

Semantic Hierarchy Emerges in Deep Generative Representations for Scene Synthesis

Ceyuan Yang* · Yujun Shen* · Bolei Zhou

Abstract Despite the success of Generative Adversarial Networks (GANs) in image synthesis, there lacks enough understanding on what generative models have learned inside the deep generative representations and 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 from synthesizing scenes from the generative representations in state-of-the-art GAN models, like StyleGAN and BigGAN. 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 activations and semantics occurring in the output image. Such a quantification identifies the human-understandable variation factors learned by GANs to compose scenes. The qualitative and quantitative results further suggest that the generative representations learned by the GANs with layer-wise latent codes are specialized to synthesize different hierarchical semantics: the early layers tend to determine the spatial layout and configuration, the middle layers control the categorical objects, and the later layers finally render the scene attributes as well as color scheme. Identifying such a set of manipulatable latent variation factors facilitates semantic scene manipulation.¹

Keywords Generative model · Scene understanding · Image manipulation · Representation interpretation · Feature visualization

* denotes equal contribution.

C. Yang, Y. Shen, B. Zhou
`{yc019, sy116, bzhou}@ie.cuhk.edu.hk`
 Department of Information Engineering, The Chinese University of Hong Kong, Hong Kong.

¹ Code and demo video can be found at <https://ceyuan.me/SemanticHierarchyEmerge>.

1 Introduction

Success of deep neural networks stems from the representation learning, which identifies the explanatory factors underlying the high-dimensional observed data [5]. Prior work has shown that many concept detectors spontaneously emerge inside the deep representations trained for the classification task. For example, Gonzalez-Garcia et al [10] shows that networks for object recognition are able to detect semantic object parts, and Bau et al [3] confirms that deep representations from classifying images learn to detect different categorical concepts at different layers.

Analyzing the deep representations and their emergent structures gives insight into the generalization ability of deep features [22] as well as the feature transferability across different tasks [37]. But current efforts on interpreting deep representations mainly focus on discriminative models [41, 10, 39, 1, 3]. Recent advance of Generative Adversarial Networks (GANs) [11, 16, 17, 6] is capable of transforming random noises into high-quality images, however, the nature of the learned generative representations and how a photo-realistic image is being composed over different layers of the generator in GAN remain much less explored.

It is known that the internal units of Convolutional Neural Networks (CNNs) emerge as object detectors when trained to categorize scenes [41]. Representing and detecting objects most informative to a specific category provides an ideal solution for classifying scenes, such as sofa and TV are representative of the living room while bed and lamp are of the bedroom. However, synthesizing a scene requires far more knowledge for the deep generative models to learn. Specifically, in order to draw highly-diverse scene images, like our humans the deep representations might be required to not only learn to generate every individual object relevant to a specific scene category, but also decide the underlying room layout as well as render various scene attributes, e.g., the lighting condition and color scheme. Recent work on

Layout**Attribute: Indoor lighting****Category: objects from bedroom to living room****Color Scheme**

Fig. 1: Manipulation results from four different abstraction levels, including *layout*, *categorical objects*, *scene attributes*, and *color scheme*. For each tuple of images, the first is the original synthesized image, the following are the ones after some degree of manipulation.

interpreting GANs [4] visualized that the internal filters at intermediate layers are specialized for generating some certain objects, but studying scene synthesis from object aspect only is far from fully understanding how GAN is able to compose a photo-realistic image, which contains multiple variation factors from layout level, category level, to attribute level. The original StyleGAN work [17] pointed out that the layer-wise latent codes actually control the synthesis from coarse to fine, but how these variation factors are composed together and how to quantify such semantic information remain unknown. Differently, this work gives a much deeper interpretation on the hierarchical generative representations in the sense that we match these layer-wise variation factors with human-understandable scene variations at multiple abstraction levels, including *layout*, *categorical object*, *attribute*, and *color scheme*. Figure 1 shows the manipulation results at such various levels when the corresponding layers are identified correctly.

Starting with the state-of-the-art StyleGAN models [17] as the example, we reveal that highly-structured semantic hierarchy emerges from the deep generative representations with layer-wise stochasticity trained for synthesizing scenes, even without any external supervision. Layer-wise representations are first probed with a broad set of visual concepts at different abstraction levels. By quantifying the causality between the layer-wise activations and the semantics occurring in the output image, we are able to identify the most relevant variation factors across different layers of a GAN model with layer-wise latent codes: the early layers specify the spatial layout, the middle layers compose the category-guided objects, and the later layers render the attributes and color scheme of the entire scene. We further show that identifying such a set of manipulatable variation factors from layouts, objects, to scene attributes and color schemes facilitates

the semantic image manipulation (as shown in Fig.1) with a large diversity. The proposed manipulation technique is further generalized to other GANs such as BigGAN [6] and ProgressiveGAN [16].

2 Related Work

Deep representations from classifying images. Many attempts have been made to study the internal representations of CNNs trained for classification tasks. Zhou et al [41] analyzed hidden units by simplifying the input image to see which context region gives the highest response, Simonyan et al [31] applied the back-propagation technique to compute the image-specific class saliency map, Bau et al [3] interpreted the hidden representations via the aid of segmentation mask, Alain and Bengio [2] trained independent linear probes to analyze the information separability among different layers. There are also some studies transferring the features of CNNs to verify how learned representations fit with different datasets or tasks [37, 1]. In addition, reversing the feature extraction process by mapping a given representation back to image space [39, 23, 21] also gives insight into what CNNs actually learn to distinguish different categories. However, these interpretation techniques developed for classification networks cannot be directly applied for generative models.

Deep representations from synthesizing images. Generative Adversarial Networks (GANs) [11] advance the image synthesis significantly. Some recent models [16, 6, 17] are able to generate photo-realistic faces, objects, and scenes, making GANs applicable to real-world image editing tasks, such as image manipulation [29, 34, 32, 36], image painting [4, 25], and image style transfer [43, 8]. Despite such a great success, it remains uncertain what GANs have actually learned to produce such diverse and realistic images. Radford

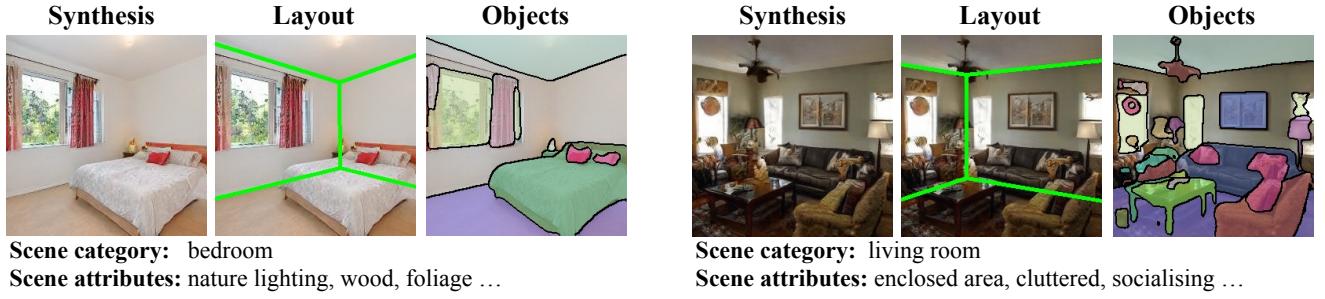


Fig. 2: Multiple levels of semantics extracted from two synthesized scenes.

et al [27] pointed out the vector arithmetic phenomenon in the underlying latent space of GAN, however, discovering what kinds of semantics exist inside a well-trained model and how these semantics are structured to compose high-quality images are still unsolved. A very recent work [4] analyzed the individual units of the generator in GAN and found that they learn to synthesize informative visual contents such as objects and textures spontaneously. Besides, concurrent work [14, 9] also explored the steerability and boosts the memorability of GANs via the learned semantics respectively. Unlike them, our work quantitatively explores the emergence of multi-level semantics inside the layer-wise generative representations.

Scene manipulation and editing. Previous efforts were also made to edit scene images. Laffont et al [18] defined 40 transient attributes and managed to transfer the appearance of a similar scene to the image for editing. Cheng et al [7] proposed verbal guided image parsing to recognize and manipulate the objects in indoor scenes. Karacan et al [15] learned a conditional GAN to synthesize outdoor scenes based on pre-defined layout and attributes. Some other work [19, 43, 13, 20] 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 interpreting the hierarchical semantics emerging from the generative representations of well-trained GANs. Besides image editing, such interpretation also gives us a better insight on how generative models are able to produce photo-realistic synthesis.

3 Variation Factors in Generative Representations

3.1 Multi-Level Variation Factors for Scene Synthesis

Imagine an artist drawing a picture of the living room. The very first step, before drawing every single object, is to choose a perspective and set up the layout of the room. After the spatial structure is set, 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 draws a scene by interpreting it from multiple

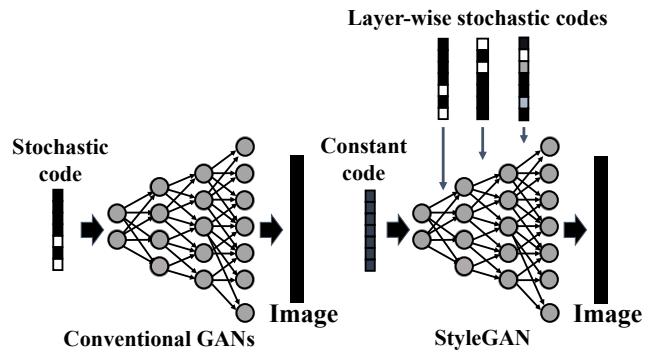


Fig. 3: Comparison between the conventional generator structure where the latent code is only fed into the very first layer and the generator in state-of-the-art GANs (*e.g.*, StyleGAN [17] and BigGAN [6]) which introduce layer-wise stochasticity by feeding latent codes to all convolutional layers.

abstraction levels. Meanwhile, given a scene image, we are able to extract multiple levels of attributes, as shown in Fig.2. As a comparison, generative models such as GANs follow a completely end-to-end training manner for synthesizing scenes, without any prior knowledge about the drawing techniques and relevant concepts. Even so, the trained GANs are able to produce photo-realistic scenes, which makes us wonder if the GANs have mastered any human-understandable drawing knowledge as well as the variation factors of scenes spontaneously.

3.2 Layer-wise Generative Representations

In general, existing generative models take a randomly sampled latent code as input and output a image synthesis as real as possible. Such a one-on-one mapping from latent codes to synthesized images is very similar to the feature extraction process in discrimination model. Accordingly, in this work, we treat the input latent code as the *generative representation* which will uniquely determine the appearance and properties of the output scene. On the other hand, the recent state-of-the-art GAN models (*e.g.*, StyleGAN [17] and BigGAN [6]) introduce layer-wise stochasticity to improve the training



Fig. 4: 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 corresponding to 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 varies in the synthesis, and use a re-scoring technique to quantitatively verify the emergence of the target concept.

stability and synthesis quality. As shown in Fig.3, compared to the conventional generator which only takes the latent code as the input of the first layer, the improved generator with layer-wise stochasticity take random latent codes into all the layers. We therefore treat them as layer-wise generative representations. It is worth mentioning that more and more latest GAN models inherit the design of using layer-wise latent codes to achieve better generation quality, such as SinGAN [28] and HoloGAN [24].

To explore how GANs are able to produce high-quality scene synthesis by learning multi-level variation factors as well as what role the generative representation of each layer plays in this process, this work aims at establishing the relationship between the variation factors and the generative representations. Karras et al [17] has already pointed out that the design of layer-wise stochasticity actually controls the synthesis from coarse to fine, however, what “coarse” and “fine” actually refer to still remains uncertain. Differently, to align the variation factors with human perception, we separate them into four abstraction levels, including *layout*, *categorical objects*, *scene attributes*, and *color scheme*. We further propose a framework in Sec.4 to quantify the causality between the input generative representations and the output variation factors. We surprisingly find that GAN synthesizes a scene in a manner highly consistent with human. Over all 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 attribute and color scheme at the later stage.

4 Identifying the Emergent Variation Factors

As described in Sec.3, we target at interpreting the latent semantics learned by scene synthesis models from four different abstraction levels. Previous efforts on several scene understanding databases [42, 33, 18, 26] enable a series of classifiers to predict scene attributes and categories. Besides,

we also employ several off-the-shelf classifiers focusing on layout detection [40] and semantic segmentation [35] to help analyze the synthesized scene images. Specially, given an image, we are able to use these classifiers to get the response scores with respect to various semantics. However, only predicting the semantic labels is far from identifying the variation factors that GANs have captured from the training data. More concretely, among all the multi-level candidate concepts, not all of them are meaningful to a particular scene synthesis model. For instance, “indoor lighting” will never happen in outdoor scenes such as bridge and tower, which “enclosed area” is always true for indoor scenes such as bedroom and kitchen. Accordingly, we come up with a method to quantitatively identify the most relevant and manipulatable variation factors that emerge inside the learned generative representation. Fig.4 illustrates the identification process which consists of two steps, *i.e.*, probing (Sec.4.1) and verification (Sec.4.2). Such identification enables the diverse scene manipulation (Sec.4.3).

4.1 Probing Latent Space

The generator of GAN, $G(\cdot)$, typically learns the mapping from latent space \mathcal{Z} to image space \mathcal{X} . Latent vectors $\mathbf{z} \in \mathcal{Z}$ can be considered as the generative representation learned by GAN. To study the emergence of variation factors inside \mathcal{Z} , we need to first extract semantic information from \mathbf{z} , which is not trivial. To solve this problem, we employ synthesized image, $\mathbf{x} = G(\mathbf{z})$, as an intermediate step and use a broad set of *off-the-shelf* image classifiers to help assign semantic scores for each sampled latent code \mathbf{z} . Taking “indoor lighting” as an example, the scene attribute classifier is able to output the probability of how an input image looks like having indoor lighting, which we use as the semantic score. Recall that we divide scene representation into layout, object (category), and attribute levels, we introduce layout estimator, scene category recognizer, and attribute classifier to predict semantic scores from these abstraction levels respectively,

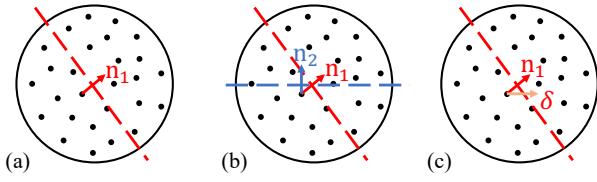


Fig. 5: Three types of manipulation: (a) *Independent* manipulation; (b) *Joint* manipulation; (c) *Jittering* manipulation.

forming a hierarchical semantic space \mathcal{S} . After establishing the one-on-one mapping from latent space \mathcal{Z} to semantic space \mathcal{S} , we search the decision boundary for each concept by treating it as a bi-classification problem, as shown in Fig.4. Here, taking “indoor lighting” as an instance, the boundary separates the latent space \mathcal{Z} to two sets, *i.e.*, presence or absence of indoor lighting.

4.2 Verifying Manipulatable Variation Factors

After probing the latent space with a broad set of candidate concepts, we still need to figure out which ones are most relevant to the generative model acting as the variation factors. The key issue is how to define “relevance”, or say, how to verify whether the learned representation has already encoded a particular variation factor. We argue that if the target concept is manipulatable from latent space perspective (*e.g.*, change the indoor lighting status of the synthesized image via simply varying the latent code), the GAN model is able to capture such variation factors during the training process.

As mentioned above, we have already got separation boundaries for each candidate. Let $\{\mathbf{n}_i\}_{i=1}^C$ denote the normal vectors of these boundaries, where C is the total number of candidates. For a certain boundary, if we move a latent code \mathbf{z} along its normal direction (positive), the semantic score should also increase correspondingly. Therefore, we propose to re-score the varied latent code to *quantify* how a variation factor is relevant to the target model for analysis. As shown in Fig.4, this process can be formulated as

$$\Delta s_i = \frac{1}{K} \sum_{k=1}^K \max \left(F_i(G(\mathbf{z}^k + \lambda \mathbf{n}_i)) - F_i(G(\mathbf{z}^k)), 0 \right), \quad (1)$$

where $\frac{1}{K} \sum_{k=1}^K$ stands for the average of K samples to make the metric more accurate. λ is a fixed moving step. To make this metric comparable among all candidates, all normal vectors $\{\mathbf{n}_i\}_{i=1}^C$ are normalized to fixed norm 1 and λ is set as 2. With this re-scoring technique, we can easily rank the score Δs_i among all C concepts to retrieve the most relevant latent variation factors.

Table 1: Description of the StyleGAN models trained on different categories. ↓ means the lower the better.

Scene Category	Indoor / Outdoor	Training Samples	FID ↓
bedroom (official)	Indoor	3M	2.65
living room	Indoor	1.3M	5.16
kitchen	Indoor	1M	5.06
restaurant	Indoor	626K	4.03
bridge	Outdoor	819K	6.42
church	Outdoor	126K	4.82
tower	Outdoor	708K	5.99
Mixed	Indoor	500K each	3.74

4.3 Manipulation with Diversity

After identifying the semantic variation factors, we propose several ways to further manipulate images. Figure 5 shows three types of scene manipulation. A simple and straightforward way, named *independent* manipulation, is to push the code \mathbf{z} along the normal vector \mathbf{n}_i of a certain semantic at the step length of λ . The manipulated code $\mathbf{z}' \leftarrow \mathbf{z} + \lambda \mathbf{n}$ is then fed into the generator to produce the new image. A second way of manipulation enables scene editing with respect to more than one variation factor jointly. We call it *joint* manipulation. Taking two variation factors (with normal vector \mathbf{n}_1 and \mathbf{n}_2) as an example, the original code \mathbf{z} is moved along the two directions simultaneously as $\mathbf{z}' \leftarrow \mathbf{z} + \lambda_1 \mathbf{n}_1 + \lambda_2 \mathbf{n}_2$. Here λ_1 and λ_2 are step parameters which control the strength of manipulation corresponding to these two semantic respectively. Since such two manipulation methods enable more precise control from multiple abstraction levels, we also introduce randomness into the manipulation process to increase the diversity, namely *jittering* manipulation. The key idea is to slightly perturb the manipulation direction with a randomly sampled noise $\delta \sim \mathcal{N}(0, 1)$. It can be then formulated as $\mathbf{z}' \leftarrow \mathbf{z} + \lambda \mathbf{n} + \delta$.

5 Experiments

In the generation process, the deep representation at each layer, especially for StyleGAN [17] and BigGAN [6], is actually directly derived from the projected latent code. Therefore, we consider the latent code as the *generative representation*, which may be slightly different from the conventional definition in the classification networks. We conduct a detailed empirical analysis of the variation factors identified across the layers of the generators in GANs. We show that the hierarchy of variation factors emerges in the deep generative representations as a result of learning to synthesize scenes.

The experimental section is organized as follows: Sec.5.1 introduces our experimental details including generative models, training datasets and the *off-the-shelf* classifiers we used. Sec.5.2 contains the layer-wise analysis on the state-of-the-art StyleGAN model [17], quantitatively and qualitatively



Fig. 6: Diverse generated samples for training decision boundary with respect to layout, objects (category), scene attributes and color scheme.

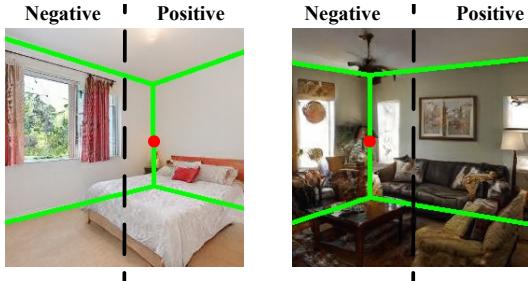


Fig. 7: The definition of layout for indoor scenes. Green lines represent for the outline prediction from the layout estimator. The dashed line indicates the horizontal center, and the red point is the center point of the intersection line of two walls. The relative position between the vertical line and the center point is used to split the dataset.

verifying that the multi-level variation factors are encoded in the latent space. In Sec.5.3, we explore the question on how GANs represent categorical information such as bedroom *v.s.* living room. We reveal that GAN synthesizes the shared objects at some intermediate layers. By controlling their activations only, we can easily overwrite the category of the output image, *e.g.* turning bedroom into living room, while preserving its original layout and high-level attributes such as indoor lighting. Sec.5.4 further shows that our approach can faithfully identify the most relevant attributes associated with a particular scene, facilitating semantic scene manipulation. Sec.5.5 conducts the ablation studies on re-scoring technique and layer-wise manipulation to show the effectiveness of our approach.

5.1 Experimental Details

Generator models. This work conducts experiments on state-of-the-art deep generative models for high-resolution scene synthesis, including StyleGAN [17], BigGAN [6], and PGGAN [16]. Among them, PGGAN employs the conventional generator structure where the latent code is only fed into the very first layer. Differently, StyleGAN and BigGAN

introduce layer-wise stochasticity by feeding latent codes to all convolutional layers as shown in Fig.3. And our layer-wise analysis sheds light on why it is effective.

Scene categories. Among the mentioned generator models, PGGAN and StyleGAN are actually trained on LSUN dataset Yu et al [38] while BigGAN is trained on Places dataset [42]. LSUN dataset consists of 7 indoor scene categories and 3 outdoor scene categories, and Places dataset contains 10 million images across 434 categories. For PGGAN model, we use the officially released models², each of which is trained to synthesize scene within a particular category of LSUN dataset. For StyleGAN, only one model related to scene synthesis (*i.e.*, bedroom) is released³. For a more thorough analysis, we use the official implementation⁴ to train some additional models on other scene categories, including both indoor scenes (living room, kitchen, restaurant) and outdoor scenes (bridge, church, tower). We also train a *mixed* model on the combination of images from bedroom, living room, and dining room with the same implementation. This model is specifically used for categorical analysis. For each StyleGAN model, Tab.1 shows the category, the number of training samples, as well as the corresponding Fréchet inception distances (FID) [12] which can reflect the synthesis quality to some extent. For BigGAN, we use the author’s officially unofficial PyTorch BigGAN implementation⁵ to train a conditional generative model by taking category label as the constraint on Places dataset [42]. The resolution of the scene images synthesized by all of the above models is 256×256 .

Semantic Classifiers. To extract semantic from synthesized images, we employ some *off-the-shelf* image classifiers to

² These PGGAN models can be found at https://drive.google.com/open?id=15hvzxt_XxuokSmj0u04xxMTMWVc0cIMU.

³ The StyleGAN model can be found at <https://drive.google.com/drive/folders/1MASQyN5m0voPcx7-9K0r5gObhvvPups7>.

⁴ The implementation of StyleGAN can be found at <https://github.com/NVlabs/stylegan>.

⁵ The implementation of BigGAN can be found at <https://github.com/ajbrock/BigGAN-PyTorch>.

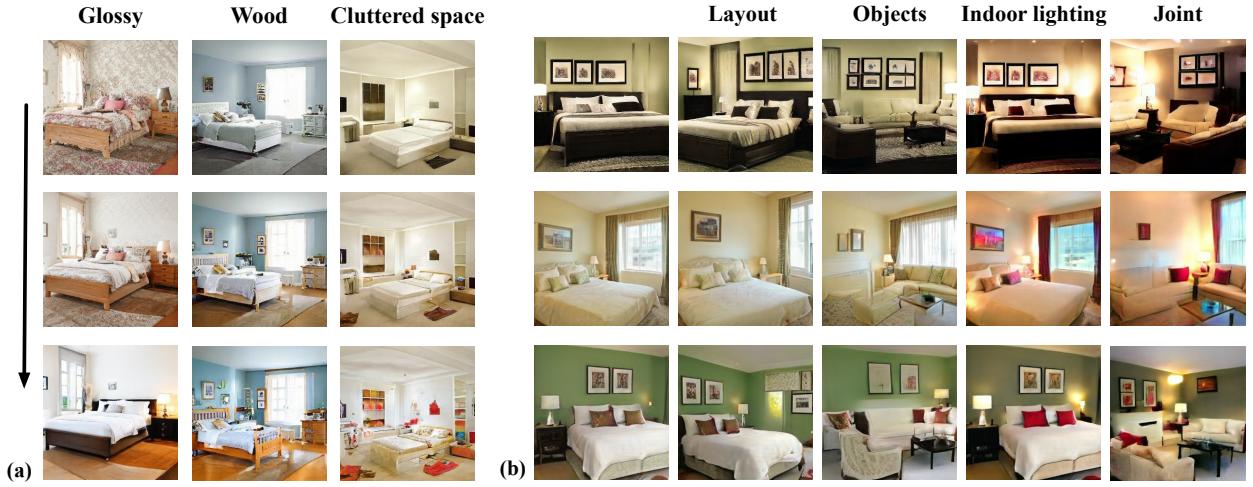


Fig. 8: (a) *Independent* attribute manipulation results on Upper layers. The middle row is the source images. We are able to both decrease (top row) and increase (bottom row) the variation factors in the images. (b) *Joint* manipulation results, where the *layout*, *objects* and *attribute* are manipulated at proper layers. The first column indicates the source images, the middle three columns are the independently manipulated images.

assign these images with semantic scores from multiple abstraction levels, including *layout*, *category*, *scene attribute*, and *color scheme*. Specifically, we use (1) a *layout estimator* [40], which predicts the spatial structure of an indoor place, (2) a *scene category classifier* [42], which classifies a scene image to 365 categories, and (3) an *attribute predictor* [42], which predicts 102 pre-defined scene attributes in SUN attribute database [26]. We also extract color scheme of a scene image through its hue histogram in HSV space. Among them, the category classifier and attribute predictor can directly output the probability of how likely an image belongs to a certain category or how likely an image has a particular attribute. As for the layout estimator, it only detects the outline structure of an indoor place, shown as the green line in Fig.7.

Semantic Probing and Verification. Given a well-trained GAN model for analysis, we first generate a collection of synthesized scene images by randomly sampling N latent codes. To ensure capturing all the potential variation factors, we set $N = 500,000$. We then use the aforementioned image classifiers to assign semantic scores for each visual concept. It is worth noting that we use the relative position between image horizontal center and the intersection line of two walls to quantify layout, as shown in Fig.7. After that, for each candidate, we select 2,000 images with the highest response as positive samples, and another 2,000 with the lowest response as negative ones. Fig.6 shows some examples, where the living room and bedroom are treated as positive and negative for scene category respectively. We then train a linear SVM by treating it as a bi-classification problem (*i.e.*, data is the sampled latent code while the label is binary indicating whether the target semantic appears in the corresponding synthesis or not) to get a linear decision

	Bottom	Lower	Upper	Top
Layout	95%	5%	0%	0%
Objects	10%	90%	0%	0%
Attributes	0%	5%	85%	5%
Color Scheme	0%	0%	25%	75%

User Study

Fig. 9: User study on how different layers correspond to variation factors from different abstraction levels.

boundary. Finally, we re-generate $K = 1,000$ samples for semantic verification as described in Sec.4.2.

5.2 Emerging Semantic Hierarchy

Humans typically interpret a scene in a hierarchy of semantics, from its layout, underlying objects, to the detailed attributes and the color scheme. Here the underlying objects refer to the set of objects most relevant to a specific category. This section shows that GAN composes a scene over the layers in a similar way with human perception. To enable analysis on layout and object, we take the *mixed* StyleGAN model trained on indoor scenes as the target model. StyleGAN [17] learns a more disentangled latent space \mathcal{W} on top of the conventional latent space \mathcal{Z} and feeds the latent code $\mathbf{w} \in \mathcal{W}$ to each convolutional layer with different transformations instead of only feeding it to the first layer. Specifically, for ℓ -th layer, \mathbf{w} is linearly transformed to layer-wise transformed latent code $\mathbf{y}^{(\ell)}$ with $\mathbf{y}^{(\ell)} = \mathbf{A}^{(\ell)}\mathbf{w} + \mathbf{b}^{(\ell)}$, where $\mathbf{A}^{(\ell)}$, $\mathbf{b}^{(\ell)}$ are the weight and

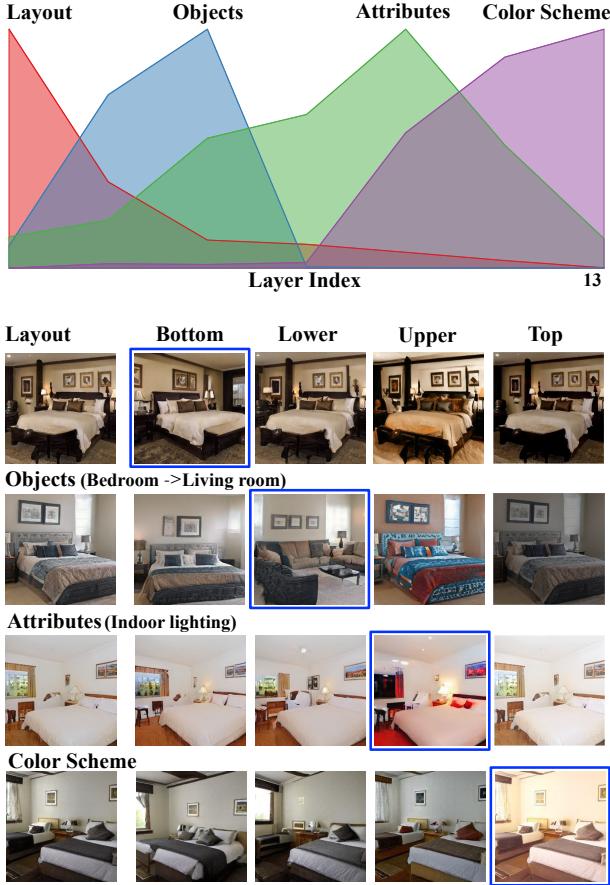


Fig. 10: Top: Four levels of visual abstractions emerge at different layers of StyleGAN. Vertical axis shows the normalized perturbation score Δs_i . Bottom: Layer-wise manipulation result. The first column is the original synthesized images, and the other columns are the manipulated images at layers from four different stages respectively. Blue boxes highlight the results from varying the latent code at the most proper layers for the target concept.

bias for style transformation respectively. We thus perform layer-wise analysis by studying $\mathbf{y}^{(\ell)}$ instead of \mathbf{z} in Eq.(1).

To quantify the importance of each layer with respect to each variation factor, we use the re-scoring technique to identify the causality between the layer-wise generative representation $\mathbf{y}^{(\ell)}$ and the semantic emergence. The normalized score in the top Fig.10 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. This is consistent with human perception. In StyleGAN model that is trained to produce 256×256 scene images, there are totally 14 convolutional layers. According to our experimental results, *layout*, *object* (*category*), *attribute*, *color scheme* correspond to *bottom*, *lower*, *upper*, and *top* layers respectively, which are actually [0, 2), [2, 6), [6, 12) and [12, 14) layers.

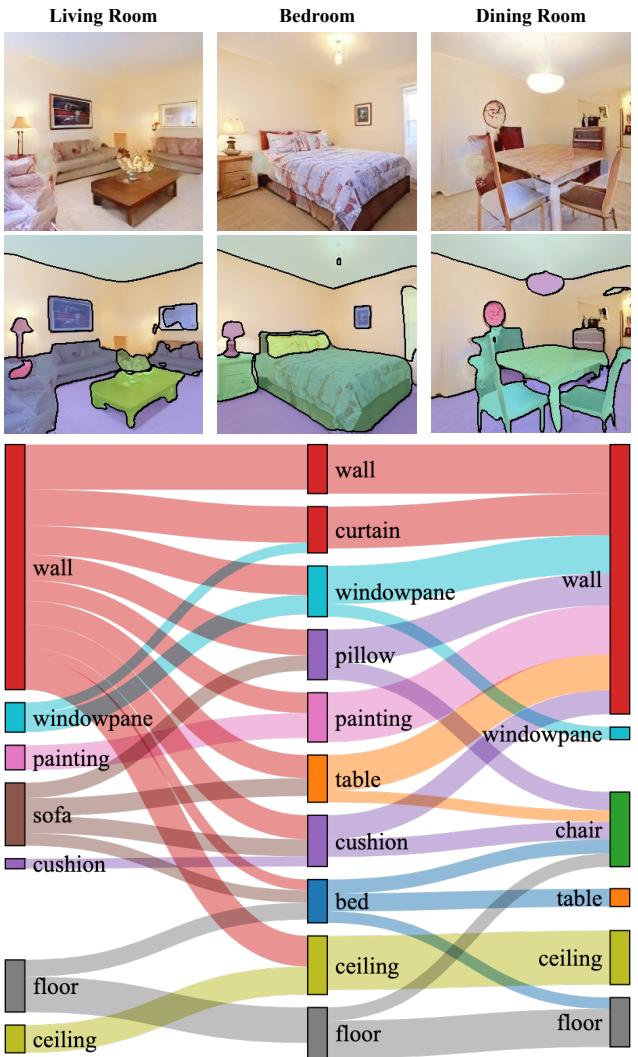


Fig. 11: Objects are transformed by GAN to represent different scene categories. On the top shows that the object segmentation mask varies when manipulating a living room to bedroom, and further to dining room. On the bottom visualizes the object mapping that appears during category transition, where 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.

To visually inspect the identified variation factors, we move the latent vector along the boundaries at different layers to show how the synthesis varies correspondingly. For example, given a boundary in regards to room layout, we vary the latent code towards the normal direction at bottom, lower, upper, and top layers respectively. The bottom of Fig.10 shows the qualitative results for several concepts. We see that the emerged variation factors follow a highly-structured semantic hierarchy, e.g., layout can be best controlled at the early stage while color scheme can only be changed at the final stage. Besides, varying latent code at the inappropriate layers may also change the image content, but the changing

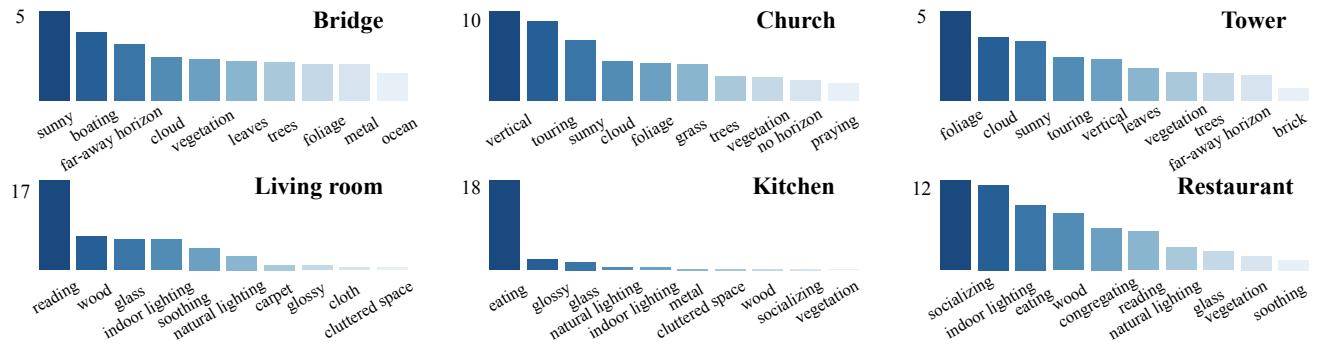


Fig. 12: Comparison of the top scene attributes identified in the generative representations learned by StyleGAN models for synthesizing different scenes. Vertical axis shows the perturbation score Δs_i .

might be inconsistent with the desired output. For example, in the second row, modulating the code at bottom layers for category only leads to a random change in the scene viewpoint.

To better evaluate the manipulability across layers, we conduct a user study. We first generate 500 samples and manipulate them with respect to several concepts on different layers. For each concept, 20 users are asked to choose the most appropriate layers for manipulation. Fig.9 shows the user study results, where most people think bottom layers best align with layout, lower layers control scene category, *etc*. This is consistent with our observations in Fig.10. It suggests that hierarchical variation factors emerge inside the generative representation for synthesizing scenes. and that our re-scoring method indeed helps identify the variation factors from a broad set of semantics.

Identifying the semantic hierarchy and the variation factors across layers facilitates semantic scene manipulation. We can simply push the latent code toward the boundary of the desired attribute at the appropriate layer. Fig.8(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) respectively. Furthermore, we can jointly manipulate hierarchical variation factors. In Fig.8(b) we simultaneously change the room layout (rotating viewpoint) at early layers, scene category (converting bedroom to living room) at middle layers, and scene attribute (increasing indoor lighting) at later layers.

5.3 What Makes a Scene?

As mentioned above, GAN models for synthesizing scenes are capable of encoding hierarchical semantics inside the generative representation, *i.e.*, from layout, object (category), to scene attribute and color scheme. One of the most noticeable properties is that the middle layers of GAN actually synthesize different objects for different scene categories. It raises the question of what makes a scene as living room

rather than bedroom. Thus we further dive into the encoding of categorical information in GANs, to quantify how GAN interprets a scene category as well as how the scene category is transformed from an object perspective.

We employ the StyleGAN model trained on the mixture of bedroom, living room, and dining room, and then search the semantic boundary between every two categories. To extract the objects from the synthesized images, we apply a semantic segmentation model [35], which can segment 150 objects (tv, sofa, *etc*) and stuff (ceiling, floor, *etc*). Specifically, we first randomly synthesize 500 living room images, and then vary the corresponding latent codes towards the “living room-bedroom” boundary and “bedroom-dining room” boundary in turn. We segment the images before and after manipulation to get the segmentation masks, as shown in Fig.11. After tracking label mapping for each pixel via the image coordinate during the manipulation process, we are able to compute the statistics on how objects are transformed along with category changing and observe how the objects change when category is transformed.

Fig.11 shows the objects mapping in the category transformation process. 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 some objects are mapped into other classes. For example, the sofa in living room is mapped to the pillow and bed in bedroom, and the bed in bedroom is further mapped to the table and chair in dining room. This phenomenon happens because sofa, bed, dining table and chair are distinguishable objects for living room, bedroom, and dining room respectively. Thus, when category is transformed, the representative objects are supposed to change. (2) Some objects are sharable between different scene categories, and the GAN model is able to spot such property and learn to generate these shared objects across different classes. For example, the lamp in living room (on the left boundary of the image) still remains after the image is converted to bedroom, especially in the same position. (3) With the ability to learn object mapping as well as share objects across different classes, we are

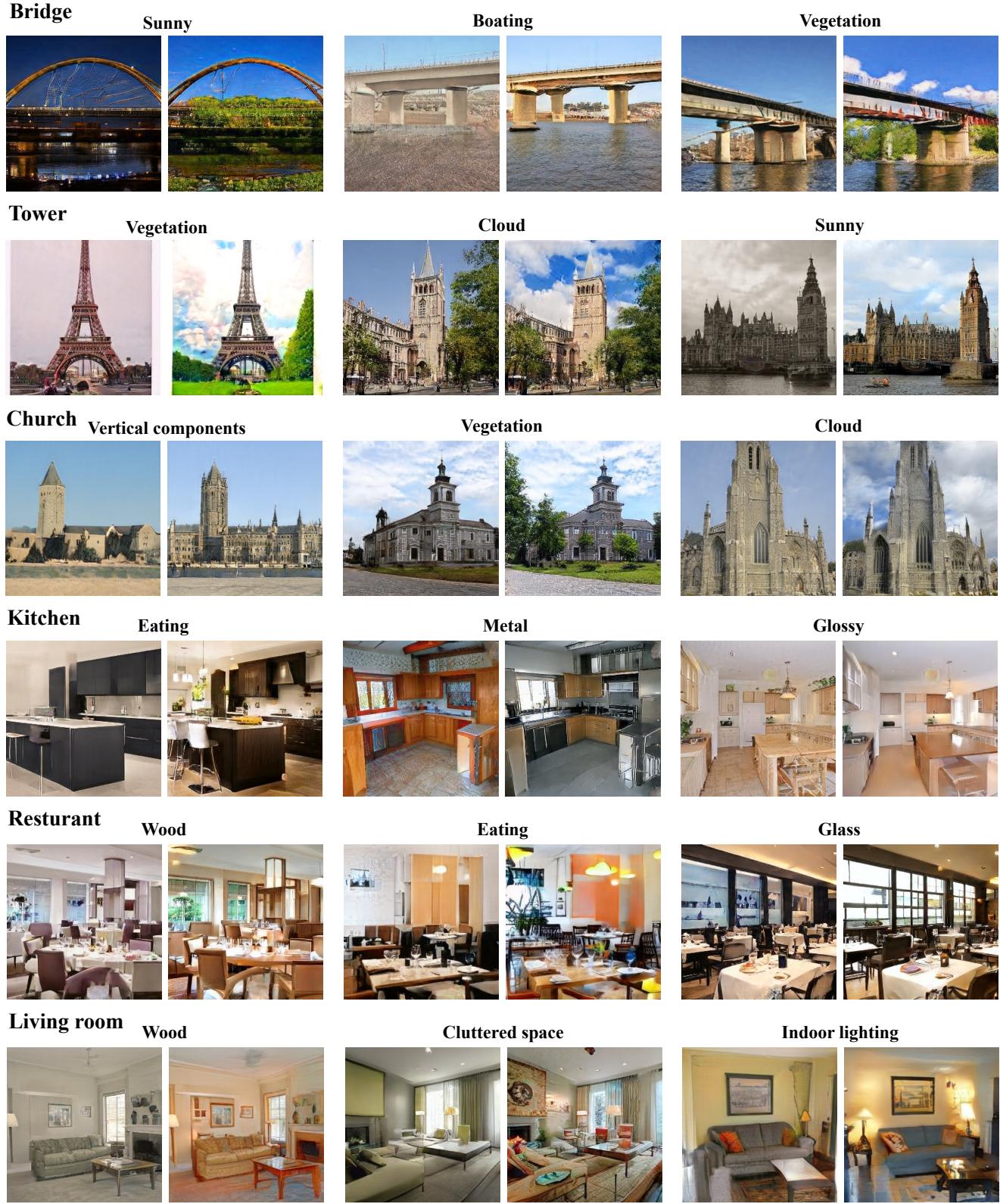


Fig. 13: *Independent* manipulation results on StyleGAN models trained for synthesizing indoor and outdoor scenes. In each pair of images, the first is the original synthesized sample and the second is the one after manipulation of a certain semantics.

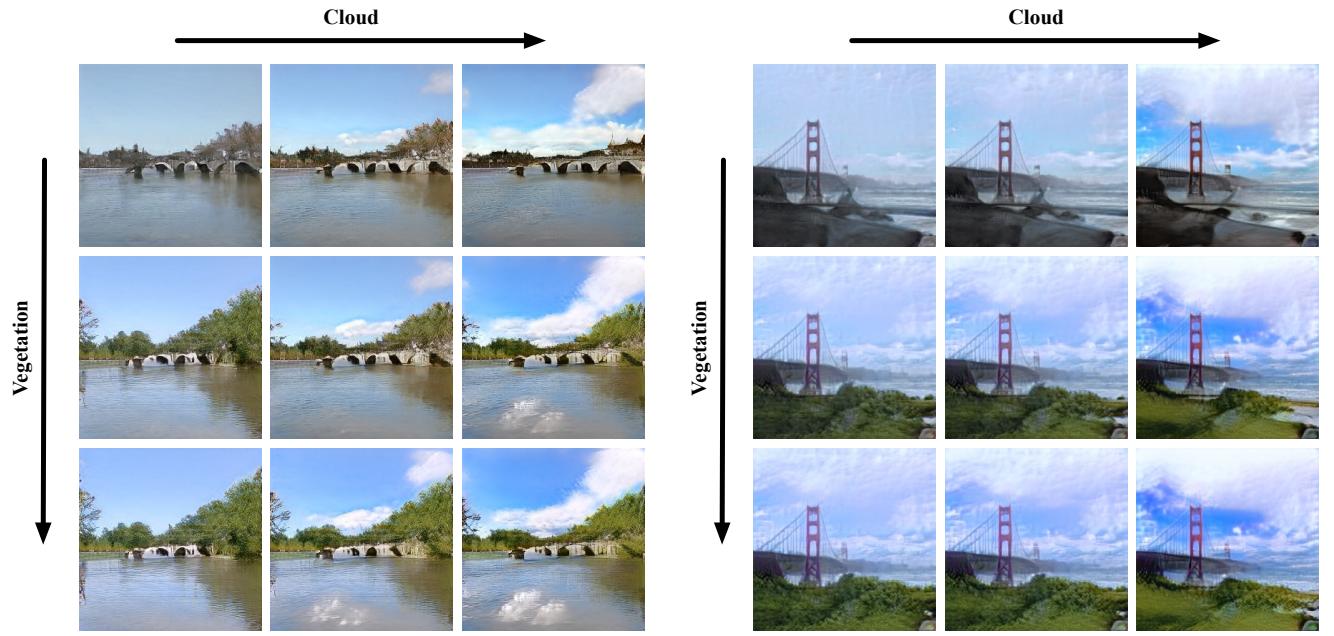


Fig. 14: *Joint* manipulation results along both *cloud* and *vegetation* boundaries with bridge synthesis model. Along the vertical and horizontal axis, the original synthesis (the central image) is manipulated with respect to *vegetation* and *cloud* attributes respectively.

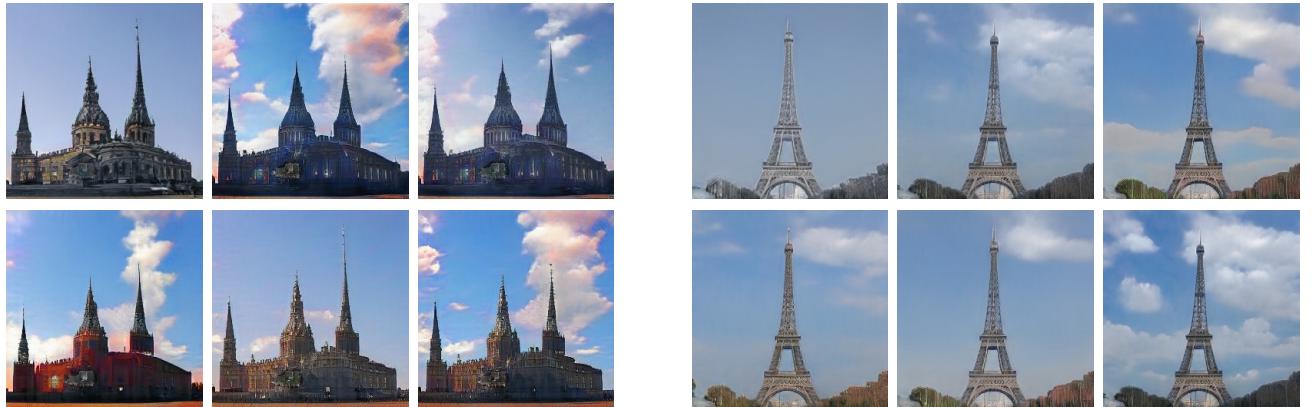


Fig. 15: *Jittering* manipulation results with tower synthesis model for *cloud* attribute. Specifically, the movement in the latent space of synthesized image is disturbed. Thus, when the cloud appears, both the shape of added cloud and appearance of the generated tower change. The top left image of two samples is the original output while the rest are the results under jittering manipulation separately.

able to turn an unconditional GAN into a GAN that can control category. Typically, to make GAN produce images from different categories, class labels have to be fed into the generator to learn a categorical embedding, like BigGAN [6]. Our result suggests an alternative approach.

5.4 Diverse Attribute Manipulation

Attribute Identification. The emergence of variation factors for scene synthesis depends on the training data. Here we apply our method to a collection of StyleGAN models, to capture a wide range of manipulatable attributes out of the 102 scene attributes pre-defined in SUN attribute database

[26]. Each styleGAN in the collection is trained to synthesize scene images from a certain category, including both outdoor (bridge, church, tower) and indoor scenes (living room, kitchen). Fig.12 shows the top-10 relevant semantics to each model. We can see that “sunny” has high scores on all outdoor categories, while “lighting” has high scores on all indoor categories. Furthermore, “boating” is identified for bridge model, “touring” for church and tower, “reading” for living room, “eating” for kitchen, and “socializing” for restaurant. These results are highly consistent with human perception, suggesting the effectiveness of the proposed quantification method.

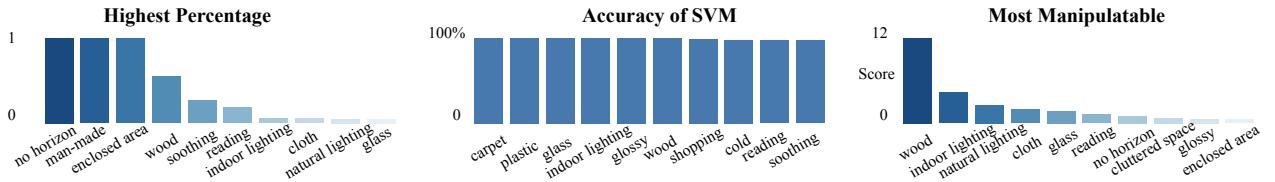


Fig. 16: Ablation study on the proposed re-scoring technique with StyleGAN model for bedroom synthesis. The left shows the percentage of scene attributes with the positive scores, the middle figure sorts by the accuracy of SVM classifiers, while the right figure sorts by our methods.

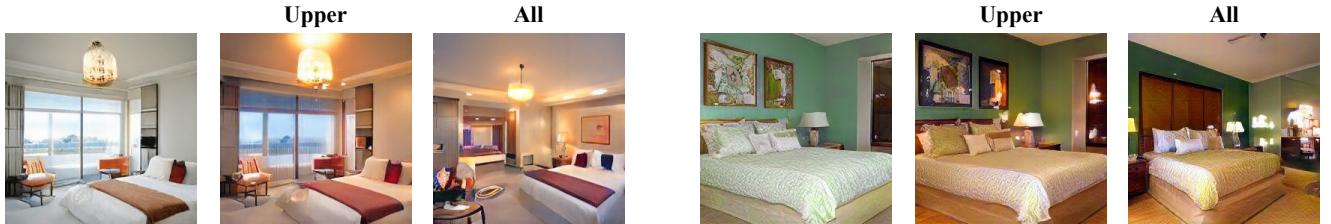


Fig. 17: Comparison results between manipulating latent codes at only upper (attribute-relevant) layers and manipulating latent codes at all layers with respect to *indoor lighting* on StyleGAN.



Fig. 18: Manipulation at the *bottom* layers in 4 different directions, along the directions of *layout*, *objects* (category), *indoor lighting*, and *color scheme* on StyleGAN.

Attribute Manipulation. Recall the three types of manipulation in Sec.4.3: *independent* manipulation, *joint* manipulation, and *jittering* manipulation. We first conduct independent manipulation on 3 indoor and 3 outdoor scenes with the most relevant scene attributes identified with our approach. Fig.13 shows the results where the original synthesis (left image in each pair) is manipulated along the positive (right) direction. We can tell that the edited images are still with high quality and the target attributes indeed change as desired. We then jointly manipulate two attributes with bridge synthesis model as shown in Fig.14. The central image of the 3×3 image grid is the original synthesis, the second row and the second column show the independent manipulation results with respect to “vegetation” and “cloud” attributes respectively, while other images on the four corners are the joint manipulation results. It turns out that we achieve good control of these two semantics and they seem to barely affect each other. However, not all variation factors show such strong disentanglement. From this point of view, our approach also provides a new metric to help measure the entanglement between two variation factors, which will be discussed in

Sec.6. Finally, we evaluate the proposed *jittering* manipulation by introducing noise into the “cloud” manipulation . From Fig.15, we observe that the newly introduced noise indeed increases the manipulation diversity. It is interesting that the introduced randomness may not only affect the shape of added cloud, but also change the appearance of the synthesized tower. But both cases keep the primary goal, which is to edit the cloudiness.

5.5 Ablation Studies

Re-scoring Technique. Before performing the proposed re-scoring technique, we have two more steps, which are (1) assigning semantic scores for synthesized samples, and (2) training SVM classifiers to search semantic boundary. We would like to verify the essentiality of the re-scoring technique in identifying manipulatable semantics. We conduct ablation study on the StyleGAN model trained for synthesizing bedrooms. As shown in Fig.16, the left figure sorts the scene attributes by how many samples are labelled as positive ones, the middle figure sorts by the accuracy of the trained

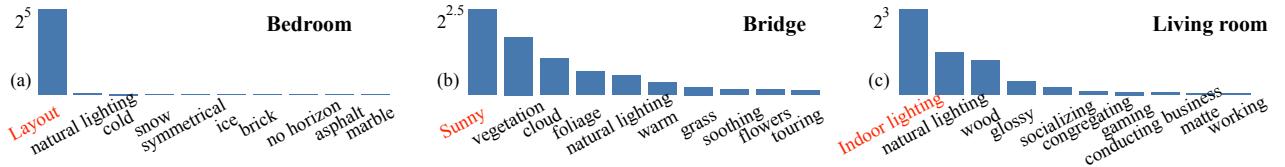


Fig. 19: Effects on scene attributes (already sorted) when varying a particular variation factor (in red color). Vertical axis shows the perturbation score Δs_i in log scale

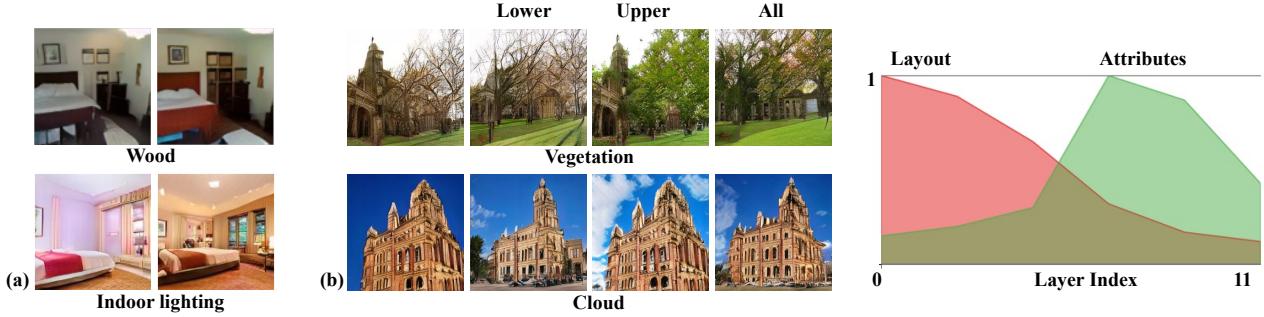


Fig. 20: (a) Some variation factors identified from PGGAN (bedroom). (b) Layer-wise analysis on BigGAN from attribute level.

SVM classifiers, while the right figure sorts by our proposed quantification metric.

In left figure, “no horizon”, “man-made”, and “enclosed area” are attributes with highest percentage. However, all these three attributes are default properties of the bedroom and thus not manipulatable. On the contrary, with the rescore technique for verification, our method successfully filters out these invariable candidates and reveals more meaningful semantics, like “wood” and “indoor lighting”. In addition, our method also manages to identify some less frequent but actually manipulatable scene attributes, such as “cluttered space”.

In the middle figure, almost all attributes get similar scores, making them indistinguishable. Actually, even the worst SVM classifier (*i.e.*, “railroad”) achieves 72.3% accuracy. That is because even some variation factors are not encoded in the latent representation (or say, not manipulatable), the corresponding attribute classifier still assigns synthesized images with different scores. Training SVM on these inaccurate data can also result in a separation boundary, even it is not expected as the target concept. Therefore, only relying on the SVM classifier is not enough to detect relevant variation factors. By contrast, our method pays more attention to the score modulation after varying the latent code, which is not biased by the initial response of attribute classifier or the performance of SVM. As a result, we are able to thoroughly yet precisely detect the variation factors in the latent space from a broad candidate set.

Layer-wise Manipulation. To further validate the emergence of semantic hierarchy, we make ablation study on layer-wise manipulation with StyleGAN model. First, we select “indoor lighting” as the target semantic, and vary the latent code

only on upper (attribute-relevant) layers *v.s.* on all layers. We can easily tell from Fig.17 that when manipulation “indoor lighting” at all layers, the objects inside the room are also changed. By contrast, manipulating latent codes only at attribute-relevant layers can satisfactorily increase the indoor lighting without affecting other factors. Second, we select bottom layers as the target layers, and select boundaries from all four abstraction levels for manipulation. As shown in Fig.18, no matter what level of semantics we choose, as long as the latent code is modified at bottom (layout-relevant) layers, only layout instead of all other semantics varies. These two experiments further verify our discovery about the emergence of the semantic hierarchy that the early layers tend to determine the spatial layout and configuration instead of other abstraction level semantics.

6 Discussions

Disentanglement of Semantics. Some variation factors we detect in the generative representation are more disentangled with each other than other semantics. Compared to the perceptual path length and linear separability described in Karras et al [17] and the cosine similarity proposed in Shen et al [30], our work offers a new metric for disentanglement analysis. In particular, we move the latent code along one semantic direction and then check how the semantic scores of other factors change accordingly. As shown in Fig.19(a), when we modify the spatial layout, all scene attributes are barely affected, suggesting that GAN learns to disentangle layout-level semantic from attribute-level. However, there are also some scene attributes (from same abstraction level) entangling with each other. Taking Fig.19(c) as an example,

when modulating “indoor lighting”, “natural lighting” also varies. This is also aligned with human perception, further demonstrating the effectiveness of our proposed quantification metric.

Application to Other GANs. We further apply our method for two other GAN structures, *i.e.*, PGGAN [16] and BigGAN [6]. These two models are trained on LSUN dataset Yu et al [38] and Places dataset [42] respectively. Compared to StyleGAN, PGGAN feeds the latent vector only to the very first convolutional layer and hence does not support layer-wise analysis. But the proposed re-scoring method can still be applied to help identify manipulatable semantics, as shown in Fig.20(a). BigGAN is the state-of-the-art conditional GAN model that concatenates the latent vector with a class-guided embedding code before feeding it to the generator, and it also allows layer-wise analysis like StyleGAN. Fig.20(b) gives analysis results on BigGAN from attribute level, where we can tell that scene attribute can be best modified at upper layers compared to lower layers or all layers. As for BigGAN model with 256×256 resolution, there are total 12 convolutional layers. As the category information is already encoded in the “class” code, we only separate the layers to two groups, which are *lower* (bottom 6 layers) and *upper* (top 6 layers). Meanwhile, the quantitative curve shows the consistent result with the discovery on StyleGAN as in Fig.10(a). These results demonstrate the generalization ability of our approach as well as the emergence of manipulatable factors in other GANs.

Limitation. There are several limitations for future improvement. First, the layout classifier can only detect the layout structure of indoor scenes. But for a more general analysis on both indoor and outdoor scene categories, there lacks a unified definition of the spatial layout. For example, our framework cannot change the layout of outdoor church images. In future work, we will leverage the computational photography tools that recover the 3D camera pose of the image, thus we can extract more universal viewpoint representation for the synthesized images. Second, our proposed re-scoring technique relies on the performances of the off-the-shelf classifiers. For some of the attributes, the classifiers are not so accurate, which leads to poor manipulation boundary. This problem could be addressed with more powerful discriminative models. Third, for simplicity we only use the linear SVM for semantic boundary search. This limits our framework from interpreting the latent semantic subspace with more complex and nonlinear structure.

7 Conclusion

In this paper, we show the emergence of highly-structured variation factors inside the deep generative representations learned by GANs with layer-wise stochasticity. In particular, the GAN model spontaneously learns to set up layout at early layers, generate categorical objects at middle layers,

and render scene attribute and color scheme at later layers when trained to synthesize scenes. A re-scoring method is proposed to quantitatively identify the manipulatable semantic concepts within a well-trained model, enabling photo-realistic scene manipulation. We will explore to extend this manipulation capability of GANs for real image editing in future work.

Acknowledgement: This work is supported by the Early Career Scheme of Hong Kong (No. 24206219) and RSFS grant from CUHK Faculty of Engineering (No. 3133233).

References

1. Agrawal P, Girshick R, Malik J (2014) Analyzing the performance of multilayer neural networks for object recognition. In: ECCV
2. Alain G, Bengio Y (2016) Understanding intermediate layers using linear classifier probes. arXiv:161001644
3. Bau D, Zhou B, Khosla A, Oliva A, Torralba A (2017) Network dissection: Quantifying interpretability of deep visual representations. In: CVPR
4. Bau D, Zhu JY, Strobelt H, Zhou B, Tenenbaum JB, Freeman WT, Torralba A (2019) Gan dissection: Visualizing and understanding generative adversarial networks. In: ICLR
5. Bengio Y, Courville A, Vincent P (2013) Representation learning: A review and new perspectives. TPAMI
6. Brock A, Donahue J, Simonyan K (2019) Large scale gan training for high fidelity natural image synthesis. In: ICLR
7. Cheng MM, Zheng S, Lin WY, Vineet V, Sturgess P, Crook N, Mitra NJ, Torr P (2014) Imagespirit: Verbal guided image parsing. ACM Trans on Graphics
8. Choi Y, Choi M, Kim M, Ha JW, Kim S, Choo J (2018) Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In: CVPR
9. Goetschalckx L, Andonian A, Oliva A, Isola P (2019) Ganalyze: Toward visual definitions of cognitive image properties. In: ICCV
10. Gonzalez-Garcia A, Modolo D, Ferrari V (2018) Do semantic parts emerge in convolutional neural networks? IJCV
11. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y (2014) Generative adversarial nets. In: NeurIPS
12. Heusel M, Ramsauer H, Unterthiner T, Nessler B, Hochreiter S (2017) Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: NeurIPS
13. Isola P, Zhu JY, Zhou T, Efros AA (2017) Image-to-image translation with conditional adversarial networks. In: CVPR

14. Jahanian A, Chai L, Isola P (2020) On the "steerability" of generative adversarial networks. ICLR
15. Karacan L, Akata Z, Erdem A, Erdem E (2016) Learning to generate images of outdoor scenes from attributes and semantic layouts. arXiv preprint arXiv:161200215
16. Karras T, Aila T, Laine S, Lehtinen J (2018) Progressive growing of gans for improved quality, stability, and variation. In: ICLR
17. Karras T, Laine S, Aila T (2019) A style-based generator architecture for generative adversarial networks. In: CVPR
18. Laffont PY, Ren Z, Tao X, Qian C, Hays J (2014) Transient attributes for high-level understanding and editing of outdoor scenes. ACM Trans on Graphics
19. Liao J, Yao Y, Yuan L, Hua G, Kang SB (2017) Visual attribute transfer through deep image analogy. arXiv preprint arXiv:170501088
20. Luan F, Paris S, Shechtman E, Bala K (2017) Deep photo style transfer. In: CVPR
21. Mahendran A, Vedaldi A (2015) Understanding deep image representations by inverting them. In: CVPR
22. Morcos AS, Barrett DG, Rabinowitz NC, Botvinick M (2018) On the importance of single directions for generalization. In: ICLR
23. Nguyen A, Dosovitskiy A, Yosinski J, Brox T, Clune J (2016) Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. In: NeurIPS
24. Nguyen-Phuoc T, Li C, Theis L, Richardt C, Yang YL (2019) Hologan: Unsupervised learning of 3d representations from natural images. In: ICCV
25. Park T, Liu MY, Wang TC, Zhu JY (2019) Semantic image synthesis with spatially-adaptive normalization. In: CVPR
26. Patterson G, Xu C, Su H, Hays J (2014) The sun attribute database: Beyond categories for deeper scene understanding. IJCV
27. Radford A, Metz L, Chintala S (2016) Unsupervised representation learning with deep convolutional generative adversarial networks. In: ICLR
28. Shaham TR, Dekel T, Michaeli T (2019) Singan: Learning a generative model from a single natural image. In: ICCV
29. Shen Y, Luo P, Yan J, Wang X, Tang X (2018) Faceid-gan: Learning a symmetry three-player gan for identity-preserving face synthesis. In: CVPR
30. Shen Y, Gu J, Tang X, Zhou B (2019) Interpreting the latent space of gans for semantic face editing. arXiv preprint arXiv:190710786
31. Simonyan K, Vedaldi A, Zisserman A (2014) Deep inside convolutional networks: Visualising image classification models and saliency maps. In: ICLR Workshop
32. Wang TC, Liu MY, Zhu JY, Tao A, Kautz J, Catanzaro B (2018) High-resolution image synthesis and semantic manipulation with conditional gans. In: CVPR
33. Xiao J, Hays J, Ehinger KA, Oliva A, Torralba A (2010) Sun database: Large-scale scene recognition from abbey to zoo. In: CVPR
34. Xiao T, Hong J, Ma J (2018) Elegant: Exchanging latent encodings with gan for transferring multiple face attributes. In: ECCV
35. Xiao T, Liu Y, Zhou B, Jiang Y, Sun J (2018) Unified perceptual parsing for scene understanding. In: ECCV
36. Yao S, Hsu TM, Zhu JY, Wu J, Torralba A, Freeman B, Tenenbaum J (2018) 3d-aware scene manipulation via inverse graphics. In: NeurIPS
37. Yosinski J, Clune J, Bengio Y, Lipson H (2014) How transferable are features in deep neural networks? In: NeurIPS
38. Yu F, Seff A, Zhang Y, Song S, Funkhouser T, Xiao J (2015) Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. arXiv preprint arXiv:150603365
39. Zeiler MD, Fergus R (2014) Visualizing and understanding convolutional networks. In: ECCV
40. Zhang W, Zhang W, Gu J (2019) Edge-semantic learning strategy for layout estimation in indoor environment. In: IEEE Transactions on Cybernetics
41. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A (2015) Object detectors emerge in deep scene cnns. In: ICLR
42. Zhou B, Lapedriza A, Khosla A, Oliva A, Torralba A (2017) Places: A 10 million image database for scene recognition. TPAMI
43. Zhu JY, Park T, Isola P, Efros AA (2017) Unpaired image-to-image translation using cycle-consistent adversarial networks. In: ICCV