SAIR: Learning Semantic-aware Implicit Representation

Canyu Zhang[1⋆], Xiaoguang Li[1⋆], Qing Guo[2 ⋆⋆], and Song Wang[1]

[1] University of South Carolina, USA
[2] IHPC and CFAR, Agency for Science, Technology and Research (A*STAR), Singapore
{canyu,xl22}@email.sc.edu, tsingqguo@ieee.org, songwang@cec.sc.edu

Abstract. Implicit representation of an image can map arbitrary coordinates in the continuous domain to their corresponding color values, presenting a powerful capability for image reconstruction. Nevertheless, existing implicit representation approaches only focus on building continuous appearance mapping, ignoring the continuities of the semantic information across pixels. Consequently, achieving the desired reconstruction results becomes challenging when the semantic information within input image is corrupted, such as when a large region is missing. To address the issue, we suggest learning *semantic-aware implicit representation (SAIR)*, that is, we make the implicit representation of each pixel rely on both its appearance and semantic information (*e.g.*, which object does the pixel belong to). To this end, we propose a framework with two modules: (1) a semantic implicit representation (SIR) for a corrupted image. Given an arbitrary coordinate in the continuous domain, we can obtain its respective text-aligned embedding indicating the object the pixel belongs. (2) an appearance implicit representation (AIR) based on the SIR. Given an arbitrary coordinate in the continuous domain, we can reconstruct its color whether or not the pixel is missed in the input. We validate the novel semantic-aware implicit representation method on the image inpainting task, and the extensive experiments demonstrate that our method surpasses state-of-the-art approaches by a significant margin.

Keywords: Image inpainting · Semantic-aware implicit representation · Visual language model

1 Introduction

Implicit neural representation has demonstrated surprising performance in the 2D image [5, 10], video [4] and novel view [29, 40, 49] reconstruction. They achieve this by extracting appearance features from 2D images using an encoder and

⋆ Canyu Zhang and Xiaoguang Li are co-first authors and contribute equally.
⋆⋆ Qing Guo is the corresponding author. Code is available at https://github.com/tsingqguo/sair.

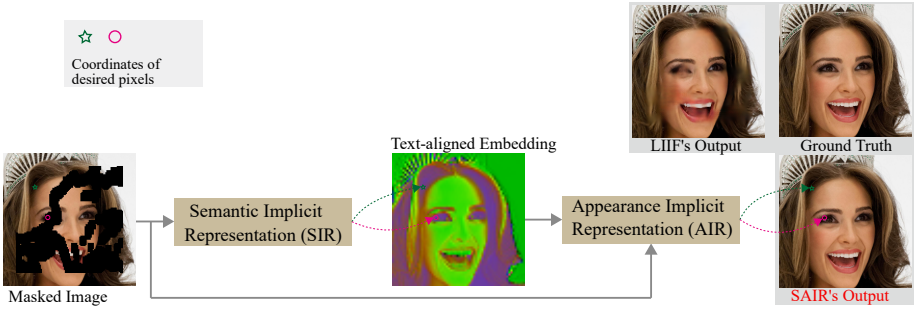

Fig. 1: Semantic-aware implicit representation (SAIR) is composed of semantic implicit representation (SIR) and appearance implicit representation (AIR). SIR processes the corrupted image and obtains its text-aligned embedding. AIR reconstructs the color.

then associating continuous coordinates with corresponding appearance features through a neural network, ultimately translating them into the RGB color space. However, these methods tend to neglect the semantic meaning of individual pixels. Reconstructed images may contain noticeable artifacts or lose crucial semantic information, especially when dealing with degraded input images, such as those with large missing regions. As shown in Fig. 1, when the local appearance information is missing around the woman's eye, previous implicit representation methods like LIIF [5] fall short in accurately reconstructing the missing pixels.

To address this issue, we propose a novel approach called Semantic-Aware Implicit Representation (SAIR). In SAIR, each pixel's implicit representation depends on both its appearance and semantic context (e.g., the object it belongs to). This integration of continuous semantic mapping mitigates the limitations of only employing appearance implicit representation. Consequently, even in cases of degraded appearance information, the network can produce high-quality outputs with the aid of semantic information. This enhancement benefits various image processing tasks, including image generation, inpainting, editing, and semantic segmentation. As illustrated in Fig. 1, our method surpasses the existing implicit neural representation approaches that rely solely on appearance mapping on the image inpainting task. Remarkably, even when confronted with severely degraded input images, *e.g.*, a large region misses, our approach still can accurately fill in the missing pixels, yielding a natural and realistic result.

The proposed semantic-aware implicit representation involved two modules: (1) a semantic implicit representation (SIR) for a corrupted image whose large regions miss. Given an arbitrary coordinate in the continuous domain, the SIR can obtain its respective text-aligned embedding indicating the object the pixel belongs to. (2) an appearance implicit representation (AIR) based on the SIR. Given an arbitrary coordinate in the continuous domain, AIR can reconstruct its color whether or not the pixel is missed in the input. Specifically, to implement the SIR, we first use the modified CLIP [32] encoder to extract the text-aligned embedding from the input image. This specific modification

(see Section 4.2) allows CLIP to output a spatial-aware embedding without introducing additional parameters and altering the feature space of CLIP. The text-aligned embedding can effectively reflect the pixel-level semantic information. However, this embedding has a smaller dimension compared to the input image. Moreover, when the input image is significantly degraded, the quality of the extracted embedding deteriorates considerably. To address those problems, we utilize the semantic implicit representation within the text-align embedding. This process not only expands the feature dimensions but also compensates for missing information when the input image is severely degraded.

To implement AIR, we utilize a separate implicit representation function that takes three inputs: the appearance embedding extracted from the input image using a CNN-based network, the enhanced text-aligned embedding by SIR (see Section 4.3), and the pixel coordinates which indicating the location information. This allows AIR to leverage both appearance and semantic information simultaneously. We validate the novel semantic-aware implicit representation (SAIR) method on the image inpainting task and conducted comprehensive experiments on the widely utilized CelebAHQ [26] and ADE20K [50] datasets. The extensive experiments demonstrate that our method surpasses state-of-the-art approaches by a significant margin.

In summary, our main contributions are listed as follows:

- We acknowledge the limitation of existing implicit representation methods that rely solely on building continuous appearance mapping, hindering their effectiveness in handling severely degraded images. To address this limitation, we introduce Semantic-Aware Implicit Representation (SAIR).
- We propose a novel framework to implement SAIR which involves two modules:(1) Semantic implicit representation (SIR) for enhancing semantic embedding, and (2) Appearance implicit representation (AIR), which builds upon SIR to leverage both semantic and appearance information simultaneously.
- Comprehensive experiments on the widely utilized CelebAHQ [26] and ADE20K [50] datasets demonstrate that our proposed method surpasses previous implicit representation approaches by a significant margin across four commonly used image quality evaluation metrics, *i.e.*, PSNR, SSIM, L1, and LPIPS.

2 Related Work

Implicit neural representation. Implicit neural functions find applications across a wide spectrum of domains, encompassing sound signals [36], 2D images [2, 5, 11, 15], videos [4], and 3D shapes [8, 12, 43, 44]. These functions offer a means to continuously parameterize signals, enabling the handling of diverse data types, such as point clouds in IM-NET [6] or video frames in NERV [3]. Implicit neural functions have demonstrated their ability to generate novel views, as exemplified by Nerf [29], which leverages an implicit neural field to synthesize new perspectives. Within the domain of image processing, methods like LIIF [5]

establish a connection between pixel features and RGB color, facilitating arbitrary-sized image super-resolution. LTE [15], a modification of LIIF, extends this concept by incorporating additional high-frequency information in Fourier space to address the limitations of a standalone MLP. Recently, However, these approaches lack explicit consideration of semantic information during training, which can result in potential inconsistencies at the semantic level.

Image inpainting. Image inpainting techniques [1, 7, 16, 17, 27, 31, 39, 45] are designed to restore corrupted image regions by leveraging information from non-missing portions, which can benefit many downstream tasks [18, 19]. Established methods such as [22, 30, 34] employ edge information or smoothed images to guide the restoration process. Another noteworthy approach, as introduced by [25], relies on valid pixels to infer the missing ones. Furthermore, [9] incorporates an element-wise convolution block to reconstruct missing regions around the mask boundary while utilizing a generative network to address other missing areas. Extending upon these techniques, [20] advances the inpainting process by implementing element-wise filtering at both feature and image levels. Feature-level filtering is tailored for substantial missing regions, while image-level filtering refines local details. However, contemporary inpainting models face challenges when confronted with substantial missing regions, as reliable neighborhood features are often lacking. In such scenarios, text prompts prove invaluable as a robust guidance mechanism, enhancing the inpainting process.

Image-text cross-model learning. Cross-model networks have gained substantial attention across various image processing domains, including image semantic segmentation [28, 41, 52], image generation [21, 33, 38, 53], and visual question answering (VQA) [42, 48]. For instance, DF-GAN [38] represents a one-stage text-to-image backbone capable of directly synthesizing high-resolution images. In the realm of image segmentation, [28] leverages latent diffusion models (LDMs) to segment text-based real and AI-generated images. In VQA, [24] incorporates explicit object region information into the answering model. Furthermore, [46] harnesses text to assist the model in generating missing regions within images, thereby pushing the boundaries of image inpainting tasks. Additionally, language models like CLIP [32] have emerged to bridge the gap between image and semantic features. In this paper, we explore the influence of semantic information within the implicit neural function on the image inpainting task. Through the integration of semantic information, our objective is to endow the model with a more profound comprehension of the semantic meaning associated with specific image coordinates.

3 Preliminary: Local Image Implicit Representation

Given an image $\mathbf{I}$, an implicit representation for the image is to map coordinates in the continuous domain to corresponding color values; that is, we have

$$c_{\mathbf{p}} = \sum_{\mathbf{q} \in \mathcal{N}_{\mathbf{p}}} \omega_{\mathbf{q}} f_\theta(\mathbf{z}_{\mathbf{q}}^{\text{app}}, \text{dist}(\mathbf{p}, \mathbf{q})), \tag{1}$$

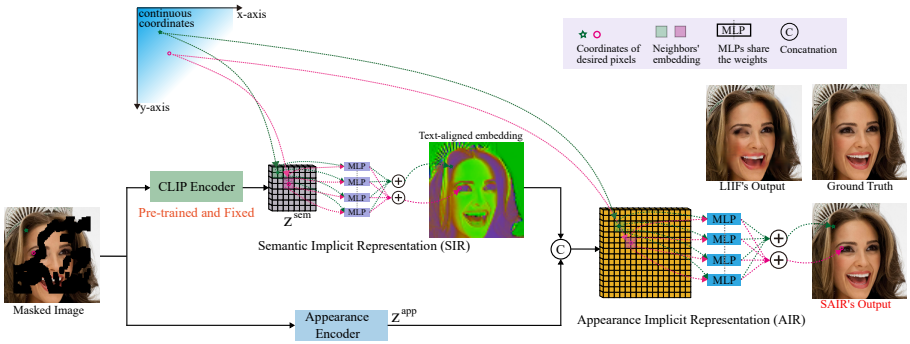

Fig. 2: Pipeline of our semantic-aware implicit representation (SAIR). The semantic implicit representation (SIR) is used to complete the missing semantic information. The appearance implicit representation (AIR) is used to complete missing details

where $\mathbf{p}$ is the continuous coordinates, the output $c_{\mathbf{p}}$ is the color of the pixel $\mathbf{p}$, $\mathcal{N}_{\mathbf{p}}$ contains all neighboring pixels of $\mathbf{p}$ within the image $\mathbf{I}$, $f_\theta(\cdot)$ is an MLP for coordinate-color mapping, $\omega_{\mathbf{q}}$ is the weight of q, and $\mathbf{z}_{\mathbf{q}}^{\text{app}}$ is the appearance feature of pixel $\mathbf{q}$. Note that, all pixels in $\mathcal{N}_{\mathbf{p}}$ are sampled from the input image $\mathbf{I}$ and their features $\{\mathbf{z}_{\mathbf{q}}^{\text{app}}\}$ are extracted through an encoder network for handling $\mathbf{I}$. Intuitively, the MLP is to transform the appearance embedding of a neighboring pixel to the color of the pixel $\mathbf{p}$ based on their spatial distance. Recent works have demonstrated that training above implicit representation via image quality loss (*e.g.*, L_1 loss) could remove noise or perform super-resolution [6, 11, 15]. However, when the neighboring pixels in $\mathcal{N}_{\mathbf{p}}$ miss, the implicit representation via Equation 1 is affected. As shown in Fig. 3, the existing implicit representation approaches cannot properly reconstruct the pixels within missing regions.

4 Semantic-aware Implicit Representation (SAIR)

4.1 Overview

To address the issue, we propose the semantic-aware implicit representation (SAIR), which contains two key modules, *i.e.*, semantic implicit representation (SIR) and appearance implicit representation (AIR). The first objective is to create a continuous semantic representation that enables us to fill in the missing semantic information within the input image. The second objective is to develop a continuous appearance representation that allows us to restore missing details.

Specifically, given an input image $\mathbf{I} \in \mathbb{R}^{H \times W \times 3}$ that may contain some missing regions indicated by a mask $\mathbf{M} \in \mathbb{R}^{H \times W}$, we aim to build the semantic implicit representation (SIR) to predict semantic embedding of an arbitrary given pixel whose coordinates could be non-integer values. The embedding could indicate the object the pixel belongs to. We formulate the process as

$$\mathbf{z}_{\mathbf{p}}^{\text{sem}} = \text{SIR}(\mathbf{I}, \mathbf{M}, \mathbf{p}), \tag{2}$$

where $\mathbf{z}_{\mathbf{p}}^{\text{sem}}$ denotes the semantic embedding of the pixel $\mathbf{p}$. Intuitively, we require the SIR to have three properties: ❶ The predicted semantic embedding should be well aligned with the extract category of the object the pixel belongs to. ❷ If the given coordinate (*i.e.*, $\mathbf{p}$) is within unlost regions but with non-integer values, SIR could estimate its semantic embedding accurately. This requires SIR to have the capability of interpolation. ❸ If the specified coordinate is within missing regions, SIR could complete the semantic embedding properly. We extend the local image implicit representation to the embedding level with text-aligned embeddings and propose the SIR in Section 4.2 to achieve the above three properties.

After getting the semantic embedding of the desired pixel, we further estimate the appearance (*e.g.*, color) of the pixel via the appearance implicit representation; that is, we have

$$c_{\mathbf{p}} = \text{AIR}(\mathbf{I}, \text{SIR}, \mathbf{p}), \tag{3}$$

where $c_{\mathbf{p}}$ denotes the color of the desired pixel $\mathbf{p}$. Intuitively, AIR is to predict the color of $\mathbf{p}$ according to the built semantic implicit representation (SIR) and input appearance. We detail the process in Section 4.3.

4.2 Semantic Implicit Representation (SIR)

We first use the modified CLIP model to extract the text-aligned embedding as the semantic embedding. Specifically, inspired by the recent work MaskCLIP [51], we remove the query and key embedding layers of the raw CLIP model and restructured the value-embedding and final linear layers into two separate 1×1 convolutional layers. This adjustment is made without introducing additional parameters or altering the feature space of CLIP, allowing the CLIP output a spatial-aware embedding tensor. Given the input image $\mathbf{I} \in \mathbb{R}^{H \times W \times 3}$, we feed it into the modified image encoder of CLIP and output a tensor $\mathbf{Z}^{\text{sem}} \in \mathbb{R}^{h \times w \times c}$ where h, w, and c are the height, width, and channel numbers. Note that $\mathbf{Z}^{\text{sem}}$ is not pixel-wise embedding with $h \ll H$ and $w \ll W$, which have much lower resolution than $\mathbf{I}$. MaskCLIP employs the naive resize operation to map the $\mathbf{Z}^{\text{sem}}$ to the same size as the input image, which cannot complete the missing semantic information. Instead, we propose to extend local image implicit representation to the text-aligned embedding and formulate the SIR as

$$\mathbf{z}_{\mathbf{p}}^{\text{sem}} = \text{SIR}(\mathbf{I}, \mathbf{M}, \mathbf{p}) = \sum_{\mathbf{q} \in \mathcal{N}_{\mathbf{p}}} \omega_{\mathbf{q}} f_\theta([\mathbf{z}_{\mathbf{q}}^{\text{sem}}, \mathbf{M}[\mathbf{q}]], \text{dist}(\mathbf{p}, \mathbf{q})), \tag{4}$$

where $\mathbf{z}_{\mathbf{q}}^{\text{sem}} = \mathbf{Z}^{\text{sem}}[\mathbf{q}]$ is the embedding of the $\mathbf{q}$ location at $\mathbf{Z}^{\text{sem}}$, and $\mathcal{N}_{\mathbf{p}}$ denotes the set of neighboring coordinates around $\mathbf{p}$. $\text{dist}(\mathbf{p}, \mathbf{q})$ measures the distance between $\mathbf{p}$ and $\mathbf{q}$. $f_\theta(\cdot)$ is a MLP with the θ being the weights. Intuitively, $f_\theta(\cdot)$ is to estimate the text-aligned embedding of the location $\mathbf{p}$ according to the known embedding of $\mathbf{q}$ and the spatial relationship between $\mathbf{p}$ and $\mathbf{q}$. Finally, all estimations based on different $\mathbf{q}$ are weightly combined through $\omega_{\mathbf{q}}$ that is also set as the area ratio of $\mathbf{p}$-$\mathbf{q}$-made rectangle in the whole neighboring area.

4.3 Appearance Implicit Representation (AIR)

With the built SIR, we aim to build the appearance implicit representation (AIR) that can estimate the colors of arbitrarily specified coordinates. In first step, we use a CNN-based appearance encoder to generate appearance feature $\mathbf{Z}^{\text{app}} = \textsc{AppEncoder}(\mathbf{I}, \mathbf{M})$, and $\mathbf{Z}^{\text{app}} \in \mathbb{R}^{H \times W \times C}$. Given a pixel's coordinates $\mathbf{p}$, we predict its color by

$$c_{\mathbf{p}} = \text{AIR}(\mathbf{I}, \mathbf{M}, \text{SIR}, \mathbf{p}) = \sum_{\mathbf{q} \in \mathcal{N}_{\mathbf{p}}} \omega_{\mathbf{q}} f_\beta([\mathbf{z}_{\mathbf{q}}^{\text{app}}, \text{SIR}(\mathbf{I}, \mathbf{M}, \mathbf{q})], \text{dist}(\mathbf{p}, \mathbf{q})), \quad (5)$$

where $\mathbf{z}_{\mathbf{q}}^{\text{app}} = \mathbf{Z}^{\text{app}}[\mathbf{q}]$ is the appearance embedding of $\mathbf{q}$-th pixel. The function f_β is a MLP with the β being its weights. Intuitively, we estimate the color of the $\mathbf{p}$-th pixel according to the appearance and semantic information of the neighboring pixels by jointly considering the spatial distance. For example, if a pixel $\mathbf{p}$ misses, the appearance feature of $\mathbf{p}$ (*i.e.*, $\mathbf{z}_{\mathbf{p}}^{\text{app}}$) is affected and tends to zero while the semantic information could be inferred from contexts. As shown in Fig. 2, even though the pixels around the left eye miss, we still know the missed pixels belong to the left eye category.

4.4 Implementation Details

Network architecture. We utilize and modify the pre-trained ViT-B/16 image encoder of CLIP model to extract the semantic embedding. And we set the AppEncoder as a convolutional neural network and detail the architecture in Table 1, which is capable of generating features of the same size as the input image. Our MLP modules $f_\alpha(\cdot)$ and $f_\beta(\cdot)$ are four-layer MLP with ReLU activation layers, and the hidden dimension is 256. **Loss functions.** During the training phase, we employ the L1 loss to measure the discrepancy between the predicted pixel color and the ground truth pixel color, which is utilized for calculating the reconstruction loss $\mathcal{L}_1$. To guarantee the feature after SIR remains in text-aligned feature space, we choose L1 loss to quantify the dissimilarity between the unmasked image's text-aligned feature $\mathbf{Z}_{\text{unmask}}^{\text{sem}} \in \mathbf{R}^{h \times w \times c}$, and the SIR reconstructed feature $\mathbf{Z}_{\text{LR}}^{\text{reconstructed}} \in \mathbf{R}^{h \times w \times c}$ without changing the resolution. The final loss function is $\mathcal{L} = \mathcal{L}_1 + \alpha \mathcal{L}_2$.

Table 1: Architecture of AppEncoder. H × W is the resolution of the input.

Output size	Operation
H × W	Conv(4, 64, 7, 1, 3), ReLU
H/2 × W/2	Conv(64, 128, 4, 2, 1), ReLU
H/4 × W/4	Conv(128, 256, 4, 2, 1), ReLU
Resnet × 8	
H/4 × W/4	Conv(256, 256, 3, 1, 1), ReLU
H/4 × W/4	Conv(256, 256, 3, 1, 1), ReLU
H/2 × W/2	Conv(256, 128, 4, 2, 1), ReLU
H × W	Conv(128, 64, 4, 2, 1), ReLU

Hyperparameters. We employ the Adam optimizer with parameters ($\beta_1 = 0.9$, $\beta_2 = 0.999$). The learning rate is set to 0.0001 and is halved every 100 epochs. Our models are trained for 200 epochs on two NVIDIA Tesla V100 GPUs, and the batch size is set to 16.

Table 2: Comparison results on the CelebAHQ dataset across varied mask ratios.

Method	0%-20%				20%-40%				40%-60%			
	PSNR↑	SSIM↑	L1 ↓	LPIPS↓	PSNR↑	SSIM↑	L1 ↓	LPIPS↓	PSNR↑	SSIM↑	L1 ↓	LPIPS↓
EdgeConnect	34.53	0.964	0.005	0.038	27.30	0.889	0.025	0.104	22.32	0.771	0.035	0.195
RFRNet	34.93	0.966	0.005	0.035	27.50	0.890	0.024	0.100	22.77	0.775	0.033	0.185
JPGNet	35.86	0.972	**0.004**	0.040	28.18	0.909	0.023	0.119	22.32	0.771	0.035	0.195
LAMA	36.04	0.973	0.008	0.024	29.14	0.932	0.020	0.029	22.94	0.854	0.033	0.152
MISF	36.32	0.976	0.012	0.019	29.85	0.932	0.021	0.055	23.91	0.868	0.042	0.133
LIIF	35.27	0.969	0.012	0.023	28.80	0.923	0.026	0.043	23.30	0.830	0.051	0.136
SAIR	**37.97**	**0.977**	0.010	**0.016**	**31.49**	**0.944**	**0.019**	**0.025**	**24.87**	**0.870**	**0.031**	**0.124**

5 Experimental Results

5.1 Setups

Datasets. We validate the effectiveness of proposed method through comprehensive experiments conducted on two widely used datasets: CelebAHQ [14] and ADE20K [50]. CelebAHQ is a large-scale dataset consisting of 30,000 high-resolution human face images, selected from the CelebA dataset [26]. These face images are categorized into 19 classes, and for our experiments, we use 25,000 images for training and 5,000 images for testing purposes. ADE20K, on the other hand, is a vast dataset comprising both outdoor and indoor scenes. It consists of 25,684 annotated training images, covering 150 semantic categories. We leverage this dataset to evaluate our method's performance on scene inpainting tasks. To create masked images for our experiments, we utilize the mask dataset [25] similar as the previous works [20]. This dataset offers over 9,000 irregular binary masks with varying mask ratios, spanning from 0% to 20%, 20% to 40%, and 40% to 60%. These masks are instrumental in generating realistic inpainting scenarios for evaluation purposes.

Baselines. We enhance our approach by incorporating semantic representations based on previous implicit neural function model LIIF [5]. By modifying image encoder and integrating semantic information, we obtain the semantic-aware implicit function, denoted as SAIR. For comparative analysis, we select state-of-the-art inpainting methods StructFlow [34], EdgeConnect [30], RFR-Net [16], JPGNet [9], LAMA [37], MISF [20], and the implicit neural function without semantic information LIIF [5] as our baselines.

Evaluation metrics. To assess the performance of all methods, we utilize four commonly employed image quality evaluation metrics: peak signal-to-noise ratio (PSNR), structural similarity index (SSIM), L1 loss, and learned perceptual image patch similarity (LPIPS) [47]. PSNR, SSIM, and L1 offer insights into the quality of the generated image, while LPIPS quantifies the perceptual distance between the restored image and the ground truth.

Table 3: Comparison results on the ADE20K dataset across varied mask ratios.

Method	0%-20%				20%-40%				40%-60%			
	PSNR↑	SSIM↑	L1 ↓	LPIPS↓	PSNR↑	SSIM↑	L1 ↓	LPIPS↓	PSNR↑	SSIM↑	L1 ↓	LPIPS↓
EdgeConnect	30.91	0.948	0.007	0.049	24.18	0.841	0.022	0.139	20.07	0.694	0.048	0.259
RFRNet	30.36	0.937	0.008	0.073	23.42	0.807	0.027	0.199	19.21	0.638	0.060	0.340
JPGNet	31.65	0.952	0.007	0.074	24.72	0.851	0.022	0.202	20.46	0.713	0.048	0.342
LAMA	31.07	0.956	0.009	0.036	24.15	0.859	0.025	0.116	20.15	0.713	0.048	0.257
MISF	31.45	0.954	0.006	**0.032**	24.97	0.859	**0.020**	0.117	20.59	0.717	0.044	0.233
LIIF	30.96	0.946	0.010	0.038	24.57	0.846	0.026	0.120	19.79	0.708	0.049	0.274
SAIR	**31.01**	**0.964**	**0.005**	0.034	**26.44**	**0.866**	0.023	**0.110**	**21.88**	**0.722**	**0.042**	**0.193**

5.2 Comparison Results

The results obtained on the CelebAHQ dataset are presented in Table 2, demonstrating a significant performance improvement achieved by incorporating semantic information into the models. For instance, SAIR outperforms MISF by 1.74 in PSNR for the 0–20% mask ratio. Moreover, SAIR surpasses LAMA by 2.35 in PSNR and 1.2% in SSIM for 20–40% ratio. In the 20–40% mask ratio, SAIR exhibits enhancements of 2.69 in PSNR and 7.1% in SSIM compared to LIIF. The results on the ADE20K dataset, as shown in Table 3, also reveal the effectiveness of incorporating semantic information. SAIR achieves a lowered LPIPS of 0.193 for the 40–60% mask ratio. And SAIR improves PSNR to 26.44 and SSIM to 86.6% in 20–40% ratio range. Notably, SAIR attains the best PSNR and SSIM performance for all mask ratios. These results demonstrate that semantic information aids in processing degraded images. Our approach overcomes the limitations imposed by noise in masked area appearance features by leveraging the guidance of semantic information.

Qualitative results from different models are presented in Fig. 3, showcasing significant enhancements achieved by our proposed method. Fig. 3 unmistakably illustrates that the implicit neural function models lacking semantic guidance tend to produce blurry reconstructions in the affected regions, often displaying a noticeable boundary between masked and unmasked areas. In contrast, models enriched with semantic information yield more visually coherent and pleasing results. As observed in the first row, it becomes apparent that traditional implicit neural functions like LIIF struggle to recover the 'eye' category when it is entirely masked. In such cases, the neighboring appearance features can only provide information about the 'face.' However, SAIR demonstrates its ability to reconstruct the 'eye' category effectively, benefitting from the restored

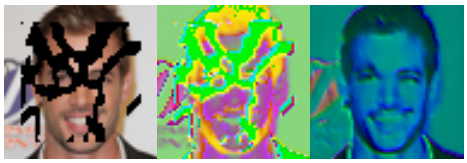

Fig. 4: From left to right, we show the input masked image, masked semantic feature after CLIP encoder, and semantic feature after SIR.

Fig. 3: Visual comparison with competitors: the first two cases are from the CelebAHQ dataset, while the last two are from the ADE20K dataset.

semantic features. Furthermore, in the last row of Fig. 3, the original implicit neural function generates unexpected regions prominently.

In Fig. 4, we present a visual representation of the image features before and after the application of our SIR module. The pre-trained CLIP encoder cannot handle the masked regions ideally. And it becomes evident that our proposed SIR module effectively reconstructs the corrupted image feature. To assess the impact of semantic information during the training process, we visually analyze the training progress of both LIIF and SAIR. The training loss curves depicted in Fig. 5 demonstrate that both models converge at a similar point. This observation suggests that the inclusion of semantic information can facilitate loss convergence without necessitating an extended training duration. Moreover, as seen in Fig. 5, the PSNR curve illustrates that the model enriched with semantic information consistently outperforms the original implicit representation model right from the outset.

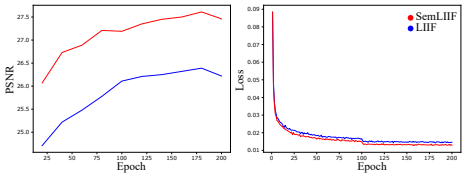

Fig. 5: PSNR vs Epoch and Training Loss vs Epoch.

5.3 Ablation Study

Study on using different image encoders. To demonstrate the compatibility of our semantic feature embedding with various image encoders, we conducted an ablation study in which we replaced our image encoder with the original LIIF encoder EDSR [23]. As indicated in Table 4, when compared to a model without the inclusion of semantic features (EDSR(wo)), the model that incorporates semantic features (EDSR(w)) also exhibited improvements, increasing the PSNR by 1.12 and the SSIM by 2.1%. These experiments provide compelling evidence that semantic information has the potential to enhance performance across different appearance feature spaces.

Table 4: Ablation study results on different image encoders and different implicit neural function models on CelebAHQ dataset.

Variant	All mask ratios PSNR↑	SSIM↑
EDSR(wo)	30.26	0.892
EDSR(w)	**31.48**	**0.913**
LTE	30.60	0.931
SemLTE	**31.97**	**0.939**

Study on using different implicit neural functions. In order to demonstrate the versatility of our semantic feature integration with various implicit neural functions, we conducted an ablation study using another implicit neural function known as LTE [15], which is specifically designed for image super-resolution tasks. In this study, we seamlessly incorporated semantic features into LTE, creating what we refer to as SemLTE. The resulting performance metrics are presented in Table 4, where SemLTE achieved significant improvements, elevating the PSNR to 31.97 and the SSIM to 93.9%. These outcomes affirm the adaptability of our proposed semantic implicit representation, showcasing its effectiveness when applied to different implicit neural functions.

Study on the models with/without SIR block. To further assess the effectiveness of SIR module, we conducted performance tests on the CLIP encoder, both with and without the SIR in the semantic segmentation task. We use masked images as inputs to generate the segmentation results, which are compared with ground truth. In the setting 'without SIR', we initially employed the CLIP text encoder to produce category features $\mathbf{CLIP\_T} \in \mathbb{R}^{\mathrm{L}\times\mathrm{C}}$ for all categories in the dataset, where L represents the number of categories. Subsequently, we used **CLIP_T** to filter the semantic feature **CLIP_I**, yielding a pixel-wise segmentation map $\mathbf{S} \in \mathbb{R}^{\mathrm{H}\times\mathrm{W}\times\mathrm{L}}$. In the setting 'with SIR', we use the SIR block to reconstruct the CLIP semantic feature **CLIP_I**, the reconstructed feature is used for segmentation. The results presented in Table 5 indicate that the inclusion of the SIR block leads to a notable increase in mIoU by 0.28, demonstrating the effective capacity of the SIR model to reconstruct semantic features.

Table 5: Semantic segmentation results from the models with/without SIR on ADE20K dataset.

Variant	mIoU
CLIP Encoder	0.17
CLIP Encoder+ SIR	**0.45**

Study on not filling the semantic feature (NFS). In the preceding section, we employed SIR to reconstruct the semantic feature of masked images. Here, we delve into an alternative scenario where we do not to fill in the masked semantic feature. In our experiments, we introduced masked semantic features into the implicit neural function alongside the appearance fea-

ture. However, as evident in the results presented in Table 6 under the label NFS, this approach yields suboptimal performance when compared to SAIR. Specifically, it leads to a noticeable decrease of 2.04 in PSNR and a 2.1% reduction in SSIM. The presence of meaningless semantic information within the masked region exerts an adverse influence on the construction of the implicit representation.

Table 6: Ablation study on *Not filling semantic feature* (NFS), *Only using semantic feature* (OUS), and SAM encoder on CelebAHQ dataset.

Variant	All mask ratios PSNR↑	SSIM↑
NFS	30.32	0.923
OUS	31.11	0.929
SAM Encoder	31.72	0.935
SAIR	**32.36**	**0.944**

Study on only using semantic feature to build implicit representation (OUS). In this section, we explore the possibility of constructing a continuous representation using only semantic features, meaning that we exclusively input semantic information into the implicit neural function. The results is shown in Table 6 as the OUS. It's worth noting that the CLIP image encoder is trained to produce features that align with textual information. In essence, this experiment underscores the significance of integrating both semantic and image-level information to attain favorable outcomes in image generation tasks.

Using other semantic embeddings. As an alternative to employing our Semantic Implicit Representation (SIR), we can also utilize existing models designed for semantic embeddings, such as the previously introduced semantic segmentation model SAM [13]. To demonstrate this, we replaced our CLIP encoder with the pre-trained SAM image encoder, and the results are presented in Table 6. Notably, it becomes evident that the CLIP encoder outperforms traditional semantic segmentation encoders in this context. This superiority is attributed to the CLIP encoder's capacity to capture rich textual information, further enhancing the inpainting task's performance.

Study on using SAIR for image superresolution. Our approach extends beyond its primary application and is equally effective for image super-resolution tasks (mask ratio is 0). SIR block can significantly enhance the resolution of semantic maps to various higher levels. In comparison to existing method LIIF, our proposed technique demonstrates superior super-resolution capabilities, as evidenced by the PSNR/SSIM results presented in Table 8 across different upsample ratios on CelebAHQ dataset. All models are trained under $\times$ 4 setting. SAIR outperforms LIIF across all mask ratios, leveraging the wealth of semantic information embedded within the images.

Importance of implicit neural function (INF). We design three variants of SAIR to validate the necessity of leveraging INR: ❶ We remove the INR from SIR by replacing Equation 4 with a bilinear operation, *i.e.*, $\mathbf{z}_{\mathbf{p}}^{\text{sem}} = \text{SIR}_{\text{woINR}}(\mathbf{I}, \mathbf{q}) = \sum_{\mathbf{q}\in\mathcal{N}_{\mathbf{p}}} \text{W}_{\mathbf{q}} \mathbf{z}_{\mathbf{q}}^{\text{sem}}$, where $\{\text{W}_{\mathbf{q}}\}$ is the set of bilinear weights. ❷ We remove the INR from AIR by replacing Equation 5 with a convolutional layer and a bilinear operation, *i.e.*, $c_{\mathbf{p}} = \text{AIR}_{\text{woINR}}(\mathbf{I}, \mathbf{M}, \text{SIR}, \mathbf{p}) = \sum_{\mathbf{q}\in\mathcal{N}_{\mathbf{p}}} \text{W}_{\mathbf{q}} f_{\text{cnn}}([\mathbf{z}_{\mathbf{q}}^{\text{app}}, \text{SIR}(\mathbf{I}, \mathbf{M}, \mathbf{q})])$. Note that, SIR remains the same.

Table 8: Ablation study results on using SAIR for image superresolution.

PSNR ↑ /SSIM ↑	Upscale ratio			
	× 2	× 4	× 6	× 8
LIIF	34.32/0.963	30.98/0.882	30.13/0.838	29.83/0.819
SAIR	**34.61**/**0.972**	**31.24**/0.881	**30.39**/**0.840**	**30.03**/**0.825**

Table 7: Ablation study results on INF.

Variants	SIR w. INR	AIR w. INR	PSNR↑	SSIM↑
❶		✔	27.29	0.912
❷	✔		29.77	0.921
❸			26.09	0.893
SAIR	✔	✔	**30.97**	**0.962**

❸ We remove INR from both SIR and AIR by performing the above replacements at the same time. The experiments are conducted on CelebAHQ dataset. As shown in Table 7, once we remove INR from SIR or AIR, the performance decreases, demonstrating its superior capability compared to alternatives such as bilinear operations and CNN layers.

5.4 Discussion

Comparison with diffusion based inpainting methods. The diffusion model is a popular topic and has been widely used in image inpainting tasks. However, our proposed method surpasses diffusion-based approaches in three key aspects. First, inference speed. The inference speed of our method (0.043s/image) is faster than that of the diffusion-based method, such as stable diffusion [35] (12s/image). Second, arbitrary resolution. Our method can generate images at arbitrary resolutions during inference, a capability lacking in diffusion-based models. Third, high fidelity. Diffusion-based models prioritize naturalness over fidelity, resulting in generated results that deviate significantly from the ground truth. We compare our method with stable diffusion on the CelebAHQ dataset under the same setting, the results are PSNR ↑ 37.97 (ours) vs. 37.60 (stable diffusion) and L1 ↓ 0.01 (ours) vs. 0.032 (stable diffusion). The diffusion model tends to generate results with low fidelity.

6 Future Work

In this study, we have demonstrated the effectiveness of the Semantic-Aware Implicit Representation (SAIR) in the domain of image inpainting. While our proposed method has shown remarkable performance in this specific task, its broader applicability to other vision-related tasks has yet to be fully explored. As part of our future research endeavors, we plan to conduct additional experiments to assess the potential of our method in addressing various vision tasks beyond inpainting, such as segmentation and denoising.

7 Conclusion

In this paper, we tackle the limitations inherent in existing implicit representation techniques, which predominantly rely on appearance information and often falter when faced with severely degraded images. To address this challenge, we introduce a novel approach: the semantic-aware implicit representation (SAIR). By seamlessly using a semantic implicit representation (SIR) to handle the pixel-level semantic feature and a appearance implicit representation (AIR) to reconstruct the image color, our method effectively mitigates the impact of potentially degraded regions. To gauge the effectiveness of our approach, we conducted comprehensive experiments on two widely recognized datasets, CelebAHQ [26] and ADE20K [50]. The results unequivocally demonstrate that our method outperforms existing implicit representation and inpainting approaches by a substantial margin across four commonly employed image quality evaluation metrics. Our model's capacity to assist the implicit neural function in processing damaged images expands its utility and applicability, offering promising prospects for various image-related tasks.

Acknowledgments. This research is supported by the National Research Foundation, Singapore, and DSO National Laboratories under the AI Singapore Programme (AISG Award No: AISG2-GC-2023-008), Career Development Fund (CDF) of Agency for Science, Technology and Research (No.: C233312028), and National Research Foundation, Singapore and Infocomm Media Development Authority under its Trust Tech Funding Initiative (No. DTC-RGC-04).

References

1. Bar, A., Gandelsman, Y., Darrell, T., Globerson, A., Efros, A.: Visual prompting via image inpainting. NeurIPS **35**, 25005–25017 (2022)
2. Cao, Y., Li, T., Cao, X., Tsang, I., Liu, Y., Guo, Q.: Irad: Implicit representation-driven image resampling against adversarial attacks. In: ICLR (2024)
3. Chen, H., He, B., Wang, H., Ren, Y., Lim, S.N., Shrivastava, A.: Nerv: Neural representations for videos. In: Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., Vaughan, J.W. (eds.) NeurIPS. vol. 34, pp. 21557–21568. Curran Associates, Inc. (2021)
4. Chen, J., Ren, X., Guo, Q., Juefei-Xu, F., Lin, D., Feng, W., Ma, L., Zhao, J.: Lrr: Language-driven resamplable continuous representation against adversarial tracking attacks. In: ICLR (2024)
5. Chen, Y., Liu, S., Wang, X.: Learning continuous image representation with local implicit image function. In: CVPR. pp. 8628–8638 (2021)
6. Chen, Z., Zhang, H.: Learning implicit fields for generative shape modeling. In: CVPR. pp. 5939–5948 (2019)
7. Feng, T., Feng, W., Li, W., Lin, D.: Cross-image context for single image inpainting. NeurIPS **35**, 1474–1487 (2022)
8. Grattarola, D., Vandergheynst, P.: Generalised implicit neural representations. arXiv preprint arXiv:2205.15674 (2022)
9. Guo, Q., Li, X., Juefei-Xu, F., Yu, H., Liu, Y., Wang, S.: Jpgnet: Joint predictive filtering and generative network for image inpainting. In: ACM International Multimedia. pp. 386–394 (2021)

10. Guo, Z., Lan, C., Zhang, Z., Chen, Z., Lu, Y.: Versatile neural processes for learning implicit neural representations. arXiv preprint arXiv:2301.08883 (2023)
11. Ho, C.H., Vasconcelos, N.: Disco: Adversarial defense with local implicit functions. arXiv preprint arXiv:2212.05630 (2022)
12. Hsu, J., Gu, J., Wu, G., Chiu, W., Yeung, S.: Capturing implicit hierarchical structure in 3d biomedical images with self-supervised hyperbolic representations. NeurIPS **34**, 5112–5123 (2021)
13. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollár, P., Girshick, R.: Segment anything. arXiv:2304.02643 (2023)
14. Lee, C.H., Liu, Z., Wu, L., Luo, P.: Maskgan: Towards diverse and interactive facial image manipulation. In: CVPR (2020)
15. Lee, J., Jin, K.H.: Local texture estimator for implicit representation function. In: CVPR. pp. 1929–1938 (June 2022)
16. Li, J., Wang, N., Zhang, L., Du, B., Tao, D.: Recurrent feature reasoning for image inpainting. In: CVPR. pp. 7760–7768 (2020)
17. Li, W., Lin, Z., Zhou, K., Qi, L., Wang, Y., Jia, J.: Mat: Mask-aware transformer for large hole image inpainting. In: CVPR. pp. 10758–10768 (2022)
18. Li, X., Guo, Q., Abdelfattah, R., Lin, D., Feng, W., Tsang, I., Wang, S.: Leveraging inpainting for single-image shadow removal. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13055–13064 (2023)
19. Li, X., Guo, Q., Cai, P., Feng, W., Tsang, I., Wang, S.: Learning restoration is not enough: Transfering identical mapping for single-image shadow removal. arXiv preprint arXiv:2305.10640 (2023)
20. Li, X., Guo, Q., Lin, D., Li, P., Feng, W., Wang, S.: Misf: Multi-level interactive siamese filtering for high-fidelity image inpainting. In: CVPR. pp. 1869–1878 (2022)
21. Li, Z., Min, M.R., Li, K., Xu, C.: Stylet2i: Toward compositional and high-fidelity text-to-image synthesis. In: CVPR. pp. 18197–18207 (2022)
22. Liao, L., Xiao, J., Wang, Z., Lin, C.W., Satoh, S.: Uncertainty-aware semantic guidance and estimation for image inpainting. IEEE Journal of Selected Topics in Signal Processing **15**(2), 310–323 (2020)
23. Lim, B., Son, S., Kim, H., Nah, S., Mu Lee, K.: Enhanced deep residual networks for single image super-resolution. In: CVPR Workshop. pp. 136–144 (2017)
24. Lin, Y., Xie, Y., Chen, D., Xu, Y., Zhu, C., Yuan, L.: Revive: Regional visual representation matters in knowledge-based visual question answering. arXiv preprint arXiv:2206.01201 (2022)
25. Liu, G., Reda, F.A., Shih, K.J., Wang, T.C., Tao, A., Catanzaro, B.: Image inpainting for irregular holes using partial convolutions. In: ECCV. pp. 85–100 (2018)
26. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: ICCV (December 2015)
27. Lu, Y., Zhou, J., McDorman, S.T., Zhang, C., Scott, D., Bukuts, J., Wilder, C., Smith, K.Y., Wang, S.: Snowvision: Segmenting, identifying, and discovering stamped curve patterns from fragments of pottery. International Journal of Computer Vision **130**(11), 2707–2732 (2022)
28. Lüddecke, T., Ecker, A.: Image segmentation using text and image prompts. In: CVPR. pp. 7086–7096 (2022)
29. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM **65**(1), 99–106 (2021)

30. Nazeri, K., Ng, E., Joseph, T., Qureshi, F., Ebrahimi, M.: Edgeconnect: Structure guided image inpainting using edge prediction. In: ICCV Workshops. pp. 0–0 (2019)
31. Ni, M., Li, X., Zuo, W.: Nuwa-lip: Language-guided image inpainting with defect-free vqgan. In: CVPR. pp. 14183–14192 (2023)
32. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
33. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 **1**(2), 3 (2022)
34. Ren, Y., Yu, X., Zhang, R., Li, T.H., Liu, S., Li, G.: Structureflow: Image inpainting via structure-aware appearance flow. In: CVPR. pp. 181–190 (2019)
35. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models (2021)
36. Su, K., Chen, M., Shlizerman, E.: Inras: Implicit neural representation for audio scenes. NeurIPS **35**, 8144–8158 (2022)
37. Suvorov, R., Logacheva, E., Mashikhin, A., Remizova, A., Ashukha, A., Silvestrov, A., Kong, N., Goka, H., Park, K., Lempitsky, V.: Resolution-robust large mask inpainting with fourier convolutions. WACV (2022)
38. Tao, M., Tang, H., Wu, F., Jing, X.Y., Bao, B.K., Xu, C.: Df-gan: A simple and effective baseline for text-to-image synthesis. In: CVPR. pp. 16515–16525 (2022)
39. Wang, Y., Tao, X., Qi, X., Shen, X., Jia, J.: Image inpainting via generative multi-column convolutional neural networks. NeurIPS **31** (2018)
40. Xie, Z., Zhang, J., Li, W., Zhang, F., Zhang, L.: S-nerf: Neural radiance fields for street views. arXiv preprint arXiv:2303.00749 (2023)
41. Xu, J., De Mello, S., Liu, S., Byeon, W., Breuel, T., Kautz, J., Wang, X.: Groupvit: Semantic segmentation emerges from text supervision. In: CVPR. pp. 18134–18144 (2022)
42. Yang, Z., Lu, Y., Wang, J., Yin, X., Florencio, D., Wang, L., Zhang, C., Zhang, L., Luo, J.: Tap: Text-aware pre-training for text-vqa and text-caption. In: CVPR. pp. 8751–8761 (2021)
43. Yariv, L., Gu, J., Kasten, Y., Lipman, Y.: Volume rendering of neural implicit surfaces. NeurIPS **34**, 4805–4815 (2021)
44. Yin, F., Liu, W., Huang, Z., Cheng, P., Chen, T., YU, G.: Coordinates are not lonely–codebook prior helps implicit neural 3d representations. arXiv preprint arXiv:2210.11170 (2022)
45. Zhang, C., Guo, Q., Li, X., Wan, R., Yu, H., Tsang, I., Wang, S.: Superinpaint: Learning detail-enhanced attentional implicit representation for super-resolutional image inpainting. arXiv preprint arXiv:2307.14489 (2023)
46. Zhang, L., Chen, Q., Hu, B., Jiang, S.: Text-guided neural image inpainting. In: ACM Multimedia. p. 1302–1310 (2020)
47. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR (2018)
48. Zhao, M., Li, B., Wang, J., Li, W., Zhou, W., Zhang, L., Xuyang, S., Yu, Z., Yu, X., Li, G., et al.: Towards video text visual question answering: Benchmark and baseline. In: Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (2022)
49. Zhenxing, M., Xu, D.: Switch-nerf: Learning scene decomposition with mixture of experts for large-scale neural radiance fields. In: ICLR (2022)

50. Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Scene parsing through ade20k dataset. In: CVPR. pp. 633–641 (2017)
51. Zhou, C., Loy, C.C., Dai, B.: Extract free dense labels from clip. In: ECCV. pp. 696–712. Springer (2022)
52. Zhou, Z., Lei, Y., Zhang, B., Liu, L., Liu, Y.: Zegclip: Towards adapting clip for zero-shot semantic segmentation. In: CVPR. pp. 11175–11185 (2023)
53. Zhu, Y., Liu, H., Song, Y., Yuan, Z., Han, X., Yuan, C., Chen, Q., Wang, J.: One model to edit them all: Free-form text-driven image manipulation with semantic modulations. NeurIPS **35**, 25146–25159 (2022)