# Learning to Synthesize Realistic Fine-grained Images by Object Parts

**Zixuan Huang**
Department of Computer Science
University of Wisconsin-Madison
NetID: zhuang356
zhuang356@wisc.edu

**Yufei Wang**
Department of Computer Science
University of Wisconsin-Madison
NetID: ywang2395
ywang2395@wisc.edu

**Yang Fang**
Department of Electrical & Computer Engineering
University of Wisconsin-Madison
NetID: fang65
fang65@wisc.edu

**Briefly explain what problem you are trying to solve.**

Learning controllable and interpretable generative model has been one of the main pursuit of image synthesis, yet the controllability and interpretability can happen at multiple faucets. Consider the examples in Fig. 1. If we want to generate these images, what is a reasonable way to generate them in a controllable fashion? We believe at least one possible solution is to synthesize these images by object parts (e.g., the head, neck, belly, legs, wings of the birds and the background). Such an approach not only provides an explicit way to synthesize fine-grained details according to each specific object part, but also brings a controllability over the synthesis procedure via independently manipulating different parts of generated images.

On the other hand, Generative Adversarial Network (GAN) [1] are tremendously successful for image generation. Current state-of-the-art methods along this direction are already able to generate photo-realistic images. By combining our idea with GANs, we hope the resulting method is able to generate both photo-realistic and controllable image by object parts.

Figure 1: Four randomly sampled images from CUB-200 [2]. Assume they are synthesized images, what is a reasonable way to generate them in a controllable fashion?

**Why is this problem important? Why are you interested in it?**

Image generation has been one of the fundamental tasks in computer vision. We believe the importance of this subarea lies in two aspects. First of all, the generated images have various application. For instance, artist can rely on these generated photo-realistic images to create art. More importantly, by generating images, we can expand existing datasets for other vision or machine learning tasks, especially unsupervised or few-shot learning.

On the other hand, the generative model itself is important, because it provides an effective intermediate representation of images. Such representation can be especially useful for other discriminative tasks, such as unsupervised image classification [3]. Meanwhile, GAN has also demonstrated an immense success in other computer vision tasks including image super-resolution [4], text to image translation [5], image inpainting [6], etc. Overall, we find it to be a quite important and interesting area to explore.

**What is the current state-of-the-art?**

To the best of our knowledge, we do not think there are any papers sharing exactly the same idea of composing natural images via part representation with GAN. However, there exists many relevant works focusing on disentangling the generation process or providing interpretable results. For instance, Info-GAN proposed to learn more interpretable and controllable models by maximizing the lower bound of the mutual information between the latent code and the generated images [7]. However, they did not model object parts explicitly, nor demonstrate any experimental results over relatively challenging datasets. Another closely related one is FineGAN [3], which proposed to model an object as a combination of background, foreground shape and foreground appearance. They also disentangled different factors by maximizing the mutual information similarly to Info-GAN. LR-GAN [8] is also highly relavant, where they composed the image by recursively synthesizing and stitching background, foreground shape and transformation.

**Are you planning on re-implementing an existing solution, or propose a new approach?**

To begin with, we will re-implement W-GAN [9] over CUB-200 as our baseline. After experimenting on this baseline, we will endeavor to improve it with our own approach.

**If you are proposing your own approach, why do you think existing approaches cannot adequately solve this problem? Why do you think you solution will work better?**

Part-based visual representation stems from the Gestalt visual perception theory and the perceptual grouping idea [10, 11]. It is believed to be a both intuitive and effective representation of images. We think if we succeed on embedding it into powerful generative models like GAN, the generation process will be more transparent and controllable compared to existing approaches. However, we admit this is a risky project. Therefore, it is possible that our proposed method cannot finally outperform the existing methods. But succeed or not, we hope our experimental results can shed some light on future research along this direction.

**How will you evaluate the performance of your solution? What results and comparisons are you eventually planning to show? Include a time-line that you would like to follow.**

We will first evaluate the performance of W-GAN quantitatively and qualitatively on CUB-200 datasets as our baseline. After implementing our own approach, we will compare it with the baseline. We plan to conduct the qualitative evaluation by showing sample outputs, and demonstrate quantitative results through commonly used Inception Score [12] or Frechet Inception Distance [13].

We will roughly follow the time-line below:

2.15 - 3.1: Finish an overall literature review and design concrete approaches to improve the baseline.
3.1 - 3.15: Implement the baseline on CUB-200 and evaluate the baseline.
3.15 - 4.15: Implement our approach and compare with the baseline. If the results are great, also compare with other state-of-the-art methods.
4.15 - 4.29: Wrap up everything and prepare for the final presentation.

## References


[1] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[2] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

[3] Krishna Kumar Singh, Utkarsh Ojha, and Yong Jae Lee. Finegan: Unsupervised hierarchical disentanglement for fine-grained object generation and discovery. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6490–6499, 2019.

[4] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017.

[5] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N. Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[6] Raymond A Yeh, Chen Chen, Teck Yian Lim, Alexander G Schwing, Mark Hasegawa-Johnson, and Minh N Do. Semantic image inpainting with deep generative models. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5485–5493, 2017.

[7] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in neural information processing systems*, pages 2172–2180, 2016.

[8] Jianwei Yang, Anitha Kannan, Dhruv Batra, and Devi Parikh. Lr-gan: Layered recursive generative adversarial networks for image generation. *arXiv preprint arXiv:1703.01560*, 2017.

[9] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017.

[10] Kurt Koffka. Perception: an introduction to the gestalt-theorie. *Psychological Bulletin*, 19(10):531, 1922.

[11] Max Wertheimer. Gestalt theory. 1938.

[12] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Advances in neural information processing systems*, pages 2234–2242, 2016.

[13] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in neural information processing systems*, pages 6626–6637, 2017.