

Text-guided Controllable Diffusion for Realistic Camouflage Images Generation

Yuhang Qian^{1,2}, Haiyan Chen², Wentong Li¹, Ningzhong Liu², Jie Qin^{1†}

¹ MoE Key Laboratory of Brain-Machine Intelligence Technology, College of Artificial Intelligence, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, Jiangsu, China

² College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, Jiangsu, China
qianyuhang@nuaa.edu.cn

Abstract

Camouflage Images Generation (CIG) is an emerging research area that focuses on synthesizing images in which objects are harmoniously blended and exhibit high visual consistency with their surroundings. Existing methods perform CIG by either fusing objects into specific backgrounds or outpainting the surroundings via foreground object-guided diffusion. However, they often fail to obtain natural results because they overlook the logical relationship between camouflaged objects and background environments. To address this issue, we propose CT-CIG, a Controllable Text-guided Camouflage Images Generation method that produces realistic and logically plausible camouflage images. Leveraging Large Visual Language Models (VLM), we design a Camouflage-Revealing Dialogue Mechanism (CRDM) to annotate existing camouflage datasets with high-quality text prompts. Subsequently, the constructed image-prompt pairs are utilized to finetune Stable Diffusion, incorporating a lightweight controller to guide the location and shape of camouflaged objects for enhanced camouflage scene fitness. Moreover, we design a Frequency Interaction Refinement Module (FIRM) to capture high-frequency texture features, facilitating the learning of complex camouflage patterns. Extensive experiments, including CLIPScore evaluation and camouflage effectiveness assessment, demonstrate the semantic alignment of our generated text prompts and CT-CIG’s ability to produce photorealistic camouflage images.

Code — <https://github.com/NikoNairre/CT-CIG>

Introduction

Camouflage is an instinctive survival mechanism that organisms utilize to blend into their surroundings, making them visually indistinct, either to evade predators or to ambush prey (Merilaita, Scott-Samuel, and Cuthill 2017). Owing to its interesting and challenging characteristics, camouflage vision perception has gained attention in research such as Camouflaged Object Detection (COD) (Fan et al. 2020; Luo et al. 2024; Yan et al. 2025) and Concealed Instance Ranking (CIR) (Lv et al. 2021). However, progress in this field is hampered by a performance bottleneck stemming from the scarcity of camouflage training datasets, underscoring the

[†]Corresponding Author.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

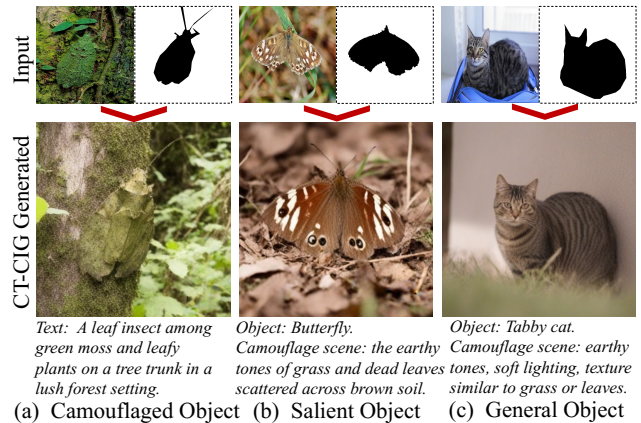


Figure 1: Example images generated by CT-CIG, which proves its ability to handle objects of different attributes.

requirement for effective data acquisition methods. The difficulty of collecting natural camouflage images—owing to intricate and environment-specific conditions—has spurred the emergence of Camouflage Images Generation (CIG) (Zhao et al. 2025), a paradigm focused on synthesizing artificial camouflage images.

The technique of generating digital camouflage images can be dated back to (Chu et al. 2010) via hand-crafted feature processing. Current CIG methods generally fall into two paradigms. The first paradigm performs camouflage by altering the color and texture patterns of objects to blend themselves harmoniously into arbitrary backgrounds (Zhang et al. 2020; Li et al. 2022; Gao, Yang, and Liu 2025), concluded as Background fitting. Despite the artistic beauty and exquisite visual concealment illusion, they fail to match the camouflage regularity that the natural world represents due to the damage to objects’ appearances and ignorance of the logic between foreground objects and background images (e.g., a tiger face in a mountain). The second paradigm, on the other hand, leverages generative models (Isola et al. 2017; Ho, Jain, and Abbeel 2020; Rombach et al. 2022) to outpaint background surroundings with visual color consistency to foreground objects guided by their inherent features (Zhang et al. 2023; Zhao et al. 2024; Das and Gopalakrishnan 2025), which can be viewed as Foreground guiding.

However, they fail to equip sufficient semantic consideration for the generation of backgrounds, which causes severe artifacts and makes the backgrounds unrealistic.

Motivated by the problems mentioned above, we aim to design a CIG method that can produce realistic and naturalistic camouflage images, and finally propose CT-CIG. Examples of generated camouflage images are displayed in Fig. 1. Specifically, CT-CIG employs stable diffusion to handle input images from existing COD datasets, with the corresponding text descriptions for the content guidance on both foreground objects and background environments. Compensating for the absence of text prompts in COD datasets, we set a Camouflage-Revealing Dialogue Mechanism (CRDM) to exploit the image perception and contextual comprehension of Large Vision Language Models (VLM) (Li et al. 2023; Wang et al. 2024a) to obtain camouflage-sensed and semantically aligned text descriptions, which are of vital significance towards realistic camouflage generation. Furthermore, a lightweight control network is utilized to handle binary masks of camouflaged objects and provide signals of objects' shapes and locations for the diffusion process. To reduce pattern artifacts raised from noisy latents and endow the generated images with greater fidelity and more intricate details, we additionally design a Frequency Interaction Refinement Module (FIRM), which is capable of capturing high-frequency textures by attention weighing on Fast Fourier Transform (FFT) features (Campbell and Robson 1968). Our key contributions can be summarized as follows:

- We leverage VLMs and develop a Camouflage-Revealing Dialogue Mechanism to generate high-quality text prompts for camouflage images, which makes the paradigm of Text-guided Camouflage Images Generation implementable.
- We propose CT-CIG, a Controllable Text-guided Camouflage Images Generation method that leverages the powerful ability of diffusion to generate camouflage images with natural realism, along with a lightweight controller for geometric guidance of camouflaged objects and a Frequency Interaction Refinement Module to ensure the texture stability and content fidelity.
- Experiments on the evaluation of metrics such as FID, KID, and CLIPScore, effectively demonstrate the fidelity of generated images and the descriptive accuracy of the generated text prompts. Related vision comparisons with previous methods validate CT-CIG's ability to produce photorealistic and logically feasible camouflage images.

Related Works

Camouflage Images Generation

The early CIG method (Chu et al. 2010) exploits hand-made features and devises a specialist algorithm for camouflage texture synthesis. DCI (Zhang et al. 2020) and LCG-Net (Li et al. 2022) leverage convolution networks (Simonyan and Zisserman 2014) on both foreground and background images and utilize feature fusion strategies inspired by style transfer (Huang and Belongie 2017) to camouflage objects into specific backgrounds. PTDiffusion (Gao, Yang, and Liu

2025) transfers a reference image containing the target object to a text-instructed scene picture in which the object hides as a form of illusion. These methods all follow the background-fitting paradigm without considering the plausibility of camouflaged objects and background surroundings, which makes camouflage a representation of vision art and optical illusion rather than a natural phenomenon.

Regarding foreground-guiding CIG methods, Generative Adversarial Networks (GAN) are used to synthesize artificial camouflage images based on the input of real images and masks from COD datasets (Zhang et al. 2023; Lamdouar, Xie, and Zisserman 2023; He et al. 2023). LAKE-RED (Zhao et al. 2024) outpaints backgrounds that have similar colors to the foreground objects by knowledge retrieval with the combination of VQVAE (Van Den Oord, Vinyals et al. 2017) and latent diffusion. FACIG (Chen et al. 2025) further refines the feature integration manner to reduce foreground distortion. Due to the lack of any components to facilitate the background's semantic understanding, their generated images exhibit severe texture artifacts that fail to realize natural realism. CamOT (Das and Gopalakrishnan 2025) alleviates this problem to some degree by constructing a representation engineering work on clip-processed embeddings (Radford et al. 2021). However, ensuring logical feasibility and content controllability remains a challenge, since their prompts contain merely the literal word "background" and object labels in COD datasets. Our CT-CIG, with VLM-produced prompts, opens the paradigm of Text-guided CIG, which enables generating logic-feasible and realistic camouflage images.

Vision Language Models

Large Vision Language Models (VLM) are capable of understanding the visual-language correlation for various tasks (e.g., object recognition, segmentation, graphic and image understanding) by learning rich vision (photo or video) text pairs that are almost infinite at web scale (Zhang et al. 2024). Therefore, VLMs excel in zero-shot prediction on different tasks. (Li et al. 2025). Pioneer VLMs such as CLIP (Radford et al. 2021) and BLIP (Li et al. 2023) are built by training the integrated vision encoder and the text encoder from scratch. In contrast, subsequent powerful VLMs like LLaVA (Liu et al. 2023) and Qwen-VL (Bai et al. 2023) exploit LLMs as a backbone and design projectors for visual-language alignment in a shared embedding space. Camobj-Llava (Ruan et al. 2025) is the first VLM that specializes in understanding camouflage scenarios by learning camouflage content produced by GPT-4o. We choose Qwen2.5-VL (Bai et al. 2025) and impose Camouflage-Revealing Dialogue Mechanism to obtain text prompts paired with camouflage images for consideration of both performance and deployment convenience.

Diffusion Models for Controllable Generation

Diffusion (Ho, Jain, and Abbeel 2020; Song, Meng, and Ermon 2020) is widely used in various image generation tasks for its powerful ability to predict and eliminate noise. LDM (Rombach et al. 2022) transforms pixel images to latent features and performs diffusion in the latent space. With

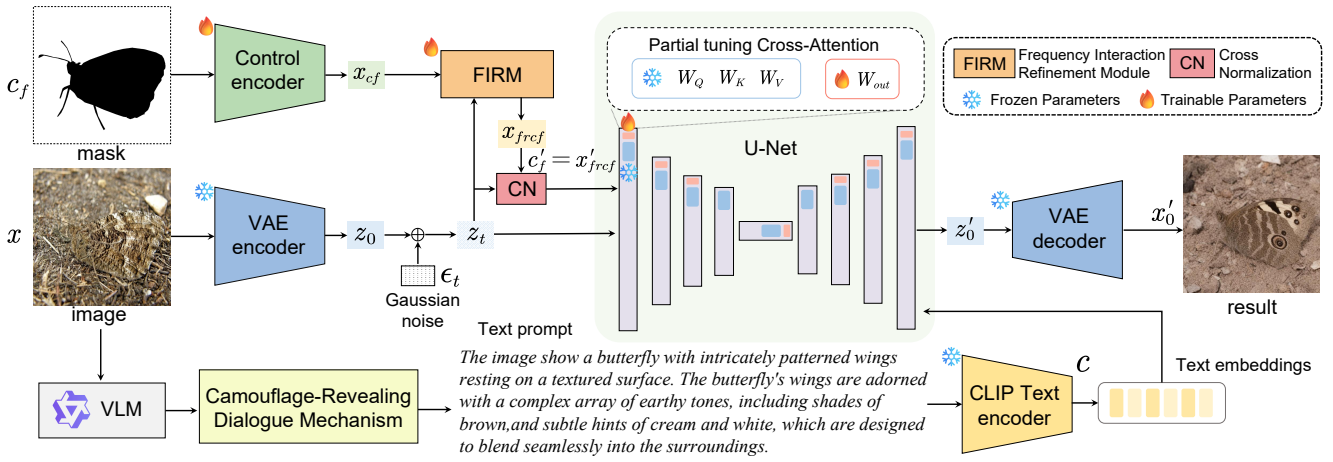


Figure 2: Overall framework of our proposed CT-CIG, which performs camouflage generation via three steps. (1) Extracting features of input images and masks through VAE and controller, followed by control augmentation through FIRM and CN. (2) Generating text prompts from the VLM through CRDM and using the CLIP encoder to obtain embeddings. (3) Performing controllable stable diffusion and generating results.

support for extra condition inputs, such as text prompts, it succeeds in producing content-controlled outcomes and pioneers downstream tasks like layout-to-image (Wang et al. 2024b), image inpainting (Chen et al. 2024) or outpainting (Eshratifar et al. 2024), and editing (Nguyen et al. 2025). ControlNet (Zhang, Rao, and Agrawala 2023) is subsequently proposed to enable the generation of images with structural consistency to the multi-modal control signals (e.g., depth map, instance segmentation, scribble) that are handled by parallelism blocks in the stable diffusion UNet encoder, adaptively grasping the related spatial distributions. ControlNext (Peng et al. 2024) furthermore processes external control signals only with a simple, lightweight network, which significantly accelerates training and serves as the foundation of CT-CIG.

Methodology

Preliminaries

The core of diffusion lies in predicting noise added to data under the paradigm of forward diffusion and reverse denoising. DDPM (Ho, Jain, and Abbeel 2020) estimates the posterior $p_\theta(x_{t-1}|x_t)$ with its corresponding noise adding process formulated as:

$$x_t = \sqrt{\alpha_t}x_{t-1} + \sqrt{1 - \alpha_t}\epsilon = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon, \quad (1)$$

where x_0 is the original image and x_t is the noisy version at timestep t , α_t denotes the diffusion constant and $\bar{\alpha}_t = \prod_{i=1}^t \alpha_i$. DDIM overcomes DDPM’s bottleneck of inference time raised from the Markov Chain to support step-skipped prediction. More details can be found in (Song, Meng, and Ermon 2020).

To reduce computation cost, LDM (Rombach et al. 2022) utilizes a pre-trained autoencoder ϵ (Kingma, Welling et al. 2013) to transform x to a smaller latent representation $z_0 = \epsilon(x_0)$ and then performs diffusion. The training objective is to optimize noise prediction $\epsilon_\theta(z_t, t)$ via MSE loss:

$$\mathcal{L}_{SD} = \mathbb{E}_{\epsilon(x), t, \epsilon \sim \mathcal{N}(0,1)} [\|\epsilon_\theta(z_t, t) - \epsilon\|_2^2]. \quad (2)$$

Extra condition c (e.g., text prompt) is optional in LDM for task-specific training, and ControlNeXt further integrates control signal c_f . The loss function can be calculated as:

$$\mathcal{L}_{SD} = \mathbb{E}_{\epsilon(x), t, c, c_f, \epsilon \sim \mathcal{N}(0,1)} [\|\epsilon_\theta(z_t, t, c, c_f) - \epsilon\|_2^2]. \quad (3)$$

Framework of CT-CIG

To effectively synthesize camouflage images, CT-CIG must understand the intricate traits of natural camouflage, including object concealment, color consistency, texture similarity, and logical plausibility between objects and their background surroundings. All traits except logical plausibility are learnable through training with authentic camouflage images. Logical plausibility is not explicitly exhibited in the pixel domain but is presented implicitly in the semantic domain. To compensate for the gap, text prompts paired with training images, which are available through VLMs and contain detailed semantic content, serve as conditional information for training CT-CIG. Inspired by ControlNeXt (Peng et al. 2024), binary masks of objects in COD datasets can be used more effectively as spatial guidance to control the location and shape of generated objects.

The general framework of CT-CIG is illustrated in Fig. 2. $x \in \mathcal{R}^{3 \times h \times w}$ and $c_f \in \mathcal{R}^{1 \times h \times w}$ are paired RGB and mask images. Text prompt that describes x is generated through Vision Question Answering (VQA) in VLM by the CRDM and subsequently fed into a CLIP Text-encoder to form text embeddings c . We use a VAE to encode the RGB camouflage image to the latent $z_0 \in \mathcal{R}^{4 \times 64 \times 64}$ and subsequently add Gaussian noise to obtain z_t , where t is randomly sampled from $[0, T]$ ($T = 1000$). A lightweight controller, comprises Resblocks (He et al. 2016), is exploited to encode c_f to get the control feature x_{cf} . FIRM then processes x_{cf} along with z_t to calculate its frequency-enhanced x_{freq} to grasp high-frequency texture details, followed by Cross Normalization (CN) to obtain the mean and variance aligned control feature x'_{freq} . After that, the denoising procedure to recover latent

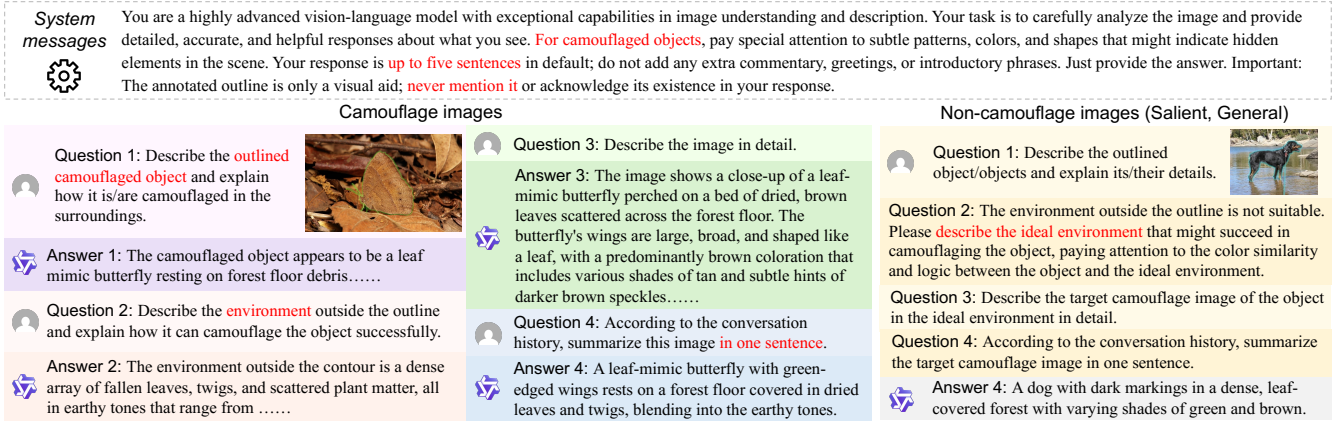


Figure 3: Intra details of the Camouflage-Revealing Dialogue Mechanism. VLMs should obey the rules in system messages to produce answers that meet our requirements. Different queries are designed for camouflage images and non-camouflage images to guide them to generate camouflage-originated responses.

noise-free z'_0 from z_t with the condition of c and x'_{fref} is conducted via UNet (Ronneberger, Fischer, and Brox 2015) based stable diffusion. Finally, a VAE decoder transforms z'_0 back to image domain and generates the result camouflage image x'_0 . Parameters in the controller, FIRM, and linear projectors in UNet’s cross-attention blocks are trainable, while the remaining parts are kept frozen.

Camouflage-Revealing Dialogue Mechanism. As shown in Fig. 3, Camouflage-Revealing Dialogue Mechanism (CRDM) leverages VLM’s ability in visual perception and contextual comprehension. We design four questions for each image with camouflage-guided instructions that gradually teach the VLM to make camouflage-oriented descriptions. All images are pre-processed by using random semi-transparent colored outlines to annotate object boundaries, where they directly adjoin the backgrounds and the key insights of camouflage lie. Thanks to the semi-transparent effect, it not only helps the VLM to localize and familiarize the camouflaged object but also preserves the related boundary pixel details to understand the camouflage paradigm. For camouflage images, questions 1 and 2 are designed to obtain the descriptions of objects and the surrounding environments, along with their relationships. Subsequently, these descriptions are reorganized to form detailed prompts through question 3. The final question 4 is defined to review all contents and summarize each prompt into one sentence. We denoted the detailed and summarized prompt as T_{detail} and T_{simple} . Despite a suitable dialogue mechanism to handle camouflage images, the descriptions of background environments act as negative knowledge regarding non-camouflage images (e.g., salient object images and general images). Therefore, we ask the VLM to imagine an ideal scenery that might successfully camouflage the object in question 2 and produce related T_{detail} and T_{simple} in later questions, as illustrated in the right part of Fig. 3. Additionally, we pre-define some system messages to aid VLMs for camouflage perception and to force VLMs to avoid producing any unrelated and redundant content. T_{simple} is uti-

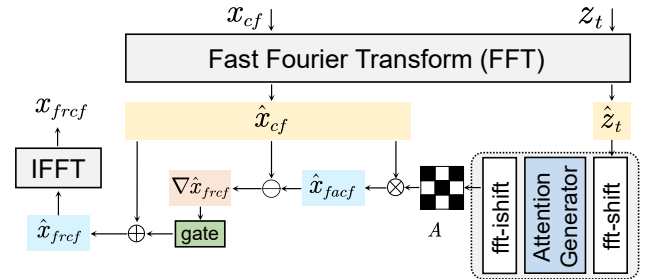


Figure 4: Frequency Interaction Refinement Module.

lized during inference for generation diversity, while T_{detail} is used for training CT-CIG because it encapsulates more intricate information, which forces CT-CIG to learn complex camouflage patterns and prevents catastrophic forgetting.

Frequency Interaction Refinement Module Compared to other controls such as depth map, infrared map, or object canny, binary masks only indicate rough position and geometric cues for target camouflaged objects, without subtle information about spatial hierarchy or objects’ intra appearance. Consequently, x_{cf} encoded by the controller is information-deficient. Through easy control convergence, it has the potential risk of producing texture artifacts and unnatural hallucinations. To address this issue, we propose the Frequency Interaction Refinement Module (FIRM), based on the Fourier Transform, to enhance the information granularity of x_{cf} . According to the theory of the Fourier Spectrum, low frequencies contribute to the image’s overall structural information, while high frequencies contribute to the image’s texture and intricate pattern information (Campbell and Robson 1968; Zhou et al. 2023). Therefore, it is possible to equip x_{cf} with detailed texture representations learned from the image latent z_t through FIRM.

Fig. 4 displays the detailed pipeline of FIRM. Designed to strengthen the granularity of the control signal guided by the input feature, x_{cf} and z_t are first applied to Fast Fourier

Transform (FFT) to obtain related spectrums \hat{x}_{cf} and \hat{z}_t in the frequency domain. An attention map A that takes magnitude $|\hat{z}_t|$ as source data is generated by an Attention Generator block, which consists of 2 convolution layers. The initial pattern of $|\hat{z}_t|$ is not continuous because high-frequency components are separated in the center and corners. To fit convolution’s local spatial perception, fftshift is employed to permute a continuous spectrum where low frequencies are around the image center, with high frequencies uniformly arranged in the outer ring. The formulation for obtaining the frequency attention weights matrix A is:

$$A = \text{ifft-shift}(\text{AG}(\text{fft-shift}(|\hat{z}_t|))), \quad (4)$$

where AG denotes Attention Generator, and ifft-shift restores the spectrum to its original pattern. Subsequently, we interact \hat{x}_{cf} with A to capture subtle texture details and calculate the attention-enhanced control spectrum \hat{x}_{facf} and the refinement gain $\nabla\hat{x}_{frcf}$. The refinement gain is then added to \hat{x}_{cf} with its intensity adaptively controlled by a learnable gate to form the frequently refined control spectrum \hat{x}_{frcf} . Related equations are presented as follows:

$$\hat{x}_{facf} = \hat{x}_{cf} \otimes A, \quad (5)$$

$$\nabla\hat{x}_{frcf} = \hat{x}_{facf} - \hat{x}_{cf}, \quad (6)$$

$$\hat{x}_{frcf} = \hat{x}_{cf} + \text{gate} \times \nabla\hat{x}_{frcf}. \quad (7)$$

Finally, Inverse Fast Fourier Transform (IFFT) is applied to transfer the refined control back to the feature domain, denoted as x_{frcf} . Compared to x_{cf} , this FIRM refined control feature is more fine-grained for ensuring robustness and facilitating the generation of complex camouflage textures.

Diffusion with Cross Normalization The discrepancy between the FIRM refined control feature and the noisy image latent poses a risk of color instability. Motivated by (Peng et al. 2024), Cross Normalization (CN) is utilized as an alternative to the "zero convolution" layers in ControlNet to enhance training robustness. Initially, channel-wise mean and standard deviation of the control feature x_{frcf} and latent z_t , notated as μ and σ , are calculated via:

$$\mu_z, \mu_{cf} = \frac{1}{n_1} \sum_{i=1}^{n_1} z_{t,i}, \frac{1}{n_2} \sum_{i=1}^{n_2} x_{cf,i}, \quad (8)$$

$$\sigma_z^2, \sigma_{cf}^2 = \frac{1}{n_1} \sum_{i=1}^{n_1} (z_{t,i} - \mu_z)^2, \frac{1}{n_2} \sum_{i=1}^{n_2} (x_{cf,i} - \mu_{cf})^2. \quad (9)$$

We subsequently perform CN that undergoes x_{frcf} standardization and z_t affine transformation formulated as:

$$x'_{frcf} = \mu_z + \frac{x_{frcf} - \mu_{cf}}{\sqrt{\sigma_{cf}^2 + \varepsilon}} \times \sigma_z, \quad (10)$$

where we define $c'_f = x'_{frcf}$ as the final control signal. This concise and ingenious strategy renders c'_f distributally consistent with z_t , effectively facilitating training.

Training optimization Apart from the condition integrated diffusion loss \mathcal{L}_{SD} , we employ an LPIPS perceptual loss (Zhang et al. 2018) to minimize the feature discrepancy of the predicted result x'_0 and the input image x_0 , which benefits the generation of natural and realistic images. The perceptual loss is calculated as:

$$\mathcal{L}_{Lpips} = \sum_l \frac{1}{h_l w_l} \sum_{h,w} \left\| \gamma_l \cdot (f_0^l - f_0'^l) \right\|_2^2, \quad (11)$$

where f_0', f_0 are VGG encoded feature stacks of x'_0 and x_0 , l and γ denote the stack id and scale factor. The overall loss function is formulated below, with perceptual loss multiplied by the corresponding weight.

$$\mathcal{L} = \mathcal{L}_{SD} + \lambda_{Lpips} \cdot \mathcal{L}_{Lpips}. \quad (12)$$

Experiments

Experimental Settings

Datasets and metrics We utilize the LAKE-RED dataset (Zhao et al. 2024), which comprises 4040 images for training, 19419 images for validation, and 5066 images for evaluation. The training dataset consists of camouflage images from COD10K (Fan et al. 2020) and CAMO (Le et al. 2019). There exist three subsets in the test dataset, including Camouflaged Objects (Lv et al. 2021), Salient Objects (Wang et al. 2017), and General Objects (Lin et al. 2014); each category contains 6473 images. We process these images through VLM to form related image-text pairs and use the CLIP Score (Hessel et al. 2021) to measure image-text alignment in both training and testing. Following previous works, we choose FID (Binkowski et al. 2018) and KID (Heusel et al. 2017) to quantify the quality of generated images, taking the 5066 camouflage images from COD10K as the evaluation benchmark.

Implementation Details To obtain text prompts, we build Qwen2.5-VL via the VLLM framework and perform VQA with 2 NVIDIA RTX 4090 GPUs. Subsequently, we implement CT-CIG based on ControlNext and take a pre-trained SDXL as the foundation of diffusion. During training, images and control masks are resized to 512×512 and then transformed to 128×128 latents with a batch size of 4. We set the control scale as 1.2 and $\lambda_{LPIPS}=1e-3$. For quick convergence, the learning rate of the lightweight controlnet and FIRM are set to $1e-4$, which facilitates control signal adaptation, while SDXL UNet’s learning rate is set to $5e-6$ for cautious finetuning. It takes around 8 hours to train our CT-CIG for 80 epochs with 4 NVIDIA RTX A5000 GPUs.

VLM Choice

VLMs vary in their ability to understand the input content and make responses. We employ CRDM to several open-source VLMs (Li et al. 2023; Liu et al. 2023; Team et al. 2025; Bai et al. 2025) and use CLIPScore to measure the quality of their produced content. As shown in Table 1, T_{detail} is likely to outperform T_{simple} under the same VLM because it contains more corpus that explains images at a finer grain. Qwen2.5-VL is chosen as our VLM backbone for its superior text-image alignment performance.

Comparison with SOTA methods

We compare our CT-CIG with 11 SOTA methods; some are specialized in CIG (Zhang et al. 2020; Li et al. 2022; Zhao et al. 2024), while others can be adapted to camouflage generation (Pérez, Gangnet, and Blake 2003; Huang and Belongie 2017; Zheng et al. 2022; Rombach et al. 2022; Lugmayr et al. 2023; Zhang, Rao, and Agrawala 2023; Eshratifar

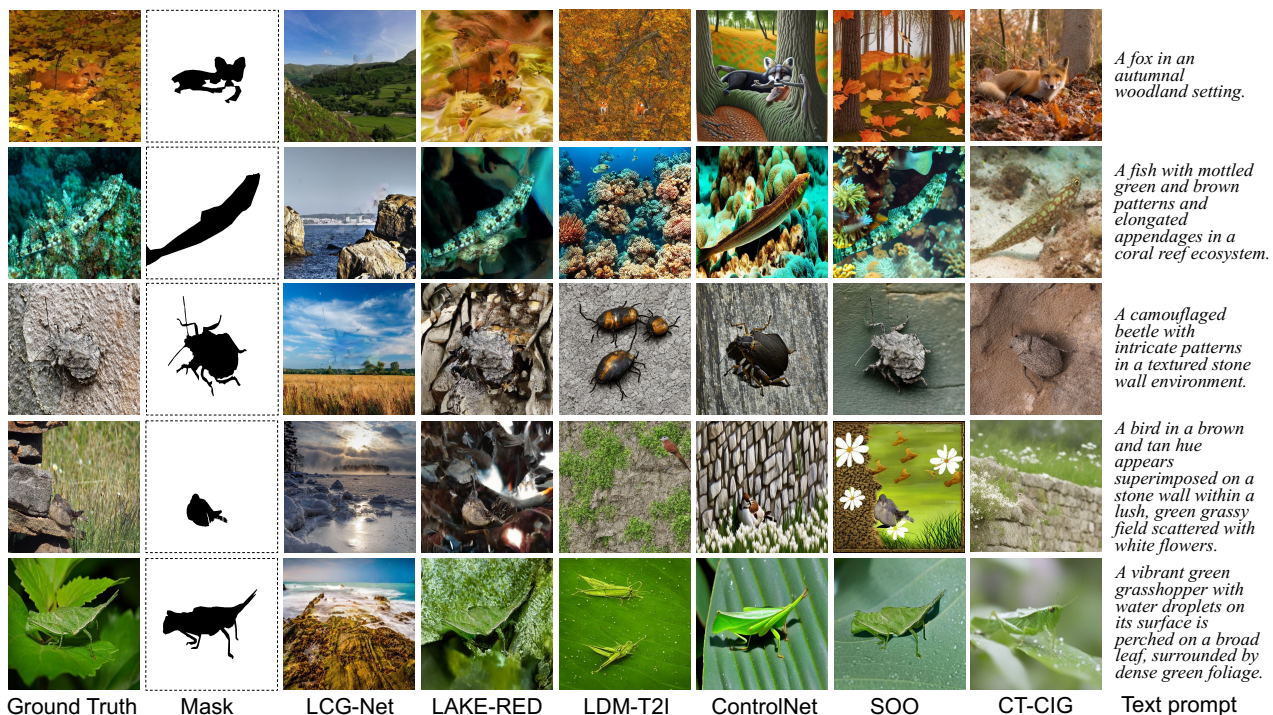


Figure 5: Results of generated images with different methods. The first two columns show the real images and paired masks in the COD datasets. Backgrounds in column 3 are randomly selected. Methods in columns 5-8 require text prompts as a condition. All methods take masks as shape guidance except LDM-T2I.

Method	CLIP simple \uparrow	CLIP detail \uparrow
BLIP2-2.7B	0.2461	0.2859
LLaVA-13B	0.2986	0.2969
Gemma3-4B	0.3127	0.3136
Qwen2.5-VL-7B	0.3183	0.3242

Table 1: CLIPScores of detailed and summarized prompts generated by employing CDRM to different VLMs.

et al. 2024), covering the paradigms of Background fitting, Foreground guiding, and Text guiding.

Qualitative Results Fig. 5 displays the generated images of different methods. The background-fitting LCGNet camouflages objects into arbitrary background images excessively, which causes severe distortion and makes them barely visible. LAKERED outpaints the background regions with similar colors guided by the foreground objects. Both of these paradigms ignore the semantic relationship between camouflaged objects and environments, which contributes to little logical plausibility. With the integration of text prompts, LDM-T2I produces images with overall semantic tendency, but their compositions are disorganized without control guidance. Taking masks as a control signal, ControlNet successfully generates objects in the correct location and shapes. However, the generative artifacts lead to a lack of photorealism, which places them in the uncanny valley. Salient Object Outpainting (SOO) is designed to fill

the background with general prompts, which can't guarantee its fitness for camouflage. Our CT-CIG, in contrast, is capable of generating photorealistic camouflage images with detailed textures, depth of field, and contextual coherence.

Quantitative Results According to the LAKERED test dataset, we split the generated images into three groups. Apart from FID and KID, we use CLIPScore in text-guided methods to evaluate the semantic alignment of generated images and text prompts. Related results are shown in Table 2. Background fitting methods perform the worst in all object types due to the object appearance destruction. Foreground guiding methods achieve some improvement by maintaining object consistency, but the textual artifacts in background regions limit their performance upperbound. Large performance enhancement can be realized with the combination of mask control signal in text guiding methods. Our method achieves the best performance on the image quality and text alignment in the overall setting, and remains highly competitive in category-specific entries. Holistically, CT-CIG presents the most compelling performance profile.

Ablation Study

As illustrated in Fig. 6 (a), the baseline model (without FIRM and CN) struggles to produce high-quality camouflage images. Adding Cross Normalization effectively reduces spurious color speckles. The integration of FIRM further mitigates the issue of missing high-frequency details and enables CT-CIG to generate camouflage images with

Paradigm	Method	Camouflaged Objects		Salient Objects		General Objects		Overall		
		FID ↓	KID ↓	FID ↓	KID ↓	FID ↓	KID ↓	FID ↓	KID ↓	CLIP ↑
Background fitting	AB (2003)	117.11	0.0645	126.78	0.0614	133.89	0.0645	120.21	0.0623	-
	AdAIN (2017)	125.16	0.0721	133.20	0.0702	136.93	0.0714	126.94	0.0703	-
	DCI (2020)	130.21	0.0689	134.92	0.0665	137.99	0.0690	130.52	0.0673	-
	LCGNet (2022)	129.80	0.0504	136.24	0.0597	132.64	0.0548	129.88	0.0550	-
Foreground guiding	TFill (2022)	63.74	0.0336	96.91	0.0453	122.44	0.0747	80.39	0.0438	-
	LDM-Inpaint (2022)	58.65	0.0380	107.38	0.0524	129.04	0.0748	84.48	0.0488	-
	RePaint-L (2023)	76.80	0.0459	114.96	0.0497	136.18	0.0686	96.14	0.0498	-
	LAKERED (2024)	39.55	0.0212	88.70	0.0428	102.67	0.0625	64.27	0.0355	-
Text guiding	LDM-T2I (2022)	51.24	0.0206	102.04	0.0217	120.41	<u>0.0304</u>	73.51	0.0261	0.2873
	ControlNet (2023)	39.67	0.0121	<u>81.72</u>	0.0303	<u>102.94</u>	0.0422	<u>59.52</u>	0.0227	0.2950
	SOO (2024)	<u>30.92</u>	0.0056	89.46	0.0267	117.31	0.0423	59.75	<u>0.0187</u>	<u>0.3043</u>
	CT-CIG (ours)	30.59	<u>0.0085</u>	81.60	<u>0.0230</u>	104.46	0.0241	52.88	0.0169	0.3243

Table 2: Quantitative evaluation results on FID, KID, and CLIPScore for generated images of CT-CIG compared with 11 SOTA methods. The best results are highlighted in bold, and the second-best results are underlined.

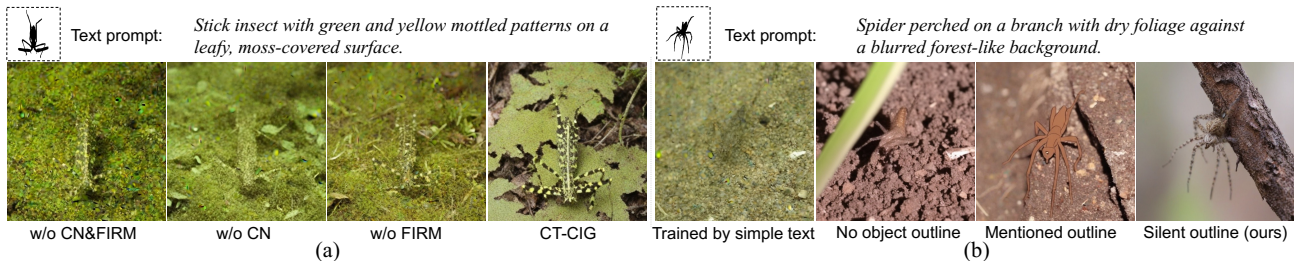


Figure 6: Visualizations of ablation studies, including (a) the role of FIRM and CN in CT-CIG, and (b) training with text prompts that were generated via different query settings in CRDM.

Configuration	FID ↓	KID ↓
w/o FIRM&CN	32.37	0.0079
w/o CN	33.99	0.0114
w/o FIRM	31.66	0.0080
CT-CIG	30.59	0.0085

Table 3: Image quality quantification for ablating FIRM and Cross Normalization. "w/o" denotes without.

clear structures and subtle textures. The results in Table 3 show the improvement in performance after incorporating FIRM and CN. Fig. 6 (b) and Table 4 show the effect of different text prompt configurations in CRDM. Silent outline means that anything directly related to object outlines is forbidden to appear in generated text prompts. Training with T_{simple} causes CT-CIG to suffer catastrophic forgetting that produces blurry results. We find that prompts generated without object outlines, even those achieving a high CLIPScore with the source image, often fail to match shape guidance. When the restrictions on outline-related content are removed, the resulting prompts tend to include explicit descriptions of outlines, which can mislead the generation process and produce line-drawing-like artifacts. These ablation studies effectively demonstrate the necessity and effectiveness of each designed component in CT-CIG.

Configuration	CLIP Score ↑	FID ↓	KID ↓
Trained by simple text	0.3183	54.92	0.0387
No object outline	0.3247	39.24	0.0112
Mentioned outline	0.3218	39.79	0.0138
Silent outline (ours)	0.3242	30.59	0.0085

Table 4: The impact of different text prompts obtaining configurations on model performance.

Conclusion

In this paper, we propose a Controllable Text-guided Camouflage Images Generation method, termed CT-CIG, which leverages a Stable Diffusion backbone with a lightweight controller to generate camouflage images with logical plausibility and natural realism. CT-CIG incorporates three key components: a Camouflage-Revealing Dialogue Mechanism that obtains text prompts for camouflage images with the aid of VLMs, a Frequency Interaction Refinement Module for capturing high-frequency texture details, and Cross Normalization to ensure generation stability. Rigorous experiments demonstrate the superior performance of CT-CIG. We hope that our work lays the groundwork for the Text-guided Camouflage Images Generation paradigm and will spur further research in this field.

Acknowledgements

This work was partially supported by the National Natural Science Foundation of China (No. U25A20533, No. 62276129), the Natural Science Foundation of Jiangsu Province (No. BK20250082), the Fundamental Research Funds for the Central Universities (No. NE2025010, No. NS2025038), and the Jiangsu Funding Program for Excellent Postdoctoral Talent (No. 2025ZB306).

References

- Bai, J.; Bai, S.; Yang, S.; Wang, S.; Tan, S.; Wang, P.; Lin, J.; Zhou, C.; and Zhou, J. 2023. Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond. arXiv:2308.12966.
- Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; Zhong, H.; Zhu, Y.; Yang, M.; Li, Z.; Wan, J.; Wang, P.; Ding, W.; Fu, Z.; Xu, Y.; Ye, J.; Zhang, X.; Xie, T.; Cheng, Z.; Zhang, H.; Yang, Z.; Xu, H.; and Lin, J. 2025. Qwen2.5-VL Technical Report. arXiv:2502.13923.
- Bińkowski, M.; Sutherland, D. J.; Arbel, M.; and Gretton, A. 2018. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*.
- Campbell, F. W.; and Robson, J. G. 1968. Application of Fourier analysis to the visibility of gratings. *The Journal of physiology*, 197(3): 551.
- Chen, P.-C.; Yao, Y.; Hsu, C.-F.; Xie, H.; Chen, H.-J.; Shuai, H.-H.; and Cheng, W.-H. 2025. Foreground Focus: Enhancing Coherence and Fidelity in Camouflaged Image Generation. *arXiv preprint arXiv:2504.02180*.
- Chen, Y.; Chen, J.; Pan, Y.; Li, Y.; Yao, T.; Chen, Z.; and Mei, T. 2024. Improving text-guided object inpainting with semantic pre-inpainting. In *European conference on computer vision*, 110–126. Springer.
- Chu, H.-K.; Hsu, W.-H.; Mitra, N. J.; Cohen-Or, D.; Wong, T.-T.; and Lee, T.-Y. 2010. Camouflage images. *ACM Trans. Graph.*, 29(4): 51–1.
- Das, B.; and Gopalakrishnan, V. 2025. Camouflage Anything: Learning to Hide using Controlled Out-painting and Representation Engineering. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 3603–3613.
- Eshratifar, A. E.; Soares, J. V.; Thadani, K.; Mishra, S.; Kuznetsov, M.; Ku, Y.-N.; and De Juan, P. 2024. Salient object-aware background generation using text-guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7489–7499.
- Fan, D.-P.; Ji, G.-P.; Sun, G.; Cheng, M.-M.; Shen, J.; and Shao, L. 2020. Camouflaged object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2777–2787.
- Gao, X.; Yang, S.; and Liu, J. 2025. PTDiffusion: Free Lunch for Generating Optical Illusion Hidden Pictures with Phase-Transferred Diffusion Model. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 18240–18249.
- He, C.; Li, K.; Zhang, Y.; Zhang, Y.; You, C.; Guo, Z.; Li, X.; Danelljan, M.; and Yu, F. 2023. Strategic Preys Make Acute Predators: Enhancing Camouflaged Object Detectors by Generating Camouflaged Objects. In *The Twelfth International Conference on Learning Representations*.
- He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.
- Hessel, J.; Holtzman, A.; Forbes, M.; Bras, R. L.; and Choi, Y. 2021. CLIPScore: A Reference-free Evaluation Metric for Image Captioning. In *EMNLP*.
- Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.
- Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.
- Huang, X.; and Belongie, S. 2017. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, 1501–1510.
- Isola, P.; Zhu, J.-Y.; Zhou, T.; and Efros, A. A. 2017. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1125–1134.
- Kingma, D. P.; Welling, M.; et al. 2013. Auto-encoding variational bayes.
- Lamdouar, H.; Xie, W.; and Zisserman, A. 2023. The making and breaking of camouflage. In *Proceedings of the IEEE/CVF international conference on computer vision*, 832–842.
- Le, T.-N.; Nguyen, T. V.; Nie, Z.; Tran, M.-T.; and Sugimoto, A. 2019. Anabran network for camouflaged object segmentation. *Computer vision and image understanding*, 184: 45–56.
- Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, 19730–19742. PMLR.
- Li, Y.; Zhai, W.; Cao, Y.; and Zha, Z.-J. 2022. Location-free camouflage generation network. *IEEE Transactions on Multimedia*, 25: 5234–5247.
- Li, Z.; Wu, X.; Du, H.; Liu, F.; Nghiem, H.; and Shi, G. 2025. A Survey of State of the Art Large Vision Language Models: Benchmark Evaluations and Challenges. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 1587–1606.
- Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, 740–755. Springer.
- Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36: 34892–34916.

- Lugmayr, A.; Danelljan, M.; Romero, A.; Yu, F.; Timofte, R.; and Van Gool, L. R. 2023. Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11461–11471.
- Luo, Z.; Liu, N.; Zhao, W.; Yang, X.; Zhang, D.; Fan, D.-P.; Khan, F.; and Han, J. 2024. Vscope: General visual salient and camouflaged object detection with 2d prompt learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 17169–17180.
- Lv, Y.; Zhang, J.; Dai, Y.; Li, A.; Liu, B.; Barnes, N.; and Fan, D.-P. 2021. Simultaneously localize, segment and rank the camouflaged objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11591–11601.
- Merilaita, S.; Scott-Samuel, N. E.; and Cuthill, I. C. 2017. How camouflage works. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 372(1724): 20160341.
- Nguyen, T.-T.; Nguyen, Q.; Nguyen, K.; Tran, A.; and Pham, C. 2025. Swiftedit: Lightning fast text-guided image editing via one-step diffusion. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 21492–21501.
- Peng, B.; Wang, J.; Zhang, Y.; Li, W.; Yang, M.-C.; and Jia, J. 2024. Controlnext: Powerful and efficient control for image and video generation. *arXiv preprint arXiv:2408.06070*.
- Pérez, P.; Gangnet, M.; and Blake, A. 2003. Poisson image editing. *ACM Transactions on Graphics*, 22(3): 313–318.
- Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.
- Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.
- Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, 234–241. Springer.
- Ruan, J.; Yuan, W.; Lin, Z.; Liao, N.; Li, Z.; Xiong, F.; Liu, T.; and Fu, Y. 2025. Mm-camobj: A comprehensive multimodal dataset for camouflaged object scenarios. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 6740–6748.
- Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.
- Song, J.; Meng, C.; and Ermon, S. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.
- Team, G.; Kamath, A.; Ferret, J.; Pathak, S.; Vieillard, N.; Merhej, R.; Perrin, S.; Matejovicova, T.; Ramé, A.; Rivière, M.; et al. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.
- Van Den Oord, A.; Vinyals, O.; et al. 2017. Neural discrete representation learning. *Advances in neural information processing systems*, 30.
- Wang, L.; Lu, H.; Wang, Y.; Feng, M.; Wang, D.; Yin, B.; and Ruan, X. 2017. Learning to detect salient objects with image-level supervision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 136–145.
- Wang, P.; Bai, S.; Tan, S.; Wang, S.; Fan, Z.; Bai, J.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; et al. 2024a. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.
- Wang, X.; Darrell, T.; Rambhatla, S. S.; Girdhar, R.; and Misra, I. 2024b. Instancediffusion: Instance-level control for image generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6232–6242.
- Yan, W.; Chen, L.; Kou, H.; Zhang, S.; Zhang, Y.; and Cao, L. 2025. UCOD-DPL: Unsupervised Camouflaged Object Detection via Dynamic Pseudo-label Learning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 30365–30375.
- Zhang, H.; Qin, C.; Yin, Y.; and Fu, Y. 2023. Camouflaged Image synthesis is all you need to boost camouflaged detection. *arXiv preprint arXiv:2308.06701*.
- Zhang, J.; Huang, J.; Jin, S.; and Lu, S. 2024. Vision-language models for vision tasks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 46(8): 5625–5644.
- Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, 3836–3847.
- Zhang, Q.; Yin, G.; Nie, Y.; and Zheng, W.-S. 2020. Deep camouflage images. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 12845–12852.
- Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 586–595.
- Zhao, P.; Fan, D.-P.; Cheng, S.; Khan, S.; Khan, F. S.; Clifton, D.; Xu, P.; and Yang, J. 2025. Deep learning in concealed dense prediction. *arXiv preprint arXiv:2504.10979*.
- Zhao, P.; Xu, P.; Qin, P.; Fan, D.-P.; Zhang, Z.; Jia, G.; Zhou, B.; and Yang, J. 2024. Lake-red: Camouflaged images generation by latent background knowledge retrieval-augmented diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4092–4101.
- Zheng, C.; Cham, T.-J.; Cai, J.; and Phung, D. 2022. Bridging global context interactions for high-fidelity image completion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11512–11522.
- Zhou, M.; Leng, H.; Fang, B.; Xiang, T.; Wei, X.; and Jia, W. 2023. Low-light image enhancement via a frequency-based model with structure and texture decomposition. *ACM Transactions on Multimedia Computing, Communications and Applications*, 19(6): 1–23.