Scene Graph to Image Synthesis via Knowledge Consensus

Yang Wu[1], **Pengxu Wei**[1,*], **Liang Lin**[1,2]

[1] School of Computer Science and Engineering, Sun Yat-sen University
[2] Key Laboratory of Information Security Technology, GuangDong Province
wuyang36@mail2.sysu.edu.cn, weipx3@mail.sysu.edu.cn, linliang@ieee.org

Abstract

In this paper, we study graph-to-image generation conditioned exclusively on scene graphs, in which we seek to disentangle the veiled semantics between knowledge graphs and images. While most existing research resorts to laborious auxiliary information such as object layouts or segmentation masks, it is also of interest to unveil the generalization of the model with limited supervision, avoiding extra cross-modal alignments. To tackle this challenge, we delve into the *causality* of the adversarial generation process, and reason out a new principle to realize a simultaneous semantic disentanglement with an alignment on target and model distributions. This principle is named *knowledge consensus*, which explicitly describes a triangle causal dependency among observed images, graph semantics and hidden visual representations. The consensus also determines a new graph-to-image generation framework, carried on several adversarial optimization objectives. Extensive experimental results demonstrate that, even conditioned only on scene graphs, our model surprisingly achieves superior performance on semantics-aware image generation, without losing the competence on manipulating the generation through knowledge graphs.

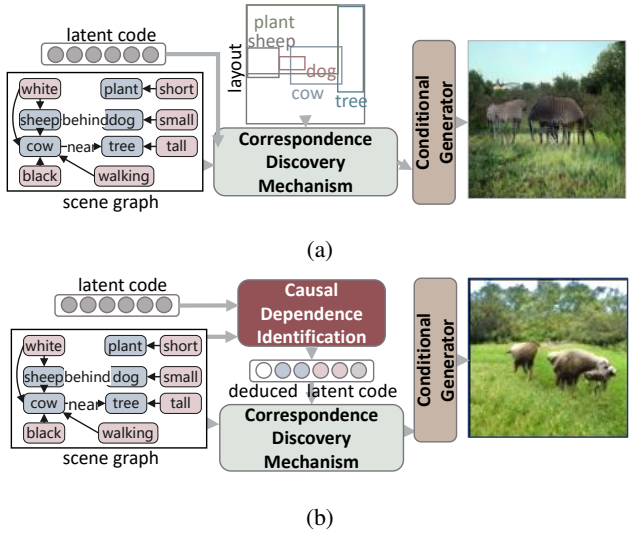


Figure 1: A comparison between (a) a typical graph-to-image generation algorithm that utilizes layout information as auxiliary, and (b) our framework.

Introduction

Conditional image generation (Isola et al. 2017) has gained popularity in computer vision for its ability of making generative modeling more controllable, as well as its potential of cognitively understanding the visual world. Previous works, for the most part, put their efforts into incorporating the condition with variety of scene descriptions, such as natural language instructions (Tan, Feng, and Ordonez 2019), bounding boxes (Zhao et al. 2019; Sun and Wu 2019; Talavera et al. 2019), semantic segmentations (Reed et al. 2016; Li et al. 2019), scene graphs (Johnson et al. 2018; Herzig et al. 2020), and many more. Although it is often reasonable to combine as many types of conditions as possible, there is also substantial motivation to use as little conditions as possible. The reasons may include but not limited to: 1) in real world scenarios, it is not always guaranteed that all the conditions required by our model are accessible; 2) controlling only one or two variables is also more user-friendly when performing human-computer interaction generation; 3) introducing extra supervision is rather than an "addition", but instead a "multiplication", since it might increase the complexity of the model.

In this paper, we follow the work of (Mittal et al. 2019) and (Herzig et al. 2020) to synthesize images conditioning only on scene graphs, since the scene graphs usually embody objects and their relationships simultaneously (compared to bounding boxes or other similar layout information), and is easier to manipulate than languages and segmentations. However, this apparently poses the challenge of only relying on the relationship between latent space and semantic factors in scene graphs, when producing images with precise semantic consistency. To mitigate this challenge, (Mittal et al. 2019) and (Herzig et al. 2020) build the relationship with relatively weak correspondence due to the ignorance of semantic consistency, so that they usually fail on the synthesis quality compared with others with extra supervision.

Distinct from those methods, our goal is to take full advantage of the information embedded in the scene graph and

*Corresponding Author

build a strong relationship with hidden space to achieve a semantic consistency from sideways. Causal inference (Pearl et al. 2009) is one of the most powerful tools that addresses our needs. Indeed, only with the help of scene graphs, causal inference has the notable feature that can help us identify the *cause and effect* relationships among graphs and their hidden space. The causal relationship is crucial for generative models, as pointed out by (Kocaoglu et al. 2018), since with a well-developed causal model, we can utilize GAN models to learn the interventional and observational distribution accorded with the generator.

With this discovery, we manage to achieve our goal by mounting causal dependence to existing graph-based generative models, such as Seg2Img from (Johnson et al. 2018) and Layout2Im from (Zhao et al. 2019). Both these works utilize segmented masks or layouts that determine the object locations in images to construct a direct mapping from latent space to graph, which can be identified as a correspondence discovery mechanism. To achieve a relatively similar and strong correspondence, our causal dependency identification manages to disentangled the random latent space to be consistent with the graph semantics. For a clear picture of our idea, we depict the difference between our framework and a typical example in Figure 1. As a consequence of using multiple auxiliary layout information, the correspondence (green box) of a regular method (in the top figure) is built directly on graph knowledge and latent code, whereas our framework (in the bottom figure), without any other conditions but scene graphs, intends to achieve the correspondence by building a causal dependence (red box) between graph knowledge and latent code.

Another challenge lies on the fact that causal dependence on generation alone might disrupt the data balance of adversarial learning, as claimed by (Mariani et al. 2018), the data imbalance between the generator and discriminator might cause the mode collapse. To avoid such risk, a thorough consideration on the whole generative framework is imperative. We thus analyse all the causal relationships of relevant variables, including the knowledge graph K, latent variable Z, real data X and synthetic data $\hat{X}$, and form a causal model described by the directed diagram. With this causal diagram, we come up with a "knowledge consensus" optimization principle that mounts the information of causal dependence on both the generative and the discriminative sides. We name this generic graph-to-image generation framework as *Knowledge Consensus Generative Model* (KCGM).

Our KCGM follows the knowledge consensus principle by integrating three components: 1) A *Graphic Information Disentanglement* (GID) module for splitting Z into separate latent codes, and subsequently, mapping them into different graphic components specified by K, which can be regarded as a direct realization of causal dependence identification; 2) A *Structured Knowledge Encoder* (SKE) module for describing the causal relation between K, real data X and synthetic data $\hat{X}$, which facilitates the information transition between the generative and discriminative model; 3) A GAN-based generation backbone for building the causal relations from Z to $\hat{X}$, X and $\hat{X}$ to K.

Our contributions are summarized as follows:

(I) Under our setting, we have proposed a novel graph-to-image generation framework that mounts causal dependency into the generative learning model, being able to generate realistic scene images *without* any auxiliary supervision.

(II) We have derived a special knowledge consensus optimization strategy by endowing the traditional adversarial learning with a directed causal diagram. The optimization guarantees an asymptotic approximation of semantic consistency that prevents the learning procedure from collapsing.

(III) Our overall model, KCGM, quantitatively achieves obvious gains on inception score and Fréchet inception distance for image generation on the Visual Genome dataset (Krishna et al. 2017), and qualitatively demonstrates superior generation capacity even for controllable image generation via graph manipulation.

Related Work

Conditional Visual Generation Ever since the invention of the conditional GAN (Mirza and Osindero 2014) that breaks through the uncontrollable problem of adversarial learning, several promising applications have arisen, such as text-to-image (Tan, Feng, and Ordonez 2019; Yin et al. 2019; Li et al. 2019), style transferring (Zhu et al. 2017), multi-scale image fusion (Jin et al. 2017), image denoising (Zhang et al. 2017) and gray photo coloring (Zhang, Isola, and Efros 2016). One of the controllable generation tasks, the graph-to-image generation (Johnson et al. 2018), recently draws plenty of attention. The definition of this generative process varies in different works (Ashual and Wolf 2019; Mittal et al. 2019; Turkoglu et al. 2019), particularly, one preferred is as an interactive process that the users compose attribute objects related to the scene. Despite that, different types of controllable generation are closely related, for instance, text-to-image (Li et al. 2019; Tan, Feng, and Ordonez 2019; Yin et al. 2019) can be transformed into graph-to-image (Deng et al. 2018).

Causal Inference Causal inference reasoning between cause and effect based on their dependencies has been comprehensively studied in (Spirtes et al. 2000; Pearl 2009). As referred by Pearl in (Pearl et al. 2009), causal analysis aims at inferring not only beliefs or probabilities under static conditions, but also the dynamics of beliefs under changing conditions. Some researchers usually make use of probabilistic graphic models (Koller and Friedman 2009; Lauritzen 1996) including generative models (Goudet et al. 2017, 2018) to implement causal relation identification. (Kocaoglu et al. 2018) reveals that GAN models can be trained to learn interventional and real distributions if the generator is accorded with a given causal graph. (Besserve et al. 2020) explores the statistics in learned latent representations for meaningful and controllable generation. (Kurutach et al. 2018) proposes a planning framework that aims at generate plausible visual plans in dynamic systems. Inspired by the spirit and effectiveness of these works, we find our way to employ causal

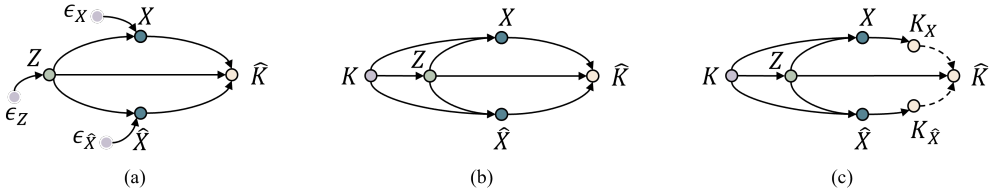


Figure 2: The evolution from (a) an initial causal diagram from to (c) our causal diagram.

relations to handle graph-to-image generation without relying on any extra layout information.

Representation Disentanglement Due to the highly entangled nature, distributed representations learned by neural models are difficult to interpret (N et al. 2017). Some works have sought to disentangle the representation by factors, in other words, the hidden spaces are trained to be semantically specified (Higgins et al. 2016; Chen et al. 2016; Karras, Laine, and Aila 2019). This strategy is proved to be beneficial for visual reasoning (Van Steenkiste et al. 2019) and controllable vision generation (Singh, Ojha, and Lee 2019; Yin et al. 2019). Related to our work, causal analysis can also be equipped to study the dependence problem among the semantic variables (Locatello et al. 2019; Yang et al. 2021), whereas they mainly investigate how to learn a generative framework that also unfolds the causality implied by the latent representations.

Assumptions and Preliminaries

Causal Relations Normally, given a real-world image X, and a fraud image $\hat{X}$, an evident causality is that a scene graph K can be recognized from X or $\hat{X}$ as an effect, denoted by $X \rightarrow K \leftarrow \hat{X}$. To avoid confusion, we represent the ground-truth scene graph as K, and the estimated scene graph as $\hat{K}$. In parallel with (Kocaoglu et al. 2018), we treat the hidden variable Z as the cause of X or $\hat{X}$, where such relation can be represented as $X \leftarrow Z \rightarrow \hat{X}$. Combining with $X \rightarrow \hat{K}$, we now have $Z \rightarrow \hat{K}$. We can also treat $Z \rightarrow \hat{K}$ as the posterior estimation of $K \rightarrow Z$. We also take outer unknown factor into consideration, also known as exogenous/unobserved factor, denoted by ϵ, that implicitly corrupts X, $\hat{X}$ or Z. These relations form the diagram in Figure 2a. Here, three assumptions are made to describe the causal discovery process and construct all the causal diagrams in Figure 2:

(I) All the functions represented by causal arrows in the causal graph is sub-optimal, namely, all the optimal function is inaccessible; All the mappings are one-to-one.

(II) We assume that each non-effect variable (*i.e.*, node with positive out-degree) is influenced by at least one exogenous factor. Furthermore, we assume that all the variables are affected by K, rather than by estimated graph $\hat{K}$.

(III) Ground-truth knowledge K is treated as the cause only in a counterfactual scenario (Pearl 2009).

We now explain these assumptions. With assumption (I), we can treat the estimated graph $\hat{K}$ as an effect of X, otherwise, ground-truth K is supposed to be the effect. With assumption (II), we can redraw the causal graph in Figure 2b from Figure 2a. We note that the current causal diagram Figure 2b is contradicted to assumption (I): There exists a function that estimates identical knowledge $\hat{K}$ from $\hat{X}$ and X, which is against to the sub-optimal assumption. To be self-consistent, we decouple the estimated knowledge graph $\hat{K}$ into three variables, $\hat{K}_X$, $\hat{K}_{\hat{X}}$ and $\hat{K}_Z$, to represent different graph distributions from three causal variables and modify the diagram from Figure 2b to Figure 2c. For simplicity, we denote $\hat{K}$ as a short hand of $\hat{K}_Z$. In fact, the cause of the real observed data X is too complicated to describe, therefore, considering ground truth K as the cause of X is merely in the counterfactual situation, as claimed by assumption (III).

Based on causal diagram Figure 2c, the objective is to minimize the divergences among the three distributions of knowledge graph $\hat{K}_X$, $\hat{K}_{\hat{X}}$ and $\hat{K}$. The optimization contains the following two processes:

(♡) $K \rightarrow Z \rightarrow \hat{K}$: the goal is to disentangle the latent variable Z into knowledge-specified factors that are consistent with K.

(♠) $\hat{K}_X \rightarrow \hat{K} \leftarrow \hat{K}_{\hat{X}}$: the goal is to make the generated samples $\hat{X}$ as realistic as the real one X by reaching consensus between estimated graph variables and ground-truth.

We call a "knowledge consensus" is satisfied if the optimal of (♠) is achieved.

Causal Model Formally, we define a causal model as $\mathcal{M}$, containing observed variables $\mathcal{V} = \{X, \hat{X}, Z, \hat{K}\}$, functional relations $\mathcal{F}$, exogenous variables $\mathcal{E} = \{\epsilon_X, \epsilon_X, \epsilon_{\hat{X}}\}$, and probability distributions $\mathcal{P}_\mathcal{E}$ over the exogenous variables. $\mathcal{M}$ is in fact a Bayesian network that induces joint probability distribution of the observable variables and their parent nodes.

Causal sufficiency (Pearl et al. 2009) suggests that every exogenous variable is a direct parent of at most one observable variable. Thus, we make Z as a parent node of X, which is a image of $X = f_x(Z, \epsilon_X)$. Similarly, let K having parent nodes X and Z, we can write $K = f_k(X, Z, \epsilon_k)$. In this sense, the causal relations illustrated in Figure 2a corresponds to the following non-parametric interpretation, with

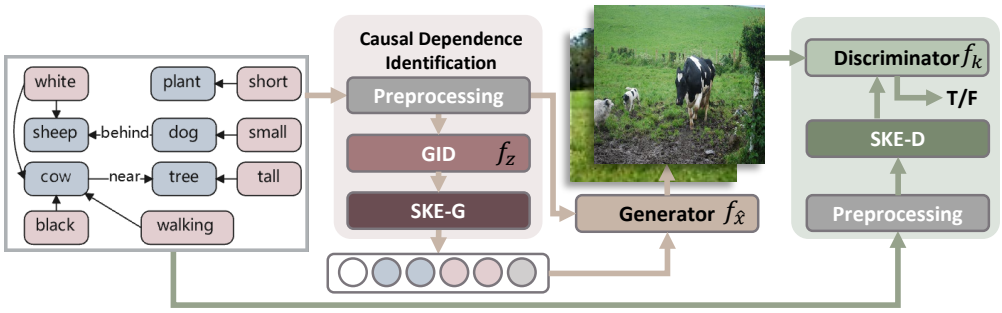


Figure 3: Overall architecture of KCGM.

each related to one of the observed variables:

$$\mathcal{M}^a = \begin{cases} X = f_x(Z, \epsilon_X), \hat{X} = f_{\hat{x}}(Z, \epsilon_{\hat{X}}), \\ Z = f_z(\epsilon_Z), \hat{K} = f_k(Z, X, \hat{X}). \end{cases} \tag{1}$$

Additionally, the exogenous variable $\hat{K}$ is beyond the consideration since $\hat{K}$ is an "effect" variable in the casual diagram based on assumption (II).

Knowledge Consensus With the causal diagram depicted in Figure 2c, we are allowed to deduce graph information from three sources (X, $\hat{X}$ and the augmented Z). Despite that, we may still face information imbalance during learning, since current causal model $\mathcal{M}^c$ cannot guarantee the inferred knowledge $\hat{K}$ to converge towards the real K. In other words, the prerequisite of (♠) is not satisfied. Fortunately, this issue can be overcame by minimizing the divergence among $\hat{K}_X$, K and $\hat{K}_{\hat{X}}$ properly, which we call this knowledge consensus optimization. The convergence of the knowledge consensus is measured by the commonly used pixel-wise *mean absolute error* (MAE):

$$\text{MAE} = \frac{1}{3n} \sum_{i=1}^{n} \left(|K^i - \hat{K}_X^i| + |K_X^i - \hat{K}_{\hat{X}}^i| + |K^i - \hat{K}_{\hat{X}}^i| \right), \tag{2}$$

where n is the total number of the samples. The reconstruction from Z to $\hat{K}$ is to minimize the divergence between K and $\hat{K}$, namely, MAE should be able to measure the divergence among $\hat{K}_X$, $\hat{K}$ and $\hat{K}_{\hat{X}}$. Further discussions are made in experiment section.

Structured Knowledge The way we specify the casual dependency $Z \to \hat{K}$ is through a structured manner. We refer to the factorization in (Singh, Ojha, and Lee 2019; Schölkopf et al. 2021), and come up with a mechanism that disentangles Z via deconstructing $K = (A, L, R, P)$. Specifically, $A \in [0, 1]^{|H| \times |H|}$ is the adjacent matrix, representing the relationships among graph nodes, where H is the set of graph vertice; L denotes the node label, which is instance-level multi-labeled in case the graph contains multiple objects; R denotes the attributes that specify the graph nodes, in a multi-label form; P is the predicate matrix with identical dimension to A.

Knowledge Consensus Generative Model

Overview The overall framework is illustrated in Figure 3, where we can see that KCGM mainly contains three components: 1) A GAN backbone for generating realistic samples from knowledge scene graphs; 2) A variational *Graphic Information Disentanglement* (GID) module for realizing causal dependency, $K \to Z \to \hat{K}$; 3) A *Structured Knowledge Encoder* (SKE) as a transition of the former two components, to fulfill knowledge consensus. In particular, our SKE is applied in both the generator and discriminator of GAN backbone (denoted as SKE-G and SKE-D respectively).

Recall the causal model $\mathcal{M}^a$ defined Eq.(1) which corresponds to the diagram in Figure 2a, the equations can then be reformulated by adding ground-truth K as a cause and applying (♠) as a restriction, according to Figure 2c:

$$\mathcal{M}^c = \begin{cases} X = f_x(Z, K, \epsilon_X), \hat{X} = f_{\hat{x}}(Z, K, \epsilon_{\hat{X}}); \\ Z = f_z(K, \epsilon_Z), \hat{K} = f_{\hat{k}}(Z); \\ \hat{K}_X = f_k(X, K), \hat{K}_{\hat{X}} = f_k(\hat{X}, K) \end{cases} \tag{3}$$

For clarity, the functions $f_{(*)}$ in Eq.(3) and the previously discussed three components can be related as follows: f_x denotes the real world sampling function; $f_{\hat{x}}$ is the generator G in GAN backbone that takes the controllable information K and latent factors Z as inputs; f_z is the structured encoding function E in SKE; f_k includes the discriminator D in GAN for extracting discriminative knowledge $\hat{K}$ and the decoder V in GAN for knowledge reconstruction from latent space. Besides, f_k serves as the main function in the knowledge consensus learning. Therefore, we can write equivalently that GAN $= (f_{\hat{x}}, f_k)$, GID $= (f_z, f_{\hat{k}})$, and we note that SKE is applied prior to $f_{\hat{x}}$ and f_k for connecting GAN and GID.

Next we will describe the three components separately.

GAN Backbone Conventional GAN models solve a minimax game W, by applying Eq.(3) to W we have our partial objective function, which is a generation process conditioned on knowledge graph label K:

$$\min_{f_{\hat{x}}} \max_{f_k} W(f_{\hat{x}}, f_k) = \underbrace{\mathbb{E}_{X \sim P_{\text{real}}}[\log f_k]}_{\mathcal{L}_D} + \underbrace{\mathbb{E}_{Z \sim \mathcal{N}(0,I)}[\log(1 - f_k(f_{\hat{x}}))]}_{\mathcal{L}_G}. \tag{4}$$

We modify f_k as an indicator function depends on whether there is a match between the sample and the graph: $f_k(X, K) = 1, f_k(\hat{X}, K) = 0$. However, the indicator f_k merely describes a partial knowledge graph distribution manifold, since it fails to discriminate two specific conditions, $f_k(\hat{X}, \cdot) = 1$ and $f_k(X, \cdot) = 0$, where '$\cdot$' could be a meaningless knowledge graph rather than the matched K. Existing methods (Odena, Olah, and Shlens 2017; Xia et al. 2018) solve this problem by adding another auxiliary classifier. However, to make the model concise, we adopt a simple modification, that the intermediate outputs, $\hat{K}_X = f_k(X, K)$ and $\hat{K}_{\hat{X}} = f_k(\hat{X}, K)$, will be compared whether identical or not. This sub-process is to minimize the KL-divergence between the distributions of K_X and $K_{\hat{X}}$, which benefits the knowledge consensus learning. Nevertheless, current optimization barely provides a constrain that guides $\hat{K}_{\hat{X}}$ to converge to $\hat{K}_X$, so that $\hat{K}_{\hat{X}}$ would have little correlation with the real knowledge graph K. With this consideration, we further introduce the other parts of the knowledge consensus learning: $\hat{K}_X \to K$, $\hat{K}_X \to K$ and $\hat{K}_{\hat{X}} \to \hat{K}_X$, which are formulated as:

$$\min_{f_{\hat{x}}, f_k} \underbrace{\mathcal{D}_{\mathrm{KL}}(P(K)\|P(\hat{K}_X))}_{\mathcal{L}_{KX}} + \underbrace{\mathcal{D}_{\mathrm{KL}}(P(K)\|P(\hat{K}_{\hat{X}}))}_{\mathcal{L}_{K\hat{X}}} + \underbrace{\mathcal{D}_{\mathrm{KL}}(P(\hat{K}_X)\|P(\hat{K}_{\hat{X}}))}_{\mathcal{L}_{X\hat{X}}}. \tag{5}$$

The final objective of knowledge consensus is to enforce independent equivalence among those variables. With this clarification, Eq.(5) can be explained by the causal relation in assumption (II). We assign each term in Eq.(5) with $\mathcal{L}_{KX}, \mathcal{L}_{K\hat{X}}$ and $\mathcal{L}_{X\hat{X}}$. To further facilitate the process of disentangled representation learning of Z, we apply InfoGAN (Chen et al. 2016) in the GAN backbone. This brings a modification to the objective: an additional term $-\lambda I(\hat{K}_X, \hat{K}_{\hat{X}})$ will be added to Eq.(4), where I computes the mutual information.

Graphic Information Disentanglement Recall that latent disentanglement is for modeling the casual dependency $K \to Z \to \hat{K}$, which involves learning functions f_z and $f_{\hat{k}}$ defined in the causal model $\mathcal{M}^c$ in Eq.(3). In practice, we disentangle the shared Z (shared with GAN) into $(Z\cdot) = (Z_a, Z_l, Z_r, Z_p)$. However, it still remains to properly map these latent codes to different graphic components (A, L, R, P) in a single network. Inspired by (Singh, Ojha, and Lee 2019; Locatello et al. 2019) and (Gao et al. 2020), for each latent code (e.g. Z_a), we apply a modified group *Variational Autoencoder* (VAE) to learn the disentanglement function f_z. This learning process resembles training a standard VAE, and separates the latent variable Z into different parts. The objective function for this module is then formulated as,

$$\min_{f_z, f_{\hat{k}}} \mathbb{E}_{K\sim P_{\mathrm{real}}}\Big[\mathbb{E}_{Z\sim f_z}[\log f_{\hat{k}}] - \mathcal{D}_{\mathrm{KL}}(f_z\|P(Z)))\Big], \tag{6}$$

where $f_{\hat{k}}$ and f_z are considered as the decoder and encoder distributions, and $P(Z) \sim \mathcal{N}(0, I)$. Note that latent codes $(Z\cdot)$ are assumed to be the couple of normal distributions with different dimensions. Furthermore, we will consider Z_ϵ in practice, the detailed relevant equations and theory will be illustrated in Appendix A.

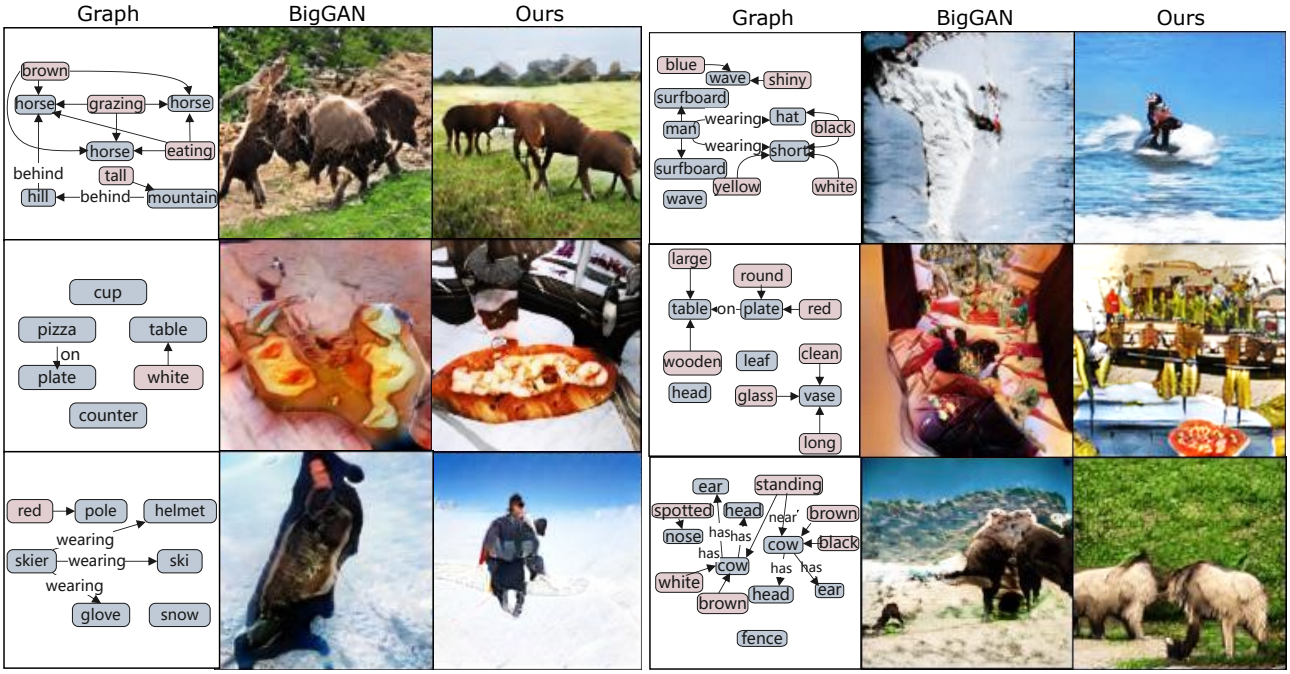
Structured Knowledge Encoding We remind that SKE is fed with graph-type data prior to the adversarial learning functions f_k and $f_{\hat{x}}$, so as to avoid data imbalance. The SKE for generator $f_{\hat{x}}$ is denoted as SKE-G, and for the discriminator f_k as SKE-D. Both SKE modules are formed by a *Graph Convolutional Network* (GCN) (Kipf and Welling 2016) for extracting features in graphic structure data. SKE-G consists of two crucial parts for *node* and *relation* features, respectively denoted as NODGCN and RELGCN. We replace the fully-connected layer of the original discriminator in GAN backbone with SKE-D, fusing the condition with visual representation by introducing another GCN module MatchGCN. See Appendix A for further details of the SKE structure.

Knowledge Consensus Optimization Our learning strategy for these three modules are provided in Appendix A, Algorithm 1, where we use C to represent all VAE models for different graph components to avoid restatement. We denote the GAN loss functions as $\mathcal{L}_G$ and $\mathcal{L}_D$; The VAE loss functions is a combination of $\mathcal{L}_{\text{relation}}$ and $\mathcal{L}_{\text{node}}$; The semantic consensus optimization losses are $\mathcal{L}_{X\hat{X}}, \mathcal{L}_{KX}, \mathcal{L}_{K\hat{X}}$ (refer to Eq.(5)). Note that the output feature $\mathbf{k}$ takes part in the computation of $\mathcal{L}_{X\hat{X}}$, $\mathcal{L}_{KX}$ and $\mathcal{L}_{K\hat{X}}$. We use lowercase letter to indicate the sample from the set, whereas the set itself is denoted by the capital letter. The parameters of each component are updated alternatively, while the parameters of other components are fixed.

Experiments

In this section, we empirically evaluate our proposed algorithm in comparison with several baseline methods. We first conduct experiments on image generation tasks conditioned on simple and complex graphs, afterwards, extend our framework to handle controllable image generation under semantic graph manipulation.

Settings We evaluate the proposed KCGM on *Visual Genome* (VG) dataset (Krishna et al. 2017). The dimension of Z_a is set to $g_M \times g_m$, where g_m and g_M are respectively the statistical minimum and maximum node number of the dataset. With this setting, when the given graph has g nodes, the rest $g_M - g$ of Z_a is non-semantic, hence we treat it as Z_ϵ. We will discuss the reasons with experiments. Note that we treat the label reconstruction as a classification task, where both the encoders and the decoders are specially designed. We apply BigGAN as the backbone network in our framework for large-scale high-resolution generation. Adam (Kingma and Ba 2015) is applied as the optimizer to learn the overall framework. The batch size during model optimization is set to be 72 for 128×128 generated image size. Each component has different learning rates and other



(a) Generations conditioned on simple graphs.

(b) Generations conditioned on complex graphs.

Figure 4: Visual examples of graph-to-image generation, where the first left column shows the input graph, and the second and third present the respective output results of the baseline model and our proposed KCGM.

hyperparameters. The GAN generator is updated every 4 iterations after the discriminator is updated. Practically, for numerically stable purpose, the KL divergence loss function in Eq.(5) is replaced by the MAE loss in Eq.(2). We pre-train the GID (including a GeVAE and β-VAEs) before jointly learning with the generative module. More details can be found in the supplementary material.

Graph-to-Image Generation The experimental results presented here are two folds: results conditioned on simple graphs with very few relations, and complex graphs with at least 3 connections.

[*Results Conditioned on Simple Graphs*] We can treat this situation as image generation conditioned on multiple labels. As an evidence in Figure 4b, the synthesized results are mainly constituted by the objects labeled in the graphs. For example, in the third row, although very subtle, the equipment of the skiers can be recognized. This experiment demonstrates the effectiveness of KCGM on generating specific objects with few labels and relations.

[*Results Conditioned on Complex Graphs*] Compared with scene generation from a simple graph, it is reasonably hard to generate results that tailor every graphic relation in a complex graph. Here we provide some complex graph generation results in Figure 4b, showing that KCGM manages to massively capture the graphic semantics endowed in the complex graph. For instance, the first sample manifests a surfing scene from both the graph and generated images.

Comparison and Ablation Study In this section, we integrally evaluate and analyse the SOTA performance of KCGM on VG dataset compared with existing approaches. The evaluation metrics we applied are the widely used *Inception Scores* (IS) and *Fréchet Inception Dis-*

Models	Seg/Img	Bbox	IS↑	FID↓
BigGAN	×	×	7.33±0.10	80.88
w/o z_ϵ	×	×	7.40±0.07	85.77
w/o SKE-GD+K	×	×	7.83±0.09	79.92
w/o GID+K	×	×	7.50±0.06	63.97
w/o SKE-D+K	×	×	8.57±0.08	55.01
w/o SKE-G+K	×	×	8.12±0.11	54.97
w/o K	×	×	9.13±0.10	55.71
KCGM	×	×	9.42±0.12	**51.27**
KCGM-T	×	×	**9.94±0.17**	57.97
Pix2pix (Isola et al. 2017)	×	√	2.7±0.02	142.86
Seg2img (Johnson et al. 2018)	√	√	6.3±0.2	74.61
WCGC (Herzig et al. 2020)	×	×	8.0±1.1	-
Layout2Im (Zhao et al. 2019)	×	√	8.1±0.1	40.07
LostGAN (Sun and Wu 2019)	×	√	11.1±0.6	29.36
OCGAN (Sylvain et al. 2021)	×	√	**12.3±0.6**	**28.6**

Table 1: Experimental comparison evaluation and ablation study on VG dataset (with label generation) for image generation. G stands for the generator, D is the discriminator, and K denotes the knowledge consensus learning. The suffix 'w/o' means 'without' certain component in KCGM for simplicity. '-T' represents two-layer GCN implementation.

Techniques	IS↑	FID↓
Raw KCGM	9.42±0.12	51.27
+Atten	9.98±0.21	46.31
+Atten+L1	10.12±0.10	38.44
+Atten+Ploss	10.17±0.14	37.76
+Atten+Ploss+L1	11.18±0.17	31.12
+Atten+Ploss+L1-T	**11.63±0.11**	**27.46**

Table 2: Results for ablation study. Ploss (Johnson, Alahi, and Fei-Fei 2016) stands for the perceptual loss, Atten (Vaswani et al. 2017) means attention mechanism, and L1 indicate the L1 pixel loss.

tance (FID) (Heusel et al. 2017). We summarize the results in Table 1 and Figure 6, where we have implemented the baselines along with BigGAN. As can be identified in Figure 6, which draws learning curves of both IS and FID, the learning curve of KCGM without the aforementioned three components (SKE, K and Z_ϵ) is more erratic throughout the training process, and sometimes even non-monotone (suggesting a collapse). Together with the figures in Table 1, showing the gap between 'w/o GID+K' and 'KCGM', demonstrates the effectiveness of disentanglement in KCGM. The quantitative results also highlight the benefits of our proposed optimization strategy and the importance of the knowledge in our framework.

Here in the ablation study, we choose BigGAN as the baseline, considering the fact that if the objects and mask predictions are completely removed, LostGAN will be reduced to a normal ResGAN with multilabels, which in turn is equivalent with our baseline BigGAN. As shown in Table 1, although LostGAN (Sun and Wu 2019) achieves better IS and FID, after removing the auxiliary layout conditions, KCGM obviously outperforms LostGAN (BigGAN). More visualizations of LostGAN without layouts are provided in Appendix B.

We also boost the performance of KCGM to be comparable to LostGAN and even to SOTA OCGAN (the latest SOTA on VG dataset, OCGAN (Sylvain et al. 2021), is built on LostGAN with auxiliary layout condition), by incorporating the pixel loss, perceptual loss and attention mechanism, in a similar way in LostGAN and other LostGAN-variations. According to the results listed in Table 2, we see that adding attention (+Atten) to the raw KCGM achieves a 0.5 increase in IS and 5 drop in FID. This means that there is still room for improving correlation in KCGM, since attention mechanism is an effective approach for aligning cross-modal features. By comparing our KCGM-T record with WCGC in Table 1, we notice that synthesized data of KCGM possess high diversity, while less similarity to the real data. We further apply pixel loss and perceptual loss to KCGM to promote the synthesis quality, and observe expected performance gain: with either pixel loss (+L1) or perceptual loss (+Ploss), KCGM respectively attains 7.9 and 8.5 decrease in FID, in the mean time, 0.2 increase in IS. It is also easy to spot that, we have significant gain on both IS and FID (IS increases to 11.18 and FID decreases to 31.12), by taking advantage of these two losses simultaneously. Eventually, we

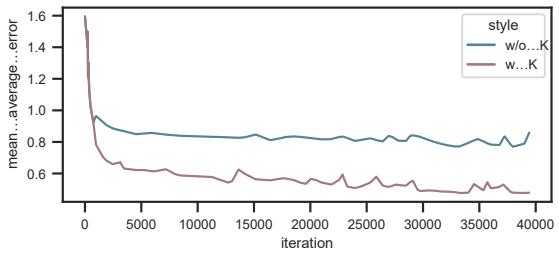
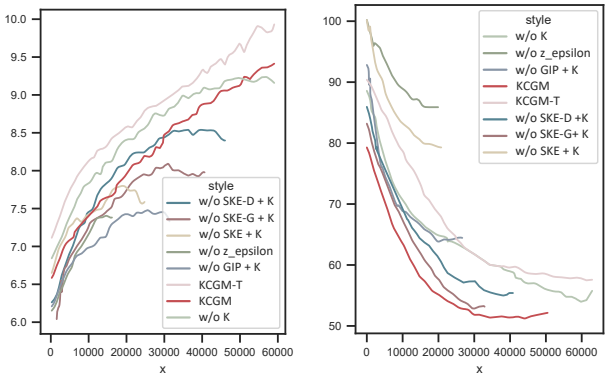


Figure 5: Evaluation on knowledge consensus learning. Two curves are the learning MAE records, where the less indicates the greater consensus achieved by the optimization. The brown curve is from an overall KCGM, and the blue is from KCGM without consensus learning loss.



(a) IS (b) FID

Figure 6: IS and FID recorded curves after 10,000 training iterations (the maximum iteration number is 65,000). The best records are listed in Table 1.

improve our performance even more by adding one more GCN layer to SKE in KCGM, since we observe in Table 1 that, implementing GCN with one more layer (KCGM-T) is beneficial.

Knowledge Consensus Verification In this experiment, we show that the "knowledge consensus" in our proposed KCGM is proved to be approximately achieved. The degree of knowledge consensus is calculated using MAE (Eq.(2)), which is the smaller the better. In Figure 5, we provide experimental ablative results of explicit knowledge consensus during the training procedure. 'w/o K' means without such explicit knowledge consensus loss, including $\mathcal{L}_{KX}$, $\mathcal{L}_{X\hat{X}}$, and $\mathcal{L}_{K\hat{X}}$. It is observed that without this loss, the MAE curve of KCGM completely flattens out by iteration 6,000 and turns unstable at iteration 30,000, whereas the MAE curve of learning with knowledge consensus remains gradually decreasing to about 0.1.

Controllable Image Generation Herein we conduct the controllable image generation via graph manipulation to further explore the validity of knowledge consensus learning, and most importantly, to verify the robustness of KCGM on achieving semantic consistency (rather than just guessing).

Specifically, we manipulate object *categories*, *attributes*, and their *relations* in the input scene graph, which are the dominant graphical information that stipulates how the images would be structured. The manipulation is strictly restricted by the principle that both the starting and ending graphs are in accordance with common sense. Otherwise, the causal relation will be counterfactual. For instance, it is allowed to modify 'pizza' with 'food' or 'vegetable', but is not allowed to replace 'food' with 'leg'.

Based on these considerations, we visualize three different types of manipulations in Appendix B, Figure 11. It can be observed that our model is able to generate desired results according to the manipulation. We highlight that manipulating in the graph will not only change the corresponding objects or their attributes, but the related objects or attributes as well. These artifacts can be seen, for example, more clouds appeared in the 'sky' after altering 'giraffe' to 'elephant'; 'elephant' morphed after changing the environment; 'parking apron' varied in color after suspending the 'airplane'. we have evaluated KCGM on counter-factual examples by the manipulation on graphs. The visualizations in Fig. 11 also show possibility for KCGM to generate counterfactuals with modified background while keeping a very similar object in the foreground. For example, given a counterfactual graph that an elephant is standing on a sandy hill, KCGM generates a matching image. These results provide evidence that our model is indeed graph-driven, rather than "memorizing" images in the dataset.

Conclusion

Graph-to-image generation is plausibly an under-explored task that demands generative modeling to guarantee the semantic consistency between images and graphs. In this work, without additional supervision of precise object detection or semantic segmentation, we explore the causal generative modeling to generate faithful images conditioned only on scene graphs. By incorporating the causal dependency, we devise a knowledge consensus generative model derived from the causal analysis in our graph-to-image generation setting. This also enables us to realize controllable generation, since our model is featured to disentangle the hidden space into factors corresponding to structured semantic components in scene graphs. This has been examined through empirical evaluations from different aspects, demonstrating that learning by reaching the "knowledge consensus" permits promising synthetic results, not to mention the latent high-level understanding of scene graphs.

Acknowledgements

This work was supported in part by National Key R&D Program of China under Grant No. 2021ZD0111601, National Natural Science Foundation of China (NSFC) under Grant No. U21A20470, U1811463, 61836012, 62006255, 61876224, 62206314, GuangDong Basic and Applied Basic Research Foundation under Grant No. 2017A030312006, 2022A1515011835. Finally, we would like to thank Xu Cai from SOC of NUS for his assistance on proofreading of the manuscript.

References

Ashual, O.; and Wolf, L. 2019. Specifying object attributes and relations in interactive scene generation. In *International Conference on Computer Vision*, 4561–4569.

Besserve, M.; Mehrjou, A.; Sun, R.; and Schölkopf, B. 2020. Counterfactuals uncover the modular structure of deep generative models. In *International Conference on Learning Representations*.

Chen, X.; Duan, Y.; Houthooft, R.; Schulman, J.; Sutskever, I.; and Abbeel, P. 2016. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in Neural Information Processing Systems*, 2172–2180.

Deng, Z.; Chen, J.; Fu, Y.; and Mori, G. 2018. Probabilistic neural programmed networks for scene generation. In *Advances in Neural Information Processing Systems*, 4028–4038.

Gao, R.; Hou, X.; Qin, J.; Chen, J.; Liu, L.; Zhu, F.; Zhang, Z.; and Shao, L. 2020. Zero-VAE-GAN: Generating Unseen Features for Generalized and Transductive Zero-Shot Learning. *IEEE Transactions on Image Processing*, 29: 3665–3680.

Goudet, O.; Kalainathan, D.; Caillou, P.; Guyon, I.; Lopez-Paz, D.; and Sebag, M. 2017. Causal Generative Neural Networks. Cite arxiv:1711.08936.

Goudet, O.; Kalainathan, D.; Caillou, P.; Guyon, I.; Lopez-Paz, D.; and Sebag, M. 2018. Learning functional causal models with generative neural networks. *Explainable and interpretable models in computer vision and machine learning*, 39–80.

Herzig, R.; Bar, A.; Xu, H.; Chechik, G.; Darrell, T.; and Globerson, A. 2020. Learning canonical representations for scene graph to image generation. In *European Conference on Computer Vision*, 210–227. Springer.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, 6629–6640.

Higgins, I.; Matthey, L.; Pal, A.; Burgess, C.; Glorot, X.; Botvinick, M.; Mohamed, S.; and Lerchner, A. 2016. beta-vae: Learning basic visual concepts with a constrained variational framework. *Urban Affairs Review*.

Isola, P.; Zhu, J.-Y.; Zhou, T.; and Efros, A. A. 2017. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1125–1134.

Jin, K. H.; McCann, M. T.; Froustey, E.; and Unser, M. 2017. Deep Convolutional Neural Network for Inverse Problems in Imaging. *IEEE Transactions on Image Processing*, 26(9): 4509–4522.

Johnson, J.; Alahi, A.; and Fei-Fei, L. 2016. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, 694–711. Springer.

Johnson, J.; Gupta, A.; Li, F.-F.; et al. 2018. Image generation from scene graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1219–1228.

Karras, T.; Laine, S.; and Aila, T. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4401–4410.

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In Bengio, Y.; and LeCun, Y., eds., *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Kipf, T. N.; and Welling, M. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Kocaoglu, M.; Snyder, C.; Dimakis, A. G.; and Vishwanath, S. 2018. CausalGAN: Learning Causal Implicit Generative Models with Adversarial Training. In *International Conference on Learning Representations*.

Koller, D.; and Friedman, N. 2009. *Probabilistic graphical models: principles and techniques*. MIT press.

Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.-J.; Shamma, D. A.; et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1): 32–73.

Kurutach, T.; Tamar, A.; Yang, G.; Russell, S. J.; and Abbeel, P. 2018. Learning Plannable Representations with Causal InfoGAN. In Bengio, S.; Wallach, H. M.; Larochelle, H.; Grauman, K.; Cesa-Bianchi, N.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, 8747–8758.

Lauritzen, S. L. 1996. *Graphical models*, volume 17. Clarendon Press.

Li, W.; Zhang, P.; Zhang, L.; Huang, Q.; He, X.; Lyu, S.; and Gao, J. 2019. Object-driven text-to-image synthesis via adversarial training. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 12174–12182.

Locatello, F.; Tschannen, M.; Bauer, S.; Rätsch, G.; Schölkopf, B.; and Bachem, O. 2019. Disentangling Factors of Variations Using Few Labels. In *International Conference on Learning Representations*.

Mariani, G.; Scheidegger, F.; Istrate, R.; Bekas, C.; and Malossi, C. 2018. Bagan: Data augmentation with balancing gan. *arXiv preprint arXiv:1803.09655*.

Mirza, M.; and Osindero, S. 2014. Conditional Generative Adversarial Nets. Cite arxiv:1411.1784.

Mittal, G.; Agrawal, S.; Agarwal, A.; Mehta, S.; and Marwah, T. 2019. Interactive Image Generation Using Scene Graphs. *CoRR*, abs/1905.03743.

N, S.; Paige, B.; van de Meent, J.-W.; Desmaison, A.; Goodman, N.; Kohli, P.; Wood, F.; and Torr, P. 2017. Learning Disentangled Representations with Semi-Supervised Deep Generative Models. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Odena, A.; Olah, C.; and Shlens, J. 2017. Conditional image synthesis with auxiliary classifier gans. In *International Conference on Machine Learning*, 2642–2651. PMLR.

Pearl, J. 2009. *Causality*. Cambridge university press.

Pearl, J.; et al. 2009. Causal inference in statistics: An overview. *Statistics surveys*, 3: 96–146.

Reed, S. E.; Akata, Z.; Mohan, S.; Tenka, S.; Schiele, B.; and Lee, H. 2016. Learning What and Where to Draw. In Lee, D.; Sugiyama, M.; Luxburg, U.; Guyon, I.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Schölkopf, B.; Locatello, F.; Bauer, S.; Ke, N. R.; Kalchbrenner, N.; Goyal, A.; and Bengio, Y. 2021. Towards Causal Representation Learning. *arXiv preprint arXiv:2102.11107*.

Singh, K. K.; Ojha, U.; and Lee, Y. J. 2019. Finegan: Unsupervised hierarchical disentanglement for fine-grained object generation and discovery. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6490–6499.

Spirtes, P.; Glymour, C. N.; Scheines, R.; and Heckerman, D. 2000. *Causation, prediction, and search*. MIT press.

Sun, W.; and Wu, T. 2019. Image synthesis from reconfigurable layout and style. In *International Conference on Computer Vision*, 10531–10540.

Sylvain, T.; Zhang, P.; Bengio, Y.; Hjelm, R. D.; and Sharma, S. 2021. Object-centric image generation from layouts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 2647–2655.

Talavera, A.; Tan, D. S.; Azcarraga, A.; and Hua, K.-L. 2019. Layout and Context Understanding for Image Synthesis with Scene Graphs. In *2019 IEEE International Conference on Image Processing (ICIP)*, 1905–1909.

Tan, F.; Feng, S.; and Ordonez, V. 2019. Text2scene: Generating compositional scenes from textual descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6710–6719.

Turkoglu, M. O.; Thong, W.; Spreeuwers, L.; and Kicanaoglu, B. 2019. A layer-based sequential framework for scene generation with gans. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 8901–8908.

Van Steenkiste, S.; Locatello, F.; Schmidhuber, J.; and Bachem, O. 2019. Are disentangled representations helpful for abstract visual reasoning? In *Advances in Neural Information Processing Systems*, 14245–14258.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, 5998–6008.

Xia, X.; Togneri, R.; Sohel, F.; and Huang, D. 2018. Auxiliary classifier generative adversarial network with soft labels in imbalanced acoustic event detection. *IEEE Transactions on Multimedia*, 21(6): 1359–1371.

Yang, M.; Liu, F.; Chen, Z.; Shen, X.; Hao, J.; and Wang, J. 2021. CausalVAE: disentangled representation learning via neural structural causal models. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 9593–9602.

Yin, G.; Liu, B.; Sheng, L.; Yu, N.; Wang, X.; and Shao, J. 2019. Semantics disentangling for text-to-image generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2327–2336.

Zhang, K.; Zuo, W.; Chen, Y.; Meng, D.; and Zhang, L. 2017. Beyond a Gaussian Denoiser: Residual Learning of Deep CNN for Image Denoising. *IEEE Transactions on Image Processing*, 26(7): 3142–3155.

Zhang, R.; Isola, P.; and Efros, A. A. 2016. Colorful image colorization. In *European Conference on Computer Vision*, 649–666. Springer.

Zhao, B.; Meng, L.; Yin, W.; and Sigal, L. 2019. Image generation from layout. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8584–8593.

Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *International Conference on Computer Vision*, 2223–2232.