Person in Context Synthesis with Compositional Structural Space

by

Weidong Yin

BSc, Fudan University, 2018

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF

Master of Science

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES
(Computer Science)

The University of British Columbia
(Vancouver)

January 2021

© Weidong Yin, 2021
The following individuals certify that they have read, and recommend to the Faculty of Graduate and Postdoctoral Studies for acceptance, the thesis entitled:

**Person in Context Synthesis with Compositional Structural Space**

submitted by Weidong Yin in partial fulfillment of the requirements for the degree of Master of Science in Computer Science.

**Examinining Committee:**

Leonid Sigal, Associate Professor, Department of Computer Science, UBC
*Supervisor*

James Little, Professor, Department of Computer Science, UBC
*Supervisory Committee Member*
Abstract

Despite significant progress, controlled generation of complex images with interacting people remains difficult. Existing layout to image 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 sparse annotations. 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 a 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 with respect to synthesis quality.
Lay Summary

It is difficult to generate complex scenes with interacting people and objects. First, generating complex scenes requires synthesizing realistic person instances and context objects. Second, a diverse set of images with different appearances need to be synthesized given same input. Third, inside the synthesized image, the person and context objects should be compatible with each other. Solving this problem may fundamentally revolutionize image search, as well as provide insights for visual inference problems. We propose a model that overcomes these limitations and shows state-of-the-art performance both quantitatively and qualitatively on two benchmark datasets.
Preface

This work presented here is conducted by the first author, Weidong Yin, with the help and advises of a second author, Dr. Ziwei Liu. This work was done under the supervision of Prof. Leonid Sigal. The first author is completely responsible for the design, implementation of the model and the experiments. Dr. Ziwei Liu has kindly shared his advice on model architectures and provided his expertise on writing a conference paper. Prof. Sigal provided feedback during each step.
# Table of Contents

Abstract ................................................................. iii

Lay Summary ............................................................ iv

Preface ........................................................................ v

Table of Contents ........................................................ vi

List of Tables .............................................................. viii

List of Figures ............................................................. ix

1 Introduction .............................................................. 1
   1.1 Contributions ..................................................... 3

2 Related Work ........................................................... 4
   2.1 Conditional Image Generation ................................. 4
   2.2 Pose Guided Image Synthesis ................................. 4
   2.3 Feature Modulation Techniques ............................... 5
   2.4 Attention Mechanisms ......................................... 5

3 Method ................................................................. 6
   3.1 The Construction of Compositional Structural Space ...... 7
   3.2 Image Synthesis from Compositional Structural Space .... 10
   3.3 Learning .......................................................... 11
4 Experiments .................................................. 12
  4.1 Benchmark Results ..................................... 12
  4.2 Comparison with State-of-the-Art Variants and Ablations .. 17
  4.3 Further Analysis ........................................ 19

5 Conclusion .................................................. 21

Bibliography .................................................. 23
List of Tables

Table 4.1  A quantitative comparison using various image generation scores on person split of COCO-Stuff and Visual Genome dataset. . . 14
Table 4.2  Qualitative results on proposed person split and person crop dataset. ¹only use layout as input. ²use both layout and key-point. ³use both layout and densepose mask.  . . . . . . . . 19
Table 4.3  User Study Results on COCO-Stuff Dataset.  . . . . . . . . . 19
List of Figures

Figure 1.1 **Generative Settings.** An illustration of the difference between layout to image synthesis, pose guided synthesis and person in context synthesis (proposed). The first column illustrates an example from [29]: The second column illustrates the result from [26]. In the third column and onward we illustrate our results.

Figure 1.2 **Person in Context Synthesis.** We present some synthesized images under the setting of person in context synthesis at 256 resolution. From left to right are input context layouts, multi-person keypoints and a set of synthesized images with diverse appearances.

Figure 3.1 **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 a Variational Autoencoder(VAE) given the cropped feature map, then fed into the person branch and context branch respectively. These two branches project layout and keypoint annotations into the 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.
Figure 3.2  **Illustration of the Two-branches.** Detailed view of the context branch and person branch respectively. The input to the context branch is label and style embeddings for different instances. Then the instance-level sparse attention mask is generated and filled with corresponding embeddings, named as instance-level context structure. The inputs to the person branch are instance-level keypoints, style embeddings, and cropped context structure. These inputs are converted into an instance-level person structure. All instance-level structures are put into locations specified by bounding boxes using differentiable bilinear warping.

Figure 4.1  **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, $64 \times 64$ images generated by Layout2im [36], $128 \times 128$ images generated by LostGAN [29] and $256 \times 256$ images by Scene Generation [3] and our method. Note that [3] only have results on COCO-Stuff.

Figure 4.2  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 embedding spaces (two randomly sampled and one extracted from ground truth images). The last three columns are visualized context, person and compositional structural space respectively.

Figure 4.3  Examples of generated images from state-of-the-art variants and ablations using both layout and person annotations. For each example we show the input layouts, input images, ground truth, two baselines from existing methods, three ablations and our method on COCO-Stuff dataset at 128 resolution.

Figure 4.4  Performance of different models under complex scenes.
Figure 4.5  Examples of generated images by interpolating poses while keeping appearance the same. For each example we show lay-outs, starting poses, interpolated image sequence, and ending poses on COCO-Stuff dataset at 256 resolution. . . . . . . . . 20
Chapter 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 are powerful representations for encoding the classes and locations of objects. For example, \[3, \] 17\] use layouts as intermediate representations 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 a separate research thread, \[4, \] 10, \] 23, \] 28\] focus on synthesizing persons with pose as a powerful guidance. Most of these works take raw images containing background as input. They then manipulate the original person(s) in that image towards the provided pose(s). There are several drawbacks for these methods: 1) they do not model the context for the corresponding person, thus either the background is simple or is provided as part of the input image, and 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 1.1 for comparison of different generative settings and
**Figure 1.1: Generative Settings.** An illustration of the difference between layout to image synthesis, pose guided synthesis and person in context synthesis (proposed). The first column illustrates an example from [29]; The second column illustrates the result from [26]. In the third column and onward we illustrate our results.

**Figure 1.2: Person in Context Synthesis.** We present some synthesized images under the setting of person in context synthesis at 256 resolution. From left to right are input context layouts, multi-person keypoints and a set of synthesized images with diverse appearances.

Figure 1.2 for inputs and outputs under person in context setting clearly at a larger scale). By specifying the input layout and keypoints inside each person box, our approach can 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 annotations with fundamentally different modalities. Previous works only deal with a single input modality. A 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 thesis, we address the above challenges by modeling layouts and keypoints using two separate neural branches, namely the context and person branches respectively, which attentively composite the respective inputs into shared compositional structural space. This learned 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 a simple linear summation. Third, the compositional structural space can be learned in a self-supervised manner from image collections and corresponding inputs. Finally, it enables high-quality and high-resolution image synthesis and shows a performance boost in FID on the proposed ‘person split’ test set.

1.1 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 propose 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. 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.
Chapter 2

Related Work

2.1 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 layouts is also a specific kind of conditional image generation task. A 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 a provided semantic map, achieving high-quality results at the expense of very laborious pixel-level user input. Unlike 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.

2.2 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 images as input and generate images with a 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 an empty, white background. Instead, our method models both complex background context and persons jointly in a unified framework.

### 2.3 Feature Modulation Techniques

Feature modulation, namely feature-wise transformation, is a simple and surprisingly effective family of conditioning mechanisms. Conditional normalization layers [8, 15] is one of the feature modulation techniques 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 transformations 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 the compositional structural space to guide final image synthesis. Thus we preserve the structural information during the generation process.

### 2.4 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.
Chapter 3

Method

Our goal is to develop a model that takes as input the context and person representations and synthesizes realistic image correspondingly. The context is represented by the layout consisting of bounding boxes and their class labels while the 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(s) in context needs 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 the 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 3.1 for illustration. Notably, all components are differentiable and trained end-to-end without any extra supervision needed, except for the ground truth images with the aforementioned annotations. We will introduce components in detail in the following sections.
Figure 3.1: 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 a Variational Autoencoder (VAE) given the cropped feature map, then fed into the person branch and context branch respectively. These two branches project layout and keypoint annotations into the 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.

3.1 The Construction of Compositional Structural Space

Person in Context Layout. The input to our model is the person in context layout. It consists of two parts, namely the context layout and multiple poses. During training, the 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 $B = \{b_1, \ldots, b_n\}$ is a set of coordinates, $b_i \in \mathbb{R}^4$, 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 $B_o = \{b_{o1}, \ldots, b_{on_o}\}$ do not contain person and $B_p = \{b_{p1}, \ldots, b_{pn_p}\}$ contain person; $n_o + n_p = n$ and $B = \{B_o, B_p\}$. For each $b_{pi}$ we have corresponding keypoints $K = \{k_1, \ldots, k_{np}\}$, where $k_i = \{\hat{x}_1, \hat{y}_1, \hat{x}_2, \hat{y}_2, \ldots, \hat{x}_m, \hat{y}_m\} \in \mathbb{R}^{2m}$.

Object Embeddings from RoIAlign. Given the ground truth image, we first extract the feature map using ResNet50 [11]. Then object embeddings corresponding
to all bounding boxes, including person and context objects, are cropped using ROIAlign [12] from the extracted feature map. The object embeddings \( o_i \in \mathbb{R}^{512} \) 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 [18] which takes extracted object embeddings \( o_i \) as input and generate corresponding style embeddings \( e_{oi} \) by sampling from the posterior \( Q(\cdot|o_i) \). At test time we sample from Gaussian prior instead to get diverse appearances for both persons and objects. The Kullback Leibler (KL) loss is introduced to regularize the network:

\[
\mathcal{L}_{KL} = \mathbb{E}[D_{KL}(Q(\cdot|o_i)||\mathcal{N}(0, I))].
\] (3.1)

**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 warping. This module is shared by the person branch and context branch. Given an instance-level structure \( f_i \) with shape \( D \times S_f \times S_f \) and the location specified by \( b_i = \{x_i^1, y_i^1, x_i^2, y_i^2\} \), the warped output \( F_i \) is of size \( D \times S_F \times S_F \) (note \( S_f < S_F \)). At each spatial location \( F_i(x, y) \), the output feature vector is calculated as

\[
F_i(x, y) = \sum_{(x',y') \in N_i(x,y)} (1 - |\alpha_i^x x + \beta_i^x - x'|)(1 - |\alpha_i^y y + \beta_i^y - y'|) f_i(x', y')
\] (3.2)

where \( \alpha_i^x, \beta_i^x, \alpha_i^y, \beta_i^y \in (0, S_f) \) and \( \alpha_i^x = \frac{S_f}{x_i^2 - x_i^1}, \beta_i^x = \frac{S_f}{x_i^2 - x_i^1}, \alpha_i^y = \frac{S_f}{y_i^2 - y_i^1}, \beta_i^y = \frac{S_f}{y_i^2 - y_i^1} \). \( N_i(x,y) \) denotes the four neighbors of \( (\alpha_i^x, \beta_i^x, \alpha_i^y, \beta_i^y) \) in \( f_i \). For other locations of \( (x,y) \) we simply pad with zeros.

After bilinear warping of \( M \) instance-level structures, we get a tensor \( F \) 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 the context branch are style embeddings \( e_{oi} \) with corresponding label embeddings \( e_{ci} \) for each bounding box \( b_{oi} \) that do not contain person. As is shown in Figure 3.2, instead of filling each bounding box
Figure 3.2: Illustration of the Two-branches. Detailed view of the context branch and person branch respectively. The input to the context branch is label and style embeddings for different instances. Then the instance-level sparse attention mask is generated and filled with corresponding embeddings, named as instance-level context structure. The inputs to the person branch are instance-level keypoints, style embeddings, and cropped context structure. These inputs are converted into an instance-level person structure. All instance-level structures are put into locations specified by bounding boxes using differentiable bilinear warping.

with \([e_{oi}, e_{ci}]\), we first generate an instance-level sparse attention mask for each context object \(m_i = \max(0, G_m(e_{ci}))\) using a mask generator \(G_m\). Given \(M = n_o\) objects, the attention masks \(M_a = \{m_1, \ldots, m_n\}\) 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 \(E = \{[e_{oi1}, e_{ci1}], \ldots, [e_{on}, e_{cn_o}]\}\) 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 a bilinear warping module to put them into correct locations and the output forms the 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 \(B_p = \{b_{p1}, \ldots, b_{pn_p}\}\) with corresponding keypoints \(K = \{k_1, \ldots, k_{n_p}\}\) inside each box, our goal is to construct the 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 $H = \{h_1, \ldots, h_n\}$ with size $M \times S_f \times S_f$. The keypoint at each location goes through a Gaussian filter with small sigma. To make persons compatible with the given context, we also crop out context structures at locations $B_p$ for different persons. Shown in Figure 3.2, 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 a 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 $M \times 1 \times S_f \times S_f$. The 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 the two branches are merged into the *compositional structural space* with a simple linear summation.

### 3.2 Image Synthesis from Compositional Structural Space

**Multi-Level Feature Modulation.** We get the *compositional structural space* $I_s$ from the two neural branches. Then we perform *multi-level feature modulation* to convert the structural space into image space. Specifically, given $I_s$ of shape $D \times S_o \times S_o$, we downsample it into multiple different scales $\{I_{s1}, \ldots, I_{sn}\}$. At each scale $S_i$ the output from previous module first goes through BatchNorm to obtain output $F_i$. Then we denormalize $F_i$:

$$F'_i = \gamma_i(I_{si}^N) * F_i + \mu_i(I_{si}^N) \quad (3.3)$$

using two convolutional layers $\gamma_i$ and $\mu_i$ 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 $I' = G_{img}(I_s)$.

**Person-Context Discriminators.** The realistic output images are generated by jointly training the two neural 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 losses $\mathcal{L}_{GAN}$ for the two discriminators are both calculated as

$$\mathcal{L}_{GAN} = \mathbb{E}_{I \sim p_{real}} \log D(I) + \mathbb{E}_{I' \sim p_{fake}} \log[1 - D(I')]$$  \hspace{1cm} (3.4)$$

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: $\mathcal{L}_{feat} = \|F(I') - F(I)\|$ penalizing the L1 difference between feature vectors of generated images and real images. The features are extracted from the discriminator and the VGG network.

2. KL divergence loss: $\mathcal{L}_{KL}$ penalizing the KL divergence of posterior distribution $Q(\cdot|o_i)$ obtained from object embedding network and the normal distribution $\mathcal{N}(0, I)$ prior.

3. Image adversarial loss: $\mathcal{L}_{GAN}$ from the discriminator encouraging the generated image patches to appear realistic. We use a hinge loss, which is a variant of GAN loss.

4. Attention Total Variation(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 $2 \times 10^{-4}$ for 100 epochs both on the COCO and the 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 the generator and the discriminator.
Chapter 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 that respect the input bounding boxes, class labels, and keypoints. As there are no existing methods that specify 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.

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 contains 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 to 12. In practice, we sort the bounding boxes in descending order of area and keep the top 12 bounding boxes with the 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, $N_{\text{max}} = 12$. To evaluate the performance of all models under the person-in-
Figure 4.1: 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, $64 \times 64$ images generated by Layout2im [36], $128 \times 128$ images generated by LostGAN [29] and $256 \times 256$ images by Scene Generation [3] and our method. Note that [3] only have results on COCO-Stuff.

context setting, we remove images in the validation/test set that do not contain any person. We name this subset 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 the 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
Table 4.1: A quantitative comparison using various image generation scores on person split of COCO-Stuff and Visual Genome dataset.

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

Figure 4.2: 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 embedding spaces (two randomly sampled and one extracted from ground truth images). The last three columns are visualized context, person and compositional structural space respectively.

Training and 5K for testing. Similarly, we use 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] generates images from scene graphs. For a fair comparison, we use the ground truth layout for them to generate images. As [3] requires mask annotation, only results on COCO-Stuff are available for this method. LostGAN [29] generate images directly from
Figure 4.3: Examples of generated images from state-of-the-art variants and ablations using both layout and person annotations. For each example we show the input layouts, input images, ground truth, two baselines from existing methods, three ablations and our method on COCO-Stuff dataset at 128 resolution.

given layout. As different methods work under different resolutions, we report results for two different resolutions at $128 \times 128$ and $256 \times 256$. Sg2im [17] and Layout2Im [36] only work under $64 \times 64$ resolution so we did not compare with those quantitatively.

**Evaluation Metrics.** We adopt multiple evaluation metrics for evaluating the generated images. Frechet Inception Distance (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 the same input. This metric 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. IS 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 4.1 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 persons are recognizable, such as heads, hands, legs, and shoes. The other methods failed to produce recognizable person appearances. These examples also show that our method generates images that respect location constraints, class constraints, and keypoints constraints. 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 the generation process.

**Diverse Sampling from the Style Embeddings.** In Figure 4.2 we demonstrate our method’s ability to generate a diverse set of images given the same layout, by sampling from different style codes that follow a Gaussian prior. Since we used a 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 a Gaussian prior are completely different from each other. The “Sample with GT Embedding” column uses embeddings extracted from ground truth images, resulting in output images that possess similar appearances as ground truth while maintaining the same structure. This disentanglement is enabled by the compositional structural space.

**Quantitative Results.** Table 4.1 compares our method with other baselines and the real test images using person splits on COCO-Stuff and Visual Genome. Our method outperforms the other methods in terms of FID and Classification Accuracy. We noticed that LostGAN achieved comparable performance to our model, and was 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 4.1, their discriminator has 57.88 million parameters, which does not scale up to higher resolutions. Instead, SG and our work borrow the discriminator from PatchGAN which requires significantly fewer 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 4.2.

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 a person in context synthesis, which specifies both the 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 the layout at 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 the person structural space with keypoints, which are concatenated directly on top of the context structural space. Similarly, “psp→dp” replaces the person structural space with densepose masks, which is a series of 2d segmentation masks that annotate the shape of different body parts. Densepose masks are available on the 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 the construction process of compositional structural space.

Person Crop Datasets. To evaluate the quality of the synthesis 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 the same as COCO. We use the training split for GauGAN to learn from scratch, and the testing split to evaluate different methods. To compare the different methods, we crop out persons from generated images at 128 resolution and resize them into $64 \times 64$ patches. The results are in Table 4.2.

**Effectiveness of Compositional Structural Space.** As is shown in Figure 4.3, the boundary between person and context looks seamless in our method (the last column), while the blended person(s) look unnatural for $[3]+[26]$ and $[29]+[26]$. This is also validated in Table 4.2, 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 the projection is decoded into an image.

**Effectiveness of Person Structural Space.** Compared with results of “psp→kp” and “psp→dp” shown in Figure 4.3 (6th and 7th column), our method shows more clear body parts and higher quality context. For example, in the 2nd row, the face of the synthesized person looks more clear, and in the 3rd row, the context looks more compatible. Further, as is shown in Table 4.2, our method achieves the lowest FID and highest IS score compared with these ablations. This validates the conclusion that stronger annotations (such as densepose masks) do not necessarily produce higher quality results. Person keypoint annotations lie in different structural spaces from context. Naive concatenation of keypoints on top of the context structural space leads to a drop in performance.

We visualized the person structural spaces with heatmaps using L1 norm of corresponding feature vectors in Figure 4.2. The visualized person features are dense around relevant body parts, highly activated around the 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 is more compatible with context representations.

**Effectiveness of Instance Level Sparse Attention.** As is shown in Figure 4.2, each instance structure (context and person), is zero at irrelevant regions. As shown in Figure 4.3 and Table 4.2, the removal of this sparse attention mask will lead to
Table 4.2: 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.

<table>
<thead>
<tr>
<th>Method</th>
<th>[29]</th>
<th>[3]</th>
<th>[29]+[26]</th>
<th>[3]+[26]</th>
<th>psp→kp</th>
<th>psp→dp</th>
<th>w/o ia</th>
<th>ours</th>
<th>ground truth</th>
</tr>
</thead>
<tbody>
<tr>
<td>Person Split</td>
<td>FID↓</td>
<td>78.20</td>
<td>95.83</td>
<td>98.07</td>
<td>100.27</td>
<td>99.75</td>
<td>100.43</td>
<td>94.74</td>
<td>77.80</td>
</tr>
<tr>
<td>IS↑</td>
<td>9.35±0.52</td>
<td>7.99±0.27</td>
<td>7.06±0.37</td>
<td>6.03±0.14</td>
<td>6.52±0.34</td>
<td>7.53±0.51</td>
<td>7.50±0.04</td>
<td>8.95±0.15</td>
<td></td>
</tr>
<tr>
<td>Person Crop</td>
<td>FID↓</td>
<td>80.60</td>
<td>81.44</td>
<td>86.84</td>
<td>86.84</td>
<td>77.74</td>
<td>75.26</td>
<td>77.57</td>
<td>52.81</td>
</tr>
<tr>
<td>IS↑</td>
<td>5.82±0.19</td>
<td>5.99±0.10</td>
<td>4.09±0.06</td>
<td>4.09±0.06</td>
<td>6.01±0.13</td>
<td>5.77±0.03</td>
<td>5.92±0.05</td>
<td>6.19±0.25</td>
<td>7.92±0.35</td>
</tr>
</tbody>
</table>

a performance drop, because: 1) different bounding boxes can affect each other in overlapping areas and 2) the shape of the instances is less accurate.

4.3 Further Analysis

Performance on Complex Scenes. We evaluate the performance of our model under scenes with multiple persons and diverse poses. In Figure 4.4(a), validation sets are divided into three groups with number of persons as criterion. When only one person is present, our model performs slightly better. As the number of person instances increase, the difference between [3, 29] and our method becomes more clear. This is because our model can deal with challenging inputs containing multiple persons. Shown in Figure 4.4(b), we cluster poses and evaluate FID on different clusters. Performance of [26] is inconsistent among clusters, while our model achieves a lower FID score and is more consistent for different types of poses.

User Study. We perform a user study to compare our method with other baselines. 20 volunteers were involved. Each volunteer was shown the synthesized images from the 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.

Table 4.3: User Study Results on COCO-Stuff Dataset.

<table>
<thead>
<tr>
<th>Method</th>
<th>Global Coherence</th>
<th>Visual Quality of Persons</th>
</tr>
</thead>
<tbody>
<tr>
<td>LostGAN [29]</td>
<td>35%</td>
<td>15%</td>
</tr>
<tr>
<td>Scene Generation [3]</td>
<td>20%</td>
<td>10%</td>
</tr>
<tr>
<td>Ours</td>
<td>45%</td>
<td>75%</td>
</tr>
</tbody>
</table>
Figure 4.4: Performance of different models under complex scenes.

Figure 4.5: Examples of generated images by interpolating poses while keeping appearance the same. For each example we show layouts, starting poses, interpolated image sequence, and ending poses on COCO-Stuff dataset at 256 resolution.

and the visual quality of persons respectively. The results reported in Table 4.3 show that our method significantly outperforms other methods, especially in terms of the visual quality of synthesized persons.

Person Reenactment under Context. In Figure 4.5 we reenact the persons in synthesized context by interpolating between starting keypoints and ending keypoints, while keeping style embeddings fixed for both the person and context. As shown in the first row, the orientation of the faces of each person is changed gradually. And in the second row, overall body structures are interpolated smoothly. This is enabled by the compositional structural space, which is a disentangled representation of person and context structures.
Chapter 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 a bounding box object layout, while pose of the person(s) by keypoints. Many unique challenges come with the new settings. To tackle these challenges, we model layouts and keypoints using two separate neural branches, namely the *context* and *person* branch respectively, which attentively composite the respective inputs into shared *compositional structural space*. We learned and adopted many recently developed techniques of generative models into this thesis, including a feature modulation strategy and a person-context discriminator. A wide range of experiments demonstrate that our framework outperforms state-of-the-art methods in synthesis quality and diversity.

This thesis opens several new directions for future work. Firstly, it might be tedious for the user to specify multi-person keypoints as input. It is possible to build another generative model which automatically suggests where to put the keypoints and how they look, so that the suggested poses are compatible with the specified context layout. For example, a person is likely to ski with snow as the context, thus we should be able to generate keypoints corresponding to possible ski poses. This kind of framework can greatly reduce the time a user needs to spend to produce an output, and improves the practicality of the person in context setting. Secondly, this thesis primarily models 2D image features, neglecting the 3D scene compo-
sition and depth order. An interesting line of future work is to reason about the depth orders of different person and context objects. We can obtain more fine-grained control over the synthesis process by moving objects back and forth, and compositing them in a different depth order.
Bibliography


25


