Compositional Transformers for Scene Generation

Drew A. Hudson
Department of Computer Science
Stanford University
dorarad@cs.stanford.edu

C. Lawrence Zitnick
Facebook AI Research
Facebook, Inc.
zitnick@fb.com

Abstract

We introduce the GANformer2 model, an iterative object-oriented transformer, explored for the task of generative modeling. The network incorporates strong and explicit structural priors, to reflect the compositional nature of visual scenes, and synthesizes images through a sequential process. It operates in two stages: a fast and lightweight *planning* phase, where we draft a high-level scene layout, followed by an attention-based *execution* phase, where the layout is being refined, evolving into a rich and detailed picture. Our model moves away from conventional black-box GAN architectures that feature a flat and monolithic latent space towards a transparent design that encourages efficiency, controllability and interpretability. We demonstrate GANformer2's strengths and qualities through a careful evaluation over a range of datasets, from multi-object CLEVR scenes to the challenging COCO images, showing it successfully achieves state-of-the-art performance in terms of visual quality, diversity and consistency. Further experiments demonstrate the model's disentanglement and provide a deeper insight into its generative process, as it proceeds step-by-step from a rough initial sketch, to a detailed layout that accounts for objects' depths and dependencies, and up to the final high-resolution depiction of vibrant and intricate real-world scenes. See https://github.com/dorarad/gansformer for model implementation.

1 Introduction

Drawing, the practice behind human visual and artistic expression, can essentially be defined as an iterative process. It starts from an initial outline, with just a few strokes that specify the spatial arrangement and overall layout, and is then gradually refined and embellished with color, depth and richness, until a vivid picture eventually emerges. These initial schematic sketches can serve as an abstract scene representation that is very concise, yet highly expressive: several lines are enough to delineate three dimensional structures, account for perspective and proportion, convey shapes and geometry, and even imply semantic information [52, 63, 67]. A large body of research in psychology highlights the importance of sketching in stimulating creative discovery [40, 58], fostering cognitive development [27], and facilitating problem solving [26, 44]. In fact, a large variety of generative and RL tasks either in the visual [5], textual [22, 78] or symbolic modalities [2], can benefit from the same strategy – prepare a high-level plan first, and then carry out the details.

At the heart of this hierarchical strategy stands the principle of compositionality – where the meaning of the whole can be derived from those of its constituents. Indeed, the world around us is highly structured. Our environment consists of a varied collection of objects, tightly interconnected through dependencies of all sorts: from close-by to long-range, and from physical and dynamic to abstract and semantic [10, 66]. As pointed out by prior literature [24, 51, 55, 57], compositional representations are pivotal to human intelligence, supporting our capabilities of reasoning [32], planning [53], learning [54] and imagination [6]. Realizing this principle, and explicitly capturing the objects and elements composing the scene, is thereby a desirable goal, that can make the generative process more explainable, controllable, versatile and efficient.

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

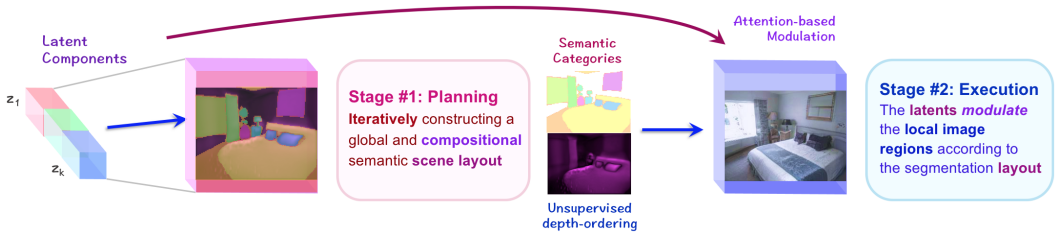

Figure 1: **Overview of the GANformer2 architecture.** The compositional model has a set of latent variables that represent the objects and entities within the generated scene. It proceeds through two stages, **planning** and **execution**: first drafting a high-level layout of the scene's structure and spatial arrangement, and then refining the layout and translating it into a photo-realistic image. Each latent corresponds to a different segment of the scene's layout, controlling both its structure during the planning stage, and style during the execution stage.

Yet, the vast majority of generative models do not directly reflect this compositionality so intrinsic to visual scenes. Rather, most networks, and GANs in particular, aim to generate the entire scene all at once from a single monolithic latent space, through a largely opaque transformation. Consequently, while they excel in producing stunning, strikingly-realistic portrayals of faces, sole centralized objects, or natural scenery, they still struggle to faithfully mimic richer or densely-packed scenes, as those common to everyday contexts, falling short of mastering their nuances and intricacies [5, 50, 76]. Likewise, while there is some evidence for property disentanglement within their latent space to independent axes of variation at the global scale [28, 36, 64, 71], most existing GANs fail to provide controllability at the level of the individual object or local region. Understanding of the mechanisms by which these models give rise to an output image, and the transparency of their synthesis process, thereby remain rather limited.

Motivated to alleviate these shortcomings and make generative modeling more compositional, interpretable and controllable, we propose GANformer2, a structured object-oriented transformer that decouples the visual synthesis task into two stages: **planning** and **execution**. We begin by sampling a set of latent variables, corresponding to the objects and entities that compose the scene. Then, at the planning stage, we transform the latents into a schematic layout – a scene representation that depicts the object segments, reflecting their shapes, positions, semantic classes and relative ordering. Its construction occurs in an iterative manner, to capture the dependencies between different objects. Next, at the execution stage, the layout is translated into the final image, with the latents cooperatively guiding the content and style of their respective segments through bipartite attention [33]. This approach marks a shift towards a more natural process of image generation, in which objects and the relations among them are explicitly modeled, while the initial sketch is successively tweaked and refined to ultimately produce a rich photo-realistic scene.

We study the model's behavior and performance through an extensive set of experiments, where it attains state-of-the-art results in both conditional and unconditional synthesis, as measured along metrics of fidelity, diversity and semantic consistency. GANformer2 reaches high versatility, demonstrated through multiple datasets of challenging simulated and real-world scenes. Further analysis illustrates its capacity to manipulate chosen objects and properties in an independent manner, achieving both spatial disentanglement and separation between structure and style. We notably observe amodal completion of occluded objects, likely driven by the sequential nature of the computation. Meanwhile, inspection of the produced layouts and their components shed more light upon the synthesis of each resulting scene, substantiating the model's transparency and explainability. Overall, by integrating strong compositional structure into the model, we can move in the right direction towards multiple favorable properties: increasing robustness for rich and diverse visual domains, enhancing controllability over individual objects, and improving the generative process's interpretability.

2 Related Work

Over the last years, the field of visual sythesis has witnessed astonishing progress, driven in particular by the emergence of Generative Adversarial Networks [23]. Tremendous strides have been made in a wide variety of tasks, from image [11] to video generation [7], to super-resolution [45], style transfer [15] and translation [35]. Meanwhile, a growing body of literature explored object discovery and sequential scene decomposition, albeit in the context of variational inference [1, 16, 19, 49]. DRAW [25] uses recurrent attention to encode and decode MNIST digits, while AIR [20], MONet [12] and

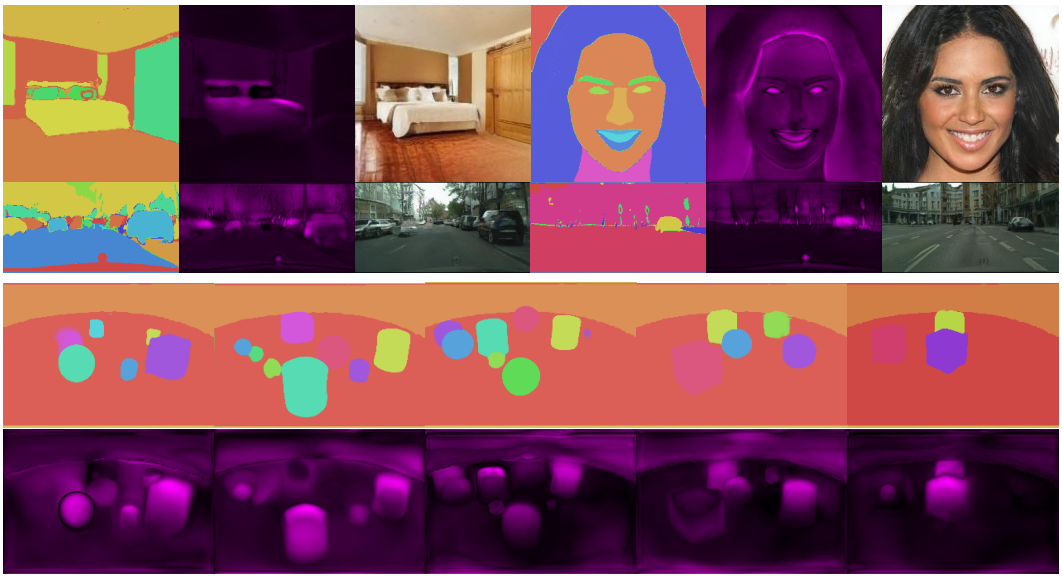

Figure 2: **Layout and image generation.** The GANformer2 model generates each image by first creating a layout that specifies its outline and object composition, which is then translated into the final image. While doing so, the model produces proxy depth maps of the objects which are used to determine their relative ordering, and are notably implicitly derived by the model in a fully unsupervised manner.

IODINE [24] reconstruct simple simulated scenes by iteratively factorizing them into semantically meaningful segments. We draw inspiration from this pioneering line of works and marry the merits of its strong structure with the scalability and fidelity obtained by GANs, yielding a compositional model that successfully creates intricate real-world scenes, while offering a more transparent and controllable synthesis along the way, conceptually meant to better resemble the causal generative process and serve as a more natural emulation of how a human might approach the creative task.

To this end, our model leverages the GANformer architecture we introduced earlier this year [33] – a *bipartite transformer* that uses soft-attention to cooperetaviely translate a set of latents into a generated image [33, 70]. In GANformer2, we take this idea a step further, and show how we can use the bipartite transformer as a backbone for a more structured process, that enables not only an implicit, but also an *explicit modeling* of the objects and constituents composing the image's scene. Our **iterative** design endows the GANformer2 model with the flexibility to consider the correlations and interactions not only among latents but between synthesized scene elements directly, circumventing the need for partial independence assumptions made by earlier works [18, 56, 69], and thereby managing to scale to natural real-world scenes. Concurrently, the explicit **object-oriented** structure enhances the model's compositionallity so to outperform past coarse-to-fine cascade [80], layered [34], and hierarchical approaches [5], which, like us, employ a two-stage layout-based procedure, but contrary to our work, do so in a non-compositional fashion, and indeed were mostly explored within limited domains, such as indoor NYUv2 [73], faces [41], pizzas [59], or bird photographs [77].

Our model links to research on conditional generative modeling [72, 85, 87], including semantic image synthesis and text-to-image generation. The seminal Pix2Pix [35] and SPADE [60] works respectively adopt either a U-Net [62] or spatially-adaptive normalization [61] to map a given input layout into an output high-resolution image. However, beyond their notable weakness of being conditional and thus unable to create scenes from scratch, their reliance on both static class embeddings for feature modulation, along with perceptual or feature-matching reconstruction losses, render their sample variation rather low [68, 86]. In our work, we mitigate this issue by proposing new purely generative structural loss functions: *Semantic-Matching* and *Segment-Fidelity*, which maintain the one-to-many relation that holds between layouts and images, and consequently, as demonstrated in section 3.3, significantly enhance the output diversity.

While GANformer2 is an unconditional model, it shares some high-level ideas with conditional text-to-image [30, 31] and scene-to-image [4, 38] techniques, which use semantic layouts as a bottleneck intermediate representation. But whereas these works focus on heavily-engineered

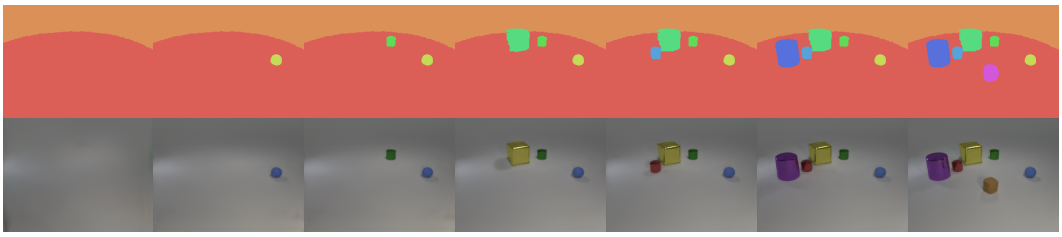

Figure 3: **Recurrent scene generation.** GANformer2 creates the layout sequentially, segment-by-segment, to capture the scene compositionality, effectively allowing us to add or remove objects from the resulting scene.

pipelines consisting of multiple pre- and post-processing stages and different sources of supervision, including textual descriptions, scene graphs, bounding boxes, segmentation masks and images, we propose a simpler, more elegant design, with a streamlined architecture and fewer supervision signals – relying on images and auto-predicted layouts only, to widen its applicability across domains. Finally, our approach relates to prior works on visual editing [8, 9, 46, 84] and image compositing [3, 13, 47, 79], but while these methods modify an existing source image, GANformer2 stands out being capable of creating structured manipulatable scenes from scratch.

3 The GANformer2 model

The GANformer2 model is a compositional object-oriented transformer for visual scene generation. It features a set of latent variables corresponding to the objects and entities composing the scene, and proceeds through two synthesis stages: (1) a sequential **planning stage**, where a scene layout – a set of compounded object segments, is being created (section **3.2**), followed by (2) a parallel **execution stage**, where the layout is being refined and translated into an output photo-realistic image (section **3.3**). Both stages are implemented using customized versions of the GANformer model (section **3.1**) – a bipartite transformer that uses multiple latent variables to modulate the evolving visual features of a generated image, so to create it in a compositional fashion.

In this paper, we explore learning from two training sets: of images and layouts. Specifically, we use panoptic segmentations [43], which specify for every segment s_i its instance identity p_i and semantic category m_i, but other types of segmentations can likewise be used. The images and layouts can either be paired or unpaired, and our training scheme accommodates both options: either by pre-training each stage independently first and then fine-tune them jointly in the paired case, or training them together from scratch in the unpaired one. See Supp F. for further details).

3.1 Generative Adversarial Transformers

The GANformer [33] is a transformer-based Generative Adversarial Network: it consists of a *generator* $G(Z) = X$ that translates a partitioned latent $Z^{k \times d} = [z_1, ..., z_k]$ into an image X^{HWc}, and a *discriminator* $D(X)$ that seeks to discern between real and fake images. The generator begins by mapping the normally-distributed latents into an intermediate space $W = [w_1, ..., w_k]$ using a feed-forward network. Then, it sets an initial 4×4 grid, which gradually evolves through convolution and upsampling layers to the final image. Contrary to traditional GANs, the GANformer also incorporates *bipartite-transformer* layers into the generator, which compute attention between the k latents and the image features, to support spatial and contentual region-wise modulation:

$$\text{Attention}(X, W) = \text{softmax}\left(\frac{q(X)k(W)^T}{\sqrt{d}}\right) v(W) \tag{1}$$

$$u(X, W) = \gamma_a(X, W) \odot LayerNorm(X) + \beta_a(X, W) \tag{2}$$

Where the equations respectively express the attention and modulation between the latents and the image; $q(\cdot), k(\cdot), v(\cdot)$ are the query, key, and value mappings; and $\gamma(\cdot), \beta(\cdot)$ compute multiplicative and additive styles (gain and bias) as a function of $\text{Attention}(X, W)$. The update rule lets the latents shape and impact the image regions as induced by the key-value attention image-to-latents assignment computed between them. This encourages visual compositionality, where different latents specialize to represent and control various semantic regions or visual elements within the generated scene.

Figure 4: **GANformer2 style variations**. Images are produced by varying the style latents of the objects while keeping their structure latents fixed, achieving high visual diversity while maintaining structural consistency.

As we will see below, for each of the two generation stages, we couple a modified version of the GANformer with novel loss functions and training schemes. One key point in which we notably depart from the original GANformer regards to the number of latent variables used by the model. While the GANformer employs a fixed number of k latent variables, GANformer2 supports a **variable number of latents**, thereby becoming more flexible in generating scenes of diverse complexity.

3.2 Planning Stage – Layout Synthesis

In the first stage, we create from scratch a schematic layout depicting the structure of the scene to be formed. Formally, we define a layout S as a collection of k segments $\{s_1, ..., s_k\}$. Each generated segment s_i is associated with the following[1]:

- **Shape & position**: specified by a spatial distribution $P_i(x, y)$ over the image grid $H \times W$.
- **Semantic category**: a soft distribution $M_i(c)$ over the semantic classes.
- **Unsupervised depth-ordering**: per-pixel values $d_i(x, y)$ indicating the order relative to other segments, internally used by the model to compose the segments into a layout (details below).

Recurrent Layout Generation. To explicitly capture conditional dependencies among segments, we construct the layout in an iterative manner, recurrently applying a parameter-lightweight GANformer G_1 over T steps (set to be around 2–4). At each step t, we generate a variable number of segments, $k_t \sim \mathcal{N}(\mu, \sigma^2)$ sampled from a trainable normal distribution[2], and gradually aggregate them into a scene layout S_t. To generate the segments at step t, we use the GANformer with k_t sampled latents, each one z_i is first mapped through a feed-forward network F_1 into an intermediate **structure latent** u_i, which then corresponds to segment s_i, guiding the synthesis of its shape, depth and position.

$$G_1(U^{k_t \times d}, S_{t-1}) = S_t \tag{3}$$

To make the GANformer recurrent, and condition the generation of new segments on ones produced in prior steps, we change the query function $q(\cdot)$ to consider not only the newly formed layout S_t (as in the standard GANformer) but also the previously aggregated one S_{t-1} (concatenating them together and passing as an input), allowing the model to explicitly capture dependencies across steps.

Object Manipulation. Notably, this construction supports a posteriori object-wise manipulation. Given a generated layout S, we can recreate a segment s_i by modifying its corresponding latent from z_i to $\hat{z}_i$ while conditioning on the other segments S_{-i}, mathematically by $G_1(\hat{z}_i^{1 \times d}, S_{-i})$. We thus reach a favorable balance between expressive and contextual modeling of relational interactions on the one hand, and stronger object disentanglement and controllability on the other.

Layout Composition. To compound the segments s_i together into a layout S_t, we overlay them according to their predicted depths $d_i(x, y)$ and shapes $P_i(x, y)$ by computing $\text{Softmax}(d_i(x, y) + \log P_i(x, y))_{i=1}^{k}$ which yields for each image position (x, y) a distribution over the segments. Intuitively, this allows resolving occlusions among overlapping segments. E.g. if a

[1] To produce the segments (instead of RGB values), we modify the output layer to predict for every segment its shape P_i, category M_i and depth d_i. Since the distributions are soft, the model stays fully differentiable

[2] Segments are synthesized in groups for computational efficiency when generating crowded scenes.

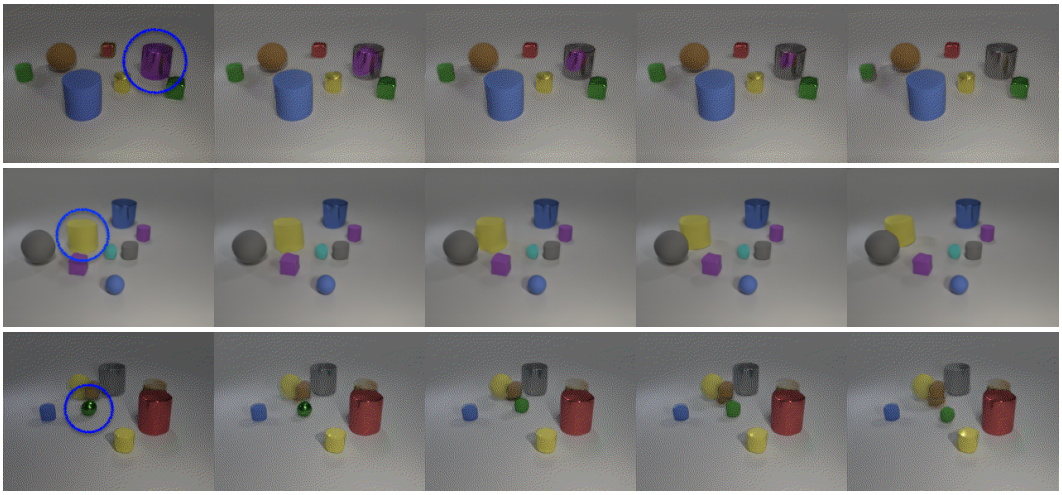

Figure 6: **Object controllability**. GANformer2 supports entity-level manipulation, and separates style from structure, enabling control over color, shape and position of the chosen objects. At each row above, we begin by generating a scene from scratch (leftmost image) and then gradually interpolate the latent vector of one object only: First modifying its **style latent**, leading to a color change, then modify its **structure latent**, resulting in a position change, and finally modify both style and structure, yielding changes along multiple dimensions.

cube segment i should appear before a sphere segment j, then $d_i(x, y) > d_j(x, y)$ within the overlap region $\{(x, y)\}$. Notably, we do not provide any supervision to the model about the true depths of the scene elements, and it rather creates its own internal depth ordering in an **unsupervised** manner, only indirectly moderated by the need to produce realistic looking layouts when stitching the segments together.

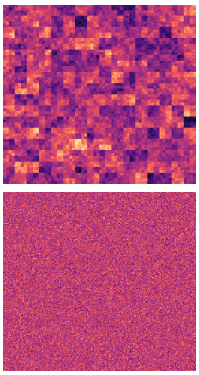

Figure 5: **Hierarchical vs. random noise.**

Noise Scheme for Discrete Generation. The segmentations data we train the planning stage over is inherently discrete, while the generated layouts are soft and continuous. To maintain stable GAN training under these conditions, and as inspired by the instance noise technique [65], we add a simple **hierarchical noise** map to the layouts fed into the discriminator, which helps ensuring its task remains challenging. We define the noise as $\sum_{r=6}^{8} \text{Upsample}(n_r^{2^r \times 2^r})$, summing up noise grids $n_r \sim \mathcal{N}(0,\, \sigma^2)$ across multiple resolutions, which allows it to be partly correlated between adjacent positions and more structured overall. This empirically enhances its effectiveness and substantially improves the quality of the generated layouts.

3.3 Execution Stage: Layout-to-Image Translation

In the second stage, we proceed form planning to execution, and translate the generated layout S into a photo-realistic image X. For each segment s_i, we first concatenate its mean depth d_i and category distribution m_i to its respective latent z_i, and map them through a feed-forward network F_2 into an intermediate **style latent** w_i, which will determine the content, texture and style of its segment. Then, we pass the resulting vectors into a second GANformer $G_2(W, S) = X$ to render the final image.

Layout Translation. As opposed to the original GANformer [33] that computes $\text{Attention}(X, W)$ on-the-fly to determine the assignment between latents and spatial regions, here we rely instead on the pre-planned layout S as a conditional guidance source to specify this association. That way, each latent w_i modulates the features within the soft region of its respective segment s_i, achieving a more explicit semantically-meaningful assignment between latents and scene elements. At the same time, since the whole image is still generated together, with all regions processed through the convolution, upsampling and modulation layers of G_2, the model can still ensure global visual coherence.

Layout Refinement. To increase the model's flexibility while rendering the output image, we optionally multiply the latent-to-segment distribution induced by the layout S with a trainable *sigmoidal gate* $\sigma(g(S, X, W))$, which considers the layout S, the evolving image features X, and

Figure 7: **Conditional generation.** Sample images generated by GANformer2 conditioned on input layouts.

the style latents W. This allows the model to locally adjust the modulation strength as it sees fit, to refine the pre-computed layout during the final image synthesis.

Semantic-Matching Loss. To promote strcutrual and semantic consistency between the layout S and image X produced by each of the synthesis stages, we complement the standard fidelity loss $\mathcal{L}(D(\cdot))$ with two new objectives: Semantic Matching and Segment Fidelity. We note that the relation between a given layout and the set of possible images that could correspond to it is a **one-to-many relation**. To encourage it, we incorporate into the discriminator a U-Net [62] to predict the semantic segmentation S' of each image, compare it using cross-entropy to the source layout $\mathcal{L}_{SM}(S, S')$, and add this loss term to the discriminator (over real samples) and to the generator (over fake ones).

This stands in a **stark contrast to the reconstruction perceptual or feature-matching losses** commonly used in conditional layout-to-image approaches [14, 35, 60], which, unjustifiably, compare the appearance of a newly generated image X to that of a natural source image X' (mathematically, through either $\mathcal{L}_P(X, X')$ or $\mathcal{L}_{FM}(f(X), f(X'))$), even though there is no reason for them to be similar in terms of style and texture, beyond sharing the same layout. Consequently, those losses severely and fundamentally hinder the diversity of the generated scenes (as shown in section 4). We circumvent this deficiency by comparing instead the layout implied by the new image with the source layout it is meant to follow, thereby staying true to the original objective of the translation task.

Segment-Fidelity Loss. To promote the semantic alignment even further, and encourage fidelity not only of the whole scene, but also of each segment it is composed of, we introduce the *Segment-Fidelity (SF) loss*. Given the layout S and the generated image X, the discriminator first concatenates and processes them together through a shared convolution-downsampling stem, to reduce their resolution. Then it partitions the image according to the layout and assesses the fidelity of each of the n segments (s_i, x_i) independently, through $\frac{1}{n}\sum \mathcal{L}_{SF}(D(s_i, x_i))$. Using the stem both reduces the image dimension before splitting it up to improve computational efficiency, and further allows for local information about the contextual surroundings of each segment to propagate into its representation. This improves over the stricter patchGAN [35] which blindly splits the image into a fixed tabular grid, while our new loss offers a more natural choice, dividing it up into semantically-meaningful regions.

Table 1: **Unconditional approaches comparison.** We compare the models on generating images from scratch, using the FID and Precision/Recall metrics for evaluation, computed over 50k sample images. COCO_p refers to a mean score over a partitioned version of the dataset, created to improve samples' visual quality (Supp. H).

	CLEVR			**Bedrooms**			**FFHQ**		
Model	FID ↓	Precision	Recall	FID ↓	Precision	Recall	FID ↓	Precision	Recall
GAN	25.02	21.77	16.76	12.16	52.17	13.63	13.18	67.15	17.64
k-GAN	28.29	22.93	18.43	69.90	28.71	3.45	61.14	50.51	0.49
SAGAN	26.04	30.09	15.16	14.06	54.82	7.26	16.21	64.84	12.26
StyleGAN2	16.05	28.41	23.22	11.53	51.69	19.42	10.83	68.61	25.45
VQGAN	32.60	46.55	63.33	59.63	55.24	28.00	63.12	67.01	29.67
SBGAN	35.13	48.12	64.41	48.75	56.14	31.02	24.52	58.32	8.17
GANformer	9.17	47.55	66.63	6.51	57.41	29.71	**7.42**	**68.77**	5.76
GANformer2	**4.70**	**64.18**	**67.03**	**6.05**	**60.93**	**37.15**	7.77	61.54	**34.45**

	COCO			**COCO_p**			**Cityscapes**		
Model	FID ↓	Precision	Recall	FID ↓	Precision	Recall	FID ↓	Precision	Recall
GAN	41.00	48.57	7.06	37.53	51.02	9.30	11.56	**61.09**	15.30
k-GAN	63.51	42.06	5.42	61.05	45.37	7.04	51.08	18.80	1.73
SAGAN	46.09	50.00	7.47	42.05	51.71	7.84	12.81	43.48	7.97
StyleGAN2	26.79	50.92	23.30	25.24	52.93	24.48	8.35	59.35	27.82
VQGAN	63.12	53.05	27.22	65.20	**55.42**	30.12	173.8	30.74	43.00
SBGAN	108.1	42.53	17.53	104.32	44.43	19.12	54.92	52.12	25.08
GANformer	25.24	**53.68**	12.31	23.81	55.29	14.48	**5.76**	48.06	33.65
GANformer2	**21.58**	49.12	**29.03**	**20.41**	51.07	**32.84**	6.21	56.12	**54.18**

4 Experiments

We investigate GANformer2's properties through a suite of quantitative and qualitative experiments. As shown in **section 4.1**, the model attains state-of-the-art performance, successfully generating images of high quality, diversity and semantic consistency, over multiple challenging datasets of highly-structured simulated and real-world scenes. In **section 4.2**, we compare the execution stage to leading layout-to-image approaches, demonstrating significantly larger output variation, possibly thanks to the new structural loss functions we incorporate into the network. Further analysis conducted in **sections 4.3 and 4.4** reveals several favorable properties that GANformer2 possesses, specifically of enhanced interpretabilty and strong spatial disentanglement, which enables object-level manipulation. In the **supplementary**, we provide additional visualizations of style and structure disentanglement as well as ablation and variation studies, to shed more light upon the model behavior and operation. Taken altogether, the evaluation offers solid evidence for the effectiveness of our approach in modeling compositional scenes in a robust, controllable and interpretable manner.

4.1 State-of-the-Art Comparison

We compare our model with both baselines as well as leading approaches for visual synthesis. Unconditional models include a baseline GAN [23], StyleGAN2 [42], Self-Attention GAN [81], the autoregressive VQGAN [21], the layerwise k-GAN [69], the two-stage SBGAN [5] and the original GANformer. Conditional approaches include SBGAN, BicycleGAN [86], Pix2PixHD [72], and SPADE [60]. We implement the unconditional methods within our public GANformer codebase and use the authors' official implementations for the conditional ones. All models have been trained with resolution of 256×256 and for an equal number of training steps, roughly spanning 10 days on a single V100 GPU per model. See Supp. G for description of baselines and competing approaches, implementation details, hyperparameter settings, data preparations and training configuration.

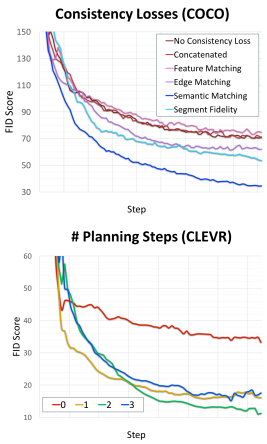

Figure 8: **Learning Curves**

As tables 1 and 3 show, the GANformer2 model outperforms prior work along most metrics and datasets in both the conditional and unconditional settings. It achieves substantial gains in terms of FID score, which correlates with visual quality [29], as is likewise reflected through the good Precision scores. Notable also are the large improvements along Recall, in the unconditional and especially the conditional settings, which indicate wider coverage of the natural image distribution. We note that the performance gains are highest for the CLEVR [37] and the COCO [48] datasets, both focusing on varied arrangements of multiple objects and high structural variance. These results serve as a strong evidence for the particular aptitude of GANformer2 for compositional scenes.

Table 2: **Disentanglement & Controllability Comparison** over sample CLEVR images, using the DCI metrics for latent-space disentanglement and our correlation-based metrics for controllability (see section 4.4).

Model	Disentanglement	Modularity	Completeness	Informativeness	Informativeness'	Object-$\rho \downarrow$	Property-$\rho \downarrow$
GAN	0.126	0.631	0.071	0.583	0.434	0.72	0.85
StyleGAN	0.208	0.703	0.124	0.685	0.332	0.64	0.49
GANformer	0.768	0.952	0.270	0.972	0.963	0.38	0.45
MONet	0.821	0.912	0.349	0.955	0.946	0.29	0.37
Iodine	0.784	**0.948**	0.382	0.941	0.937	0.27	0.35
GANformer2	**0.852**	0.946	**0.413**	**0.974**	**0.965**	**0.19**	**0.32**

4.2 Scene Diversity & Semantic Consistency

Beyond visual fidelity, we study the models under the potentially competing aims of content variability on the one hand and semantic consistency on the other. To assess consistency, we compare the input layouts S with those implied by the synthesized images, using standard metrics of mean IoU, pixel accuracy and ARI. Following prior works [60, 86], the inferred layouts S' are obtained by a pre-trained segmentation model that we use for evaluation purposes. Then, to measure diversity, we sample $n = 20$ images X_i for each input layout S, and compute the mean LPIPS pairwise distance [82] over image pairs X_i, X_j. Ideally, we expect a good model to achieve strong alignment between the input layout and the one derived from the output synthesized image, while still maintaining high variance between samples that are conditioned on the same source layout.

We see in table 3 that GANformer2 produces images of significantly better diversity, as reflected both through the LPIPS and Recall scores. Other conditional approaches such as Pix2Pix, BicycleGAN and SPADE that rely on pixel-wise perceptual or feature matching, reach high precision at the expense of considerably reduced variability, revealing the weakness of those loss functions. We further note that our model achieves better consistency than prior work, especially for LSUN-Bedrooms and COCO. These results demonstrate the benefits of our new structural losses (section 3.3), as is also corroborated by the model's learning efficiency when evaluated with different objectives (figure 8). We compare the Semantic-Matching and Segment-Fidelity losses with several alternatives (Supp. C), and observe improved performance and accelerated learning when using the new objectives, possibly thanks to the local semantic alignment and finer-detail fidelity they respectively encourage.

4.3 Transparency & Interpretability

An advantage of our model compared to traditional GANs, is that we incorporate explicit structure into the visual generative process, which both makes it more explainable, while providing means for controlling individual scene elements (section 4.4). In figure 3, we see how GANformer2 gradually constructs the image layout, recurrently adding new object segments into the scene. The gains of iterative over non-iterative generation are also demonstrated by figure 8 and Supp. figure 14. We note that different datasets benefit from a different number of recurrent synthesis steps, reflecting their respective degree of structural compositionality (Supp. D). Figure 2 further shows that the model develops an internal notion for segments' depth, which aids it in determining the most effective order to compile them together. Remarkably, no explicit information is given to the network about the depth of entities within the scene, and it rather learns it on its own in an unsupervised manner, through the indirect need to find an order that will enable creating feasible and compelling layouts.

4.4 Controllability & Disentanglement

The iterative nature of GANformer2 and its compositional latent space enhance its disentanglement along two main dimensions: **spatially**, enabling independent control over chosen objects, and **semantically**, separating between structure and style – the former is considered during planning while the latter at the execution. Figure 6 features qualitative examples where we modify an object of interest, changing its color, material, shape, or position, all without negatively impacting its surroundings. Figures 4 and 10-12 (in Supp.) further illustrate how GANformer2 can produce images of diverse textures, styles and color schemes, while still conforming to a source layout. Finally, as illustrated by figure 3, we can even add or remove objects to the synthesized scene, thanks to the sequential nature of the synthesis.

To quantify our model's degree of disentanglement, we refer to the classic DCI and modularity metrics [17, 83], which measure the extant to which there is 1-to-1 correspondence between latent

Table 3: **Conditional approaches comparison.** We compare conditional generative models that map an input layout into an output image. Multiple metrics are used for different purposes: FID, Inception (IS) and Precision scores for visual quality, class-weighted mean IoU , Pixel Accuracy (pAcc) and Adjusted Rand Index (ARI) for layout – image consistency, Recall for coverage of the natural image distribution, and LPIPS for measuring the output visual diversity given an input layout. All metrics are computed over 50k samples per model.

	CLEVR							
Model	FID ↓	IS ↑	Precision ↑	Recall ↑	mIoU ↑	pAcc ↑	ARI ↑	LPIPS ↑
Pix2PixHD	4.88	2.19	72.49	60.40	98.18	99.06	97.90	6e-8
SPADE	5.74	2.23	70.25	61.96	**98.30**	99.12	98.01	0.12
BicycleGAN	35.62	2.29	6.27	10.11	88.15	92.78	90.21	0.10
SBGAN	18.02	2.22	52.29	48.20	95.27	97.81	92.34	8e-7
GANformer2	**0.99**	**2.34**	**73.26**	**78.38**	97.83	**99.15**	**98.46**	**0.19**
	Bedrooms							
Model	FID ↓	IS ↑	Precision ↑	Recall ↑	mIoU ↑	pAcc ↑	ARI ↑	LPIPS ↑
Pix2PixHD	32.44	2.27	70.46	2.41	64.54	76.62	75.48	2e-8
SPADE	17.06	2.33	**80.20**	7.04	70.27	80.63	79.44	0.07
BicycleGAN	23.34	2.52	49.68	10.45	58.47	70.76	72.73	0.22
SBGAN	29.35	2.28	56.21	7.93	65.31	71.39	73.24	3e-8
GANformer2	**5.84**	**2.55**	54.92	**41.94**	**79.06**	**86.62**	**83.00**	**0.50**
	CelebA							
Model	FID ↓	IS ↑	Precision ↑	Recall ↑	mIoU ↑	pAcc ↑	ARI ↑	LPIPS ↑
Pix2PixHD	54.47	2.44	57.48	0.73	68.04	80.18	63.75	4e-9
SPADE	33.42	2.35	**85.08**	2.72	67.99	80.11	63.80	0.07
BicycleGAN	56.56	2.53	52.93	2.36	66.95	79.41	62.50	0.15
SBGAN	50.92	2.44	59.28	3.62	67.42	80.12	61.42	4e-9
GANformer2	**6.87**	**3.17**	57.32	**37.06**	**68.88**	**80.66**	**64.65**	**0.38**
	Cityscapes							
Model	FID ↓	IS ↑	Precision ↑	Recall ↑	mIoU ↑	pAcc ↑	ARI ↑	LPIPS ↑
Pix2PixHD	47.74	1.51	50.53	0.34	73.20	83.49	78.48	3e-8
SPADE	22.98	1.46	83.32	9.55	75.96	85.25	**80.74**	3e-3
BicycleGAN	25.02	1.56	63.01	9.49	67.24	78.43	72.20	7e-4
SBGAN	45.82	1.52	51.38	5.39	72.04	81.90	79.32	4e-8
GANformer2	**5.95**	**1.63**	**84.12**	**37.72**	**76.11**	**85.58**	80.12	**0.27**
	COCO							
Model	FID ↓	IS ↑	Precision ↑	Recall ↑	mIoU ↑	pAcc ↑	ARI ↑	LPIPS ↑
Pix2PixHD	18.91	17.97	**80.91**	16.35	58.75	71.94	68.20	1e-8
SPADE	22.80	16.29	79.00	10.63	55.72	69.55	67.73	0.19
BicycleGAN	69.88	9.16	40.05	5.83	33.58	47.98	56.55	0.18
SBGAN	45.94	14.31	58.93	10.27	47.82	54.58	62.29	4e-8
GANformer2	**17.39**	**18.01**	70.20	**24.86**	**74.93**	**84.37**	**76.41**	**0.38**

factors and image attributes. Following the protocol in [33, 75], we compute them over a sample set of synthesized CLEVR images, using a pre-trained object detector to extract and summarize their visual and semantic attributes. To further assess the level of controllability, we perform random perturbations over the latent variables used to generate the CLEVR scenes, and estimate the mean correlation between the resulting semantic changes among object and property pairs. Intuitively, we expect a disentangled representation to yield uncorrelated changes among objects and attributes. As shown in table 2, GANformer2 attains higher disentanglement and stronger controllability than leading GAN-based and variational approaches, quantitatively confirming the efficacy of our approach.

5 Conclusion

In this paper, we have tackled the task of visual synthesis, and introduced the GANformer2 model, an iterative transformer for transparent and controllable scene generation. Similar to how a person may approach the creative task, it first sketches a high-level outline and only then fills in the details. By equipping the network with a strong explicit structure, we endow it with the capacity to plan ahead its generation process, capture the dependencies and interactions between objects within the scene, and reason about them directly, moving a step closer towards compositional generative modeling. A limitation of our work is its reliance on training layouts, but we believe it could be mitigated by future work to integrate the generative efforts with unsupervised object discovery and scene understanding. Further exploration into disentanglement and means to encourage it may also be fruitful. From a broader perspective, enhancing properties of interpretability and controllability can make content generation more reliable, accessible, and easy to interact with, for the benefit of artists, graphic designers, the film industry and even the general society, and may open the door for new and promising avenues and opportunities.

6 Acknowledgements

We performed the experiments for the paper on AWS cloud, thanks to Stanford HAI credit award. Drew A. Hudson (Dor) is a PhD student at Stanford University and C. Lawrence Zitnick is a research scientist at Facebook FAIR.

Dor wishes to thank Prof. Christopher D. Manning for providing the kind support towards her PhD that allowed this work to happen!

References

[1] Titas Anciukevicius, Christoph H Lampert, and Paul Henderson. Object-centric image generation with factored depths, locations, and appearances. *arXiv preprint arXiv:2004.00642*, 2020.

[2] Jacob Andreas, Dan Klein, and Sergey Levine. Modular multitask reinforcement learning with policy sketches. In *International Conference on Machine Learning*, pp. 166–175. PMLR, 2017.

[3] Relja Arandjelović and Andrew Zisserman. Object discovery with a copy-pasting GAN. *arXiv preprint arXiv:1905.11369*, 2019.

[4] Oron Ashual and Lior Wolf. Specifying object attributes and relations in interactive scene generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4561–4569, 2019.

[5] Samaneh Azadi, Michael Tschannen, Eric Tzeng, Sylvain Gelly, Trevor Darrell, and Mario Lucic. Semantic bottleneck scene generation. *arXiv preprint arXiv:1911.11357*, 2019.

[6] Samaneh Azadi, Deepak Pathak, Sayna Ebrahimi, and Trevor Darrell. Compositional GAN: Learning image-conditional binary composition. *International Journal of Computer Vision*, 128 (10):2570–2585, 2020.

[7] Amir Bar, Roei Herzig, Xiaolong Wang, Anna Rohrbach, Gal Chechik, Trevor Darrell, and Amir Globerson. Compositional video synthesis with action graphs. *arXiv preprint arXiv:2006.15327*, 2020.

[8] David Bau, Hendrik Strobelt, William Peebles, Jonas Wulff, Bolei Zhou, Jun-Yan Zhu, and Antonio Torralba. Semantic photo manipulation with a generative image prior. *arXiv preprint arXiv:2005.07727*, 2020.

[9] David Bau, Alex Andonian, Audrey Cui, YeonHwan Park, Ali Jahanian, Aude Oliva, and Antonio Torralba. Paint by word. *arXiv preprint arXiv:2103.10951*, 2021.

[10] Irving Biederman. Recognition-by-components: a theory of human image understanding. *Psychological review*, 94(2):115, 1987.

[11] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *7th International Conference on Learning Representations, ICLR*, 2019.

[12] Christopher P Burgess, Loic Matthey, Nicholas Watters, Rishabh Kabra, Irina Higgins, Matt Botvinick, and Alexander Lerchner. MONet: Unsupervised scene decomposition and representation. *arXiv preprint arXiv:1901.11390*, 2019.

[13] Lucy Chai, Jonas Wulff, and Phillip Isola. Using latent space regression to analyze and leverage compositionality in GANs. *arXiv preprint arXiv:2103.10426*, 2021.

[14] Qifeng Chen and Vladlen Koltun. Photographic image synthesis with cascaded refinement networks. In *Proceedings of the IEEE international conference on computer vision*, pp. 1511–1520, 2017.

[15] Yunjey Choi, Min-Je Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pp. 8789–8797. IEEE Computer Society, 2018. doi: 10.1109/CVPR.2018.00916.

[16] Zhiwei Deng, Jiacheng Chen, Yifang Fu, and Greg Mori. Probabilistic neural programmed networks for scene generation. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pp. 4032–4042, 2018.

[17] Cian Eastwood and Christopher KI Williams. A framework for the quantitative evaluation of disentangled representations. In *International Conference on Learning Representations*, 2018.

[18] Sébastien Ehrhardt, Oliver Groth, Aron Monszpart, Martin Engelcke, Ingmar Posner, Niloy J. Mitra, and Andrea Vedaldi. RELATE: physically plausible multi-object scene synthesis using structured latent spaces. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[19] Martin Engelcke, Adam R. Kosiorek, Oiwi Parker Jones, and Ingmar Posner. GENESIS: generative scene inference and sampling with object-centric latent representations. In *8th International Conference on Learning Representations, ICLR*, 2020.

[20] S. M. Ali Eslami, Nicolas Heess, Theophane Weber, Yuval Tassa, David Szepesvari, Koray Kavukcuoglu, and Geoffrey E. Hinton. Attend, infer, repeat: Fast scene understanding with generative models. In Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems*, pp. 3225–3233, 2016.

[21] Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis. *arXiv preprint arXiv:2012.09841*, 2020.

[22] Angela Fan, Mike Lewis, and Yann Dauphin. Hierarchical neural story generation. *arXiv preprint arXiv:1805.04833*, 2018.

[23] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *arXiv preprint arXiv:1406.2661*, 2014.

[24] Klaus Greff, Raphaël Lopez Kaufman, Rishabh Kabra, Nick Watters, Christopher Burgess, Daniel Zoran, Loic Matthey, Matthew Botvinick, and Alexander Lerchner. Multi-object representation learning with iterative variational inference. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2424–2433. PMLR, 2019.

[25] Karol Gregor, Ivo Danihelka, Alex Graves, Danilo Jimenez Rezende, and Daan Wierstra. DRAW: A recurrent neural network for image generation. In Francis R. Bach and David M. Blei (eds.), *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pp. 1462–1471. JMLR.org, 2015.

[26] Fanny Guérin, Bernadette Ska, and Sylvie Belleville. Cognitive processing of drawing abilities. *Brain and Cognition*, 40(3):464–478, 1999.

[27] Fanny Guérin, Bernadette Ska, and Sylvie Belleville. Cognitive processing of drawing abilities. *Brain and Cognition*, 40(3):464–478, 1999.

[28] Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. GANspace: Discovering interpretable GAN controls. *arXiv preprint arXiv:2004.02546*, 2020.

[29] 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 Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 6626–6637, 2017.

[30] Tobias Hinz, Stefan Heinrich, and Stefan Wermter. Generating multiple objects at spatially distinct locations. *arXiv preprint arXiv:1901.00686*, 2019.

[31] Seunghoon Hong, Dingdong Yang, Jongwook Choi, and Honglak Lee. Inferring semantic layout for hierarchical text-to-image synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7986–7994, 2018.

[32] Drew A Hudson and Christopher D Manning. Learning by abstraction: The neural state machine. *arXiv preprint arXiv:1907.03950*, 2019.

[33] Drew A Hudson and C. Lawrence Zitnick. Generative adversarial transformers. *Proceedings of the 38th International Conference on Machine Learning, ICML*, 2021.

[34] Daniel Jiwoong Im, Chris Dongjoo Kim, Hui Jiang, and Roland Memisevic. Generating images with recurrent adversarial networks. *arXiv preprint arXiv:1602.05110*, 2016.

[35] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pp. 5967–5976. IEEE Computer Society, 2017. doi: 10.1109/CVPR.2017.632.

[36] Ali Jahanian, Lucy Chai, and Phillip Isola. On the" steerability" of generative adversarial networks. *arXiv preprint arXiv:1907.07171*, 2019.

[37] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross B. Girshick. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pp. 1988–1997, 2017.

[38] Justin Johnson, Agrim Gupta, and Li Fei-Fei. Image generation from scene graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1219–1228, 2018.

[39] Asako Kanezaki. Unsupervised image segmentation by backpropagation. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 1543–1547. IEEE, 2018.

[40] Andrea Kantrowitz. The man behind the curtain: what cognitive science reveals about drawing. *Journal of Aesthetic Education*, 46(1):1–14, 2012.

[41] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *6th International Conference on Learning Representations, ICLR*, 2018.

[42] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of styleGAN. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pp. 8107–8116. IEEE, 2020. doi: 10.1109/CVPR42600.2020.00813.

[43] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9404–9413, 2019.

[44] D Lane. Drawing and sketching: Understanding the complexity of paper–pencil interactions within technology education. *Handbook of technology education. Springer international publishing AG*, pp. 385–402, 2018.

[45] Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew P. Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[46] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. MaskGAN: Towards diverse and interactive facial image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5549–5558, 2020.

[47] Chen-Hsuan Lin, Ersin Yumer, Oliver Wang, Eli Shechtman, and Simon Lucey. St-GAN: Spatial transformer generative adversarial networks for image compositing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9455–9464, 2018.

[48] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

[49] Zhixuan Lin, Yi-Fu Wu, Skand Vishwanath Peri, Weihao Sun, Gautam Singh, Fei Deng, Jindong Jiang, and Sungjin Ahn. SPACE: unsupervised object-oriented scene representation via spatial attention and decomposition. In *8th International Conference on Learning Representations, ICLR*, 2020.

[50] Xihui Liu, Guojun Yin, Jing Shao, Xiaogang Wang, and Hongsheng Li. Learning to predict layout-to-image conditional convolutions for semantic image synthesis. *arXiv preprint arXiv:1910.06809*, 2019.

[51] Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. *arXiv preprint arXiv:2006.15055*, 2020.

[52] Catarina Silva Martins. From scribbles to details. *A Political Sociology of Educational Knowledge: Studies of Exclusions and Difference*, pp. 103, 2017.

[53] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.

[54] Li Nanbo, Cian Eastwood, and Robert Fisher. Learning object-centric representations of multi-object scenes from multiple views. *Advances in Neural Information Processing Systems*, 33, 2020.

[55] Charlie Nash, SM Ali Eslami, Chris Burgess, Irina Higgins, Daniel Zoran, Theophane Weber, and Peter Battaglia. The multi-entity variational autoencoder. In *NIPS Workshops*, 2017.

[56] Thu Nguyen-Phuoc, Christian Richardt, Long Mai, Yong-Liang Yang, and Niloy Mitra. BlockGAN: Learning 3d object-aware scene representations from unlabelled images. *arXiv preprint arXiv:2002.08988*, 2020.

[57] Michael Niemeyer and Andreas Geiger. Giraffe: Representing scenes as compositional generative neural feature fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11453–11464, 2021.

[58] Takeshi Okada and Kentaro Ishibashi. Imitation, inspiration, and creation: Cognitive process of creative drawing by copying others' artworks. *Cognitive science*, 41(7):1804–1837, 2017.

[59] Dim P Papadopoulos, Youssef Tamaazousti, Ferda Ofli, Ingmar Weber, and Antonio Torralba. How to make a pizza: Learning a compositional layer-based GAN model. In *proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8002–8011, 2019.

[60] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 2337–2346. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPR.2019.00244.

[61] Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron C. Courville. Film: Visual reasoning with a general conditioning layer. In Sheila A. McIlraith and Kilian Q. Weinberger (eds.), *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pp. 3942–3951, 2018.

[62] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

[63] RA Salome and BE Moore. The five stages of development in children's art, 2010.

[64] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of GANs for semantic face editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9243–9252, 2020.

[65] Casper Kaae Sønderby, Jose Caballero, Lucas Theis, Wenzhe Shi, and Ferenc Huszár. Amortised map inference for image super-resolution. *arXiv preprint arXiv:1610.04490*, 2016.

[66] Elizabeth S Spelke and Katherine D Kinzler. Core knowledge. *Developmental science*, 10(1): 89–96, 2007.

[67] Bob Steele. *Draw me a story: An illustrated exploration of drawing-as-language*. Portage & Main Press, 1998.

[68] Vadim Sushko, Edgar Schönfeld, Dan Zhang, Juergen Gall, Bernt Schiele, and Anna Khoreva. You only need adversarial supervision for semantic image synthesis. *arXiv preprint arXiv:2012.04781*, 2020.

[69] Sjoerd van Steenkiste, Karol Kurach, Jürgen Schmidhuber, and Sylvain Gelly. Investigating object compositionality in generative adversarial networks. *Neural Networks*, 130:309–325, 2020.

[70] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems*, pp. 5998–6008, 2017.

[71] Andrey Voynov and Artem Babenko. Unsupervised discovery of interpretable directions in the GAN latent space. In *International Conference on Machine Learning*, pp. 9786–9796. PMLR, 2020.

[72] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8798–8807, 2018.

[73] Xiaolong Wang and Abhinav Gupta. Generative image modeling using style and structure adversarial networks. In *European conference on computer vision*, pp. 318–335. Springer, 2016.

[74] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019.

[75] Zongze Wu, Dani Lischinski, and Eli Shechtman. StyleSpace analysis: Disentangled controls for StyleGAN image generation. *arXiv preprint arXiv:2011.12799*, 2020.

[76] Kun Xu, Haoyu Liang, Jun Zhu, Hang Su, and Bo Zhang. Deep structured generative models. *arXiv preprint arXiv:1807.03877*, 2018.

[77] 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.

[78] Denis Yarats and Mike Lewis. Hierarchical text generation and planning for strategic dialogue. In *International Conference on Machine Learning*, pp. 5591–5599. PMLR, 2018.

[79] Fangneng Zhan, Jiaxing Huang, and Shijian Lu. Hierarchy composition GAN for high-fidelity image synthesis.

[80] 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 *Proceedings of the IEEE international conference on computer vision*, pp. 5907–5915, 2017.

[81] Han Zhang, Ian J. Goodfellow, Dimitris N. Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pp. 7354–7363. PMLR, 2019.

[82] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018.

[83] Bo Zhao, Bo Chang, Zequn Jie, and Leonid Sigal. Modular generative adversarial networks. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 150–165, 2018.

[84] Haitian Zheng, Zhe Lin, Jingwan Lu, Scott Cohen, Jianming Zhang, Ning Xu, and Jiebo Luo. Semantic layout manipulation with high-resolution sparse attention. *arXiv preprint arXiv:2012.07288*, 2020.

[85] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *IEEE International Conference on Computer Vision, ICCV*, pp. 2242–2251, 2017.

[86] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. *arXiv preprint arXiv:1711.11586*, 2017.

[87] Peihao Zhu, Rameen Abdal, Yipeng Qin, and Peter Wonka. SEAN: image synthesis with semantic region-adaptive normalization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, pp. 5103–5112. IEEE, 2020.