Person-in-Context Synthesis with Compositional Structural Space

Weidong Yin[1] Ziwei Liu[2] Leonid Sigal [1]
[1]University of British Columbia [2]Nanyang Technological University
{wdyin,lsigal}@cs.ubc.ca zwliu.hust@gmail.com

Abstract

Despite significant progress, controlled generation of complex images with interacting people remains difficult. Existing layout generation methods fall short of synthesizing realistic person instances; while pose-guided generation approaches focus on a single person and assume simple or known backgrounds. To tackle these limitations, we propose a new problem, ***Persons in Context Synthesis****, which aims to synthesize diverse person instance(s) in consistent contexts, with user control over both. The context is specified by the bounding box object layout which lacks shape information, while pose of the person(s) by keypoints which are sparsely annotated. To handle the stark difference in input structures, we proposed two separate neural branches to attentively composite the respective (context/person) inputs into shared "compositional structural space", which encodes shape, location and appearance information for both context and person structures in a disentangled manner. This structural space is then decoded to the image space using multi-level feature modulation strategy, and learned in a self supervised manner from image collections and their corresponding inputs. Extensive experiments on two large-scale datasets (COCO-Stuff [6] and Visual Genome [19]) demonstrate that our framework outperforms state-of-the-art methods w.r.t. synthesis quality.*

1. Introduction

Learning to synthesize complex scenes with multiple persons and objects is one of the core problems in computer vision. Such technology may fundamentally revolutionize image search, as well as provide insights for visual inference problems. Many recent works tackle the problem using layouts, which is a powerful structured representations for encoding the classes and locations of objects. For example, [3, 17] use layout as an intermediate representation between scene graphs and images. Alternatively, [29, 36] directly take layout as input to generate images. While being able to generate limited objects with simple structures, existing works fail to model 'person' faithfully, see Fig. 1-left. This observation is also supported in [5], where GANs fail to reconstruct the person of the original image. Presumably the challenge is the diversity of human articulation and appearance.

In separate research thread, [4, 10, 23, 28] focus on synthesizing persons with pose as a powerful guidance. Most of these works take raw image containing background as input. They then manipulate the original person(s) in that image towards provided pose(s). There are several drawbacks for these methods: 1) they do not model the context for corresponding person, thus either the background is simple or is provided as part of the input image. 2) they can only model one person per image, lacking the interactions between different person instances.

To overcome these limitations, we propose a new task called **Persons in Context Synthesis**, which aims to synthesize diverse person instance(s) in the specified layout context (see Figure 1 for illustration). By specifying the input layout and keypoints inside each person box, our approach is able to generate a high-resolution realistic image that contains the desired context and compatible person instance(s). In this manner, we jointly model the interactions between and among persons and objects, within a unified framework.

Several unique challenges arise with this new task. First, layouts and keypoints are fundamentally different modalities. Previous works only deal with single input modality. Naive combination of these two research streams does not yield satisfactory results. Second, the information conveyed by layouts and keypoints is limited. Unlike semantic image generation tasks that leverage masks, the input here contains limited spatial information. The actual shape and appearance of object(s) and person(s) should be determined by not only the locations, labels and keypoints, but also their interactions and compatibility in the scene. A good generative model should take all of these factors into consideration.

In this work, we address the above challenges by modeling layouts and keypoints using two separate neural branches, namely *context* and *person* branch respectively, which attentively composite the respective inputs into shared *compositional structural space*. This learned

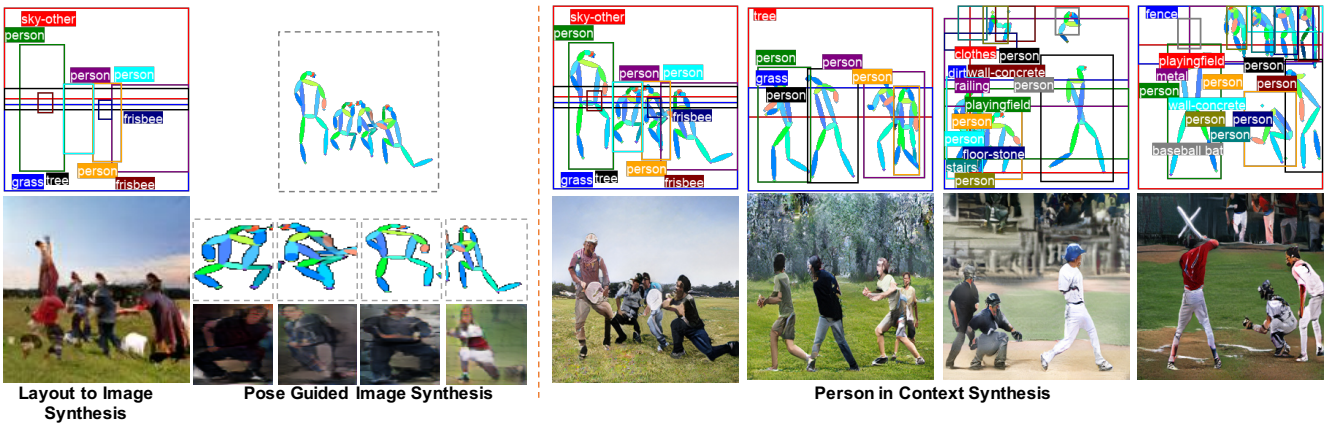

Figure 1: **Generative Settings.** An illustration of the difference between layout to image synthesis, pose guided synthesis and person in context synthesis (proposed). First column illustrates an example from [29]; The second column illustrates result from [26]. In third column and onward we illustrate our results.

structural space is beneficial for final synthesis in many aspects. First, the shape, location and appearance of each person, or context object, is represented and encoded in a disentangled manner. Second, the person and context structures are compatible with each other and can be composited in this mid-level space with simple linear summation. Third, the compositional structural space can be learned in a self supervised manner from image collections and corresponding inputs, with proposed multi-level feature modulation strategy and person-context discriminator. Finally, it enables high-quality and high-resolution image synthesis, and shows performance boost in FID on proposed 'person split' test set.

Contributions. Our contributions are three-fold: **1)** We propose a new task called persons in context synthesis, which takes both keypoints and layouts as input, and aims to synthesize diverse person instances as well as varying contexts that are visually compatible with the synthesized person(s). **2)** To handle the stark difference in input structures,we proposed two separate neural branches to attentively composite the respective (context/person) inputs into shared "compositional structuralspace", which encodes shape, location and appearance information for both context and person structures in a disentangled manner. **3)** We performed extensive evaluations on two large-scale datasets (COCO-Stuff [6] and Visual Genome [19]) to demonstrate that our framework outperforms state-of-the-art methods in synthesis quality and diversity.

2. Related Work

Conditional Image Generation. Conditional image generation approaches generate images conditioned on additional input information, including semantic maps [16, 26, 31], image captions [20, 33], sketches [7, 22] and input images [21, 37, 38]. Generating images from layout is also a specific kind of conditional image generation task. Layout is often used as an intermediate representation during the generation process, e.g., when generating from text [20] or scene graphs [17]. However, such approaches fail to generate images of high quality. In contrast, [17] generate images from provided semantic map, achieving high quality results at the expense of very laborious pixel-level user input. Different from these works, we try to generate images directly from the given layout and keypoints, which is a novel and fundamentally different paradigm for image generation.

Pose Guided Image Synthesis. Recently, several GAN-based models [4, 10, 23, 28] have been proposed for pose guided image synthesis. Most of these works take raw image as input and generate images with different pose by borrowing information from the raw input image. In contrast, [24] use sampling, in the disentangled latent space, to generate person images. However, these approaches learn to predict a person in a new pose on top of the specified training background or even require empty, white background. Instead, our method models both complex background context and persons jointly in a unified framework.

Feature Modulation Techniques. Conditional normalization layers [15, 8] were first proposed in the task of style transfer, and then applied to other kinds of tasks. Most of these conditional normalization layers work by first normalizing the layer activations into zero mean and unit variance. Then they are denormalized into different mean and variance using learned affine transformaitons conditioned on external data such as class labels. The earlier normalization techniques produce uniform normalization parameters across spatial locations, washing away class information across different spatial locations. For these reasons we adopt the spatial adaptive normalization layer [26]. In our work, the normalization parameters are generated from compositional structural canvas to provide guidance

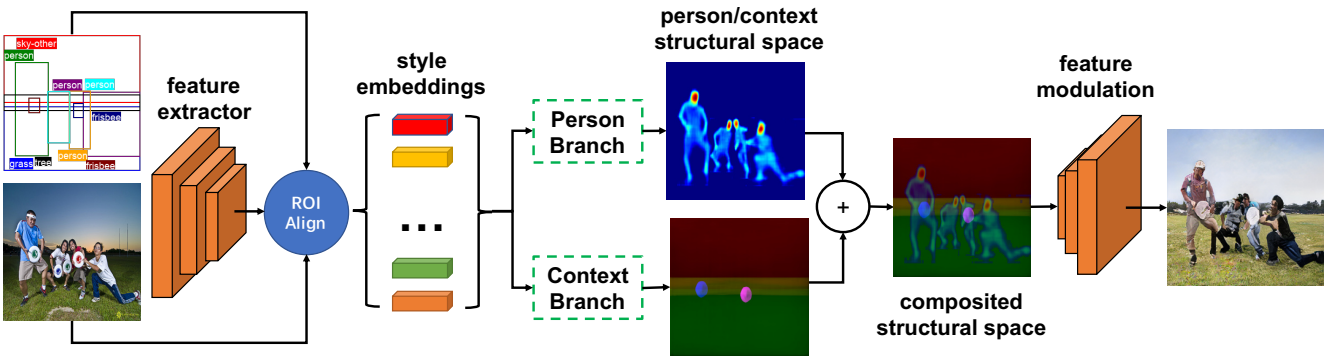

Figure 2: **Overview of our framework.** The input to our model (in training) is the ground truth image with its layout and keypoints. High level feature maps are first extracted from ground truth using ResNet50. Then we use ROIAlign to crop out feature maps for different instances including persons and objects. Style embedding for each object is generated using VAE given the cropped feature map, then fed into person branch and context branch respectively. These two branches project layout and keypoint annotations into shared *compositional structural space* conditioned on the style embeddings. Finally, we perform multi-level feature modulation to decode this structural space to a final image.

towards final image synthesis. Thus we preserve the structural information during the generation process.

Attention Mechanisms. Attention was first proposed in machine translation and then widely applied in various vision tasks such as classification [14, 30], image captioning [2] and generative models [25, 34]. Most attention mechanisms work by generating attention masks and then aggregating features with these provided masks. The resulting dynamic feature aggregation strategies enhance traditional neural networks. In this work, we proposed instance-level attention to better model the diverse shapes and varying appearances of different objects.

3. Our Approach

Our goal is to develop a model which takes as input the context and person representations and synthesizes realistic image correspondingly. The context is represented by layout consisting of bounding boxes and their class labels while person(s) are specified by keypoints in corresponding bounding boxes. The primary challenges are as follows: First, the layout as context representation is coarse and synthesized images need to respect the location of bounding boxes, class labels and style embeddings specified by the input. Second, the synthesized person instances need to be diverse and respect the pose(s). Finally, the synthesized image of person in context need to be compatible and realistic with natural interactions between and among person(s) and object(s).

To address these challenges, we introduce two key components in our framework, namely *person branch* and *context branch*. These two branches are used to model two different types of annotations separately and project them into the same *compositional structural space*, which undergoes multi-level feature modulation in decoding to obtain a synthesized image. See Figure 2 for illustration. Notably, all components are differentiable and trained end-to-end without any extra supervision needed, except for the ground truth images with aforementioned annotations. We will introduce components in detail in following sections.

3.1. The Construction of Compositional Structural Space

Person in Context Layout. The input to our model is person in context layout. It consists of two parts, namely context layout and multiple poses. During training, ground truth image is also needed. Specifically, given a set of object categories C, a person in context layout L is a tuple (O, B, K) where $O = \{c_1, \ldots, c_n\}$ is a set of objects with class types $c_i \in C$, and $B = \{\mathbf{b}_1, \ldots, \mathbf{b}_n\}$ is a set of coordinates, $\mathbf{b}_i \in \mathbb{R}^4$, of the form (x_1, y_1, x_2, y_2), where $(x_1, y_1), (x_2, y_2)$ is the upper left corner and lower right corner of the corresponding bounding boxes respectively. Bounding boxes are divided into two types, where $B_o = \{\mathbf{b}_{o1}, \ldots, \mathbf{b}_{on_o}\}$ do not contain person and $B_p = \{\mathbf{b}_{p1}, \ldots, \mathbf{b}_{pn_p}\}$ contain person; $n_o + n_p = n$ and $B = \{B_o, B_p\}$. For each $\mathbf{b}_{pi}$ we have corresponding keypoints $K = \{\mathbf{k}_1, \ldots, \mathbf{k}_{n_p}\}$, where $\mathbf{k}_i = \{(\hat{x}_1, \hat{y}_1), (\hat{x}_2, \hat{y}_2), ..., (\hat{x}_m, \hat{y}_m)\} \in \mathbb{R}^{2m}$.

Object Embeddings from RoIAlign. Given the ground truth image, we first extract feature map using ResNet50 [12]. Then object embeddings corresponding to all bounding boxes, including person and context objects, are cropped using ROIAlign [11] from the extracted feature map. The object embeddings $\mathbf{o}_i \in \mathbb{R}^{512}$ are used to model and control appearance (color/texture) of different objects.

Diverse Style Embeddings. The extracted object embeddings, by default, do not follow any distribution that can be easily sampled at test time. To be able to sample diverse images with different styles of objects, we introduced a VAE

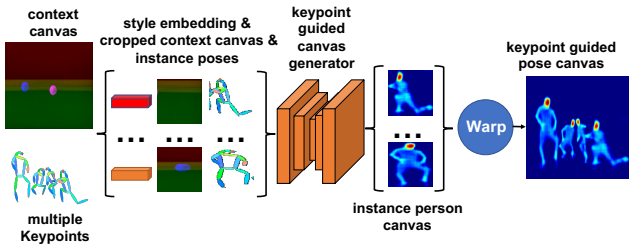

Figure 3: **Illustration of person branch.** The inputs to person branch are instance-level keypoints, style embeddings and cropped context structure. These inputs are converted into instance-level person structure. All instance-level structures are put into locations specified by bounding boxes using differentiable bilinear warpping.

[18] which takes extracted object embeddings $\mathbf{o}_i$ as input and generate corresponding style embeddings $\mathbf{e}_{oi}$ by sampling from the posterior $Q(\cdot|\mathbf{o}_i)$. At test time we sample from Gaussian prior instead to get diverse appearances for both persons and objects. KL loss is introduced to regularize the network:

$$\mathcal{L}_{KL} = \mathbb{E}[D_{KL}(Q(\cdot|\mathbf{o}_i)\|\mathcal{N}(0, I))]. \quad (1)$$

Location Retargeting by Bilinear Warping. To put different instance-level structures into locations specified by bounding boxes B in a fully differentiable manner, we used differentiable bilinear warpping. This module is shared by person branch and context branch. Given an instance-level structure $\mathbf{f}_i$ with shape $D \times S_f \times S_f$ and the location specified by $\mathbf{b}_i = \{x_1^i, y_1^i, x_2^i, y_2^i\}$, the warped output $\mathbf{F}_i$ is of size $D \times S_F \times S_F$ (note $S_f < S_F$). At each spatial location $\mathbf{F}_i(x, y)$, the output feature vector is calculated as

$$\mathbf{F}_i(x,y) = \sum_{(x',y')\in N_i(x,y)} (1 - |\alpha_x^i x + \beta_x^i - x'|) (1 - |\alpha_y^i y + \beta_y^i - y'|)\mathbf{f}_i(x', y') \quad (2)$$

where $\alpha_x^i x + \beta_x^i \in (0, S_f), \alpha_y^i y + \beta_y^i \in (0, S_f)$ and $\alpha_x^i = \frac{S_f}{x_2^i - x_1^i}, \beta_x^i = \frac{S_f x_1^i}{x_1^i - x_2^i}, \alpha_y^i = \frac{S_f}{y_2^i - y_1^i}, \beta_y^i = \frac{S_f y_1^i}{y_1^i - y_2^i}$. $N_i(x, y)$ denotes the four neighbors of $(\alpha_x^i x + \beta_x^i, \alpha_y^i y + \beta_y^i)$ in $\mathbf{f}_i$. For other locations of (x, y) we simply pad with zeros.

After bilinear warping of M instance-level structures, we get a tensor $\mathbf{F}$ of shape $M \times D \times S_F \times S_F$. Then we sum along the first dimension to compose these features together, resulting in the structural space of shape $D \times S_F \times S_F$.

Context Branch. The inputs to context branch are style embeddings $\mathbf{e}_{oi}$ with corresponding label embeddings $\mathbf{e}_{ci}$ for each bounding box $\mathbf{b}_{oi}$ that do not contain person. As is shown in Figure 4, instead of filling each bounding boxes with $[\mathbf{e}_{oi}, \mathbf{e}_{ci}]$, we first generate an instance-level sparse attention mask for each context object $\mathbf{m}_i = \max(0, G_m(\mathbf{e}_{ci}))$ using a mask generator G_m. Given $M = n_o$ objects, the attention masks $M_a = \{\mathbf{m}_1, \ldots, \mathbf{m}_{n_o}\}$ are of shape $M \times S_f \times S_f$ where S_f is spatial size of each mask. Then we fill them with embeddings $E = \{[\mathbf{e}_{o1}, \mathbf{e}_{c1}], \ldots, [\mathbf{e}_{on_o}, \mathbf{e}_{cn_o}]\}$ of shape $M \times D$ by cross product and the outputs are $M = n_o$ instance-level structures each of shape $D \times S_f \times S_f$. Then we use bilinear warping module to put them into correct locations and the output forms context structural space, which is of shape $D \times S_F \times S_F$.

Person Branch. Given $M = n_p$ (with slight abuse of notation) bounding boxes of person $B_p = \{\mathbf{b}_{p1}, \ldots, \mathbf{b}_{pn_p}\}$ with corresponding keypoints $K = \{\mathbf{k}_1, \ldots, \mathbf{k}_{n_p}\}$ inside each box, our goal is to construct person structural space from these inputs similar to that in context branch. To achieve this goal, we first convert the keypoints K into pose heatmaps $H = \{\mathbf{h}_1, \ldots, \mathbf{h}_{n_p}\}$ with size $M \times S_f \times S_f$. The keypoint at each location goes through Gaussian filter with small sigma. To make persons compatible with given context, we also crop out context structures at locations B_p for different persons. Shown in Figure 4, given pose heatmaps, cropped context structures and style embeddings for each person, we concatenate them together and introduce a neural person structure generator to get converted person representation C_p of shape $M \times D \times S_f \times S_f$ and sparse attention masks for every person as M_p of shape $M \times 1 \times S_f \times S_f$. Instance-level person structure is constructed as $C'_p = C_p \times M_p$. Given C'_p of shape $M \times D \times S_f \times S_f$ and bounding boxes B_p, we use the same bilinear warping module to put them into correct locations, and the constructed person structural space is of shape $D \times S_F \times S_F$.

The person and context structural spaces from two branches are merged into *compositional structural space* with simple linear summation.

3.2. Image Synthesis from Compositional Structural Space

Multi-Level Feature Modulation. We get *compositional structural space* $\mathbf{I}_s$ from two neural branches. Then we perform *multi-level feature modulation* to convert the structural space into image space. Specifically, given $\mathbf{I}_s$ of shape $D \times S_o \times S_o$, we downsample it into multiple different scales $\{\mathbf{I}_s^{S_1}, \ldots, \mathbf{I}_s^{S_n}\}$. At each scale S_i the output from previous module first goes through BatchNorm to obtain output $\mathbf{F}_i$. Then we denormalize $\mathbf{F}_i$:

$$\mathbf{F}'_i = \gamma_i(\mathbf{I}_s^{S_i}) * \mathbf{F}_i + \mu_i(\mathbf{I}_s^{S_i}) \quad (3)$$

using two convolutional layers γ_i and μ_i which takes S_i as input. Then the denormalized output is fed into next Residual block as input. Thus the final image is synthesized as $\mathbf{I}' = G_{img}(\mathbf{I}_s)$.

Person-Context Discriminators. The realistic output images are generated by jointly training the two neural

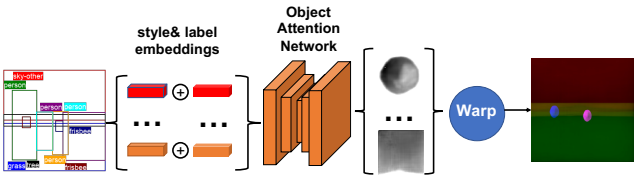

Figure 4: **Illustration of context branch.** The input to context branch are label and style embeddings for different instances. Then instance-level sparse attention mask is generated and filled with corresponding embeddings, named as instance-level context structure.

branches and feature modulation parameters against two discriminators D_{cxt} and D_{person}. D_{cxt} operates on the whole image while D_{person} operates on cropped person image patches to provide more training signal for person branch. We used the same patch-based discriminator as pix2pixHD[31] at three different scales. The adversarial loss $\mathcal{L}_{GAN}$ for two discriminators are both calculated as

$$\mathcal{L}_{GAN} = \mathbb{E}`_{\mathbf{I}\sim p_{\text{real}}} \log D(\mathbf{I}) + \mathbb{E}_{\mathbf{I}'\sim p_{\text{fake}}} \log[1 - D(\mathbf{I}')] \quad (4)$$

3.3. Learning

Training Objectives. We jointly train the two branches, feature modulation parameters G_{img} and the discriminators D_{cxt}, D_{person}. The generation network is trained to minimize the weighted sum of following losses:

1. *Feature matching loss:* $\mathcal{L}_{feat} = \|F(\mathbf{I}') - F(\mathbf{I})\|$ penalizing the L1 difference between feature vectors of generated images and real images. The features are extracted from discriminator and VGG network.
2. *KL divergence loss:* $\mathcal{L}_{KL}$ penalizing the KL divergence of posterior distribution $Q(\cdot|\mathbf{o}_i)$ obtained from object embedding network and the normal distribution $\mathcal{N}(0, I)$ prior.
3. *Image adversarial loss:* $\mathcal{L}_{GAN}$ from discriminator encouraging the generated image patches to appear realistic. We use a hinge loss, which is a variant of GAN loss.
4. *Attention TV loss:* $\mathcal{L}_{attn} = \sum_i \|\nabla\Phi^{x_i}\|^2 + \|\nabla\Phi^{y_i}\|^2$ on instance level sparse attention $\mathbf{m}_i$ both for person and context to regularize the attention mask to be smooth with fewer holes.

Implementation Details. We train all models using Adam with learning rate 2×10^{-4} for 100 epochs both on COCO and Visual Genome dataset. We use batch size 8 for each GPU at 256 resolution and 32 at 128 resolution. We use 4 Tesla P100 in parallel and the model converges in 5 days at 256 and 1 day at 128 resolution. We use LeakyReLU for both generator and discriminator.

4. Experiments

We evaluated our model at two different resolutions on Visual Genome and COCO-Stuff datasets. In our experiments we aim to show that our method generates images of complex layouts which respect the input bounding boxes, class labels and keypoints. As there's no existing methods that specifies both layouts and keypoints as input, we divide our comparison into two sections. In the first section we compare with all standard baselines. In the second section we compare with state-of-the-art variants and ablations that specify both layout and person annotation as input for a detailed analysis. We will release the code upon acceptance.

4.1. Benchmark Results

Datasets. We perform experiments on the 2017 COCO-Stuff [6] dataset, which augments a subset of the COCO dataset with additional stuff categories. The dataset annotates 40K train and 5K validation images with bounding boxes for 182 categories in total.

We set the maximum number of bounding boxes to appear in one image as 12. In practice, we sort the bounding boxes in a descending order of area and keep the top 12 bounding boxes with largest area, removing the rest. We also remove images with objects covering less than 70% of the area, and those without any bounding boxes containing keypoints, leaving around $55K$ images for training. For COCO, $N_{max} = 12$. To evaluate the performance of all models under person-in-context setting, we remove images in the validation/test set that do not contain any person. We name it as "person split" for COCO which gives us around 1K images. We will release the corresponding splits.

We also used Visual Genome [19] version 1.4 which comprises around $110K$ images annotated with bounding boxes. We divide the data into 80% train, 10% val and 10% test using same splits as [17]. Also we use the same label set as [17], except that we use one label 'person' for all instances of 'woman', 'man', etc. We remove small bounding boxes and images with little object coverage following the same procedure as for COCO. Finally, we use AlphaPose [9, 32] to detect keypoints automatically in all images. The original split gives us around $60K$ images for training and $5K$ for testing. And similarly, we evaluate on the "person split" of Visual Genome which contains around 2K images.

Standard Comparison Methods. We compare our approach with several existing state-of-the-art image synthesis methods. Scene Generation(SG) [3] generate images from scene graphs. For fair comparison, we use ground truth layout for them to generate images. The [3] requires mask annotation so only results on COCO-Stuff are available for this method. LostGAN [29] generate images directly from given layout. As different methods work under different resolutions, we report results for two different resolutions at 128×128 and 256×256. Sg2im[17] and Layout2Im[36] only works under 64×64 resolution so we did not compare with those quantitatively.

Evaluation Metrics. We adopt multiple evaluation metrics for evaluating the generated images. Frechet Inception Dis-

Figure 5: **Examples of generated images from complex layouts.** Results on COCO-Stuff and Visual Genome obtained by our method and the baselines. For each example we show input layout with keypoints, ground truth, 64 × 64 images generated by Layout2im [36], 128 × 128 images generated by LostGAN [29] and 256 × 256 images by Scene Generation [3] and our method. Note that [3] only have results on COCO-Stuff.

Table 1: A quantitative comparison using various image generation scores on person split of COCO-Stuff and Visual Genome dataset.

Datasets		COCO-Stuff				Visual Genome				Param Num	
Resolution	Method	IS	FID	Acc	DS	Inception	FID	Acc	DS	G	D
128x128	Real Im	17.30±0.14	0.00	58.51	-	17.41±0.16	0.00	63.24	-	-	-
	SG[3]	9.17±0.66	85.83	39.92	0.35±0.08	-	-	-	-	183.07	**1.50**
	LostGAN[29]	**9.35±0.52**	78.20	41.10	**0.40±0.09**	**8.26±0.35**	62.10	40.94	**0.43±0.09**	36.30	57.88
	Ours	8.95±0.15	**77.80**	**50.17**	0.33±0.12	7.68±0.46	**58.74**	**57.49**	0.32±0.09	**22.70**	4.40
256x256	Real Im	20.22±0.77	0.00	61.77	-	22.63±0.23	0.00	65.82	-	-	-
	SG[3]	10.33±0.43	103.80	37.84	**0.48±0.09**	-	-	-	-	183.07	**1.50**
	Ours	**10.92±0.41**	**76.10**	**51.08**	0.38±0.09	10.61±0.43	**60.86**	**58.88**	**0.36±0.10**	**35.10**	4.40

tance (FID) [13] is employed to measure the distribution distance between generated images and real images. The lower the better. Diversity Score(DS) [35] is used to measure the distance between pairs of images generated given same input. It is based on the perceptual similarity between two images. The higher the better. Inception Score (IS) [27] is also used to evaluate the quality of generated images. It uses an ImageNet classification model to encourage recognizable objects within images and diversity across images. Classification Accuracy (Acc) is used to evaluate whether the generated objects are recognizable. The higher the better. We trained a ResNet50 classifier on real images with two different scales to serve as an oracle.

Qualitative Results. Figure 5 shows generated images using our method as well as the baselines. As can be seen we can generate complex images with multiple objects at high resolution and with realistic details. For example, in column two our method generates three persons with diverse textures, and different parts of the person are recognizable, such as heads, hands, legs and shoes. The other

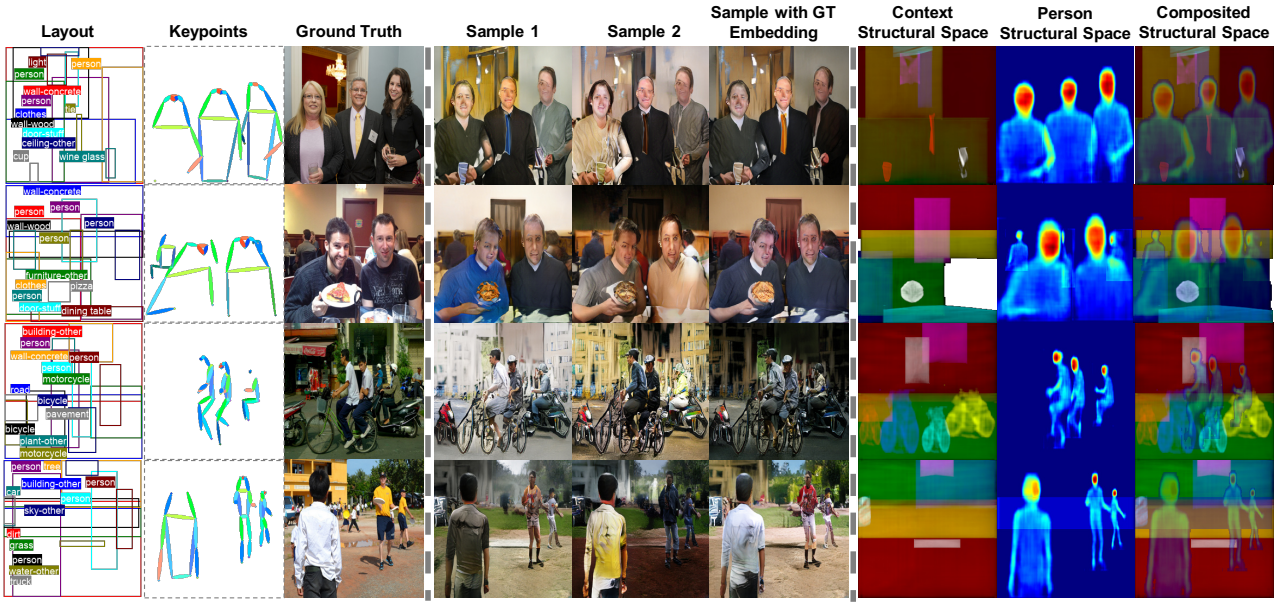

Figure 6: Examples of generated images from different style embeddings and the corresponding visualized structural space. The first three columns are ground truth layouts, keypoints and images. The next three columns are synthesized images at 256 resolution from different style embeddings(two randomly sampled and one extracted from ground truth images). The last three columns are visualized context, person and compositional structural space respectively.

methods failed to produce recognizable person appearances. These examples also show that our method generates images which respect the location constraint, class constraint and keypoints constraint. This is due to the superiority of our combination of compositional structural space and feature modulation techniques, which projects annotations with different modalities into shared structural space such that they are compatible during generation process.

Diverse Sampling from Style Embeddings. In Figure 6 we demonstrate our method's ability to generate a diverse set of images given the same layout, by sampling from different style codes which follow Gaussian prior. Since we used VAE to construct the latent space of style codes, we can easily manipulate the style of different objects by providing different style codes. For example, in column "Sample 1" and "Sample 2", the sampled style embeddings from Gaussian prior are completely different from each other. And the "Sample with GT Embedding" column use embeddings extracted from ground truth images, resulting in output images that possess similar appearances as ground truth while maintaining same structures. This disentanglement is enabled by compositional structural space.

Quantitative Results. Table 1 compares our method with other baselines and the real test images using person splits on COCO-Stuff and Visual Genome. Our method outperforms other method in terms of FID and Classification Accuracy. We noticed that LostGAN achieved comparable performance as our model, and even better in terms of Inception Score. This is due to their discriminator which has an order of magnitude higher number of parameters. As is shown in Table 1, their discriminator has 57.88 millon parameters, which does not scale up to higher resolutions. Instead, SG and our work borrow discriminator from patchGAN which requires significantly less parameters (1.5 and 4.4 million respectively). As a result, our method is more stable during training, requires less computational cost and scales to higher resolutions. With the same patchGAN based discriminator, our method beats SG by a large margin. Our diversity score is not as good as some of the other baselines. This is because our method respects the input specified by compositional structural space, and the diversity sampling will only change the texture of generated images instead of the structure as is shown in Figure 6.

4.2. Comparison with State-of-the-Art Variants and Ablations

State-of-the-Art Variants. There is no existing method that addresses the problem of person in context synthesis, which specifies both layout and keypoint as input. Thus we proposed several variants, which require both layout annotations and person annotations such as keypoints or densepose masks[1]. Two variants([26]+[29],[26]+[3]) are proposed based on existing state-of-the-art. GauGAN [26] specifies one pose heatmap as input and synthesizes one single person each time. We trained it from scratch for keypoint guided pose synthesis. Then we combine results with [29] and [3], respectively, by blending the synthesized person image patches with synthesized images from layout at

Table 2: Qualitative results on proposed person split and person crop dataset. [1] only use layout as input. [2] use both layout and keypoint. [3] use both layout and densepose mask.

Method		[29][1]	[3][1]	[29]+[26][2]	[3]+[26][2]	psp→kp[2]	psp→dp[3]	w/o ia [2]	ours[2]	ground truth
Person Split	FID↓	78.20	85.83	98.07	100.27	99.75	100.43	94.74	**77.80**	0
	IS↑	**9.35±0.52**	7.39±0.27	7.08±0.37	6.03±0.34	6.52±0.34	7.53±0.51	7.50±0.04	8.95±0.15	17.00±0.28
Person Crop	FID↓	80.60	81.44	86.84	86.84	77.74	75.26	77.57	**52.81**	0
	IS↑	5.82±0.19	5.99±0.10	4.09±0.06	4.09±0.06	6.01±0.13	5.77±0.03	5.92±0.05	**6.19±0.25**	7.92±0.35

corresponding person box locations using Poisson blending.

We also demonstrate that a naive combination of context and pose annotations does not succeed, neither for sparse keypoints nor dense segmentation masks, by providing three ablations that take both of these annotations. "psp→kp" replaces person structural space with keypoints, which is concatenated directly on top of context structural space. Similarly, "psp→dp" replaces person structural space with densepose masks, which is a series of 2d segmentation masks that annotates the shape of different body parts. Densepose masks are available on COCO dataset. Note that these masks are more powerful and expensive annotations as compared with 2d keypoints used by us. "w/o ia" removes the instance-level sparse attention during construction process of compositional structural space.

Person Crop Datasets. To evaluate the synthesize quality of person images, we construct another dataset named 'person crop'. It is constructed from COCO images and each person crop is resized into 64×64 patch. The training and testing split for person crop is same as COCO. We use the training split for GauGAN to learn from scratch, and the testing split to evaluate different methods. To compare with, we crop out persons from generated images at 128 resolution and resize them into 64×64 patches. Results are in Table 2.

Effectiveness of Compositional Structural Space. As is shown in Table 2, where our method achieves the lowest FID on both 'person split' and 'person crop'. If we look at the performance difference between [29] and [29]+[26], or [3] and [3]+[26], there is a performance drop with [26] added. This leads to the conclusion that modeling layouts and keypoints separately in image space will decrease the performance after blending. By projecting them into the same compositional structural space, we get more coherent and compatible results when it is decoded into an image.

Effectiveness of Person Structural Space. We observe that in Table 2, our method achieves the lowest FID and highest IS score compared with these ablations. This validates the conclusion that stronger annotations (such as densepose mask) does not necessarily produce higher quality results. Person keypoint annotations lie in a different structural spaces from context. Naive concatenation of keypoints on top of context structural space leads to performance drop.

Table 3: User Study Results on COCO-Stuff Dataset at 128×128 resolution.

Method	Global Coherence	Visual Quality of Persons
LostGAN [29]	35%	15%
Scene Generation [3]	20%	10%
Ours	45%	75%

We visualized person structural spaces with heatmaps using L1 norm of corresponding feature vectors in Figure 6. The visualized person features are dense around relevant body parts, highly activated around head (shown in red) and joints (shown in green) and not activated in irrelevant regions. This learned representation has richer structures than raw annotations such as keypoints or densepose masks, and are more compatible with context representations.

Effectiveness of Instance Level Sparse Attention. As is shown in Figure 6, each instance structure (context and person), is zero at irrelevant regions. As shown in Table 2, the removal of this sparse attention mask will lead to performance drop, because: 1) different bounding boxes can affect each other in overlapping areas and 2) the shape of the instances is less accurate.

User Study. We perform a user study to compare with other baselines. 20 volunteers were involved. Each volunteer was shown the synthesized images from COCO-Stuff dataset at 256 resolution and was asked to select the preferable images in terms of the global coherence of both context and persons, and the visual quality of persons respectively. The results reported in Table 3 show that our method significantly outperforms other methods, especially in terms of visual quality of synthesized persons.

5. Conclusion

We proposed a novel problem called **Persons in Context Synthesis**, which aims to synthesize 1) diverse person instances, as well as 2) varying contexts that are visually compatible with the synthesized persons. The context is specified by bounding box object layout, while pose of the person(s) by keypoints. This difference in input modalities motivate the use of separate neural branches that attentively project the respective (context/person) inputs into "compositional structural space", where person and context representations are compatible with each other. Extensive experiments on two large-scale datasets (COCO-Stuff and Visual Genome) demonstrate that our approach outperforms state-of-the-art in synthesis quality and diversity.

References

[1] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7297–7306, 2018.

[2] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6077–6086, 2017.

[3] Oron Ashual and Lior Wolf. Specifying object attributes and relations in interactive scene generation. *arXiv preprint arXiv:1909.05379*, 2019.

[4] Guha Balakrishnan, Amy Zhao, Adrian V. Dalca, Frédo Durand, and John V. Guttag. Synthesizing images of humans in unseen poses. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8340–8348, 2018.

[5] David Bau, Jun-Yan Zhu, Jonas Wulff, William Peebles, Hendrik Strobelt, Bolei Zhou, and Antonio Torralba. Seeing what a gan cannot generate. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4502–4511, 2019.

[6] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Cocostuff: Thing and stuff classes in context. In *Computer vision and pattern recognition (CVPR), 2018 IEEE conference on*. IEEE, 2018.

[7] Wengling Chen and James Hays. Sketchygan: Towards diverse and realistic sketch to image synthesis. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9416–9425, 2018.

[8] Henk Jan de Vries, Florian Strub, Jérémie Mary, Hugo Larochelle, Olivier Pietquin, and Aaron C. Courville. Modulating early visual processing by language. In *NIPS*, 2017.

[9] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. RMPE: Regional multi-person pose estimation. In *ICCV*, 2017.

[10] Xintong Han, Zuxuan Wu, Zhe Wu, Ruichi Yu, and Larry S. Davis. Viton: An image-based virtual try-on network. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7543–7552, 2017.

[11] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[13] 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 *Advances in Neural Information Processing Systems*, pages 6626–6637, 2017.

[14] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7132–7141, 2017.

[15] Xun Huang and Serge J. Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 1510–1519, 2017.

[16] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5967–5976, 2016.

[17] 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*, pages 1219–1228, 2018.

[18] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[19] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017.

[20] Wenbo Li, Pengchuan Zhang, Lei Zhang, Qiuyuan Huang, Xiaodong He, Siwei Lyu, and Jianfeng Gao. Object-driven text-to-image synthesis via adversarial training. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12174–12182, 2019.

[21] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. *ArXiv*, abs/1703.00848, 2017.

[22] Yongyi Lu, Shangzhe Wu, Yu-Wing Tai, and Chi-Keung Tang. Image generation from sketch constraint using contextual gan. In *ECCV*, 2017.

[23] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. Pose guided person image generation. In *Advances in Neural Information Processing Systems*, pages 406–416, 2017.

[24] Liqian Ma, Qianru Sun, Stamatios Georgoulis, Luc Van Gool, Bernt Schiele, and Mario Fritz. Disentangled person image generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 99–108, 2018.

[25] Youssef A. Mejjati, Christian Richardt, James Tompkin, Darren Cosker, and Kwang In Kim. Unsupervised attention-guided image-to-image translation. In *NeurIPS*, 2018.

[26] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Gaugan: semantic image synthesis with spatially adaptive normalization. In *SIGGRAPH '19*, 2019.

[27] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Advances in neural information processing systems*, pages 2234–2242, 2016.

[28] Aliaksandr Siarohin, Enver Sangineto, Stéphane Lathuilière, and Nicu Sebe. Deformable gans for pose-based human image generation. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3408–3416, 2017.

[29] Wei Sun and Tianfu Wu. Image synthesis from reconfigurable layout and style. *arXiv preprint arXiv:1908.07500*, 2019.

[30] Feng Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Yen Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. Residual attention network for image classification. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6450–6458, 2017.

[31] 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. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8798–8807, 2017.

[32] Yuliang Xiu, Jiefeng Li, Haoyu Wang, Yinghong Fang, and Cewu Lu. Pose Flow: Efficient online pose tracking. In *BMVC*, 2018.

[33] Guojun Yin, Bin Liu, Lu Sheng, Nenghai Yu, Xiaogang Wang, and Jing Shao. Semantics disentangling for text-to-image generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2327–2336, 2019.

[34] Han Zhang, Ian J. Goodfellow, Dimitris N. Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In *ICML*, 2018.

[35] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

[36] Bo Zhao, Lili Meng, Weidong Yin, and Leonid Sigal. Image generation from layout. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8584–8593, 2019.

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

[38] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A. Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. In *NIPS*, 2017.