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, Gwangju, South Korea
{cyheo001, yj.cho}@jnu.ac.kr

Abstract

In this paper, we propose a semantic-guided framework to address the challenging problem of large-mask image inpainting, where essential visual content is missing and contextual cues are limited. To compensate for the limited context, we leverage a pretrained Amodal Completion (AC) model to generate structure-aware candidates that serve as semantic priors for the missing regions. We introduce Context-Semantic Fusion Network (CSF-Net), a transformer-based fusion framework that fuses these candidates with contextual features to produce a semantic guidance image for image inpainting. This guidance improves inpainting quality by promoting structural accuracy and semantic consistency. CSF-Net can be seamlessly integrated into existing inpainting models without architectural changes and consistently enhances performance across diverse masking conditions. Extensive experiments on the `Places365` *and* `COCOA` *datasets demonstrate that CSF-Net effectively reduces object hallucination while enhancing visual realism and semantic alignment. The code for CSF-Net is available at https://github.com/chaeyeonheo/CSF-Net.git*

1. Introduction

Image inpainting is an important vision task that restores missing regions with semantic consistency to the surrounding context. It plays a crucial role in various applications, including photo editing, object removal, and scene reconstruction. Early approaches relied on low-level cues such as color similarity and texture propagation, but struggled with large or complex missing regions due to limited semantic understanding. Recent advances in generative models, such as Generative Adversarial Networks (GANs) and diffusion models, have significantly improved image inpainting. Among these, GAN-based methods [15, 17] have enhanced inpainting by learning semantic structures and incorporating structural priors, leading to more coherent and context-aware results. Diffusion-based approaches [4, 23] have further advanced inpainting by synthesizing detailed textures and improving inference efficiency.

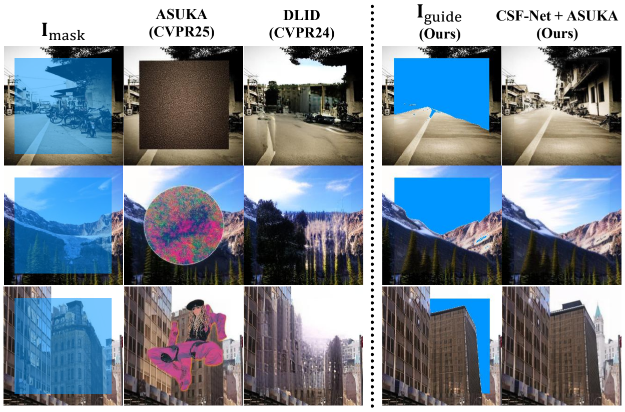

$\mathbf{I}_{\text{mask}}$
ASUKA (CVPR25)
DLID (CVPR24)
$\mathbf{I}_{\text{guide}}$ (Ours)
CSF-Net + ASUKA (Ours)

Figure 1. Large-mask inpainting comparison. Existing methods [3, 23] often produce structural errors or object hallucinations in challenging cases. Incorporating our semantic guidance ($\mathbf{I}_{\text{guide}}$) yields more accurate and semantically coherent inpainting results.

However, despite recent progress, inpainting models still struggle in large-mask scenarios where much of the image is missing and contextual cues are limited. As shown in Fig. 1, even state-of-the-art models often fail to generate semantically accurate content when key structural regions are missing. This results in object hallucination, where the generated content appears visually plausible but is inconsistent with the surrounding context due to insufficient information about the missing areas.

To address this challenge, we incorporate semantic guidance by leveraging a pretrained amodal completion model [16], which generates structure-aware candidates for the missing regions through object-level reasoning. These candidates recover both shape and appearance, providing strong priors that guide semantically coherent inpainting. Since the model learns to reason at the object level, it com-

*Corresponding author

pletes missing regions by extending only the visible parts of each object. This prevents the generation of unrelated content and ensures that the results remain semantically consistent with each object instance.

In this paper, we propose the Context-Semantic Fusion Network (CSF-Net), a transformer-based framework that generates a semantic guidance image to support structure-aware inpainting. CSF-Net consists of three main components: a candidate generation module that selects plausible completions from amodal outputs (Sec. 4.1); a dual encoder and a fusion decoder that extract and integrate contextual and semantic features (Sec. 4.2); and a pixel selection module (Sec. 4.3) that selects the final pixel values to construct the semantic guidance image. As illustrated in Fig.2, the masked image $\mathbf{I}_{\text{mask}}$ and structure-aware candidates are encoded and fused to produce the semantic guidance image $\mathbf{I}_{\text{guide}}$. This guidance provides high-level structural cues, enabling the inpainting model to restore missing regions with greater semantic accuracy and visual consistency.

We evaluate CSF-Net on the `Places365` [26] and `COCOA` [27] datasets under challenging masking scenarios, including Center Box (50%, 80%) and RandomBrush (50–80%). Extensive experiments confirm that CSF-Net consistently improves the performance of diverse inpainting models, including LaMa [22], MAT [12], DLID [3], and ASUKA [23]. Our method achieves consistent gains across multiple metrics (FID, LPIPS, C@m), demonstrating both the generality and effectiveness of the proposed approach. Importantly, CSF-Net can be seamlessly integrated into existing inpainting architectures without requiring any modifications, making it a versatile enhancement module.

Our main contributions are as follows:

- We introduce a semantic guidance strategy that leverages amodal completion to compensate for missing contextual cues in large-mask inpainting.
- We propose CSF-Net, a transformer-based fusion framework that unifies contextual cues and semantic priors to generate the a semantic guidance image for inpainting.
- We show that CSF-Net improves the performance of diverse inpainting baselines under various masking conditions, without requiring any architectural modifications.

To the best of our knowledge, CSF-Net is the first attempt to leverage amodal completion for guiding image inpainting.

2. Related Works

Image Inpainting. Image inpainting aims to restore missing or corrupted regions in images with content that is both visually coherent and semantically meaningful. Traditional approaches such as [1, 5] propagated neighboring information or copied similar patches to fill in the gaps. Structure propagation methods extended edges or contours from surrounding regions to fill missing areas. Patch-based techniques copied similar patches from known regions into the target holes. While effective for small or textured regions, these methods often produced blurry or repetitive artifacts in large, semantically complex areas.

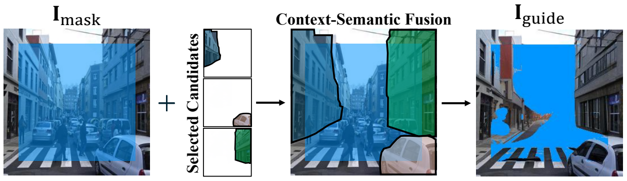


Figure 2. Overview of semantic guidance image ($\mathbf{I}_{\text{guide}}$) generation of CSF-Net. This image incorporates object-level semantic priors and serves as an input to the inpainting model.

Generative Adversarial Networks (GANs) [7] significantly advanced image inpainting by enabling the generation of semantically coherent content in missing regions. Context Encoders [17] first applied GANs to predict missing regions from high-level semantic features. Later methods such as EdgeConnect [15] and StructureFlow [18] incorporated structural priors, including edges and semantic layouts, to guide the inpainting process. These works demonstrated the benefits of guiding inpainting with explicit structure.

More recently, MaskGIT [2] improved inference speed through parallel iterative decoding, offering an efficient alternative to autoregressive models. Diffusion model [19, 21] has recently shown strong performance in image synthesis but struggle with artifacts near mask boundaries in large-mask settings. To address this, a post-processing method, ASUKA [23] applies pretrained priors and decoding to reduce hallucinations and color inconsistencies. LatentPaint [4] further improves upon this by performing inpainting directly in the latent space of pretrained diffusion models. Nonetheless, generative approaches remain limited when the visible context lacks sufficient semantic cues.

Parallel to the development of generative models such as GANs and diffusion, several methods have been proposed to address challenges of large-mask image inpainting. LaMa [22] employs Fast Fourier Convolutions (FFC) to establish global receptive fields early in the network, facilitating the completion of repetitive patterns. Transformer-based methods such as MAT [12] and DLID [3] further improve performance by modeling long-range dependencies and learning latent codes from visible regions. However, these models still rely heavily on visible context, resulting in limited performance when semantic cues are sparse [25].

Amodal Completion. Amodal completion aims to infer the full shape and appearance of partially occluded objects, including their invisible parts. In contrast to inpainting methods that fill arbitrary masks, amodal completion focuses on object-level reasoning to reconstruct complete en-

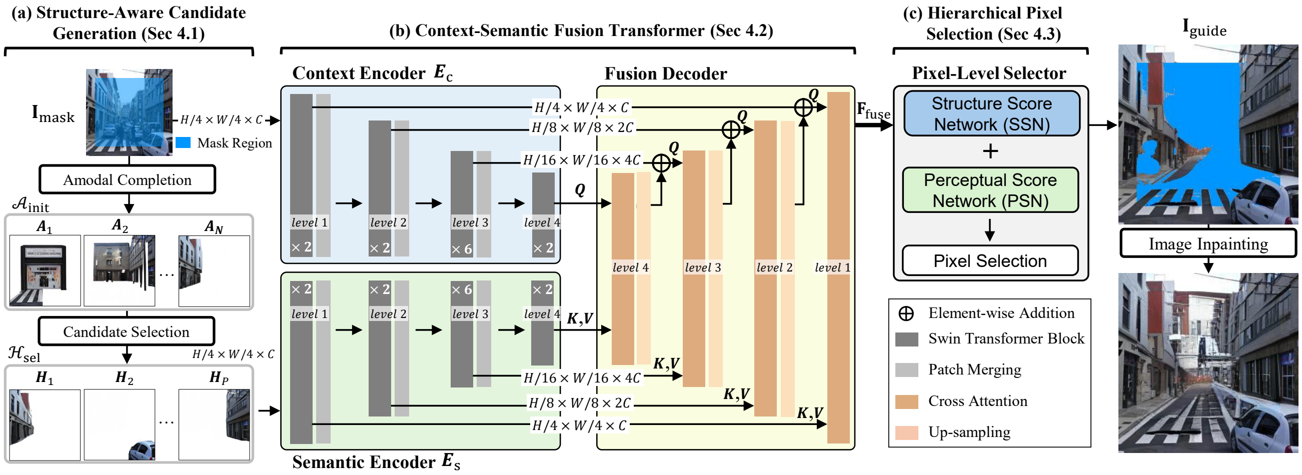

(a) Structure-Aware Candidate Generation (Sec 4.1)
(b) Context-Semantic Fusion Transformer (Sec 4.2)
(c) Hierarchical Pixel Selection (Sec 4.3)
$\mathbf{I}_{\text{guide}}$
$\mathbf{I}_{\text{mask}}$
$H/4 \times W/4 \times C$
Mask Region
Amodal Completion
$\mathcal{A}_{\text{init}}$
A_1
A_2
A_N
Candidate Selection
$\mathcal{H}_{\text{sel}}$
$H/4 \times W/4 \times C$
H_1
H_2
H_P
Context Encoder E_c
Semantic Encoder E_s
Fusion Decoder
level 1
level 2
level 3
level 4
× 2
× 6
$H/4 \times W/4 \times C$
$H/8 \times W/8 \times 2C$
$H/16 \times W/16 \times 4C$
Q
K,V
$\mathbf{F}_{\text{fuse}}$
Pixel-Level Selector
Structure Score Network (SSN)
+
Perceptual Score Network (PSN)
Pixel Selection
Image Inpainting
⊕ Element-wise Addition
Swin Transformer Block
Patch Merging
Cross Attention
Up-sampling

Figure 3. Overview of the proposed CSF-Net. (a) A pretrained amodal completion model generates multiple object completions, and context-inconsistent candidates are filtered out. (b) Dual Swin-Transformer encoders extract multi-scale features from the masked image and selected candidates, which are fused via a cross-attention fusion decoder. (c) Hierarchical pixel selection is performed using structural and perceptual scores to generate the final semantic guidance image $\mathbf{I}_{\text{guide}}$.

tities from partial observations. Early methods usually relied on geometric cues, while deep learning approaches predicted complete binary masks [11], focusing primarily on shape recovery. Built on generative modeling, SeGAN [6] jointly performed segmentation and appearance synthesis. More recently, Pix2Gestalt [16] employs diffusion models to synthesize complete object appearances using large-scale pretrained generative priors.

In this paper, we exploit this capability to generate complete object-level representations, providing high-level semantic guidance for inpainting. This enables our method to overcome the limitations of context-only models, particularly in large-mask scenarios.

3. Motivation and Main Ideas

The large mask image inpainting problem is inherently ill-posed, especially when large regions are missing and lack reliable visual cues. In such cases, recent diffusion-based inpainting methods [14, 23] often hallucinate content that is semantically inconsistent with the surrounding context. This is mainly due to their training objective, which prioritizes generating visually plausible textures over understanding high-level semantics. As a result, their performance significantly degrades in large missing areas, where accurate content prediction requires deeper semantic reasoning.

To address this limitation, we propose the Context-Semantic Fusion Network (CSF-Net), which guides the inpainting process using semantic guidances. Conventional inpainting models often fill missing regions by relying only on the surrounding visible pixels, without reasoning about the underlying object structure. In contrast, our approach leverages a pretrained Amodal Completion (AC) model [16] that learns to infer the complete shape of partially occluded objects from visible cues. This model generates multiple completions that reflect strong object-level semantics. Among these, only contextually consistent candidates are selected and then fused to form a coherent semantic representation. This fused result, denoted as $\mathbf{I}_{\text{guide}}$, serves as a guidance for image inpainting.

4. Proposed Methods

The proposed CSF-Net generates a guidance image $\mathbf{I}_{\text{guide}}$, which can serve as structural guidance for existing image inpainting methods. The overall framework consists of three stages as shown in Fig. 3. First, we generate structure-aware semantic candidates by leveraging a pretrained Amodal Completion (AC) model that infers object shapes from visible regions (Sec. 4.1). Next, we fuse features from the masked image and the semantic candidates through a pair of encoders and a fusion decoder (Sec. 4.2). Finally, we select the optimal candidate for each pixel to compose $\mathbf{I}_{\text{guide}}$ using a multi-scale pixel selection (Sec. 4.3).

4.1. Structure-Aware Candidate Generation

In this section, we generate multiple amodal completions using a pretrained AC model. We then evaluate each completion based on its structural and contextual consistency with the visible region, and select only the reliable ones as semantic candidates as shown in Fig. 3(a). The AC model was originally developed to reconstruct the invisible parts of partially occluded objects in natural scenes. We apply it to the inpainting task by treating the artificial mask as an occluder and reconstructing the occluded structures near its boundary. Given a masked image $\mathbf{I}_{\text{mask}}$, the AC model generates an initial set of N completions, denoted by $\mathcal{A}_{\text{init}} = \{\mathbf{A}_i\}_{i=1}^{N}$, while preserving candidate diversity through its segment-wise generation. At this stage, the visible areas are pre-segmented and each segment is completed individually, ensuring that structurally distinct completion candidates can emerge. Each completion $\mathbf{A}_i$ serves as a structurally plausible hypothesis for the missing region. However, some completions are semantically inconsistent with the visible context, which can degrade the inpainting quality when included in the fusion.

To mitigate this, we evaluate initial completions using a combined consistency score, defined as $S_{\text{valid}} = S_{\text{MSE}} + S_{\text{LPIPS}}$, where S_{MSE} measures pixel-wise similarity and S_{LPIPS} quantifies perceptual similarity based on the LPIPS metric [24][1]. Each $\mathbf{A}_i$ is evaluated by comparing its content against the visible region of the masked image. The consistency score reflects how well each completion aligns with the contextual information contained in the visible region of $\mathbf{I}_{\text{mask}}$. Completions with pixel-wise or perceptual inconsistencies with the visible region receive lower scores. We rank all completions by their S_{valid} scores and select the top P most consistent ones to form a selected set $\mathcal{H}_{\text{sel}} = \{\mathbf{H}_i\}_{i=1}^{P}$, referred to semantic candidates for inpainting. This selection process reduces semantic ambiguity and narrows the candidate space for more useful fusion.

4.2. Context-Semantic Fusion Transformer

The proposed CSF-Net integrates two sources of information: a masked image $\mathbf{I}_{\text{mask}}$ and a set of selected semantic candidates $\mathcal{H}_{\text{sel}}$. It aims to produce a fused representation that jointly captures structural and semantic cues. To process these inputs, CSF-Net employs a dual-encoder architecture based on Swin Transformer blocks [13], followed by a fusion decoder. The overall structure of CSF-Net is designed to follow a U-Net–style [20] architecture with multi-resolution skip connections, enabling the network to capture long-range dependencies while preserving fine-grained spatial details. An overview architecture is shown in Fig. 3(b).

Context and Semantic Encoders. The encoders in CSF-Net independently process the masked input image $\mathbf{I}_{\text{mask}}$ and the semantic candidates $\mathcal{H}_{\text{sel}}$, extracting multi-scale features. The masked image is passed through the context encoder E_c to extract contextual features at different resolution levels:

$$\mathbf{F}_{\text{ctx}}^{(l)} = E_c\left(\mathbf{I}_{\text{mask}}\right)^{(l)}, \tag{1}$$

where $l = \{1, 2, \ldots, L\}$ denotes the resolution level within the encoder (e.g., from fine to coarse). These features encode the global layout and boundary-sensitive structures of the visible region, providing spatial cues. In parallel, each semantic candidate $\mathbf{H}_i$ is passed through the semantic encoder E_s as follows:

$$\mathbf{F}_{\text{sem},i}^{(l)} = E_s\left(\mathbf{H}_i\right)^{(l)}. \tag{2}$$

Both E_c and E_s adopt the same Swin Transformer architecture, composed of stacked transformer blocks and patch merging. These layers gradually downsample the features, enabling the model to capture multi-scale representations essential for reconstructing missing regions.

Fusion Decoder. The fusion decoder performs hierarchical fusion from coarse to fine resolutions, starting at level the final level L. At this level, the context feature serves as the query, while the semantic features from all candidates serve as the keys and values in a cross-attention:

$$\mathbf{F}_{\text{fuse}}^{(L)} = \text{CrossAttn}\left(Q = \mathbf{F}_{\text{ctx}}^{(L)},\ K = V = \mathbf{F}_{\text{sem},i}^{(L)}\right). \tag{3}$$

Here, $\mathbf{F}_{\text{sem},i}^{(L)}$ denotes the semantic features from each candidate, treated as individual tokens in the attention. This initial output of the decoder is a set of fused feature maps that unify contextual and semantic information across multiple scales.

For finer resolution levels $l = \{L-1, L-2, \ldots, 1\}$, the decoder upsamples the fused output from the previous level and adds it element-wise to the corresponding context feature to form the query:

$$\mathbf{F}_{\text{fuse}}^{(l)} = \text{CrossAttn}\left(Q = U(\mathbf{F}_{\text{fuse}}^{(l+1)}) + \mathbf{F}_{\text{ctx}}^{(l)},\ K = V = \mathbf{F}_{\text{sem},i}^{(l)}\right), \tag{4}$$

where $U(\cdot)$ denotes bilinear upsampling. These fused features $\mathbf{F}_{\text{fuse}}^{(l)}$ serve as a strong representation for selecting plausible pixels in the masked regions during the final reconstruction stage.

4.3. Hierarchical Pixel Selection

To generate the semantic guidance image $\mathbf{I}_{\text{guide}}$, we introduce a hierarchical pixel selection that determines the most plausible content for each masked pixel. This module selects pixel values from the candidate set $\mathbf{H}_i$ based on their consistency with the fused feature representation. Each semantic candidate $\mathbf{H}_i$ has the same spatial resolution as the input image but contains valid pixel values only within its completed region, while the rest remains undefined. Because multiple candidates may overlap at the same pixel location, the model must evaluate which candidate provides the most suitable content for each pixel. To address this,

[1] In practice, S_{MSE} and S_{LPIPS} are first computed as pixel-wise error and perceptual similarity (LPIPS), respectively, and then inverted. The two inverted values are summed to obtain the final score. Therefore, a higher S_{valid} indicates greater consistency.

the semantic candidates $\mathbf{H}_i$ are evaluated in a coarse-to-fine manner across multiple resolution levels l.

The masked image $\mathbf{I}_{\text{mask}}$ serves as a contextual reference, providing spatial and structural cues for evaluating semantic candidates. To assess the quality of each candidate, we introduce two complementary scoring networks, as shown in Fig. 4(a). The Structure Score Network (SSN), implemented as a lightweight convolutional network, estimates structural plausibility. In parallel, the Perceptual Score Network (PSN) leverages a pretrained VGG-19 [9] to evaluate perceptual quality and texture realism. The score maps for each network are computed as follows:

$$S_i^{(l)}(x,y) = \text{SSN}\left(\left[\mathbf{F}_{\text{fuse}}^{(l)}(x,y), \mathbf{H}_i^{(l)}(x,y), \mathbf{I}_{\text{mask}}^{(l)}(x,y)\right]\right), \quad (5)$$

$$P_i^{(l)}(x,y) = \text{PSN}\left(\left[\mathbf{F}_{\text{fuse}}^{(l)}(x,y), \mathbf{H}_i^{(l)}(x,y), \mathbf{I}_{\text{mask}}^{(l)}(x,y)\right]\right), \quad (6)$$

where (x, y) denotes pixel coordinates, and i is the index of the candidate. Both the maksed image $\mathbf{I}_{\text{mask}}$ and candidate completions $\mathbf{H}_i$ are downsampled to match the spatial resolution of the l-th level.

The aggregated score map for each candidate at each level is computed as the average of the structural and perceptual score maps:

$$C_i^{(l)}(x,y) = \frac{1}{2}\left(S_i^{(l)}(x,y) + P_i^{(l)}(x,y)\right). \quad (7)$$

To improve spatial consistency across resolution levels, we refine $C_i^{(l)}$ by blending it with the upsampled score from the coarser level $l+1$, weighted by a learnable $\beta^{(l)}$, as follows:

$$\tilde{C}_i^{(l)}(x,y) = (1-\beta^{(l)})C_i^{(l)}(x,y) + \beta^{(l)}U\left(C_i^{(l+1)}(x,y)\right), \quad (8)$$

where $U(\cdot)$ denotes bilinear upsampling. The coefficients $\beta^{(l)}$ are jointly optimized during training to adaptively control multi-scale information flow. This refinement is applied recursively from coarse to fine levels, resulting in the final confidence score map $C_i^{\text{final}}(x,y)$ at the highest resolution.

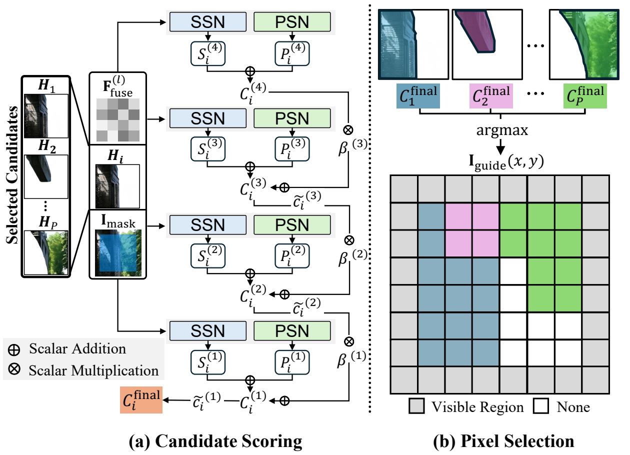
To construct the semantic guidance image $\mathbf{I}_{\text{guide}}$, we perform discrete pixel-wise selection using the refined confidence scores $C_i^{\text{final}}(x,y)$. At each location (x,y), the model selects the candidate index with the highest confidence:

$$i^*(x,y) = \underset{i\in\{1,2,\ldots,P\}}{\text{argmax}}\; C_i^{\text{final}}(x,y). \quad (9)$$

If the highest score is below a predefined threshold, the pixel remains masked. Using the selected indices $i^*(x,y)$, the final guidance image is assembled as:

$$\mathbf{I}_{\text{guide}}(x,y) = \mathbf{H}_{i^*(x,y)}(x,y). \quad (10)$$

This selection strategy enables the model to choose the most suitable candidate for each pixel, even though the argmax operation in Eq. 10 is non-differentiable. The loss is computed on the final guidance image, and the gradients are propagated to the SSN and the fusion transformer, which learn to assign higher scores to consistent candidates.



(a) Candidate Scoring **(b) Pixel Selection**

Figure 4. Hierarchical Pixel Selection in the CSF-Net. (a) The Structure Score Network (SSN) and Perceptual Score Network (PSN) compute confidence scores at each scale using fused features and the masked input. Multi-scale consistency is enforced via learnable coefficients β. (b) At the finest scale, the highest-scoring candidate is selected for each pixel to form the semantic guidance image $\mathbf{I}_{\text{guide}}$.

4.4. Loss Functions

Our model is trained end-to-end, jointly optimizing the Context-Semantic Feature Fusion Transformer (Sec. 4.2) and the Hierarchical Pixel Selection (Sec. 4.3). While the fused semantic feature map $\mathbf{F}_{\text{fuse}}$ is not directly supervised, it is implicitly guided by the loss applied to the guidance image $\mathbf{I}_{\text{guide}}$.

To ensure that the guidance image $\mathbf{I}_{\text{guide}}$ appears natural and aligns well with the ground truth in the filled regions, we define a reconstruction loss $\mathcal{L}_{\text{recon}}$. Note that the loss is applied only to the pixels actually filled by the proposed CSF-Net within the masked area. It consists of three components as follows:

$$\mathcal{L}_{\text{recon}} = \mathcal{L}_1 + \mathcal{L}_{\text{perc}} + \mathcal{L}_{\text{smooth}}, \quad (11)$$

where $\mathcal{L}_1$ measures pixel-wise differences in the inpainted region, $\mathcal{L}_{\text{perc}}$ encourages perceptual similarity to the ground truth by comparing deep features from a pretrained VGG network [9]. $\mathcal{L}_{\text{smooth}}$ discourages abrupt or inconsistent candidate switching between neighboring pixels. Overall, $\mathcal{L}_{\text{recon}}$ guides $\mathbf{I}_{\text{guide}}$ to closely match the ground truth in terms of both pixel accuracy and perceptual quality, while ensuring a smooth and coherent inpainting result.

To promote consistency across decoder levels, we introduce a hierarchical consistency loss $\mathcal{L}_{\text{hier}}$ that encourages finer-scale reconstructions to align with coarser-scale outputs:

$$\mathcal{L}_{\text{hier}} = \frac{1}{L-1}\sum_{l=1}^{L-1}\left\|\mathbf{I}^{(l)} - D(\mathbf{I}^{(l-1)})\right\|_1, \quad (12)$$

Methods	Places365 [26]									References
	Center Box (80%)			Center Box (50%)			Random (50–80%)			
	FID↓	LPIPS↓	C@m↑	FID↓	LPIPS↓	C@m↑	FID↓	LPIPS↓	C@m↑	
MaskGIT [2]	21.66	0.425	0.630	6.842	0.175	0.726	44.32	0.428	0.684	CVPR 2022
LaMa [22]	16.72	0.346	0.688	5.858	0.131	0.777	11.55	0.355	0.743	WACV 2022
MAT [12]	17.50	0.364	0.670	4.492	0.139	0.758	10.77	0.350	0.742	CVPR 2022
DLID [3]	26.16	0.405	0.680	6.660	0.148	0.795	17.12	0.384	0.725	CVPR 2024
ASUKA [23]	10.10	0.377	0.701	4.408	0.143	0.755	5.835	0.332	0.802	CVPR 2025
CSF-Net + ASUKA (Ours)	**9.434**	**0.332**	**0.702**	**3.612**	**0.105**	**0.796**	**5.324**	**0.325**	**0.803**	WACV 2026

Table 1. Quantitative comparison between state-of-the-art inpainting methods and our CSF-enhanced model (shown with ASUKA integration) at 256 × 256 resolution on Places365. **Bold** indicates the best performance for each metric.

where $\mathbf{I}^{(l)}$ denotes the reconstructed guidance image at decoder level l, and $D\left(\mathbf{I}^{(l-1)}\right)$ represents the bilinearly downsampled output from the previous coarser level. This loss encourages pixel-wise consistency across scales, helping the decoder produce stable and coherent results throughout the coarse-to-fine refinement process.

The total loss is defined as:

$$\mathcal{L}_{\text{total}} = \lambda \cdot \mathcal{L}_{\text{recon}} + (1-\lambda) \cdot \mathcal{L}_{\text{hier}}, \tag{13}$$

where each term contributes to reconstruction quality and hierarchical consistency, respectively.

5. Experimental Results

5.1. Evaluation Settings

Datasets and Settings. We conduct our experiments on the Places365 [26], a widely used scene-centric benchmark for image inpainting, and additionally evaluate on COCOA [27] as a test set to verify the effectiveness of our method in object-centric scenarios. Another representative benchmark is CelebA-HQ [10], which primarily consists of aligned human face images. Since our framework emphasizes object-level amodal completion, CelebA-HQ [10] is not suitable for our setting. Therefore, we focus on Places365 [26], COCOA [27], which contain diverse scene categories with complex structures, providing a more appropriate testbed for evaluating our method. We randomly sample 30 images per category (365 categories in total), resulting in 10,950 training images. The official validation set of Places365 [26] and 1,450 images from COCOA [27] are used as the evaluation and test sets, respectively. All images are resized to a resolution of 256 × 256. Since CSF-Net is designed to fuse outputs from pretrained models rather than training an entire network, it requires significantly fewer training samples compared to conventional end-to-end inpainting approaches. We further define three mask settings for our experiments: (1) Center Box 50%, (2) Center Box 80%, and (3) RandomBrush masks ranging from 50–80% area coverage with mixed brush strokes and rectangles, following the strategy used in MAT [12]. The Center Box masks simulate structured occlusions by masking a fixed-size box at the image center, while RandomBrush generates more irregular and organic missing regions. We train our model using the AdamW optimizer with a learning rate of 2×10^{-4}, a batch size of 12, and 200 training epochs. All experiments are conducted on a single NVIDIA A100 GPU.

Methods	Places365 [26]					
	Random (50–80%)			Center Box (80%)		
	FID↓	LPIPS↓	C@m↑	FID↓	LPIPS↓	C@m↑
LaMa [22]	11.55	0.355	0.743	16.72	0.346	0.688
+CSF-Net(Ours)	**11.17**	**0.309**	**0.761**	**15.93**	**0.326**	**0.701**
MAT [12]	10.77	0.350	0.742	17.50	0.364	0.670
+CSF-Net(Ours)	**10.61**	**0.325**	**0.744**	**16.08**	**0.338**	**0.743**
DLID [3]	17.12	0.384	0.725	26.16	0.405	0.680
+CSF-Net(Ours)	**15.45**	**0.362**	**0.732**	**21.13**	**0.400**	**0.693**
ASUKA [23]	5.835	0.332	0.802	10.10	0.377	0.701
+CSF-Net(Ours)	**5.324**	**0.325**	**0.803**	**9.434**	**0.332**	**0.702**

Methods	COCOA [27]					
	Random (50–80%)			Center Box (80%)		
	FID↓	LPIPS↓	C@m↑	FID↓	LPIPS↓	C@m↑
LaMa [22]	64.98	0.375	0.719	83.56	0.428	0.653
+CSF-Net(Ours)	**62.55**	**0.334**	**0.720**	**71.20**	**0.373**	**0.661**
MAT [12]	60.24	0.351	0.723	78.37	0.459	0.652
+CSF-Net(Ours)	**58.07**	**0.298**	**0.780**	**75.65**	**0.394**	**0.688**
DLID [3]	71.90	0.395	0.708	90.37	0.441	0.641
+CSF-Net(Ours)	**63.93**	**0.373**	**0.742**	**72.04**	**0.417**	**0.665**
ASUKA [23]	69.77	0.446	0.742	50.90	0.399	0.666
+CSF-Net(Ours)	**54.56**	**0.387**	**0.786**	**49.77**	**0.376**	**0.708**

Table 2. Performance comparison of models with CSF-Net integration on Places365 [26] and COCOA [27].

Evaluation Metrics. Our main evaluation objectives are to evaluate the visual quality, perceptual fidelity, and semantic consistency of the inpainted images. To this end, we adopt the following three metrics: FID [8] to assess global realism, LPIPS [24] for perceptual similarity, and C@m [23] for evaluating object-level semantic consistency

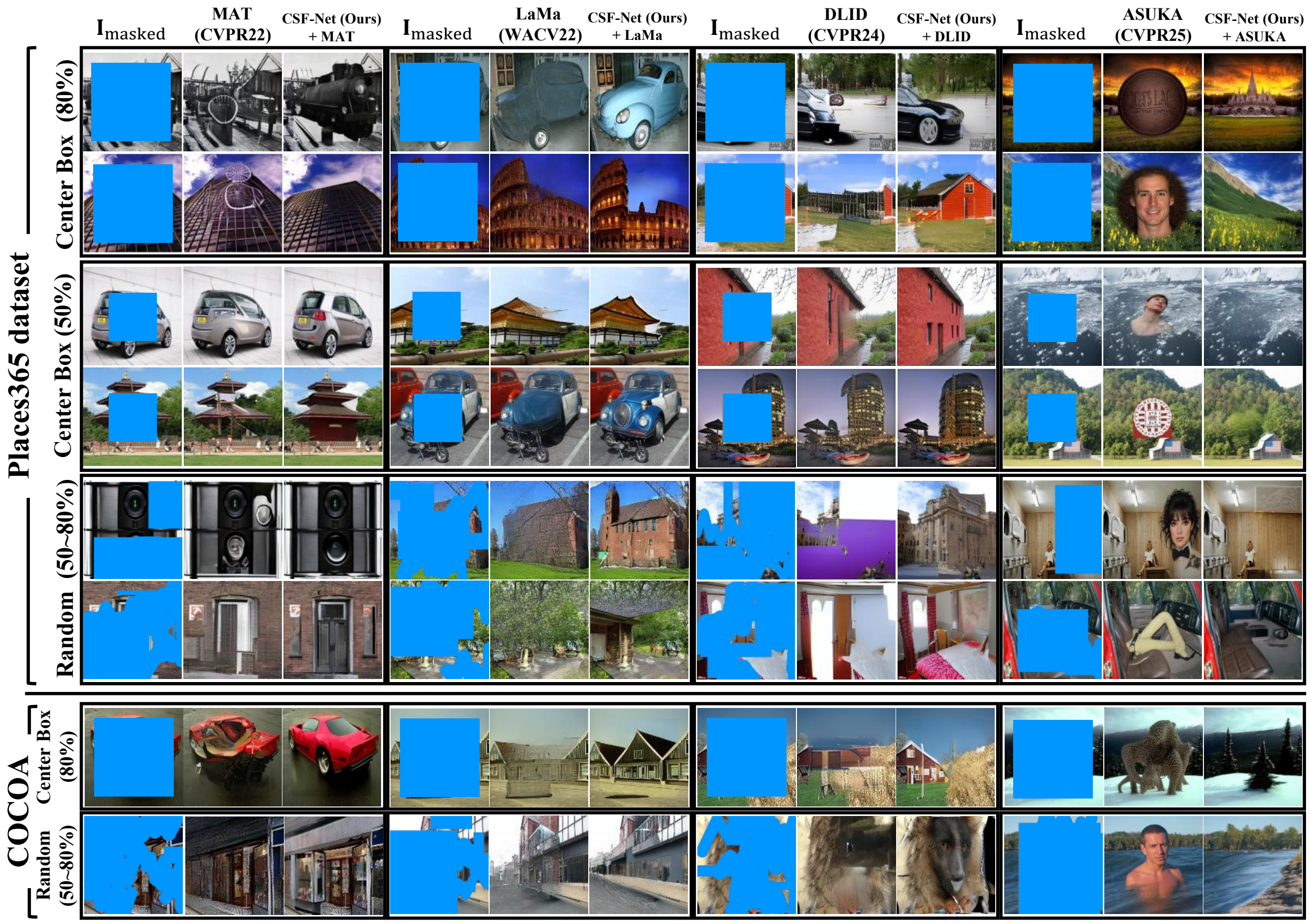


Figure 5. Comprehensive qualitative comparison under different mask configurations using the `Places365` [26] (evaluated on all three mask types) and `COCOA` [27] (evaluated on Center Box 80% and RandomBrush 50–80%).Our CSF-Net consistently generates clearer and more coherent results compared to baseline methods across diverse scenes and mask types, effectively reducing object hallucination.

within masked regions. C@m is a CLIP-based metric that measures the similarity between the restored region and the ground truth, effectively capturing object-level semantic consistency. As CSF-Net aims to restore semantically meaningful content within masked regions, C@m offers a more appropriate evaluation by measuring object-level alignment using CLIP-based features. To complement the quantitative evaluation and provide a more comprehensive assessment, we present qualitative results illustrating visual fidelity and semantic plausibility of the inpainted outputs.

5.2. Comparison with State-of-the-Art Methods

We compare CSF-Net with state-of-the-art inpainting methods. Table 1 presents the performance comparison on the `Places365` [26]. While CSF-Net can be integrated into various inpainting frameworks, we adopt ASUKA as the baseline in this experiment.

To evaluate the general applicability of CSF-Net, we integrate it into four state-of-the-art inpainting models as shown in Tab 2. The results demonstrate that performance gains are particularly notable under challenging Center Box masks (50% and 80%), while improvements with RandomBrush masks (50–80%) are somewhat smaller than those for Center Box masks, due to their irregular patterns allowing richer use of surrounding context. Nevertheless, CSF-Net consistently improves the performance of all base models across various masking conditions and evaluation metrics. Note that the `COCOA` [27] dataset contains relatively fewer samples, which leads to higher FID values overall. However, even under this limitation, CSF-Net yields consistent performance improvements, highlighting its robustness. These results demonstrate that CSF-Net provides effective object-aware guidance and consistently improves performance regardless of the underlying inpainting model. Moreover, a key advantage of CSF-Net is that it does not require any architectural modifications to the baseline inpainting frameworks. The proposed method can be applied simply by generating a guidance image from the input mask.

Figure 5 illustrates the visual differences between baseline methods and their CSF-Net–integrated versions. Our method consistently produces sharper and more structurally coherent inpainting results across various mask types. While ASUKA generates perceptually plausible outputs, it often suffers from object hallucination under large-mask conditions due to the lack of explicit semantic guidance. By leveraging semantic cues and surrounding structural context, CSF-Net alleviates this issue through its fusion strategy, resulting in more realistic and semantically faithful completions.

5.3. Ablation Study

All ablation studies are conducted using the LaMa under the Center Box (80%) masking condition to ensure consistent analysis. In Sec. 4.1, we proposed a candidate generation based on Amodal Completion and selected the most reliable candidates using consistency scores. Table 3 shows the inpainting performance under different scoring strategies, showing that the combined metric $S_{\text{MSE}} + S_{\text{LPIPS}}$ achieves the best overall performance.

We investigate the effect of varying the number of candidate completions (P) in CSF-Net. As shown in Tab. 4, using P=3 achieves the best performance across all metrics. Increasing P to 5 or 10 results in degraded inpainting performance. This is because top-ranked candidates already provide sufficient semantic diversity, while adding more may introduce redundant or low-quality results that interfere with the fusion process. We also compare the case of P=1, which directly selects the top-ranked amodal completion without going through the Fusion Transformer and pixel selection process in CSF-Net. Simply using the top-ranked candidate without fusion leads to lower performance in all metrics. This shows that the fusion process in CSF-Net is essential for generating better and coherent results.

We further examine the effect of the Pixel Selection described in Sec. 4.3. As shown in Tab. 5, incorporating pixel selection improves all evaluation metrics, demonstrating its importance in refining the guidance image with spatially coherent content. Finally, we evaluate the impact of the encoder design proposed in Sec. 4.2. We tested a single-encoder design that takes the concatenated inputs $[\mathbf{I}_{\text{mask}}, \mathcal{H}_{\text{sel}}]$ along the channel dimension. As shown in Tab. 6, the dual-encoder design (i.e., separate context and semantic branches) significantly outperforms the single-encoder across all metrics. This confirms that separating contextual and candidate inputs enables better semantic disentanglement and leads to improved feature representations.

Scores for Candidate Generation	FID↓	LPIPS↓	C@m↑
MSE only (S_{MSE})	25.91	0.379	0.676
LPIPS only (S_{LPIPS})	18.82	0.361	0.688
MSE+LPIPS ($S_{\text{MSE}} + S_{\text{LPIPS}}$)	**15.93**	**0.326**	**0.701**

Table 3. Effect of consistency scores for candidate selection.

# of candidates (P)	FID↓	LPIPS↓	C@m↑
P = 1	18.67	0.411	0.679
P = 3	**15.93**	**0.326**	**0.701**
P = 5	16.37	0.394	0.690
P = 10	16.51	0.406	0.682

Table 4. Ablation study on the number of candidate completions (P) and the impact of fusion in CSF-Net.

Encoder design	FID↓	LPIPS↓	C@m↑
w/o Pixel Selection	18.54	0.421	0.684
Ours	**15.93**	**0.326**	**0.701**

Table 5. Ablation study on Pixel Selection in CSF-Net.

Encoder design	FID↓	LPIPS↓	C@m↑
Single-encoder	21.08	0.430	0.689
Ours	**15.93**	**0.326**	**0.701**

Table 6. Ablation study on encoder design in CSF-Net.

6. Conclusions and Future Work

In this paper, We proposed CSF-Net, a transformer-based fusion framework that introduces object-level semantic guidance into the image inpainting process. By leveraging a pretrained amodal completion model, CSF-Net generates multiple structure-aware semantic candidates, which are fused with contextual information to produce a semantic guidance image ($\mathbf{I}_{\text{guide}}$). This guidance enables more accurate and semantically aligned inpainting, particularly in challenging large-mask scenarios where contextual cues are limited. CSF-Net can be seamlessly integrated into various inpainting backbones without any architectural modifications, demonstrating both its generality and practicality for real-world applications. Extensive experiments on the `Places365` [26] and `COCOA` [27] datasets demonstrate that CSF-Net consistently improves performance across multiple inpainting baselines and evaluation metrics.

While CSF-Net shows strong results and wide compatibility, there are still areas for improvement. First, the performance of CSF-Net can be affected by the quality of the amodal completion candidates. Although we apply a filtering step (Sec. 4.1) to remove noisy or irrelevant outputs, the model may struggle when all candidates lack meaningful semantic information. Second, it is used independently from the amodal completion and inpainting models. This modular design makes it easy to apply to other systems, but jointly training all components could improve the results.

References

[1] Marcelo Bertalmio, Guillermo Sapiro, Vincent Caselles, and Coloma Ballester. Image inpainting. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 417–424, 2000.

[2] Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T. Freeman. Maskgit: Masked generative image transformer. In *CVPR*, pages 11315–11325, 2022.

[3] Haiwei Chen and Yajie Zhao. Don't look into the dark: Latent codes for pluralistic image inpainting. In *CVPR*, pages 7591–7600, 2024.

[4] Ciprian Corneanu, Raghudeep Gadde, and Aleix M. Martinez. Latentpaint: Image inpainting in latent space with diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 4334–4343, 2024.

[5] Antonio Criminisi, Patrick Pérez, and Kentaro Toyama. Region filling and object removal by exemplar-based image inpainting. *IEEE Transactions on image processing*, 13(9): 1200–1212, 2004.

[6] Kiana Ehsani, Roozbeh Mottaghi, and Ali Farhadi. Segan: Segmenting and generating the invisible. In *CVPR*, 2018.

[7] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2014.

[8] 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. In *Advances in Neural Information Processing Systems*, pages 6626–6637, 2017.

[9] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016.

[10] 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.

[11] Ke Li and Jitendra Malik. Amodal instance segmentation. In *European Conference on Computer Vision*, pages 677–693. Springer, 2016.

[12] Wenbo Li, Zhe Lin, Kun Zhou, Lu Qi, Yi Wang, and Jiaya Jia. Mat: Mask-aware transformer for large hole image inpainting. In *CVPR*, pages 10758–10768, 2022.

[13] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[14] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *CVPR*, pages 11461–11471, 2022.

[15] Kamyar Nazeri, Eric Ng, Tony Joseph, Faisal Qureshi, and Mehran Ebrahimi. Edgeconnect: Structure guided image inpainting using edge prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, 2019.

[16] Ege Ozguroglu, Ruoshi Liu, Dídac Surís, Dian Chen, Achal Dave, Pavel Tokmakov, and Carl Vondrick. pix2gestalt: Amodal segmentation by synthesizing wholes. In *CVPR*, pages 3931–3940, 2024.

[17] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A. Efros. Context encoders: Feature learning by inpainting. In *CVPR*, 2016.

[18] Yurui Ren, Xiaoming Yu, Ruonan Zhang, Thomas H Li, Shan Liu, and Ge Li. Structureflow: Image inpainting via structure-aware appearance flow. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 181–190, 2019.

[19] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[20] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[21] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. pmlr, 2015.

[22] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2149–2159, 2022.

[23] Yikai Wang, Chenjie Cao, Junqiu Yu, Ke Fan, Xiangyang Xue, and Yanwei Fu. Towards enhanced image inpainting: Mitigating unwanted object insertion and preserving color consistency. In *CVPR*, pages 23237–23248, 2025.

[24] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pages 586–595, 2018.

[25] Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai. Pluralistic image completion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1438–1447, 2019.

[26] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017.

[27] Yan Zhu, Yuandong Tian, Dimitris Metaxas, and Piotr Dollár. Semantic amodal segmentation. In *CVPR*, pages 1464–1472, 2017.