NÜWA-LIP: Language-guided Image Inpainting with Defect-free VQGAN

Minheng Ni\textsuperscript{1}  Xiaoming Li\textsuperscript{1}  Wangmeng Zuo\textsuperscript{1,2}
\textsuperscript{1}Harbin Institute of Technology  \textsuperscript{2}Peng Cheng Laboratory
mhni@stu.hit.edu.cn  csxmli@gmail.com  wmzuo@hit.edu.cn

This supplemental material mainly contains:

• Discussion with Partial Convolution in Section I
• Details of DF-VQGAN and MP-S2S in Section II
• Details of proposed datasets in Section III
• User Study in Section IV
• More comparisons with other models in Section V
• More comparisons of DF-VQGAN in Section VI
• More comparisons with VQGAN in Section VII
• More inpainting results in Section VIII
• Analyses of failure case in Section IX
• Broader impact and limitations in Section X

I. Difference with Partial Convolution

We note that Liu et al. \cite{6} propose a partial convolutional layer (PConv) where the convolution operation is masked and renormalized to be conditioned on only non-defective pixels for image inpainting task. It is defined as:

\[
P_{\text{Conv}}(x) = \begin{cases} W^T (x \odot m) \text{softmax}(m) + b, & \text{if } \text{sum}(m) > 0 \\ 0, & \text{otherwise} \end{cases}, \tag{1}
\]

where \(x\) is the defective image and \(m\) is the mask matrix.

In contrast, we simplify the formulation of our defect-free operation in DF-VQGAN by removing the symmetrical connection, and the defect-free operations on convolution, normalization, and attention layers are defined as:

\[
\begin{align*}
\text{Conv}'(y) &= \text{Conv}^{DF}(x) \odot (1 - m) + \text{Conv}(y) \odot m, \\
\text{Norm}'(y) &= \text{Norm}^{DF}(x) \odot (1 - m) + \text{Norm}(y) \odot m, \\
\text{Attn}'(y) &= \text{Attn}^{DF}(x) \odot (1 - m) + \text{Attn}(y) \odot m,
\end{align*}
\]

where \(y\) represents the ground-truth image and \(x = y \odot m\). \(\text{Conv}^{DF}\), \(\text{Norm}^{DF}\), and \(\text{Attn}^{DF}\) are the defect-free operations, which are defined as:

\[
\begin{align*}
\text{Conv}^{DF}(x) &= W_e^T (x \odot m) + b, \\
\text{Conv}(y) &= W_e^T (y) + b, \\
\text{Norm}^{DF}(x) &= \frac{x - \frac{1}{N} \sum_{i=1}^{N} \text{Var}[x_i]}{\sqrt{\frac{1}{N-1} \sum_{i=1}^{N} \text{Var}[x_i]}} + \epsilon, \\
\text{Norm}(y) &= \frac{y - \sum_{i=1}^{N} \text{Var}[y_i]}{\sqrt{\sum_{i=1}^{N} \text{Var}[y_i]}} + \epsilon, \\
\text{Attn}^{DF}(x) &= \text{Softmax}(x^T W_a x) \odot m \odot x, \\
\text{Attn}(y) &= \text{Softmax}(y^T W_a y) \odot y,
\end{align*}
\]

where \(N\) and \(N_m\) are the numbers of all pixels and defective pixels in \(x\), respectively. \(x'\) is the revised \(x\) with the defective region fulfilling with \(\frac{1}{N} \sum_{i=1}^{N} \text{E}[x_i]\). \text{E}[] and \text{Var}[] denote expectation and variance, respectively. \(W_e (W_a)\) is shared parameters between \(\text{Conv}^{DF}\) and Conv (\(\text{Attn}^{DF}\) and Attn).

Comparing Eqn. (1) with Eqns. (2-3), our DF-VQGAN has two differences with PConv: (1) DF-VQGAN is a VAE model, which is trained by reconstructing a full image \(\hat{y}\) from a full image \(y\). For adopting VAE in the inpainting task, we need to introduce mask \(m\) carefully without destroying the schema of VAE. However, PConv is not a VAE model, and it takes the defective image \(x\) as input to predict a inpainted result \(\hat{y}\). (2) Instead of performing on the convolution layer only, our DF-VQGAN focuses on three operations, which may easily lead to receptive spreading. This allows our DF-VQGAN effectively learn the valid features from defective input and reconstruct results with high fidelity. To validate the effectiveness of DF-VQGAN, we add the PConv.
We use Conceptual Captions \[10\] as the pre-training corpus. We randomly select 10,000 samples from the MaskCOCO and MaskFlickr. As for MaskVG, we randomly select 10,000 samples from the VG dataset. For each image-text pair, the original image and corresponding caption are considered ground-truth images and text descriptions. Each image will be cropped and resized to the resolution of 256 × 256. The defective image is generated by masking with either one bounding box of the object or a random irregular region. The details of the proposed three datasets are listed in Tab. A. We will release them under a Creative Commons Attribution 4.0 License.

### III. Details of Proposed Datasets

We follow \[5\] and select the test sets of MSCOCO and Flickr to build our MaskCOCO and MaskFlickr. As for MaskVG, we randomly select 10,000 samples from the VG dataset. For each image-text pair, the original image and corresponding caption are considered ground-truth images and text descriptions. Each image will be cropped and resized to the resolution of 256 × 256. The defective image is generated by masking with either one bounding box of the object or a random irregular region. The details of the proposed three datasets are listed in Tab. A. We will release them under a Creative Commons Attribution 4.0 License.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Image-Text Pairs</th>
<th>Mask Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>MaskCOCO</td>
<td>5000</td>
<td>31.5%</td>
</tr>
<tr>
<td>MaskFlickr</td>
<td>1000</td>
<td>48.3%</td>
</tr>
<tr>
<td>MaskVG</td>
<td>10000</td>
<td>14.6%</td>
</tr>
</tbody>
</table>

### IV. User Study

To further evaluate the quality of our NUWA-LIP and baselines, we conduct a user study from real human perception. We randomly select 500 samples from the MaskCOCO dataset and compare with NUWA-LIP, NUWA, and GLIDE on two aspects, i.e., visual quality and semantic consistency. The visual quality focuses on evaluating the structures whether they are photo-realistic or contain distorted details. The semantic consistency assesses whether the inpainted results have semantically consistent content with the language guidance. Volunteers with a computer vision background are required to give a choice about which one has better quality. From Fig. B, we can observe that compared with these competing methods (i.e., NUWA and GLIDE), our NUWA-LIP has obvious better performance in both visual quality and semantic consistency, which indicates that our NUWA-LIP can generate more photo-realistic and consistent results.

### V. Comparisons with Other Models

To explore the effectiveness of NUWA-LIP, we conduct additional experiments with other inpainting models. Specifically, we compare NUWA-LIP with TDANET \[13\], a popular non-pre-trained language-guided inpainting model, and MASKGIT \[2\] and LAMA \[11\], which are class-conditional and unconditional inpainting models pre-trained on large-scale data, respectively. For MASKGIT, we use CLIP to classify the class of the ground-truth image as the input. As shown in Tab. B, NUWA-LIP outperforms all these models, showing its effectiveness and the essentials of the language.

**Discussion with Stable Diffusion**

STABLE DIFFUSION is an effective model for visual synthesis tasks. Fig. C shows the difference between STABLE DIFFUSION and most prior image inpainting works \[2, 11, 13\]. In general image inpainting settings, the input image for the inpainting model is defective or damaged. However, the input image of STABLE DIFFUSION should be normal images without defective...
Figure C. Comparison of STABLE DIFFUSION and NÜWA-LIP pipelines. Different from most prior works, the input image of STABLE DIFFUSION needs to be well-formed, which is called image editing in most related works.

Table B. Comparison with different models on MaskCOCO. †denotes trained or finetuned on COCO.

<table>
<thead>
<tr>
<th>MODEL</th>
<th>FID†</th>
<th>CLIP Score†</th>
</tr>
</thead>
<tbody>
<tr>
<td>STABLE DIFFUSION [9] (IMAGE EDITING)</td>
<td>10.9</td>
<td>30.38</td>
</tr>
<tr>
<td>TDANET† [13]</td>
<td>27.2</td>
<td>27.90</td>
</tr>
<tr>
<td>MASKIT [2]</td>
<td>15.5</td>
<td>27.20</td>
</tr>
<tr>
<td>NÜWA-LIP (w/o PRETRAIN)†</td>
<td>11.0</td>
<td>28.74</td>
</tr>
<tr>
<td>NÜWA-LIP</td>
<td>12.0</td>
<td>29.34</td>
</tr>
<tr>
<td>NÜWA-LIP (FINETUNE)†</td>
<td>10.5</td>
<td>29.65</td>
</tr>
</tbody>
</table>

Table C. More quantitative comparisons of DF-VQGAN on ImageNet. DF-VQGAN outperforms VQGAN or VQGAN-P on both image reconstruction (IMG.REC) and oracle inpainting (ORC.INP).

<table>
<thead>
<tr>
<th>MODEL</th>
<th>RESOLUTION</th>
<th>TOKEN LENGTH</th>
<th>VOCAB SIZE</th>
<th>IMG.REC</th>
<th>ORC.INP</th>
</tr>
</thead>
<tbody>
<tr>
<td>VQGAN</td>
<td>256 → 16</td>
<td>256</td>
<td>12288</td>
<td>6.03</td>
<td>7.15</td>
</tr>
<tr>
<td>VQGAN-P</td>
<td>256 → 16</td>
<td>256</td>
<td>12288</td>
<td>6.03</td>
<td>3.77</td>
</tr>
<tr>
<td>PARTIAL CONV.</td>
<td>256 → 16</td>
<td>256</td>
<td>12288</td>
<td>6.83</td>
<td>7.14</td>
</tr>
<tr>
<td>TS-VQGAN</td>
<td>256 → 16</td>
<td>256</td>
<td>12288</td>
<td>5.89</td>
<td>6.47</td>
</tr>
<tr>
<td>DF-VQGAN/NS</td>
<td>256 → 16</td>
<td>256</td>
<td>12288</td>
<td>5.14</td>
<td>5.44</td>
</tr>
<tr>
<td>DF-VQGAN</td>
<td>256 → 16</td>
<td>256</td>
<td>12288</td>
<td>5.16</td>
<td>2.95</td>
</tr>
</tbody>
</table>

VI. More Comparisons of DF-VQGAN

To validate whether relative estimation can avoid receptive spreading of defective regions, we compare VQGAN with DF-VQGAN/s, which is DF-VQGAN without symmetrical connection. In the upper part of Tab. C, we can find that we significantly reduce the FID score from 7.15

---

1 https://github.com/huggingface/diffusers

2 We use the best SD-V1-4 checkpoint.
to 5.44 in the oracle inpainting task. Besides, the gain in image reconstruction can be ascribed to the usage of relative estimation in improving the robustness of the model.

We further validate whether symmetrical connection can protect the information of non-defective regions. We compare DF-VQGAN with VQGAN-P, which directly copies and pastes the non-defective region of the image into the generated results. In the bottom part of Tab. C, we can find that we achieve a better FID score (i.e., 2.95 v.s. 3.77) in oracle inpainting task, which indicates that our symmetrical connection can make a significantly better transition between the non-defective region and inpainted part. Meanwhile, we obtain a comparable FID score of 5.16 in the image reconstruction task. The comparison with DF-VQGAN/s shows the benefits of combining relative estimation and symmetrical connection in image inpainting task.

Finally, we conduct the comparison with TS-VQGAN [1], which is used in conditional image synthesis to encode an image without defective regions. The goal of TS-VQGAN is to avoid information leaking, which means results are more similar to reference images rather than the target condition. Different from TS-VQGAN, DF-VQGAN works in the image inpainting scenario, in which defective regions and non-defective regions exist at the same time in an image. From Tab. C, we can observe that our approach still outperforms TS-VQGAN with a large margin in oracle inpainting.

VII. More Comparisons with VQGAN

We provide more visual results in Fig. D to compare our DF-VQGAN with vanilla VQGAN and analyze their performance on both defective and non-defective regions. We can find that our DF-VQGAN can well capture the semantic details and generate consistent structures in defective regions. More importantly, our DF-VQGAN can well keep the non-defective content unchanged.

VIII. More Inpainting Results

We provide more inpainting results to show the effectiveness of NUWA-LIP in Fig. F. We can observe that NUWA-LIP can leverage the text guidance well and generate results with higher fidelity and better consistency.

IX. Failure Case

Although NUWA-LIP shows effectiveness in most cases, we find that it may fail in some cases like Fig. E, which shows fine-tuning will cause failure in inpainting some rare objects. In most cases, fine-tuning brings impressive improvement in the quality of the inpainted images but may fail in some objects which occurs very little in the fine-tuning dataset. We will balance the distribution of each object and augment these with fewer samples.

X. Broader Impact and Limitations

NUWA-LIP, which is an effective model for language-guided image inpainting, can provide the potential for users to edit and manipulate an image, which may lead to destructive behaviors, e.g., fake images may be abused in some cases like news reporting. We will explore a more trustworthy model to prevent such abuse cases. Besides, as a common issue of autoregressive models, handling an extremely large image would have a much higher computational cost, and may be easy for users to retrain this model.
<table>
<thead>
<tr>
<th>Defective Image</th>
<th>GLIDE</th>
<th>NÜWA</th>
<th>NÜWA-LIP</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1" alt="Image" /></td>
<td><img src="image2" alt="Image" /></td>
<td><img src="image3" alt="Image" /></td>
<td><img src="image4" alt="Image" /></td>
</tr>
</tbody>
</table>

- *A kitchen with a bright window and house plants.*
- *A herd of cattle sitting in front of a church with a steeple.*
- *A person with an orange blanket covering them, sleeping on a wooden park bench.*
- *A yellow and blue train is next to an overhang.*
- *Man on a contraption, surrounded by a bicycle.*
- *A cat eating a bird it has caught.*
- *A fork rests on a plate next to a piece of cake.*
- *A man in the black suit.*
- *A man in the black sports wear.*
- *A man in the blue suit.*
- *A light house near the sea.*
- *A light house in the sea.*
- *A light house on the grass land.*
- *A man in the black suit.*
- *A man in the black sports wear.*
- *A man in the blue suit.*
- *A woman sitting on the grass.*
- *A kid looking for something.*
- *A little bonfire.*
- *A stack of books.*
- *A white bird.*
- *A toy car.*
- *A stone block.*

Figure F. More inpainting results. NÜWA-LIP can effectively complete the defective image under the guidance of different texts.
References


