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 *sketching* phase, where we plan a high-level scene layout, followed by an attention-based *painting* phase, where the layout is evolving into a rich and detailed output image. GANformer2 moves away from conventional black-box GAN architectures that feature a flat monolithic latent space, towards a transparent design that encourages efficiency, controllability and interpretability. We demonstrate the model'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 image quality, consistency and diversity. Further quantitative and qualitative experiments illustrate the additional merits of our approach in terms of adaptability, versatility and disentanglement, and provide a deeper insight into its generation 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.

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 global layout, and is then gradually refined and embellished with color, depth and richer details, until a vivid picture eventually emerges. These 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 proportions, convey shapes and geometry, and even imply semantic information [51, 61, 65]. A large body of psychology research highlights the importance of sketching in stimulating creative discovery [38, 56], fostering cognitive development [25], and facilitating problem solving [24, 43]. In fact, a large variety of generative and RL tasks either in the visual [5], textual [20, 76] or embodied modalities [2], can benefit from the same strategy – have a high-level plan first, and then focus on the details.

At the heart of this hierarchical strategy stands the principle of compositionality. 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, from physical and dynamical to abstract and semantic [9, 64]. As pointed out by prior literature [22, 50, 54], compositional representations are indeed at the core of human intelligence, supporting our capabilities of reasoning [30], imagining [6], planning [52] and learning [53]. Realizing this principle and explicitly capturing the different objects and entities composing the scene, is thereby a desirable goal, that can turn the generative modeling to be more explainable, controllable, generalizable and efficient.

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

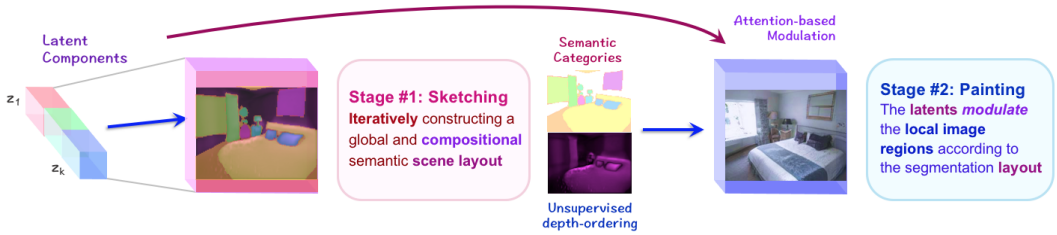

Figure 1: **An overview of the GANformer2 architecture.** The compositional model has a set of latent variables to control the various scene elements. Its generation proceeds through two stages, **sketching** and **painting** – first planning a high-level layout of the scene structure and arrangement, and then refining the layout and translating it into a photo-realistic image.

Yet, the vast majority of generative models do not explicitly 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 sole centralized objects, faces 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, 49, 74]. Likewise, while there is some evidence for property disentanglement within their latent space to independent axes of variation at the global scale [26, 34, 62, 70], existing approaches fail to provide controllability at the level of the individual object or local region. The overall understanding of the mechanisms by which these models give rise to the output image, and the transparency and adaptability of their synthesis process, thereby remain rather limited.

Motivated to alleviate these shortcomings and make generative modeling more adjustable, interpretable and compositional, we propose GANformer2, a structured object-oriented network that decouples the visual synthesis task into two steps: **sketching** and **painting**. First, a parameter-lightweight transformer creates a schematic layout – a scene representation that depicts the objects' shapes, positions, semantic classes and unsupervised depth-ordering. The layout construction occurs in an iterative manner, so as to capture the dependencies and relations between different objects. Once the it is ready, we move to the next stage, where a second transformer translates the layout into an output image, using it as a guidance for its operation while supporting its adaptive attention-based refinement during the process. This approach marks a shift towards a more natural form of image formation, in which the dependencies between the scene elements are explicitly considered, as an 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 layout-conditional and unconditional synthesis, measured along several metrics including fidelity, diversity and semantic consistency, and as demonstrated throughmultiple challenging datasets from CLEVR [35] to indoor and outdoor scenes of LSUN-Bedrooms [77] Cityscapes [15] up to the rich COCO dataset [47]. Further analysis illustrates the capacity of the model to manipulate particular scene objects in an independent manner, achieving both spatial disentanglement as well as separation between style and structure. For the CLEVR dataset, we even see indications for potentially developing capability for amodal completion. Meanwhile, inspection of the produced layouts, depth maps, and transformer attentions shed more light upon the process behind each generated image, substantiating the model's transparency and explainability.

Overall, our contributions are summarized as follows: (1) We introduce a new iterative transformer for compositional scene generation. (2) Propose a novel but simple noise scheme fir synthesizing discrete panoptic layouts, while integrating depth-awareness in a 2.5D-representation. (3) Present new purely generative structural loss functions to encourage segment fidelity and semantic layout consistency. Overall, by integrating a strong structure into the model, we can successfully move towards multiple favorable properties: robustness for diverse visual domains, controllability over individual objects, and interpretability of the generative process.

2 Related Work

In the last years, the field of visual synthesis witnessed astonishing progress driven by the emergence of Generative Adversarial Networks [21]. Tremendous strides have been made in a wide variety of

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

tasks, from image [10, 40] to video generation [67], to super-resolution [44], style transfer [14] and translation [33]. Meanwhile, a growing body of literature has explored object decomposition and sequential scene generation, albeit in the context of Variational Autoencoders [1, 17, 48, 54]. DRAW [23] uses recurrent attention to encode and decode MNIST images, while AIR [18], MONet [11] and IODINE [22] reconstruct simple simulated 3d scenes by iteratively factorizing them into semantically meaningful segments. We draw inspiration from this pioneering line of works and marry its merits with the scalability and robustness of GANs, yielding a compositional model that successfully creates real-world scenes, while offering a transparent and controllable synthesis, conceptually meant to resemble a more natural emulation of how a human might approach the generative task.

To this end, we leverage the GANformer [31], a new architecture that we believe heads in the right direction, producing diverse scene using transformers' soft-attention [69]. In our work, we take it a step further, and show how we can use the GANformer as a backbone for a structured generation, extending it to enable not only an implicit, but also an *explicit modeling* of compositional scenes. Our **iterative** design endows the model with the flexibility to consider the correlations and interactions between scene elements, circumventing the need for independence assumptions made by earlier works [16, 55, 68] which hence fail to translate to natural scenes. Concurrently, the object-centered structure enhances the model's **compositionallity** so to outperform past coarse-to-fine cascade [79], layered [32], or 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 in limited domains, such as indoor NYUv2 [72], faces [39], pizzas [57], or bird [75] photographs.

Our model links to research on conditional generative modeling [71, 83, 85], including semantic image synthesis and text-to-image generation. The seminal Pix2pix [33] and Spade [58] works adopt either a U-Net [60] or spatially-adaptive normalization [59] to translate a given input layout into an output high-resolution image. However, beyond their notable weakness of being conditional and thereby 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 [66, 84] (see section 3.3.1 for details). In our work, we mitigate this issue by proposing novel *Semantic-Matching* and *Segment-Fidelity* losses which maintain the one-to-many relation that holds between layouts and images, to enhance output diversity.

While GANformer2 is an unconditional model, it shares some high-level ideas with text-to-image [28, 29] and scene-to-image [4, 36] techniques, of using a semantic layout as a bottleneck intermediate representation, but whereas these works focus at conditional heavily-engineered pipelines consisting

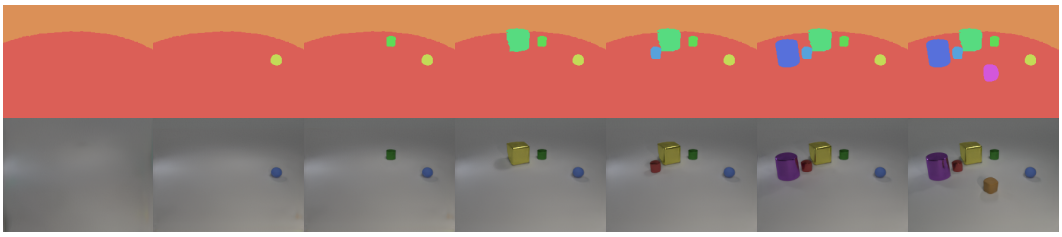

Figure 3: **Iterative scene generation.** GANformer2 creates the layout sequentially, segment-by-segment, to capture the scene compositionality. Each segment has matching style and structure latents, used while translating the layout into the final image, effectively allowing us to add or remove objects from the resulting scene.

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 panoptic segmentations only) to widen its applicability across visual domains. Finally, our approach relates to prior works on visual editing [7, 8, 45, 82], and image compositing [3, 12, 46, 78], but while these methods modify an existing source image, GANformer2 stands out being capable of creating manipulatable and structured scenes from scratch.

3 The GANformer2 model

GANformer2 is a compositional object-centric transformer for visual scene generation. It proceeds through two stages: (1) parameter lightweight sequential **sketching stage**, where a scene layout – a set of compounded object segments, is being created (section 3.2),followed by a parallel (2) **painting stage**, where the layout is further refined and translated into an output photo-realistic image (section 3.3). Both stages are implemented using customized versions of the recently proposed GANformer (section 3.1) – a bipartite transformer that uses a set of latent variables to modulate the evolving visual feature of a generated image, so to create it in a compositional fashion.

We assume having training sets of images and panotpic segmentations [42] (specifying instance and cateogry information). However, in contrast to the common practice of prior works to use scarcer hand-annotated ground-truth maps, and to demonstrate the model robustness, we instead intentionally explore training on auto-predicted segmentations (either produced in an unsupervised manner [37] for CLEVR [35] or by pre-trained segmentor [73] otherwise). The images and layouts can either be paired or unpaired, and our training scheme accommodates either option.

3.1 Generative Adversarial Transformers

The GANformer [31] is a transformer-based Generative Adversarial Network: it consists of a *generator* that maps a partitioned latent $Z^{k \times d} = [z_1, ..., z_k]$ into an image $X^{(HW) \times d}$, and a *discriminator* 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 feed-forward and self-attention layers. Then, it sets an initial 4×4 image grid, which gradually evolves through convolution and upsampling layers up to the final resolution. However, in contrast to the traditional, the GANformer also incorporates *bipratite transformer* layers into the generator, which compute attention between the k latents and the image features, to support semantic and spatial region-wise modulation:

$$\text{Attention}(X, Z) = \text{softmax}\left(\frac{q(X)k(Z)^T}{\sqrt{d}}\right) v(Z)$$
$$u(X, Z) = \gamma_a(X, Z) \odot LayerNorm(X) + \beta_a(X, Z)$$

Where the equations respectively express the attention and modulation between the latents and the image; $q(\cdot), k(\cdot), v(\cdot)$ are the queries, keys, and values mapping; and $\gamma(\cdot), \beta(\cdot)$ compute multiplicative and additive styles (gain and bias) as a function of Attention(X, Z). The update rule lets the latents to shape and impact the image regions as induced by the key-value attention computed between them, encouraging visual compositionality, where different latents specialize to represent and control various visual objects or elements within the generated scene.

Figure 4: **GANformer2 style variations.** for CLEVR and LSUN-Bedrooms. Images are produced by varying the style latents of the objects while keeping the structure latents fixed.

For each of the two stages, we make key modifications and adjustments to the GANformer architecture, and couple it with novel loss functions, as discussed below. One key point in which we notably depart from the original GANformer regards to the number of latent variables used by the model. While originally the GANformer employs a fixed number k of latent variables, we allow for a variable normally-distributed number of latents (capped by a set maximum) so to make the model more flexible in generating scenes of diverse complexity.

3.2 Sketching Stage – Layout Planning

In the first stage, we create 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 segment s_i is associated with (1) a spatial distribution over the image grid $H \times W$ specifying the segment's shape and position (2) a distribution reflecting the segment's category (semantic class), and (3) per-pixel values $d_i(x, y)$ indicating its depth-ordering, relative to other segments, so to support object occlusion in case of segments overlap. For instance, if there is one segment i of a cube and another segment j of a sphere in front of it, then $d_j(x, y) > d_i(x, y)$ within the overlap region $\{(x, y)\}$. Notably, we do not provide supervision to the model about the true depths of the scene elements, and it rather creates its own internal depth-ordering in an unsupervised fashion, indirectly supervised by the need to produce realistic looking layouts when compounding the segments together.

3.2.1 Iterative Layout Generation

To enable object-wise controllability and enhance compositionality on the one hand, while capturing the object dependencies on the other, we construct the scene layout in an iterative fashion, by recurrently applying a fast parameter-light GANformer model over T timesteps (empirically set in the range of 2–4). At each step t, we generate a variable number of segments $k_t \sim \mathcal{N}(\mu, \sigma^2)$ using a trainable normal distribution[1], gradually aggregated into a scene layout S_t. To generate the segments, we use the GANformer with the same number of k_t latent variables z_i, each corresponding to a segment to be formed. The GANformer then uses the latents to guide the feature modulation of the synthesized layout, following the formulation presented above.

However, to make the GANformer recurrent, and condition its computation upon the aggregated segmentation generated in prior steps, we change the *key function* $k(\cdot)$ to consider not only the currently generated layout S_t (as in the standard GANformer), but also the prior aggregated layout S_{t-1} (by simply concatenating them together and passing them to the key function). This way, the attention-based modulation accounts for both the new segments being formed as well as the prior ones, allowing the model to capture dependencies between segments from different steps. Meanwhile, this construction supports the option to manipulate posteriori a segment s_i, by modifying its corresponding latent z_i, achieving a favorable balance between expressive and contextual modeling of relational interactions on the one hand and object disentanglement on the other. Finally, in its output layer, instead of producing RGB values, we modify the GANformer to output probability distributions that reflect the segment identity, category and depths at each spatial position.

[1]Segments are grouped and created together for computational efficiency in generating crowded scenes.

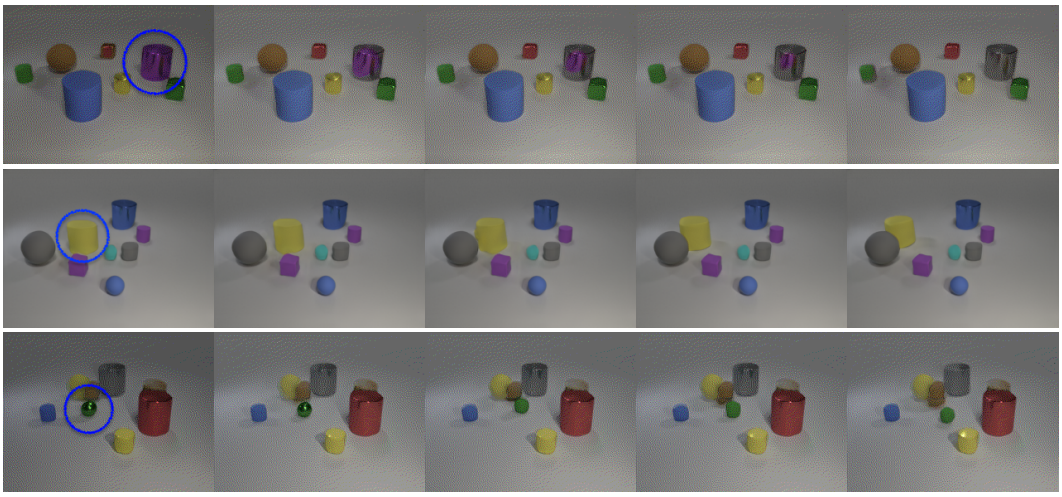

Figure 5: **Object controllability**. GANformer2 supports entity-level manipulation, and separates style from structure, enabling control over color, shape and position of particular chosen objects. At each row we start by generating a scene (leftmost image) and then gradually interpolate the latent vectors of one object only. In the first row we modify its style vector which leads to a color change, in the second row we modify its structure vectors, resulting in a position change, and in the third row we modify both its style and structure, yielding a change both in the object's position and material.

3.2.2 Layout Composition

The segments s_i are gradually merged into the generated scene layout S_t, up to the final layout S with k segments overall. We can stitch the segments by overlaying them according to their depth ordering $d_i(x, y)$, specifically by computing *softmax* over their depths: $\text{Softmax}(d(x, y)_i)_{i=1}^{k}$, yielding for each image position (x, y) a distribution over the segments s_i, which results overall in a 2D panoptic segmentation map. Notably, this layout is soft, consisting of probability distributions over segment classes and identities, making the sketching stage fully differentiable.

3.2.3 Discriminator Noise Scheme

We train the sketching stage over a set of panoptic segmentations [42], which are inherently discrete, while the generated segmentations, as mentioned above, are soft and continuous. To maintain a stable GAN training under these conditions, and inspired by the instance noise technique [63], we add a simple *hierarchical noise* map to the layouts given to the discriminator, which helps ensuring its task remains challenging. We define the noise as $\sum_{r=6}^{8} Upsample(n_r^{2^r \times 2^r})$, summing noise grids of multiple resolutions, sampled from $\mathcal{N}(0,\, \sigma^2)$, allowing for noise to be partly correlated between adjacent positions, empirically improving its effectiveness. See supplementary for details.

3.3 Painting Stage – Translation & Refinement

In the second stage, we proceed by translating and refining the layout we compiled earlier into a photo-realistic image. For each segment, we sample a latent y_i to encode its texture and style, and then pass it along with the mean depth value and class distribution of each segment to a second GANformer to produce the scene image. As opposed to the original GANformer that computes attention $A(X, Y)$ on-the-fly to determine the assignment between latents and spatial regions, here we replace the computed *softmax* by using the pre-determined soft layout as a conditional guidance source instead, such that each latent modulates the region of its corresponding pre-assigned segment, leading to a more explicit semantically-meaningful association between latents and scene elements.

At the same time, to increase the model's flexibility and leave room for layout refinement, we multiply the latent-to-segment distribution, induced by our layout, with the output of a trainable **sigmoidal gating mechanism** $\sigma(g(X, Y))$, allowing the GANformer to adjust the modulation strength online

Figure 6: Samples of images generated by GANformer2 when conditioned on an input layout.

as it sees fit, as determined by a function $g(\cdot)$ of both the evolving image features X and the style latent Y. In the supplementary, we empirically validate the effectiveness of this extension.

3.3.1 Objective Functions for Semantic Consistency

To encourage semantic consistency between the layout compiled by the sketching phase and the image produced by the painting stage, we complement the standard fidelity loss by two new objective functions: Semantic-Matching and Segment-Fidelity, used during the painting stage. 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 [60] to predict a semantic segmentation given a generated image, compare it using cross-entropy to the source layout, and finally add this **Semantic-Matching (SM) loss** term both to the generator and the discriminator.

This stands in a stark contrast to the typical reconstruction feature-matching or perceptual losses commonly used in conditional layout-to-image approaches [13, 33, 58], which, unjustifiably, compare the appearance of a newly generated image to that of a natural source image, even though there is no reason for these 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. We avoid this deficiency by comparing instead the layout implied by the new image with the original layout that the image is meant to follow, thereby staying true to objective of the translation task.

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 and the generated image, the discriminator first concatenates them together, processes them through a shared convolutional-downsampling stem, and then partitions the image according to the layout and asses the fidelity of each segment independently. The stem both reduces the image dimension before splitting it up for computational efficiency, and further allows for local information about the contextual surroundings to propagate into each segment. This improves over the stricter

Table 1: **Unconditional Approaches Comparison.** We compare the models on generating images from scratch, using the FID and Precision/Recall metrics for evaluation, averaged over 10 random with 50k sample images. COCO' refers to a mean score over a partitioned version of the dataset. See supplementary material for details.

	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	12.85	**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'**			**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	65.2	**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**

patchGAN [33] which blindly splits the image into a fixed tabular grid, while the SF loss offers a more natural choice, dividing it up into semantically-meaningful regions.

3.4 Model Training

To build the complete end-to-end GANformer2 model, we chain the two stages together: first performing the sketching, unconditionally mapping sampled latent variables $z_1, ..., z_k$ to a schematic layout, then proceeding to the conditional painting and translating the layout into a final output image. In terms of training, each stage is practically a separated generator-discriminator pair. We therefore first train each stage independently, and then fine-tune them jointly. While this procedure assumes a paired dataset of layouts and images (to train the conditional stage), we note that the model could in fact be seamlessly trained in the unpaired case too, by training the two connected stages of the model jointly from scratch. See supplementary for further details about its efficacy in training stabilization.

4 Experiments

We investigate GANformer2's properties through a suite of quantitative and qualitative experiments. As we will see, the model attains state-of-the-art performance, successfully generating images of high quality, diversity and semantic consistency, over multiple datasets of challenging simulated and real-world scenes. In section 4.1, we study the diversity of the images by the model, and compare it especially to conditional layout-to-image approaches, demonstrating larger output variation, possibly thanks to the new structural loss functions we incorporate into the network. Further analysis conducted in section 4.2 reveals several favorable properties that GANformer2 posses, including enhanced interpretabilty and stronger spatial disentanglement, allowing for object-level manipulation. In the Supplementary, we provide additional ablation and variation studies, visualizations of style and structure disentanglement, and an exploration of the model's internal latent space, to shed more light upon its behavior and operation. Taken altogether, the evaluation offers solid evidence for the effectiveness of our approach in modeling compsitional scenes in an interpretable manner.

Baselines. We compare our model with multiple approaches including both baselines and leading models for image synthesis: First we inspect unconditional methods which synthesize images from scratch: (1) A baseline GAN [21], (2) The StyleGAN2 [41] model, (3) SAGAN [80] which utilizes self-attention, (4) k-GAN [68] that blends together k generated images through alpha-composition, (5) VQGAN [19], an autoregessive transformer image autoencoder, and (6) SBGAN [5], a two-stages image synthesis approach. We then proceed to compare our painting stage to leading conditional semantic generation approaches including both the aforementioned SBGAN as well as (7) Pix2PixHD [71] that uses a U-Net [60] to translate source to target images, (8) BicycleGAN [84] that leverages cycle consistency for conditional generation, and finally (9) SPADE [58], which performs spatial modualtion using a fixed set of trainable semantic category vectors.

Experimental settings. We implement all the unconditinal methods within our public GANformer codebase [31]. This ensures the models are tested under comparable conditions in terms of training details, model sizes, and optimization scheme. For the conditional models we use the authors' official implementations, likewise implemented as extensions to the Pix2PixHD repository for conditional generative modeling. All approaches have been trained with images of 256×256 resolution and for the same number of training steps, roughly spanning 10 days on 1 NVIDIA V100 GPUs per model. As mentioned in section 3.4, we assume access to a training set of image and panoptic segmentations [42], indicating the segment unique identity and its semantic class. Contrary to prior conditional works which rely on costly ground-truth segmentations, we instead use auto-predicted segmentations predicted by a pre-trained segmentor [73]. We will publicly release both our code and processed data for the benefit of the community at `https://github.com/dorarad/gansformer`. See supplementary for implementation details, hyperparameter settings, training configuration and training set preparations and statistics.

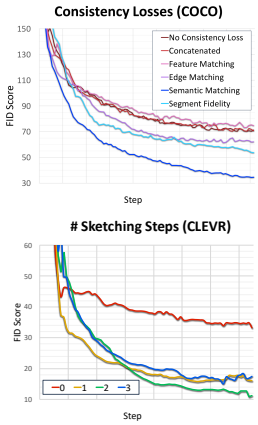

Figure 7: Learning Curves

Results. As shown in tables 1 and 2, the GANformer2 model in most cases matches or outperforms prior work in the conditional and unconditional settings respectively. It achieves substantial gains in terms of FID score, which correlates with image quality [27], as well as the Precision and Recall (P&R). We note that the performance boost is significantly large for the CLEVR dataset [35] which focuses on compositional multi-object scenes, obtaining 4.70 in the unconditional case and the strong 0.98 FID score when conditioning on a ground-truth layout. This result serves as an indication for the particular aptitude of the model for compositional scenes.

4.1 Scene Diversity & Semantic Consistency

Beyond image fidelity, we study the models further under the poentially competing aims of content*diversity* on the one hand and semantic *consistency* to its source layout on the other. To measure these qualities, we follow prior works [58, 84], and use mean IoU and pixel Accuracy between the source layout and the one implied from the generated scene to measure their alignment. For image diversity, we use the LPIPS [81] mean paired-distance metric, to assess the image variability when conditioned on the same layout. We see in table 2 that GANformer2 generates images of a significantly better diversity, as is reflected both by the LPIPS and the Recall scores. Other conditional approaches such as Pix2Pix, BicycleGAN and SPADE, which rely on either pixel-wise perceptual or feature matching, reach strong precision at the expense of significantly reduced diversity, revealing the weakness of these loss functions. We further see that our model performs oftentimes better in terms of visual consistency compared to prior approaches. A possible explanation for these results is the new structured losses we use as discussed earlier. This is indeed corroborated by the model's learning curves when evaluated with different objective functions, shown in figure 7. We compare our proposed Semantic-Matching and Segment-Fidelity losses with several alternatives, defined in the supplementary, and observe improved performance and faster learning when in our approach.

4.2 Transparency & Controllability

An advantage of our approach compared to traditional GANs is that we incorporate explicit structure into the visual generative process, which both turns it more explainable, while providing means for controlling individual scene elements. In figure 3, we can see how the model proceeds as it is gradually constructing the scene layout, adding segments and element iteratively. We see in figure 2that it also computes a proxy internal representation for "depth" that helps the model in determining the most effective manner to compile the different segments together. Interestingly and remarkably, no explicit information is given to the network about the depth of the entities in the scene, and it impressively learns that on its own, in an unsupervised manner, through the indirect need to find a segments ordering that will support feasible and compelling output layouts.

Figure 6 shows how we can modify an object of interest, changing its color, material, shape, or position, all without negatively impacting its surroundings. The model's iterative synthesis and the fact it allocates different latent variables to control and create each of the scene segments, endow it with the ability to manipulate particular objects and properties, and explicitly separate structure from

Table 2: **Conditional Approaches Comparison.** We compare conditional models that generate an output image given an input layout. Multiple metrics are used for different purposes: FID and Precision for the visual quality, mIoU (class-weighted) and pAcc (pixel accuracy) for semantic consistency between the layouts and image, Recall for the coverage of the natural image distribution, and LPIPS for output visual diversity given an input layout. All metrics are computed over 50k samples per model.

	CLEVR					
Model	FID ↓	Precision ↑	Recall ↑	mIoU ↑	pAcc ↑	LPIPS ↑
Pix2PixHD	4.88	72.49	60.40	**97.63**	98.29	3e-8
SPADE	5.74	70.25	61.96	79.91	84.78	0.08
BicycleGAN	35.62	6.27	10.11	91.24	94.62	0.07
SBGAN	18.02	52.29	48.20	83.74	82.91	8e-7
GANformer2	**0.98**	**73.26**	**78.38**	97.39	**98.60**	**0.13**
	Bedrooms					
Model	FID ↓	Precision ↑	Recall ↑	mIoU ↑	pAcc ↑	LPIPS ↑
Pix2PixHD	32.44	70.46	2.41	31.23	38.53	2e-8
SPADE	17.06	**80.20**	7.04	32.11	39.47	0.05
BicycleGAN	23.34	49.68	10.45	30.98	37.21	0.17
SBGAN	29.35	56.21	7.93	29.25	36.94	3e-8
GANformer2	**5.84**	54.92	**41.94**	**32.69**	**40.12**	**0.42**
	CelebA					
Model	FID ↓	Precision ↑	Recall ↑	mIoU ↑	pAcc ↑	LPIPS ↑
Pix2PixHD	54.47	57.48	0.73	35.57	52.40	4e-9
SPADE	33.42	**85.08**	2.72	34.70	49.33	0.07
BicycleGAN	56.56	52.93	2.36	33.43	51.22	0.15
SBGAN	50.92	59.28	3.62	32.49	47.93	4e-9
GANformer2	**6.87**	57.32	**37.06**	**35.65**	**53.36**	**0.38**
	Cityscapes					
Model	FID ↓	Precision ↑	Recall ↑	mIoU ↑	pAcc ↑	LPIPS ↑
Pix2PixHD	47.74	50.53	0.34	86.24	91.49	2e-8
SPADE	23.98	83.32	9.55	68.48	74.53	2e-3
BicycleGAN	24.02	63.01	9.49	83.49	89.13	3e-8
SBGAN	45.82	51.38	5.39	87.92	90.53	4e-8
GANformer2	**6.87**	**84.12**	**57.32**	**89.58**	**92.84**	**0.17**
	COCO					
Model	FID ↓	Precision ↑	Recall ↑	mIoU ↑	pAcc ↑	LPIPS ↑
Pix2PixHD	18.91	**80.91**	16.35	29.19	36.54	1e-8
SPADE	22.79	79.00	10.63	27.69	31.98	0.15
BicycleGAN	69.88	40.05	5.83	28.69	35.25	0.15
SBGAN	45.94	58.93	10.27	26.38	34.74	4e-8
GANformer2	**17.39**	70.2	**24.86**	**30.84**	**37.15**	**0.29**

style (the former is considered in the sketching phase while the latter in the painting). We can even add or remove entities, illustrating how the model performs *amodal completion* of occluded objects.

5 Discussion & Conclusion

In this paper, we have tackled the task of visual synthesis, and introduced the GANformer2 model, an iterative and compositional transformer for transparent and controllable scene generation, the draws inspiration from the way humans think and commonly approach creative tasks – first sketching a high-level plan, and then filling in the details. By equipping the network with a strong sequential structure, we allow the model to plan ahead its synthesis process, capture the dependencies and interactions between the objects within the scene, and reason about them explicitly, moving a step closer towards compositional generative modeling. One limitation of our work is its reliance on training layouts, but we believe that this could be mitigated by future work that will integrate the generative efforts with unsupervised object discovery and scene understanding. From a broader perspective, the field of visual synthesis has potentially far-reaching societal impact, positive and negative. It could aid graphic designers and artists in their creative work, and even help the common user, improving human-machine interaction. Reliable, explainable and controllable techniques for content generation, as those we seek to find in this work, could substantially affect the movie and video industries, among others. While, as most technologies, visual generation could also be used for negative goals, mainly through the synthesis of fake or unauthentic content, we believe that turning the process to be more transparent and explainable will help ameliorate the risk, and open up the door for positive avenues and opportunities.

6 Acknowledgements & Funding Disclosure

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

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] 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.

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

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

[10] 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, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.

[11] 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.

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

[13] 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.

[14] 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.

[15] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pp. 3213–3223. IEEE Computer Society, 2016. doi: 10.1109/CVPR.2016.350.

[16] 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.

[17] 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, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

[18] 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 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pp. 3225–3233, 2016.

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

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

[21] 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.

[22] 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.

[23] 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.

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

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

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

[27] 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.

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

[29] 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.

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

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

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

[33] 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.

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

[35] 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 *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pp. 1988–1997. IEEE Computer Society, 2017. doi: 10.1109/CVPR.2017.215.

[36] 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.

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

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

[39] 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, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

[40] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 4401–4410. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPR.2019.00453.

[41] 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.

[42] 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.

[43] 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.

[44] 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 *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pp. 105–114. IEEE Computer Society, 2017. doi: 10.1109/CVPR.2017.19.

[45] 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.

[46] 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.

[47] 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.

[48] 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, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

[49] 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.

[50] 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.

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

[52] 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.

[53] 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.

[54] 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.

[55] 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.

[56] 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.

[57] 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.

[58] 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.

[59] 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, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pp. 3942–3951. AAAI Press, 2018.

[60] 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.

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

[62] 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.

[63] 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.

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

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

[66] 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.

[67] Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. Mocogan: Decomposing motion and content for video generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1526–1535, 2018.

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

[69] 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 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 5998–6008, 2017.

[70] 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.

[71] 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.

[72] 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.

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

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

[75] 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.

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

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

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

[79] 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.

[80] 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.

[81] 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.

[82] 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.

[83] 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 2017, Venice, Italy, October 22-29, 2017*, pp. 2242–2251. IEEE Computer Society, 2017. doi: 10.1109/ICCV.2017.244.

[84] 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.

[85] Peihao Zhu, Rameen Abdal, Yipeng Qin, and Peter Wonka. SEAN: image synthesis with semantic region-adaptive normalization. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pp. 5103–5112. IEEE, 2020. doi: 10.1109/CVPR42600.2020.00515.