scott and Neil A. Dodgson. 2021. Example-based terrain synthesis with pit removal. *Comp. Graph.* 99 (2021), 43–53. https://doi.org/10.1016/j.cag.2021.06.012
Yinxuan Shi, Jan Ondřej, He Wang, and Carol O'Sullivan. 2017. Shape up! Perception based body shape variation for data-driven crowds. In *2017 IEEE Virtual Humans and Crowds for Immersive Environments (VHCIE'17)*. 1–7. https://doi.org/10.1109/VHCIE.2017.7935623
Kaleigh Smith, Pierre-Edouard Landes, Joëlle Thollot, and Karol Myszkowski. 2008. Apparent greyscale: A simple and fast conversion to perceptually accurate images and video. *Comp. Gr. Forum* 27, 2 (2008), 193–200.
Ondřej St'ava, Bedrich Benes, Matthew Brisbin, and Jaroslav Křivánek. 2008. Interactive terrain modeling using hydraulic erosion. In *Proc. of the SCA (SCA'08)*. Eurographics Association, 201–210.
Robert M. W. Travers. 1984. *Human Information Processing*. ERIC.
Aurelie Tremblet. 2016. The mountain sublime of Philip James de Loutherbourg and Joseph Mallord William Turner. *J. of Alpine Res.* 45–54 (2016), 104–102. https://doi.org/10.4000/rga.3395
Mari Sundli Tveit, Asa Ode Sang, and Caroline M. Hagerhall. 2012. Scenic beauty: Visual landscape assessment and human landscape perception. In *Environmental Psychology: An Introduction*. BPS Blackwell, Chapter 4, 37–46.
Kiwon Um, Xiangyu Hu, and Nils Thuerey. 2017. Perceptual evaluation of liquid simulation methods. *ACM Trans. on Graph.* 36, 4 (2017), 143.
He Wang, Jan Ondřej, and Carol O'Sullivan. 2017. Trending paths: A new semantic-level metric for comparing simulated and real crowd data. *IEEE TVCG* 23, 5 (May 2017), 1454–1464. https://doi.org/10.1109/TVCG.2016.2642963
Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. 2004. Image quality assessment: From error visibility to structural similarity. *Trans. Img. Proc.* 13, 4 (April 2004), 600–612.
Martin Weier, Michael Stengel, Thorsten Roth, Piotr Didyk, Elmar Eisemann, Martin Eisemann, Steve Grogorick, Andr Hinkenjann, Ernst Kruijff, Marcus Magnor, Karol Myszkowski, and Philipp Slusallek. 2017. Perception-driven accelerated rendering. *Comp. Gr. Forum* 36, 2 (2017), 611–643. https://doi.org/10.1111/cgf.13150
Stefan Winkler and Praveen Mohandas. 2008. The evolution of video quality measurement: From PSNR to hybrid metrics. *IEEE Trans. on Broadcasting* 54, 3 (Sept. 2008), 660–668. https://doi.org/10.1109/TBC.2008.2000733
Krzysztof Wolski, Daniele Giunchi, Nanyang Ye, Piotr Didyk, Karol Myszkowski, Radoslaw Mantiuk, Hans-Peter Seidel, Anthony Steed, and Rafat K. Mantiuk. 2018. Dataset and metrics for predicting local visible differences. *ACM Trans. on Graph.* 37, 5, Article 172 (2018), 14 pages. https://doi.org/10.1145/3196493
Jinliang Wu, Xiaoyong Shen, Wei Zhu, and Ligang Liu. 2013. Mesh saliency with global rarity. *Graph. Models* 75, 5 (Sept. 2013), 255–264.
Peng Ye, Jayant Kumar, and David Doermann. 2014. Beyond human opinion scores: Blind image quality assessment based on synthetic scores. In *IEEE CVPR*. 4241–4248. https://doi.org/10.1109/CVPR.2014.540
Howard Zhou, Jie Sun, Greg Turk, and James M. Rehg. 2007. Terrain synthesis from digital elevation models. *Trans. on Visual. Comp. Graph.* 13, 4 (2007), 834–848.
Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *IEEE ICCV*.

Received August 2020; revised November 2021; accepted January 2022