Amodal Content Completion via Gated Convolution and Latent Diffusion Features

Kaziwa Saleh[a,b], Sándor Szénási[b,c], Zoltán Vámossy[b]

[a] Doctoral School of Applied Informatics and Applied Mathematics, Obuda University, Budapest, Hungary
[b] John von Neumann Faculty of Informatics, Obuda University, Budapest, Hungary
[c] Faculty of Economics and Informatics, J. Selye University, Komarno, Slovakia
kaziwa.saleh@nik.uni-obuda.hu, szenasi.sandor@nik.uni-obuda.hu, vamossy.zoltan@nik.uni-obuda.hu

Abstract: *Amodal content completion is the reconstruction of the invisible region(s) of occluded objects by inferring their missing appearance, style, and texture from the available visual context. Despite recent advances in image completion, existing methods often struggle to accurately recover the content of the occluded areas and fail to generate contextually plausible completions or correctly capture long-range dependencies. To address these limitations, this paper proposes a novel amodal content completion network, that integrates semantic priors from a pre-trained diffusion model with local structural representations extracted by a gated convolutional encoder. A hierarchical transformer applies self-attention on the multi-resolution features to guide the decoder to reconstruct the occluded content. Experimental results on the COCOA dataset and two of its subsets demonstrate the effectiveness of the proposed approach in generating visually plausible and structurally coherent completion compared with the baseline methods.*

Keywords: amodal completion; occlusion handling; de-occlusion; image completion; RGBA completion; generative AI; computer vision

1 Introduction

The ability to infer information that is not directly visible in an image is an essential aspect of visual understanding. In many real-world scenes, objects are partially covered by other objects, causing only a part of their visual information to be seen

[1]. Recovering the gestalt and content of the hidden regions of such objects is referred to as amodal completion. By estimating information that is not directly visible, it provides a more complete representation of the scene and can support downstream tasks that require knowledge of object shape and appearance. One aspect of amodal completion is amodal content completion (or amodal appearance completion), which focuses on reconstructing the appearance of the occluded regions [2].

The field of image generation and completion has progressed considerably in recent years. Early approaches mainly relied on convolutional architectures and generative adversarial networks, while recent approaches mostly utilize diffusion models due to their ability to synthesize high-fidelity and visually realistic images [3]. Their ability to learn visual patterns from training on large-scale data has made models useful for a wide range of image reconstruction and editing tasks.

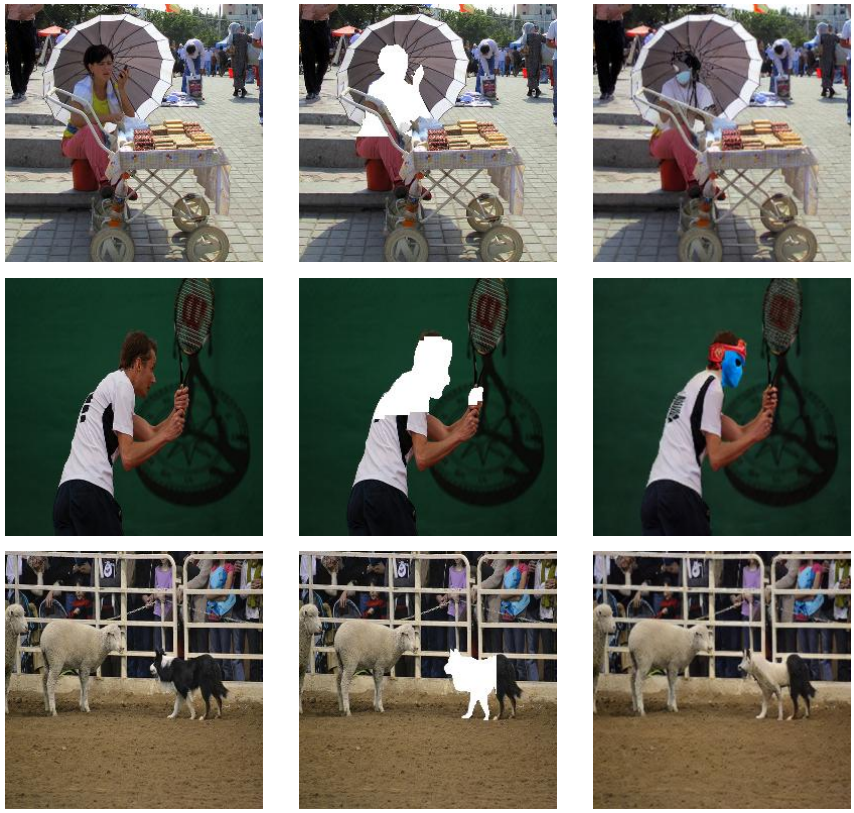
Despite these developments, recovering the appearance of the hidden region(s) of partially visible objects remains challenging. Unlike conventional image inpainting, where the missing regions may correspond to arbitrary areas of an image, the invisible region of occluded objects is constrained by the surrounding scene and by the visible part of the object itself [4]. As a result, the completion process requires the model to generate content that is visually plausible and semantically correct, i.e., it is consistent with the visual information present in the image and the object.

Several studies [5], [6], [7], [8] have explored the use of pre-trained vision language models (VLMs) such as CLIP [9], Stable Diffusion (SD) [3], GPT-4o [10], and Segment Anything [11], for amodal completion. These models provide rich representations learned from a large amount of visual data and provide useful contextual information that can complement the limited information available in amodal datasets. Nevertheless, directly utilizing Stable Diffusion for amodal completion does not always produce satisfactory results, its limited ability to reason about occlusion may lead to reconstructions that are inconsistent with the visible image cues [12], as can be seen in the samples of Figure 1.

In this paper, motivated by the work in [13], we utilize the representation power of a pre-trained Stable Diffusion model indirectly by using it as a feature extractor. We propose an amodal content completion network that combines local image representations with high-level features extracted from a pre-trained Stable Diffusion model. The proposed architecture consists of four main components: a gated convolutional encoder, a frozen Stable Diffusion feature extraction branch, a hierarchical transformer module, and a decoder with skip connections. The gated convolutional encoder captures local information from the visible regions of the image, while the frozen Stable Diffusion branch provides semantic representations learned from large-scale visual data. These features are subsequently refined and integrated by a hierarchical transformer that facilitates information exchange across multiple resolutions through self-attention. Finally, the decoder reconstructs the missing regions by jointly utilizing the refined features from the transformer module

and the local spatial details provided via skip connections from the encoder. This way, the proposed architecture aims to generate reconstructions that remain visually consistent with the partially visible object and its surroundings.

We implement a self-supervised training strategy [14] where synthetic occlusions are created by overlaying randomly selected objects onto other objects in the dataset. The modal mask of the occluded object and the mask corresponding to the synthetically occluded region are combined to form the target completion region. The proposed model is then trained to generate the visual content within this region using the available image context.



(a) Ground truth (b) Masked image (c) SD output

Figure 1

Samples of reconstructed images with Stable Diffusion v1.5.

We evaluate the proposed method on the COCOA [15] dataset and two of its subsets. The experimental results demonstrate that combining local image representations with semantic features extracted from a pre-trained diffusion model

and contextual information provided by the hierarchical transformer produces visually coherent and consistent completions. We also conduct additional experiments to investigate the effect of using the weighted mask [16] instead of the binary mask as input to the Stable Diffusion model.

In summary, the contributions of our work are:

1. We propose an amodal content completion network that combines a gated convolutional encoder, a frozen Stable Diffusion feature extraction branch, a hierarchical transformer, and a decoder to reconstruct the occluded area of partially visible objects.
2. We demonstrate that combining semantic features from a pre-trained diffusion model with local image features produces visually coherent reconstructions and achieves superior performance compared with baseline methods.

The remainder of this paper is organized as follows. Section 2 reviews existing works related to amodal appearance completion methods based on convolutional and generative models. Section 3 presents the proposed architecture, the training losses, and the datasets used for training and evaluation. Section 4 reports and discusses the quantitative and qualitative experimental results. Finally, Section 5 provides the conclusions to our work.

2 Related works

Several studies have investigated the problem of amodal content completion. A number of methods developed dedicated architectures trained specifically to reconstruct the occluded regions of partially visible objects. Ehsani et al. introduced SeGAN [17], which is trained on a synthetic dataset of photo-realistic images to estimate the appearance of occluded regions using a conditional generative adversarial network. To reduce the dependence on manually annotated amodal data, PCNets [14] introduced a self-supervised learning framework called partial completion. In this approach, synthetic occlusions are generated by randomly overlaying objects, and the resulting visible mask of the occluder and the occludee are used alternately to train the network without requiring explicit amodal annotations. The framework operates progressively by recovering the occlusion ordering between objects and performing amodal segmentation. The predicted amodal masks are then utilized by PCNet-C to reconstruct the RGB appearance of the hidden regions of partially occluded objects. Furthermore, Saleh et al. [16] proposed a network based on gated convolutional layers [18], which reduces the influence of missing regions during the completion process. To place greater emphasis on the visible portions of partially occluded objects, a weighted mask is provided as an additional input alongside the image. To further reinforce the effect

of the weighted mask throughout the encoder, the gated convolutional layers are replaced with an enhanced variant that explicitly incorporates the weighted mask information at each layer. This prevents the influence of the mask to diminish as the layers deepen.

More recently, pre-trained vision models have also been explored for amodal completion. Ozguroglu et al. [5] adapt large conditional diffusion models using a synthetic dataset containing various occlusion scenarios to address amodal segmentation, object recognition, and 3D reconstruction simultaneously. Similarly, Ao et al. [7] employ a pre-trained diffusion model to develop a training-free framework for open-world content completion, where textual prompts guide the reconstruction process. Fan et al. [8] combine multiple agents that perform occlusion reasoning, mask refinement, and language-based description generation to guide image completion using information extracted from visible masks and attention maps. Likewise, Li et al. [19] propose a diffusion-based framework that first synthesizes a large-scale amodal dataset through self-supervised learning and human-guided refinement and then trains a text-conditioned diffusion model to reconstruct the complete appearance of partially occluded objects. Another line of research focuses on exploiting latent representations learned by diffusion models. Liu et al. [20] propose a self-supervised framework that learns to transform latent representations of partially visible objects into their corresponding complete representations. During inference, the generated latent representation is decoded to recover the complete appearance of the target object.

In contrast to the previously mentioned works, we do not employ the diffusion model directly for the completion task. Instead, we utilize the representations learned by the pre-trained model as semantic features and integrate them with a hierarchical transformer that applies self-attention to refine these features and guide the decoder during image reconstruction.

3 Methodology

3.1 Gated Convolution

Standard convolution operations designed for complete images cannot be directly applied when parts of the image are missing. To handle regions with invalid pixels, such as occluded areas, either hard-gating approaches like partial convolution [21] or soft-gating approaches like gated convolution [18] are used to restrict feature extraction to valid pixels only. Gated convolution achieves this by concatenating the input image with a mask of the invalid region and learning a dynamic gating mechanism from both.

In amodal content completion, the goal is to reconstruct a specific object defined by its amodal mask. Pixels within the occluded region are treated as invalid and excluded from computation, while only the visible pixels outside the occluded patch contribute to feature extraction. Therefore, to limit the effect of the missing regions during feature extraction and decoding the features, gated convolutional layers are employed in both the encoder and decoder of the proposed model.

3.2 Model Architecture

The proposed network is designed for amodal completion, which aims to reconstruct the occluded or missing image regions by jointly utilizing local structural cues and semantic priors learned from large-scale diffusion models. The architecture includes four main components: a gated convolutional encoder, a frozen Stable Diffusion feature extractor, a hierarchical transformer fusion module, and a decoder with multi-scale skip connections (Figure 2).

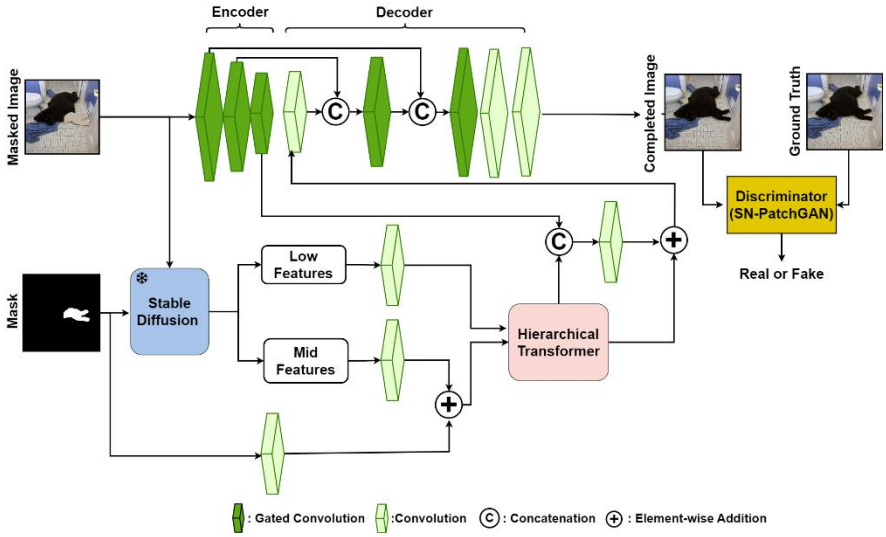


Figure 2

Detailed architecture of the proposed amodal content completion network consisting of a gated convolutional encoder, a frozen Stable Diffusion feature extractor, a hierarchical transformer module, and a multi-scale decoder.

3.2.1 Gated Convolutional Encoder

The masked input image is first processed by a main encoder composed of three gated convolutional layers. Unlike conventional convolutions, gated convolutions learn dynamic feature selection masks that modulate the contribution of each spatial location and channel, enabling the network to distinguish valid image regions from missing areas. The encoder progressively downsamples the input and produces multi-scale feature maps that capture local structures and object layouts. These feature representations are then reused by the decoder through skip connections to preserve fine spatial details.

3.2.2 Stable Diffusion Feature Extractor

In parallel with the convolutional encoder, a frozen Stable Diffusion inpainting model extracts semantic guidance from the masked image. The feature extractor consists of a pre-trained Variational Autoencoder (VAE) [22], a diffusion scheduler based on the denoising diffusion process [23], and a U-Net [24] whose parameters are fixed during training. The masked image is encoded into latent representations, which are then perturbed according to the diffusion process and processed by the U-Net together with the corresponding binary mask. Although the U-Net normally receives text embeddings as part of its input interface, empty embeddings are provided during training and inference, therefore no explicit textual conditioning is employed. Intermediate feature maps are extracted from the first two U-Net upsampling blocks, providing high-level semantic representations that capture object appearance, textures, and contextual relationships learned from large-scale image data.

3.2.3 Hierarchical Transformer Fusion

The extracted Stable Diffusion features are projected into a common embedding space and fused with an embedded binary mask representation. Similar to hierarchical vison Transformers [25], which process visual features at multiple spatial resolutions, the hierarchical transformer module refines the representations at low and mid resolutions. Unlike a single-scale Transformer, the module allows the two feature levels to exchange information by feature downsampling, upsampling, and residual feature addition. The multi-head self-attention at each resolution allows the network to learn long-range dependencies and contextual relationships over the entire scene. Then the resulting semantic guidance is combined with the deepest encoder features through the feature concatenation followed by residual refinement, such that the local structural information and global semantic priors complement each other.

3.2.4 Multi-Scale Decoder

The fused representation is progressively upsampled to reconstruct the completed image. The decoder employs skip connections from the encoder to recover high-frequency spatial information lost during downsampling. The upsampled representations are combined with the encoder features through gated convolutional layers in the intermediate decoding stages. This allows a smooth merging of preserved background details and hallucinated object content. Finally, conventional convolutional layers map the decoded features to the RGB image space to produce the reconstructed output image.

Finally, the output image is passed to a spectral normalized PatchGAN (SN-PatchGAN) [18] discriminator, which evaluates its realism against the corresponding ground-truth image during adversarial training. The discriminator operates at the patch level and produces a spatial score map where each value represents the realism of the corresponding image patch. This encourages the model to produce visually consistent textures while maintaining structural coherence.

3.3 Dataset

One of the main issues when training models to complete amodal content is the lack of labeled datasets that contain ground-truth information for occluded regions. This is because the task requires the model to predict what occluded parts of objects look like, meaning that it needs RGB data for regions that are not visible, and such data is normally not available.

Therefore, the proposed method uses the self-supervised method from Zhan et al. [14]. The method generates synthetic occlusion by placing an eraser mask over an object instance mask and shifting it relative to the object. Instead of choosing the horizontal and vertical shifts independently, the position of the eraser is controlled by a randomly selected overlap ratio. Assuming that the object and the eraser have the same dimensions, the overlap between them can be approximated as:

$$o = (1 - |\mathrm{off}_x|)\left(1 - \left|\mathrm{off}_y\right|\right)$$

where o is the desired overlap ratio, and off_x and off_y are the normalized horizontal and vertical shifts, respectively. The algorithm first selects a horizontal shift and then computes the corresponding vertical shift required to achieve the desired overlap. This allows different occlusion patterns to be generated while keeping the amount of overlap approximately the same.

After the eraser is positioned, the amount of the object that becomes hidden is measured by computing the intersection between the object mask M and the shifted eraser mask E:

$$r = \frac{|M \cap E|}{|M|}$$

where r is the occlusion ratio, $M \cap E$ represents the pixels shared by both masks, and $|M|$ is the number of pixels belonging to the target object. The placement process is repeated until the measured occlusion ratio falls within a predefined range. The occluded area is kept below 50% of the size of the object. Therefore, the resulting eraser mask covers a controlled portion of the object, producing synthetic partially occluded instances that can be used for training and evaluation.

We train the model on the COCOA [15] dataset. There are 22,163 object instances in the training set and 12,753 instances in the validation set.

We construct two additional subsets from COCOA, focusing only on animals. The first one, COCOA-animal-M, was constructed by manually checking each image in the dataset. An image was kept only if it had at least one animal that was clearly visible. Images were removed if they had no animals, if the animals present were not clearly visible, or if the animals were already heavily obstructed. To enlarge this set, we applied common data augmentation methods, such as flipping, rotation, cropping and shifting. This resulted in a total of 5,264 training examples.

The second subset, COCOA-animal-cls, was generated using a simpler method of category labels at the time of data loading. All other categories were discarded, and only annotations with animal labels (e.g., dog, cat or zebra) were retained. Here, unlike the first subset, there was no manual checking. This method resulted in a larger training set of 13,640 instances.

The augmentation was not applied for the validation data, so the validation sets for these two subsets are smaller, with only 342 and 529 instances respectively.

3.4 Loss Functions

The proposed model is optimized by a combination of hole [26], valid [21], Learned Perceptual Image Patch Similarity (LPIPS) [27], style [28], and adversarial losses [29] to ensure pixel-level accuracy and visual realism. The total training objective is defined as:

$$\mathcal{L}_{total} = \lambda_h \mathcal{L}_{hole} + \lambda_v \mathcal{L}_{valid} + \lambda_l \mathcal{L}_{LPIPS} + \lambda_s \mathcal{L}_{style} + \lambda_g \mathcal{L}_{GAN}$$

where λ_h, λ_v, λ_l, λ_s, λ_g are weighting coefficients to balance the contribution of each loss term. Empirically, these were set to 5.0, 10.0, 1.0, 5.0, and 1.0, respectively.

The hole loss is calculated only on the missing region, and it encourages the model to properly reconstruct the occluded content:

$$\mathcal{L}_{hole} = \frac{1}{N} \left\| \left(I_{gt} - I_{pred}\right) \odot (1 - M) \right\|_1$$

where I_{gt} and I_{pred} are the ground-truth and predicted images, respectively, M is the binary mask, $\odot$ is the element-wise multiplication, and N is the total number of pixels in the image.

The valid loss is computed as the mean absolute error between the reconstructed image and the ground-truth image over the visible regions indicated by the binary mask. This loss penalizes modifications in visible areas, encouraging the network to retain the original content of the image, while concentrating the completion process on the missing regions.

$$\mathcal{L}_{valid} = \frac{1}{N} \left\| (I_{gt} - I_{pred}) \odot M \right\|_1$$

Reconstruction losses encourage pixel-wise similarity but do not explicitly encourage high-level semantic consistency. Therefore, the LPIPS loss computes the perceptual distance between the reconstructed and ground-truth images using deep feature representations extracted from a pre-trained VGG [30] network:

$$\mathcal{L}_{LPIPS} = \sum_l \frac{1}{H_l W_l} \left\| w_l \odot \left(\phi_l(I_{gt}) - \phi_l(I_{pred}) \right) \right\|_2^2$$

where $\phi_l(\cdot)$ denotes the feature representation extracted from the l-th layer of a pre-trained network, w_l represents learned channel-wise weights, H_l and W_l are the spatial dimensions of the feature map at layer l. By comparing images in the feature space rather than at the pixel level, LPIPS encourages the generation of perceptually realistic structures, textures, and semantic content.

To further improve the texture reconstruction and enforce realistic local patterns, the style loss is utilized by comparing the Gram matrices of the extracted feature maps:

$$\mathcal{L}_{style} = \sum_i \left\| G\left(\varphi_i(I_{gt})\right) - G\left(\varphi_i(I_{pred})\right) \right\|_1$$

where $\varphi_i(\cdot)$ is the feature map extracted from the i-th VGG layer, and $G(.)$ is the Gram matrix, encoding the correlations between feature channels and therefore the texture statistics of the image.

Finally, the adversarial loss promotes the generation of visually plausible completed regions. The generator objective is defined by the hinge loss as:

$$\mathcal{L}_{GAN} = -\mathbb{E}\left[D\left(I_{pred}\right)\right]$$

where $D(\cdot)$ is the discriminator network.
Through combining all loss terms using empirically determined weighting coefficients, the model balances pixel-level accuracy with perceptual and textural fidelity.

4 Results and Discussion

Experiments are conducted on the COCOA dataset and its subsets using images of resolution 256×256 pixels. The proposed network contains 6.55 million trainable parameters and is optimized using Adam [31] with β1 = 0.5 and β2 = 0.999. A constant learning rate of 2×10^{-4} is employed for both the generator and discriminator. The network weights are initialized according to the Kaiming method [32] with a gain factor of 0.02. Training is performed for 100 epochs using a batch size of 8 on an NVIDIA GRID A100-20C GPU with 16 GB of memory.

The proposed model is quantitatively evaluated against the baseline architectures, DeepFill [18] and PCNet-C [14], using pixel-level errors, mean $\mathcal{L}_1$ and $\mathcal{L}_2$ errors, and fidelity indices, namely Peak Signal-to-Noise Ratio (PSNR), and Structural Similarity Index (SSIM) metrics. To assess the stability of the results, each experiment was performed five times. The reported results correspond to the mean computed across all runs. The proposed model is also compared against mask gated frameworks [16] [33], and the pre-trained Stable Diffusion model (Table 1, 2, 3).

Table 1

Quantitative comparison of the proposed method against baseline approaches on the validation set of COCOA-animal-M. Results are reported as mean across five runs.

Description	**$\mathcal{L}_1$error ↓**	**$\mathcal{L}_2$ error ↓**	**PSNR ↑**	**SSIM ↑**
PCNet-C	0.0986	0.0148	19.5851	0.7893
DeepFill	0.0624	0.0082	21.0661	0.8365
Mask gated completion	0.0265	**0.0020**	27.4376	0.9381
Enhanced mask gated completion	**0.0243**	**0.0020**	27.8587	0.9444
Pre-trained SD	0.0616	0.0103	30.2942	0.6658
Proposed model	0.0271	0.0033	**35.4633**	**0.9695**

Table 2

Performance comparison of the proposed method and baseline methods on the COCOA-animal-cls validation set.

Description	**$\mathcal{L}_1$error ↓**	**$\mathcal{L}_2$ error ↓**	**PSNR ↑**	**SSIM ↑**
PCNet-C	0.0855	0.0121	20.8871	0.7965
DeepFill	0.0431	0.0036	24.6027	0.8722
Mask gated completion	0.0341	0.0029	25.6855	0.8956
Enhanced mask gated completion	0.0227	**0.0016**	28.7437	0.9397
Pre-trained SD	0.0536	0.0083	31.4068	0.6898
Proposed model	**0.0192**	0.0025	**36.9390**	**0.9775**

Table 3

Comparative quantitative analysis of the proposed method and baseline approaches on the COCOA validation set.

Description	$\mathcal{L}_1$**error** ↓	$\mathcal{L}_2$ **error** ↓	**PSNR** ↑	**SSIM** ↑
PCNet-C	0.0786	0.0106	21.8377	0.8216
DeepFill	0.0346	0.0026	26.0692	0.8825
Mask gated completion	0.0400	0.0078	22.6790	0.8817
Enhanced mask gated completion	0.0186	**0.0014**	30.0610	0.9461
Pre-trained SD	0.0386	0.0049	33.8090	0.7769
Proposed model	**0.0121**	0.0016	**36.1740**	**0.9854**

The empirical results demonstrate consistent performance improvement by the proposed model. It achieves a noticeable improvement in both PSNR and SSIM metrics across all three datasets, as shown in Figure 3 and 4. In COCOA-animal-M dataset, the model attains a PSNR of 35.4633 dB and an SSIM of 0.9695, outperforming convolutional methods such as DeepFill (21.0661 dB) and Enhanced mask gated completion (27.8587 dB). The same pattern is observed in COCOA-animal-cls, with the proposed method obtaining the best reconstruction quality with a PSNR of 36.939 dB and an SSIM of 0.9775. On COCOA, the model achieves its highest fidelity scores, proving its generalizing capability across different object classes.

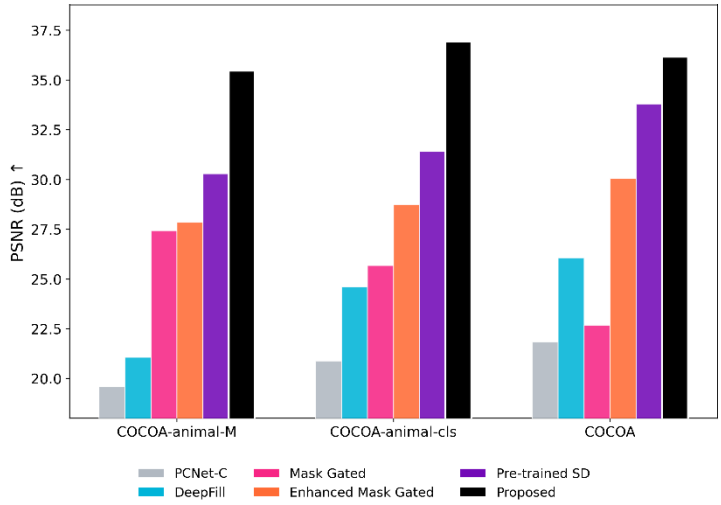


Figure 3

Comparison of PSNR results across the completion methods evaluated on the COCOA-animal-M, COCOA-animal-cls, and COCOA validation sets. Higher values indicate greater pixel-level fidelity.

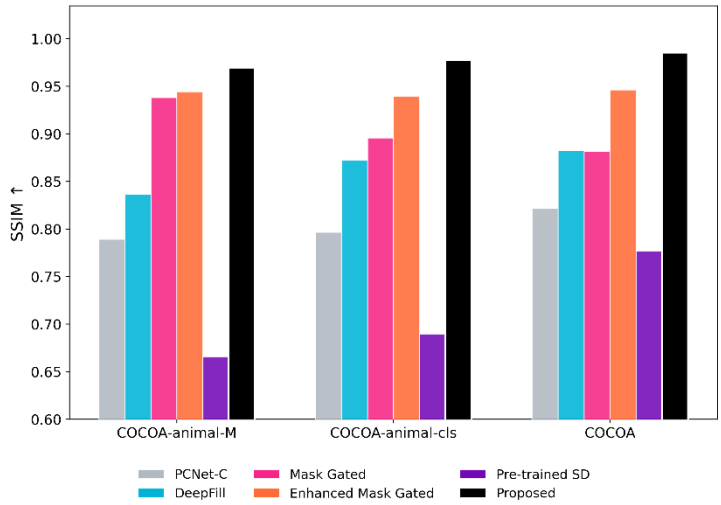


Figure 4

SSIM scores for all completion methods across the three validation sets. Higher values indicate better perceptual and structural coherence with the ground truth.

Concurrently, the proposed model significantly reduces reconstruction errors. On COCOA validation set, it achieves the lowest $\mathcal{L}_1$ error of 0.0121, which demonstrates a clear improvement over baseline methods. Although the enhanced mask gated completion model achieves marginally lower $\mathcal{L}_2$ error across three datasets, these improvements however, do not result in better image quality. Its PSNR remains 7.5 to 8.1 dB lower than that of the proposed model.

Convolution-based baseline methods have limited ability to capture long-range dependencies between image regions. As a result, PCNet-C and DeepFill show lower performance in complex content completion tasks, achieving PSNR values of only about 21-26 dB.

On the other hand, the pre-trained Stable Diffusion model achieves relatively high PSNR values (30-33 dB) but produces noticeably lower SSIM scores, which reaches 0.6658 in Table 1. These results indicate that, although the model can generate realistic textures, it struggles to maintain the structure consistency and spatial relationships required for accurate amodal content completion.

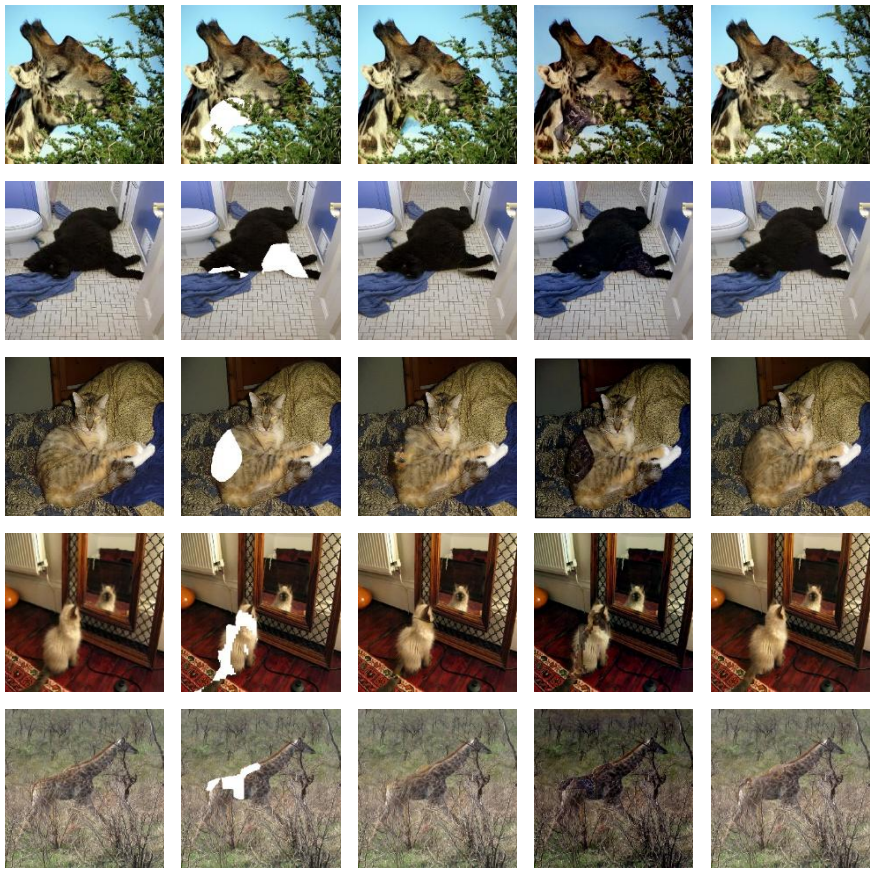
The proposed architecture addresses these limitations by combining a structural aware feature extraction branch and a latent diffusion refinement branch. This design preserves structural consistency while maintaining high visual quality.

The qualitative results, shown in Figure 5, support the quantitative findings. Although the proposed model does not always produce satisfactory reconstructions, it consistently generates more realistic and structurally coherent completions than

the baseline methods. In many cases, DeepFill and PCNet-C produce outputs with visible artifacts, distorted textures, or semantically incorrect content in the occluded regions. In contrast, the model is generally able to preserve the object structure and generate visually plausible content that better matches the surrounding context.

Figure 5

Qualitative evaluation of completed images from the COCOA-animal-cls validation set.



(a) Ground truth (b) Masked image (c) DeepFill (d) PCNet-C (e) Proposed model

We also consider the design possibility of utilizing the weighted mask instead of the binary mask as input to the Stable Diffusion model and the gated convolutional layers, as in mask gated completion method. In contrast to the conventional binary mask, which assigns 0 to missing regions and 1 to valid pixels, the weighted mask [16] assigns three distinct values: 0 to occluded (invalid) pixels, 1 to the visible portion of the target object, and 0.5 to the remaining background regions (Figure 6).

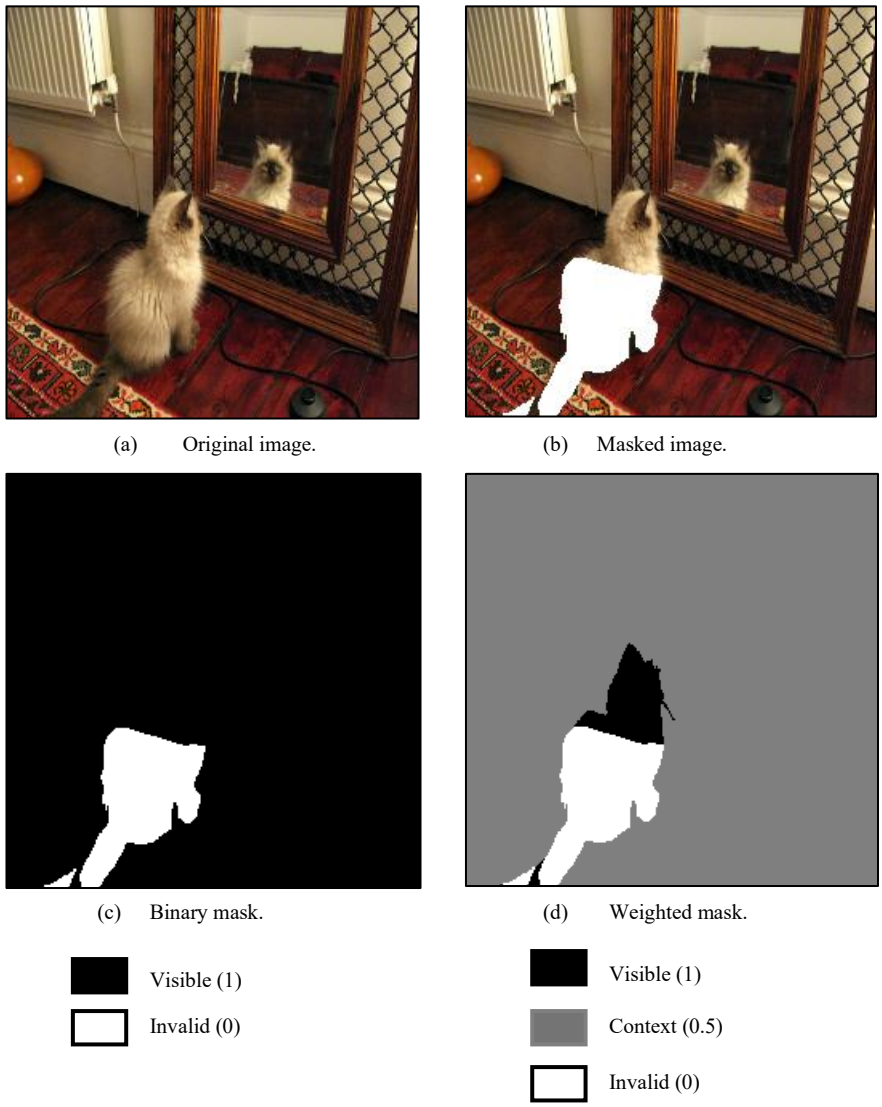


Figure 6

Comparison between the conventional binary mask and the weighted mask. For visualization purposes, the visible region is illustrated in black and the missing region in white, despite their assigned numerical values.

This formulation reflects the assumption that the visible part of the object provides richer semantic cues than the background, particularly for objects with relatively

homogeneous textures, such as animals, where different parts of the same object are more similar than they are to surrounding regions.

To evaluate the effect of the masking strategy, an ablation study was performed by comparing the two masks on all three validation datasets. The results summarized in Table 4 suggest that the binary mask generally provides better performance for the proposed architecture. It produces lower reconstruction errors while maintaining stable and consistent performance across five runs.

On the two animal-focused datasets, the binary mask outperforms the weighted mask. On the COCOA validation set, however, a small trade-off is observed. The weighted mask obtains marginally higher PSNR values; however, this improvement comes with significantly higher reconstruction errors and lower SSIM. Since accurate amodal content completion requires both structural consistency and low reconstruction error, the binary mask is therefore the preferred masking strategy for the proposed model.

Table 4

Quantitative ablation study of two masking strategies on the validation sets of COCOA, COCOA-animal-M, and COCOA-animal-cls. Results are reported as mean ± standard deviation of five independent runs.

COCOA-animal-M				
Description	**$\mathcal{L}_1$error ↓**	**$\mathcal{L}_2$ error ↓**	**PSNR ↑**	**SSIM ↑**
Binary mask	0.0271 ± 0.0003	0.0033 ± 0.0001	35.4633 ± 0.1808	0.9695 ± 0.0007
Weighted mask	0.0475 ± 0.0003	0.0061 ± 0.0002	32.1579 ± 0.1135	0.9296 ± 0.0010
COCOA-animal-cls				
Description	**$\mathcal{L}_1$error ↓**	**$\mathcal{L}_2$ error ↓**	**PSNR ↑**	**SSIM ↑**
Binary mask	0.0192 ± 0.0003	0.0025 ± 0.0002	36.939 ± 0.2627	0.9775 ± 0.0008
Weighted mask	0.2170 ± 0.0004	0.0030 ± 0.0003	36.0491 ± 0.3002	0.9748 ± 0.0008
COCOA				
Description	**$\mathcal{L}_1$error ↓**	**$\mathcal{L}_2$ error ↓**	**PSNR ↑**	**SSIM ↑**
Binary mask	0.0121 ± 6e-05	0.0016 ± 5e-05	36.174 ± 0.1195	0.9854 ± 0.0002
Weighted mask	0.0263 ± 6e-05	0.0025 ± 4e-05	36.523 ± 0.0548	0.9622 ± 0.0002

We also compare the computational complexity of the proposed model with the baseline methods in terms of the number of trainable parameters, total parameters, computational complexity (FLOPs), inference time, and peak GPU memory consumption. The reported values in Table 5 were measured during inference using the same hardware configuration and evaluation settings. For the proposed model, the total parameter counts, FLOPs, inference time, and GPU memory include the frozen Stable Diffusion feature extractor, whereas the trainable parameter count corresponds only to the learnable components optimized during training. For

PCNet-C, the difference between trainable and total parameters is due to the frozen mask-update convolution layers used in partial convolution. Mask gated method and DeepFill utilize the same underling architecture, therefore, only DeepFill was included in the table.

Among the evaluated models, PCNet-C exhibits the highest number of trainable parameters (25.783 million), while DeepFill and the enhanced mask gated require 16.175, and 17.349 million, respectively. On the other hand, the proposed model requires only 6.552 million. This indicates that the proposed architecture requires significantly fewer learnable paraments than the baseline models.

Despite its smaller trainable network, the proposed method has the largest total parameter count, FLOPs, inference time, and GPU memory consumption compared to convolution-based models. This is mainly due to the incorporation of the frozen Stable Diffusion feature extractor, which is employed during inference to provide high-level semantic guidance but is not optimized during training. However, compared with directly using the pre-trained Stable Diffusion model, the proposed model achieves lower inference time. This is because the pre-trained Stable Diffusion pipeline performs iterative denoising over multiple steps.

Table 5

Computational complexity comparison of the proposed model and the baseline methods.

Model	**Trainable Parameters (M)**	**Total Parameters (M)**	**FLOPs (G)**	**Inference Time (ms/image)**	**Peak GPU Memory (MB)**
PCNet-C	25.783	51.557	304.990	1.257	1652.85
DeepFill	16.175	16.186	1239.668	7.975	2051.52
Enhanced Mask Gated	17.349	17.360	1349.650	8.204	2107.60
Pre-trained SD	0.000	1066.250	4291.498	2185.137	4801.94
Proposed	6.552	949.741	2115.276	14.684	12730.73

In summary, these results indicate a trade-off between computational efficiency and reconstruction capability. While the proposed model has a high inference cost due to the pretrained diffusion backbone, it maintains the smallest trainable architecture among all evaluated methods while leveraging the rich semantic representations provided by the frozen diffusion model.

5 Conclusions

This paper presents a model that combines the mask-aware feature selection through gated convolutions, semantic representations extracted from a pre-trained latent diffusion model, and a hierarchical transformer for multi-scale feature fusion and long-range contextual reasoning to reconstruct the hidden regions of partially occluded objects.

Experimental results demonstrate that the proposed approach effectively generates realistic and perceptually coherent reconstructions, outperforming both convolution-based methods and diffusion models. These findings indicate that, although the Stable Diffusion model is sometimes unable to produce satisfactory results when completing a complex occlusion scene, its representational power can provide an effective guidance for content completion.

Despite these, the proposed model faces challenges when reconstructing objects with highly complex or articulated shapes, particularly human figures, where strong semantic priors are required to produce plausible completions.

Future work will focus on improving the modelling of such challenging scenarios.

Acknowledgements

We acknowledge the support of the Doctoral School of Applied Informatics and Applied Mathematics at Obuda University and the “Examining Occlusions Using Deep Machine Learning” project for access to the HUN-REN Cloud [34] [35], which helped us in conducting the experiments of this work.

References

[1] K. Saleh, S. Szénási and Z. Vámossy, "Occlusion handling in generic object detection: A review," *2021 IEEE 19th World Symposium on Applied Machine Intelligence and Informatics (SAMI),* pp. 000477-000484, 2021.

[2] J. Ao, Q. Ke and K. A. Ehinger, "Image amodal completion: A survey," *Computer Vision and Image Understanding,* vol. 229, p. 103661, 2023.

[3] R. Rombach, A. Blattmann, D. Lorenz, P. Esser and B. Ommer, "High-resolution image synthesis with latent diffusion models," *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition,* pp. 10684-10695, 2022.

[4] W. Xiong, J. Yu, Z. Lin, J. Yang, X. Lu, C. Barnes and J. Luo, "Foreground-aware image inpainting," *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition,* pp. 5840-5848, 2019.

[5] E. Ozguroglu, R. Liu, D. SurÍs, D. Chen, A. Dave, P. Tokmakov and C. Vondrick, "pix2gestalt: Amodal segmentation by synthesizing wholes," *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 3931-3940, 2024.

[6] K. Xu, L. Zhang and J. Shi, "Amodal completion via progressive mixed context diffusion," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,* pp. 9099-9109, 2024.

[7] J. Ao, Y. Jiang, Q. Ke and K. A. Ehinger, "Open-world amodal appearance completion," *Proceedings of the Computer Vision and Pattern Recognition Conference,* pp. 6490-6499, 2025.

[8] H. Fan, L. Wang, H. Chen, Z. Huang, J. Wu and L. Sheng, "Multi-agent amodal completion: Direct synthesis with fine-grained semantic guidance," *Proceedings of the 33rd ACM International Conference on Multimedia,* pp. 9911-9919, 2025.

[9] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark and others, "Learning transferable visual models from natural language supervision," *International conference on machine learning,* pp. 8748-8763, 2021.

[10] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat and others, "GPT-4 technical report," *arXiv preprint arXiv:2303.08774,* 2023.

[11] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo and others, "Segment anything," *Proceedings of the IEEE/CVF international conference on computer vision,* pp. 4015-4026, 2023.

[12] G. Zhan, C. Zheng, W. Xie and A. Zisserman, "What does stable diffusion know about the 3d scene?," *arXiv preprint arXiv:2310.06836,* 2023.

[13] G. Zhan, C. Zheng, W. Xie and A. Zisserman, "Amodal ground truth and completion in the wild," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,* pp. 28003-28013, 2024.

[14] X. Zhan, X. Pan, B. Dai, Z. Liu, D. Lin and C. C. Loy, "Self-supervised scene de-occlusion," *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition,* pp. 3784-3792, 2020.

[15] Y. Zhu, Y. Tian, S. Shankar and S. Savarese, "Semantic amodal segmentation," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 1464-1472, 2017.

[16] K. Saleh, S. Szénási and Z. Vámossy, "Mask guided gated convolution for amodal content completion," *2024 IEEE 22nd Jubilee International Symposium on Intelligent Systems and Informatics (SISY),* pp. 000321-000326, 2024.

[17] K. Ehsani, R. Mottaghi and A. Farhadi, "Segan: Segmenting and generating the invisible," *Proceedings of the IEEE conference on computer vision and pattern recognition,* pp. 6144-6153, 2018.

[18] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu and T. S. Huang, "Free-form image inpainting with gated convolution," *Proceedings of the IEEE/CVF international conference on computer vision,* pp. 4471-4480, 2019.

[19] X. Li, C. Yi, J. Lai, M. Lin, Y. Qu, S. Zhang and L. Cao, "SynergyAmodal: Deocclude anything with text control," *Proceedings of the 33rd ACM International Conference on Multimedia,* p. 9444–9453, 2025.

[20] Z. Liu, Q. Liu, C. Chang, J. Zhang, D. Pakhomov, H. Zheng, Z. Lin, D. Cohen-Or and C.-W. Fu, "Object-level scene deocclusion," *ACM SIGGRAPH 2024 Conference Papers,* 2024.

[21] G. Liu, F. A. Reda, K. J. Shih, T.-C. Wang, A. Tao and B. Catanzaro, "Image inpainting for irregular holes using partial convolutions," *Proceedings of the European conference on computer vision (ECCV),* pp. 85-100, 2018.

[22] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *International Conference on Learning Representations (ICLR),* 2014.

[23] J. Ho, A. Jain and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems (NeurIPS),* pp. 6840-6851, 2020.

[24] O. Ronneberger, P. Fischer and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," *Medical Image Computing and Computer-Assisted Intervention (MICCAI),* pp. 234-241, 2015.

[25] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin and B. Guo, "Swin Transformer: Hierarchical vision transformer using shifted windows," *Proceedings of the IEEE/CVF international conference on computer vision (ICCV),* pp. 10012-10022, 2021.

[26] C. Yang, X. Lu, Z. Lin, E. Shechtman, O. Wang and H. Li, "High-resolution image inpainting using multi-scale neural patch synthesis," *Proceedings of the IEEE conference on computer vision and pattern recognition,* pp. 6721-6729, 2017.

[27] R. Zhang, P. Isola, A. A. Efros, E. Shechtman and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," *Proceedings of the IEEE conference on computer vision and pattern recognition,* pp. 586-595, 2018.

[28] W. Xian, P. Sangkloy, V. Agrawal, A. Raj, J. Lu, C. Fang, F. Yu and J. Hays, "Texturegan: Controlling deep image synthesis with texture patches," *Proceedings of the IEEE conference on computer vision and pattern recognition,* pp. 8456-8465, 2018.

[29] T. Miyato, T. Kataoka, M. Koyama and Y. Yoshida, "Spectral normalization for generative adversarial networks," *International Conference on Learning Representations (ICLR),* 2018.

[30] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *International Conference on Learning Representations (ICLR),* 2015.

[31] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *International Conference on Learning Representations (ICLR),* 2015.

[32] K. He, X. Zhang, S. Ren and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," *Proceedings of the IEEE international conference on computer vision,* pp. 1026-1034, 2015.

[33] K. Saleh, S. Szénási and Z. Vámossy, "Amodal animal completion via mask guided enhanced gated convolution," *2026 IEEE 13th International Joint Conference on Cybernetics and Computational Cybernetics, Cyber-Medical Systems (ICCC),* pp. 000305-000310, 2026.

[34] M. Héder, E. Rigó, D. Medgyesi, R. Lovas, S. Tenczer, A. Farkas, M. B. Emődi, J. Kadlecsik and P. Kacsuk, "The past, present and future of the ELKH cloud," *Információs Társadalom: Társadalomtudományi Folyóirat,* vol. 22, pp. 128-137, 2022.

[35] M. B. Emődi, K. Bánfi and J. Kovács, "SLURM deployment in cloud environments: Enhancing utilization and scalability," *Acta Polytechnica Hungarica,* vol. 22, p. 261275, 2025.