# Thesis Work Special – Related Research Papers

1. AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks
2. Photographic Text-to-Image Synthesis with a Hierarchically-nested Adversarial Network
3. StackGAN++: Realistic Image Synthesis with Stacked Generative Adversarial Networks
4. Progressive Growing of GANs for Improved Quality, Stability, and Variation
5. Semantic Image Synthesis via Adversarial Learning
6. StackGAN: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks
7. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks
8. Photographic Image Synthesis with Cascaded Refinement Networks
9. Image-to-Image Translation with Conditional Adversarial Networks
10. Generative Adversarial Text to Image Synthesis
11. Learning Deep Representations of Fine-Grained Visual Descriptions
12. Generative Adversarial Nets
13. Stacked Generative Adversarial Networks
14. StoryGAN: A Sequential Conditional GAN for Story Visualization
15. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks
16. BEGAN: Boundary Equilibrium Generative Adversarial Networks
17. LR-GAN: Layered recursive generative adversarial networks for image generation
18. CanvasGAN: A simple baseline for text to image generation by incrementally patching a canvas
19. Improved Techniques for Training GANs
20. Sequential Attention GAN for Interactive Image Editing via Dialogue
21. Language-Based Image Editing with Recurrent Attentive Models
22. Conversational Image Editing: Incremental Intent Identification in a New Dialogue Task
23. Generative Image Modeling using Style and Structure Adversarial Networks
24. Generative Multi-Adversarial Network (GMAN)
25. Deep Generative Image Models using a Laplacian Pyramid of Adversarial Networks
26. Mask R-CNN
27. 
28. 
29. 
30.

### 1. AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks

*Tao Xu 1, Pengchuan Zhang 2, Qiuyuan Huang 2, Han Zhang 3, Zhe Gan 4, Xiaolei Huang 1, Xiaodong He 5*

*Abstract* - In this paper, we propose an Attentional Generative Adversarial Network (AttnGAN) that allows attention-driven, multi-stage refinement for fine-grained text-to-image generation. With a novel attentional generative network, the AttnGAN can synthesize fine-grained details at different sub-regions of the image by paying attentions to the relevant words in the natural language description. In addition, a deep attentional multimodal similarity model is proposed to compute a fine-grained image-text matching loss for training the generator. The proposed AttnGAN significantly out-performs the previous state of the art, boosting the best reported inception score by 14.14% on the CUB dataset and 170.25% on the more challenging COCO dataset. A detailed analysis is also performed by visualizing the attention layers of the AttnGAN. It for the first time shows that the layered attentional GAN is able to automatically select the condition at the word level for generating different parts of the image.

### 2. Photographic Text-to-Image Synthesis with a Hierarchically-nested Adversarial Network

*Zizhao Zhang ∗, Yuanpu Xie ∗, Lin Yang*

*Abstract* - This paper presents a novel method to deal with the challenging task of generating photographic images conditioned on semantic image descriptions. Our method introduces accompanying hierarchical-nested adversarial objectives inside the network hierarchies, which regularize mid-level representations and assist generator training to capture the complex image statistics. We present an extensile single-stream generator architecture to better adapt the jointed discriminators and push generated images up to high resolutions. We adopt a multi-purpose adversarial loss to encourage more effective image and text information usage in order to improve the semantic consistency and image fidelity simultaneously. Furthermore, we introduce a new visual-semantic similarity measure to evaluate the semantic consistency of generated images. With extensive experimental validation on three public datasets, our method significantly improves previous state of the arts on all datasets over different evaluation metrics.

### 3. StackGAN++: Realistic Image Synthesis with Stacked Generative Adversarial Networks

*Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Senior Member, IEEE, Xiaogang Wang, Member, IEEE, Xiaolei Huang, Member, IEEE, Dimitris N. Metaxas∗, Fellow, IEEE*

*Abstract* - Although Generative Adversarial Networks (GANs) have shown remarkable success in various tasks, they still face challenges in generating high quality images. In this paper, we propose Stacked Generative Adversarial Networks (StackGANs) aimed at generating high-resolution photo-realistic images. First, we propose a two-stage generative adversarial network architecture, StackGAN-v1, for text-to-image synthesis. The Stage-I GAN sketches the primitive shape and colors of a scene based on a given text description, yielding low-resolution images. The Stage-II GAN takes Stage-I results and the text description as inputs, and generates high-resolution images with photo-realistic details. Second, an advanced multi-stage generative adversarial network architecture, StackGAN-v2, is proposed for both conditional and unconditional generative tasks. Our StackGAN-v2 consists of multiple generators and multiple discriminators arranged in a tree-like structure; images at multiple scales corresponding to the same scene are generated from different branches of the tree. StackGAN-v2 shows more stable training behavior than StackGAN-v1 by jointly approximating multiple distributions. Extensive experiments demonstrate that the proposed stacked generative adversarial networks significantly outperform other state-of-the-art methods in generating photo-realistic images.

### 4. Progressive Growing of GANs for Improved Quality, Stability, and Variation

*Tero Karras, Timo Aila, Samuli Laine, Jaakko Lehtinen*

*Abstract* – We describe a new training methodology for generative adversarial networks. The key idea is to grow both the generator and discriminator progressively: starting from a low resolution, we add new layers that model increasingly fine details as training progresses. This both speeds the training up and greatly stabilizes it, allowing us to produce images of unprecedented quality, e.g., CELEBA images at 1024 2. We also propose a simple way to increase the variation in generated images, and achieve a record inception score of 8.80 in unsupervised CIFAR10. Additionally, we describe several implementation details that are important for discouraging unhealthy competition between the generator and discriminator. Finally, we suggest a new metric for evaluating GAN results, both in terms of image quality and variation. As an additional contribution, we construct a higher-quality version of the CELEBA dataset.

### 5. Semantic Image Synthesis via Adversarial Learning

*Hao Dong ∗, Simiao Yu ∗, Chao Wu, Yike Guo*

*Abstract* – In this paper, we propose a way of synthesizing realistic images directly with natural language description, which has many useful applications, e.g. intelligent image manipulation. We attempt to accomplish such synthesis: given a source image and a target text description, our model synthesizes images to meet two requirements: 1) being realistic while matching the target text description; 2) maintaining other image features that are irrelevant to the text description. The model should be able to disentangle the semantic information from the two modalities (image and text), and generate new images from the combined semantics. To achieve this, we proposed an end-to-end neural architecture that leverages adversarial learning to automatically learn implicit loss functions, which are optimized to fulfill the aforementioned two requirements. We have evaluated our model by conducting experiments on Caltech-200 bird dataset and Oxford-102 flower dataset, and have demon strated that our model is capable of synthesizing realistic images that match the given descriptions, while still maintain other features of original images.

### 6. StackGAN: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks

*Han Zhang 1, Tao Xu 2, Hongsheng Li 3, Shaoting Zhang 4, Xiaogang Wang 3, Xiaolei Huang 2, Dimitris Metaxas 1*

*Abstract* – Synthesizing high-quality images from text descriptions is a challenging problem in computer vision and has many practical applications. Samples generated by existing text-to-image approaches can roughly reflect the meaning of the given descriptions, but they fail to contain necessary details and vivid object parts. In this paper, we propose Stacked Generative Adversarial Networks (StackGAN) to generate 256×256 photo-realistic images conditioned on text descriptions. We decompose the hard problem into more manageable sub-problems through a sketch-refinement process. The Stage-I GAN sketches the primitive shape and colors of the object based on the given text description, yielding Stage-I low-resolution images. The Stage-II GAN takes Stage-I results and text descriptions as inputs, and generates high-resolution images with photo-realistic details. It is able to rectify defects in Stage-I results and add compelling details with the refinement process. To improve the diversity of the synthesized images and stabilize the training of the conditional-GAN, we introduce a novel Conditioning Augmentation technique that encourages smoothness in the latent conditioning manifold. Extensive experiments and comparisons with state-of-the-arts on benchmark datasets demonstrate that the proposed method achieves significant improvements on generating photo-realistic images conditioned on text descriptions.

### 7. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks

*Jun-Yan Zhu∗, Taesung Park∗, Phillip Isola, Alexei A. Efros*

*Abstract* – Image-to-image translation is a class of vision and graphics problems where the goal is to learn the mapping between an input image and an output image using a training set of aligned image pairs. However, for many tasks, paired training data will not be available. We present an approach for learning to translate an image from a source domain X to a target domain Y in the absence of paired examples. Our goal is to learn a mapping $G : X \rightarrow Y$ such that the distribution of images from $G(X)$ is indistinguishable from the distribution Y using an adversarial loss. Because this mapping is highly under-constrained, we couple it with an inverse mapping $F : Y \rightarrow X$ and introduce a cycle consistency loss to enforce $F(G(X)) \approx X$ (and vice versa). Qualitative results are presented on several tasks where paired training data does not exist, including collection style transfer, object transfiguration, season transfer, photo enhancement, etc. Quantitative comparisons against several prior methods demonstrate the superiority of our approach.

### 8. Photographic Image Synthesis with Cascaded Refinement Networks

*Qifeng Chen †, Vladlen Koltun †*

*Abstract* – We present an approach to synthesizing photographic images conditioned on semantic layouts. Given a semantic label map, our approach produces an image with photographic ppearance that conforms to the input layout. The approach thus functions as a rendering engine that takes a two-dimensional semantic specification of the scene and produces a corresponding photographic image. Unlike recent and contemporaneous work, our approach does not rely on adversarial training. We show that photographic images can be synthesized from semantic layouts by a single feed-forward network with appropriate structure, trained end-to-end with a direct regression objective. The presented approach scales seamlessly to high resolutions; we demonstrate this by synthesizing photographic images at 2-megapixel resolution, the full resolution of our training data. Extensive perceptual experiments on datasets of out-door and indoor scenes demonstrate that images synthesized by the presented approach are considerably more realistic than alternative approaches.

### 9. Image-to-Image Translation with Conditional Adversarial Networks

*Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, Alexei A. Efros*

*Abstract* – We investigate conditional adversarial networks as a general-purpose solution to image-to-image translation problems. These networks not only learn the mapping from input image to output image, but also learn a loss function to train this mapping. This makes it possible to apply the same generic approach to problems that traditionally would require very different loss formulations. We demonstrate that this approach is effective at synthesizing photos from label maps, reconstructing objects from edge maps, and colorizing images, among other tasks. Moreover, since the release of the pix2pix software associated with this paper, hundreds of twitter users have posted their own artistic experiments using our system. As a community, we no longer hand-engineer our mapping functions, and this work suggests we can achieve reasonable results without hand-engineering our loss functions either.

### 10. Generative Adversarial Text to Image Synthesis

*Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran 1, Bernt Schiele, Honglak Lee*

*Abstract* – Automatic synthesis of realistic images from text would be interesting and useful, but current AI systems are still far from this goal. However, in recent years generic and powerful recurrent neural network architectures have been developed to learn discriminative text feature representations. Meanwhile, deep convolutional generative adversarial networks (GANs) have begun to generate highly compelling images of specific categories, such as faces, album covers, and room interiors. In this work, we develop a novel deep architecture and GAN formulation to effectively bridge these advances in text and image modeling, translating visual concepts from characters to pixels. We demonstrate the capability of our model to generate plausible images of birds and flowers from detailed text descriptions.

### 11. Learning Deep Representations of Fine-Grained Visual Descriptions

*Scott Reed 1, Zeynep Akata 2, Honglak Lee 1 and Bernt Schiele 2*

*Abstract* – State-of-the-art methods for zero-shot visual recognition formulate learning as a joint embedding problem of images and side information. In these formulations the current best complement to visual features are attributes: manually-encoded vectors describing shared characteristics among categories. Despite good performance, attributes have limitations: (1) finer-grained recognition requires commensurately more attributes, and (2) attributes do not provide a natural language interface. We propose to overcome these limitations by training neural language models from scratch; i.e. without pre-training and only consuming words and characters. Our proposed models train end-to-end to align with the fine-grained and category-specific content of images. Natural language provides a flexible and compact way of encoding only the salient visual aspects for distinguishing categories. By training on raw text, our model can do inference on raw text as well, providing humans a familiar mode both for annotation and retrieval. Our model achieves strong performance on zero-shot text-based image retrieval and significantly outperforms the attribute-based state-of-the-art for zero-shot classification on the Caltech-UCSD Birds 200-2011 dataset.

**12. Generative Adversarial Nets**

*Ian J. Goodfellow, Jean Pouget-Abadie∗, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair†, Aaron Courville, Yoshua Bengio‡*

*Abstract* – We propose a new framework for estimating generative models via an adversarial process, in which we simultaneously train two models: a generative model G that captures the data distribution, and a discriminative model D that estimates the probability that a sample came from the training data rather than G. The training procedure for G is to maximize the probability of D making a mistake. This framework corresponds to a minimax two-player game. In the space of arbitrary functions G and D, a unique solution exists, with G recovering the training data distribution and D equal to ½ everywhere. In the case where G and D are defined by multilayer perceptrons, the entire system can be trained with backpropagation. There is no need for any Markov chains or unrolled approximate inference networks during either training or generation of samples. Experiments demonstrate the potential of the framework through qualitative and quantitative evaluation of the generated samples.

**13. Stacked Generative Adversarial Networks**

*Xun Huang 1, Yixuan Li 2, Omid Poursaeed 2, John Hopcroft 1, Serge Belongie 1, 3*

*Abstract* – In this paper, we propose a novel generative model named Stacked Generative Adversarial Networks (SGAN), which is trained to invert the hierarchical representations of a bottom-up discriminative network. Our model consists of a top-down stack of GANs, each learned to generate lower-level representations conditioned on higher-level representations. A representation discriminator is introduced at each feature hierarchy to encourage the representation manifold of the generator to align with that of the bottom-up discriminative network, leveraging the powerful discriminative representations to guide the generative model. In addition, we introduce a conditional loss that encourages the use of conditional information from the layer above, and a novel entropy loss that maximizes a variational lower bound on the conditional entropy of generator outputs. We first train each stack independently, and then train the whole model end-to-end. Unlike the original GAN that uses a single noise vector to represent all the variations, our SGAN decomposes variations into multiple levels and gradually resolves uncertainties in the top-down generative process. Based on visual inspection, Inception scores and visual Turing test, we demonstrate that SGAN is able to generate images of much higher quality than GANs without stacking.

**14. StoryGAN: A Sequential Conditional GAN for Story Visualization**

*Yitong Li∗1, Zhe Gan 2, Yelong Shen 4, Jingjing Liu 2, Yu Cheng 2, Yuexin Wu 5, Lawrence Carin 1, David Carlson 1 and Jianfeng Gao 3*

*Abstract* – In this work we propose a new task called Story Visualization. Given a multi-sentence paragraph, the story is visualized by generating a sequence of images, one for each sentence. In contrast to video generation, story visualization focuses less on the continuity in generated images (frames), but more on the global consistency across dynamic scenes and characters – a challenge that has not been addressed by any single-image or video generation methods. Therefore, we propose a new story-to-image-sequence generation model, StoryGAN, based on the sequential conditional GAN framework. Our model is unique in that it consists of a deep Context Encoder that dynamically tracks the story flow, and two discriminators at the story and image levels, respectively, to enhance the image quality and the consistency of the generated sequences. To evaluate the model, we modified existing datasets to create the CLEVR-SV and Pororo-SV datasets. Empirically, StoryGAN outperformed state-of-the-art models in image quality, contextual consistency metrics, and human evaluation.

### 15. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks

*Alec Radford & Luke Metz:* ***indico Research;*** *Soumith Chintala****: Facebook AI Research***

*Abstract* – In recent years, supervised learning with convolutional networks (CNNs) has seen huge adoption in computer vision applications. Comparatively, unsupervised learning with CNNs has received less attention. In this work we hope to help bridge the gap between the success of CNNs for supervised learning and unsupervised learning. We introduce a class of CNNs called deep convolutional generative adversarial networks (DCGANs), that have certain architectural constraints, and demonstrate that they are a strong candidate for unsupervised learning. Training on various image datasets, we show convincing evidence that our deep convolutional adversarial pair learns a hierarchy of representations from object parts to scenes in both the generator and discriminator. Additionally, we use the learned features for novel tasks - demonstrating their applicability as general image representations.

### 16. BEGAN: Boundary Equilibrium Generative Adversarial Networks

*David Berthelot, Thomas Schumm, Luke Metz:* ***Google***

*Abstract* – We propose a new equilibrium enforcing method paired with a loss derived from the Wasserstein distance for training auto-encoder based Generative Adversarial Networks. This method balances the generator and discriminator during training. Additionally, it provides a new approximate convergence measure, fast and stable training and high visual quality. We also derive a way of controlling the trade-off between image diversity and visual quality. We focus on the image generation task, setting a new milestone in visual quality, even at higher resolutions. This is achieved while using a relatively simple model architecture and a standard training procedure.

### 17. LR-GAN: Layered recursive generative adversarial networks for image generation

*Jianwei Yang*, Anitha Kannan, Dhruv Batra∗ and Devi Parikh∗*

*Abstract* – We present LR-GAN: an adversarial image generation model which takes scene structure and context into account. Unlike previous generative adversarial networks (GANs), the proposed GAN learns to generate image background and foregrounds separately and recursively, and stitch the foregrounds on the background in a contextually relevant manner to produce a complete natural image. For each foreground, the model learns to generate its appearance, shape and pose. The whole model is unsupervised, and is trained in an end-to-end manner with gradient descent methods. The experiments demonstrate that LR-GAN can generate more natural images with objects that are more human recognizable than DCGAN. The code is available at https://github.com/jwyang/lr-gan.pytorch.

## 18. CanvasGAN: A simple baseline for text to image generation by incrementally patching a canvas

*Amanpreet Singh and Sharan Agrawal:* ***New York University***

*Abstract* – We propose a new recurrent generative model for generating images from text captions while attending on specific parts of text captions. Our model creates images by incrementally adding patches on a "canvas" while attending on words from text caption at each timestep. Finally, the canvas is passed through an upscaling network to generate images. We also introduce a new method for generating visual-semantic sentence embeddings based on self-attention over text. We compare our model's generated images with those generated Reed et al. [25]'s model and show that our model is a stronger baseline for text to image generation tasks. **Keywords:** *image generation, GAN, conditional generation*

## 19. Improved Techniques for Training GANs

*Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, Xi Chen*

*Abstract* – We present a variety of new architectural features and training procedures that we apply to the generative adversarial networks (GANs) framework. We focus on two applications of GANs: semi-supervised learning, and the generation of images that humans find visually realistic. Unlike most work on generative models, our primary goal is not to train a model that assigns high likelihood to test data, nor do we require the model to be able to learn well without using any labels. Using our new techniques, we achieve state-of-the-art results in semi-supervised classification on MNIST, CIFAR-10 and SVHN. The generated images are of high quality as confirmed by a visual Turing test: our model generates MNIST samples that humans cannot distinguish from real data, and CIFAR-10 samples that yield a human error rate of 21.3%. We also present ImageNet samples with unprecedented resolution and show that our methods enable the model to learn recognizable features of ImageNet classes.

## 20. Sequential Attention GAN for Interactive Image Editing via Dialogue

*Yu Cheng 1, Zhe Gan 1, Yitong Li 2, Jingjing Liu 1, Jianfeng Gao 3:* ***Microsoft***

*Abstract* – In this paper, we introduce a new task - interactive image editing via conversational language, where users can guide an agent to edit images via multi-turn dialogue in natural language. In each dialogue turn, the agent takes a source image and a natural language description from the user as the input, and generates a target image following the textual description. Two new datasets are created for this task, Zap-Seq and DeepFashion-Seq, collected via crowdsourcing. For this task, we propose a new Sequential Attention Genrative Adversarial Network (SeqAttnGAN) framework, which applies a neural state tracker to encode both source image and textual descriptions, and generates high quality images in each dialogue turn. To achieve better region specific text-to-image generation, we also introduce an attention mechanism into the model. Experiments on the two datasets, including quantitative evaluation and user study, show that our model outperforms state-of-the-art approaches in both image quality and text-to-image consistency.

### 21. Language-Based Image Editing with Recurrent Attentive Models

*Jianbo Chen∗, Yelong Shen†, Jianfeng Gao†, Jingjing Liu†, Xiaodong Liu†*

*Abstract* – We investigate the problem of Language-Based Image Editing (LBIE). Given a source image and a natural language description, we want to generate a target image by editing the source image based on the description. We propose a generic modeling framework for two sub-tasks of LBIE: language-based image segmentation and image colorization. The framework uses recurrent attentive models to fuse image and language features. Instead of using a fixed step size, we introduce for each region of the image a termination gate to dynamically determine after each inference step whether to continue extrapolating additional information from the textual description. The effectiveness of the framework is validated on three datasets. First, we introduce a synthetic dataset, called CoSaL, to evaluate the end-to-end performance of our LBIE system. Second, we show that the framework leads to state-of-the-art performance on image segmentation on the ReferIt dataset. Third, we present the first language-based colorization result on the Oxford-102 Flowers dataset.

### 22. Conversational Image Editing: Incremental Intent Identification in a New Dialogue Task

*Ramesh Manuvinakurike 1, Trung Bui 2, Walter Chang 2, Kallirroi Georgila 1:* ***Adobe***

*Abstract* – We present "conversational image editing", a novel real-world application domain combining dialogue, visual information, and the use of computer vision. We discuss the importance of dialogue incrementality in this task, and build various models for incremental intent identification based on deep learning and traditional classification algorithms. We show how our model based on convolutional neural networks outperforms models based on random forests, long short term memory networks, and conditional random fields. By training embeddings based on image-related dialogue corpora, we outperform pre-trained out-of-the-box embeddings, for intention identification tasks. Our experiments also provide evidence that incremental intent processing may be more efficient for the user and could save time in accomplishing tasks.

### 23. Generative Image Modeling using Style and Structure Adversarial Networks

*Xiaolong Wang, Abhinav Gupta*

*Abstract* – Current generative frameworks use end-to-end learning and generate images by sampling from uniform noise distribution. However, these approaches ignore the most basic principle of image formation: images are product of: (a) Structure: the underlying 3D model; (b) Style: the texture mapped onto structure. In this paper, we factorize the image generation process and propose Style and Structure Generative Adversarial Network (S2-GAN). Our S2-GAN has two components: the Structure-GAN generates a surface normal map; the Style-GAN takes the surface normal map as input and generates the 2D image. Apart from a real vs. generated loss function, we use an additional loss with computed surface normals from generated images. The two GANs are first trained independently, and then merged together via joint learning. We show our S2-GAN model is interpretable, generates more realistic images and can be used to learn unsupervised RGBD representations.

### 24. Generative Multi-Adversarial Networks

*Ishan Durugkar∗, Ian Gemp∗, Sridhar Mahadevan*

*Abstract* – Generative adversarial networks (GANs) are a framework for producing a generative model by way of a two-player minimax game. In this paper, we propose the Generative Multi-Adversarial Network (GMAN), a framework that extends GANs to multiple discriminators. In previous work, the successful training of GANs requires modifying the minimax objective to accelerate training early on. In contrast, GMAN can be reliably trained with the original, untampered objective. We explore a number of design perspectives with the discriminator role ranging from formidable adversary to forgiving teacher. Image generation tasks comparing the proposed framework to standard GANs demonstrate GMAN produces higher quality samples in a fraction of the iterations when measured by a pairwise GAM-type metric.

### 25. Deep Generative Image Models using a Laplacian Pyramid of Adversarial Networks

*Emily Denton∗, Soumith Chintala∗, Arthur Szlam, Rob Fergus*

*Abstract* – In this paper we introduce a generative parametric model capable of producing high quality samples of natural images. Our approach uses a cascade of convolutional networks within a Laplacian pyramid framework to generate images in a coarse-to-fine fashion. At each level of the pyramid, a separate generative con-vnet model is trained using the Generative Adversarial Nets (GAN) approach [12]. Samples drawn from our model are of significantly higher quality than alternate approaches. In a quantitative assessment by human evaluators, our CIFAR10 samples were mistaken for real images around 40% of the time, compared to 10% for samples drawn from a GAN baseline model. We also show samples from models trained on the higher resolution images of the LSUN scene dataset.

### Mask R-CNN

*Kaiming He, Georgia Gkioxari, Piotr Doll´ar, Ross Girshick: **Facebook AI Research (FAIR)***

*Abstract* – We present a conceptually simple, flexible, and general framework for object instance segmentation. Our approach efficiently detects objects in an image while simultaneously generating a high-quality segmentation mask for each instance. The method, called Mask R-CNN, extends Faster R-CNN by adding a branch for predicting an object mask in parallel with the existing branch for bounding box recognition. Mask R-CNN is simple to train and adds only a small overhead to Faster R-CNN, running at 5 fps. Moreover, Mask R-CNN is easy to generalize to other tasks, e.g., allowing us to estimate human poses in the same framework. We show top results in all three tracks of the COCO suite of challenges, including instance segmentation, bounding-box object detection, and person keypoint detection. Without bells and whistles, Mask R-CNN outperforms all existing, single-model entries on every task, including the COCO 2016 challenge winners. We hope our simple and effective approach will serve as a solid baseline and help ease future research in instance-level recognition. Code has been made available at:

https://github.com/facebookresearch/Detectron.