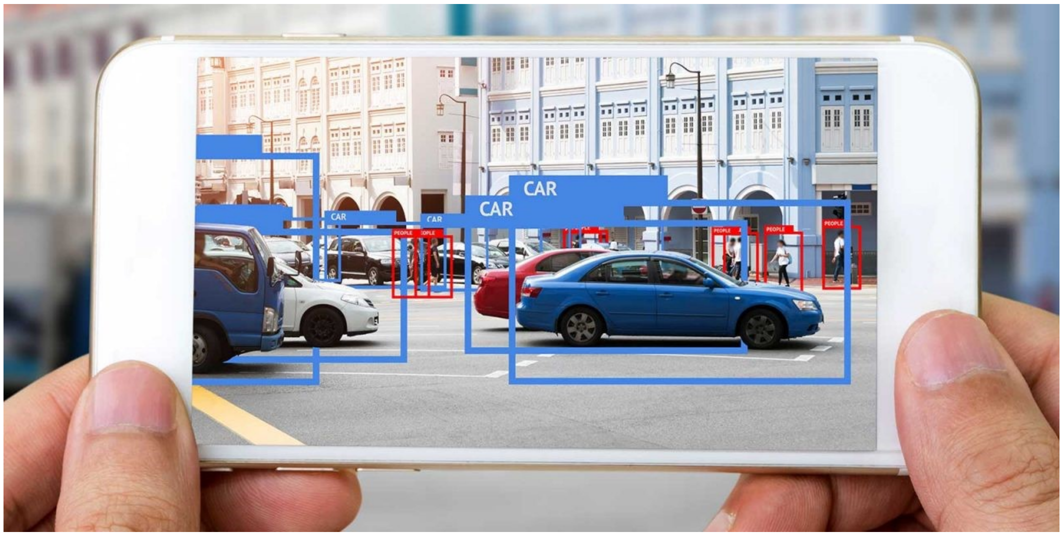
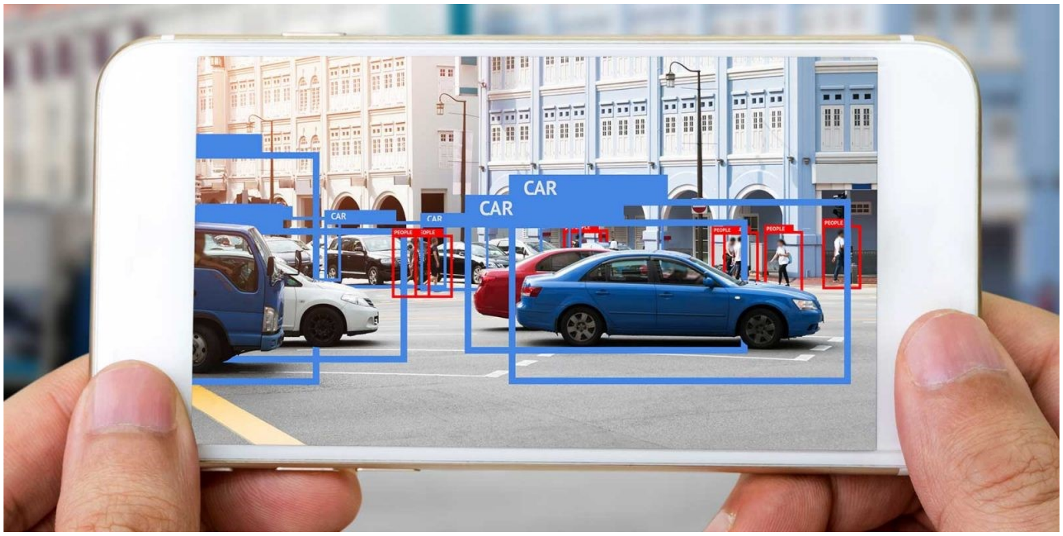


10 Cutting Edge Research Papers In Computer Vision & Image Generation

www.topbots.com
25 mins read



Ever since convolutional neural networks began outperforming humans in specific image recognition tasks, research in the field of computer vision has proceeded at breakneck pace.

The basic architecture of CNNs (or ConvNets) was developed in the 1980s. Yann LeCun improved upon the original design in 1989 by using backpropagation to train models to recognize handwritten digits.

We've come a long way since then.

In 2018, we saw novel architecture designs that improve upon performance benchmarks and also expand the range of media that machine learning models can analyze.

We also saw a number of breakthroughs with media generation which enable photorealistic style transfer, high-resolution image generation, and video-to-video synthesis.

Due to the importance and prevalence of computer vision and image generation for applied and enterprise AI, we did feature some of the papers below in our previous article summarizing the top overall machine learning papers of 2018. Since you might not have read that previous piece, we chose to highlight the vision-related research ones again here.

We've done our best to summarize these papers correctly, but if we've made any mistakes, please contact us to request a fix. Special thanks also goes to computer vision specialist Rebecca BurWei for generously offering her expertise in editing and revising drafts of this article.

If these summaries of scientific AI research papers are useful for you, you can subscribe to our AI Research mailing list at the bottom of this article to be alerted when we release new summaries. We're planning to release summaries of important papers in computer vision, reinforcement learning, and conversational AI in the next few weeks.

If you'd like to skip around, here are the papers we featured:

Important Computer Vision Research Papers of 2018

1. Spherical CNNs, by Taco S. Cohen, Mario Geiger, Jonas Koehler, and Max Welling

Original Abstract

Convolutional Neural Networks (CNNs) have become the method of choice for learning problems involving 2D planar images. However, a number of problems of recent interest have created a demand for models that can analyze spherical images. Examples include omnidirectional vision for drones, robots, and autonomous cars, molecular regression problems, and global weather and climate modelling. A naive application of convolutional networks to a planar projection of the spherical signal is destined to fail, because the space-varying distortions introduced by such a projection will make translational weight sharing ineffective.

In this paper we introduce the building blocks for constructing spherical CNNs. We propose a definition for the spherical cross-correlation that is both expressive and rotation-equivariant. The spherical correlation satisfies a generalized Fourier theorem, which allows us to compute it efficiently using a generalized (non-commutative) Fast Fourier Transform (FFT) algorithm. We demonstrate the computational efficiency, numerical accuracy, and effectiveness of spherical CNNs applied to 3D model recognition and atomization energy regression.

Our Summary

Omnidirectional cameras that are already used by cars, drones, and other robots capture a spherical image of their entire surroundings. We could analyze such spherical signals by projecting them to the plane and using CNNs. However, any planar projection of a spherical signal results in distortions. To overcome this problem, the group of researchers from the University of Amsterdam introduces the theory of spherical CNNs, the networks that can analyze spherical images without being fooled by distortions. The approach demonstrates its effectiveness for classifying 3D shapes and Spherical MNIST images as well as for molecular energy regression, an important problem in computational chemistry.

What's the core idea of this paper?

- Planar projections of spherical signals result in significant distortions as some areas look larger or smaller than they really are.
- Traditional CNNs are ineffective for spherical images because as objects move around the sphere, they also appear to shrink and stretch (think maps where Greenland looks much bigger than it actually is).
- The solution is to use a spherical CNN which is robust to spherical rotations in the input data. By preserving the original shape of the input data, spherical CNNs treat all objects on the sphere equally without distortion.

What's the key achievement?

- Introducing a mathematical framework for building spherical CNNs.

- Providing easy to use, fast and memory efficient PyTorch code for implementation of these CNNs.
- Providing the first empirical support for the utility of spherical CNNs for rotation-invariant learning problems:
    - classification of Spherical MNIST images
    - classification of 3D shapes,
    - molecular energy regression.

What does the AI community think?

- The paper won the Best Paper Award at ICLR 2018, one of the leading machine learning conferences.

What are future research areas?

- Development of a Steerable CNN for the sphere to analyze sections of vector bundles over the sphere (e.g., wind directions).
- Expanding the mathematical theory from 2D spheres to 3D point clouds for classification tasks that are invariant under reflections as well as rotations.

What are possible business applications?

- Models that can analyze spherical images are important for:
    - the omnidirectional vision for drones, robots, and autonomous cars;
    - molecular regression problems in computational chemistry;
    - global weather and climate modeling.

Where can you get implementation code?

- The authors provide the original implementation for this research paper on GitHub.

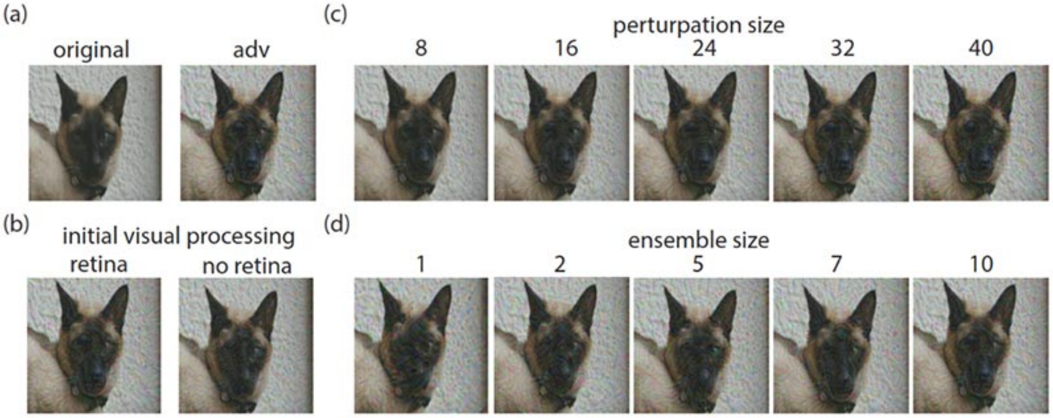
2. Adversarial Examples that Fool both Computer Vision and Time-Limited Humans, by Gamaleldin F. Elsayed, Shreya Shankar, Brian Cheung, Nicolas Papernot, Alex Kurakin, Ian Goodfellow, Jascha Sohl-Dickstein

Original Abstract

Machine learning models are vulnerable to adversarial examples: small changes to images can cause computer vision models to make mistakes such as identifying a school bus as an ostrich. However, it is still an open question whether humans are prone to similar mistakes. Here, we address this question by leveraging recent techniques that transfer adversarial examples from computer vision models with known parameters and architecture to other models with unknown parameters and architecture, and by matching the initial processing of the human visual system. We find that adversarial examples that strongly transfer across computer vision models influence the classifications made by time-limited human observers.

Our Summary

Google Brain researchers seek an answer to the question: do adversarial examples that are not model-specific and can fool different computer vision models without access to their parameters and architectures, can also fool time-limited humans? They leverage key ideas from machine learning, neuroscience, and psychophysics to create adversarial examples that do in fact impact human perception in a time-limited setting. Thus, the paper introduces a new class of illusions that are shared between machines and humans.



What's the core idea of this paper?

- As the first step, the researchers use the black box adversarial example construction techniques that create adversarial examples without access to the model's architecture or parameters.
- Then, they adapt computer vision models to mimic the initial visual processing of humans. This includes:
  - prepending each model with a retinal layer that pre-processes the input to incorporate some of the transformations performed by the human eye;
  - performing an eccentricity-dependent blurring of the image to approximate the input which is received by the visual cortex of human subjects through their retinal lattice.

- Classification decisions of humans are evaluated in a time-limited setting to detect even subtle effects in human perception.

What's the key achievement?

- Showing that adversarial examples that transfer across computer vision models do also successfully influence the perception of humans.
- Demonstrating the similarity between convolutional neural networks and the human visual system.

What does the AI community think?

- The paper is widely discussed by the AI community. While most of the researchers are stunned by the results, some argue that we need a stricter definition of adversarial image because if humans classify the perturbated picture of a cat as a dog than it's probably already a dog, not a cat.

What are future research areas?

- Researching which techniques are crucial for the transfer of adversarial examples to humans (i.e., retinal preprocessing, model ensembling).

What are possible business applications?

- Practitioners should consider the risk that imagery could be manipulated to cause human observers to have unusual

> reactions because adversarial images can affect us below the horizon of awareness.

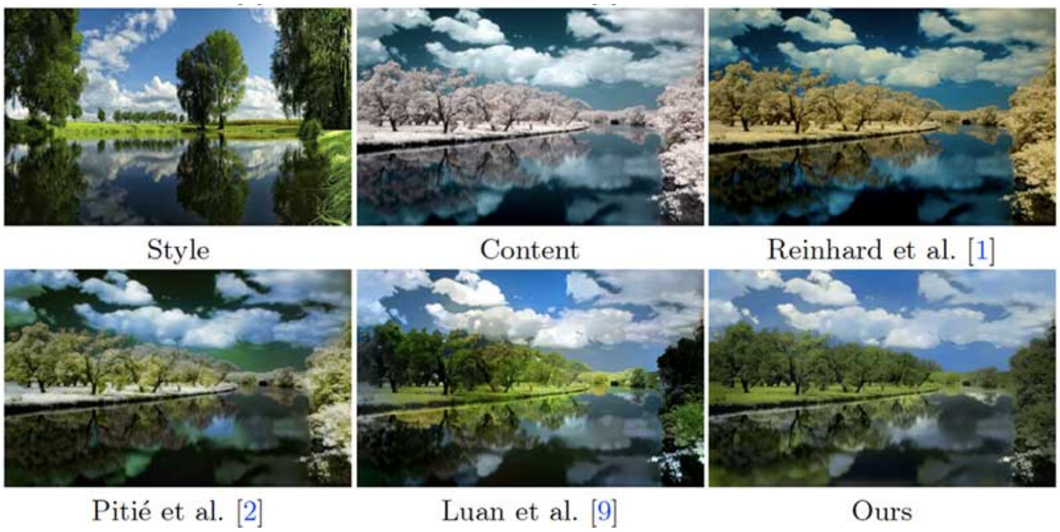
3. A Closed-form Solution to Photorealistic Image Stylization, by Yijun Li, Ming-Yu Liu, Xueting Li, Ming-Hsuan Yang, Jan Kautz

Original Abstract

Photorealistic image stylization concerns transferring style of a reference photo to a content photo with the constraint that the stylized photo should remain photorealistic. While several photorealistic image stylization methods exist, they tend to generate spatially inconsistent stylizations with noticeable artifacts. In this paper, we propose a method to address these issues. The proposed method consists of a stylization step and a smoothing step. While the stylization step transfers the style of the reference photo to the content photo, the smoothing step ensures spatially consistent stylizations. Each of the steps has a closed-form solution and can be computed efficiently. We conduct extensive experimental validations. The results show that the proposed method generates photorealistic stylization outputs that are more preferred by human subjects as compared to those by the competing methods while running much faster. Source code and additional results are available at https://github.com/NVIDIA/FastPhotoStyle.

Our Summary

The team of scientists at NVIDIA and the University of California, Merced propose a new solution to photorealistic image stylization, FastPhotoStyle. The method consists of two steps: stylization and smoothing. Extensive experiments show that the suggested approach generates more realistic and compelling images than previous state-of-the-art. Even more, thanks to the closed-form solution, FastPhotoStyle can produce the stylized image 49 times faster than traditional methods.



What's the core idea of this paper?

- The goal of photorealistic image stylization is to transfer style of a reference photo to a content photo while keeping the stylized image photorealistic.
- The task is split into the stylization and smoothing steps:
    - The stylization step is based on the whitening and coloring transform (WCT), which processes images via feature projections. However, WCT was developed for artistic image stylizations, and thus, often generates structural artifacts for photorealistic image stylization. To overcome this problem, the paper introduces PhotoWCT method, which replaces the upsampling layers in the WCT

    with unpooling layers, and so, preserves more spatial information.
  - The smoothing step is required to solve spatially inconsistent stylizations that could arise after the first step. Smoothing is based on a manifold ranking algorithm.

- Both steps have a closed-form solution, which means that the solution can be obtained in a fixed number of operations (i.e., convolutions, max-pooling, whitening, etc.). Thus, computations are much more efficient compared to the traditional methods.

What's the key achievement?

- Introducing a novel image stylization approach, FastPhotoSyle, which:
  - outperforms artistic stylization algorithms by rendering much fewer structural artifacts and inconsistent stylizations, and
  - outperforms photorealistic stylization algorithms by synthesizing not only colors but also patterns in the style photos.

- The experiments demonstrate that users prefer FastPhotoStyle results over the previous state-of-the-art in terms of both stylization effects (63.1%) and photorealism (73.5%).
- FastPhotoSyle can synthesize an image of 1024 x 512 resolution in only 13 seconds, while the previous state-of-the-art method needs 650 seconds for the same task.

What does the AI community think?

- The paper was presented at ECCV 2018, leading European Conference on Computer Vision.

What are future research areas?

- Finding the way to transfer small patterns from the style photo as they are smoothed away by the suggested method.
- Exploring the possibilities to further reduce the number of structural artifacts in the stylized photos.

What are possible business applications?

- Content creators in the business settings can largely benefit from photorealistic image stylization as the tool basically allows you to automatically change the style of any photo based on what fits the narrative.
- The photographers also discuss the tremendous impact that this technology can have in real estate photography.

Where can you get implementation code?

- NVIDIA team provides the original implementation for this research paper on GitHub.

4. Group Normalization, by Yuxin Wu and Kaiming He

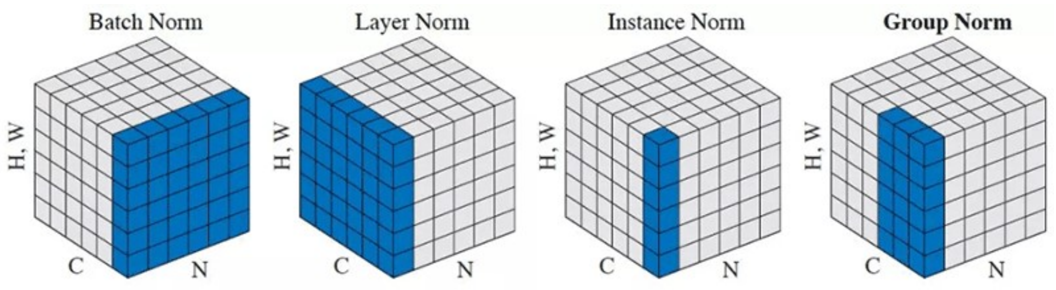
Original Abstract

Batch Normalization (BN) is a milestone technique in the development of deep learning, enabling various networks to train. However, normalizing along the batch dimension introduces problems – BN's error increases rapidly when the batch size becomes smaller, caused by inaccurate batch statistics estimation. This limits BN's usage for training larger models and transferring features to computer vision tasks including detection, segmentation, and video, which require small batches constrained by memory consumption. In this paper, we present Group Normalization (GN) as a simple alternative to BN. GN divides the channels into groups and computes within each group the mean and variance for normalization. GN's computation is independent of batch sizes, and its accuracy is stable in a wide range of batch sizes. On ResNet-50 trained in ImageNet, GN has 10.6% lower error than its BN counterpart when using a batch size of 2; when using typical batch sizes, GN is comparably good with BN and outperforms other normalization variants. Moreover, GN can be naturally transferred from pre-training to fine-tuning. GN can outperform its BN-based counterparts for object detection and segmentation in COCO, and for video classification in Kinetics, showing that GN can effectively replace the powerful BN in a variety of tasks. GN can be easily implemented by a few lines of code in modern libraries.

Our Summary

Facebook AI research team suggest Group Normalization (GN) as an alternative to Batch Normalization (BN). They argue that BN's error increases dramatically for small batch sizes. This limits the usage of BN when working with large models to solve computer vision tasks that require small batches due to memory constraints. On the contrary, Group Normalization is independent of batch sizes as it divides the channels into groups and computes the mean and variance for normalization within each group. The experiments confirm that

GN outperforms BN in a variety of tasks, including object detection, segmentation, and video classification.



What's the core idea of this paper?

- Group Normalization is a simple alternative to Batch Normalization, especially in the scenarios where batch size tends to be small, for example, computer vision tasks, requiring high-resolution input.
- GN explores only the layer dimensions, and thus, its computation is independent of batch size. Specifically, GN divides channels, or feature maps, into groups and normalizes the features within each group.
- Group Normalization can be easily implemented by a few lines of code in PyTorch and TensorFlow.

What's the key achievement?

- Introducing Group Normalization, new effective normalization method.
- Evaluating GN's behavior in a variety of applications and showing that:
    - GN's accuracy is stable in a wide range of batch sizes as its computation is independent of batch size. For example, GN demonstrated a 10.6%

    lower error rate than its BN-based counterpart for ResNet-50 in ImageNet with a batch size of 2.
    - GN can be also transferred to fine-tuning. The experiments show that GN can outperform BN counterparts for object detection and segmentation in COCO dataset and video classification in Kinetics dataset.

What does the AI community think?

- The paper received an honorable mention at ECCV 2018, leading European Conference on Computer Vision.
- It is also the second most popular paper in 2018 based on the people's libraries at Arxiv Sanity Preserver.

What are future research areas?

- Applying group normalization to sequential or generative models.
- Investigating GN's performance on learning representations for reinforcement learning.
- Exploring if GN combined with a suitable regularizer will improve results.

What are possible business applications?

- Business applications that rely on BN-based models for object detection, segmentation, video classification and other computer vision tasks that require high-resolution input may benefit from moving to GN-based models as they are more accurate in these settings.

Where can you get implementation code?

- Facebook AI research team provides Mask R-CNN baseline results and models trained with Group Normalization.
- PyTorch implementation of group normalization is also available on GitHub.

5. Taskonomy: Disentangling Task Transfer Learning, by Amir R. Zamir, Alexander Sax, William Shen, Leonidas J. Guibas, Jitendra Malik, and Silvio Savarese

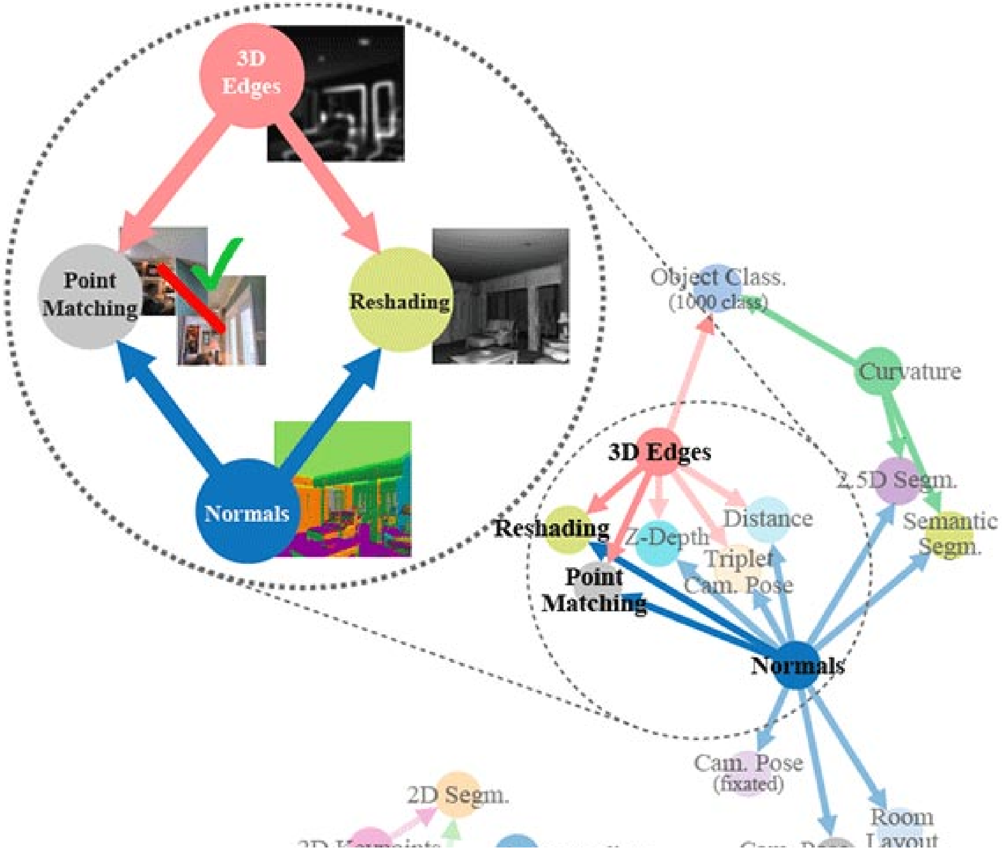
Original Abstract

Do visual tasks have a relationship, or are they unrelated? For instance, could having surface normals simplify estimating the depth of an image? Intuition answers these questions positively, implying existence of a structure among visual tasks. Knowing this structure has notable values; it is the concept underlying transfer learning and provides a principled way for identifying redundancies across tasks, e.g., to seamlessly reuse supervision among related tasks or solve many tasks in one system without piling up the complexity.

We proposes a fully computational approach for modeling the structure of space of visual tasks. This is done via finding (first and higher-order) transfer learning dependencies across a dictionary of twenty six 2D, 2.5D, 3D, and semantic tasks in a latent space. The product is a computational taxonomic map for task transfer learning. We study the consequences of this structure, e.g. nontrivial emerged relationships, and exploit them to reduce the demand for labeled data. For example, we show that the total number of labeled datapoints needed for solving a set of 10 tasks can be reduced by roughly 2/3 (compared to training independently) while keeping the performance nearly the same. We provide a set of tools for computing and probing this taxonomical structure including a solver that users can employ to devise efficient supervision policies for their use cases.

Our Summary

Assertions of the existence of a structure among visual tasks have been made by many researchers since the early years of modern computer science. And now Amir Zamir and his team make an attempt to actually find this structure. They model it using a fully computational approach and discover lots of useful relationships between different visual tasks, including the nontrivial ones. They also show that by taking advantage of these interdependencies, it is possible to achieve the same model performance with the labeled data requirements reduced by roughly ⅔.



What’s the core idea of this paper?

- A model aware of the relationships among different visual tasks demands less supervision, uses less computation, and behaves in more predictable ways.
- A fully computational approach to discovering the relationships between visual tasks is preferable because it avoids imposing prior, and possibly incorrect, assumptions: the priors are derived from either human intuition or analytical knowledge, while neural networks might operate on different principles.

What’s the key achievement?

- Identifying relationships between 26 common visual tasks.
- Showing how this structure helps in discovering types of transfer learning that will be most effective for each visual task.
- Creating a new dataset of 4 million images of indoor scenes including 600 buildings annotated with 26 tasks.

What does the AI community think?

- The paper won the Best Paper Award at CVPR 2018, the key conference on computer vision and pattern recognition.
- The results are very important as for the most real-world tasks large-scale labeled datasets are not available.

What are future research areas?

- To move from a model where common visual tasks are entirely defined by humans and try an approach where human-defined visual tasks are viewed as observed samples which are composed of computationally found latent subtasks.
- Exploring the possibility to transfer the findings to not entirely visual tasks, e.g. robotic manipulation.

What are possible business applications?

- Relationships discovered in this paper can be used to build more effective visual systems that will require less labeled data and lower computational costs.

Where can you get implementation code?

- The authors provide the original implementation for this research paper on GitHub.

6. Self-Attention Generative Adversarial Networks, by Han Zhang, Ian Goodfellow, Dimitris Metaxas, Augustus Odena

Original Abstract

In this paper, we propose the Self-Attention Generative Adversarial Network (SAGAN) which allows attention-driven, long-range dependency modeling for image generation tasks. Traditional convolutional GANs generate high-resolution details as a function of only spatially local points in lower-resolution feature maps. In SAGAN, details can be generated using cues from all feature locations. Moreover, the discriminator can check that highly detailed features in distant portions of the image are consistent with each other. Furthermore, recent work has shown that generator conditioning affects GAN performance. Leveraging this insight, we apply spectral normalization to the GAN generator and find that this improves training dynamics. The proposed SAGAN achieves the state-of-the-art results, boosting the best published Inception score from 36.8 to 52.52 and reducing Frechet Inception distance from 27.62 to 18.65 on the challenging ImageNet dataset. Visualization of the attention layers shows that the generator leverages neighborhoods that correspond to object shapes rather than local regions of fixed shape.

Our Summary

Traditional convolutional GANs demonstrated some very promising results with respect to image synthesis. However, they have at least one important weakness – convolutional layers alone fail to capture geometrical and structural patterns in the images. Since convolution is a local operation, it is hardly possible for an output on the top-left position to have any relation to the output at bottom-right. The paper introduces a simple solution to this problem – incorporating the self-attention mechanism into the GAN framework. This solution combined with several stabilization techniques helps the Senf-Attention Generative Adversarial Networks (SAGANs) achieve the state-of-the-art results in image synthesis.

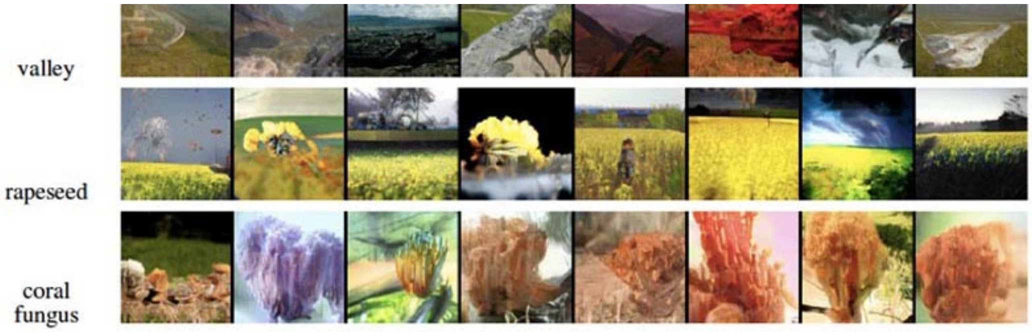


Figure 6: 128×128 example images generated by SAGAN for different classes. Each row shows samples from one class.

What's the core idea of this paper?

- Convolutional layers alone are computationally inefficient for modeling long-range dependencies in images. On the contrary, a self-attention mechanism incorporated into the GAN framework will enable both the generator and the discriminator to efficiently model relationships between widely separated spatial regions.
- The self-attention module calculates response at a position as a weighted sum of the features at all positions.
- The following techniques help to stabilize GAN training on challenging datasets:
    - Applying spectral normalization for both generator and discriminator – the researchers argue that not only the discriminator but also the generator can benefit from spectral normalization, as it can prevent the escalation of parameter magnitudes and avoid unusual gradients.
    - Using separate learning rates for the generator and the discriminator to compensate for the problem of slow learning in a regularized discriminator and make it possible to use fewer generator steps per discriminator step.

What’s the key achievement?

- Showing that self-attention module incorporated into the GAN framework is, in fact, effective in modeling long-range dependencies.
- Demonstrating the effectiveness of the proposed stabilization techniques for GAN training. In particular, showing that:
    - spectral normalization applied to the generator stabilizes GAN training;
    - utilizing imbalanced learning rates speeds up training of regularized discriminators.
- Achieving state-of-the-art results in image synthesis by boosting the Inception Score from 36.8 to 52.52 and reducing Fréchet Inception Distance from 27.62 to 18.65.

What does the AI community think?

- “The idea is simple and intuitive yet very effective, plus easy to implement.” – Sebastian Raschka, assistant professor of Statistics at the University of Wisconsin-Madison.

What are future research areas?

- Exploring the possibilities to reduce the number of weird samples generated by GANs.

What are possible business applications?

- Image synthesis with GANs can replace expensive manual media creation for advertising and e-commerce purposes.

Where can you get implementation code?

- PyTorch and TensorFlow implementations of Self-Attention GANs are available on GitHub.

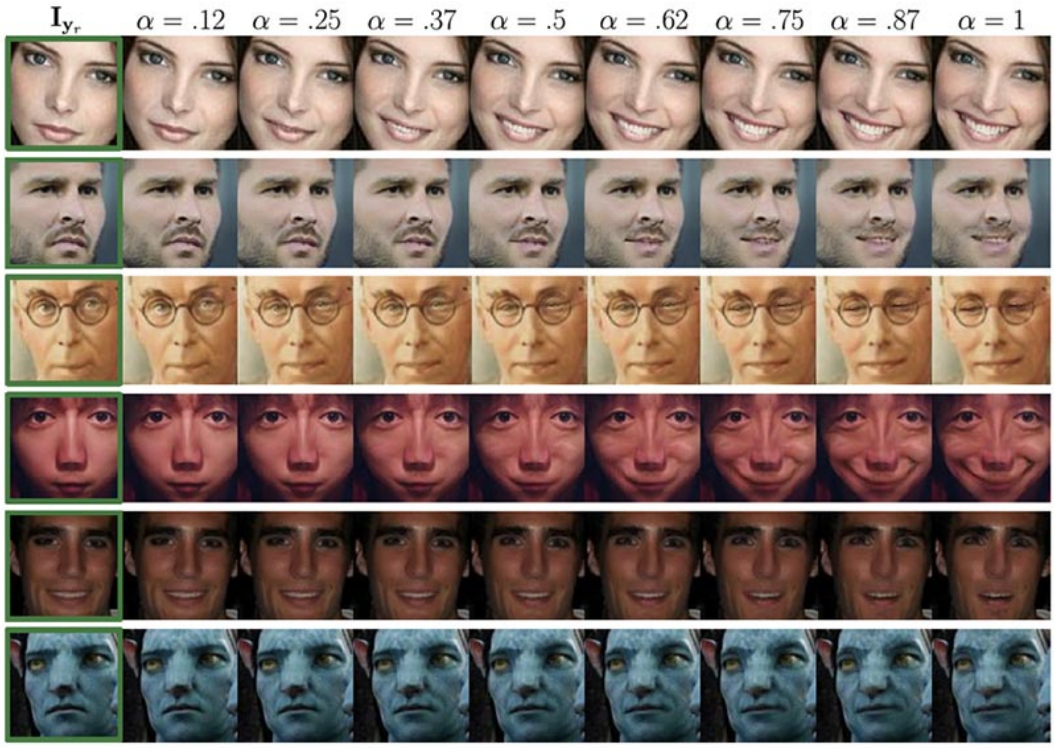
7. GANimation: Anatomically-aware Facial Animation from a Single Image, by Albert Pumarola, Antonio Agudo, Aleix M. Martinez, Alberto Sanfeliu, Francesc Moreno-Noguer

Original Abstract

Recent advances in Generative Adversarial Networks (GANs) have shown impressive results for task of facial expression synthesis. The most successful architecture is StarGAN, that conditions GANs generation process with images of a specific domain, namely a set of images of persons sharing the same expression. While effective, this approach can only generate a discrete number of expressions, determined by the content of the dataset. To address this limitation, in this paper, we introduce a novel GAN conditioning scheme based on Action Units (AU) annotations, which describes in a continuous manifold the anatomical facial movements defining a human expression. Our approach allows controlling the magnitude of activation of each AU and combine several of them. Additionally, we propose a fully unsupervised strategy to train the model, that only requires images annotated with their activated AUs, and exploit attention mechanisms that make our network robust to changing backgrounds and lighting conditions. Extensive evaluation show that our approach goes beyond competing conditional generators both in the capability to synthesize a much wider range of expressions ruled by anatomically feasible muscle movements, as in the capacity of dealing with images in the wild.

Our Summary

The paper introduces a novel GAN model that is able to generate anatomically-aware facial animations from a single image under changing backgrounds and illumination conditions. It advances current works, which had only addressed the problem for discrete emotions category editing and portrait images. The approach renders a wide range of emotions by encoding facial deformations as Action Units. The resulting animations demonstrate a remarkably smooth and consistent transformation across frames even with challenging light conditions and backgrounds.



What’s the core idea of this paper?

- Facial expressions can be described in terms of Action Units (AUs), which anatomically describe the contractions of specific facial muscles. For example, the facial expression for ‘fear’ is generally produced with the following activations: Inner Brow Raiser (AU1), Outer Brow Raiser (AU2), Brow Lowerer (AU4), Upper Lid Raiser (AU5), Lid Tightener (AU7), Lip Stretcher (AU20) and Jaw Drop (AU26). The magnitude of each AU defines the extent of emotion.
- A model for synthetic facial animation is based on the GAN architecture, which is conditioned on a one-dimensional vector indicating the presence/absence and the magnitude of each Action Unit.
- To circumvent the need for pairs of training images of the same person under different expressions, a bidirectional generator is used to both transform an image into a desired expression and transform the synthesized image back into the original pose.
- To handle images under changing backgrounds and illumination conditions, the model includes an attention layer that focuses the action of the network only in those regions of the image that are relevant to convey the novel expression.

What’s the key achievement?

- Introducing a novel GAN model for face animation in the wild that can be trained in a fully unsupervised manner and generate visually compelling images with remarkably smooth and consistent transformation across frames even with challenging light conditions and non-real world data.

- Demonstrating how a wider range of emotions can be generated by interpolating between emotions the GAN has already seen.

What does the AI community think?

- The paper received an honorable mention at ECCV 2018, leading European Conference on Computer Vision.

What are future research areas?

- Applying the introduced approach to video sequences.

What are possible business applications?

- The technology that automatically animates the facial expression from a single image can be applied in several areas including the fashion and e-commerce business, the movie industry, photography technologies.

Where can you get implementation code?

- The authors provide the original implementation of this research paper on GitHub.

8. Video-to-Video Synthesis, by Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, Bryan Catanzaro

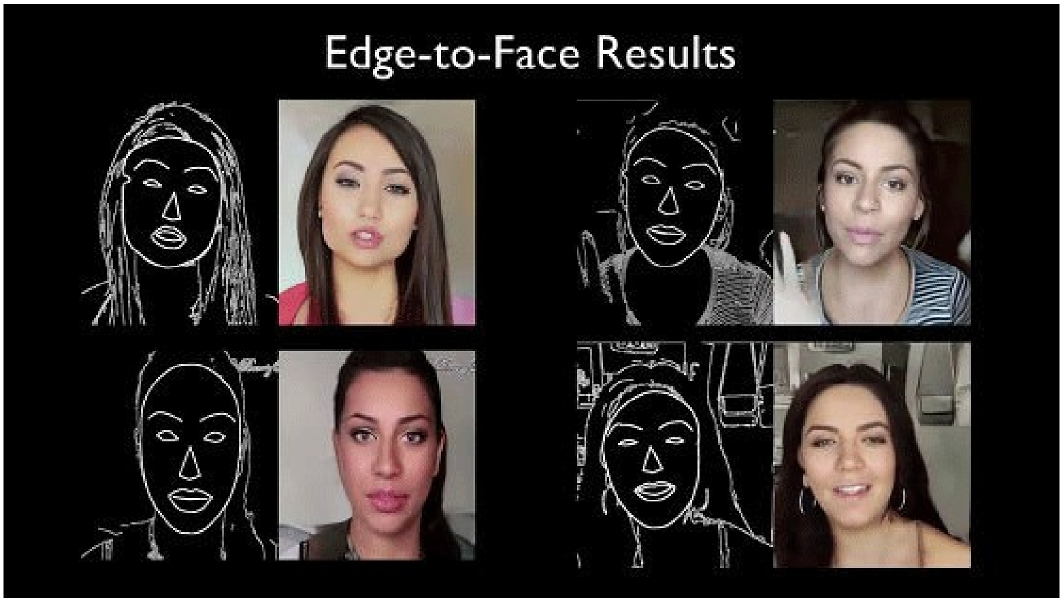
Original Abstract

We study the problem of video-to-video synthesis, whose goal is to learn a mapping function from an input source video (e.g., a sequence of semantic segmentation masks) to an output photorealistic video that precisely depicts the content of the source video. While its image counterpart, the image-to-image synthesis problem, is a popular topic, the video-to-video synthesis problem is less explored in the literature. Without understanding temporal dynamics, directly applying existing image synthesis approaches to an input video often results in temporally incoherent videos of low visual quality. In this paper, we propose a novel video-to-video synthesis approach under the generative adversarial learning framework. Through carefully-designed generator and discriminator architectures, coupled with a spatio-temporal adversarial objective, we achieve high-resolution, photorealistic, temporally coherent video results on a diverse set of input formats including segmentation masks, sketches, and poses. Experiments on multiple benchmarks show the advantage of our method compared to strong baselines. In particular, our model is capable of synthesizing 2K resolution videos of street scenes up to 30 seconds long, which significantly advances the state-of-the-art of video synthesis. Finally, we apply our approach to future video prediction, outperforming several state-of-the-art competing systems.

Our Summary

Researchers from NVIDIA have introduced a novel video-to-video synthesis approach. The framework is based on conditional GANs.

Specifically, the method couples carefully-designed generator and discriminator with a spatio-temporal adversarial objective. The experiments demonstrate that the suggested vid2vid approach can synthesize high-resolution, photorealistic, temporally coherent videos on a diverse set of input formats including segmentation masks, sketches, and poses. It can also predict the next frames with far superior results than the baseline models.



What's the core idea of this paper?

- Video frames can be generated sequentially, and the generation of each frame only depends on three factors:
    - current source frame;
    - past two source frames;
    - past two generated frames.
- Using multiple discriminators can mitigate the mode collapse problem during GANs training:

    - Conditional image discriminator ensures that each output frame resembles a real image given the same source image.
    - Conditional video discriminator ensures that consecutive output frames resemble the temporal dynamics of a real video given the same optical flow.

- Foreground-background prior in the generator design further improves the synthesis performance of the proposed model.
- Using a soft occlusion mask instead of binary allows to better handle the "zoom in" scenario: we can add details by gradually blending the warped pixels and the newly synthesized pixels.

What's the key achievement?

- Outperforming the strong baselines in video synthesis:

    - Generating high-resolution (2048x2048), photorealistic, temporally coherent videos up to 30 seconds long.
    - Outputting several videos with different visual appearances depending on sampling different feature vectors.

- Outperforming the baseline models in future video prediction.
- Open-sourcing a PyTorch implementation of the technique. This code can be used for:

    - Converting semantic labels into realistic real-world videos.

  - Generating multiple outputs of talking people from edge maps.
  - Generating an entire human body given a pose.

What does the AI community think?

- "NVIDIA's new vid2vid is the first open-source code that lets you fake anybody's face convincingly from one source video. [...] interesting times ahead...", Gene Kogan, an artist and a programmer.
- The paper has also received some criticism over the concern that it can be used to create deepfakes or tampered videos which can deceive people.

What are future research areas?

- Using object tracking information to make sure that each object has a consistent appearance across the whole video.
- Researching if training the model with coarser semantic labels will help reduce the visible artifacts that appear after semantic manipulations (e.g., turning trees into buildings).
- Adding additional 3D cues, such as depth maps, to enable synthesis of turning cars.

What are possible business applications?

- Marketing and advertising can benefit from the opportunities created by the vid2vid method (e.g., replacing the face or even the entire body in the video). However, this should be used with caution, keeping in mind the ethical considerations.

Where can you get implementation code?

- NVIDIA team provides the original implementation of this research paper on GitHub.

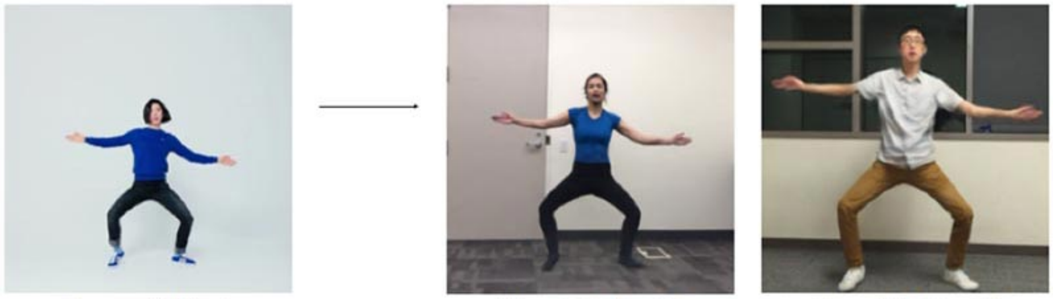
9. Everybody Dance Now, by Caroline Chan, Shiry Ginosar, Tinghui Zhou, Alexei A. Efros

Original Abstract

This paper presents a simple method for “do as I do” motion transfer: given a source video of a person dancing we can transfer that performance to a novel (amateur) target after only a few minutes of the target subject performing standard moves. We pose this problem as a per-frame image-to-image translation with spatio-temporal smoothing. Using pose detections as an intermediate representation between source and target, we learn a mapping from pose images to a target subject’s appearance. We adapt this setup for temporally coherent video generation including realistic face synthesis. Our video demo can be found at https://youtu.be/PCBTZh41Ris.

Our Summary

UC Berkeley researchers present a simple method for generating videos with amateur dancers performing like professional dancers. If you want to take part in the experiment, all you need to do is to record a few minutes of yourself performing some standard moves and then pick up the video with the dance you want to repeat. The neural network will do the main job: it solves the problem as a per-frame image-to-image translation with spatio-temporal smoothing. By conditioning the prediction at each frame on that of the previous time step for temporal smoothness and applying a specialized GAN for realistic face synthesis, the method achieves really amazing results.



Source Subject

Target Subject 1

Target Subject 2

What's the core idea of this paper?

- "Do as I do" motion transfer is approached as a per-frame image-to-image translation with the pose stick figures as an intermediate representation between source and target:
    - A pre-trained state-of-the-art pose detector creates pose stick figures from the source video.
    - Global pose normalization is applied to account for differences between the source and target subjects in body shapes and locations within the frame.
    - Normalized pose stick figures are mapped to the target subject.
- To make videos smooth, the researchers suggest conditioning the generator on the previously generated frame and then giving both images to the discriminator. Gaussian smoothing on the pose keypoints allows to further reduce jitter.
- To generate more realistic faces, the method includes an additional face-specific GAN that brushes up the face after the main generation is finished.

What's the key achievement?

- Suggesting a novel approach to motion transfer that outperforms a strong baseline (pix2pixHD), according to both qualitative and quantitative assessments.
- Demonstrating that face-specific GAN adds considerable detail to the output video.

What does the AI community think?

- "Overall I thought this was really fun and well executed. Looking forward to the code release so that I can start training my dance moves.", Tom Brown, member of technical staff at Google Brain.
- "'Everybody Dance Now' from Caroline Chan, Alyosha Efros and team transfers dance moves from one subject to another. The only way I'll ever dance well. Amazing work!!!", Soumith Chintala, AI Research Engineer at Facebook.

What are future research areas?

- Replacing pose stick figures with temporally coherent inputs and representation specifically optimized for motion transfer.

What are possible business applications?

- "Do as I do" motion transfer might be applied to replace subjects when creating marketing and promotional videos.

Where can you get implementation code?

- PyTorch implementation of this research paper is available on GitHub.

10. Large Scale GAN Training for High Fidelity Natural Image Synthesis, by Andrew Brock, Jeff Donahue, and Karen Simonyan

Original Abstract

Despite recent progress in generative image modeling, successfully generating high-resolution, diverse samples from complex datasets such as ImageNet remains an elusive goal. To this end, we train Generative Adversarial Networks at the largest scale yet attempted, and study the instabilities specific to such scale. We find that applying orthogonal regularization to the generator renders it amenable to a simple "truncation trick", allowing fine control over the trade-off between sample fidelity and variety by truncating the latent space. Our modifications lead to models which set the new state of the art in class-conditional image synthesis. When trained on ImageNet at 128×128 resolution, our models (BigGANs) achieve an Inception Score (IS) of 166.3 and Frechet Inception Distance (FID) of 9.6, improving over the previous best IS of 52.52 and FID of 18.65.

Our Summary

DeepMind team finds that current techniques are sufficient for synthesizing high-resolution, diverse images from available datasets such as ImageNet and JFT-300M. In particular, they show that

Generative Adversarial Networks (GANs) can generate images that look very realistic if they are trained at the very large scale, i.e. using two to four times as many parameters and eight times the batch size compared to prior art. These large-scale GANs, or BigGANs, are the new state-of-the-art in class-conditional image synthesis.

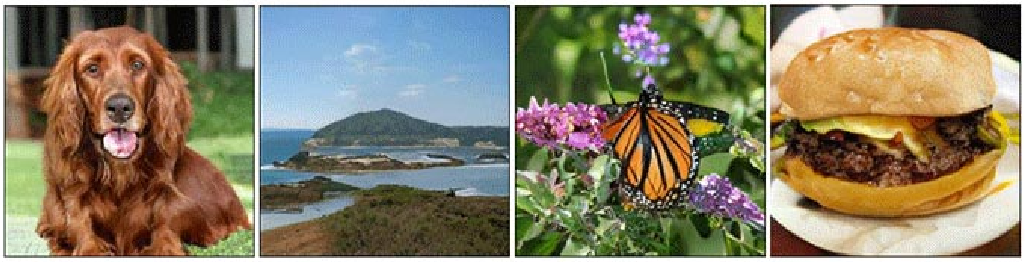


Figure 1: Class-conditional samples generated by our model.

What's the core idea of this paper?

- GANs perform much better with the increased batch size and number of parameters.
- Applying orthogonal regularization to the generator makes the model responsive to a specific technique ("truncation trick"), which provides control over the trade-off between sample fidelity and variety.

What's the key achievement?

- Demonstrating that GANs can benefit significantly from scaling.
- Building models that allow explicit, fine-grained control of the trade-off between sample variety and fidelity.
- Discovering instabilities of large-scale GANs and characterizing them empirically.

- BigGANs trained on ImageNet at 128×128 resolutions achieve:
    - an Inception Score (IS) of **166.3** with the previous best IS of 52.52;
    - Frechet Inception Distance (FID) of **9.6** with the previous best FID of 18.65.

What does the AI community think?

- The paper is under review for next ICLR 2019.
- After BigGAN generators become available on TF Hub, AI researchers from all over the world are playing with BigGANs to generate dogs, watches, bikini images, Mona Lisa, seashores and many more.

What are future research areas?

- Moving to larger datasets to mitigate GAN stability issues.
- Exploring the possibilities to reduce the number of weird samples generated by GANs.

What are possible business applications?

- Replacing expensive manual media creation for advertising and e-commerce purposes.

Where can you get implementation code?

- A BigGAN demo implemented in TensorFlow is available to use on Google's Colab tool.

- Aaron Leong has a Github repository for BigGAN implemented in PyTorch.

Want Deeper Dives Into Specific AI Research Topics?

Due to popular demand, we've released several of these easy-to-read summaries and syntheses of major research papers for different subtopics within AI and machine learning.

Enjoy this article? Sign up for more AI research updates.

We'll let you know when we release more summary articles like this one.

- Email Address*
- Name*
  First Last
- Company*
- What areas of AI research are you interested in? Select all that apply*
  - Natural Language Processing (NLP)
  - Chatbots & Conversational AI
  - Computer Vision
  - Ethics & Safety
  - Robotics
  - Machine Learning
  - Deep Learning

  - Reinforcement Learning
  - Generative Models
  - Other (Please Describe Below)

- What is your biggest challenge with AI research?*

■