Introduction to the Special Issue on Realistic Synthetic Data: Generation, Learning, Evaluation

This is the foreword of our special issue volume on Realistic Synthetic Data: Generation, Learning, Evaluation organized with the ACM Transactions on Multimedia Computing, Communications, and Applications. It presents the target of the special issue that relates to synthetic data for various modalities, e.g., signals, images, volumes, audio, and so on, controllable generation for learning from synthetic data, transfer learning and generalization of models, causality in data generation, addressing bias, limitations and trustworthiness in data generation, evaluation measures/protocols and benchmarks to assess quality of synthetic content, open synthetic datasets and software tools, and ethical aspects of synthetic data. The call for papers received a record number of 40 submissions out of which 15 were finally accepted for publication. This introduction provides an overview of the topics of each of the articles.

In the current context of **Machine Learning (ML)** and **Deep Learning (DL)**, data and especially high-quality data are central for ensuring proper training of the networks. It is well known that DL models require an important quantity of annotated data to be able to reach their full potential. Annotating content for models is traditionally made by human experts or at least by typical users, e.g., via crowdsourcing. This is a tedious task that is time-consuming and expensive—massive resources are required, content has to be curated, and so on. Moreover, there are specific domains where data confidentiality makes this process even more challenging, e.g., in the medical domain where patient data cannot be made publicly available, easily.

With the advancement of neural generative models such as **Generative Adversarial Networks (GAN)**, or, recently Diffusion Models, a promising way of solving or alleviating such problems that are associated with the need for domain-specific annotated data is to go toward realistic synthetic data generation. These data are generated by learning specific characteristics of the classes of target data. The advantage is that these networks allow for infinite variations within these classes while producing realistic outcomes, typically hard to distinguish from the real data. These data usually

CCS Concepts: • **Information systems** → **Information retrieval**; **Data management systems**; **Multimedia information systems**; **Digital libraries and archives**;

Additional Key Words and Phrases: synthetic data, Generative Adversarial Networks, Diffusion Models, data augmentation, data generation, datasets

ACM Reference format:
Bogdan Ionescu, Ioannis Patras, Henning Müller and Alberto Del Bimbo. 2024. Introduction to the Special Issue on Realistic Synthetic Data: Generation, Learning, Evaluation. *ACM Trans. Multimedia Comput. Commun. Appl.* 21, 1, Article 1 (December 2024), 7 pages.
https://doi.org/10.1145/3703593

The special issue was endorsed by the AI4Media project, a European Excellence Centre for Media, Society and Democracy, H2020 ICT-48-2020, grant #951911.

ACM 1551-6865/2024/12-ART1
https://doi.org/10.1145/3703593

have no proprietary or confidentiality restrictions and seem a viable solution to generate new datasets or augment existing ones, for example, by generating more examples for less represented classes. Existing results show very promising results for signal generation, images, and so on.

Nevertheless, there are some limitations that need to be overcome so as to advance the field. For instance, how can one control/manipulate the latent codes of GANs, or the diffusion process, so as to produce the desired classes and the desired variations like real data in the output? In many cases, results are not of high quality and selection should be made by the user, which is similar to manual annotation. Bias may intervene in the generation process due to the bias in the input dataset. Are the networks trustworthy? Is the generated content violating data privacy? In some cases one can predict the actual data source based on a generated image used for training the network. Would it be possible to train the networks to produce new classes and learn causality of the data? How do we objectively assess the quality of the generated data? These are just a few research questions that need to be addressed.

In this context, we organized the ACM Transactions on Multimedia Computing, Communications, and Applications special issue on Realistic Synthetic Data: Generation, Learning, Evaluation, to trigger interesting discussions and to seek solutions to the aforementioned challenges of synthetic data.

The special issue called for innovative algorithms and approaches addressing the following topics (but not limited to):

—Synthetic data for various modalities, e.g., signals, images, volumes, audio, and so on.
—Controllable generation for learning from synthetic data.
—Transfer learning and generalization of models.
—Causality in data generation.
—Addressing bias, limitations, and trustworthiness in data generation.
—Evaluation measures/protocols and benchmarks to assess quality of synthetic content.
—Open synthetic datasets and software tools.
—Ethical aspects of synthetic data.

The topic is of high interest for the community and the call attracted a lot of interest, receiving a high number of 40 submissions, out of which 15 have been accepted, therefore a 38% acceptance rate. Each article was reviewed by at least three peer reviewers and underwent at least one revision round.

The overview of the accepted articles is presented as below.

The article on "Synthetic Data for Object Detection with Neural Networks: State of the Art Survey of Domain Randomisation Techniques" analyzes the research done in the area of Domain Randomisation applied to Neural Networks predominant in object detection. It proposes a set of criteria for comparison of previously published works and utilizes these criteria to make conclusions about various trends in the area, similarities/differences, and key discoveries made since conception. The purpose of this work is to advise practitioners on leading solutions and help researchers gain a better understanding of the landscape. The key takeaways from this analysis show that current state-of-the-art solutions within the mid-quartile range allow object detection with typically about 1–25% performance decrease in comparison to manually annotated datasets; while the top performant approaches above the upper quartile gain about 2–32% lead over real data training in their specific application areas.

Another article "GANs in the Panorama of Synthetic Data Generation Methods" focuses on the creation and evaluation of synthetic data to address the challenges of imbalanced datasets in ML applications, using fake news detection as a case study. It conducts a thorough literature review on GANs for tabular data, synthetic data generation methods, and synthetic data quality assessment.

By augmenting a public news dataset with synthetic data generated by different GAN architectures, it demonstrates the potential of synthetic data to improve ML models' performance in fake news detection. It also modifies and extends a data usage approach to evaluate the quality of synthetic data and investigates the relationship between synthetic data quality and data augmentation performance in classification tasks. There is a positive correlation between synthetic data quality and performance in the under-represented class, highlighting the importance of high-quality synthetic data for effective data augmentation.

The article on "Synthesized Image Training Techniques: On Improving Model Performance Using Confusion" seeks to improve model performance on limited labeled datasets by reducing confusion. It observes that misclassification (or confusion) between classes is usually more prevalent between specific classes and accordingly it introduces SIT2, a novel confusion-based training framework that identifies pairs of classes with high confusion and synthesizes not-sure images from these pairs. The not-sure images are utilized in three new training strategies as follows: (i) The not-sure training strategy pre-trains a model using not-sure images and the original training images, (ii) The sure-or-not strategy pre-trains with only synthesized images into the sure or not-sure class, and (iii) The multi-label strategy also synthesizes and pre-trains with only synthesized images but predicts the original class(es) of the synthesized images. Finally, the pre-trained model is finetuned on the original training dataset. An extensive evaluation is proposed on five open medical and non-medical datasets. Several improvements are statistically significant, which shows the promising future of our confusion-based training framework.

The article on "GAN-Assisted Road Segmentation from Satellite Imagery" proposes a GAN-assisted training scheme for road segmentation from high-resolution RGB color satellite images, which includes three critical components: (i) synthetic training sample generation, (ii) synthetic training sample selection, and (iii) assisted training strategy. Apart from the GeoPalette and cSinGAN image generators introduced in the previous literature, this work explains how to generate new training pairs using OpenStreetMap and introduces a new set of evaluation metrics for selecting synthetic training pairs from a pool of generated samples. Extensive quantitative and qualitative experiments are conducted to compare different image generators and training strategies. The experiments on the downstream road segmentation task show that the proposed metrics are more aligned with the trained model performance compared to commonly used GAN evaluation metrics, such as the Fréchet inception distance, as well as the fact that by using synthetic data with the best training strategy, the model performance, mean Intersection over Union is improved from 60.92% to 64.44%, when 1,000 real training pairs are available for learning, which reaches a similar level of performance as a model that is standard-trained with 4,000 real images (64.59%), e.g., enabling a four-fold reduction in real dataset size.

"GANonymization: A GAN-Based Face Anonymization Framework for Preserving Emotional Expressions" introduces GANonymization, a novel face anonymization framework with facial expression-preserving abilities. The proposed approach is based on a high-level representation of a face which is synthesized into an anonymized version based on GAN. The effectiveness of the approach is assessed by evaluating its performance in removing identifiable facial attributes to increase the anonymity of the given individual face. Additionally, the performance of preserving facial expressions was evaluated on several affect recognition datasets and outperformed the state-of-the-art methods in most categories. Finally, the proposed approach was analyzed for its ability to remove various facial traits, such as jewelry, clothing, and hair and demonstrated reliable performance in such tasks.

Another article is on "4D Facial Expression Diffusion Model" and introduces a generative framework for generating 3D facial expression sequences, i.e., 4D faces, that can be conditioned on different inputs to animate an arbitrary 3D face mesh. It is composed of two tasks: (i) Learning the

generative model that is trained over a set of 3D landmark sequences, and (ii) Generating 3D mesh sequences of an input facial mesh driven by the generated landmark sequences. The generative model is based on a Denoising Diffusion Probabilistic Model, which has achieved remarkable success in generative tasks of other domains. While it can be trained unconditionally, its reverse process can still be conditioned by various condition signals. This allows efficient development of several downstream tasks involving various conditional generation, by using expression labels, text, partial sequences, or simply a facial geometry. To obtain the full mesh deformation, the authors introduced a landmark-guided encoder-decoder to apply the geometrical deformation embedded in landmarks on a given facial mesh. Experiments show that the proposed model has learned to generate realistic, quality expressions solely from the dataset of relatively small size, improving over the state-of-the-art methods.

The contribution on "Text-Guided Synthesis of Masked Face Images" addresses the solutions to tackle pandemic face recognition considering solutions such as: (i) Train the face recognition systems to identify the person with the upper face features, (ii) Reconstruct the complete face of the person with a generative model, and (iii) Train the model with a dataset of the masked faces of the people. It explores the scope of generative models for image synthesis and uses stable diffusion to generate masked face images of popular celebrities on various text prompts. A realistic dataset of 15k masked face images of 100 celebrities is generated and denoted as the Realistic Synthetic Masked Face Dataset. This seems to be the largest masked face recognition dataset with realistic images. The generated images were tested on popular deep face recognition models and achieved significant results. The dataset is also trained and tested on some of the famous image classification models, and the results are competitive.

Another article is on "Double Reference Guided Interactive 2D and 3D Caricature Generation" which proposes the first geometry and texture (double) referenced interactive 2D and 3D caricature generating and editing method. The main challenge of caricature generation lies in the fact that it not only exaggerates the facial geometry but also refreshes the facial texture. This challenge is addressed by utilizing the semantic segmentation maps as an intermediary domain, removing the influence of photo texture while preserving the person-specific geometry features. Specifically, the proposed method consists of two main components: 3D-CariNet and CariMaskGAN. 3D-CariNet uses sketches/caricatures to exaggerate the input photo into several types of 3D caricatures. To generate a CariMask, it geometrically exaggerates the photos using the projection of exaggerated 3D landmarks, after which CariMask is converted into a caricature by CariMaskGAN. In this step, users can edit and adjust the geometry of caricatures freely. Moreover, it proposes a semantic detail pre-processing approach that considerably increases the details of generated caricatures and allows modification of hair strands, wrinkles, and beards. By rendering high-quality 2D caricatures as textures, it produces 3D caricatures with a variety of texture styles.

"RSUIGM: Realistic Synthetic Underwater Image Generation with Image Formation Model" proposes to synthesize realistic underwater images with a novel image formation model, considering both downwelling depth and **Line of Sight (LoS)** distance as cue, denoted as the **Realistic Synthetic Underwater Image Generation Model (RSUIGM)**. The light interaction in the ocean is a complex process and demands specific modeling of direct and backscattering phenomena to capture the degradations. Most of the image formation models rely on complex radiative transfer models and *in situ* measurements for synthesizing and restoration of underwater images. Typical image formation models consider only LOS distance z and ignore downwelling depth d in the estimation of effect of direct light scattering. The authors derive the dependencies of downwelling irradiance in direct light estimation for generation of synthetic underwater images unlike state-of-the-art image formation models. They propose to incorporate the derived downwelling irradiance in estimation of direct light scattering for modeling the image formation process and

generate realistic synthetic underwater images with the proposed RSUIGM. The quality of restored images is compared with state-of-the-art methods using benchmark real underwater image datasets and achieve improved results. In addition, the distribution of realistic synthetic underwater images versus real underwater images is validated both qualitatively and quantitatively.

The article "Parents and Children: Distinguishing Multimodal DeepFakes from Natural Images" pioneers a systematic study on deepfake detection generated by state-of-the-art diffusion models. Firstly, it conducts a comprehensive analysis of the performance of contrastive and classification-based visual features, respectively extracted from CLIP-based models and ResNet or ViT-based architectures trained on image classification datasets. The results demonstrate that fake images share common low-level cues, which render them easily recognizable. Further, it devises a multi-modal setting wherein fake images are synthesized by different textual captions, which are used as seeds for a generator. Under this setting, the performance of fake detection strategies is quantified and introduces a contrastive-based disentangling method that lets us analyze the role of the semantics of textual descriptions and low-level perceptual cues. Finally, it releases a new dataset, called COCOFake, containing about 1.2M images generated from the original COCO image-caption pairs using two recent text-to-image diffusion models, namely Stable Diffusion v1.4 and v2.0.

Another article is on "New Metrics and Dataset for Biological Development Video Generation." The authors propose a new dataset, denoted GoldenDOT, which tracks the evolution of apples cut in parallel over 10 days, allowing to observe their progress over time while remaining static. Four new metrics are proposed that provide different analyses of the generated videos as a whole and individually. The proposed dataset and measures are used to study three state-of-the-art video generative models and their feasibility for video generation with biological development: **TemporalGAN (TGANv2)**, Low Dimensional Video Discriminator GAN, and Video Diffusion Model. Among them, the TGANv2 model proved to obtain the best results in the vast majority of metrics, including those already known in the state of the art, demonstrating the viability of the new proposed metrics and their congruence with these standard measures.

"Generating and Evaluating Data of Daily Activities with an Autonomous Agent in a Virtual Smart Home" shifts the attention to the identification of human behavior. Due to the high variety of home environments and occupant behaviors, collecting datasets that are representative of all possible homes is a major challenge. In addition, privacy and cost are major hurdles to collect real home data. One solution consists of training these models using purely synthetic data, which can be generated through the simulation of home and their occupants. Two challenges arise from this approach: (i) Designing a methodology with a simulation able to generate credible simulated data, and (ii) Evaluating this credibility. The article explains the methodology used to generate diversified synthetic data of daily activities, through the combination of an agent model to simulate an occupant, and a simulated 3D house enriched with sensors and effectors to produce such data. It demonstrates the credibility of the generated synthetic data by comparing their efficacy for training human context understanding models against the efficacy generated by real data. It replicates a real dataset collection setting with a smart home simulator. The occupant is replaced by an autonomous agent following the same experimental protocol used for the real dataset collection. This agent is a BDI-based model enhanced with a scheduler designed to offer a balance between control and autonomy. This balance is useful in synthetic data generation since strong constraints can be imposed on the agent to simulate desired situations while allowing autonomous behaviors outside these constraints to generate diversified data. The simulated sensors and effectors were configured to react to the agent's behaviors similarly to the real ones. It experimentally shows that data generated from this simulation are valuable for two human context understanding tasks: (i) Current human activity recognition, and (ii) Future human activity prediction.

The contribution on "Autoregressive GAN for Semantic Unconditional Head Motion Generation" addresses the task of unconditional head motion generation to animate still human faces in a low-dimensional semantic space from a single reference pose. Different from traditional audio-conditioned talking head generation that seldom puts emphasis on realistic head motions, it devises a GAN-based architecture that learns to synthesize rich head motion sequences over long duration while maintaining low error accumulation levels. In particular, the autoregressive generation of incremental outputs ensures smooth trajectories, while a multi-scale discriminator on input pairs drives generation toward better handling of high- and low-frequency signals and less mode collapse. It experimentally demonstrates the relevance of the proposed method and shows its superiority compared to models that attained state-of-the-art performances on similar tasks.

The article on "DashReStreamer: Framework for Creation of Impaired Video Clips under Realistic Network Conditions" addresses **Quality of Experience (QoE)**. To evaluate user QoE, subjective quality assessment, where people watch and grade videos, and objective quality assessment in which videos are graded using one or many objective metrics, are to be conducted. While there is a plethora of video databases available for subjective and objective video quality assessment, these videos are artificially infused with various temporal and spatial impairments. Videos being assessed are artificially distorted with startup delay, bitrate changes, and stalls due to rebuffering events. To conduct a more credible quality assessment, a reproduction of original user experiences while watching different types of streams on different types and quality of networks is needed. To help current efforts in bridging the gap between the mapping of objective video QoE metrics to user experience, it develops the DashReStreamer, an open source framework for re-creating adaptively streamed video in real networks. DashReStreamer also calculates popular video quality metrics like PSNR, SSIM, MS-SSIM, and VMAF. Finally, DashReStreamer allows creating impaired video sequences from the popular streaming platform, YouTube. As a demonstration of framework usage it creates a database of 234 realistic video clips, based on video logs collected from real mobile and wireless networks. Every video clip is supplemented with bandwidth trace and video logs used in its creation and also with objective metrics calculation reports.

Another article is on "Exploring Generative Adversarial Networks for Augmenting Network Intrusion Detection Tasks." It proposes using the power of GANs to create and augment flow-based network datasets. It evaluates a series of GAN architectures, including Wasserstein, conditional, energy-based, gradient penalty, and LSTM GANs. It trains and tests their performances on a set of flow-based network traffic data, collected from 16 subjects that used their computers for home, work, and study purposes. The performance of these GAN architectures is described according to metrics that involve networking principles, data distribution among a collection of flows, and temporal data distribution. Given the tendency of network intrusion detection datasets and benchmarks to have a very imbalanced data distribution, i.e., a large number of samples in the "normal traffic" category and a comparatively low number of samples assigned to the "intrusion" categories, it tests GANs by augmenting the intrusion data and checking whether this helps intrusion detector networks in their task.

In conclusion, the special issue attracted a lot of interest from the community, highlighting again the high importance of the subject as well as its impact on shaping the field of AI.

Bogdan Ionescu
National University of Science and Technology Politehnica Bucharest, Romania

Ioannis Patras
Queen Mary University of London, UK

Henning Müller
University of Applied Sciences Western Switzerland, Delémont, Switzerland

Alberto Del Bimbo
University of Florence, Florence, Italy