Nested Scale-Editing for Conditional Image Synthesis

Lingzhi Zhang∗ Jiancong Wang∗ Yinshuang Xu Jie Min
Tarmily Wen James C. Gee Jianbo Shi

University of Pennsylvania

Abstract

We propose an image synthesis approach that provides stratified navigation in the latent code space. With a tiny amount of partial or very low-resolution image, our approach can consistently out-perform state-of-the-art counterparts in terms of generating the closest sampled image to the ground truth. We achieve this through scale-independent editing while expanding scale-specific diversity. Scale-independence is achieved with a nested scale disentanglement loss. Scale-specific diversity is created by incorporating a progressive diversification constraint. We introduce semantic persistency across the scales by sharing common latent codes. Together they provide better control of the image synthesis process. We evaluate the effectiveness of our proposed approach through various tasks, including image outpainting, image superresolution, and cross-domain image translation.

1. Introduction

Imagine that we want to identify a person based on the appearance of their eyes and nose, or a lower resolution image, as shown in figure 1. One solution may be to outpaint the entire face, conditioned on the partial information available.

∗equal contribution.

Figure 1: Our approach enables scale-specific visual editings in conditional image synthesis. We can choose to surgically manipulate coarse-level structural information or fine-level details in Cat2Dog translation and image outpainting tasks.

We want to be as imaginative and as detailed as possible to give us a greater chance of success in finding the right person. These tasks are multimodal in nature, i.e., a single input corresponds to many plausible outputs.

Conditional image synthesis approaches aim to solve this problem by sampling stochastic latent codes to generate images in a GAN setting. However, these image synthesis methods of sampling operate as uncontrollable “black boxes”. During inference, we can only hope that a sampled random variable generates the ideal image we desire; otherwise, we need to keep sampling.

We propose a steerable conditional image synthesis approach. Inspired by the steerable filtering in the wavelet process [48], we wish to ‘steer’ the image synthesis across the spatial scales consistently. While in steerable filtering we are concerned with angular edge orientation, in our domain, we focus on object semantics. Specifically, we aim to create visual information from a coarse-level structure to fine-level texture. The key objectives are 1) scale-independence: we learn disentangled representations that model scale-specific visual details, and 2) diversity/mode covering: we ensure that the decoder covers diverse variations presented on ground truth images.

To implement the scale-independent objective, we take inspiration from the Laplacian image pyramid decomposi-
tion: our algorithm essentially learns to generate progressively more refined image along spatial scales, with each level of refinement independent of each other. To implement the diversity objective, we extend a successful diversity constraint [36] to multi-scale and ensure scale-specific diversity.

Unlike current multi-scale noise injection methods [7, 25], our multi-scale injected noises share same latent variable during training. This introduces semantic persistency, meaning the decoder expects latent variables on different scales to have similar semantic meaning. Semantic-persistence can play a major role in search efficiency, since it enables stratified navigation in the latent code space. Fig. 2 illustrates the stratified navigation process for face superresolution. We first coarsely sample a widespread set of latent random variables to find an image roughly matching the ground truth. Because of our scale-independent representation, we can efficiently edit the image by adjusting any of the latent variable at a specific scale and edit the image information at the corresponding scale. Therefore we can generate a refined image by adjusting the existing latent variable at next scale and repeat, until final scale is reached. This is the ideal steering behavior we seek.

In summary, we highlight our contributions as follows:

- We are the first to propose a multi-scale feature disentanglement loss and a progressive diversification regularization to achieve scale-specific control for conditional image synthesis.
- To the best of our knowledge, our work is the first to utilize diverse conditional image generation for identity recovery. We developed three evaluation metrics for identity recovery in diverse conditional synthesis scenarios.
- We evaluate our aforementioned development on tasks of image outpainting, image superresolution, and multimodal image translation. Our methods achieves competitive image quality and diversity compared to state-of-the-art counterparts, while consistently outperforming them in terms of identity recovery.

2. Related Work

2.1. Multimodal Conditional Generation

Deep generative models have been widely used in many conditional image synthesis tasks, such as super-resolution [9, 28, 54, 60–62], inpainting missing regions [21, 35, 43, 44, 56, 58, 59, 65], style transfer [15, 18, 22, 34, 39], image blending [40, 49, 55, 66], and text-to-image [23, 32, 46]. The majority of these tasks are in nature multimodal, where single input condition may correspond to multiple plausible outputs. BicycleGAN [68] first proposed to model this one-to-many distribution by explicitly encoding the target domain into a compact Gaussian latent distribution from which the generator samples. During inference, the generator maps a random variable drawn from the latent distribution, combined with the given input, to the output. StackedGAN and its variant StackedGAN++ [63, 64] proposed to use a hierarchical generator which incorporates conditional code on multiple scales and were able to generate high quality synthetic images. DRIT [31] and MUNIT [20] proposed to disentangle the features into domain-invariant content codes and domain-specific style codes for unpaired image-to-image translations. During inference, the sampled style codes combined with the content code can be transformed into many plausible outputs.

Although the above approaches can generate multimodal outputs given a conditional input, there is no explicit constraint to prevent the generator from mapping the various sampled random variables to similar outputs, which is known as mode collapse. Two concurrent works aim to al-
leviate this issue by proposing diversity regularization techniques for generative training. Mode Seeking GAN (MSGAN) [41] proposes to maximize the ratio of two sampled images over the corresponding latent variables. Normalized Diversification (NDiv) [36] proposes to enforce the generator to preserve the normalized pairwise distance between the sparse samples from a latent distribution to the corresponding high-dimensional output space.

In addition to image synthesis, other applications with multimodal predictions include but are not limited to predicting uncertain motion flows in the future [14, 51, 57], hallucinating diverse body pose affordance in 2D [52] and 3D [33] scenes.

Different from previous work, our focus is to achieve scale-specific control for image synthesis and progressively inject stochasticity into different scales of the synthesized image. Therefore, our proposed techniques can be readily added into these orthogonal previous approaches. We demonstrate that our techniques work with face outpainting, face superresolution, and multimodal animal translation with modified MUNIT [20].

### 2.2. Feature Disentanglement

Recently there is increasing interest in disentanglement of distinct image characteristics for image synthesis. [26, 29, 53] attempt to decouple image style and content, while [12, 13, 38] target object shape and appearance. These approaches explicitly incorporate two codes that denote the two characteristics, respectively, into the generative model and introduce a guided loss or incorporate the invariance constraint to orthogonalize the two codes, while our method disentangles the scale-specific variations from global structural feature to local texture feature through hierarchical input of latent variables into the generative model.

Besides the disentanglement of two specific characteristics, several prior efforts [4, 5, 8, 10, 17] have explored partially or fully interpretable representations of independent factors for generative modeling. Some current work learn representations of the specific attribute by supervised learning [2, 45] with a conditioning discriminator. Our method focuses on multi-scale disentanglement through unsupervised learning, which is distinct from the concept of disentangling specific semantic factors.

[24] has first proposed synthesize images from low-resolution scale to higher-resolution scale in a progressive manner. Similarly, other closely related work is [7] and [25]. [7] pioneered the use of the hierarchical generator with latent codes injected at each level for multi-scale control of image synthesis and further advanced in [25]. Both [7] and [25], however, only target unconditional image synthesis and do not explicitly enforce diversity of outputs. Our work extends to conditional synthesis and incorporates explicit diversity constraints.

Since the purpose and definition of disentanglement in our method are different from previous work, the existing metrics for evaluating disentanglement [4, 11, 17, 25, 27] are not appropriate for measuring feature disentanglement for our method. We therefore develop a new means to quantify the hierarchical disentanglement for our approach.

### 3. Methods

Multimodal conditional image synthesis combines a given input conditional code with sampled latent codes drawn from a compact latent space (usually a standard normal or uniform distribution) and decodes the combination into an output image. Unlike previous efforts [20, 36, 41, 68], we propose a cascading disentangled decoder inspired by Laplacian image pyramid [1]. With a central multi-scale backbone, it generates output images at every feature spatial scale through a single convolutional layer. We enforce the generated images at every scale to be average-pooled back into a lower-resolution generated image. By doing so, we distill features at each spatial scale to only focus on image details — similar to a Laplacian image — at the corresponding spatial resolution. With these scale-independent features, we can inject random variables into each scale of the image features to model the scale-specific stochasticity of image details.

### 3.1. Multiscale Disentanglement

To enable scale-specific editing of visual contents, we need a model that has a disentangled latent code representation. This implies that visual content on specific scales can be modified by changing corresponding latent codes, while visual contents on other scales remain unaffected.

In a decoder network that receives only a single latent code at the coarsest scale, the single latent code controls image generation at all scales and hence changing the code will affect all scales. This motivates the multi-latent code design: injecting latent codes on all spatial scales and allowing each individual latent code to control the image on corresponding scales. Let $Z_0$ denote the random base latent code, and $Z_1 = A_1(Z_0), Z_2 = A_2(Z_0), \ldots, Z_k = A_k(Z_0)$, and the latent variable at each scale level $i \in \{1, \ldots, k\}$, where $A_i$ is an affine mapping matrix. Intuitively, the latent code at coarse scale may mostly affect global structure while latent code at fine scale is more likely to alter local texture and respective scale. Such behavior can be seen in [7] and [25].

However, this design does not guarantee that visual information represented by the latent code at different scales are disentangled. For example, latent code at coarse scale might control texture and color that are also affected by latent code at fine scale. With such a decoder, it is still difficult to edit the scale-specific visual information while keeping information on other scales unchanged.
Figure 3: Our decoder network takes a conditional code and a single random variable as inputs. A single latent variable is injected into multi-scale feature representations through the first several shared layers of MLPs. Then that output is injected into different affine transformation layers with Adaptive Instance Normalization (AdaIN) [19]. At every spatial scale, a single convolutional layer is used to decode a real image at each corresponding resolution. Our proposed disentanglement loss enforces the generated images at every scale to be averaged-pooled back into a lower-resolution generated image. At each iteration, we sample four latent variables and generate four images at each spatial scale, where we also enforce the pairwise distance between sampled images and latent variables to encourage diverse synthesized outputs.

Therefore, we propose a simple but effective approach to disentangle features at each layer of the decoder to only control the visual information at the corresponding spatial scale. At each layer in the decoder, we add a single convolution layer to synthesize an image at the corresponding spatial resolution. Then, we enforce that each synthesized image, when downsampled, matches the synthesized image at the previous spatial scale. We call this constraint a multi-scale disentanglement loss $L_{disent}$. Specifically, we use average pooling to downsample the synthesized image and pixel-wise Euclidean distance to constraint the down-sampling consistency. Our intuition is that, by doing so, the features at each layer are not allowed to change any visual information at the previous or deeper layers. In this way, we distill each level of features to only edit the visual information at its corresponding spatial scale.

Formally, we denote $S$ as the downsampling operation on image $x$, specifically, average $2 \times 2$ pooling with stride of 2. The loss function of progressive downsampling consistency is defined as follows,

$$L_{disent} = \sum_{i=1}^{n-1} d(S(G_{i+1}(c, z)), G_i(c, z)), \quad (1)$$

where $n$ is the number of resolution scales, $c$ is the conditional code, $z$ is random variable, and $G_i$ is the generator, whose subscript refers to the network layers that are responsible for synthesizing images at each scale. For $G_i(2 \leq i \leq n)$, they have iterative format:

$$G_i(c, z) = U[G_{i-1}(c, z), A_i(z)] \quad (2)$$

and,

$$G_1(c, z) = U[c, A_1(z)]. \quad (3)$$

where $U$ denotes the Upsampling Module.

At each spatial scale, we also applied conditional GAN to synthesize photo-realistic images, where the loss functions are as follows,

$$L_{GAN} = E_{x \sim p_{data}} \left[ \sum_{i=1}^{n} \log(D_i(S^{n-i}(x)|c)) \right] + E_{z \sim p(z)} \left[ \sum_{i=1}^{n} \log(1 - D_i(G_i(z, c))) \right]. \quad (4)$$

The similar multi-scale adversarial loss has been applied in SinGAN [47] as well.

3.2. Progressive Diversification

To avoid mode collapse and increase diversity of the synthesized images, we leverage the normalized diversification [36], which forces the normalized pairwise distance of generative outputs to be at least as large as that of the corresponding latent variables.
Table 1: Quantitative Comparison with state-of-the-art approaches in multimodal image outpainting task.

<table>
<thead>
<tr>
<th>Method</th>
<th>Quality ↓</th>
<th>Diversity ↑</th>
<th>Shortest Distance ↓</th>
<th>Recovery Count(%) ↑</th>
<th>Landmark Alignment ↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>BicycleGAN</td>
<td>64.133</td>
<td>0.093</td>
<td>0.233</td>
<td>15.34</td>
<td>5.914</td>
</tr>
<tr>
<td>MSGAN</td>
<td>56.998</td>
<td>0.232</td>
<td>0.237</td>
<td>28.93</td>
<td>4.754</td>
</tr>
<tr>
<td>Ndiv</td>
<td>68.855</td>
<td>0.319</td>
<td>0.256</td>
<td>20.95</td>
<td>5.126</td>
</tr>
<tr>
<td>Ours</td>
<td>55.854</td>
<td>0.333</td>
<td>0.228</td>
<td>34.78</td>
<td>4.540</td>
</tr>
</tbody>
</table>

Here we introduce normalized diversification in a progressive manner, that is, we add normalized diversification loss at each layer of the hierarchical decoder. In comparison, previous work [36,41,42] applied the diversity penalty only at the final scale output of the model, which enforces in a brute force way the final output diversity but does not prevent individual levels of the model from mode collapse (for a 3-layer model where latent code $z$ is injected at each level, previous efforts only enforce that the final output varies by $z$, but allow individual layers to collapse. For example, the first layer of the model utilizes the $z$ while second and third layer ignore $z$ entirely).

Mode collapses at individual levels prevent us from exactly controlling the diversity on a specific level of structure or texture for synthesized data. Thus, we propose progressive diversification, effectively unfolding manifold topology for different scales. In this way, we achieve not only independent multi-scale control during the generative process but also guarantee latent that the code $z$ introduces variation on every scale, likely from structure to texture.

The inserted progressive normalized diversification can be formulated as loss function 5.

$$\mathcal{L}_{N_{div}} = \sum_{k=1}^{N} \frac{1}{N^2 - N} \sum_{i=1}^{N} \sum_{j=1}^{N} \max(\alpha D_{ij}^z - D_{ij}^{G_k(z,c)}),$$

(5)

where $N$ is the number of samples to calculate the normalized pairwise distance matrix and $D_{ij}^z, D_{ij}^{G_k(z,c)}$ are defined as elements in normalized pairwise distance matrix $D^z, D^{G_k(z,c)} \in \mathbb{R}^{N \times N}$ of $x_i \sim p(z)$:

$$D_{ij}^z = \frac{d(z_i - z_j)}{\sum_j d(z_i - z_j)},$$

$$D_{ij}^{G_k(z,c)} = \frac{d(G_k(z,c) - G_k(z,c)_j)}{\sum_j d(G_k(z,c) - G_k(z,c)_j)}.$$  

(6)

Here, $d$ the latent variable is the Euclidean distance, and for generative outputs is the pixel-wise Euclidean distance.

4. Experiments

The proposed approaches were evaluated through their performance on various tasks, including image outpainting, image superresolution, and dog2cat/cat2dog translation. In
addition to conventional quality and diversity assessment, we propose to evaluate the extent to which diverse sampling can improve identity recovery, especially in the context of facial recognition. We believe that this is a first attempt that aims to apply diverse synthesis for better recognition. Our premise is that given a conditional code containing only partial information of the ground truth image, a decoder capable of generating diverse output can produce at least one or more results that is close to or recovers the ground truth image, as long as sufficient latent code is sampled. This property would be useful in many difficult recognition situations, such as identifying criminals in largely occluded or very low-resolution images. Diverse sampling would provide a set of candidates, which can then be narrowed down further by human reviewers. We describe next three newly proposed evaluation metrics in this work for identity recovery.

4.1. Evaluation Metrics

To perform evaluation of our approach, we use the following metrics.

**FID.** We use FID [16] to evaluate the quality of generated data. This metric applies the Inception Network [50] to extract features from real and synthesized data, and then calculates the Frechet distance between the two distributions of collected real and synthesized features, respectively. A lower FID score indicates less discrepancy between real and synthesized data and hence higher quality.

**LPIPS.** We apply LPIPS [67] to quantify the diversity, which calculates the pairwise average feature distance across the whole generated dataset. We use AlexNet [30] pretrained on ImageNet [6] to extract features. Larger values of the pairwise LPIPS score indicate increased image diversity.

**Identity Recovery.** To quantify how well diverse sampling recovers the true facial identity, we propose to evaluate the distance between the most similar sampled output and the ground truth image. First, we compute the shortest embedding distance between a set of sampled outputs and the ground truth, where the embedding distance is given by LPIPS. We average this shortest embedding distance across all training examples, which we denote as **Shortest Distance** in Table (1) and (2). We also count the chance that each method generates the most similar outputs under evaluation of this embedding distance, and denote this as **Recovery Count.** We also evaluate the identity recovery of

<table>
<thead>
<tr>
<th>Method</th>
<th>Quality ↓</th>
<th>Diversity ↑</th>
<th>Shortest Distance ↓</th>
<th>Recovery Count(%) ↑</th>
<th>Landmark Alignment ↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>BicycleGAN</td>
<td>40.837</td>
<td>0.019</td>
<td>0.129</td>
<td>26.49</td>
<td>7.062</td>
</tr>
<tr>
<td>MSGAN</td>
<td>49.647</td>
<td>0.0438</td>
<td>0.131</td>
<td>29.48</td>
<td>5.907</td>
</tr>
<tr>
<td>Ndiv</td>
<td>81.421</td>
<td><strong>0.0758</strong></td>
<td>0.143</td>
<td>5.60</td>
<td>5.938</td>
</tr>
<tr>
<td>Ours</td>
<td>46.346</td>
<td>0.0677</td>
<td><strong>0.125</strong></td>
<td><strong>38.43</strong></td>
<td><strong>4.261</strong></td>
</tr>
</tbody>
</table>

Table 2: Quantitative Comparison with state-of-the-art approaches in multimodal image superresolution task.
face images using facial landmarks, which are obtained using a pretrained facial landmark detector [3]. Specifically, the similarity between a sampled output and the ground truth is given by the mean squared error distance between the corresponding 68 facial landmark locations for the sampled and ground-truth images. We refer to this evaluation as Landmark Alignment.

4.2. Image Outpainting

We first present experimental results in an image outpainting task, where the goal is to fill in large areas of missing pixels in a highly occluded image, which may have many different solutions for a given input. In terms of model implementations, we only need to add our proposed multiscale disentanglement and diversification into a standard conditional encoder-decoder. The experiment is conducted over the CelebA dataset [37] with cropped 128x128 images.

We compare our model with the current state-of-the-art multimodal conditional synthesis approaches, including BicycleGAN [68], MSGAN [41], and NDIV [36]. The experimental results show that our model can generate the best results in terms of image quality, diversity as well as identity, as shown in Table 1. In our qualitative comparison figure 4, we show that one of the sampled image could best recover the ground truth facial identity. We think that this is because our model can sample not only very diverse but also realistic images.

4.3. Image Super-Resolution

Another multimodal conditional synthesis we run on face data is super-resolution. While most of other super-resolution approaches model this task as a deterministic image-to-image process, we consider super-resolution from a very low resolution image as a one-to-many process because of its uncertainty in nature. For example, the 16x16 low resolution in Figure 5 can be very difficult to identified. Even for human, it is very difficult to tell what is the ground true high-resolution image is. Or, every human would probably have a different answer. In this task, we use bilinearly downsampled 16x16 image as low-resolution input and 128x128 image as high-resolution output in CelebA dataset [37].

Our implementation of superresolution model is similar to image outpainting, but we start to decode image at scale of 16x16, which is the same as input resolution. The disentanglement and diversification is added at every other higher resolution scale. As seen from Table (2), even though our approach does not reach the best quality or diversity, but it achieves the best identity score. We think that the simply measuring quality or diversity separately is not sufficient. BicycleGAN has the best quality but lack diversity, so it is difficult to “hit” the ground truth image. NDIV has the largest diversity but lacks quality, and thus it is also very difficult to recover the realistic ground truth image. In contrast, our model can produce reasonably large diversity and good quality, and thus has the highest chance to recover the ground truth image. Our implementation of super-resolution model is similar to image outpainting, but we start to decode image at scale of 16x16, which is the same as input resolution. The disentanglement and diversification is added at every other higher resolution scale. As seen from Table. (2), even though our approach does not reach the best quality or diversity, but it achieves the best identity score.

Table 3: Quantitative comparison with state-of-the-art approaches on the cross-modal image-to-image translation task.

<table>
<thead>
<tr>
<th>Method</th>
<th>Quality ↓</th>
<th>Diversity ↑</th>
<th>Shortest Distance ↓</th>
<th>Recovery Count(%) ↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>MUNIT [20]</td>
<td>15.74</td>
<td>0.533</td>
<td>0.445</td>
<td>15.0</td>
</tr>
<tr>
<td>Ours</td>
<td>21.22</td>
<td>0.547</td>
<td>0.393</td>
<td>85.0</td>
</tr>
</tbody>
</table>
ground truth image. NDiv [36] has the largest diversity but lacks quality, and thus it is also very difficult to recover the realistic ground truth image. In contrast, our model can produce reasonably large diversity and good quality, and thus has the highest chance to recover the ground truth image.

4.4. Dog to Cat Image Translation

We conduct an unpaired cross-modal image translation using the MUNIT backbone [20] on the cat and dog dataset from [31]. This dataset contains 1264 training/100 testing dog and 771 training/100 testing cat images. We modify the MUNIT backbone for multi latent codes injection and implemented the disentanglement loss and progressive diversification loss as mentioned in section 3.1 and 3.2. Detailed network architecture before and after modification can be found in the supplemental material. We compare FID and LPIPS of our model against the original MUNIT model trained on the cat and dog dataset and report the performance in Table 3. For identity recovery evaluation, we pass the ground truth image through the content encoder from the MUNIT framework and derive a content code, which is recombined with sampled latent codes and decoded into images. Quantitatively, we evaluate identity recovery using the LPIPS metric from 4.1. Qualitative results of identity recovery is demonstrated in Fig. 6.

4.5. Multi-scale Disentanglement Evaluation

To quantify the multiscale disentanglement, we evaluate perceptual variations on output images against different noise varying scales. In specific, for scale $k$ in our $n$ scale decoder, a given fixed input condition code $c$, we sample a center latent code $z$. We fixed the input latent codes on every scale to be $z$ except for for scale $k$, where we sample 10 latent codes centered around $z$ and generate 10 different output images. Pairwise perceptual distances (using LPIPS) among these 10 images are calculated. We calculate the pairwise perceptual distance for scale $k$ across 1000 different input condition codes and averaged them. We plotted the averaged perceptual distance against the scale on Fig. 7. For a general decoder (left) without the the disentanglement constraint, the perceptual variation were not monotonically decreasing along the scale. Our scale disentangled decoder (right) on the other hand achieves multiscale disentanglement: latent codes at finer scales monotonically introduce less variation to the image, therefore editing latent code at finer level has little effect on visual information on coarse scale.

5. Conclusion

We develop a conditional image synthesis network that enables scale-specific and diverse control of image content. We instantiate our design with a cascading decoder network. We couple it with multi-scale feature disentanglement constraints and a progressive diversification regularization. In addition, we gain semantic persistency in the decoder by sharing latent code across scales during training. This allows for stratified navigation and search within latent code space, and motivate the task of identity recovery. We propose three evaluation metrics for identity recovery within conditional image synthesis scenarios. On tasks of image outpainting, image superresolution, and multimodel image translation, our method consistently outperforms state-of-the-art counterparts in terms of identity recovery, while having competitive image quality and diversity. Hence we believe our method may potentially be useful in extreme image recognition situations, such as recognizing criminals in largely occluded or very low-resolution images, and finding lost pets from low quality surveillance images.

Acknowledgement

We gratefully appreciate the support from Honda Research Institute Curious Minded Machine Program. We also gratefully acknowledge a GPU donation from NVIDIA.
References


