Enhancing Person Synthesis in Complex Scenes via Intrinsic and Contextual Structure Modeling

Xi Tian[1]
xt275@bath.ac.uk
Yong-Liang Yang[1]
y.yang@cs.bath.ac.uk
Qi Wu[2]
qi.wu01@adelaide.edu.au

[1] Department of Computer Science
University of Bath
Bath, UK

[2] Department of Computer Science
University of Adelaide
Adelaide, Australia

Abstract

The Generative Adversarial Network (GAN) and its variations have enabled high-quality image generation. However, generating reasonable persons in complex scenes (such as MS-COCO images) remains challenging. We propose a novel structure-based and context-aware approach to enhance the person synthesis in complex scenes. The method can successfully predict the person pose and face structures while respecting the weak layout-based context, then leverage the structures to refine the person appearance. Our method involves three parts. First, a memory-based model is used to encode person intrinsic structures including pose and face keypoints. Second, a context-aware model infers the conditional person structures from the layout context. Third, the structure-guided person appearance refiners further enhance the final image generation. Our experiments present convincing person generation results in layout-to-image tasks on a challenging dataset. Person-related evaluations demonstrate our method achieves state-of-the-art performance, especially on person accuracy and face detection metrics.

1 Introduction

The advent of Generative Adversarial Nets (GANs) [9] largely facilitates high-quality image generation. In terms of person-related generation, recent models have been successfully applied to various applications including photo-realistic face generation [4, 19], person pose transfer [28, 38, 44], and virtual clothes try-on [7, 20], *etc*. However, existing works are always task-oriented using hand-crafted datasets that contain clean and aligned persons [25] or faces [19, 24]. Besides, heavy annotations are frequently used for guiding person synthesis, *e.g*., using the segmentation masks or UV maps of human parts for person generation [10, 33]. Despite the realistic results, these settings significantly constrain the person generation in the wild and hinder further applications in real life.

In this paper, we focus on *person-in-scene* synthesis, a task for generating persons in natural scenes with complicated context but weak conditions. It is more challenging because there are various objects in the context and the person pose and appearance are also

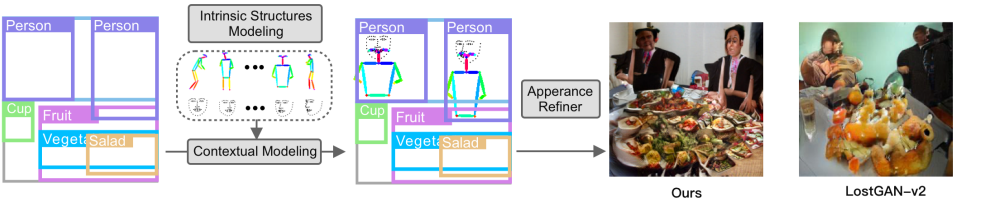

Figure 1: Illustration of our approach improving the person generation by three steps: intrinsic structure modeling, contextual structure prediction, and appearance refinement. Existing methods like [36] fail to generate plausible persons in complex scenes because they lack prior knowledge of person structures that are important to guide quality person generation.

diverse. The existing layout-to-image (L2I) generation approach using only bounding boxes as input conditions has witnessed the challenges in generating reasonable persons on COCO dataset [3, 23]. Figure 1-Right shows an image synthesis example from the state-of-the-art (LostGAN-v2 [36]), where the generated persons are difficult to recognize due to the lack of clear limbs and face features. Therefore, generating quality persons with reasonable poses and faces in complex scenes remains a largely unsolved problem.

The difficulty of the present problem has also been pointed out previously. Bau *et al*. [2] found that GANs tend to ignore persons though persons are frequent in the datasets. Sun *et al*. [36] discussed the limitations of existing L2I tasks that generating satisfactory persons are more difficult, with the argument that persons are more articulated compared with other common objects. We speculate there are two main challenges that hinder existing methods [1, 35, 36] to generate better persons in complex scenes. First, the lack of person structure knowledge. Persons are internally structure-based, despite the size and appearance variety. The structures, such as pose and face keypoints, can indicate the shape, action, and facial expression for persons. We call them *intrinsic structures*, which exist as innate characteristics of persons, independent of the complex scenes. However, existing methods ignore such prior knowledge of persons. Second, person structures are diverse and contextual. The diversity of persons is much higher in the wild. Person poses and appearances change depending on different environments. However, existing approaches directly generate persons along with other objects. They ignore the spatial relationship between persons and the layout context. Mixing persons with other texture-based classes, such as snowfields and trees, can easily lead to distorted synthesized results.

In this work, we provide novel solutions to alleviate the above problems. We target at the layout-to-image generation task where the scene layout is given as input. First, we propose the person *intrinsic structure* model. It encodes the person pose and face keypoints as structures in a memory-based network, providing sufficient prior knowledge that aids the further synthesis. Also, the person structures are independently learned, thus avoiding distortions that are usually caused by convolution-based generation. Second, we design a unique person-centered graph neural network to capture the contextual person features from the layout and infer the structure from this context. Third, with the predicted person structure, we use two person appearance refiners to generate a spatial intermediate semantic map from the person structure keypoints. The semantic feature map captures richer person structure and appearance features for person generation. Figure 1 shows the method pipeline.

We carry out experiments on the challenging COCO dataset [3]. The results demonstrate that our method achieves state-of-the-art performance on multiple person-specific metrics including person classification accuracy, person FID, face detection precision and recall. As

shown in Figure 1, our method is able to infer the person pose and face keypoints from the context and synthesize reasonable persons.

2 Related Work

Image Generation from Layouts. The layouts comprising bounding boxes of objects are first used as an intermediate step for the text-to-image generation task [15, 21]. Zhao *et al*. [42] firstly proposed the layout-to-image task for generating images from layouts containing bounding boxes and their corresponding object classes. The simple yet flexible and rich form of layouts attracts more following works. Sun *et al*. [35, 36] proposed LostGANs that adopt ISLA-Norm layers for better and higher resolution image generation. Other works focused on improving the image quality through generating better masks [22] or learning context-aware object representation [13], *etc*. The layout-to-image generation also acts as a sub-process in image generation from scene graphs [1]. However, all the aforementioned methods do not consider the complexity of the person class but only treat person as a common object. The only exception is [40], which provides keypoints for each person and constructs a compositional space for a better person-in-context generation. It requires person keypoints as input for both training and inference, making it less flexible.

Human Pose Transfer. Human pose transfer aims to generate a human image with the target pose from a source image and a source pose. It is related because of the use of pose information to guide the transformation. Ma *et al*. [27] firstly introduced this challenging task and adopted a coarse-to-fine method by directly concatenating source image, source pose, and target pose to model the output image. Others used a two-branch-based framework to separately deal with pose and image to alleviate misalignment. Zhu *et al*. [44] proposed a local attention mechanism to progressively transfer information from source pose to target pose. Tang *et al*. [38] introduced co-attention blocks to model the shape and appearance of persons. These methods use sparse keypoints as the pose representation to guide the transformation, but it is hard to build the correspondence between pose and image using the sparse representation. Following research [6, 10, 28] proposed to first generate a person parsing segmentation map and then render the image. A recent work [8] proposed a new application that generates a person in an image with other persons by predicting natural pose and semantic map of the generated person. However, it requires intensive inputs such as accurate body parts, face, and appearance parsing maps.

3 Method

3.1 Method Design

Our main idea is to design a structure-based generation approach to avoid generating distorted person, such that during inference, the extra person structure information can be inferred from the context and then used to guide better person generation. In the layout-to-image scenario, an intuitive way is to use a conditional generative model (like GAN) to predict the person structure from the layout. However, we argue this is not a fully conditional problem because of the particularity of persons for two concerns: first, although the context can affect the persons behavior to some extent, persons are actually free to have any

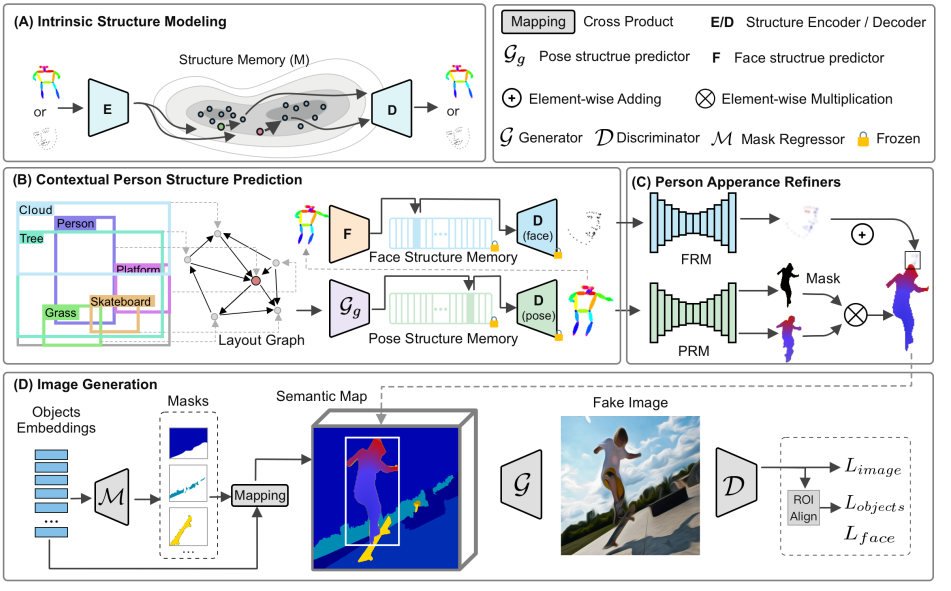

Figure 2: Method overview. (A) Learning the intrinsic structures including pose and face keypoints into a memory-based VAE. (B) Predicting the person pose from the memory conditioned on the layout graph, followed by the face keypoints prediction. (C) Two person refiners PRM and FRM for generating intermediate semantic maps. (D) The person semantic map fused with other object feature maps for final image generation.

pose/face structure as long as they follow some pattern, which we term as *intrinsic structure*. Second, we expect a model to infer reasonable person structures under weak or complicated layout contexts. With these considerations, we thus hope to model the person structure to fulfill its intrinsic as well as contextual requirements: person structures are naturally owned and have diversity in different contexts.

Figure 1 shows the illustration of the three-step pipeline, which reflects our design ideas. (1) **Intrinsic Structure Modeling**. Encoding person pose and face keypoints into a memory-based model is essential to maintain the structure distributions. (2) **Contextual Structure Prediction**. The layout context is used as hints to infer what person pose or face keypoints would be. We propose a layout-graph-based model to predict the possible pose/face pairs from the pre-learned structure memory. (3) **Appearance Refinement**. We convert person structures to pixel space to refine person generation. The idea is to learn an intermediate semantic map for each person and fuse it with other non-person objects. Figure 2 shows the procedure in detail.

3.2 Intrinsic Structure Model

The person structures consist of pose and face keypoints. We propose to use a memory-based design to maintain person intrinsic structures. Concretely, we introduce the person intrinsic structure model (ISM) by using two auto-encoders with a memory bank to save the most representative pose and face keypoints, respectively. As shown in Figure 2-A, we learn an auto-encoder consisting of an encoder E, a decoder D, and a structure memory

bank M to represent person structures with the structure memory bank. The memory bank is of size $N_M \times N_e$, where N_M is the number of stored features and N_e is the dimension of memory vectors. With the encoding $f_e = E(k) \in \mathbb{R}^{N_e}$ given the structure keypoints k, f_e is then mapped onto its closest latent code m_i in the memory:

$$f_e = m_i, i = \arg\min_l \|f_e - m_l\|_2^2, \tag{1}$$

where $l \in \{0, 1, ..., N_M\}$. This is also called vector quantization [29, 30]. Then the reconstructed structure is obtained through the decoding $\hat{k} = D(f_e)$. In training the learning objective is to minimize the loss:

$$L_{structure} = \left\|\hat{k} - k\right\|_2^2 + \ \|\text{sg}[f_e] - m\|_2^2 + \|\text{sg}[m] - f_e\|_2^2 \ . \tag{2}$$

where $\left\|\hat{k} - k\right\|_2^2$ is the reconstruction loss L_{rec} and the rest is the latent loss L_{latent}. $\text{sg}[\cdot]$ is the stop-gradient operation because Eqn. 1 is not differentiable in backpropagation, so a gradient estimation way is used to copy the gradients from the decoder to the encoder.

3.3 Contextual Structure Predictor

We propose the Contextual Structure Predictor (CSP), a person-centered graph neural network ($\mathcal{G}_g$) to infer the most reasonable person pose from the memory while respecting the layout context. Given the predicted pose, we then use a face structure predictor (F) to infer the face structure, as shown in Figure 2-C.

Graph Representation. The layout implies a *layout graph*, in which the objects are represented by the nodes and their relationships are defined on the edges. We first define the relationship categories $\mathcal{R}$. Given a layout L with N_o objects O, the nodes in the layout graph are represented by $O = \{(o_i)_{i=1}^{N_o}\}$ and edges are represented by triples T in the form (o_i, r, o_j), where $o_i \in O$, $o_j \in O$ and $r \in \mathcal{R}$. Specifically, $\mathcal{R}$ includes seven relationship categories: *left of*, *right of*, *above*, *below*, *inside*, *surrounding* and *in image*. The representation for object node o_i is concatenated object category embedding o_i^c, object location feature o_i^{loc} and size feature o_i^{size}. Different from *scene graph* [18], *layout graph* is parsed from the coarse bounding boxes instead of precise segmentation masks of the objects.

Model architecture. In each CSP layer, there are three linear modules g_s, g_o, g_r to separately encode the subjective nodes, the objective nodes, and the relationships. For each edge (o_i, r, o_j) in T, we have $f_r' = g_r(f_i, f_r, f_j)$, where f_i and f_j are inputs for objects and f_r for relationship. As an object can either be subjective or objective, the graph output for each object needs collect features from both directions, so $f_i' = p(\{g_s(f_i, f_r^s, f_m)\} \cup \{g_o(f_n, f_r^o, f_i)\})$, where f_m, f_n and f_r^s, f_r^o are objects and relationships when f_i are subjective and objective, respectively. $(o_i, r^s, o_m) \in T$, $(o_n, r^o, o_i) \in T$. p is an average function to map the collected values to a single vector. The proposed CSP has two unique features: (1) The graph inputs are based on coarse bounding boxes. (2) The model is person-centered, *i.e.*, only person endpoints exist in the last layer despite information flows to each node in intermediate layers.

Contextual Structure Prediction. We use CSP to predict the pose representation from the pretrained memory bank (Section 3.2). Then followed by the pretrained ISM pose decoder,

a pose structure can then be recovered. We further introduce a linear-based pose encoder F to predict the face structure in a similar way since we speculate face structure is more relevant to the pose. We use Cross-Entropy loss (L_{CE}^{pose} and L_{CE}^{face}) as querying error from the pose/face memory banks.

3.4 Person Appearance Refiners and Image Generation

We follow a common layout-to-mask-to-image procedure [1, 35, 36] for final image generation. As shown in Figure 2-D, a shared mask regressor first generates the masks for individual non-person objects, then place them to their layout positions specified by the bounding boxes to create a segmentation mask, or a *semantic map* $\in \mathbb{R}^{C\times H\times W}$, where C is the channel. The *semantic map* is then used for generating the whole image. To fuse person structures into the *semantic map*, we design two structure-guided appearance refiners, *i.e.*, Pose Refinement Module (PRM) and Face Refinement Module (FRM).

Pose and Face Refinement Modules. In this step, we first convert the coordinate-based structure keypoints to their spatial keypoints heatmaps, denoted as $S_p^{map} \in \mathbb{R}^{N_p\times h_p\times w_p}$ for pose and $S_f^{map} \in \mathbb{R}^{N_f\times h_f\times w_f}$ for face. The channels are N_p and N_f, which are the number of keypoints coordinates for pose and face, respectively. h_p/h_f and w_p/w_f denote the spatial height and width. Specifically, each channel of the heatmap encodes a point in the spatial position defined by its coordinate. Next, we use the convolutional-based refiners to generate their semantic maps. For the pose, formulated by $S_p, S_p^M = PRM(S_p^{map}, z_p)$, z_p is the Gaussian noise, $S_p \in \mathbb{R}^{C\times h_p\times w_p}$ and $S_p^M \in \mathbb{R}^{1\times h_p\times w_p}$ are learnt person pose semantic map and mask. Then final result is obtained by $S_p = S_p \bigotimes S_p^M$, where $\bigotimes$ denotes element-wise multiplication. Similarly, face semantic map $S_f = FRM(S_f^{map}, z_f) \in \mathbb{R}^{C\times h_f\times w_f}$, where z_f is the Gaussian noise. Note the output semantic share the same channel dimension C. In practice, the PRM and FRM model are based on fully convolutional network [26] with skip connections [31].

Final Image Generation. Given image-level Gaussian noise z_{img} and the semantic map, we use a generator $\mathcal{G}$ to generate the final image. Specifically, $\mathcal{G}$ is constructed by ResNet blocks (ResBlock) [11] that adopts the self-modulation methods [5]. The objectives in image generation training include: (1) Adversarial Loss. We use the hinge loss [39] for image-level and object-level penalty, where the object features are extracted using ROI Align [12]. (2) $L1$ and Perceptual Loss [17]. We use $L1$ loss and VGG19 [34] as perceptual loss for pixel-level and feature-level matching, respectively. (3) Total Variation (TV) Loss. This loss is used to suppress high-frequency and isolated parts in mask generation. (4) Face Loss. We take face as an object class and calculate its object-level GAN loss. Detailed model structures and losses explanation can be found in the supplementary.

4 Experiments and Results

4.1 Experiments Setup

Datasets. We conduct experiments at two different resolutions 128×128 and 256×256 on COCO-Stuff dataset [3]. We exploit the pose keypoints from COCO [23] and face keypoints from COCO WholeBody [16]. As we focus more on person (in the context) generation, and

	Methods	IS↑	FID↓	DS↑	PFID↓	PAcc(%)↑	FaceAP↑	FaceAR↑
128x128	Real images	14.40±0.89	-	-	-	84.36	0.621	0.580
	Grid2Im (GT Layout) [1]	6.75±0.18	77.26	0.42	45.12	61.82	0.044	0.047
	LostGAN-v2 [36]	**8.24±0.47**	47.73	**0.53**	28.09	73.26	0.146	0.168
	Ours (Pred. KP)	8.06±0.36	**44.73**	0.52	**24.29**	**84.91**	**0.351**	**0.315**
	Ours (GT KP)	8.47±0.35	43.98	0.51	21.73	80.34	0.511	0.491
256x256	Real images	20.82±1.13	-	-	-	84.77	0.771	0.905
	Grid2Im (GT Layout) [1]	8.01±0.23	100.47	0.56	62.32	60.83	0.174	0.212
	LostGAN-v2 [36]	**10.78±0.36**	53.67	0.62	34.24	76.02	0.312	0.373
	Ours (Pred. KP)	10.16±0.33	**53.21**	0.62	**29.42**	**87.5**	**0.638**	**0.581**
	Ours (GT KP)	10.76±0.58	51.57	0.62	26.99	81.41	0.733	0.734

Table 1: The comparison results with the state-of-the-arts on multiple metrics using ground truth person pose and face keypoints (GT KP) and predicted person pose and face keypoints (Pred. KP). ↑ (↓) means the higher (lower) value is better.

need person keypoints for training our model, we construct a new subset of COCO called COCO-Person by following the similar dataset preparation as in [18, 35]. The new dataset includes images containing up to 12 objects that have a minimum size ratio of 0.02 and at least one person with pose keypoint annotation. The same preparation is used to build the testing set. Finally, the training set and testing set contain around 51k and 2.1k images, respectively. For fair comparisons, we retrain the baselines Grid2Im [1] and LostGAN-v2 [36] using their official implementation with their default settings until convergence.

Evaluation Metrics. Commonly used metrics include Inception Score (IS) [32], Fréchet Inception Distance (FID) [14] and Diversity Score (DS) [41] for evaluating image-level generation performance in terms of image quality, distribution, and the diversity degree, respectively. In addition, we design person-specific evaluation metrics for pose and face evaluation. The Person FID (PFID), derived from [37] focuses on the crops of generated persons, and measures the object-level statistical distributions between generated and real persons. Person Accuracy (PAcc) uses a pre-trained 101-layer ResNet classifier for person classification measurement. We then use pre-trained face detector TinaFace [43] to report the face detection Average Accuracy (FaceAP) and Average Recall (FaceAR) over 10 Intersection over Union (IOU) thresholds evenly ranging in [0.5, 0.7]. The ground truth face bounding box is given by [16], as a person bounding box does not indicate where the face will be generated. We lower the IOU thresholds to [0.1, 0.3] for a fair comparison. For methods predicting face positions, the predicted face bounding boxes are used as ground truth.

4.2 Results

Image-level Synthesis Quality. Table 1 summarizes the quantitative comparisons between our method and the state-of-the-arts Grid2Im [1] and LostGAN-v2 [36]. For 128 × 128 resolution, our model using ground truth person keypoints achieves the best performance in image-level metrics Inception Score (IS) and FID. For 256 × 256 resolution, we also have the best FID and competitive IS. Our model using predicted keypoints also performs better than [36] on both resolutions. We also achieve competitive diversity scores.

User Study	Better Person	Better Face	Pred. KP Matching
LostGANv2 [36]	25.42%	9.78%	-
Ours (Pred. KP)	74.58%	90.22%	76.4%

Table 2: Summary of the user studies

Person Synthesis Quality. Table 1 also reflect the person generation quality. Our method achieves better person classification accuracy and FID. Interestingly, our model using predicted person keypoints leads to the best person accuracy which is even higher than real images (84.91 against 84.36 and 87.5 against 84.77 on 128×128 and 256×256 resolution, respectively). We speculate the reason is that the learned person structures in the memory provide the most representative person pose and face priors, which make the generated persons easier to be recognized. Our method using ground truth or predicted keypoints also achieves very high face detection performance in Average Recall and Average Precision.

Qualitatively, Figure 3 presents our generation results. Our method can generate persons with reasonable pose structures, appearances, and faces, while the existing methods [1, 36] failed to generate recognizable persons. The synthesised persons have clear body shapes (*e.g.*, columns A and B) and face details (*e.g.*, columns E and F). This shows that person keypoints are important for guiding person generation in complex scenes and the proposed person appearance refiners can successfully convert the person structural keypoints to semantic parts, such as body, limbs, and faces, to improve person generation.

Context-Aware Person Keypoints Prediction. Given the layouts, our method can infer person keypoints complying with the pre-encoded structures in the Intrinsic Structure Model. As shown in Figure 3, the inferred pose keypoints can successfully cope with the context and even other persons. For example, the predicted person poses in the snowfield context in layout (A) are skiing poses. Layout (C) mainly includes a motorcycle surrounded by two persons, and the inferred results are riding poses. Layout (F) indicates two persons with pizzas and our method can infer two half-body poses near the table. Also, the predicted face keypoints are compatible with the predicted poses. For example, all face keypoints respect the poses, especially in columns (B), (C), and (D) where the face-facing directions are the same as the various facing directions of poses. Furthermore, when there are not many hints from the context, the inferred person poses can also be reasonable and even contain interactions in-between, as shown in columns (B) and (D). However, there still exist limitations when the person bounding box is too challenging. For example, the bottom-left corner of (D) has a small person box but our method tries to infer a full pose. We provide more results and analysis in the supplementary.

User Study. We conducted two user studies to evaluate the human preferences between our results and those generated by state-of-the-art LostGAN-v2 [36]. The first one focuses on the generated person and face quality. The second one evaluates whether the inferred person pose and face keypoints are compatible with the corresponding layouts. We showed 2k randomly sampled testing images (or inferred layouts) evenly to 10 users. Table 2 demonstrates that our method achieves much higher person and face generation ratings. Also, 76.4% of the predicted pose and face keypoints are considered reasonable and compatible with the layouts. We provide more details in the supplementary.

Figure 3: Comparison with existing methods. The proposed approach infers the pose and face keypoints given the layout context, and can generate persons with clear pose structure and recognizable face features. The inferred face keypoints are magnified for a better view.

4.3 Ablation Study

This part evaluates the effectiveness of the proposed three modules and the losses of our approach. The ablation studies are conducted on COCO-Person using 128 $\times$ 128 image resolution. Table 3 summarizes comprehensive results using image-level and person-specific metrics. The *base* approach is a vanilla ResBlock based generator, and the upper part shows the ablation study on Person Appearance Refiners with default ISM and CSP. The lower part shows the effectiveness evaluation of ISM and CSP under the same refiners and loss settings.

Effectiveness of Intrinsic Structure Model. The memory design in the Intrinsic Structure Model is important to maintain person structures. To validate this, we compare with two models: (a) VanillaAE: an MLP based Auto Encoder with the same model architecture but without memory design. (b) VanillaGAN: we directly fed the output features from the Contextual Structure Predictor to a generator, training in an end-to-end way with adversarial

Modules	FID↓	IS↑	DS↑	PAcc(%)↑	PFID↓	FaceAP↑	FaceAR↑
Ours Base	51.77	6.98±0.28	0.52	75.17	33.72	0.014	0.043
+ PRM	46.35	7.92±0.24	0.52	79.62	25.91	0.138	0.220
+ PRM + TV Loss	44.02	8.06±0.18	0.51	81.40	24.79	0.143	0.212
+ PRM + TV Loss + FRM	45.82	7.83±0.34	0.51	**82.44**	23.14	0.246	0.272
+ PRM + TV Loss + FRM + Face Loss (0.01)	**43.39**	8.25±0.47	0.51	80.40	22.50	0.479	0.469
+ PRM + TV Loss + FRM + Face Loss (0.03)	43.98	**8.47±0.35**	0.51	80.34	**21.73**	**0.511**	**0.491**
- ISM + VanillaAE	46.32	7.96±0.21	0.52	76.74	25.10	0.325	0.336
- ISM + VanillaGAN	45.14	8.33±0.25	0.51	74.82	26.38	0.274	0.310
- CSP + SA	45.87	7.79±0.23	0.51	75.18	26.60	0.262	0.280

Table 3: Ablation study on different modules. +/- means adding or removing a module/loss. - followed by + denotes replacing the former module with the later one.

losses. The results (`- ISM + VanillaAE` / `- ISM + VanillaGAN`) are worse than ours. The reasons could be (1) the keypoints have higher variety, and (2) weak or noisy layout condition signals can lead to worse generated keypoints. Our memory-based model overcomes the above limitations and is necessary for keeping the keypoints within a desirable distribution. We present more results in the supplementary.

Effectiveness of Contextual Structure Predictor. We replace the Contextual Structure Predictor with a self-attention-based module (SA) as used in [13]. The layout context is encoded using cross attention of each object's label embedding, and person-specific features are used to predict the structures. The results (`- CSP + SA`) are inferior to ours. The reason is that [13] focuses more on enhancing appearance in a local context while ignoring the relationships between objects in terms of classes, sizes, and positions in the global layout.

Effectiveness of Appearance Refiner and Losses. Using the ground truth person keypoints, we study the ablations of the person pose and face refiners and losses. It shows that the PRM improves person FID, classification accuracy, face detection performance, and even the image-level generation quality (better FID and IS). The TV loss slightly benefits the overall results. FRM further contributes to better face detection AP and AR. The face loss greatly improves the performance in face metrics, and also enhances FID and IS. Although the person classification accuracy is slightly lower, the person FID is further improved.

5 Conclusion

We propose a novel approach to improve the person generation quality in layout-to-image tasks. Firstly, we model the person intrinsic structures including pose and face keypoints using a memory-based model. Secondly, we introduce the context-aware structure predictor from the memory model using a person-centered graph neural network. Thirdly, we employ person refinement modules that fuse person structural information with other objects for final image generation. We show that the contextual structure predictor can predict reasonable person keypoints from layouts. Our method achieves state-of-the-art results in multiple person-related metrics and can generate reasonable person poses and better face regions.

Acknowledgements. This work is supported by RCUK grant CAMERA (EP/M023281/1, EP/T022523/1), Centre for Augmented Reasoning (CAR) at the Australian Institute for Machine Learning, and a gift from Adobe.

References

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

[2] 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/CVF International Conference on Computer Vision*, pages 4502–4511, 2019.

[3] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1209–1218, 2018.

[4] Shu-Yu Chen, Wanchao Su, Lin Gao, Shihong Xia, and Hongbo Fu. Deepfacedrawing: Deep generation of face images from sketches. *ACM Transactions on Graphics (TOG)*, 39(4):72–1, 2020.

[5] Ting Chen, Mario Lucic, Neil Houlsby, and Sylvain Gelly. On self modulation for generative adversarial networks. *arXiv preprint arXiv:1810.01365*, 2018.

[6] Haoye Dong, Xiaodan Liang, Ke Gong, Hanjiang Lai, Jia Zhu, and Jian Yin. Soft-gated warping-gan for pose-guided person image synthesis. *arXiv preprint arXiv:1810.11610*, 2018.

[7] Haoye Dong, Xiaodan Liang, Xiaohui Shen, Bochao Wang, Hanjiang Lai, Jia Zhu, Zhiting Hu, and Jian Yin. Towards multi-pose guided virtual try-on network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9026–9035, 2019.

[8] Oran Gafni and Lior Wolf. Wish you were here: Context-aware human generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7840–7849, 2020.

[9] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, page 2672–2680, 2014.

[10] Xintong Han, Xiaojun Hu, Weilin Huang, and Matthew R Scott. Clothflow: A flow-based model for clothed person generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10471–10480, 2019.

[11] 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.

[12] 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.

[13] Sen He, Wentong Liao, Michael Ying Yang, Yongxin Yang, Yi-Zhe Song, Bodo Rosenhahn, and Tao Xiang. Context-aware layout to image generation with enhanced object appearance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15049–15058, 2021.

[14] 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. *Advances in neural information processing systems*, 30, 2017.

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

[16] Sheng Jin, Lumin Xu, Jin Xu, Can Wang, Wentao Liu, Chen Qian, Wanli Ouyang, and Ping Luo. Whole-body human pose estimation in the wild. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[17] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016.

[18] 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.

[19] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019.

[20] Zorah Lahner, Daniel Cremers, and Tony Tung. Deepwrinkles: Accurate and realistic clothing modeling. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 667–684, 2018.

[21] 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/CVF Conference on Computer Vision and Pattern Recognition*, pages 12174–12182, 2019.

[22] Zejian Li, Jingyu Wu, Immanuel Koh, Yongchuan Tang, and Lingyun Sun. Image synthesis from layout with locality-aware mask adaption. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13819–13828, 2021.

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

[24] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[25] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1096–1104, 2016.

[26] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[27] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. Pose guided person image generation. *arXiv preprint arXiv:1705.09368*, 2017.

[28] Yifang Men, Yiming Mao, Yuning Jiang, Wei-Ying Ma, and Zhouhui Lian. Controllable person image synthesis with attribute-decomposed gan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5084–5093, 2020.

[29] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. *arXiv preprint arXiv:1711.00937*, 2017.

[30] Ali Razavi, Aaron van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. In *Advances in neural information processing systems*, pages 14866–14876, 2019.

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

[32] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29:2234–2242, 2016.

[33] Kripasindhu Sarkar, Lingjie Liu, Vladislav Golyanik, and Christian Theobalt. Humangan: A generative model of human images. In *2021 International Conference on 3D Vision (3DV)*, pages 258–267. IEEE, 2021.

[34] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[35] Wei Sun and Tianfu Wu. Image synthesis from reconfigurable layout and style. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10531–10540, 2019.

[36] Wei Sun and Tianfu Wu. Learning layout and style reconfigurable gans for controllable image synthesis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (to appear), 2021.

[37] Tristan Sylvain, Pengchuan Zhang, Yoshua Bengio, R Devon Hjelm, and Shikhar Sharma. Object-centric image generation from layouts. *arXiv preprint arXiv:2003.07449*, 1(2):4, 2020.

[38] Hao Tang, Song Bai, Li Zhang, Philip HS Torr, and Nicu Sebe. Xinggan for person image generation. In *European Conference on Computer Vision*, pages 717–734. Springer, 2020.

[39] Dustin Tran, Rajesh Ranganath, and David M Blei. Hierarchical implicit models and likelihood-free variational inference. *arXiv preprint arXiv:1702.08896*, 2017.

[40] Weidong Yin, Ziwei Liu, and Leonid Sigal. Person-in-context synthesis with compositional structural space. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2827–2836, 2021.

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

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

[43] Yanjia Zhu, Hongxiang Cai, Shuhan Zhang, Chenhao Wang, and Yichao Xiong. Tinaface: Strong but simple baseline for face detection. *arXiv preprint arXiv:2011.13183*, 2020.

[44] Zhen Zhu, Tengteng Huang, Baoguang Shi, Miao Yu, Bofei Wang, and Xiang Bai. Progressive pose attention transfer for person image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2347–2356, 2019.