1 Network Architectures

**TS-VQGAN.** The architecture of TS-VQGAN used in pose guiding (Penn [11] and synthesis Deep-Fashion [7]) is follow the VQGAN [4] with its ImageNet pretrained weights. The TS-VQGAN for face datasets (FFHQ [6]) is simplified as in Tab. 1. The channels of codebook are 256. For the pose guiding and face editing, we set the total codebook number as 16384 and 2048 respectively.

**Transformer.** We use the transformer with a ‘base’ scale [3] of channels 768, attention head numbers 12, transformer layers 12, which is trained with dropout rate 0.1.

2 Supplementary Experiment Details

Table 1: TS-VQGAN for face datasets, where Two-stream (TS) convolutions are only used in the encoder.

<table>
<thead>
<tr>
<th>Encoder ResidualBlock</th>
<th>Decoder ResidualBlock</th>
<th>DownBlock</th>
<th>UpBlock</th>
</tr>
</thead>
<tbody>
<tr>
<td>TS-Conv2d (256 × 256 × 64)</td>
<td>Conv2d (16 × 16 × 512)</td>
<td>ResidualBlock×4 (16 × 16 × 512)</td>
<td>InstanceNorm+SWISH (16 × 16 × 512)</td>
</tr>
<tr>
<td>ResidualBlock×2+DownBlock (128 × 128 × 64)</td>
<td>ResidualBlock×2+UpBlock (32 × 32 × 384)</td>
<td>ResidualBlock×2+UpBlock (128 × 128 × 128)</td>
<td>InstanceNorm+SWISH (256 × 256 × 64)</td>
</tr>
<tr>
<td>ResidualBlock×2+DownBlock (64 × 64 × 128)</td>
<td>ResidualBlock×2+UpBlock (64 × 64 × 256)</td>
<td>ResidualBlock×2+UpBlock (256 × 256 × 64)</td>
<td>Conv2d+Tanh (256 × 256 × 3)</td>
</tr>
<tr>
<td>ResidualBlock×2+DownBlock (32 × 32 × 256)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ResidualBlock×2+DownBlock (16 × 16 × 384)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ResidualBlock×4 (16 × 16 × 512)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>InstanceNorm+SWISH (16 × 16 × 512)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TS-Conv2d (16 × 16 × 256)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

2.1 Masks for the Penn Action Dataset

Since the Penn Action (PA) [11] dataset includes various scenes videos, many handheld objects influence the generated results, e.g., baseball bats, golf clubs, and Barbells, etc. We get the related masks with various dilated rates according to the image scenes as shown in Fig. [1].
2.2 Synthesis for the DeepFashion and Places2

The synthesis DeepFashion \cite{7} (SDF) is composed of DeepFashion and Places2 \cite{12} as foregrounds and backgrounds respectively. We show the combined results in Fig. 2.
Table 2: Human evaluation in face editing.

<table>
<thead>
<tr>
<th>Method</th>
<th>Taming</th>
<th>Taming*</th>
<th>SC-FEGAN</th>
<th>iLAT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average Scores</td>
<td>0.33</td>
<td>3.33</td>
<td>17.33</td>
<td>38.33</td>
</tr>
</tbody>
</table>

3 More Experiment Results

3.1 Human Evaluation

For more comprehensive comparisons, 68 images selected from FFHQ and CelebA are compared by 3 uncorrelated volunteers. Particularly, volunteers need to choose the best one from randomly shuffled 4 compared methods, which include Taming [4], Taming* (Taming with LA mask), SC-FEGAN [5], and our iLAT. Outputs of all these methods are combined with the references. Then, they give one score to the best approach. If all methods work roughly the same for a certain sample, it will be ignored for the final decision. The average scores are shown in Tab. 2.

3.2 Combining Outputs and References

Although our proposed iLAT can maintain the global semantics in most cases, it still suffers from few unexpected changes in some details, e.g., facial features (Fig. 3 rows 1 and 3 (c)), words in backgrounds (Fig. 3 row 2 (c)), and hats (Fig. 3 row 4 (c)) as shown in Fig. 3. This problem is similar to the one in most GAN inverse based methods [1, 9]. To this end, these methods combine their outputs with original images, and achieve the final results. We also show the combination of our method in face dataset in Fig. 3. We simply resize the input masks to 128×128, and resize them back to 256×256 for the smoothness. Note that combinations of Taming and references have obvious boundaries caused by the inconsistent semantics (Fig. 3(d)), while our iLAT outputs can be seamlessly combined with the references, and achieve much better performance.

3.3 More Qualitative Results

We show more related qualitative experiment results in face dataset (Fig. 4), PA (Fig. 5), SDF (Fig. 6), and sequential generations of pose guiding (Fig. 7).

3.4 Alternative Version of the Attention Mask

As discussed in the main paper, the proposed local autoregressive (LA) attention mask $M_{LA}$ enjoys both global semantic information and local AR generation.

In contrast, we also consider another possible version for attention mask as illustrated in Fig. 8. For the tokens that are not needed to predict any other tokens, such as $\{t_1, t_2, t_3, t_5, t_7, t_8\}$, their respective fields can be further shifted to right until the masked tokens, as the red points shown in Fig. 8. However, technically it is very hard to implement this version processed in parallel; and thus it is impossible to efficiently utilize this version in our iLAT. We will make it as the future work of extending this alternative of attention mask.

References


Figure 3: Comparisons of face dataset, where images in (d) are combinations of Taming and references, and images in (e) are combinations of iLAT and references. Please zoom-in for details.


Figure 4: More results from CelebA-HQ and FFHQ. From left to right, references, targets with sketches, Taming [4], Taming with LA mask (Taming*), SC-FEGAN [5], our iLAT, iLAT combined with references.


Figure 5: More results from PA. From left to right, references, targets, PATN [13], PN-GAN [8], PoseWarp [2], MR-Net [10], Taming [4], our iLAT.

Figure 6: More qualitative results in SDF compared between Taming and iLAT.
Figure 7: Sequential generation results in PA compared between Taming and iLAT.

Figure 8: The illustration of the further extended local autoregressive (LA) attention mask $\tilde{M}_{LA}$, which is hard to achieve parallelly without significant improvements.