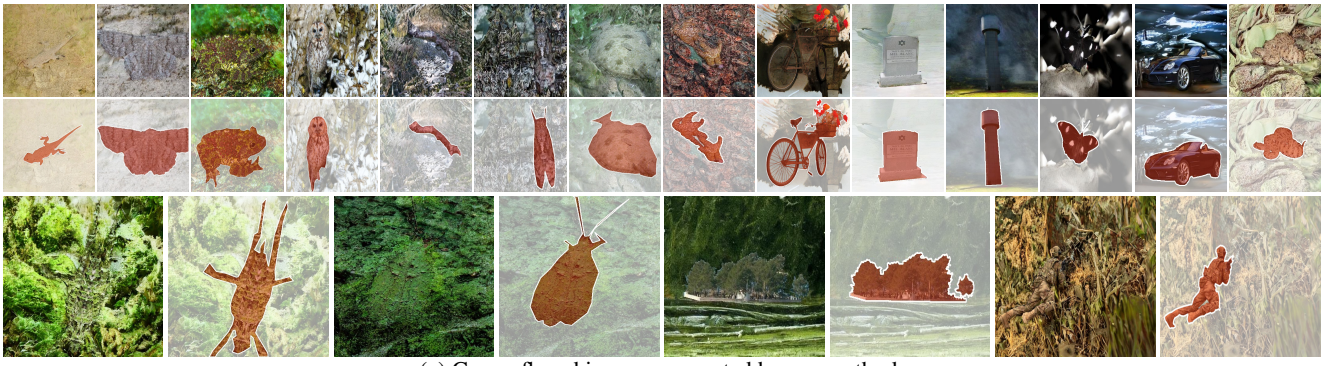
LAKE-RED: Camouflaged Images Generation by Latent Background Knowledge Retrieval-Augmented Diffusion

Pancheng Zhao[1,2] Peng Xu[3*] Pengda Qin[4] Deng-Ping Fan[2,1]
Zhicheng Zhang[1,2] Guoli Jia[1] Bowen Zhou[3] Jufeng Yang[1,2]
[1] VCIP & TMCC & DISSec, College of Computer Science, Nankai University
[2] Nankai International Advanced Research Institute (SHENZHEN· FUTIAN)
[3] Department of Electronic Engineering, Tsinghua University [4] Alibaba Group
pc.zhao99@gmail.com, peng_xu@tsinghua.edu.cn, pengda.qpd@alibaba-inc.com, dengpfan@gmail.com
gloryzzc6@sina.com, exped1230@gmail.com, zhoubowen@tsinghua.edu.cn, yangjufeng@nankai.edu.cn



(a) Camouflaged images generated by our method.

(b) Generating variants for existing camouflaged images. (c) Transferring images from general to camouflaged.

Figure 1. **LAKE-RED synthesizes realistic camouflaged images for a given foreground object by a knowledge retrieval-augmented diffusion model.** Without any human-specified background; the model automatically generates a background sufficient to conceal the foreground objects. (b) and (c) shows the image generation process of our method in two application scenarios.

Abstract

Camouflaged vision perception is an important vision task with numerous practical applications. Due to the expensive collection and labeling costs, this community struggles with a major bottleneck that the species category of its datasets is limited to a small number of object species. However, the existing camouflaged generation methods require specifying the background manually, thus failing to extend the camouflaged sample diversity in a low-cost manner. In this paper, we propose a Latent Background Knowledge Retrieval-Augmented Diffusion (LAKE-RED) for camouflaged image generation. To our knowledge, our contributions mainly include: (1) For the first time, we propose a camouflaged generation paradigm that does not need to receive any background inputs. (2) Our LAKE-RED is the first knowledge retrieval-augmented method with interpretability for camouflaged generation, in which we propose an idea that knowledge retrieval and reasoning enhancement are separated explicitly, to alleviate the task-specific challenges. Moreover, our method is not restricted to specific foreground targets or backgrounds, offering a potential for extending camouflaged vision perception to more diverse domains. (3) Experimental results demonstrate that our method outperforms the existing approaches, generating more realistic camouflage images. ***Our source code is released on https://github.com/PanchengZhao/LAKE-RED.***

*Corresponding Author.

1. Introduction

• **Background.** Camouflaged vision perception [14] is a challenging problem (*e.g.*, camouflaged object detection [13]) aiming to perceive the concealed complex patterns and extensively applied in various fields such as pest detection [11], healthcare [19, 22, 51], and autonomous driving [3, 20, 28, 32, 44]. It has made significant progress in recent years. However, these kinds of overly complex visual scenes and patterns make it extremely time-consuming and labor-intensive to annotate the pixel-wise masks. There is a fact that an instance-level annotation in the COD10K dataset took an average of 60 minutes [12], far longer than the 3 minutes in the COCO-Stuff dataset [5], clearly illustrating this issue. Thus, this community struggles with a major bottleneck in that the species category of its datasets is limited to a small number of object species, *e.g.*, animals.

• **Existing Technical Limitations.** Recently, the rapid development in the AIGC community, particularly generative models based on GAN [8] and Diffusion [18], has revealed the potential of using synthetic data to address data scarcity. DatasetGAN [55] and BigDatasetGAN [27] train a shallow decoder to generate pixel-level annotations from the feature space of pre-trained GANs. DiffuMask [49] is inspired by the attention map in the Diffusion Model and obtains pixel-level annotations from the cross-attention process of the text and image. However, the above method is designed for generic scenarios, and the generated data has a significant domain gap with the training data for the camouflage vision perception task. Moreover, as shown in Fig. 2, the existing camouflaged generation methods require specifying the background manually, thus failing to extend the camouflaged sample diversity in a low-cost manner.

• **Motivation.** Our idea is to make full use of the domain-specific traits of camouflaged scenes to implement a low-cost solution. As shown in Fig. 2, the level of target camouflage depends largely on its surrounding environmental context. Furthermore, we observed that a majority of camouflaged images utilize a background-matching perceptual deception strategy, where the concealed object blends seamlessly into the surrounding background. In this scenario, the foreground and background regions of the camouflaged image exhibit remarkable visual perceptual consistency. For instance, the frog concealed in the grass surface displays a mottled pattern of green and brown just like the grass and ground. This feature convergence between foreground and background makes it possible to retrieve and reason about background through foreground features.

• **Method Overview.** Inspired by the above motivation, we introduce LAKE-RED, a pipeline that automatically generates high-quality camouflage images and pixel-level segmentation masks. The model accepts a foreground object input to achieve object-to-background image inpainting. Specifically, the model first perceives features from the foreground and utilizes them as queries to *retrieve latent background knowledge* from a pre-constructed external knowledge base. Then, the model learns to *reason* from foreground objects to background scenes via using the retrieved knowledge to conduct the camouflaged background reconstruction. This helps the model achieve a richer condition-guided background generation. Simultaneously, this synthesis preserves the precise foreground annotation and prevents boundary blurring caused by mask generation. Fig. 1 illustrates pairs of camouflaged images generated by our LAKE-RED, along with examples of two application scenarios. Without the need for manually specified background inputs, the proposed model can efficiently produce high-quality camouflaged images at a low cost.

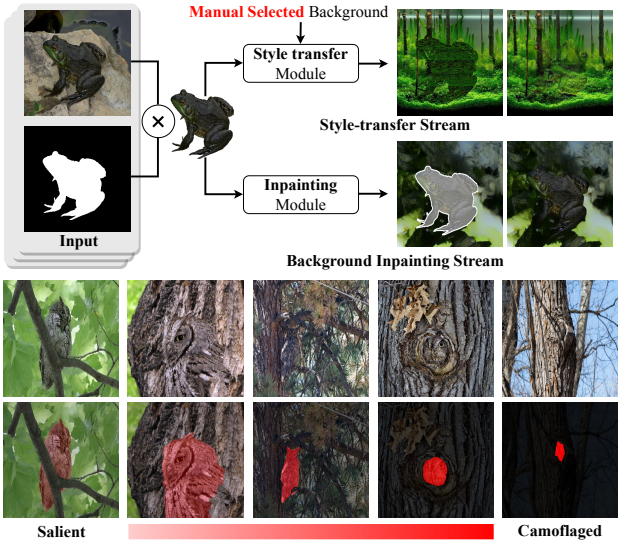


Figure 2. **Comparison of Frameworks for Camouflage Image Generation.** Existing methods rely on manually specified backgrounds, which not only receive limitations in diversity and scope from the human's own cognition but also result in expensive image generation on a large scale. Without changing the texture of itself, the same target can be camouflaged to different degrees in different environments. Inspired by it, we synthesize camouflaged images through a background inpainting stream, hiding by automatically choosing a suitable background for the object.

• **Contribution.** (1) For the first time, we propose a camouflaged generation paradigm without any background inputs. (2) Our LAKE-RED is the first knowledge retrieval-augmented method with interpretability for camouflaged generation, in which we propose an idea that knowledge retrieval and reasoning enhancement are separated explicitly, to alleviate the task-specific challenges. Moreover, our method is not restricted to specific foregrounds or backgrounds, offering a potential for extending camouflaged vision perception to more diverse domains. (3) Experimental results demonstrate our method outperforms the existing approaches, generating more realistic camouflage images.

2. Related Work

• **Synthetic Dataset Generation.** Synthetic data has gained significant attention as one of the primary approaches to tackle data bottlenecks in deep learning methods due to its low cost [24, 40]. Previous research on synthetic datasets has mainly focused on producing high-quality simulated scenes in 3D environments and generating data from them, which has been extensively employed for tasks such as recognition [23, 46, 47, 62], segmentation [6, 34, 48, 57], object tracking [33, 58], image and video understanding [52, 56, 59–61, 63], optical flow estimation [4, 35], and 3D reconstruction [66–68]. The considerable disparity between the distribution of synthetic data through simulated scenarios and real data restricts their validity. Significant progress in generative modeling has recently enabled the reduction of the domain gap between synthetic and real data. With realistic image data generated by advanced generative models (*e.g.*, GAN, DALL-E2, and Stable Diffusion), some research has attempted to investigate the potential of synthetic data as a replacement for real data [15, 16, 30]. Specifically, DatasetGAN [55] and BigDatasetGAN [27] excel in generating a significant quantity of synthetic images with segmentation masks with limited labeled data. On the other hand, Diffumask [49] relies exclusively on textual supervision to extract semantic labels from the cross-attention maps of text and images.

• **Camouflage Image Generation.** Camouflage images are different from regular images as they contain one or more concealed objects [12]. Although the concept of camouflage can be traced back to Darwin's theory of evolution [9, 39, 42] and has long been used in various fields, the task of camouflage image generation was not proposed until 2010 by Chu *et al.* [7]. The proposed model gets a specified foreground and background as input and uses hand-crafted features to give the foreground textural details similar to the background, making the concealed objects difficult for humans to recognize. Recent advancements in deep learning methods for style transfer and image composing have provided new ideas for generating camouflage images. Subsequent models, such as Zhang's [53] and Li's [29], have further improved camouflage image generation by composing the foreground with the background through style transfer and structure alignment. However, the use of artificially specified backgrounds increases the cost of data acquisition and limits the diversity of generated images due to human cognitive limitations. These limitations make it impossible to generate large-scale datasets, greatly reducing the application value of the generated images.

3. Methodology

Our objective is to generate camouflaged images by automatically complementing the background region for a specific foreground object, resulting in a realistic image where the object is concealed in the generated background. While there have been advancements in camouflage image generation methods, manually specifying the background is not practical due to the high human cost and limited cognitive range. Through our observation of the camouflage phenomenon, we have noticed that the background region of a camouflaged image often shares similar image features with the surface of the foreground object. This suggests that a suitable camouflage background may already exist within the foreground image itself. Formally, given a source image $\mathbf{x}_s \in \mathbb{R}^{H\times W\times 3}$, containing an object with an irregular shape. The object's location is precisely indicated by a binary mask $\mathbf{m}$ with the same size as the original image $\mathbf{x}_s$, where $\mathbf{m}_{i,j} = 0$, with $i \in [0, H]$ and $j \in [0, W]$, represents the object region that needs to be maintained in subsequent operations, and $\mathbf{m}_{i,j} = 1$ represents the editable background region. The model takes $\{\mathbf{x}_s, \mathbf{m}\}$ as input, and outputs a camouflaged image $\mathbf{x}_c$. The objective is to obtain a prior from the foreground $\mathbf{x}_s \odot \bar{\mathbf{m}}$ to generate a suitable background that replaces the original one. The foreground should harmoniously match the new background.

3.1. Preliminaries

• **Revisiting Latent Diffusion Models.** Aiming to generate high-quality camouflage images, our proposed method is based on classic Latent Diffusion Models (LDM) [41]. Similar to other probabilistic models, LDM learns the probability distribution $p(x)$ of a given image set x through self-supervised training and achieves high-quality image generation by reversing a Markov forward process. Specifically, the forward process adds a sequence noise to the original images $\mathbf{y}_0 = \mathbf{x_s}$ to obtain a noisy image $\{\mathbf{y}_t \mid t \in [1, T]\}$, where $\mathbf{y}_t = \alpha_t \mathbf{y}_0 + (1 - \alpha_t)\epsilon$. As α_t decreases with time step t, more Gaussian noise ϵ is introduced into $\mathbf{y}_0$. The generation process can be described as a sequence of denoising autoencoders $\epsilon_\theta(\mathbf{y}_t, \mathbf{c}, t)$ to predict a denoised variant of input $\mathbf{y}_t$. Furthermore, in order to decrease the computational demands of high-resolution image synthesis for the model, a pre-trained autoencoder ε is employed to encode $\mathbf{y}$ into a latent representation $\mathbf{z} = \varepsilon(\mathbf{y})$, where $\mathbf{z} \in \mathbb{R}^{h\times w\times c}$. So the training objective can be defined as the following loss function:

$$\mathcal{L} = \mathbb{E}_{t,\varepsilon(\mathbf{y}),\epsilon} \left\| \epsilon_\theta(\mathbf{z}_t, \mathbf{c}, t) - \epsilon \right\|_2^2 . \tag{1}$$

For the inpainting stream, the condition $\mathbf{c}$ includes $\mathbf{x}_s \odot \bar{\mathbf{m}}$ to indicate the remaining area. Once T steps have been completed, the model predicts the latent representation $\mathbf{z}'_0$, of which the noise ϵ has been entirely removed.

Finally, to reconstruct a high-resolution image from the latent representation, a VQVAE [43] based decoder $\mathcal{D}$ is utilized in the final stage. The visual information from the code book $\mathbf{e}$ is embedded into the latent representation by incorporating a quantization layer ν into the decoder, which

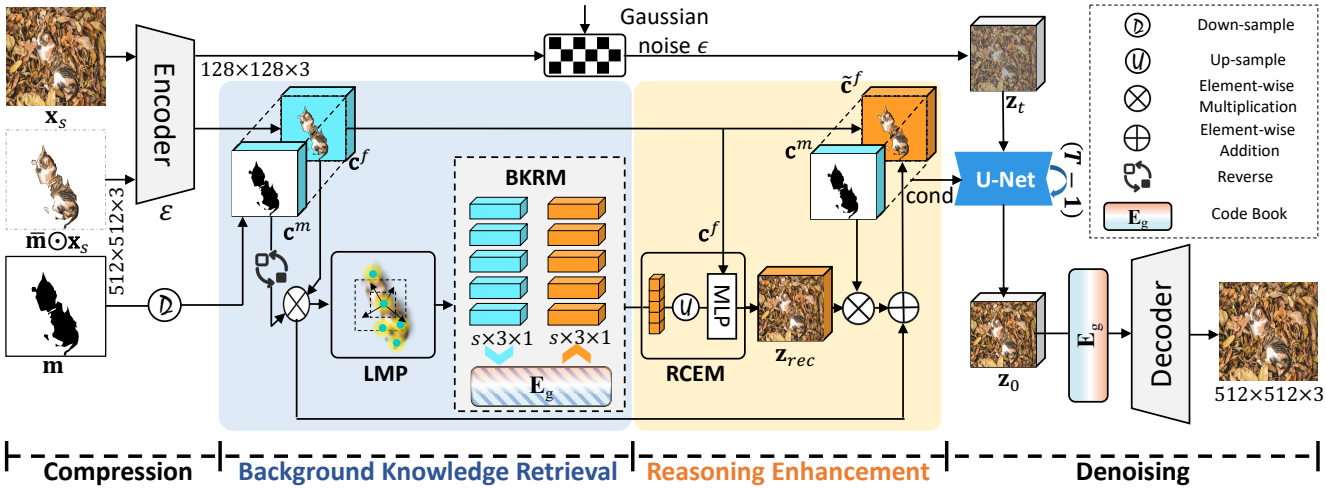


Figure 3. **The pipeline of our camouflaged images generation framework LAKE-RED.** Our framework mainly includes three steps: (1) Extracting visual representations of foreground areas by Localized Masked Pooling (LMP). (2) The Background Knowledge Retrieval Module (BKRM) is utilized to retrieve background-related features from the codebook. (3) The Reasoning-Driven Condition Enhancement module (RCEM) allows the model to learn foreground-to-background reasoning through a background reconstruction.

can be yielded as:

$$\mathbf{y}_0' = \mathcal{D}\left(\nu(\mathbf{e}, \mathbf{z}_0')\right), \tag{2}$$

where $\mathbf{e} \in \mathbb{R}^{K\times D}$, K, and D denote the size of the discrete latent space and the dimensionality of each latent embedding vector, respectively.

3.2. Model Designs

Current image inpainting methods accept a conditional input $\mathbf{c}$ that includes known image regions and indicates editable regions, which can be defined as:

$$\begin{aligned} \mathbf{c}^f, \mathbf{c}^m &= \varepsilon(\mathbf{I}_{known}), downsample(\mathbf{m}, f), \\ \mathbf{c} &= Concat\left(\mathbf{c}^f, \mathbf{c}^m\right), \end{aligned} \tag{3}$$

where $\mathbf{I}_{known} = \mathbf{x}_s \odot \bar{\mathbf{m}}$, and $\mathbf{m}$ is down sampled by a factor $f = 2^n$, with $n \in \mathbb{N}$. However, they tend to prioritize preserving the structural continuity of the object in the image and infer to fill in the missing areas. The inference of the model is constrained when the non-edited region forms a complete object that lacks structural continuity with the background. This means that the current condition is not enough to facilitate the model in making accurate inferences from the foreground object to the background scene. To mitigate the negative impact of this performance bottleneck on the results, as shown in Fig. 3, we focus on retrieving richer background knowledge and develop a reasoning-based background reconstruction task that enables the model to explicitly learn the relationship between the foreground and background of a camouflaged image. The reconstructed features can then be used to enhance existing conditions and provide the model with richer guidance information.

3.2.1 Background Knowledge Retrieval

As mentioned before, inferring from object to background is a significant challenge for image inpainting models. However, unlike general images, camouflage images are primarily characterized by background matching, where the background and the object exhibit a high degree of consistency in terms of texture. This implies that it becomes feasible to retrieve background knowledge using foreground features. The training framework for reconstructing backgrounds through masked ground truth (GT) implicitly models the relationship between the object and background, which results in the model paying insufficient attention to the texture consistency of the object and background. Explicitly retrieving background features aligned with the object features is a viable option to provide richer guidance for the denoising process. In order to obtain feature representations about the background texture, we take inspiration from the autoencoder and decoder used by LDM, which is based on VQ-VAE.

VQ-VAE constructs a code book $\mathbf{e}$ in the embedding space between the encoder and the decoder during the training process. The codebook can be injected with features into the representation of the latent space by vector quantized operation before the decoder to obtain a better reconstruction performance. To address the issue of missing background features of the condition, the pre-trained codebook is replicated and shifted to the denoising process as a global visual embedding $\mathbf{E}_g = \mathbf{e}^{\mathrm{T}} \in \mathbb{R}^{D\times K}$. The process of obtaining background features $\mathbf{x}^b$ using a latent space codebook E_g can be summarized as:

$$\begin{aligned} \mathbf{x}^b &= \text{Concat}(\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_H)\mathbf{W}^{f\to b}, \\ \mathbf{h}_i &= a_i \mathbf{x}^f \mathbf{W}_i^V, \\ a_i &= \text{softmax}\left(\frac{\left[\mathbf{x}^f \mathbf{W}_i^Q\right] \cdot \left[\mathbf{E}_g \mathbf{W}_i^K\right]^{\mathrm{T}}}{\sqrt{d_k}}\right). \end{aligned} \tag{4}$$

We feed the foreground feature $\mathbf{x}^f$ into the Multi-Head At-

tention (MHA) layer with H heads, as the query, for retrieving the related background content from codebook $\mathbf{E}_g$, and obtain the background aligned visual feature $\mathbf{x}^b$.

3.2.2 Localized Masked Pooling

We introduce a simple and efficient latent background knowledge retrieval module, denoted as $\mathcal{B}\left(\mathbf{x}^f, \mathbf{E}_g\right)$, that retrieves background-aligned visual features $\mathbf{x}^b$ from codebook $\mathbf{E}_g$ using foreground features $\mathbf{x}^f$. The richness of the feature representation $\mathbf{x}^f$ extracted from $\mathbf{c}^f$ directly impacts the validity of features that can be retrieved from the codebook. Thus, the foreground feature representation $\mathbf{x}_f$ can become another potential performance bottleneck. To exclude features in the background region during feature extraction, a straightforward approach is to follow [54] using Masked Averaged Pooling (MAP), to obtain representative vectors of foreground features as:

$$\mathbf{x}_i^f = \Phi\left(\mathbf{c}_i^f, \mathbf{c}^m\right) = \frac{\sum_{x=1,y=1}^{w,h} \mathbf{c}_{i,x,y}^f * \bar{\mathbf{c}}_{x,y}^m}{\sum_{x=1,y=1}^{w,h} \bar{\mathbf{c}}_{x,y}^m}, \tag{5}$$

where $i \in \{1, 2, \ldots, \vartheta\}$ indicates the channel number. The MAP treats the foreground as a whole and compresses it into a unified representation, which can lead to a significant loss of information. In particular, the encoder $\varepsilon(\cdot)$ maintains the channel number of the feature to be 3, resulting in $\mathbf{x}^f \in \mathbb{R}^{3\times 1}$. This simple representation is insufficient to capture the rich features of the foreground and can limit the effectiveness of latent background knowledge retrieval.

Foreground objects in camouflaged images often display intricate visual features, which we define as a combination of s sub-features. The higher the value of s, the more intricate and detailed the corresponding feature is. To extract richer foreground features, we shift our focus from global to local and employ the SLIC algorithm [1] to cluster the foreground regions into s superpixels. The above process can be reformulated as:

$$\mathbf{p}_1^i, \mathbf{p}_2^i, \cdots, \mathbf{p}_s^i = \mathcal{S}(\mathbf{c}_i^f, \mathbf{c}^m),$$
$$\mathbf{x}_{i,j}^f = \Phi_s\left(\mathbf{p}_j^i, \mathbf{c}^m\right) = \frac{\sum_{x=1,y=1}^{w,h} \mathbf{c}_{i,x,y}^f * \mathbf{p}_{j,x,y}^i}{\sum_{x=1,y=1}^{w,h} \mathbf{p}_{j,x,y}^i}. \tag{6}$$

3.2.3 Reasoning-Driven Condition Enhancement

Additionally, we upsample the obtained background knowledge features $\mathbf{x}^b$ and combine them with the foreground features $\mathbf{c}^f$ to reconstruct the GT image features $\mathbf{z}_0 = \varepsilon(\mathbf{y}_0), \mathbf{z}_0 \in \mathbb{R}^{h\times w\times c}$. The reconstruction feature can be computed as:

$$\mathbf{z}_{rec} = \mathbf{MLP}(Concat(\mathbf{c}^f, upsample(\mathbf{x}^b, f))). \tag{7}$$

Then, $\mathbf{z}_{rec}$ is utilized to refine the initial condition of the input. To emphasize the background features, we created a feature reconstruction task that enhances the model's ability to reason about real background features using background knowledge. Specifically, we populate the background region of $\mathbf{c}^f$ with the reconstructed $\mathbf{z}_{rec}$ to strengthen the information embedded in the condition while reserving the foreground areas. The strategy for enhancing the condition can be formulated as:

$$\begin{aligned} \tilde{\mathbf{c}}^f &= \mathbf{c}^f \cdot (1 - \mathbf{c}^m) + \mathbf{z}_{rec} \cdot \mathbf{c}^m, \\ \tilde{\mathbf{c}} &= Concat\left(\tilde{\mathbf{c}}^f, \mathbf{c}^m\right). \end{aligned} \tag{8}$$

For the loss of background reconstruction, we have:

$$\mathcal{L}_{bgrec} = \frac{1}{h \times w} \sum_{i=1}^{h} \sum_{j=1}^{w} (\mathbf{z}_{rec} \cdot \mathbf{c}^m - \mathbf{z}_0 \cdot \mathbf{c}^m)^2. \tag{9}$$

Then, the overall loss can be reformulated as:

$$\begin{aligned} \mathcal{L} &= \mathcal{L}_{diff} + \mathcal{L}_{bgrec} \\ &\propto \left\|\epsilon_\theta\left(\mathbf{z}_t, \tilde{\mathbf{c}}, t\right) - \epsilon\right\|_2^2 + \left\|\mathbf{z}_{rec} \cdot \mathbf{c}^m - \mathbf{z}_0 \cdot \mathbf{c}^m\right\|^2. \end{aligned} \tag{10}$$

By leveraging the properties of the camouflaged image, we refine and enhance the input condition $\mathbf{c}$. While defining the image features of the foreground area, the enhanced condition $\tilde{\mathbf{c}}$ guides the generation of background. The implicit and explicit constraints work together to help the model learn the texture consistency between the foreground object and the background, resulting in high-quality camouflage image generation.

4. Experiments

4.1. Experimental Setups

• **Datasets.** Following the previous works [13] for COD, 4,040 images (3,040 from COD10K [12], 1,000 from CAMO [26]) are used as real data for training the model. To verify the generative performance, we collected image-mask pairs from various fields to construct a test data set, including three subsets: Camouflaged Objects (CO), Salient Objects (SO), and General Objects (GO). In CO, there are 6,473 pairs of images from CAMO [26], COD10K [12], and NC4K [38]. Then we randomly selected 6473 images from the well-known salient object detection datasets (DUTS [45], DUT-OMRON [50], etc.) and the segmentation dataset (COCO2017 [31]) to evaluate the performance of the model on open domain data.

• **Metrics.** Following the good practices of AIGC [27, 41] and COD [25, 37], we choose the InceptionNet-based metrics FID [2] and KID [17] to measure the quality of generated camouflaged images. Once the image features are extracted by InceptionNet, the distance between them is computed to indicate the level of resemblance between the model's output and the target dataset. Different from the general images, well-synthesized camouflaged images should not include easily identifiable objects, and it is more challenging to extract discriminative features [37]. A smaller value indicates that the generated image is more similar to the real camouflaged image.

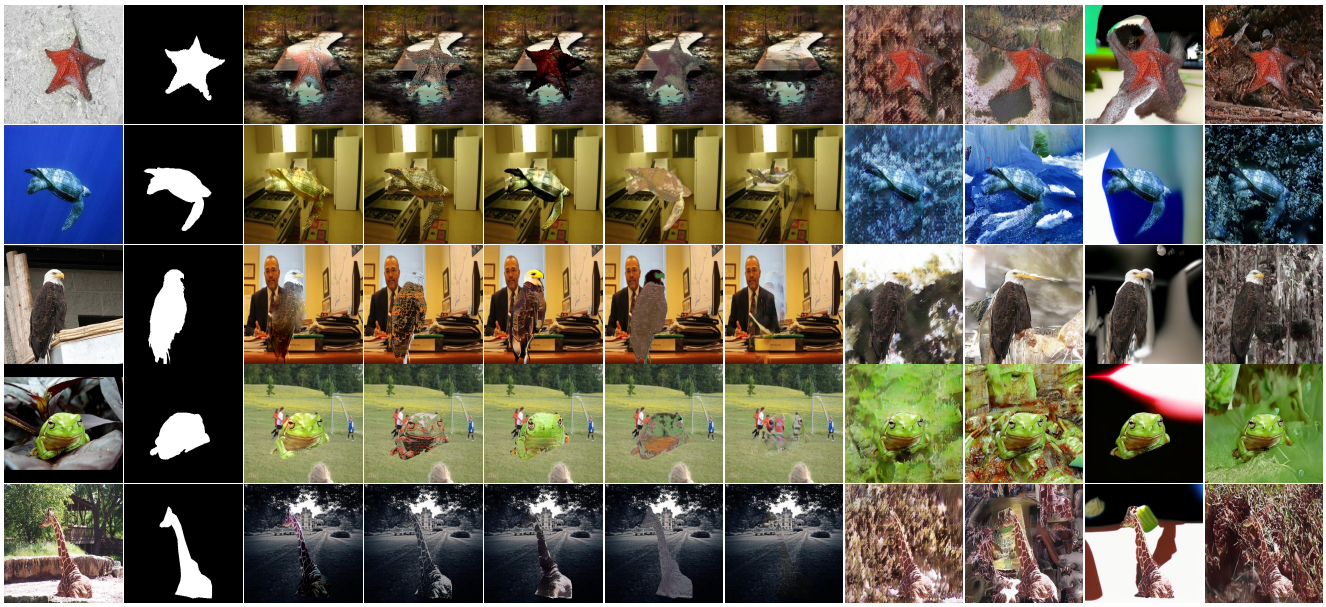


Figure 4. **Comparison with existing methods in transferring general images into camouflaged images.** The first two columns are the input images and we provide camouflaged images generated by nine methods for the comparison. Note that the methods in columns 3 to 7 additionally share a randomly sampled background image as input.

• **Implementation Details.** To generate camouflaged images by given foreground images, we utilize a powerful Latent Diffusion Model [41] pre-trained in the inpainting task as initialization. The model is designed to handle images and masks of size 512×512 and is compressed to a latent space of $128 \times 128 \times 3$ using a pre-trained VQ-VAE [43]. During training, we focus on training the denoising U-Net and do not fine-tune the auto-encoder and decoder components. We refine and enhance the existing condition through the proposed module in this paper. The parameters optimization such as initialization, data augmentation, and batch size are set similar to the original paper. Finally, the model generates the camouflaged image and resizes it to align with the original input. We conduct all the experiments by PyTorch and GeForce RTX 3090 GPUs are used for all experiments.

4.2. Comparison with the State-of-the-art Methods

Previous camouflage image generation methods are based on image blending or style transfer, which differ fundamentally from the method proposed in this paper. Thus, for each solution, we select cutting-edge methods for comparison. For the image blending and style transfer schemes, the model requires a manually specified background image when accepting a foreground input. We used Places365 [65], a large-scale scene dataset, as the source of background images. For a given foreground input, we randomly sampled a background image from Places365, resized it, and then performed image synthesis process. To facilitate comparison between different methods, all methods shared the same background image for a given foreground input. For the image inpainting scheme, the model only accepts one foreground input and generates a camouflaged image as output.

• **Qualitative analysis.** Fig. 4 presents a comparison of the quality of camouflaged images generated by our method and other methods from a general image. The results show that methods such as AB and CI are highly influenced by the background image input, despite the foreground features being processed to align with the background. As a result, the foreground exhibits conflicts with the background scenes and objects, such as the eagle and turtle in the room (2nd and 3rd rows), and the larger-than-life frog (4th row). LCGNet performs the best in hiding the objects, with their features being barely visible. Camouflaged objects in nature are seamlessly embedded in the background rather than being completely invisible. On the other hand, image inpainting methods only require foreground object input and adaptive background generation can meet the requirements of large-scale generation. However, existing methods suffer from issues such as lack of authenticity of the background (TFill), low degree of camouflage (LDM), and failure of background complementation (Repaint-L). In contrast, our method naturally integrates the given target into the generated background, preserving all the target's features while achieving overall camouflage of the image.

• **Quantitative analysis.** A large-scale test set is constructed to evaluate the quality of camouflage image generation, which includes three types of foreground objects to assess the model's adaptability to different image domains. The salient objects subset and the general objects subset are sampled from datasets in the salient object detection and

Table 1. **Quantitative performance.** The proposed camouflaged image generation method is subjected to a quantitative evaluation, wherein it is compared with state-of-the-art (SOTA) methods. The evaluation involved specific foreground objects sampled from camouflaged images, salient images, and general images. The proposed method shows excellent performance.

Methods		Input	Camouflaged Objects		Salient Objects		General Objects		Overall	
			FID↓	KID↓	FID↓	KID↓	FID↓	KID↓	FID↓	KID↓
Image Blending	AB [10]$_{03}$	$\mathcal{F}+\mathcal{B}$	117.11	0.0645	126.78	0.0614	133.89	0.0645	120.21	0.0623
	CI [7]$_{10}$	$\mathcal{F}+\mathcal{B}$	124.49	0.0662	136.30	0.7380	137.19	0.0713	128.51	0.0693
	AdaIN [21]$_{17}$	$\mathcal{F}+\mathcal{B}$	125.16	0.0721	133.20	0.0702	136.93	0.0714	126.94	0.0703
	DCI [53]$_{20}$	$\mathcal{F}+\mathcal{B}$	130.21	0.0689	134.92	0.0665	137.99	0.0690	130.52	0.0673
	LCGNet [29]$_{22}$	$\mathcal{F}+\mathcal{B}$	129.80	0.0504	136.24	0.0597	132.64	**0.0548**	129.88	0.0550
Image Inpainting	TFill [64]$_{22}$	$\mathcal{F}$	63.74	0.0336	96.91	0.0453	122.44	0.0747	80.39	0.0438
	LDM [41]$_{22}$	$\mathcal{F}$	58.65	0.0380	107.38	0.0524	129.04	0.0748	84.48	0.0488
	RePaint-L [36]$_{22}$	$\mathcal{F}$	76.80	0.0459	114.96	0.0497	136.18	0.0686	96.14	0.0498
	Ours$_{23}$	$\mathcal{F}$	**39.55**	**0.0212**	**88.70**	**0.0428**	**102.67**	0.0625	**64.27**	**0.0355**

image segmentation domains, respectively, with the number of images kept consistent with the COD test set. The distance between the generated results and the real COD benchmarks is measured using FID and KID, and the results are presented in Tab. 1.

The results on the three subsets display a step-wise distribution, indicating that the model performance was strongly influenced by the image domain gap, with general objects being more challenging to transform than salient objects. The image blending-based methods produce large results because they mechanically shift the foreground features towards being consistent with the background features, resulting in image visual features that are primarily determined by the background image. When the background image is randomly sampled, the related indexes also exhibit some degree of randomness. On the other hand, image inpainting-based schemes tended to generate a suitable background for the object and generally show better performance.

In addition, we observe outliers in the validation results of LCGNet on the subset of General Objects, which are caused by a combination of the following reasons. First, the difficulty of synthesizing increases in three subsets. The camouflaged object comes from a concealed scene and is easy to hide. The salient object is of moderate size and position and usually has a complete structure. The general object has a rich variety of classes and diverse sizes, making it challenging to find suitable camouflage environments for it. As the complexity rises, these approaches progressively struggle to conceal general objects flawlessly, leading to a decline in performance within that particular subset. In this case, LCGNet maximally discards foreground features, and the results mainly depend on the randomly sampled backgrounds (Fig. 4). It is least affected by the negative influence from the foreground and is introduced to randomness by the background, thus resulting in anomalous results. However, our method achieved optimal performance on the overall test set.

• **User Study.** Since both image generation quality and camouflage effectiveness require human perception, we conducted user studies to obtain subjective human judgments on the generated results. To this end, we followed the previous work on camouflage image generation to randomly select 20 sets of foreground images and applied various methods to generate the results. For style transfer-based methods, we used an additional image randomly sampled from Places365 as the background input, which was kept consistent for all methods. We invited 20 participants to rate the results based on three questions:

-Q#1: Which result is the hardest to find?
-Q#2: Which is the most visually natural result?
-Q#3: Which result appears closest to the real camouflaged image dataset?

For each question, participants need to select their top 3 choices, with 1 being the highest. The results of the user survey are presented in Fig. 5. Although LCGNet received more votes in **Q#1** due to the almost invisible foreground in the generated results, our method was considered to produce more natural and visually closer results to the real dataset in terms of visual presentation.

Table 2. **Quantitative Ablation study.** We progressively add each module to the base model to compare their impact on the quality of the generated results and costs. The result shows that the method we proposed is effective and almost cost-free.

Module			Prams(M)↓	MAC(G)↓	FPS(Hz)↑	Overall	
BKRM	RCEM	LMP				FID↓	KID↓
✗	✗	✗	440.46	577.97	0.2482	96.14	0.0498
✓	✗	✗	440.47	577.99	0.2442	69.80	0.0417
✓	✓	✗	440.47	577.99	0.2438	69.52	0.0412
✓	✓	✓	440.47	577.99	0.2008	64.27	0.0355

4.3. Ablation Study

We conduct the ablation study by gradually adding modules to the base LDM to evaluate the effectiveness of each component in our proposed method. As shown in Tab. 2, the quality of the generated camouflage images gradually improves with the introduction of the modules proposed in this paper, demonstrating the effectiveness. When all three modules are applied simultaneously, the model performance reaches its peak, achieving improvements of 33.14%

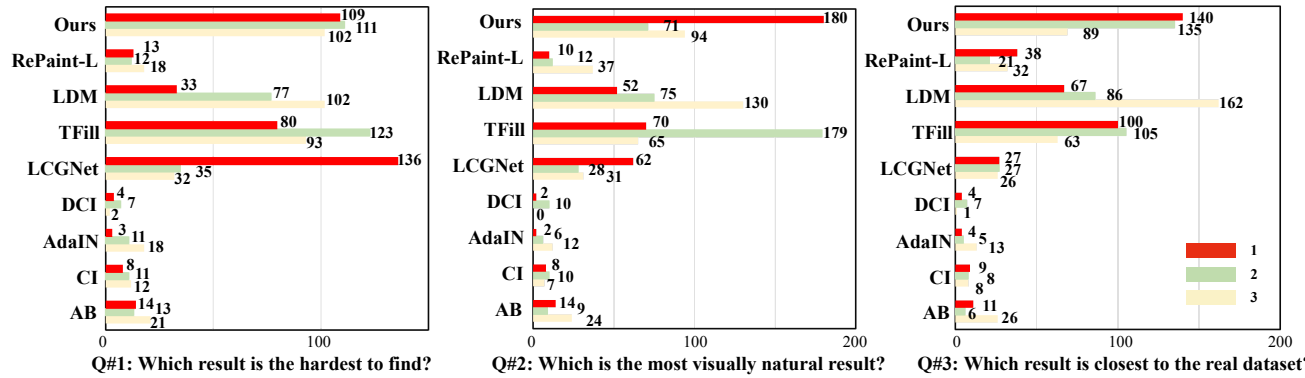


Figure 5. **User study about subjective ratings of the camouflaged image generated by 9 different methods.** Our method is considered to produce the most natural and visually closest results to the real camouflage image.

Figure 6. **The visualization of ablation study.** We visualize the samples during the ablation experiments to visualize the effectiveness of the modules we proposed.

and 28.71% in the FID and KID metrics, respectively. At this point, the introduction of the three modules only adds about 0.01M parameters and 0.02G of computation to the model, with the inference speed reduced by only 0.04Hz. These results clearly indicate that our method is effective and comes at almost no additional cost.

We further visualize the samples during the ablation experiments to show the effectiveness of these modules. Two sets of results are shown in Fig. 6. The LDM faces challenges in focusing on the camouflage properties during inpainting from the foreground object to the background. It also struggles to generate the background in certain regions, resulting in black color blocks due to the complexity of the task. By incorporating a latent background knowledge retrieval module (BKRM), the model is explicitly constrained to learn foreground and background similarity, resulting in a closer alignment of the generated background with the foreground. Furthermore, the reasoning-driven condition enhancement module (RCEM) enhances the realism of scenes by incorporating a background reconstruction loss that compels the model to reason and reconstruct the background features accurately. Finally, the introduction of localized masked pooling (LMP) shifted the model's attention from global to local foreground features, enhancing the texture diversity of the generated background.

5. Conclusion

We propose a latent background knowledge retrieval-augmented diffusion (LAKE-RED) for camouflaged image generation. Unlike existing methods, our generation paradigm is background-free. By knowledge retrieval and reasoning enhancement, we get a strong background condition from the foreground, resulting in synthetic images that surpass those generated by other SOTA camouflaged image generation methods. Our approach is not restricted to specific foreground targets or human-selected backgrounds. This enables us to generate camouflage images on a large scale and offers the potential for extending camouflaged vision perception to more diverse domains in the future.

6. Acknowledgments

This work was supported by the Natural Science Foundation of Tianjin, China (NO.20JCJQJC00020), the National Natural Science Foundation of China (NO.62306162), Fundamental Research Funds for the Central Universities, and Supercomputing Center of Nankai University (NKSC).

References

[1] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Süsstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *TPAMI*, 34 (11):2274–2282, 2012. 5

[2] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. In *ICLR*, 2018. 5

[3] Keenan Burnett, Sepehr Samavi, Steven Waslander, Timothy Barfoot, and Angela Schoellig. autotrack: A lightweight object detection and tracking system for the sae autodrive challenge. In *CRV*, 2019. 2

[4] Daniel J Butler, Jonas Wulff, Garrett B Stanley, and Michael J Black. A naturalistic open source movie for optical flow evaluation. In *ECCV*, 2012. 3

[5] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Cocostuff: Thing and stuff classes in context. In *CVPR*, 2018. 2

[6] Yajie Chen, Xin Yang, and Xiang Bai. Confidence-weighted mutual supervision on dual networks for unsupervised cross-modality image segmentation. *SCIS*, 66(11):210104, 2023. 3

[7] Hung-Kuo Chu, Wei-Hsin Hsu, Niloy J Mitra, Daniel Cohen-Or, Tien-Tsin Wong, and Tong-Yee Lee. Camouflage images. *TOG*, 29(4):51–1, 2010. 3, 6, 7

[8] Antonia Creswell, Tom White, Vincent Dumoulin, Kai Arulkumaran, Biswa Sengupta, and Anil A Bharath. Generative adversarial networks: An overview. *SPM*, 35(1):53–65, 2018. 2

[9] IC Cuthill. Camouflage. *J ZOOL*, 308(2):75–92, 2019. 3

[10] J. Matías Di Martino, Gabriele Facciolo, and Enric Meinhardt-Llopis. Poisson Image Editing. *IPOL*, 6:300–325, 2016. 6, 7

[11] MA Ebrahimi, Mohammad Hadi Khoshtaghaza, Saeid Minaei, and Bahareh Jamshidi. Vision-based pest detection based on svm classification method. *COMPAG*, 137:52–58, 2017. 2

[12] Deng-Ping Fan, Ge-Peng Ji, Guolei Sun, Ming-Ming Cheng, Jianbing Shen, and Ling Shao. Camouflaged object detection. In *CVPR*, 2020. 2, 3, 5

[13] Deng-Ping Fan, Ge-Peng Ji, Ming-Ming Cheng, and Ling Shao. Concealed object detection. *TPAMI*, 44(10):6024–6042, 2021. 2, 5

[14] Deng-Ping Fan, Ge-Peng Ji, Peng Xu, Ming-Ming Cheng, Christos Sakaridis, and Luc Van Gool. Advances in deep concealed scene understanding. *VI*, 1(1):16, 2023. 2

[15] Yunhao Ge, Jiashu Xu, Brian Nlong Zhao, Laurent Itti, and Vibhav Vineet. Dall-e for detection: Language-driven context image synthesis for object detection. *arXiv preprint arXiv:2206.09592*, 2022. 3

[16] Ruifei He, Shuyang Sun, Xin Yu, Chuhui Xue, Wenqing Zhang, Philip Torr, Song Bai, and Xiaojuan Qi. Is synthetic data from generative models ready for image recognition? In *ICLR*, 2023. 3

[17] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, 2017. 5

[18] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 2

[19] Duojun Huang, Xinyu Xiong, De-Jun Fan, Feng Gao, Xiao-Jian Wu, and Guanbin Li. Annotation-efficient polyp segmentation via active learning. *arXiv preprint arXiv:2403.14350*, 2024. 2

[20] Duojun Huang, Xinyu Xiong, Jie Ma, Jichang Li, Zequn Jie, Lin Ma, and Guanbin Li. Alignsam: Aligning segment anything model to open context via reinforcement learning. In *CVPR*, 2024. 2

[21] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, 2017. 6, 7

[22] Ge-Peng Ji, Jing Zhang, Dylan Campbell, Huan Xiong, and Nick Barnes. Rethinking polyp segmentation from an out-of-distribution perspective. *MIR*, pages 1–9, 2024. 2

[23] Guoli Jia and Jufeng Yang. S 2-ver: Semi-supervised visual emotion recognition. In *ECCV*, 2022. 3

[24] Amlan Kar, Aayush Prakash, Ming-Yu Liu, Eric Cameracci, Justin Yuan, Matt Rusiniak, David Acuna, Antonio Torralba, and Sanja Fidler. Meta-sim: Learning to generate synthetic datasets. In *ICCV*, 2019. 3

[25] Hala Lamdouar, Weidi Xie, and Andrew Zisserman. The making and breaking of camouflage. In *ICCV*, 2023. 5

[26] Trung-Nghia Le, Tam V Nguyen, Zhongliang Nie, Minh-Triet Tran, and Akihiro Sugimoto. Anabranch network for camouflaged object segmentation. *CVIU*, 184:45–56, 2019. 5

[27] Daiqing Li, Huan Ling, Seung Wook Kim, Karsten Kreis, Sanja Fidler, and Antonio Torralba. Bigdatasetgan: Synthesizing imagenet with pixel-wise annotations. In *CVPR*, 2022. 2, 3, 5

[28] Jiaming Li, Jiacheng Zhang, Jichang Li, Ge Li, Si Liu, Liang Lin, and Guanbin Li. Learning background prompts to discover implicit knowledge for open vocabulary object detection. In *CVPR*, 2024. 2

[29] Yangyang Li, Wei Zhai, Yang Cao, and Zheng-Jun Zha. Location-free camouflage generation network. *TMM*, 25: 5234–5247, 2023. 3, 6, 7

[30] Ziyi Li, Qinye Zhou, Xiaoyun Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Open-vocabulary object segmentation with diffusion models. In *ICCV*, 2023. 3

[31] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 5

[32] Gengxin Liu, Oliver van Kaick, Hui Huang, and Ruizhen Hu. Active self-training for weakly supervised 3d scene semantic segmentation. *CVMJ*, pages 1–14, 2024. 2

[33] Xin Liu and Jufeng Yang. Progressive neighbor consistency mining for correspondence pruning. In *CVPR*, 2023. 3

[34] Xianglong Liu, Shihao Bai, Shan An, Shuo Wang, Wei Liu, Xiaowei Zhao, and Yuqing Ma. A meaningful learning method for zero-shot semantic segmentation. *SCIS*, 66(11): 210103, 2023. 3

[35] Xin Liu, Guobao Xiao, Riqing Chen, and Jiayi Ma. Pgfnet: Preference-guided filtering network for two-view correspondence learning. *TIP*, 32:1367–1378, 2023. 3

[36] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *CVPR*, 2022. 6, 7

[37] Xue-Jing Luo, Shuo Wang, Zongwei Wu, Christos Sakaridis, Yun Cheng, Deng-Ping Fan, and Luc Van Gool. Camdiff: Camouflage image augmentation via diffusion. *AIR*, 2: 9150021, 2023. 5

[38] Yunqiu Lv, Jing Zhang, Yuchao Dai, Aixuan Li, Bowen Liu, Nick Barnes, and Deng-Ping Fan. Simultaneously localize, segment and rank the camouflaged objects. In *CVPR*, 2021. 5

[39] Sami Merilaita, Nicholas E Scott-Samuel, and Innes C Cuthill. How camouflage works. *Philos T R Soc B*, 372 (1724):20160341, 2017. 3

[40] Alhassan Mumuni, Fuseini Mumuni, and Nana Kobina Gerrar. A survey of synthetic data augmentation methods in machine vision. *MIR*, pages 1–39, 2024. 3

[41] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 3, 5, 6, 7

[42] Martin Stevens and Sami Merilaita. Animal camouflage: current issues and new perspectives. *Philos T R Soc B*, 364 (1516):423–427, 2009. 3

[43] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. In *NeurIPS*, 2017. 3, 6

[44] Junyi Wang and Yue Qi. Multi-task learning and joint refinement between camera localization and object detection. *CVMJ*, pages 1–19, 2024. 2

[45] Lijun Wang, Huchuan Lu, Yifan Wang, Mengyang Feng, Dong Wang, Baocai Yin, and Xiang Ruan. Learning to detect salient objects with image-level supervision. In *CVPR*, 2017. 5

[46] Lijuan Wang, Guoli Jia, Ning Jiang, Haiying Wu, and Jufeng Yang. Ease: Robust facial expression recognition via emotion ambiguity-sensitive cooperative networks. In *ACM MM*, 2022. 3

[47] Changsong Wen, Guoli Jia, and Jufeng Yang. Dip: Dual incongruity perceiving network for sarcasm detection. In *CVPR*, 2023. 3

[48] Magnus Wrenninge and Jonas Unger. Synscapes: A photorealistic synthetic dataset for street scene parsing. *arXiv preprint arXiv:1810.08705*, 2018. 3

[49] Weijia Wu, Yuzhong Zhao, Mike Zheng Shou, Hong Zhou, and Chunhua Shen. Diffumask: Synthesizing images with pixel-level annotations for semantic segmentation using diffusion models. In *ICCV*, 2023. 2, 3

[50] Chuan Yang, Lihe Zhang, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang. Saliency detection via graph-based manifold ranking. In *CVPR*, 2013. 5

[51] Li Yuan, Xinyi Liu, Jiannan Yu, and Yanfeng Li. A full-set tooth segmentation model based on improved pointnet++. *VI*, 1(1):21, 2023. 2

[52] Yingjie Zhai, Guoli Jia, Yu-Kun Lai, Jing Zhang, Jufeng Yang, and Dacheng Tao. Looking into gait for perceiving emotions via bilateral posture and movement graph convolutional networks. *TAFFC*, 2024. 3

[53] Qing Zhang, Gelin Yin, Yongwei Nie, and Wei-Shi Zheng. Deep camouflage images. In *AAAI*, 2020. 3, 6, 7

[54] Xiaolin Zhang, Yunchao Wei, Yi Yang, and Thomas S Huang. Sg-one: Similarity guidance network for one-shot semantic segmentation. *TCYB*, 50(9):3855–3865, 2020. 5

[55] Yuxuan Zhang, Huan Ling, Jun Gao, Kangxue Yin, Jean-Francois Lafleche, Adela Barriuso, Antonio Torralba, and Sanja Fidler. Datasetgan: Efficient labeled data factory with minimal human effort. In *CVPR*, 2021. 2, 3

[56] Zhicheng Zhang and Jufeng Yang. Temporal sentiment localization: Listen and look in untrimmed videos. In *ACM MM*, 2022. 3

[57] Zhicheng Zhang, Song Chen, Zichuan Wang, and Jufeng Yang. Planeseg: Building a plug-in for boosting planar region segmentation. *TNNLS*, pages 1–15, 2023. 3

[58] Zhicheng Zhang, Shengzhe Liu, and Jufeng Yang. Multiple planar object tracking. In *ICCV*, 2023. 3

[59] Zhicheng Zhang, Lijuan Wang, and Jufeng Yang. Weakly supervised video emotion detection and prediction via cross-modal temporal erasing network. In *CVPR*, 2023. 3

[60] Zhicheng Zhang, Junyao Hu, Wentao Cheng, Danda Paudel, and Jufeng Yang. Extdm: Distribution extrapolation diffusion model for video prediction. In *CVPR*, 2024.

[61] Zhicheng Zhang, Pancheng Zhao, Eunil Park, and Jufeng Yang. Mart: Masked affective representation learning via masked temporal distribution distillation. In *CVPR*, 2024. 3

[62] Sicheng Zhao, Guoli Jia, Jufeng Yang, Guiguang Ding, and Kurt Keutzer. Emotion recognition from multiple modalities: Fundamentals and methodologies. *SPM*, 38(6):59–73, 2021. 3

[63] Sicheng Zhao, Xingxu Yao, Jufeng Yang, Guoli Jia, Guiguang Ding, Tat-Seng Chua, Björn W. Schuller, and Kurt Keutzer. Affective image content analysis: Two decades review and new perspectives. *TPAMI*, 44(10):6729–6751, 2022. 3

[64] Chuanxia Zheng, Tat-Jen Cham, Jianfei Cai, and Dinh Phung. Bridging global context interactions for high-fidelity image completion. In *CVPR*, 2022. 6, 7

[65] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *TPAMI*, 40(6):1452–1464, 2018. 6

[66] Shihao Zhou, Mengxi Jiang, Qicong Wang, and Yunqi Lei. Towards locality similarity preserving to 3d human pose estimation. In *ACCV*, 2020. 3

[67] Shihao Zhou, Mengxi Jiang, Shanshan Cai, and Yunqi Lei. Dc-gnet: Deep mesh relation capturing graph convolution network for 3d human shape reconstruction. In *ACM MM*, 2021.

[68] Shihao Zhou, Duosheng Chen, Jinshan Pan, Jinglei Shi, and Jufeng Yang. Adapt or perish: Adaptive sparse transformer with attentive feature refinement for image restoration. In *CVPR*, 2024. 3