Semantic Hierarchy Emerges in Deep Generative Representations for Scene Synthesis

Ceyuan Yang, Yujun Shen, Bolei Zhou
The Chinese University of Hong Kong
{yc019,sy116,bzhou}@ie.cuhk.edu.hk

Figure 1: Manipulation results from four different abstraction levels, including *layout*, *categorical objects*, *scene attributes*, and *color scheme*. For each image tuple, the first is the original synthesis and the followings are the results with some degree of manipulation.

Abstract

Although Generative Adversarial Networks (GANs) have significantly advanced image synthesis, it remains unknown how photo-realistic images are able to be composed of the layer-wise stochasticity introduced in recent GANs. In this work, we show that highly-structured semantic hierarchy emerges as variation factors for synthesizing scenes. By probing the layer-wise representations with a broad set of semantics at different abstraction levels, we are able to quantify the causality between the input activation and the semantics occurring in the output image. Such a quantification identifies the human-understandable variation factors learned by GANs to compose scenes. We find that the layer-wise latent codes used in GANs are specialized to synthesize hierarchical semantics: the early layers determine the spatial layout and configuration, the middle layers control the categorical objects, and the later layers render the scene attributes as well as color scheme. Identifying such a set of manipulable latent variation factors facilitates semantic scene manipulation.

1. Introduction

Success of deep neural networks stems from the representation learning, which identifies the explanatory factors underlying the high-dimensional observed data [4]. But current efforts on interpreting deep representations mainly focus on discriminative models [30, 9, 28, 1, 2]. Recent advance of Generative Adversarial Networks (GANs) [10, 14] is capable of transforming random noises into high-quality images and layer-wise latent codes are introduced into GANs [15, 5] to control the synthesis from coarse to fine. However, how these variation factors are composed together and how to quantify such semantic information remain unknown. In this work, we deeply study the hierarchical generative representations and find that GANs actually learn human-understandable scene variations at multiple abstraction levels, including *layout*, *categorical object*, *attribute*, and *color scheme*. A re-scoring technique is proposed to quantitatively identity these latent semantics, which can be further used to precisely control the generation process, as shown in Fig.1. Codes are available at https://genforce.github.io/higan/.

Figure 2: Method for identifying the emergent variation factors in generative representation. By deploying a broad set of *off-the-shelf* image classifiers as scoring functions, $F(\cdot)$, we are able to assign a synthesized image with semantic scores associated with each candidate variation factor. For a particular concept, we learn a decision boundary in the latent space by considering it as a binary classification task. Then we move the sampled latent code towards the boundary to see how the semantic score varies correspondingly, and use a re-scoring technique to quantitatively verify the emergence of the target concept.

1.1. Related Work

Deep Representations for Image Synthesis. Generative Adversarial Networks (GANs) [10] advance the image synthesis significantly. Some recent models [14, 5, 15] are able to generate photo-realistic faces, objects, and scenes, making GANs applicable to real-world image editing tasks [22, 24, 23, 26, 3, 19, 32, 7]. Despite such a great success, it remains uncertain what GANs have actually learned to produce such diverse and realistic images. Bau *et al.* [3] analyzed the individual units of the generator in a GAN and some concurrent work [12, 8, 21] interpret the latent semantics learned by GANs. Unlike them, our work *quantitatively* identifies the emergence of *multi-level* semantics inside the layer-wise generative representations.

Semantic Scene Editing. Laffont *et al.* [16] pre-defined 40 transient attributes and managed to transfer the appearance across scenes. Cheng *et al.* [6] proposed verbal guided image parsing to recognize and manipulate the objects in indoor scenes. Karacan *et al.* [13] learned a conditional GAN to synthesize outdoor scenes based on pre-defined layout and attributes. Some other work [17, 32, 11, 18] studied image-to-image translation and can be used to transfer the style of one scene to another. Different from them, we achieve scene manipulation by identifying the hierarchical semantics emerging from the generative representations of *fixed* GANs. We can also precisely control the editing from different abstraction levels, including *layout*, *categorical object*, *attribute*, and *color scheme*.

2. Semantic Hierarchy in Scene Representation

In this section, we introduce the semantic hierarchy in the deep generative representations for scene synthesis.

Multi-level Semantics. Imagine an artist drawing a picture of the living room. The very first step is to choose a perspective and set up the room layout. After the spatial structure is decided, the next step is to add objects that typically occur in a living room, such as a sofa and TV. Finally, the artist will refine the details of the picture with specified decoration styles, *e.g.*, warm or cold, natural lighting or indoor lighting. The above process reflects how a human interprets a scene to draw it. As a comparison, generative models such as GANs follow a completely end-to-end training for synthesizing scenes, without any prior knowledge about the drawing techniques and relevant concepts. This work explores how the semantics learned by GANs align with the human-understandable variation factors. We surprisingly find that GAN synthesizes a scene in a manner highly consistent with the human. Over the convolutional layers, GAN manages to compose these multi-level abstractions hierarchically. In particular, GAN constructs the spatial layout at the early stage, synthesizes category-specified objects at the middle stage, and renders the scene attributes and color scheme at the later stage.

Identifying the Emergent Variation Factors. We use some off-the-shelf classifiers to extract multi-level semantics (*e.g.*, layout and scene attributes) from the synthesized image as candidates. We then propose a re-scoring technique to quantitatively identify the emerging variation factors that are most relevant to the scene synthesis model. Fig.2 illustrates the identification process which consists of two steps, *i.e.*, probing and verification. In the probing phase, we construct a training set with randomly sampled latent codes (data) and the corresponding semantic scores (label). Then we train a linear boundary in the latent space for each candidate by solving a bi-classification task. In the verification phase, we re-sample some latent codes and move these codes towards the latent boundary to see how the semantic scores change accordingly. Hence, we have

$$\Delta s_i = \frac{1}{K}\sum_{k=1}^{K}\Big(F_i\big(G(\mathbf{z}^k+\lambda\mathbf{n}_i)\big) - F_i\big(G(\mathbf{z}^k)\big)\Big)^+, \quad (1)$$

where $\frac{1}{K}\sum_{k=1}^{K}$ stands for the average of K samples to make the metric more accurate. $(\cdot)^+$ denotes the zero-truncation operator and $\lambda = 2$ is a fixed moving step. Then we can easily rank the score Δs_i among all candidates to retrieve the most relevant latent variation factors.

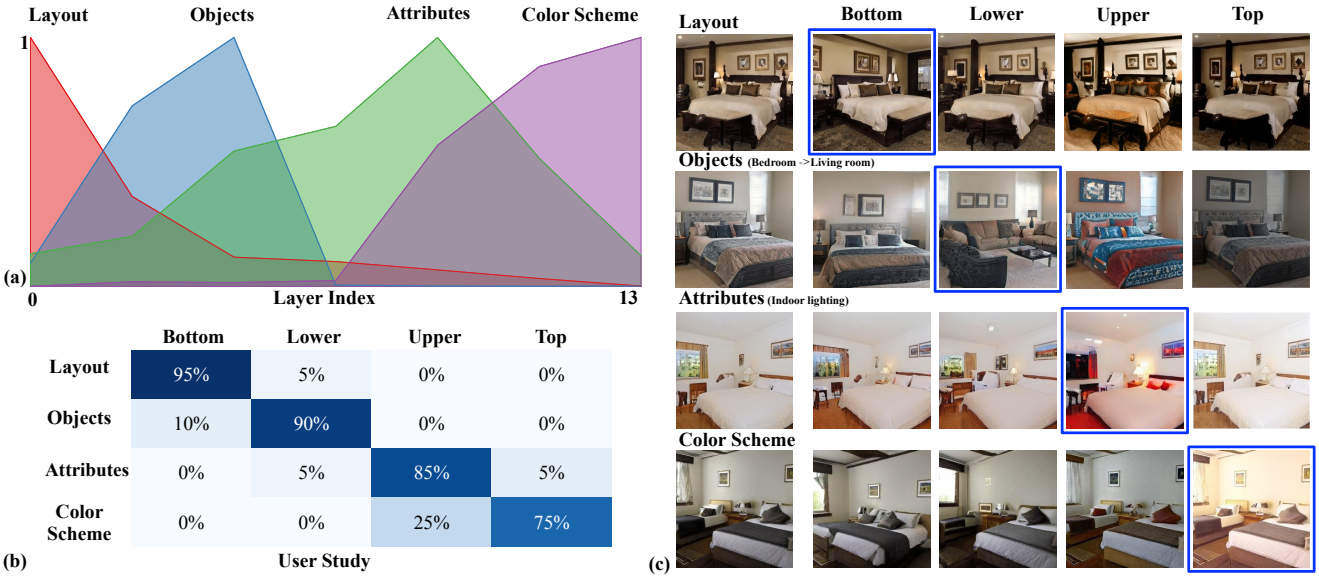

Figure 3: (a) Four levels of visual abstractions emerge at different layers of StyleGAN. Vertical axis shows the normalized Δs_i. (b) User study on how different layers correspond to variation factors from different abstraction levels. (c) Layer-wise manipulation result. The first column is the original synthesized images, and the other columns are the manipulated images at layers from four different stages respectively. Blue boxes highlight the results from varying the latent code at the most proper layers for the target concept.

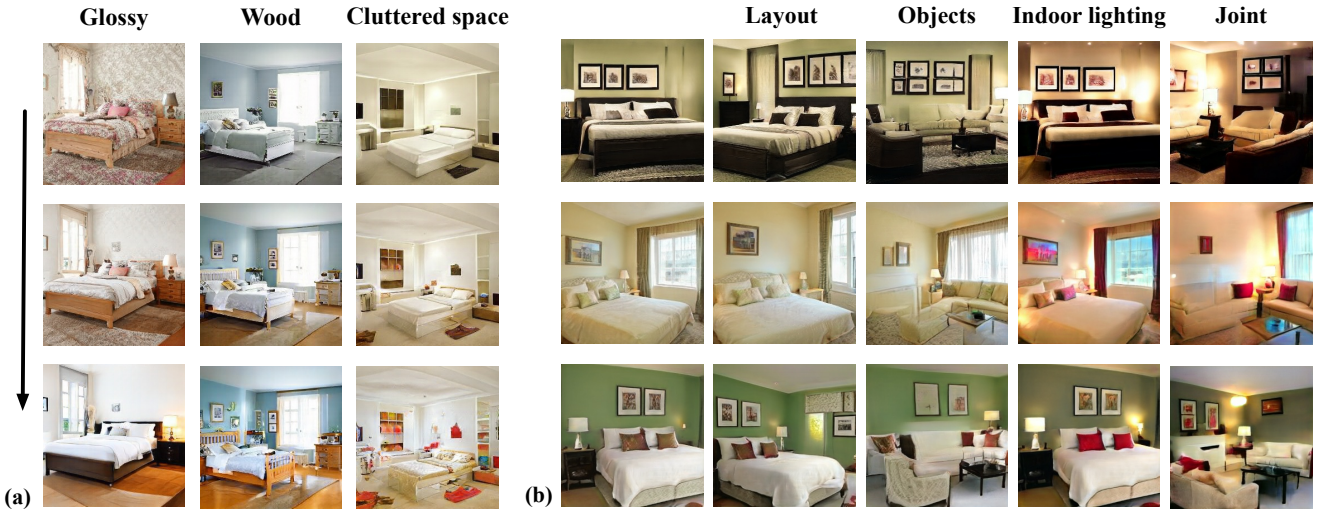

Figure 4: (a) Independent attribute manipulation results. The middle row is the source images. We are able to both decrease (top row) and increase (bottom row) the variation factors in the images. (b) Joint manipulation results, where the *layout* is manipulated at the early layers, the *categorical objects* are manipulated at the middle layers, while indoor lighting *attribute* is manipulated at the later layers. The first column indicates the source images and the middle three columns are the independently edited images.

3. Experiments

We do experiments on a StyleGAN [15] model trained with LSUN [27] dataset. We use a layout estimator ([29]), a scene category recognizer ([31]), and an attribute classifier ([31]) trained on SUN attribute database ([20]) as the *off-the-shelf* predictors.

3.1. Emerging Semantic Hierarchy

We apply the proposed re-scoring technique on the layer-wise latent codes used in StyleGAN. Fig.3 (a) shows that the layers of the generator in GAN are specialized to compose semantics in a hierarchical manner: the bottom layers determine the layout, the lower layers and upper layers control category-level and attribute-level variations respectively, while color scheme is mostly rendered at the top layers. This is consistent with human perception. To visually inspect the identified variation factors, we move the latent vector towards the semantic boundaries at different layers to show how the synthesis varies correspondingly. We conduct user study on the manipulated images to

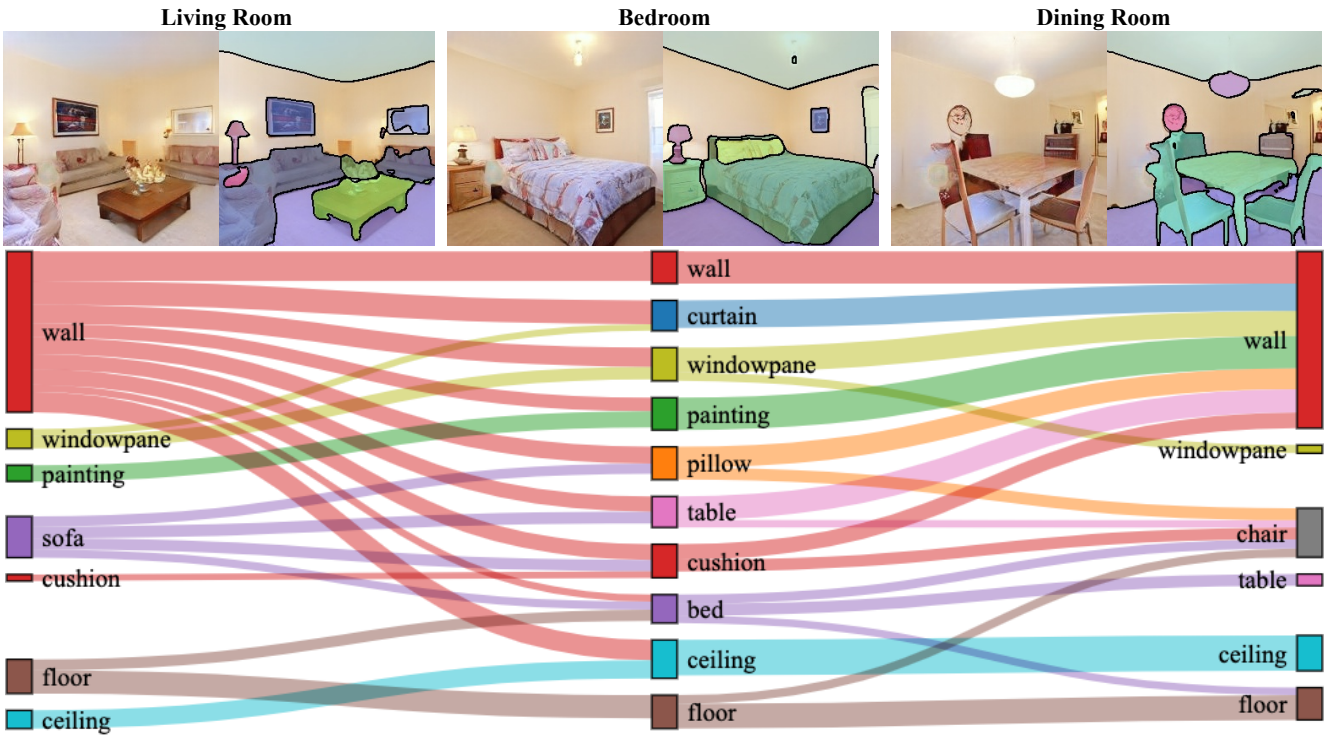

Figure 5: Objects are transformed by GAN to represent different scene categories. On the top shows that the object segmentation mask varies when manipulating a living room to bedroom, and further to dining room. On the bottom visualizes the object mapping that appears during category transition, where pixels are counted only from object level instead of instance level. GAN is able to learn shared objects as well as the transformation of objects with similar appearance when trained to synthesize scene images from more than one category.

validate our discovery about the hierarchical semantics emerging in the generative representations. The results are shown in Fig.3 (b), where we can get same conclusion as from Fig.3 (a) Fig.3 (c) shows some qualitative results from various abstraction levels.

3.2. Semantic Scene Editing

Identifying the hierarchical variation factors across different layers significantly facilitates scene manipulation. We can push the latent code towards the boundary of the desired attribute at the most *appropriate* layer. Fig.4 (a) shows that we can change the decoration style (crude to glossy), the material of furniture (cloth to wood), or even the cleanliness (tidy to cluttered). Furthermore, we can jointly manipulate these multi-level semantics. In Fig.4 (b), we simultaneously change the room layout (rotating viewpoint) at early layers, the scene category (converting bedroom to living room) at middle layers, and the scene attribute (increasing indoor lighting) at later layers.

3.3. Categorical Analysis

From all four abstraction levels (*i.e.*, layout, object (category), scene attribute, and color scheme), one of the most noticeable things is that GANs are able to transfer the scene from one category to another by changing the objects contained in the image. To deeply study the object mapping in this transferring process, we employ a semantic segmentation model ([25]), which can segment 150 objects (tv, sofa, *etc*) and stuff (ceiling, floor, *etc*). From Fig.5, we can see that (1) When an image is manipulated among different categories, most of the stuff classes (*e.g.*, ceiling and floor) remain the same, but the categorical objects may be mapped from one to another. For example, the sofa (living room) is mapped to a bed (bedroom) and further mapped to a table (dinning room). (2) Some objects are shared between different scene categories. For example, the lamp in living room (on the left edge of the image) remains after being converted to bedroom. (3) With the capacity of learning the object mapping as well as sharing objects across different categories, we are able to turn an unconditional GAN into a GAN that can control category.

4. Conclusion

In this paper, we propose a re-scoring method to quantitatively study the emergence of highly-structured variation factors inside the deep generative representations learned by GANs with layer-wise stochasticity. In particular, GANs spontaneously learn to set up the layout at early layers, generate categorical objects at middle layers, and render scene attributes and color scheme at later layers. Such semantic hierarchy enables photo-realistic scene manipulation.

References

[1] Pulkit Agrawal, Ross Girshick, and Jitendra Malik. Analyzing the performance of multilayer neural networks for object recognition. In *ECCV*, 2014. 1

[2] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *CVPR*, 2017. 1

[3] David Bau, Jun-Yan Zhu, Hendrik Strobelt, Bolei Zhou, Joshua B. Tenenbaum, William T. Freeman, and Antonio Torralba. Gan dissection: Visualizing and understanding generative adversarial networks. In *ICLR*, 2019. 2

[4] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *TPAMI*, 2013. 1

[5] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. In *ICLR*, 2019. 1, 2

[6] Ming-Ming Cheng, Shuai Zheng, Wen-Yan Lin, Vibhav Vineet, Paul Sturgess, Nigel Crook, Niloy J Mitra, and Philip Torr. Imagespirit: Verbal guided image parsing. *ACM Trans. on Graphics*, 2014. 2

[7] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *CVPR*, 2018. 2

[8] Lore Goetschalckx, Alex Andonian, Aude Oliva, and Phillip Isola. Ganalyze: Toward visual definitions of cognitive image properties. In *ICCV*, 2019. 2

[9] Abel Gonzalez-Garcia, Davide Modolo, and Vittorio Ferrari. Do semantic parts emerge in convolutional neural networks? *IJCV*, 2018. 1

[10] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014. 1, 2

[11] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017. 2

[12] Ali Jahanian, Lucy Chai, and Phillip Isola. On the"steerability" of generative adversarial networks. In *ICLR*, 2020. 2

[13] Levent Karacan, Zeynep Akata, Aykut Erdem, and Erkut Erdem. Learning to generate images of outdoor scenes from attributes and semantic layouts. *arXiv preprint arXiv:1612.00215*, 2016. 2

[14] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *ICLR*, 2018. 1, 2

[15] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 1, 2, 3

[16] Pierre-Yves Laffont, Zhile Ren, Xiaofeng Tao, Chao Qian, and James Hays. Transient attributes for high-level understanding and editing of outdoor scenes. *ACM Trans. on Graphics*, 2014. 2

[17] Jing Liao, Yuan Yao, Lu Yuan, Gang Hua, and Sing Bing Kang. Visual attribute transfer through deep image analogy. *arXiv preprint arXiv:1705.01088*, 2017. 2

[18] Fujun Luan, Sylvain Paris, Eli Shechtman, and Kavita Bala. Deep photo style transfer. In *CVPR*, 2017. 2

[19] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *CVPR*, 2019. 2

[20] Genevieve Patterson, Chen Xu, Hang Su, and James Hays. The sun attribute database: Beyond categories for deeper scene understanding. *IJCV*, 2014. 3

[21] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing. In *CVPR*, 2020. 2

[22] Yujun Shen, Ping Luo, Junjie Yan, Xiaogang Wang, and Xiaoou Tang. Faceid-gan: Learning a symmetry three-player gan for identity-preserving face synthesis. In *CVPR*, 2018. 2

[23] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *CVPR*, 2018. 2

[24] Taihong Xiao, Jiapeng Hong, and Jinwen Ma. Elegant: Exchanging latent encodings with gan for transferring multiple face attributes. In *ECCV*, 2018. 2

[25] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *ECCV*, 2018. 4

[26] Shunyu Yao, Tzu Ming Hsu, Jun-Yan Zhu, Jiajun Wu, Antonio Torralba, Bill Freeman, and Josh Tenenbaum. 3d-aware scene manipulation via inverse graphics. In *NeurIPS*, 2018. 2

[27] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015. 3

[28] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *ECCV*, 2014. 1

[29] Weidong Zhang, Wei Zhang, and Jason Gu. Edge-semantic learning strategy for layout estimation in indoor environment. In *IEEE Transactions on Cybernetics*, 2019. 3

[30] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Object detectors emerge in deep scene cnns. In *ICLR*, 2015. 1

[31] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *TPAMI*, 2017. 3

[32] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017. 2