Multi-scale Contrastive Learning for Complex Scene Generation

Hanbit Lee Youna Kim Sang-goo Lee
Seoul National University, Seoul, Korea
{skcheon,anna9812,sglee}@europa.snu.ac.kr

Abstract

Recent advances in Generative Adversarial Networks (GANs) have enabled photo-realistic synthesis of single object images. Yet, modeling more complex distributions, such as scenes with multiple objects, remains challenging. The difficulty stems from the incalculable variety of scene configurations which contain multiple objects of different categories placed at various locations. In this paper, we aim to alleviate the difficulty by enhancing the discriminative ability of the discriminator through a locally defined self-supervised pretext task. To this end, we design a discriminator to leverage multi-scale local feedback that guides the generator to better model local semantic structures in the scene. Then, we require the discriminator to carry out pixel-level contrastive learning at multiple scales to enhance discriminative capability on local regions. Experimental results on several challenging scene datasets show that our method improves the synthesis quality by a substantial margin compared to state-of-the-art baselines.

1. Introduction

In recent years, generative adversarial networks (GAN) [11] have achieved significant improvements due to extensive studies on network structures [35, 53, 3, 24, 25, 38], objective functions [30, 1, 27], and regularization techniques [13, 32, 31]. Now GAN models can produce high-quality images that are almost indistinguishable from real ones, showing impressive results in the wide range of object classes including human faces [24], animals [3, 38], and cars [25]. Despite these successes, when it comes to more complex images such as scenes with multiple objects, they easily fail to achieve the same level of realism as in single object images [4, 10].

In single object images, there is a common layout of each component, allowing it easier for the discriminator to supervise where and how each component should be synthesized to result in a realistic image. For instance, each component of dog's face, e.g., eyes, nose, and mouth, may vary in shapes and proportions, but remain in a common layout that forms the face. On the other hand, natural scene images exhibit much more diverse and complex distributions as they include a collection of objects in various sizes, shapes, and spatial locations [4, 40, 18]. Therefore, it is much harder for the discriminator to learn multi-layered differences between real and fake images from local semantic structures, such as objects, to overall scene layouts [39, 10]. As a result, even state-of-the-art GAN models produce unsatisfactory results of limited distribution coverage and low synthesis quality with messy layouts and incomplete internal objects.

In this work, we explore a way to improve discriminative ability on such complex scenes through a self-supervised pretext task assigned to the discriminator. Self-supervised representation learning has been extensively studied in recent years and shown to yield beneficial representations for various downstream tasks [5, 14, 12]. The progress continues to generative models and recent studies have shown GAN models also can be improved by leveraging various self-supervised pretext tasks such as rotation prediction [6, 41, 17], consistency regularization [54, 56] and contrastive learning [57, 21, 49]. While successful, existing studies mainly focus on enhancing image-level global representations especially for single-object images, thus the improvement tend to be limited for more complex data distributions, such as scene images containing various local objects.

To better model complex local semantic structures in the scene images, we propose to enhance local representations as well as the global representation with auxiliary pretext tasks locally defined and at multiple scales. To this end, we design a multi-scale discriminator having multi-level branches where each branch processes local patches of different sizes. Branch at each scale produces per-pixel auxiliary representations as well as per-pixel discriminator logits. These auxiliary representations are used to perform pixel-level contrastive learning to enhance per-pixel classification task. Both tasks are defined for each scale level and jointly optimized across all scales, thereby the discriminator could improve local-to-global discriminative ability to better model local structures in complex scenes at various scales.

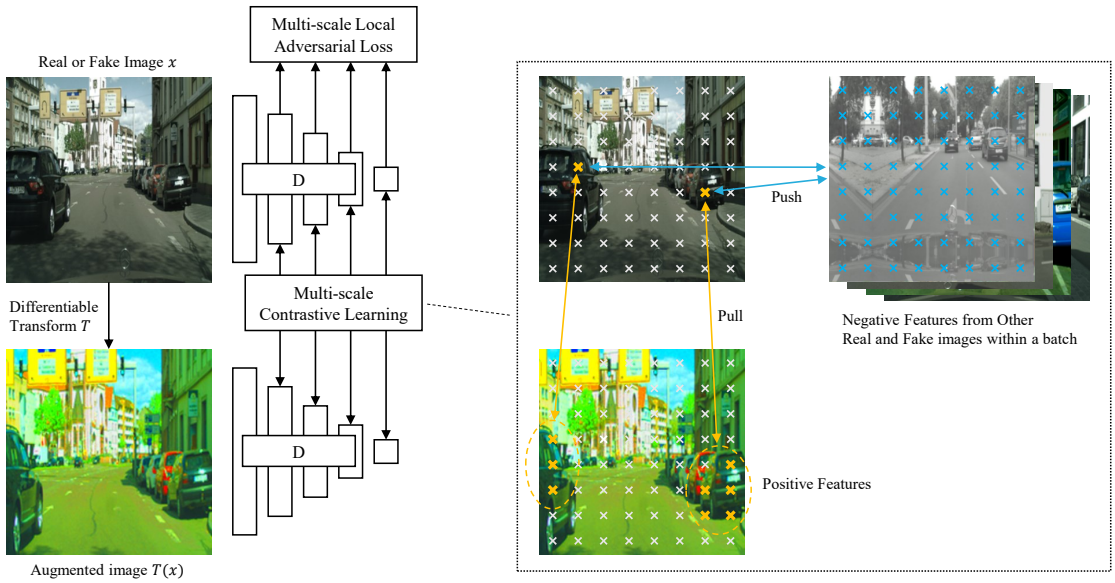


Figure 1. **Overview of the proposed method.** Our approach improves the discriminative ability by two means. First, we train the model through multi-scale local adversarial feedback generated from feature pyramid of backbone network. To further enhance the feedback, the discriminator performs multi-scale contrastive learning which aims to distinguish between positive features from the augmented image $T(x)$ and negative features from other irrelevant images.

We evaluate our method on several challenging scene image datasets with metrics for both scene-level and object-level synthesis quality. Compared to recent state-of-the-art GAN models, our method consistently achieves better results in terms of visual quality and diversity. In particular, our method significantly improves synthesis quality of individual objects in the scene, demonstrating that multi-scale representation learning effectively enhances the adversarial feedback to better model local semantic structures.

2. Related Work

2.1. Discriminator Design for GAN

Discriminator's ability to distinguish between real and fake images plays a critical role in GAN training, since the generator entirely relies on the feedback signal passed from the discriminator. Such ability has been significantly improved with the advances in discriminator architectures, from multi-layer perceptrons [11] to convolutional networks [35, 22], residual networks [32, 24], and self-attention based models [53, 3, 52]. However, even state-of-the-art models still struggle in modeling complex scenes, since they rely solely on global discriminator feedback therefore missing high frequency details. To alleviate the problem, we redesign the discriminator to utilize local feedback on multiple scales.

Local discriminator feedback has been used in various conditional image generation tasks [58, 19, 33, 8, 51] in the form of PatchGAN discriminator [20]. To cover multiple scales, Wang et al. [43] propose to use multiple PatchGAN discriminators to process each image interpolated at different resolutions. These architectures have been helpful for modeling high frequency patterns, but they rely on explicit conditions such as segmentation maps or input images, to model global layouts. In contrast, our method allows to model local-to-global structures by utilizing multi-scale feedback which emerges from natural hierarchy inherent in the pyramidal features of backbone network. Recently proposed ProjectedGAN [37] has also verified the usefulness of multi-scale features, but they focus on mixing multiple levels of pretrained features rather than utilizing local feedback.

2.2. Self-supervised Learning for GAN

Self-supervised learning has been recognized as one of the most influential methodologies in recent years as it can learn informative representations from a large amount of unlabeled data. Recent studies have shown that GAN training can also benefit from various self-supervised pretext tasks. A group of works [6, 41, 17] have shown that the rotation prediction task prevents catastrophic forgetting in GAN and leads to better results. Consistency regularization [54, 56] stabilizes GAN training by imposing consistency of discriminator output between a clean image and its augmented version. More recently, several studies have explored the use of the instance discrimination task [45, 14, 5] as an auxiliary task to further enhance the discriminator [57, 21, 49]. The self-supervised pretext tasks generally

involve various image transformation functions to acquire different views of an image. In GAN training, differentiable image transformations [23, 55] applied on both real and fake images have shown to stabilize the training in limited data regimes and improve the data efficiency. Our work relies on previous findings on improved GAN training with self-supervised pretext tasks. However, while all previous studies focus on enhancing global representation space by integrating image-level tasks, in this work, we seek to enhance region-level representations to improve discriminative ability on local features.

2.3. Dense Representation Learning

Recent studies on self-supervised representation learning mainly focus on image-level representations for object-centric images, i.e., ImageNet [9]. Despite their success, the image-level global representations are often sub-optimal for general vision tasks defined on complex scenes, as globally pooled representations lose spatial information of local objects. Therefore, more recent works attempt to learn pixel-level [34, 48, 44] or region-level [36, 46, 47] representations and have achieved meaningful improvements in dense prediction downstream tasks such as object detection and instance segmentation. We repurpose the dense representation learning as a mean to aid the real-fake discrimination on multiple scales, thereby validate its efficacy on improving synthesis quality of local objects in complex scenes.

3. Method

In this section, we describe the proposed method, namely, Multi-scale Contrastive Discriminator (***MsConD***) in detail. First, we briefly introduce the image synthesis methodology of standard GAN in Section 3.1. We then describe the improved discriminator architecture in Section 3.2, followed by multi-scale pixel-level contrastive learning that further enhances the discriminator in Section 3.3 and finally the full objective function which optimizes the entire network in Section 3.4.

3.1. Generative Adversarial Networks

A standard GAN involves a minimax optimization between two networks, a generator G and a discriminator D as follows:

$$\min_G \max_D \mathcal{L}_{adv}(G, D) = \\ \mathbb{E}_{x\sim p_{data}}\left[\log D(x)\right] + \mathbb{E}_{z\sim p_z}\left[\log(1 - D(G(z)))\right], \quad (1)$$

where p_{data} is an empirical data distribution and p_z is a known prior distribution. D aims to distinguish between real images and generated images while G aims to synthesize realistic-looking images so that they can be distinguished as real ones by D. Intuitively, since G is optimized by the criteria presented by D, the synthesis quality is limited to the discriminative ability of D. Therefore, this work focuses on improving the discriminative ability by two means: redesigning the discriminator architecture and introducing an effective auxiliary task for it.

3.2. Multi-scale Discriminator with Multi-level Branches

In unconditional image synthesis, a discriminator is typically equipped with several sub-sampling layers that progressively downsample the input high-resolution images into lower resolution features constructing pyramidal feature maps [35, 3, 53, 24]. To enable discrimination of each local feature in the feature maps, we use branches for each scale l to translate the intermediate features into corresponding local outputs. Each branch consists of three components: residual blocks ϕ^l_{res}, a classification head ϕ^l_{disc}, and a projection head ϕ^l_{proj}. All components are implemented with 1×1 convolution layers to process each local feature individually. Figure 2 (left) shows the proposed discriminator design.

Concretely, our discriminator D is composed of backbone network F and per-scale branch networks $\phi^l = \{\phi^l_{res}, \phi^l_{disc}, \phi^l_{proj}\}$. Given an input image, the backbone network F produces multi-scale feature maps. We denote the feature map at scale level l as f_l. f_l is first transformed into h_l of the same shape by ϕ^l_{res} and then h^l is processed by two separate head networks, a real/fake classification head ϕ^l_{disc} and a projection head ϕ^l_{proj} to produce two outputs U^l and V^l.

$$h_l = \phi^l_{res}(f_l) \in \mathbb{R}^{H_l \times W_l \times C_h} \quad (2)$$

$$U_l = \phi^l_{disc}(h_l) \in \mathbb{R}^{H_l \times W_l \times 1} \quad (3)$$

$$V_l = \phi^l_{proj}(h_l) \in \mathbb{R}^{H_l \times W_l \times C_p}, \quad (4)$$

where C_p is number of channels of the projection output.

We denote the classification head output U_l and the projection output V_l for an input image x as $D^l_{disc}(x)$ and $D^l_{proj}(x)$, respectively. $D^l_{disc}(x)$ is used to compute per-pixel adversarial loss at l-th scale while $D^l_{proj}(x)$ is used to perform pixel-level contrastive learning which will be described in the following section. The adversarial loss at l-th scale is computed by averaging all per-pixel adversarial losses as follows:

$$\mathcal{L}^l_{adv}(G, D) = \mathbb{E}_x\left[\frac{1}{H_l W_l}\sum_{i,j}\log\left[D^l_{disc}(x)\right]_{i,j}\right] \\ + \mathbb{E}_z\left[\frac{1}{H_l W_l}\sum_{i,j}\log\left(1 - \left[D^l_{disc}(G(z))\right]_{i,j}\right)\right], \quad (5)$$

where $\left[D^l_{disc}(x)\right]_{i,j}$ refers to the classification output at pixel (i, j). As shown in Figure 2 (left), the global rep-

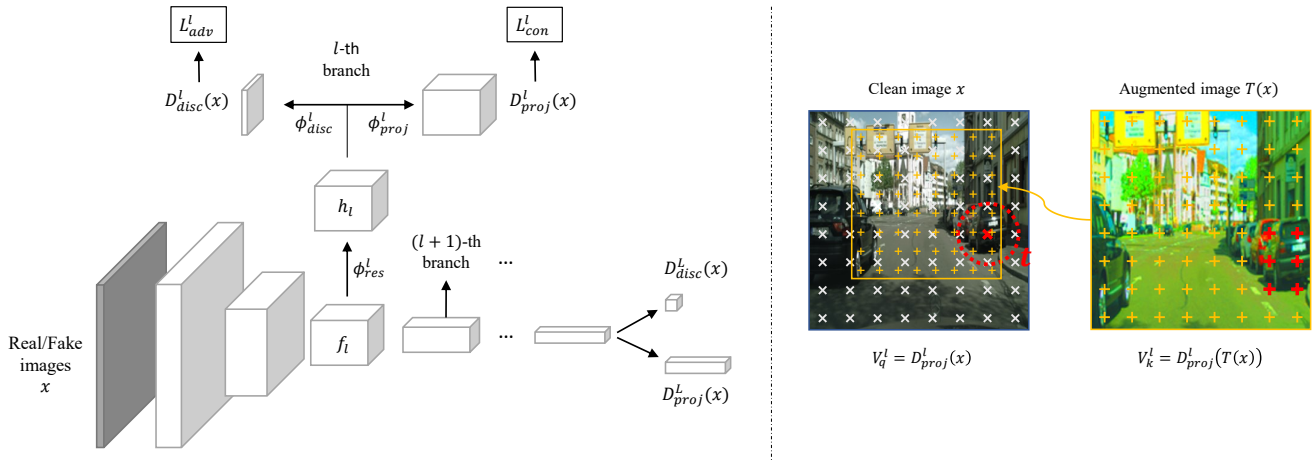


Figure 2. **Discriminator architecture (left).** Our discriminator produces multi-scale outputs from the intermediate features at different layers via layer-wise branches. At each layer, the intermediate feature map f_l is mapped into two different outputs: the discriminator output $D^l_{disc}(x)$ and the projection output $D^l_{proj}(x)$. Two outputs are used to compute the per-pixel adversarial loss $\mathcal{L}^l_{adv}$ and the pixel-level contrastive loss $\mathcal{L}^l_{con}$, respectively. **Spatially consistent Pixel-level contrastive learning (right).** For each per-pixel feature (red $\times$) in the projection output of the clean image x, the positive feature set (red $+$) in the augmented image $T(x)$ is defined with a predefined distance threshold t. On the other hand, the negative feature set are constructed with features from images in the same mini-batch.

resentation at the top of the backbone network is likewise mapped to the global discriminator output and the global projection output, which are used to compute the adversarial and contrastive losses, respectively. See Section 1 in Supplementary Material for more details.

3.3. Multi-scale Contrastive Learning for GAN

The redesigned discriminator learns to differentiate between real and fake images based on local-to-global region-level decisions. To further enhance the discriminative ability, we propose to assign the discriminator an auxiliary self-supervised task designed to enrich the region-level representation on which each decision is performed.

Given a clean image x, its augmented view $T(x)$ is obtained by applying a differentiable transformation T. Then the respective projection outputs V^l_q and V^l_k at l-th scale are extracted through the projection branch:

$$V^l_q = D^l_{proj}(x) \in \mathbb{R}^{H_l W_l \times C_p} \tag{6}$$

$$V^l_k = D^l_{proj}\left(T(x)\right) \in \mathbb{R}^{H_l W_l \times C_p}. \tag{7}$$

Instance discrimination task [45, 14, 5] is a widely adopted pretext task in self-supervised representation learning. Typically, it conducts training by contrasting the positive views of an instance from the negative views which are irrelevant to the instance. In image-level instance discrimination task, the positive features can be easily obtained by simply applying random transformations to an image. However, our objective is to learn local representations to support real-fake decision on individual local features, thereby an instance for the task no longer represents the whole image but local regions of an image. In this case, the positive features should be cautiously identified to ensure sufficient overlap between the regions represented by the features. Otherwise, it can interfere with representation learning by associating areas that are completely unrelated to each other in the image.

In this work, we identify two feature vectors from V^l_q and V^l_k as a positive pair if they are close enough to contain the same region in the image [48]. The spatial closeness is measured by the Euclidean distance between the coordinates of two feature vectors in the image space. Figure 2 (right) shows an example. Concretely, we warp the pixels in V^l_k into the clean image space to obtain the reference coordinates and compute all-pair Euclidean distances between the coordinates of feature vectors in the two feature maps V^l_q and V^l_k. For each feature vector $v_q \in \mathbb{R}^{C_p}$ in V^l_q, we define the positive feature set from V^l_k as follows:

$$pos(v_q) = \{v_k \in V^l_k : dist(v_q, v_k) < t\}, \tag{8}$$

where $dist(v_q, v_k)$ denotes the Euclidean distance between the coordinates of feature vectors v_q and v_k in the clean image space, and t is predefined distance threshold.

On the other hand, we construct the negative feature set $neg(v_q)$ with the same level features from other images in the same mini-batch. It is worth noting that we use both real and fake images for negative features in order to construct larger negative set. We empirically observed that this leads to a slight performance improvement. With positive and negative feature sets, the contrastive loss at l-th layer can be

formulated as:

$$\mathcal{L}^l_{con}(x, T(x)) = \sum_{v_q \in V^l_q} -\log \frac{\sum_{v_k \in pos(v_q)} e^{v_q \cdot v_k/\tau}}{\sum_{v_k \in pos(v_q)} e^{v_q \cdot v_k/\tau} + \sum_{v_k \in neg(v_q)} e^{v_q \cdot v_k/\tau}}, \tag{9}$$

where τ is a temperature hyper-parameter which is set to 0.3. We normalize the feature vector v_q and v_k before computing the contrastive loss thus the dot product between them assesses the cosine similarity between the vectors.

We demand the discriminator to solve the same task for fake images $G(z)$ and their augmented views $T(G(z))$. In this way, the discriminator can learn from infinite samples generated by the model beyond the limited amount real images [49]. Finally, the contrastive loss at l-th scale is computed using contrastive losses applied on both real and fake sample as follows:

$$\begin{aligned}\mathcal{L}^l_{con}(D) =& \mathbb{E}_x \left[\mathcal{L}^l_{con}(x, T(x))\right] + \\ & \mathbb{E}_z \left[\mathcal{L}^l_{con}\left(G(z), T(G(z))\right)\right]. \end{aligned} \tag{10}$$

3.4. Full objective

The total loss for MsConD is calculated using adversarial loss and contrastive loss summed on all scales as follows:

$$\mathcal{L}_{adv}(G, D) = \sum_l \mathcal{L}^l_{adv}(G, D) \tag{11}$$

$$\mathcal{L}_{con}(D) = \sum_l \mathcal{L}^l_{con}(D) \tag{12}$$

$$\min_G \max_D \mathcal{L}_{adv}(G, D) - \lambda \mathcal{L}_{con}(D), \tag{13}$$

where λ controls the strength of contrastive loss. We found that $\lambda = 0.2$ gives desirable balance between the two loss terms, and we use this value for all experiments.

3.5. Implementation and Training

The MsConD is implemented upon the resnet-based discriminator of StyleGAN2 [25]. We adopt the training techniques used in StyleGAN2 including lazy R1 regularization and path length regularization. For augmentation T, we use differentiable transformations including pixel blitting, geometric and color transformations following StyleGAN2-ADA [23]. One notable difference is that StyleGAN2 computes the R1 regularization loss using the global discriminator output, whereas MsConD computes the R1 losses for each branch output and regularizes the network with the sum of the losses. We use Adam optimizer with batch size of 32, learning rate of 0.002, $\beta_1 = 0.0$ and $\beta_2 = 0.99$. All models including the baselines have been trained for the same number of training steps (10 million images).

4. Experiments

Datasets. We evaluate the proposed method on three challenging scene image datasets. Cityscapes [7] contains 25k images of street scenes recorded from a driving car in 50 cities. LSUN [50] is a large collection of scene images covering wide range of indoor and outdoor scenes. Among them, we choose livingroom and kitchen dataset as benchmark datasets since they exhibit highly complex data distributions derived from diverse scene layouts with various objects. Livingroom and kitchen datasets contain 1.3 million and 2.2 million scene images, respectively. All images used in the experiments are resized to 256×256 resolution.

Evaluation metrics. To quantitatively evaluate the synthesis quality, we use Frechet inception distance [15], Kernel inception distance [2], Precision, and Recall [26]. Following the previous works [16, 23], all metrics are calculated using 50,000 fake images and all training images.

Perceptual quality of scene images is largely determined by synthesis quality of individual objects within the scene. Since there is no object-level label in the evaluation dataset, we employ a pretrained object detector to identify objects depicted in both real and generated scenes. Then we calculate FID scores using the crops of detected objects for each object category. The object crops detected from 50,000 real images are used to obtain per-category real distributions. For a fair comparison, we calculate the FID using the same number of object crops from each model. We use YOLOR [42] object detector trained on MS-COCO [28].

Comparison methods. We use several recent competitive models as our baselines. UnetGAN [38] and StyleGAN2 [25] are utilized to compare different discriminator architectures. ADA [23] uses differentiable data augmentations, while InsGen [49] applies image-level instance discrimination upon ADA. ProjectedGAN [37] is a parallel state-of-the-art study using multiple discriminators to leverage multi-scale features from pretrained networks. We use officially released code base of baseline methods except for UnetGAN where we employ better backbone of StyleGAN2.

We use the same StyleGAN2 generator for all methods to fairly compare the discriminator ability except for ProjectedGAN where the lighter generator, i.e., FastGAN [29] generator, has been reported to perform better. Since we observed that most methods are highly sensitive to the R1 penalty term [31], we carefully explored the best performing R1 weights in the range of 1 to 50 for each method. For ADA and InsGen, we use the same set of image transformations consisting of pixel blitting, geometric and color transformations, which have shown the most stable results in the literature. For other hyper-parameters, we use the same values as originally proposed in each paper.

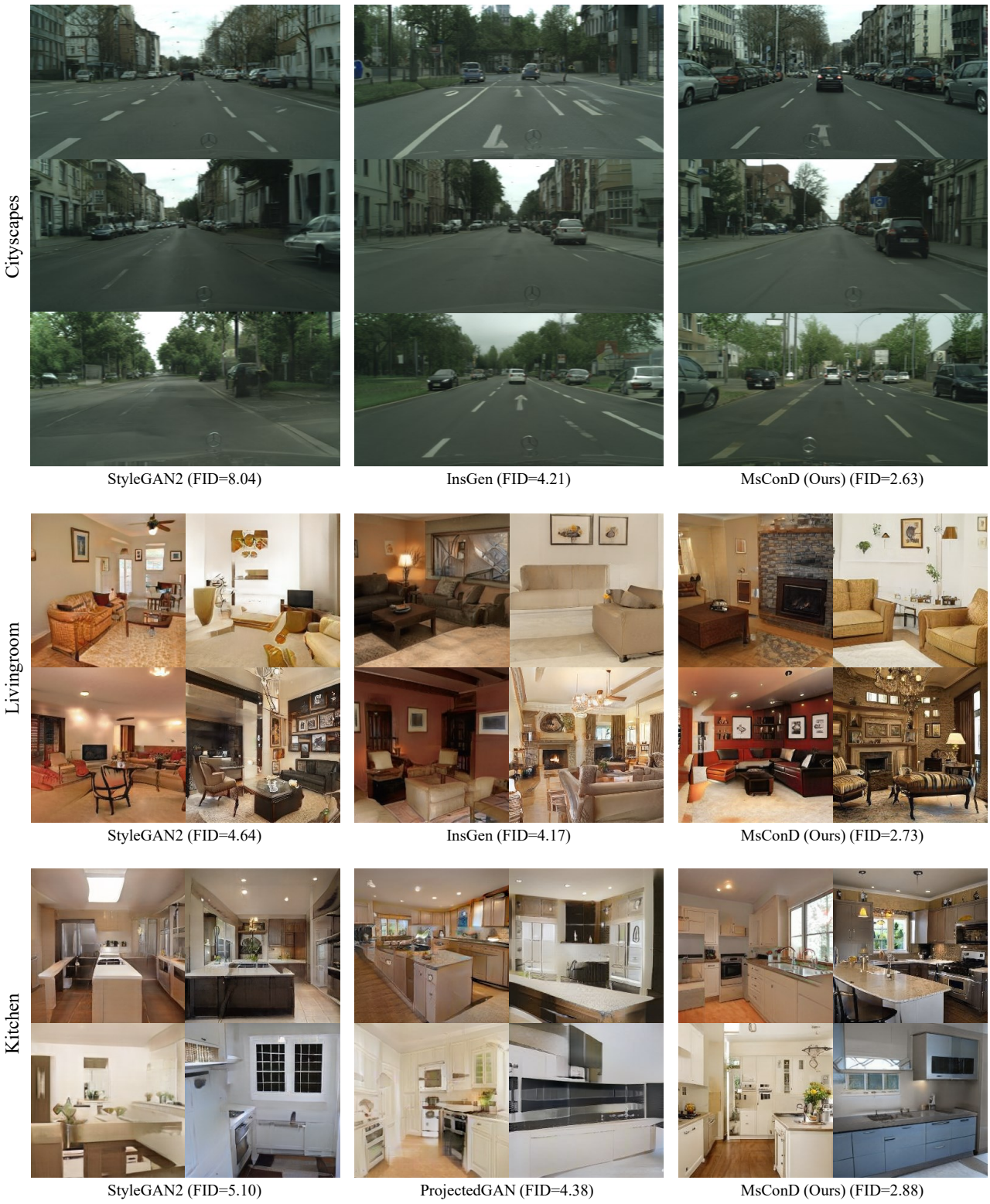


Figure 3. **Comparison of generated samples.** We demand pre-trained generators to reconstruct the same real image query to compare aligned results. Zoom in for details. Compared to the baselines, MsConD shows better results with more realistic scene components, such as cars, buildings, tables, sofas, lamps, sink, drawers, etc. See Section 3 in Supplementary Material for more samples.

Method	Cityscapes				Livingroom				Kitchen			
	FID↓	KID↓	Prec↑	Rec↑	FID↓	KID↓	Prec↑	Rec↑	FID↓	KID↓	Prec↑	Rec↑
UnetGAN [38]	14.47	8.41	0.434	0.132	6.73	3.92	0.518	0.265	6.71	4.13	0.528	0.290
StyleGAN2 [25]	8.04	5.27	0.539	0.260	4.64	2.22	0.512	0.268	5.10	2.58	0.530	0.305
ADA [23]	5.03	1.86	0.604	0.221	4.95	2.34	0.507	0.267	6.47	3.62	0.484	0.272
InsGen [49]	4.21	1.64	0.583	0.349	4.17	2.09	0.556	0.318	5.76	2.57	0.535	0.312
ProjectedGAN [37]	5.07	1.94	**0.620**	0.270	5.51	2.36	**0.571**	0.273	4.38	2.11	**0.587**	0.250
MsConD (Ours)	**2.63**	**0.99**	0.605	**0.485**	**2.73**	**1.14**	0.538	**0.431**	**2.88**	**0.97**	0.544	**0.429**

Table 1. **Comparison result on Scene-level generation metrics.** FID, KID, Precision, and Recall are reported as evaluation metrics. We highlight the best in **bold** and second best with underline. For KID, we report KID $\times 10^3$.

Cityscapes	car		person		traffic light		truck		bus	
	FID↓	KID↓	FID↓	KID↓	FID↓	KID↓	FID↓	KID↓	FID↓	KID↓
StyleGAN2 [25]	10.96	7.38	26.75	16.96	86.71	51.06	34.06	16.46	61.45	24.51
InsGen [49]	7.89	4.72	26.04	14.52	81.52	40.53	36.65	15.50	64.01	25.30
ProjectedGAN [37]	20.12	11.12	32.59	30.41	96.51	32.78	57.14	12.39	76.50	37.46
MsConD (Ours)	**4.57**	**2.64**	**17.02**	**8.60**	**48.16**	**17.50**	**23.13**	**6.14**	**52.80**	**18.90**
Livingroom	couch		chair		potted plant		tv		vase	
	FID↓	KID↓	FID↓	KID↓	FID↓	KID↓	FID↓	KID↓	FID↓	KID↓
StyleGAN2 [25]	11.21	8.06	14.58	8.64	16.09	8.14	12.22	9.02	40.19	6.19
InsGen [49]	9.57	7.01	14.22	9.03	14.20	7.42	14.62	11.23	39.44	5.75
ProjectedGAN [37]	8.60	4.68	21.77	10.18	22.16	12.03	12.76	7.12	42.87	6.44
MsConD (Ours)	**4.30**	**2.19**	**8.64**	**3.66**	**10.15**	**2.92**	**9.51**	**4.47**	**35.86**	**3.52**
Kitchen	oven		chair		microwave		potted plant		refrigerator	
	FID↓	KID↓	FID↓	KID↓	FID↓	KID↓	FID↓	KID↓	FID↓	KID↓
StyleGAN2 [25]	8.91	4.98	19.89	11.32	11.82	7.57	21.77	10.89	19.11	10.05
InsGen [49]	9.15	4.87	19.70	10.49	11.74	6.14	20.18	7.89	17.45	8.83
ProjectedGAN [37]	13.78	6.01	21.74	9.42	20.06	12.29	25.95	8.24	21.55	13.60
MsConD (Ours)	**5.01**	**1.36**	**13.28**	**5.54**	**8.85**	**3.33**	**15.17**	**3.40**	**12.36**	**4.86**

Table 2. **Object-level metrics for each object category.** We compute FID and KID scores on the crops of detected objects for each object category. Different object categories are detected according to different data domains. For KID, we report KID $\times 10^3$.

4.1. Comparison to State-of-the-Art

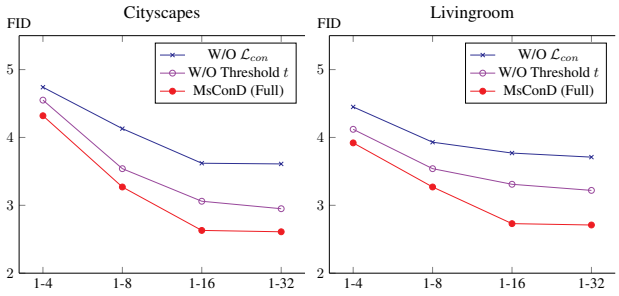
Scene-level Metrics. Table 1 shows the quantitative comparison result using standard GAN metrics. In terms of FID, our method outperforms all other baselines, achieving 37%, 35% and 33% relative improvements in each dataset compared to the best baseline methods. Our method achieves significantly improved recall across all datasets, demonstrating its capability to synthesize diverse scene images. Albeit ProjectedGAN achieves the highest precision, we empirically observed that it produces larger fraction of artifacts than other methods. This is also verified by its inferior object synthesis quality in Table 2. We speculate that the pretrained feature space learned on object-centric images, i.e., ImageNet, may not be best suited for learning more complex data distributions.

Object-level Metrics. To validate if our method improves the synthesis quality of individual objects in the scene, we measure FID, KID, and IS scores for top 5 most frequent object categories detected in each data domain. Table 2 shows the comparative result. In all object categories, our method achieves significantly improved metric scores over the baselines. These results validate that the proposed MsConD effectively incentivizes the generator to improve local details and produce more realistic objects in the scene images. Figure 3 provides visual comparison between samples generated by different methods. As shown, our method produces more realistic scene details with a well arranged layout over other methods. See Section 3 in Supplementary Material for more result.

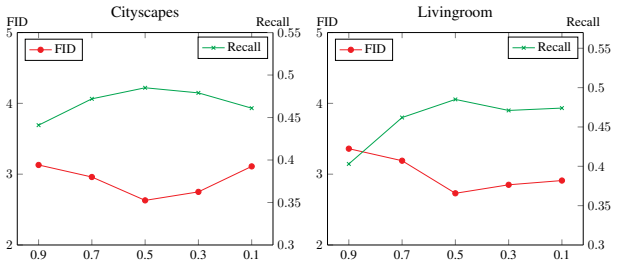
4.2. Ablation study

In this section, we conduct an ablation study to investigate how each component of MsConD contributes to the generation performance. Figure 4 summarizes the ablation results. Figure 4 (a) shows scene-level FID when MsConD is trained with different scales of feature maps. We compare the model with/without multi-scale contrastive loss $\mathcal{L}_{con}$ to

Figure 4. **Quantitative Ablation Result.** (a) Comparison results in terms of scene-level FID under different model configurations. For each configuration, we plot the results when different scales of feature maps are utilized. For example, '1-8' means that the model uses feature maps whose height is from 1 to 8. (b) The effect of distance threshold t using 1-16 full MsConD.



(a) Ablating each component of MsConD with varying scales.



(b) Effect of distance threshold t.

validate its efficacy. As shown in the result, in both cases, the generation performance increases as more feature maps are utilized, yet the performance is prominently boosted through multi-scale contrastive learning. We also report the result with contrastive learning but without distance threshold t to verify the effectiveness of our strategy for positive feature sampling. In this case, we use all pairs of local features from augmented images as positive samples without any spatial constraints. The result shows that the performance gain is far limited without the distance threshold, since semantically irrelevant local features impede the representation learning.

To further investigate the effect of distance threshold, we report FID and Recall with varying thresholds in Figure 4 (b). The performance deteriorates if t is too high or too low. When t is too low, only a narrow range of features are utilized as positive features, degrading the sample diversity. On the other hand, if the t is too high, irrelevant features could be treated as the positive features, and possibly hinder the learning.

Figure 5 shows samples generated by MsConD trained under different configurations. When the model is trained without multi-scale adversarial loss (W/O MS Adv.), local objects tend to be incomplete and discontinued as the generator is not provided with local feedback. On the other hand,

Figure 5. **Qualitative ablation result on Livingroom dataset.** Comparison of representative samples under different model configurations. Zoom in for details.

when the model is trained only with multi-scale adversarial loss (W/O MS Con.), we observe repetitive patterns often appear in the generated images, which are known to be a common side-effect of PatchGAN discriminator. These artifacts are prominently mitigated in the results of MsConD, resulting in more realistic local objects. See Section 2 in Supplementary Material for additional ablation study and analysis.

5. Conclusion

Despite recent advances of GANs, challenges still remain in modeling more complex data distributions. One of these challenges lies in learning complex and diverse local structures, such as individual objects in scene images. To mitigate the difficulty, we redesign the discriminator to leverage local feedback from multi-scale features through multi-level branches. In addition, we propose to enrich the multi-scale representations through contrastive learning in order to further enhance the multi-scale GAN feedback. Experimental results show our method improves the local-to-global discriminative ability, thus effectively incentivizes the generator to synthesize diverse scene images with realistic details.

Acknowledgements

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korean government (MSIT) (No.2021-0-00302). This work was also supported in part by the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT, and Future Planning through the Basic Science Research Program under Grant 2020R1F1A1075952.

References

[1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017.

[2] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. In *International Conference on Learning Representations*, 2018.

[3] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *International Conference on Learning Representations*, 2019.

[4] Arantxa Casanova, Michal Drozdzal, and Adriana Romero-Soriano. Generating unseen complex scenes: are we there yet? *arXiv preprint arXiv:2012.04027*, 2020.

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[6] Ting Chen, Xiaohua Zhai, Marvin Ritter, Mario Lucic, and Neil Houlsby. Self-supervised gans via auxiliary rotation loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12154–12163, 2019.

[7] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016.

[8] Ugur Demir and Gozde Unal. Patch-based image inpainting with generative adversarial networks. *arXiv preprint arXiv:1803.07422*, 2018.

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[10] Raghudeep Gadde, Qianli Feng, and Aleix M Martinez. Detail me more: Improving gan's photo-realism of complex scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13950–13959, 2021.

[11] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[12] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

[13] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In *Advances in Neural Information Processing Systems*, 2017.

[14] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020.

[15] 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.

[16] 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.

[17] Liang Hou, Huawei Shen, Qi Cao, and Xueqi Cheng. Self-supervised GANs with label augmentation. In *Advances in Neural Information Processing Systems*, 2021.

[18] Tianyu Hua, Hongdong Zheng, Yalong Bai, Wei Zhang, Xiao-Ping Zhang, and Tao Mei. Exploiting relationship for complex-scene image generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1584–1592, 2021.

[19] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 172–189, 2018.

[20] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[21] Jongheon Jeong and Jinwoo Shin. Training gans with stronger augmentations via contrastive discriminator. In *International Conference on Learning Representations*, 2021.

[22] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018.

[23] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. In *Proc. NeurIPS*, 2020.

[24] 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.

[25] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8110–8119, 2020.

[26] Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. *Advances in Neural Information Processing Systems*, 32:3927–3936, 2019.

[27] Jae Hyun Lim and Jong Chul Ye. Geometric gan. *arXiv preprint arXiv:1705.02894*, 2017.

[28] 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.

[29] Bingchen Liu, Yizhe Zhu, Kunpeng Song, and Ahmed Elgammal. Towards faster and stabilized {gan} training for high-fidelity few-shot image synthesis. In *International Conference on Learning Representations*, 2021.

[30] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2794–2802, 2017.

[31] Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. Which training methods for gans do actually converge? In *International conference on machine learning*, pages 3481–3490. PMLR, 2018.

[32] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations*, 2018.

[33] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2337–2346, 2019.

[34] Pedro O Pinheiro, Amjad Almahairi, Ryan Y Benmalek, Florian Golemo, and Aaron C Courville. Unsupervised learning of dense visual representations. In *NeurIPS*, 2020.

[35] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[36] Byungseok Roh, Wuhyun Shin, Ildoo Kim, and Sungwoong Kim. Spatially consistent representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1144–1153, 2021.

[37] Axel Sauer, Kashyap Chitta, Jens Müller, and Andreas Geiger. Projected gans converge faster. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[38] Edgar Schonfeld, Bernt Schiele, and Anna Khoreva. A u-net based discriminator for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8207–8216, 2020.

[39] Edgar Schönfeld, Vadim Sushko, Dan Zhang, Juergen Gall, Bernt Schiele, and Anna Khoreva. You only need adversarial supervision for semantic image synthesis. In *International Conference on Learning Representations*, 2021.

[40] Tristan Sylvain, Pengchuan Zhang, Yoshua Bengio, R Devon Hjelm, and Shikhar Sharma. Object-centric image generation from layouts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2647–2655, 2021.

[41] Ngoc-Trung Tran, Viet-Hung Tran, Bao-Ngoc Nguyen, Linxiao Yang, and Ngai-Man Man Cheung. Self-supervised gan: Analysis and improvement with multi-class minimax game. *Advances in Neural Information Processing Systems*, 32:13253–13264, 2019.

[42] Chien-Yao Wang, I-Hau Yeh, and Hong-Yuan Mark Liao. You only learn one representation: Unified network for multiple tasks. *arXiv preprint arXiv:2105.04206*, 2021.

[43] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8798–8807, 2018.

[44] Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. Dense contrastive learning for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3024–3033, 2021.

[45] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3733–3742, 2018.

[46] Tete Xiao, Colorado J Reed, Xiaolong Wang, Kurt Keutzer, and Trevor Darrell. Region similarity representation learning. *arXiv preprint arXiv:2103.12902*, 2021.

[47] Enze Xie, Jian Ding, Wenhai Wang, Xiaohang Zhan, Hang Xu, Peize Sun, Zhenguo Li, and Ping Luo. Detco: Unsupervised contrastive learning for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8392–8401, 2021.

[48] Zhenda Xie, Yutong Lin, Zheng Zhang, Yue Cao, Stephen Lin, and Han Hu. Propagate yourself: Exploring pixel-level consistency for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16684–16693, 2021.

[49] Ceyuan Yang, Yujun Shen, Yinghao Xu, and Bolei Zhou. Data-efficient instance generation from instance discrimination. *Advances in Neural Information Processing Systems*, 34:9378–9390, 2021.

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

[51] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Free-form image inpainting with gated convolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4471–4480, 2019.

[52] Ning Yu, Guilin Liu, Aysegul Dundar, Andrew Tao, Bryan Catanzaro, Larry S Davis, and Mario Fritz. Dual contrastive loss and attention for gans. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6731–6742, 2021.

[53] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In *International conference on machine learning*, pages 7354–7363. PMLR, 2019.

[54] Han Zhang, Zizhao Zhang, Augustus Odena, and Honglak Lee. Consistency regularization for generative adversarial networks. In *International Conference on Learning Representations*, 2020.

[55] Shengyu Zhao, Zhijian Liu, Ji Lin, Jun-Yan Zhu, and Song Han. Differentiable augmentation for data-efficient gan

training. *Advances in Neural Information Processing Systems*, 33:7559–7570, 2020.

[56] Zhengli Zhao, Sameer Singh, Honglak Lee, Zizhao Zhang, Augustus Odena, and Han Zhang. Improved consistency regularization for gans. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 11033–11041, 2021.

[57] Zhengli Zhao, Zizhao Zhang, Ting Chen, Sameer Singh, and Han Zhang. Image augmentations for gan training. *arXiv preprint arXiv:2006.02595*, 2020.

[58] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.