# Dual-CLIP Guided Diffusion Model For Image Object Removal

**Tianyi Shang[1], Weijun HU[1], Han Huang[1], Zihan Ruan[1], Rongen Guo[1], Haoyan Zhong[1]**
*[1]Maynooth International Engineering College, Fuzhou University, Fujian, China*

***Abstract—The removal and restoration of objects in images is a highly challenging task, requiring the deletion of existing objects within user-specified masked regions, followed by filling the deleted areas with background content based on the image's contextual information. The model may encounter confusion or fail to completely remove the original objects due to unclear boundary definitions. Additionally, it may struggle to fully leverage the global contextual information, resulting in less appropriate background generation. When the missing region is large, the increased size of the masked area significantly adds to the difficulty of background filling. Previous researchers have employed various Convolutional Neural Networks (CNNs)* [1] *and Transformer* [2][3] *models in an attempt to capture global contextual information for restoration tasks. In recent years, the latest advancements in diffusion models[4][5] have notably driven progress in this field.***

***In this study, we propose an improved diffusion model called DCGD (Dual-CLIP Guided Diffusion Model), aiming to more effectively remove masked objects and ensure that the generated content fully considers the global background. Inspired by the revolutionary work of CLIP[6][7] (Contrastive Language-Image Pre-training Model), we innovatively use CLIP to separately describe the masked region Tmask and the unmasked region Tinfo in the image. We then use these two text descriptions to guide the diffusion model. The description of the masked region Tmask allows the model to accurately understand the object that needs to be removed, while the description of the unmasked region Tinfo provides global contextual information of the image to generate appropriate background content.***

***This dual-guidance approach makes object removal more precise and accurate, and the generated background better considers the overall integrity of the image, avoiding potential confusion and inconsistencies.Extensive experimental results demonstrate that our DCGD model performs excellently in practical applications. It is capable of intelligently removing objects and generating restoration content that is highly consistent with the global background, resulting in seamless and realistic images. Our approach shows great potential for various applications in image editing and restoration.***

***Keywords: Cross Model，Image Implating，Image Object Removal , Computer Vision***

## I. Introduction

The core of image inpainting[8] lies in filling user-specified masked regions with plausible content while ensuring that the generated results are consistent with the surrounding environment. This technology has found broad applications in various real-world scenarios, such as virtual background replacement[9], photo restoration[10], and style transfer[11].

Early image inpainting methods mainly relied on geometric structures to infer missing content. However, with the rapid development of deep learning in computer vision, Convolutional Neural Networks (CNNs) have been widely adopted for feature extraction. Among these, U-Net[12], with its encoder-decoder structure, has become a fundamental model in the field of image inpainting. The core idea behind U-Net is to extract features through downsampling and then reconstruct the output image through upsampling.

The use of U-Net in image inpainting has had a profound impact on the field. With the emergence of text-to-image models such as Stable Diffusion, researchers identified that diffusion models also adopt a similar encoder-decoder structure. These models gradually restore images from noise through multiple iterative steps, making them particularly well-suited for object removal tasks. The transformer-based architecture used in diffusion models allows them to capture global features more effectively. More importantly, this approach has significantly accelerated the development of semantic-guided image inpainting.

In semantic-guided image inpainting, language input provides valuable feedback and guidance for the generated results. Through text descriptions, users can not only correct errors in image inpainting but also generate specific objects within masked regions and specify the desired style. This interactive guidance makes the inpainting process more intelligent and offers powerful support for personalized generation.

In this study, we focus on the challenging subtask of object removal within the field of image inpainting. While we aim to leverage semantic information to guide the image generation process, object removal differs fundamentally from tasks like, "Please draw a dog in the middle of the image." In the case of object removal, human language descriptions offer limited value, as the task is solely to remove objects from the image and restore the background. All the necessary information for this process is already present within the image itself.

To enhance diffusion-based object removal using the semantic information embedded within the image, we propose a novel DCGD (Dual-CLIP Guided Diffusion) model, which integrates a fine-tuned CLIP model for object removal tasks.

Specifically, DCGD is a diffusion-based text-to-image generation model, but it differs from conventional approaches: the text information used to guide the diffusion process is not provided externally by human descriptions but is instead extracted directly from the image itself. For a given image, we employ a pre-trained CLIP model to generate two textual descriptions. These descriptions are used as follows: Tmask – This description corresponds to the masked region, representing the object to be removed. We use this as a negative prompt to ensure that the diffusion process avoids regenerating the object when filling in the masked area. Tinfo – This description captures the unmasked portion of the image, containing global contextual information. It serves as a positive prompt, guiding the diffusion model to generate a background that aligns seamlessly with the overall style of the image, ensuring the masked area is filled coherently.

Through this approach, DCGD reduces errors where objects fail to be removed entirely while also leveraging the global information of the image to enhance background generation quality and accuracy. This makes DCGD a highly competitive solution for object removal tasks.

Our key contributions are as follows:

1. Innovative integration of CLIP with diffusion models: We creatively utilize semantic information inherent to the image rather than relying on external human descriptions, enhancing the effectiveness of object removal.

2. Dual-region descriptions as negative and positive prompts: We develop a novel method of generating distinct textual descriptions for different regions of the image—using the masked region's description as a negative prompt and the unmasked region's description as a positive prompt.

3. Superior performance in object removal: Our DCGD model demonstrates strong performance in object removal tasks, improving both the precision and quality of generated content.

## II. Literature Review

### A. *Visual Language Model*

With the rapid development of computer vision and natural language processing, aligning visual information with semantic information and uncovering the relationships between them has become a critical research topic. In this context, vision-language pre-trained models have emerged. These models are typically trained on large-scale datasets, mapping image and text information into the same latent semantic space to capture their deep correlations. Once pre-trained, these models can be applied to various downstream tasks, such as visual question answering (VQA) and image generation.

CLIP is one of the most representative vision-language pre-trained models, developed by OpenAI. Its architecture is straightforward: a large number of image-text pairs are used for training. The text is processed by a Transformer-based text encoder, while the image is processed by a Transformer-based image encoder. The encoded text and image features are then optimized through contrastive learning[13]: the model minimizes the distance between matching image-text pairs while increasing the distance between non-matching ones. This process ensures that semantically similar images and texts are effectively clustered together.

Although CLIP adopts a simple approach, it delivers remarkable performance and exhibits strong generalization capabilities. This generalization power primarily stems from training the model on a large-scale dataset. To develop CLIP, OpenAI built a dataset containing 400 million image-text pairs, all collected from publicly available resources on the internet. Such an extensive dataset enables the model to accurately recognize new images and texts that it has never encountered before.

Thanks to its outstanding generalization ability, CLIP plays an essential role in various downstream tasks and demonstrates great potential for wide applications in both the vision and language domains.

In the field of image segmentation, Lseg[14] leverages CLIP to achieve powerful zero-shot capabilities. In object detection, ViLD[15] not only identifies labels present in the training dataset but also extends to recognizing unseen new vocabulary. Similarly, GLIP[16] follows CLIP's approach by adopting semi-supervised learning, exhibiting strong generalization capabilities as well. In the image generation domain, CLIPasso[17] aligns sketches with their original images by ensuring maximum similarity through CLIP's image encoder.

In our work, we utilize CLIP's core functionality to generate corresponding text descriptions from images. CLIP's strong generalization ability enables it to provide accurate descriptions across a wide variety of images.

### B. *Image Implating*

Current image inpainting tasks cover multiple aspects, including object removal, text erasure, irregular mask inpainting, and old photo restoration. Initially, Pathak et al. proposed a foundational encoder-decoder model—the Context Encoder[18]. This model downsamples images via an encoder to extract features, then progressively upsamples the compressed features through a decoder, ultimately generating a repaired image consistent with the original size. However, the Context Encoder can only inpaint occluded regions located at the center of the image, which presents certain limitations.

To address these issues, researchers introduced a dual-branch approach[19], processing the global and local information of the image separately to enhance inpainting results. Subsequently, Liu et al. improved the visual inpainting effect for large occluded regions by introducing Partial Convolutions[20]. For large-hole inpainting in specific scenarios, especially when multiple plausible inpainting results are required, Zheng et al. proposed a solution using parallel GAN networks[21], generating progressive texture inpainting results by learning the prior distribution of missing regions. Additionally, some researchers have explored multi-stage network architectures from coarse to fine, incrementally refining the image inpainting process. To further optimize the inpainting effect of multi-stage networks, Yu et al. and Yi et al. proposed improved gated convolutions[22], addressing the issue where traditional convolutions treat all inputs as valid pixels. This method dynamically selects features, significantly enhancing the quality and flexibility of the inpainted images.

However, the aforementioned methods are only suitable for low-resolution images and perform poorly on high-resolution images. To address this, CRA[23] improved the quality of high-resolution inpainting by aggregating multiple high-level features. Yu and Lin et al.[24] adopted a two-stage inpainting method; in the second stage, they generated higher-resolution inpainted images by iterative feedback, upsampling, and considering confidence.

Although these techniques have improved inpainting results to some extent, the generated images often suffer from over-smoothing. To solve this issue, Kamyar Nazeri proposed the EdgeConnect method. This method first detects the edge information of the image and uses the edge map as prior knowledge to guide the subsequent image generation process.

With the rise of text-to-image diffusion models, diffusion-based methods have surpassed previous approaches in the field of image inpainting. By introducing techniques such as structure guidance, latent space inpainting, and 3D

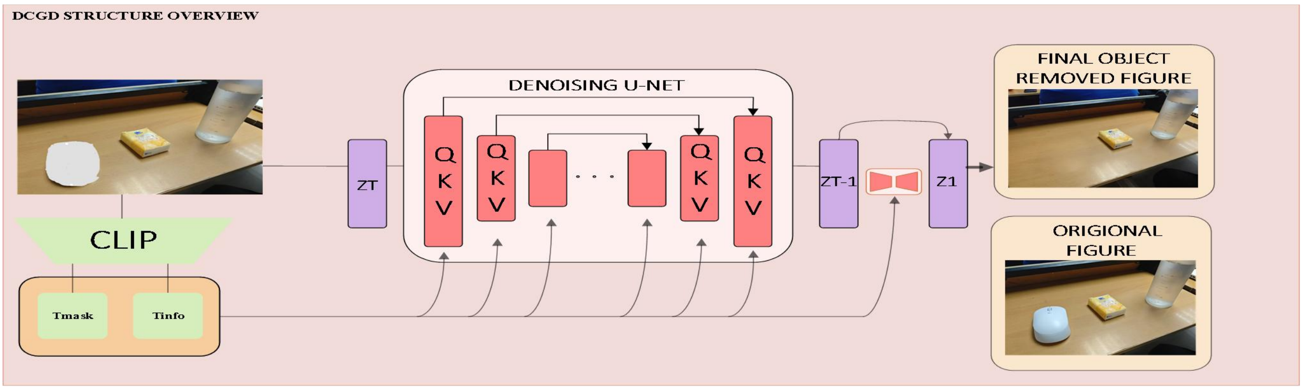

Fig1.The schematic of our method uses CLIP to generate a positive and a negative promote to guide the image reconstruction of the diffusion model

perception, the inpainting results have been effectively improved. LatentPaint[25] operates in the latent space to reduce computational costs; RenderDiffusion[26] achieves multi-view consistent inpainting through 3D representations. Additionally, methods combining Controllable Probabilistic Models (TPM)[27] and global structure guidance have further enhanced the generation quality and consistency in image and text inpainting.

In our method, we fine-tune a pre-trained diffusion model and employ trainable textual guidance to achieve precise image object erasure.

## III. Methodology

### *A. Diffusion models*

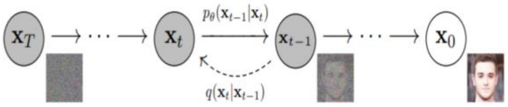
$\mathbf{x}_T$
$\mathbf{x}_t$
$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$
$\mathbf{x}_{t-1}$
$\mathbf{x}_0$
$q(\mathbf{x}_t|\mathbf{x}_{t-1})$

Fig2. The basic process of a diffusion model

As Shown in Fig2. diffusion models are a class of generative models. Their fundamental principle is to gradually add Gaussian noise to image data until a purely Gaussian-distributed image is obtained. Then, a model is trained to progressively remove the Gaussian noise, ultimately producing a new image

Forward Diffusion Process： We first need to define a forward process, gradually adding Gaussian noise：

$$q(x_t|x_{t-1}) = \mathcal{N}\left(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I\right) \tag{1}$$

Where：

$\beta_t$ is a small positive variance schedule (usually predefined).

$\mathcal{N}$ denotes a Gaussian distribution.

$I$ is the identity matrix.

This process gradually transforms the original data into pure Gaussian noise as t approaches T.

Reverse Diffusion Process： The goal of the Reverse Diffusion Process is to predict the noise that needs to be removed from the current image, which can be regarded as the reverse Process of the Forward Diffusion Process.We model the reverse process as:

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}\left(x_{t-1}; \mu_\theta(x_t,t), \Sigma_\theta(x_t,t)\right) \tag{2}$$

Where:

$\theta$ represents the parameters of the neural network.

$\mu_\theta$ and $\Sigma_\theta$ are the mean and covariance functions learned by the model.

The model is trained to predict the noise added at each step by minimizing the difference between the true noise and the noise predicted by the model. The simplified loss function is:

$$L(\theta) = \mathbb{E}_{t,x_0,\epsilon}[\|\epsilon - \epsilon_\theta(x_t,t)\|^2] \tag{3}$$

To generate new data, we start with pure Gaussian noise $x_T \sim \mathcal{N}(0,I)$, then iteratively apply the learned denoising steps:

$$x_{t-1} = \frac{1}{\sqrt{1-\beta_t}}\left(x_t - \beta_t\epsilon_\theta(x_t,t)\right) + \sqrt{\beta_t}z \tag{4}$$

Where：

For $t > 1, z \sim \mathcal{N}(0,I)$ is Gaussian noise; for $t = 1, z = 0$. $\epsilon_\theta(x_t,t)$ is the model's estimate of the noise at time $t$.

By recursively applying this reverse process, the model gradually transforms the noisy data into a sample that closely resembles the true data distribution.

More specifically, a diffusion model is a latent variable model that maps to the latent space using a MarkovChain (MC). Through the Markov chain, noise is gradually added to the data xi in each time step t to obtain a posterior probability q(x1:T | x0), where x1,... xT represents that the input data is also latent space, that is, the latent space of the Diffusion Models has the same dimension as the input data.

A Markov chain is a random process in a state space that goes through a transition from one state to another. This process requires a "memory-free" nature: the probability distribution of the next state can only be determined by the current state, and the events preceding it in the time series have nothing to do with it. This particular type of "memorylessness" is called the Markov property. The Diffusion Models are divided into the forward diffusion process and the reverse reverse diffusion process. From x0 to the final xT is a Markov chain, representing the random process in the state space through the transition from one state to another.

### B. CLIP

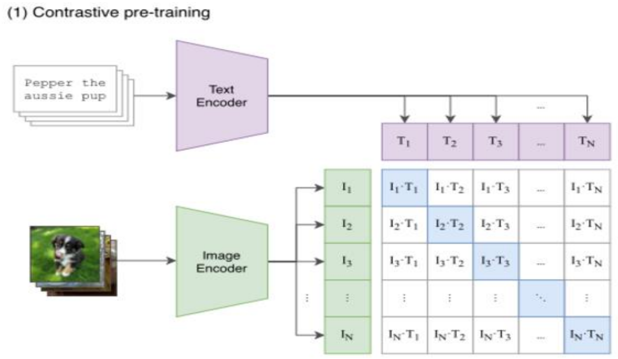

Fig3. training stage of CLIP

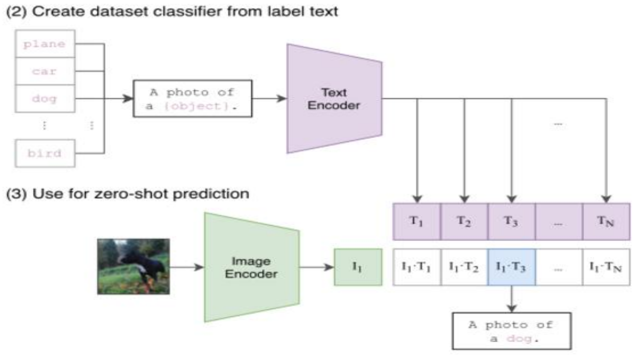

Fig4. Inference Stage of CLIP

CLIP (Contrastive Language - Image Pre-training) establishes zero-shot transfer, natural language supervision and comparative learning. CLIP is a pre-trained model that also acts as a backbone, its function is to input a piece of text (or an image) and output a vector representation of the text (image). The difference between CLIP and other pre-trained models is that CLIP is multi-modal and can process the information of two modes at the same time, and then analyze the correlation between the two input information, so as to carry out the next step of processing, while the pre-trained models such as VIT,DINO and BERT can only process the information of a single mode. In this article, CLIP is used to process the ability of image-text cross-modal information, which is not available in other models

The basic starting point of CLIP is: since the language and image processing mode are both based on transformer, are the vectors extracted between the two relevant? If the two are correlated, then can we get rid of the limitation of the number of image data sets and train a cross-modal model on a large image-text data set using text information as a supervisory signal? This is the core of the CLIP method: using natural language supervision signals to train a better visual model.

Using text signals as a monitoring method is undoubtedly very effective, and the first benefit here is the ease of building data sets. In previous image datasets, such as ImageNet, images were divided into 1000 categories, and the trained model could only identify 1000 categories at most. However, in the image-text data set that CLIP uses, the text data is just a short paragraph, such as: "A dog in a hat" or "an electronics engineer doing research", this category of description becomes extremely large, and the data set only needs to be crawled from the Internet, making the data set of picture-text collection becomes easy, OpenAI specifically to collect a data set WIT (WebImageText), With 400 million image-text pairs, this huge amount of data is one of the most critical factors in clip's success.

Another obvious advantage of this kind of training through text information supervision is that the model trained in this way is a multi-modal model. As shown in Figure 2, CLIP is divided into text branches and image branches, which are encoded respectively, and the features of the encoded image-text pair are highly similar. This makes clip a pre-trained text-picture cross-modal large model for the ages. In addition, as we can see above, clip's text description is open, which enables clip to identify all relevant words that have appeared in the 400 million training corpus, thus making clip obtain unprecedented zero-shot capability. Making clip's potential for downstream tasks almost endless, a large number of cross-modal related papers have used clip as a text encoder.

The most important point of Clip in this training process is the use of contrast learning as a loss function， as shown in Fig3. The author initially tried the loss function of conditional text generation, that is, to predict a corresponding text description for each picture, but the same picture can correspond to multiple descriptions, using the loss function of this predictive shape will make the training unstable. If we use contrast learning, we only need to reason whether the given picture and text are related, which simplifies the difficulty of model reasoning, improves the efficiency of training reasoning, and in the face of noise in data sets, this method is also more conducive to convergence. The actual meaning of contrast learning is to narrow the distance of similar image text pairs, and to distance all other image samples that are not correctly matched. With this loss function, clip performs this training very efficiently on large data sets. As shown in Figure 4, we can see that clip calculates the similarity between the query image and all text prompts respectively in the inference stage, and finally obtains the most similar image-text pair, completing the classification task。

In our article, we want to use two different text prompt guided diffusion models for image reconstruction. Essentially, what we need is cross-modal interaction between image and text. CLIP's powerful cross-modal properties are very important to us, and we try to fine-tune CLIP's text encoder to encode text in this way. We hope that the encoded text embedding can be as close as possible to the image embedding space of the diffusion model, so that the diffusion model can fully understand the text guidance information in the reconstruction process. In our experiment, it is proved that the idea of fine-tuning clip is very effective, which greatly affects the quality of object removal in the image, which is crucial for our experiment.

### C. Dual-CLIP Guided Diffusion Model

In this paper, we propose a novel method called the Dual-CLIP Guided Diffusion Model (DCGD) for image restoration, with a particular focus on removing target objects from complex backgrounds. The DCGD framework is based on the diffusion process, which we extend by refining the input structure and incorporating two prompts generated by CLIP. These prompts guide the diffusion process, resulting in enhanced object removal capabilities.

To enhance the model's understanding of both masked regions and overall image details, we modify the input structure during the diffusion process. The input is defined as:

$$x_t^{'} = \mathrm{x_t} \oplus (\mathrm{x_0} \circ \mathrm{o}) \oplus (\mathrm{x_0} \circ (1 - \mathrm{o})) \quad (5)$$

Where:

$\mathrm{x_t}$ denotes the noisy latent variable.

$\mathrm{x_0}$ represents the original image.

o corresponds to the masked region, and $1 - o$ represents the unmasked region.

The symbol $o$ indicates element-wise multiplication.

$\oplus$ denotes concatenation.

Based on this input structure, we initiate the diffusion process. However, a significant challenge arises if we aim to remove an object from a background that frequently appears in the image, as it could lead to inaccurate restoration or even the regeneration of removed objects. To address this challenge, we introduce the dual CLIP guidance, utilizing two text prompts to assist the denoising task: the Positive Prompt $\mathrm{T_{info}}$ describes the overall content of the image, helping maintain consistency in the restored scene. The Negative Prompt $\mathrm{T_{mask}}$ provides detailed information of the masked region, guiding the model to suppress features in that area and preventing the reconstruction of the removed object.

Similar to traditional diffusion models, our DCGD model uses a U-Net architecture for the denoising network. With dual CLIP-guided prompting, the noise estimation is defined as:

$$\hat{\epsilon}_\mathrm{t} = \mathrm{w} \cdot \epsilon_\theta\left(\mathrm{x}_\mathrm{t}^{'}, \tau_\theta(\mathrm{T_{info}}), \mathrm{t}\right) + (1 - \mathrm{w}) \cdot \epsilon_\theta\left(\mathrm{x}_\mathrm{t}^{'}, \tau_\theta(\mathrm{T_{mask}}), \mathrm{t}\right) \quad (6)$$

Where:

w is a weight parameter balancing the contributions of the positive and negative prompts.

$\epsilon_\theta$ represents the output of the denoising network.

$\tau_\theta$is the text embedding produced by the CLIP encoder.

During training, we optimize the model by minimizing the mean squared error (MSE) between the predicted noise and the ground truth noise:

$$\mathrm{L} = \mathbb{E}_{\mathrm{x_0},0,\mathrm{t},\mathrm{T_{info}},\mathrm{T_{mask}},\epsilon_\mathrm{t}}[||\epsilon_\mathrm{t} - \hat{\epsilon}_\mathrm{t}||^2] \quad (7)$$

Where:

$\epsilon_\mathrm{t}$ denotes the ground truth noise.

$\hat{\epsilon}_\mathrm{t}$ is the predicted noise.

By minimizing this loss function, the model is better able to reconstruct image details and effectively remove noise.

We train and validate our model using this approach. Our method performs well across multiple real-world tasks, achieving satisfactory results in object removal while demonstrating superior performance on varinus datasets.

## IV. EXPERIMENT AND RESULT

### A. Experiment Settings

We have carried out full and detailed experiments on the trained DCGD. Our experiments were verified on two datasets of Paris StreetView[28] and Places[29], and compared with EdgeConnect[30], Global-Local[31]methods. Our training took place on an Nvidia V100 GPU where we set the batch size to 64, the learning rate to 1e-4, and used the adam optimizer for fine tuning. Our experiment was divided into three parts, namely visualization part, quantization result comparison part and ablation experiment part. Through our complete experiment, we fully and effectively proved the effectiveness of the DCGD model, and our double-flow CLIP structure significantly enhanced our performance in the image erasure part.

### B. Visualization analysis

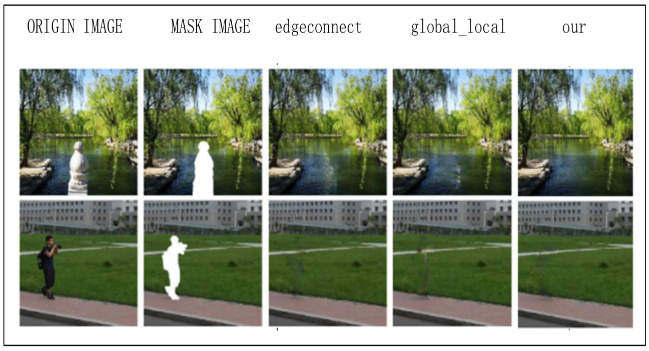

Fig5. Visualization results on StreetView dataset

As shown in Fig5, we found that our method was first tested on the test set of StreetView dataset. We compared our method with the advanced methods EdgeConnect[30] and Global-Local[31]. Through the image, we could find that Given the object erased by the image, our method effectively reduces the artifacts after object removal. Compared with the other two methods, the residual traces of object removal are slighter than those of the comparison method, which effectively proves the effectiveness of our method in object removal. In the first line of fig5, we can see that there are obvious white traces in the middle of the river recreated by EdgeConnect and Global-Local, which is the artifact of the original white sculpture. However, in our method, the river generated is very natural, and the white artifact in the middle is basically invisible. A similar analysis can be performed on the erase results of the second row. Finally, we demonstrated the validity of our method in the visualized results of this dataset, and our method is very effective in this basic erasure task.

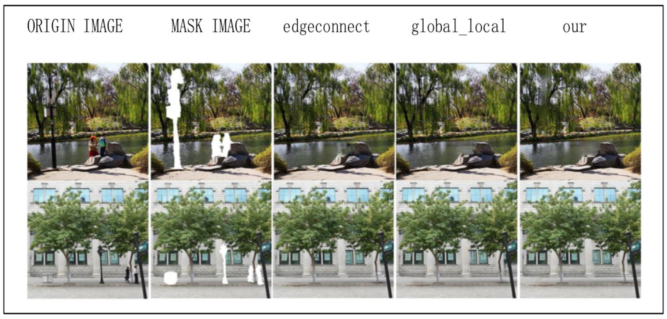

Fig6. Visualization results on Places Dataset

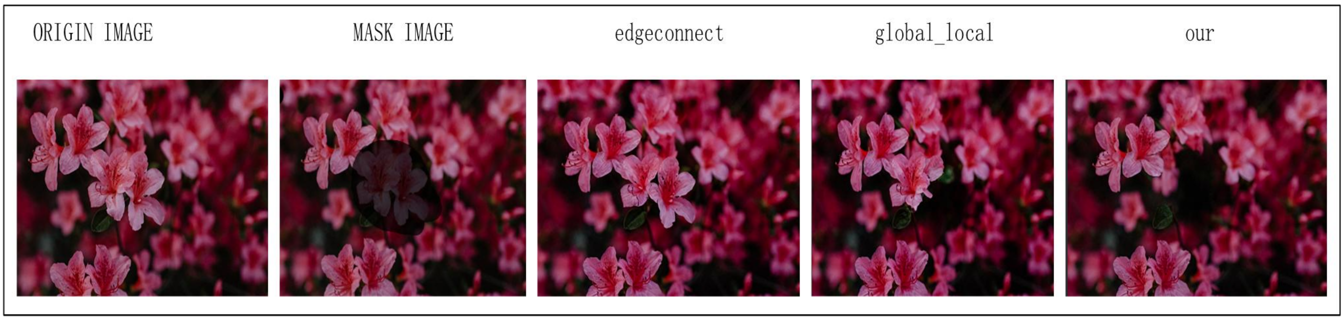

Fig7. When surrounded by very similar backgrounds, our DCGD approach shows clear advantage

In fig6, we show the effect of our method on the test set of the Places Dataset. The significance of this result is that our method has good performance on both datasets. The specific analysis of images is the same as that on the StreetView dataset, so it is not necessary to go into details

In fig7, the biggest advantage of our DCGD can be intuitively felt: our method performs well in an environment surrounded by similar objects. Other methods, such as edgeconnect and global_local, will use repeated objects in the background to fill the background under the condition of dense repetition. In our figure, we mask the two flowers in the middle and want to erase them, but there are also a large number of similar and repeated flowers in the background. edgeconnect and global_local fill with flowers in the background, resulting in a complete failure to erase objects. However, our DCGD unique design uses clip to describe the masked object, and then uses this description as a negative prompt to guide the model to inhibit the generation of the masked object (in this case, an iris) during background reconstruction. Fig7 results strongly demonstrate the effectiveness of our design and represent the biggest innovation of our model.

### *C. Quantization result comparison*

| | Mask percent | EdgeConnect | Global-Local | OUR |
|---|---|---|---|---|
| PSNR | 0.01 – 0.1 | 32.08 | 32.37 | **32.62** |
| | 0.1 – 0.2 | 27.28 | 27.33 | **27.39** |
| | 0.2 – 0.3 | 23.71 | 23.82 | **23.89** |
| | 0.3 – 0.4 | 22.28 | 22.43 | **22.79** |
| SSIM | 0.01 – 0.1 | **0.950** | 0.949 | 0.950 |
| | 0.1 – 0.2 | 0.873 | **0.875** | 0.872 |
| | 0.2 – 0.3 | 0.770 | **0.770** | 0.763 |
| | 0.3 – 0.4 | 0.694 | **0.699** | 0.693 |
| MAE | 0.01 – 0.1 | 0.0044 | **0.0037** | 0.0049 |
| | 0.1 – 0.2 | 0.0102 | 0.0090 | **0.0088** |
| | 0.2 – 0.3 | 0.0205 | 0.0204 | **0.0195** |
| | 0.3 – 0.4 | 0.0275 | 0.0274 | **0.0258** |
| LPIPS | 0.01 – 0.1 | 0.0643 | 0.0608 | **0.0580** |
| | 0.1 – 0.2 | 0.1708 | 0.1622 | **0.1516** |
| | 0.2 – 0.3 | 0.3423 | 0.3396 | **0.3265** |
| | 0.3 – 0.4 | 0.4677 | 0.4594 | **0.4314** |

Table1. Quantitative analysis on the Places dataset

As shown in Table1, we conducted detailed experiments on the Places dataset to compare the performance of our method with other methods.

PSNR measures the quality of image reconstruction. A higher value indicates that the reconstructed image is closer to the original image. SSIM measures image similarity, considering brightness, contrast, and structure. A value closer to 1 indicates higher similarity. MAE measures the average absolute difference between predicted and true values. A lower value indicates smaller errors. LPIPS measures perceptual image quality, based on deep learning models. A lower value indicates higher perceptual quality.

From Table1, we can see that our model exceeds the baseline model in a wide range of image object removal tasks, and our model achieves comprehensive outperformance in the three data set performance indicators of PSNR, MAE, and LPIPS, which proves the effectiveness of our method. In particular, the Places dataset is an open scene dataset, which does not have many repeated scenes of a single similar object, so it only validates the validity of our approach under general real-world conditions. There are currently no specialized datasets for situations where a single similar object is repeated many times, and our model is designed specifically for such situations. For the effectiveness of our model at this point, we visualized it in detail in fig7.

## V. CONCLUSION

In this paper, we introduce a novel approach called the Dual-CLIP Guided Diffusion Model to address the challenging problem of object removal in images. Our method innovatively utilizes positive and negative prompts generated by CLIP to guide the diffusion module, facilitating both object removal and seamless background generation. By fine-tuning and retraining the diffusion model, our model not only achieves parity with existing state-of-the-art (SOTA) methods in both natural and realistic environments but also demonstrates superior capabilities in complex scenarios where the object to be removed appears multiple times within the background. This advancement effectively mitigates errors that typically arise from selecting identical backgrounds as the object during the filling process, ensuring more accurate and visually coherent results. Comprehensive experiments have validated that our model performs exceptionally well under these challenging conditions, consistently meeting and often exceeding our expectations. This work represents a significant step forward in the field of image editing, offering a powerful tool for applications requiring precise object removal and background reconstruction.

## VI. References

[1] Z. Li, F. Liu, W. Yang, S. Peng, and J. Zhou, ‘A survey of convolutional neural networks: Analysis, applications, and prospects,’ IEEE Trans. Neural Netw. Learn. Syst., early access, Jun. 10, 2021, doi: 10.1109/TNNLS.2021.3084827.

[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N.Gomez, Kaiser, and I. Polosukhin, “ Attention is all you need,” in Proc. 31st Int. Conf. Neural Information Processing Systems, Long Beach, USA, 2017, pp. 6000–6010.

[3] Z. Liu et al., “Swin transformer: Hierarchical vision transformer using shifted windows,” in Proc. IEEE Int. Conf. Comput. Vis., 2021, pp. 9992– 10002.

[4] F.-A. Croitoru, V. Hondru, R. T. Ionescu, and M. Shah, “Diffusion models in vision: A survey,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 45, no. 9, pp. 10850–10869, Sep. 2023.

[5] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, “Highresolution image synthesis with latent diffusion models,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 10 684–10 695.

[6] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, “Learning transferable visual models from natural language supervision,” in Proc. 38th Int. Conf. Machine Learning, 2021, pp. 8748–8763.

[7] M. J. B. Duff, “Review of the CLIP image processing system,” in Proc. Nut. Comput. Conf., 1978, pp. 1055-1060.

[8] Y. Zeng, J. Fu, H. Chao, and B. Guo, ‘‘Learning pyramid-context encoder network for high-quality image inpainting,’’ in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 1486–1494.

[9] R. Pandey et al., “Total relighting: Learning to relight portraits for background replacement,” ACM Trans. Graph., vol. 40, 2021, Art. no. 43.

[10] Z. Wan, B. Zhang, D. Chen, P. Zhang, D. Chen, F. Wen, and J. Liao, ‘‘Old photo restoration via deep latent space translation,’’ IEEE Trans. Pattern Anal. Mach. Intell., vol. 45, no. 2, pp. 2071–2087, Feb. 2023, doi: 10.1109/TPAMI.2022.3163183.

[11] Y. Jing, Y. Yang, Z. Feng, J. Ye, Y. Yu, and M. Song, ‘‘Neural style transfer: A review,’’ IEEE Trans. Vis. Comput. Graphics, vol. 26, no. 11, pp. 3365–3385, Nov. 2020.

[12] Y. Lei, J. Harms, T. Wang, Y. Liu, H. Shu, A. B. Jani, W. J. Curran, H. Mao, T. Liu, and X. Yang, ‘‘MRI-only based synthetic CT generation using dense cycle consistent generative adversarial networks,’’ Med. Phys., vol. 46, no. 8, pp. 3565–3581, Aug. 2019

[13] S. Olusegun, “Constructivism learning theory: A paradigm for teaching and learning,” J. Res. Method Educ., vol. 5, no. 6, pp. 66–70, 2015, doi: DOI: 10.9790/7388-05616670.

[14] B. Li, K. Q. Weinberger, S. Belongie, V. Koltun, and R. Ranftl, “Language-driven semantic segmentation,” in Proc. IEEE/CVF Int. Conf. Learn. Representations, 2023, pp. 11207–11216.

[15] X. Gu, T.-Y. Lin, W. Kuo, and Y. Cui, “Open-vocabulary object detection via vision and language knowledge distillation,” 2021, arXiv:2104.13921

[16] L. H. Li et al., “Grounded language-image pre-training,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2022, pp. 10955–10965.

[17] Y. Vinker et al., “CLIPasso: Semantically-aware object sketching,” 2022, arXiv:2202.05822

[18] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, “Context encoders: Feature learning by inpainting,” in Proc. Comput. Vision Pattern Recognit., 2016, pp. 2536–2544.

[19] S. Iizuka, E. Simo-Serra, and H. Ishikawa, ‘‘Globally and locally consistent image completion,’’ ACM Trans. Graph., vol. 36, no. 4, pp. 1–14, Jul. 2017.

[20] G. Liu, F. A. Reda, K. J. Shih, T.-C. Wang, A. Tao, and B. Catanzaro, “Image inpainting for irregular holes using partial convolutions,” in Proc. Eur. Conf. Comput. Vision, 2018, pp. 85–100.

[21] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, ‘‘Generative image inpainting with contextual attention,’’ in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 5505–5514.

[22] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. Huang, ‘‘Free-form image inpainting with gated convolution,’’ in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 4471–4480.

[23] Z. Yi, Q. Tang, S. Azizi, D. Jang, and Z. Xu, ‘‘Contextual residual aggregation for ultra high-resolution image inpainting,’’ in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 7505–7514, doi: 10/gg9969.

[24] Y. Zeng, Z. Lin, J. Yang, J. Zhang, E. Shechtman, and H. Lu, “High-resolution image inpainting with iterative confidence feedback and guided upsampling,” in Proc. Eur. Conf. Comput. Vis. (ECCV), Aug. 2020, pp. 1–17

[25] C. Corneanu, R. Gadde, and A. M. Martinez, ‘‘LatentPaint: Image inpainting in latent space with diffusion models,’’ in Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV), Jan. 2024, pp. 4334–4343.

[26] T. Anciukevicius, Z. Xu, M. Fisher, P. Henderson, H. Bilen, N. J. Mitra, and P. Guerrero, ‘‘RenderDiffusion: Image diffusion for 3D reconstruction, inpainting and generation,’’ in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2023, pp. 12608–12618

[27] Ju, Xuan, et al. "Brushnet: A plug-and-play image inpainting model with decomposed dual-branch diffusion." arXiv preprint arXiv:2403.06976 (2024).

[28] Doersch C, Singh S, Gupta A, et al. What makes Paris look like Paris[C].international conference on computer graphics and interact ive techniques,2012, 31(4).

[29] Zhou B, Lapedriza A, Khosla A, et al. Places: A 10 Million Image Database for Scene Recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(6): 1452-1464.

[30] Nazeri, Kamyar, Ng, Eric, Joseph, Tony, et al. EdgeConnect: Generative Image Inpainting with Adversarial Edge Learning[J]. arXiv preprint arXiv:1901.00212, 2019.

[31] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. Globally and

[32] locally consistent image completion. ACM Transactions on Graphics (TOG),pages 107:1–107:14, 2017.

[33] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas SHuang. Free-form image inpainting with gated convolution Proceedings of the IEEE International Conference on Computer Vision. 2019: 4471-4480.