CSF-Net: Context-Semantic Fusion Network for Large Mask Inpainting

Chae-Yeon Heo and Yeong-Jun Cho*
Department of Artificial Intelligence Convergence
Chonnam National University, South Korea
{cyheo001, yj.cho}@jnu.ac.kr

1. Detailed Ablation Study Implementation

This appendix provides detailed implementation descriptions for the ablation study settings presented in the main paper. All experiments were conducted using the LaMa model as the baseline under Center Box (80%) masking conditions on the `Places365` dataset.

1.1. Single Encoder Architecture Variant

The baseline CSF-Net architecture processes the masked image $\mathbf{I}_{\text{mask}}$ and semantic candidates $\mathcal{H}_{\text{sel}} = \{\mathbf{H}_i\}_{i=1}^{P}$ through separate encoders. This enables disentangled learning of contextual and semantic information. In contrast, the single encoder variant processes $\mathbf{I}_{\text{mask}}$ and each candidate $\mathbf{H}_i$ simultaneously within a single shared encoder. The input is formed by concatenating $\mathbf{I}_{\text{mask}}$ and each individual candidate $\mathbf{H}_i$ along the channel dimension for separate processing: $\mathbf{I}_{\text{in},i} = \text{Concat}(\mathbf{I}_{\text{mask}}, \mathbf{H}_i)$, resulting in input tensors of shape $\mathbf{I}_{\text{in},i} \in \mathbb{R}^{B\times 6\times H\times W}$ (3 channels from masked RGB image + 3 channels from candidate RGB image). Each candidate-context pair is processed independently through the same shared encoder with identical weights.

In the dual encoder structure, features extracted from $\mathbf{I}_{\text{mask}}$ are used as Queries, while those from each candidate $\mathbf{H}_i$ are used as Keys and Values in the cross-attention mechanism (see Sec 4.2 of the main paper). In contrast, the single encoder structure has only one input, so all attention components (Q, K, V) are derived from the same feature representation using self-attention.

$$Q = K = V = \text{Linear}(\mathbf{F}_{\text{shared}}). \tag{1}$$

where $\mathbf{F}_{\text{shared}}$ is the shared feature extracted from the single encoder and $\text{Linear}(\cdot)$ denotes learnable linear transformations. This architecture does not clearly separate contextual and semantic information, making it difficult to learn meaningful differences between candidates. As shown in Tab 6 of the main paper, this structural limitation leads to significantly worse performance, with FID 21.08 and LPIPS 0.430, compared to the dual encoder structure.

1.2. P=1 Direct Paste without Network

This ablation study uses only the single candidate $\mathbf{H}_{\text{top}}$ with the highest consistency score and performs reconstruction without any network components, including the Context and Semantic Encoders (E_c, E_s), fusion decoder, or hierarchical pixel selection module. The visible regions in $\mathbf{I}_{\text{mask}}$ are preserved, while the masked regions are directly replaced with pixels from $\mathbf{H}_{\text{top}}$:

$$\mathbf{I}_{\text{guide}}(x,y) = \begin{cases} \mathbf{I}_{\text{mask}}(x,y) & \text{if } (x,y) \in \text{visible region} \\ \mathbf{H}_{\text{top}}(x,y) & \text{if } (x,y) \in \text{masked region.} \end{cases} \tag{2}$$

This approach skips all learning-based processes. The implementation simply copies pixels from a single candidate image into the masked regions, without considering contextual cues or candidate diversity. As a result, the reconstruction may misalign with object structures and produce visually inconsistent artifacts in high-frequency regions. As shown in Tab 4, this method yields FID 18.67 and LPIPS 0.411, indicating degraded performance compared to the proposed P=3 structure.

1.3. No Pixel Selection: Direct Top-P Paste

In this setting, we bypass the learning-based pixel-wise source selection and confidence weighting performed by the Structure Score Network (SSN) and Perceptual Score Network (PSN). Instead, the semantic guidance image($\mathbf{I}_{\text{guide}}$) is constructed by sequentially pasting valid pixels from the top P candidate images ranked by the combined consistency score $S_{\text{valid}} = S_{\text{MSE}} + S_{\text{LPIPS}}$. This approach omits the hierarchical scoring mechanism described in Sec 4.3 of the main paper.

Instead, all learning-based fusion components, including feature extraction, attention-based integration, and the hierarchical consistency refinement with learnable coefficients $\beta^{(l)}$, are skipped. The implementation relies on simple conditional logic and iterative pixel copying, without spatial reasoning that accounts for candidate diversity or spatial

*Corresponding author

coherence. As a result, it may produce visually inconsistent artifacts near object boundaries or in high-frequency regions. According to Tab 5 in the main paper, this method yields FID 18.54 and LPIPS 0.421, indicating degraded performance compared to the proposed CSF-Net.

1.4. Candidate Selection

To complement the quantitative evaluation of candidate selection presented in Tab 3, we also provide a visual comparison in Fig 1. It qualitatively illustrates how the combined consistency score $S_{\text{valid}} = S_{\text{MSE}} + S_{\text{LPIPS}}$ yields more reliable completions by jointly capturing structural and perceptual consistency, validating the effectiveness of the proposed ranking strategy.

1.5. Experimental Configuration Summary

The ablation experiments were designed to independently assess: (1) the effectiveness of the dual encoder architecture (E_c and E_s) for separating contextual and semantic features, (2) the necessity of the Context-Semantic Fusion Transformer for multi-candidate integration, (3) the impact of learning-based hierarchical pixel selection using SSN and PSN, and (4) the effectiveness of the candidate selection strategy based on the combined consistency score $S_{\text{valid}} = S_{\text{MSE}} + S_{\text{LPIPS}}$. Results show that the full CSF-Net, with all components enabled, achieves the best performance across FID, LPIPS, and C@m metrics, confirming that each component contributes to performance gains in large-mask inpainting.

2. Algorithm of CSF-Net

To clarify the full inference pipeline of CSF-Net, we provide pseudocode describing how semantic guidance image ($\mathbf{I}_{\text{guide}}$) is generated from a masked input image ($\mathbf{I}_{\text{mask}}$) and a set of structure-aware semantic candidates. The procedure includes candidate generation, context-semantic feature fusion, and hierarchical pixel selection, as detailed in Alg 1.

Algorithm 1 An Algorithm of the Proposed CSF-Net

Input: Masked image $\mathbf{I}_{\text{mask}}$
Output: Semantic guidance image $\mathbf{I}_{\text{guide}}$

Stage 1: Structure-Aware Candidate Generation
$\mathcal{A}_{\text{init}} \leftarrow \text{AC}(\mathbf{I}_{\text{mask}})$
// AC: pretrained Amodal Completion model
Compute consistency scores S_{valid} for each $\mathbf{A}_i \in \mathcal{A}_{\text{init}}$
Select top-P candidates $\mathcal{H}_{\text{sel}} = \{\mathbf{H}_1, ..., \mathbf{H}_P\}$

Stage 2: Context-Semantic Fusion Transformer
$\mathbf{F}_{\text{ctx}}^{(L)}, ..., \mathbf{F}_{\text{ctx}}^{(1)} \leftarrow E_c(\mathbf{I}_{\text{mask}})$
for each $\mathbf{H}_i$ in $\mathcal{H}_{\text{sel}}$ **do**
  $\mathbf{F}_{\text{sem},i}^{(L)}, ..., \mathbf{F}_{\text{sem},i}^{(1)} \leftarrow E_s(\mathbf{H}_i)$
end for
for $l = L$ to 1 **do**
  if $l = L$ **then**
    $\mathbf{F}_{\text{fuse}}^{(L)} \leftarrow \text{CrossAttn}(\mathbf{F}_{\text{ctx}}^{(L)}, \{\mathbf{F}_{\text{sem},i}^{(L)}\})$
  else
    $\mathbf{F}_{\text{fuse}}^{(l)} \leftarrow \text{CrossAttn}(U(\mathbf{F}_{\text{fuse}}^{(l+1)}) + \mathbf{F}_{\text{ctx}}^{(l)}, \{\mathbf{F}_{\text{sem},i}^{(l)}\})$
  end if
end for

Stage 3: Hierarchical Pixel Selection
for $l = L$ to 1 **do**
  for $i = 1$ to P **do**
    $S_i^{(l)} \leftarrow \text{SSN}(\mathbf{F}_{\text{fuse}}^{(l)}, \mathbf{H}_i^{(l)}, \mathbf{I}_{\text{mask}}^{(l)})$
    $P_i^{(l)} \leftarrow \text{PSN}(\mathbf{F}_{\text{fuse}}^{(l)}, \mathbf{H}_i^{(l)}, \mathbf{I}_{\text{mask}}^{(l)})$
    $C_i^{(l)} \leftarrow \frac{1}{2}(S_i^{(l)} + P_i^{(l)})$
    $\widetilde{C}_i^{(l)} \leftarrow (1 - \beta^{(l)})C_i^{(l)} + \beta^{(l)}U(C_i^{(l+1)})$
  end for
end for
$C_i^{\text{final}} \leftarrow \widetilde{C}_i^{(1)}$ *// final confidence scores at finest level*

for each pixel (x, y) **do**
  $i^*(x, y) \leftarrow \arg\max_{i \in \{1,2,...,P\}} C_i^{\text{final}}(x, y)$
  $\mathbf{I}_{\text{guide}}(x, y) \leftarrow \mathbf{H}_{i^*(x,y)}(x, y)$
end for
return $\mathbf{I}_{\text{guide}}$

3. Object Hallucination Problem

Large-mask inpainting poses significant challenges for existing methods, particularly diffusion-based approaches that often generate semantically inconsistent content. Figure 2 demonstrates examples of object hallucination in diffusion-based inpainting models, highlighting the need for semantic guidance in our proposed approach.

4. Additional Qualitative Results

We provide additional qualitative results to complement the quantitative evaluation in the main paper. Figures 3, 4, and 5 show comprehensive comparisons under Center Box (80%), Center Box (50%), and RandomBrush (50–80%) masking conditions, respectively. By incorporating object-level semantic guidance, CSF-Net consistently produces better visual quality and reduces object hallucination compared to baseline methods.

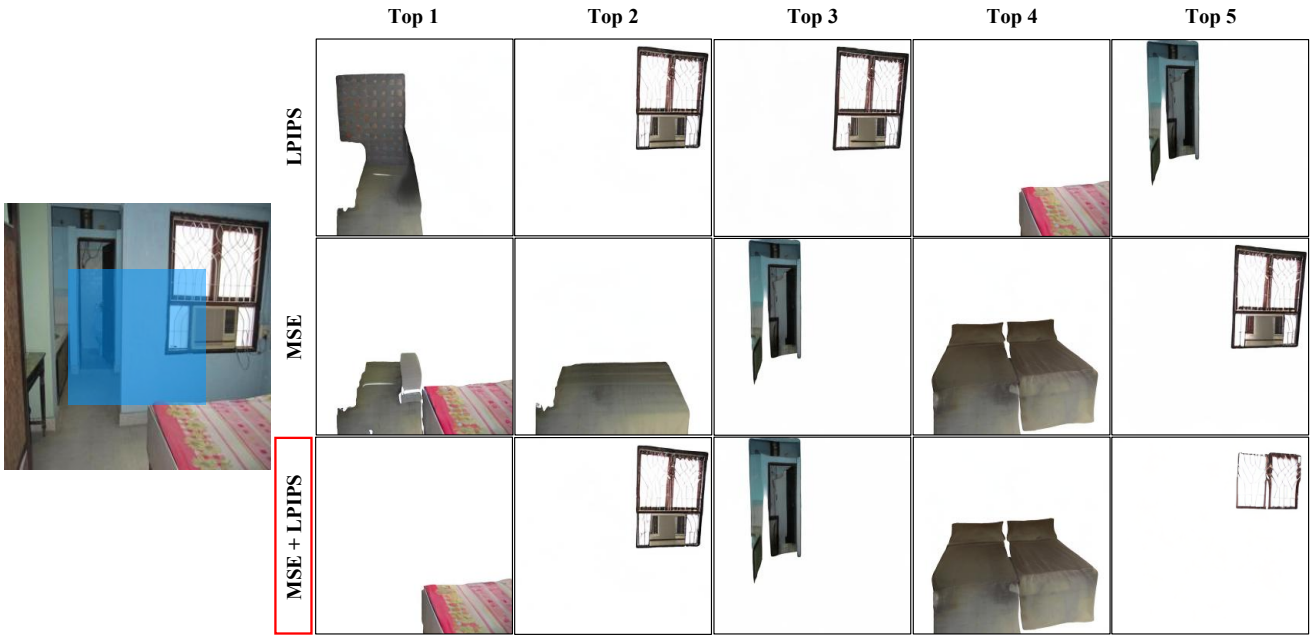

Figure 1. Visual demonstration of candidate selection comparison corresponding to Tab 3 in the main paper. Top row shows LPIPS-only ranking, middle row shows MSE-only ranking, and bottom row (red box) shows the combined $S_{\text{valid}} = S_{\text{MSE}} + S_{\text{LPIPS}}$ ranking. The combined approach yields higher quality candidate selection, visually confirming the quantitative results in Tab 3 that demonstrate the effectiveness of jointly considering structural and perceptual consistency.

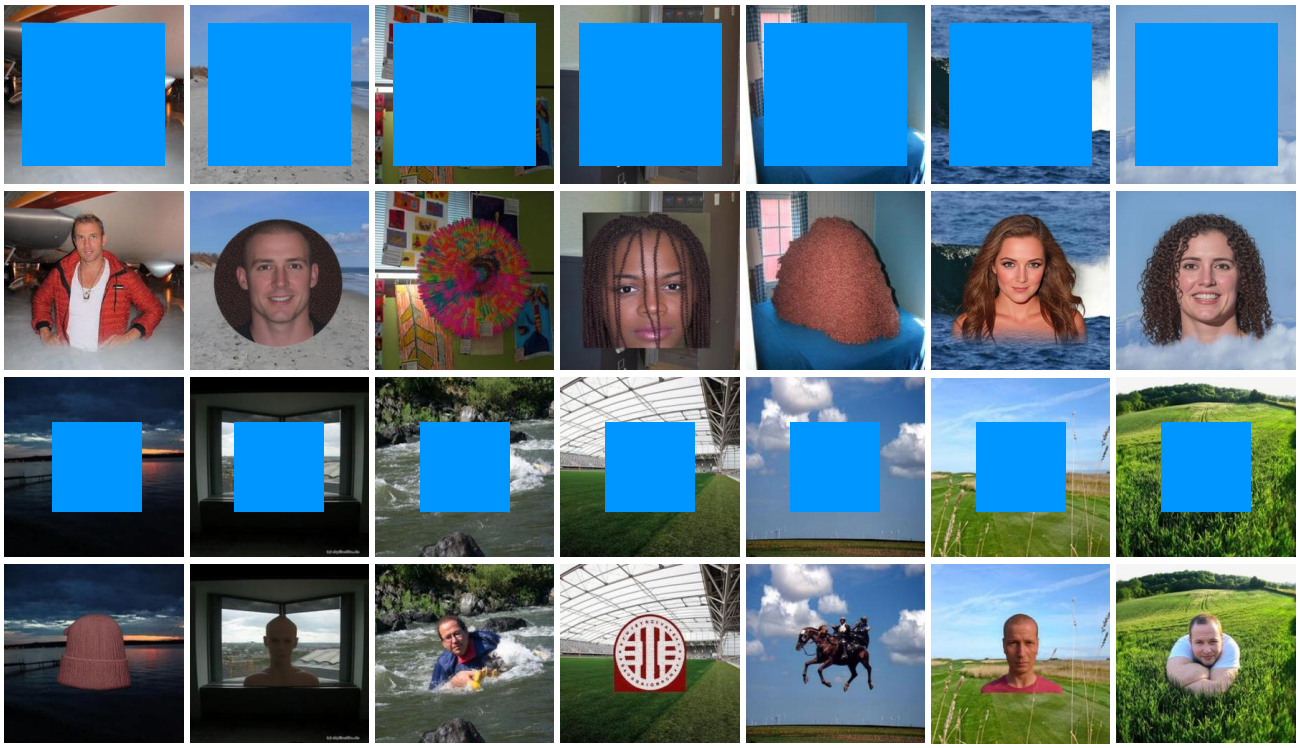

Figure 2. Examples of object hallucination in diffusion-based inpainting models.

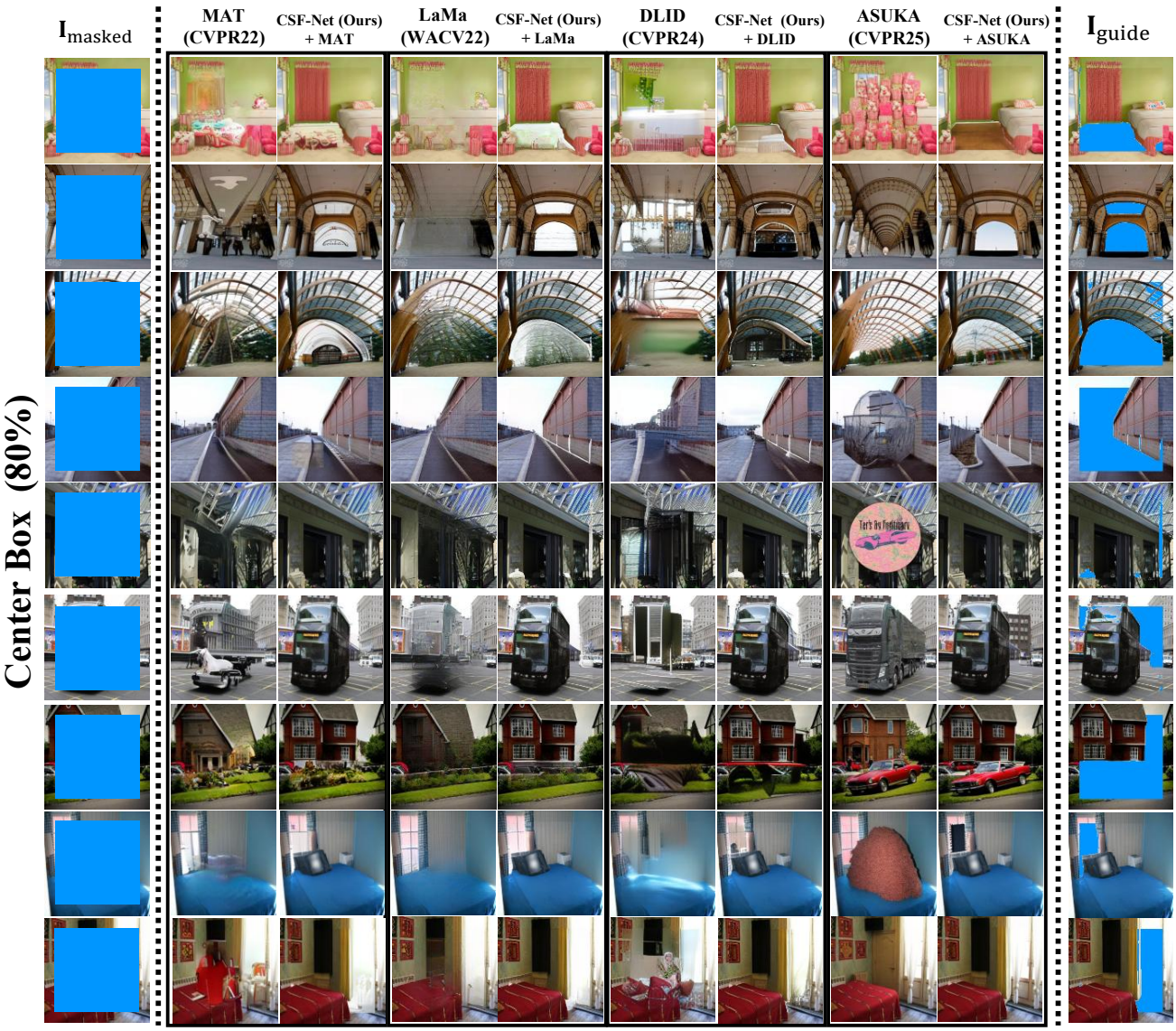

Figure 3. Additional qualitative results on Center Box (80%) masking condition.

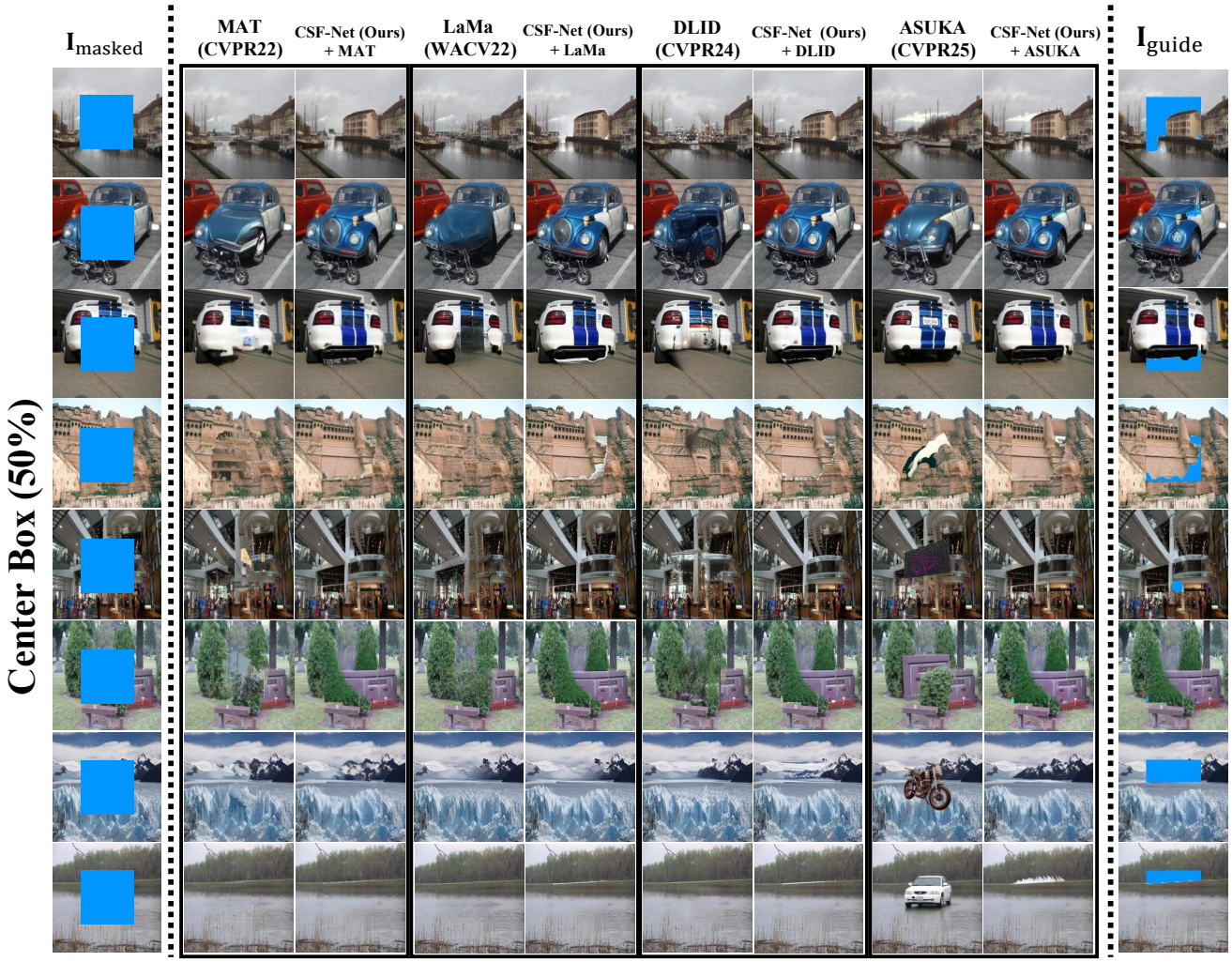

Figure 4. Additional qualitative results on Center Box (50%) masking condition.

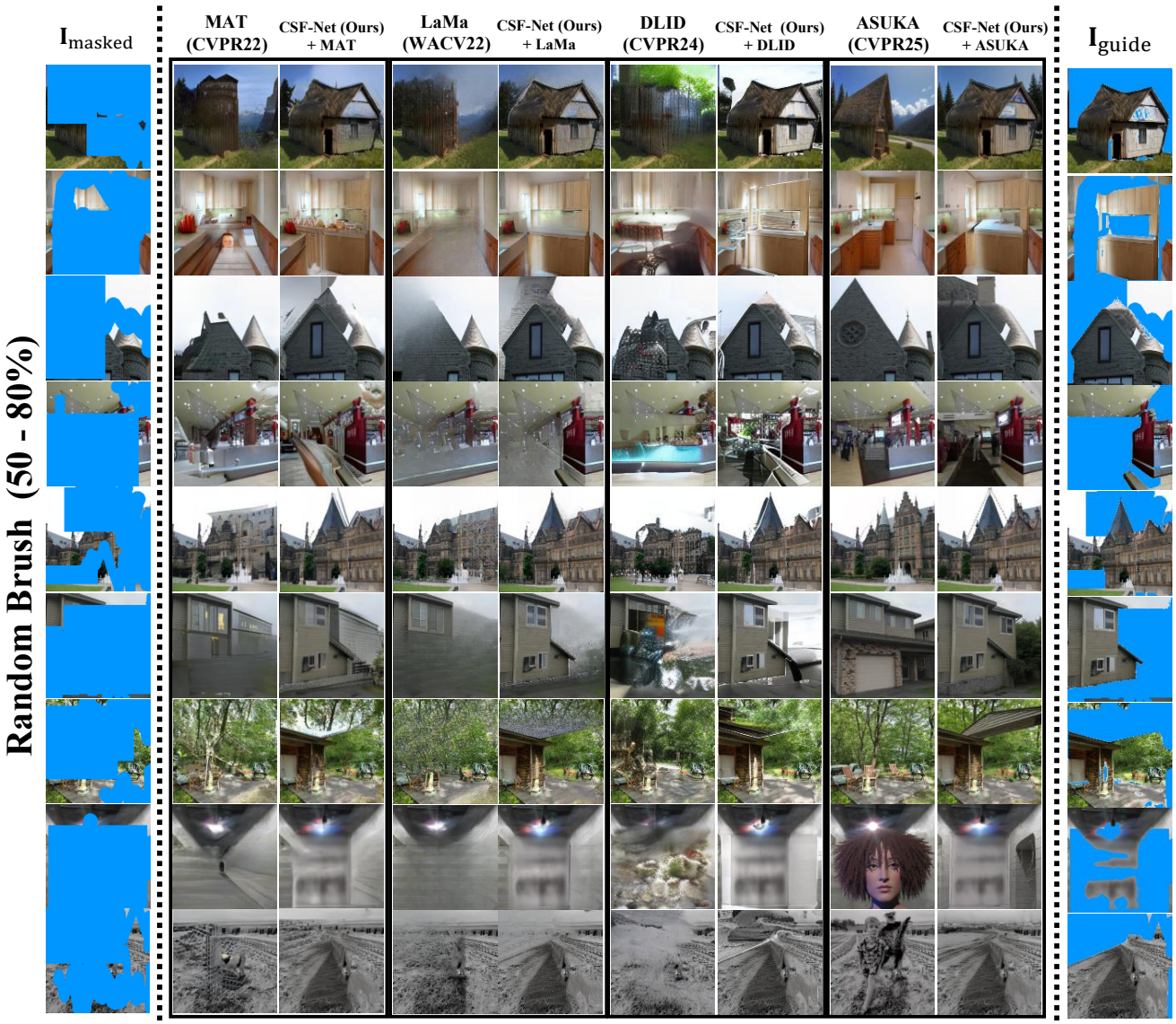

Figure 5. Additional qualitative results on Random Brush(50–80%) masking condition.