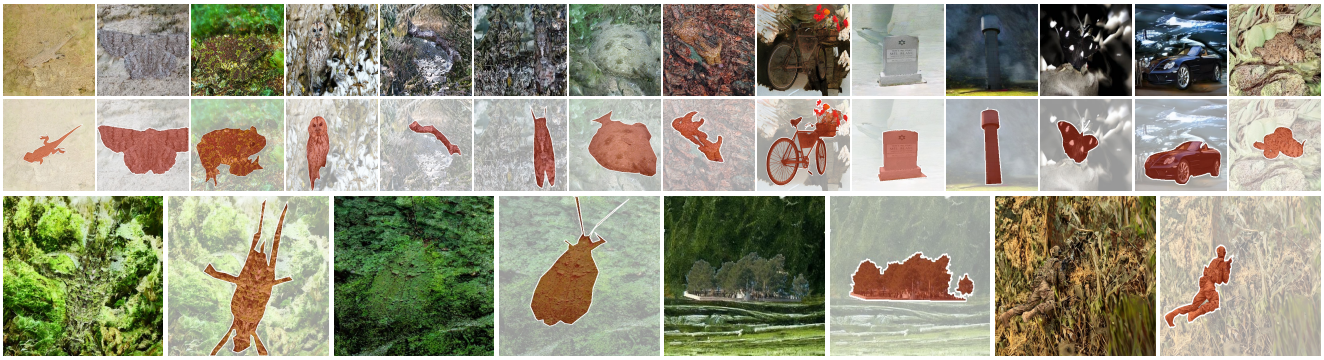
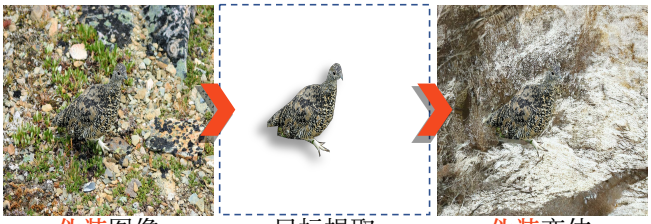
基于背景知识检索和增强的伪装图像生成隐空间扩散模型

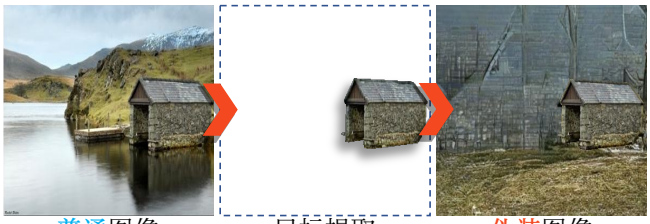
赵攀诚 [1,2]　徐　鹏 [3†]　秦鹏达 [4]　范登平 [2,1]
张知诚 [1,2]　贾国力 [1]　周伯文 [3]　杨巨峰 [1,2]
[1] 南开大学计算机学院　[2] 南开国际先进研究院（深圳 · 福田）
[3] 清华大学电子工程系　[4] 阿里巴巴



(a) 本文方法生成的伪装图像。



(b) 由伪装图像生成新的伪装图像。



(c) 将普通图像转化为伪装图像。

图 1. **基于背景知识检索和增强的扩散模型，由给定前景区域生成逼真的伪装图像**。(a) 本文提出的方法不需要任何人为指定的背景，即可自动生成足以隐藏前景物体的伪装图像。(b) 和 (c) 展示了本文方法在两种应用场景下的伪装图像生成过程。

摘要

伪装视觉感知是一项具有许多实际应用的重要视觉任务。由于昂贵的数据收集和标注成本，该领域面临一个主要瓶颈，即数据集中的物体仅限于少数物种类别。然而，现有的伪装图像生成方法需要人为指定背景，因此未能以低成本方式扩展伪装样本的多样性。本文中提出了一种用于伪装图像生成的隐空间背景知识检索增强扩散方法（*Latent Background Knowledge Retrieval-Augmented Diffusion, LAKE-RED*）。本文的贡献主要包括：（*1*）本文首次提出了一种无需接收任何背景输入的伪装图像生成范式。（*2*）*LAKE-RED* 是第一个具有可解释性的知识检索增强方法，本文提出将知识检索和推理增强明确解耦的方案，以减轻任务特定的挑战。此外，本文的方法不局限于特定的前景目标或背景，为将伪装视觉感知扩展到更多领域提供了可能性。（*3*）实验结果表明，本文的方法优于现有方法，可以生成更真实的伪装图像。本文的代码公开在 *https://github.com/PanchengZhao/LAKE-RED*。[1]

[1]本文是 CVPR 2024 论文 [62] 的中文译稿。通讯作者：徐鹏；译者：赵攀诚、刘兴宇；校稿：范登平。

1. 引言

• **背景** 伪装视觉感知 [14] 是一个具有挑战性的问题 (例如：伪装物体检测 [13])，它旨在感知复杂的伪装模式，并广泛地应用到各个领域，如害虫检测 [11]，医疗保健 [19, 22, 51]，以及自动驾驶 [3, 20, 28, 32, 44]。尽管伪装视觉感知领域在近年来取得了显著的进展，然而，对这类过于复杂的视觉场景和模式进行以像素级别的掩码标注是非常耗时费力的。事实上，在 COD10K 数据集中，为一个数据标注实例级的标签大约需要 60 分钟 [12]，远大于 COCO-Stuff 数据集单个标签标注所需的 3 分钟 [5]，这个例子清楚地说明了这个问题。因此，该社区面临着一个重大瓶颈，即目前数据集的物种类别仅限于少数目标物种，大部分为动物类别。

• **现有的技术限制** 最近，人工智能生成内容 (AIGC) 社区的快速发展，特别是基于 GAN 模型 [8] 和扩散模型 [18] 的生成式模型，揭示了使用合成数据来解决数据稀缺问题的潜力。DatasetGAN [55] 和 BigDatasetGAN [27] 通过训练一个浅层解码器，从预训练 GAN 的特征空间中生成像素级标签。DiffuMask [49] 受到了扩散模型中注意力图的启发，通过文本和图像的交叉注意力过程获取像素级别的标签。然而，上述方法是为了通用场景而设计的，所以生成的数据与伪装视觉感知任务的训练数据存在显著的领域差距。此外，如图2所示，现有的伪装图像生成方法需要手动指定背景，因此无法以低成本方式扩展伪装样本的多样性。

• **研究动机** 本文的想法是充分利用伪装场景的特定领域特征来实现低成本的解决方案。如图2所示，目标伪装的程度在很大程度上取决于其周围的环境背景。此外，大多数伪装图像采用了一种背景匹配的感知欺骗策略，即隐藏的物体无缝嵌入周围的背景。在这种场景下，伪装图像的前景和背景呈现出显著的视觉感知一致性。例如，藏在草地表面的青蛙呈现出绿色和棕色的斑驳图案，与草地和地面一致。前景和背景之间的这种特征趋同性质使得通过前景特征检索和推理背景图像成为可能。

• **方法概述** 受到以上动机的启发，本文提出了 LAKE-RED，这是一个能够自动生成高质量伪装图像和像素级分割掩码的框架。该模型接受一个前景物体作为输入，实现物体到背景的图像修复。具体而言，模型首先从前景中感知特征，并将它们用作查询，从预先构建的外部知识库中检索潜在的背景知识。然后，模型通过使用检索到的知识进行伪装背景重建，学习从前景物体到背景场景的推理。这有助于模型实现更丰富的条件引导背景生成。同时，这种合成方法保留了精确的前景标注，并防止了由生成掩码而引起的边界模糊。图1展示了 LAKE-RED 生成的伪装图像对，以及两个应用场景的示例。在无需手动指定背景输入的情况下，本文提出的模型能够以较低成本高效地生成高质量的伪装图像。

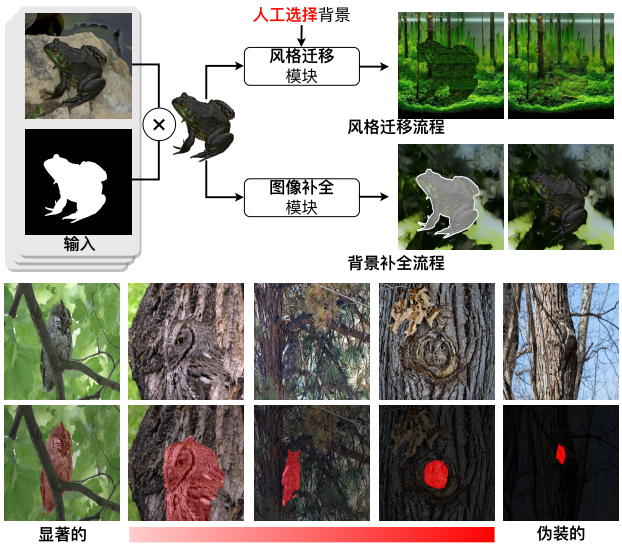


图 2. **伪装图像生成框架比较**。现有的方法依赖于人工指定的背景，不仅在多样性和范围上受到人类自身认知的的限制，还导致了大规模图像生成的高昂成本。本文观察到，即使不改变自身的颜色和纹理，同一个目标在不同环境中也可以呈现出不同程度的伪装。受此启发，本文通过一个背景修复流来合成伪装图像，通过自动选择适合的背景对给定前景物体进行隐藏。

• **本文贡献** (1) 本文首次提出了一种不需要任何背景输入的伪装生成范式。(2) 本文提出的 LAKE-RED 是第一个用于伪装生成的具有解释性的知识检索增强方法。其中，提出了将知识检索和推理增强明确分开的想法，以减轻特定任务所面临的挑战。此外，本文的方法并不局限于特定的前景或背景，为将伪装视觉感知扩展到更多领域提供了可能性。(3) 实验结果表明，本文的方法优于现有方法，能够生成更逼真的伪装图像。

2. 相关工作

• 合成数据集生成 合成数据因其低成本而受到广泛关注，是解决深度学习方法中数据瓶颈的主要方法之一[24, 40]。以前关于合成数据集的研究主要集中在制作高质量的 3 维环境模拟场景，并从中生成图像或视频数据。这些数据已被广泛用于诸如识别 [23, 46, 47, 63]、分割 [6, 34, 48, 57]、目标跟踪 [33, 58]、图像和视频理解 [52, 56, 59–61, 64]、光流估计 [4, 35] 以及 3 维重建 [67–69] 等任务。合成数据在模拟场景中的分布与真实数据之间存在的巨大差异限制了它们的有效性。最近，生成模型技术取得了重大进展，使得合成数据与真实数据之间的领域差距得以缩小。利用先进的生成模型（如：GAN、DALL-E2 和稳定扩散模型）生成的逼真图像数据，一些研究尝试探索合成数据替代真实数据的潜力 [15, 16, 30]。具体而言，DatasetGAN [55] 和 BigDatasetGAN [27] 在用有限标注数据生成大量带有分割掩码的合成图像方面表现出色。另一方面，Diffumask [49] 完全依赖文本监督，从文本和图像的交叉注意力图中提取语义标签。

• 伪装图像生成 伪装图像与普通图像不同，因为它们包含一个或多个隐藏的物体 [12]。尽管伪装的概念可以追溯到达尔文的进化理论 [9, 39, 42]，并且长期以来一直在各个领域中使用，但伪装图像生成的任务直到 2010 年才被朱等人才提出 [7]。他们所提出的模型接收指定的前景和背景作为输入，并使用手工设计的特征来赋予前景与背景相似的纹理细节，从而使人类难以识别被隐藏的物体。最近深度学习方法在风格转移和图像合成方面的进展为生成伪装图像提供了新的思路。后续的模型，如张的 [53] 和李的 [29]，通过风格转移和结构对齐的方式将前景与背景组合，进一步提高了伪装图像的生成质量。然而，由于人类认知的局限性，使用人工指定的背景会增加数据采集的成本，并限制生成图像的多样性。这些限制使得无法生成大规模的数据集，极大地降低了生成图像的应用价值。

3. 方法介绍

本文的目标是通过自动补充特定前景物体的背景区域来生成伪装图像，从而生成前景物体隐藏在生成背景中的逼真图像。尽管在伪装图像生成方法方面取得了进展，但由于高昂的人力成本和有限的认知范围，手动指定背景并不实际。通过对伪装现象的观察，本文注意到伪装图像的背景区域通常与前景物体的表面具有相似的图像特征。这表明一个合适的伪装背景可能已经存在于前景图像本身。形式上，给定原图像 $\mathbf{x}_s \in \mathbb{R}^{H\times W\times 3}$，其中包含一个不规则形状的物体。物体的位置由一个与原始图像 $\mathbf{x}_s$ 大小相同的二值掩码 $\mathbf{m}$ 精确指出，其中 $\mathbf{m}=0$ 表示需要在后续操作中需要保留的前景物体区域，$\mathbf{m}=1$ 表示可编辑的背景区域。该模型以 $\{\mathbf{x_s}, \mathbf{m}\}$ 作为输入，并输出一幅伪装图像 $\mathbf{x_c}$. 目标是从前景 $\mathbf{x_s}\odot\mathbf{m}$ 中获取一个先验，生成一个合适的伪装背景来替换原始背景。前景 $\mathbf{x_s}\odot\mathbf{m}$ 应与新背景和谐匹配，并整体呈现出伪装效果。

3.1. 序言

• 潜在扩散模型。 为了生成高质量的伪装图像，本文提出的方法基于经典的潜在扩散模型（Latent Diffusion Models, LDM）[41]。与其他概率模型类似，潜在扩散模型通过自监督训练学习给定图像集 x 的概率分布 $p(x)$，并通过逆转马尔可夫前向过程实现高质量图像生成。具体而言，前向过程会在原始图像 $\mathbf{y_0}=\mathbf{x_s}$ 中添加一个序列噪声从而获得加噪的图像集合 $\{\mathbf{y_t} \mid t\in[1,T]\}$，其中 $\mathbf{y}_t=\alpha_t\mathbf{y}_0+(1-\alpha_t)\epsilon$。随着时间步长 t 的增加，α_t 减小，$\mathbf{y_0}$ 中会被引入更多的高斯噪声 ϵ。图像的生成过程可以被描述为一系列用于去噪的自动编码器 $\epsilon_\theta\left(\mathbf{y}_t, \mathbf{c}, t\right)$，用于预测输入 $\mathbf{y}_t$ 的去噪变体。此外，为了降低高分辨率图像合成对模型的计算能力的要求，使用了一个预先训练好的自编码器 ε 将 $\mathbf{y}$ 将输入图像编码成一个隐空间的表示 $\mathbf{z}=\varepsilon(\mathbf{y})$，其中 $\mathbf{z}\in\mathbb{R}^{h\times w\times c}$。因此，训练目标可以定义为以下损失函数：

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

对于一个图像修复模型，条件 $\mathbf{c}$ 包含 $\mathbf{x_s}\odot\mathbf{m}$ 和 $\mathbf{m}$ 来指定保留区域。经过 T 步去噪后，模型预测出隐空间的表示 $\mathbf{z}_0$，其中的噪声 ϵ 已被完全去除。

最后，为了从隐空间表示中重建高分辨率图像，在模型的最后阶段使用了基于 VQVAE [43] 的解码器 $\mathcal{D}$。通过在解码器中加入一个量化层 ν，将来自码本 $\mathbf{e}$ 中的视觉信息嵌入隐空间的特征中，过程表示为：

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

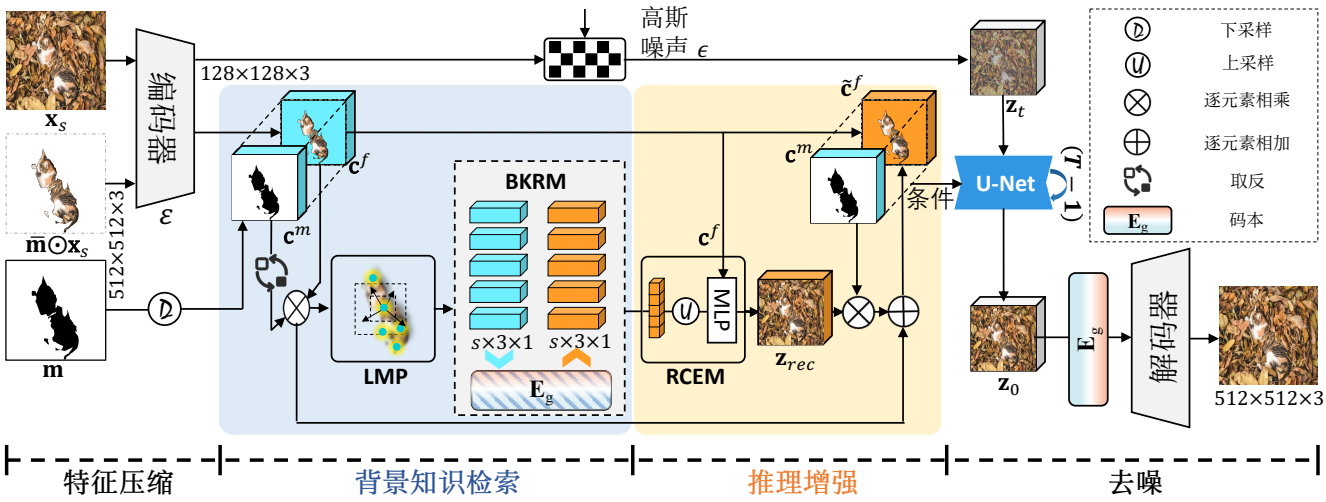


图 3. **伪装图像生成框架 LAKE-RED**。本文的框架主要包含三部分：（1）局部掩码池化（Localized Masked Pooling, LMP）用于提取前景区域的视觉表示。（2）背景知识检索模块（Background Knowledge Retrieval Module, BKRM）用于从码本中检索与背景相关的特征。（3）推理驱动的条件增强模块（Reasoning-Driven Condition Enhancement module, RCEM）使得模型通过背景重建，学习前景到背景的推理。

其中 $e \in \mathbb{R}^{K \times D}$，$K$ 和 D 分别表示离散的隐空间大小和每个潜在嵌入向量的维度。

3.2. 模型设计

当前的图像修复方法接受一个条件输入 $\mathbf{c}$，其中包括已知的图像区域并指出哪些区域是可编辑的，定义为如下表示：

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

其中 $I_{known} = \mathbf{x_s} \odot \mathbf{m}$，同时 $\mathbf{m}$ 被下采样因子 $f = 2^n$ 缩小到与隐空间特征表示一样的大小。不过，它们往往会优先考虑保持图像中物体结构的连续性，并推断填补缺失区域。当不可编辑区域为一个完整的物体，且与背景缺乏结构连续性时，模型性能就会受到限制。这意味着当前的条件输入不足以帮助模型从前景物体到背景场景做出准确的推断。为了减轻这一性能瓶颈对结果的负面影响，如 图3 所示，本文将重点放在检索更丰富的背景知识上，并开发了基于推理增强的背景重建任务，使模型能够显式地学习伪装图像的前景和背景之间的关系。重建后的特征可用于优化现有条件输入，并为模型去噪过程提供更丰富的指导信息。

3.2.1 背景知识检索

如前文所述，从物体到背景的推理是图像修复模型面临的一大挑战。然而，与通用图像不同，伪装图像的主要特点是背景匹配，即背景和物体在纹理方面表现出高度的一致性。这意味着利用前景特征检索背景知识变得可行。目前的训练框架通过对真实伪装图像进行背景掩码再重建的操作隐式地建模物体与背景之间的关系，这导致模型对二者间纹理一致性的关注不足。显式构建与前景物体特征对齐的背景特征是一种可行的选择，可为去噪过程提供更丰富的指导。为了获得有关背景纹理的特征表示，本文从 LDM 使用的基于 VQVAE 的自动编码器和解码器中汲取了灵感。

在训练过程中，VQVAE 会在编码器和解码器之间的嵌入空间中构建一个码本 e 。在解码器之前，可以通过矢量量化操作向隐空间的表示中注入码本的特征，以获得更好的重建性能。为了解决条件输入中背景特征缺失的问题，将预先训练好的码本作为全局视觉嵌入 $E_g = e^T \in \mathbb{R}^{D \times K}$ 复制并前移到迭代去噪的过程中。使用隐空间码本 E_g 获取背景特征 x^b 的过程可概括为：

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

本文将前景特征 x^f 输入有 H 个头的多头注意力层（Multi-Head Attention，MHA）中作为查询，从码本 E_g 中检索与之相关的背景特征，从而获得与背景对齐的视觉特征 x^b。

3.2.2 局部掩码池化

本文引入了一个简单高效的隐空间背景知识检索模块，记为 $\mathcal{B}\left(x^f, E_g\right)$，该模块利用前景特征 x^f 从码本 E_g 中检索背景对齐的视觉特征 x^b。从 c^f 提取到的特征表示 x^f 的丰富程度会直接影响从码本中提取的特征的有效性。由此，前景特征表示 x_f 可能成为另一个潜在的性能瓶颈。为了在特征提取过程中排除背景区域的特征，一种简单直接的方法是使用掩码平均池化（Masked Averaged Pooling, MAP）[54]，以获取前景特征的代表向量，具体方法如下：

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

其中 $i \in \{1, 2, \ldots, \vartheta\}$ 指示通道编号。MAP 将前景特征视为一个整体，并将其压缩为一个统一的表示，这会导致信息的大量丢失。特别是，操作 $\varepsilon(\cdot)$ 保持特征的通道数为 3，这导致 $x^f \in \mathbb{R}^{3\times 1}$。这种简单的表示法不足以捕捉前景的丰富特征，会限制隐空间背景知识检索的效果。

伪装图像中的前景物体往往表现出复杂的视觉特征，本文将其定义为 s 个子特征的组合。s 的值越高，相应的特征就越复杂和细致。为了提取更丰富的前景特征，本文将重点从全局转向局部，并采用简单线性迭代聚类 SLIC（Simple Linear Iterative Clustering）[1] 算法将前景区域聚类为 s 个超像素区域，过程表示为：

$$\mathbf{p}_1^i, \mathbf{p}_2^i, \cdots, \mathbf{p}_s^i = \mathcal{S}(\mathbf{c}_i^f, \mathbf{c}^m),$$

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

3.2.3 推理驱动的条件增强

进一步，本文对获得的背景知识特征 x^b 进行上采样，然后与原始前景特征 c^f 融合，尝试推理重建真值 GT 的图像特征 $\mathbf{z}_0 = \varepsilon(\mathbf{y}_0), \mathbf{z}_0 \in \mathbb{R}^{h\times w\times c}$。重建过程可以表示为：

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

然后，利用 $\mathbf{z}_{rec}$ 完善和增强输入的初始条件。为了强调背景特征，本文创建了一个特征重构任务，以增强模型利用背景知识推理真实背景特征的能力。具体来说，本文用重建的 $\mathbf{z}_{\mathbf{rec}}$ 填充 c^f 的背景区域，以加强条件中蕴含的信息，同时保持前景区域不变。增强条件的策略可表述为：

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

因此，背景重建损失定义为：

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

模型训练的整体损失被定义为：

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

利用伪装图像的特性，本文完善并增强了输入条件 $\mathbf{c}$。在保留前景区域图像特征的同时，使用增强条件 $\mathbf{e}$ 引导背景的生成。隐式和显式约束共同帮助模型学习前景物体和背景之间的纹理一致性，从而生成高质量的伪装图像。

4. 实验

4.1. 实验设置

• **数据集** 与现有的工作 [13] 相同，对于伪装目标检测领域，有 4040 对伪装图像和掩码对被用于训练模型，其中 3040 张图像来自数据集 COD10K [12]，1000 张来自于数据集 CAMO [26]。为了验证模型的生成性能，本文收集了来自不同领域的图像掩码对，构建了一个大规模测试数据集，其中包括三个子集：伪装物体（CO）、显著物体（SO）和通用物体（GO）。在 CO 中，有 6473 对来自 CAMO [26]、COD10K [12] 和 NC4K [38] 三个常见伪装目标检测数据集的真实伪装图像。然后本文从常见的显著物体检测数据集 DUTS [45]，DUT-OMRON [50] 等中选取了相同数量的显著物体图像对，从分割数据集 COCO2017 [31] 中选取了相同数量的通用图像对，以评估模型在开放领域数据上的性能。

• **评价指标** 遵循 AIGC [27, 41] 和 COD [25, 37] 的良好实践，本文选择基于 InceptionNet 的指标 FID [2] 和 KID [17] 来衡量生成伪装图像的质量。首先使用预训练的 InceptionNet 提取出图像特征，然后计算它们

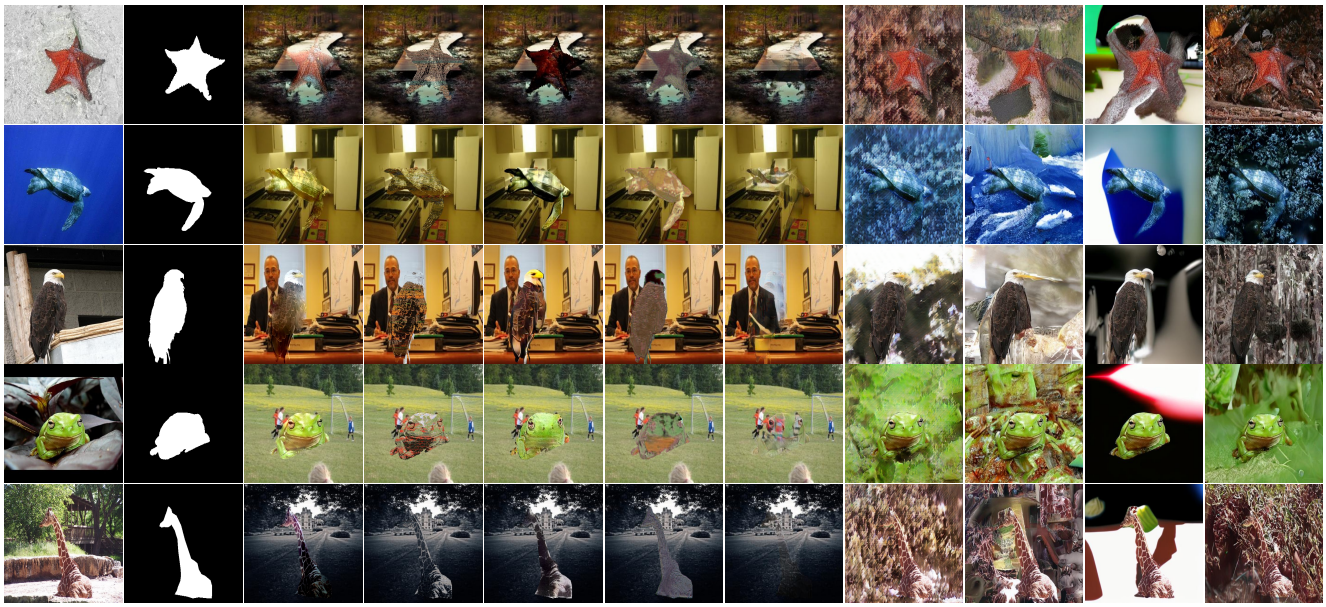


图 4. **与现有方法的性能比较.** 前两列是输入图像对，本文提供了九种方法生成的伪装图像进行比较。请注意，第 3 至第 7 列中的方法中的方法还共用了一张随机取样的背景图片作为输入。

之间的距离以显示模型输出与目标数据集之间的相似程度。与一般图像不同，合成良好的伪装图像不应包含易于识别的物体，因此提取具有鉴别力的特征更具挑战性 [37] 。数值越小，表示生成的图像与真实伪装图像越相似。

• **实现细节** 为了给指定的前景图像生成伪装图像，本文使用了一个强大的隐空间扩散模型 [41] 作为初始化，该模型在图像修复任务中经过了预先训练。该模型被设计用于处理大小为 512×512 的图像和掩码，并使用预先训练的 VQVAE [43] 将其压缩到 $128 \times 128 \times 3$ 的潜空间。在训练过程中，本文专注于训练用于去噪过程的 U-Net，而不对自动编码器和解码器组件进行微调。通过本文提出的模块，完善并增强了现有条件输入。模型初始化、数据扩充和批量大小等参数优化设置与原论文类似。最后，模型生成伪装图像并调整其大小，使其与原始输入图像保持一致。本文使用 PyTorch 和 GeForce RTX 3090 GPU 进行所有实验。

4.2. 与现有方法的比较

以往的伪装图像生成方法基于图像融合或风格迁移，与本方法存在差异。因此，对于每种解决方案，本文都选择了最先进的方法进行比较。对于图像融合和风格迁移的方案，模型在接受前景输入时需要手动指定背景图像。本文使用大规模场景数据集 Places365 [66] 作为背景图片的来源。对于给定的前景输入，本文从 Places365 中随机抽取一张背景图片，将其调整为与前景图像相同的大小，然后进行图像合成。为了便于不同方法之间的比较，所有方法都对给定的前景输入共享相同的背景图片。对于图像修复的方案，模型只接受一个前景输入，并生成一个伪装图像作为输出。

• **定性分析** 图4比较了本文的方法和其他方法从普通图像生成的伪装图像时的结果质量。结果表明，尽管前景特征经过处理后与背景保持一致，但 AB 和 CI 等方法受背景图像输入的影响很大。因此，前景目标会与背景场景和物体发生冲突，例如房间里的老鹰和乌龟（第 2 行和第 3 行），以及比真人还大的青蛙（第 4 行）。LCGNet 在隐藏物体方面表现最佳，肉眼几乎看不到物体的特征。但自然界中的伪装物体会无缝嵌入背景中，而不是完全隐形。另一方面，基于图像修复的方法只需要输入前景物体，自适应背景生成可以满足大规模生成的要求。然而，现有方法存在背景缺乏真实性（TFill）、伪装程度低（LDM）和背景修补失败（Repaint-L）等问题。相比之下，本文的方法能将给定目标自然融入生成的背景中，保留目标的所有特征，同时实现图像的整体伪装。

表 1. **定量性能比较**。本文对所提出的伪装图像生成方法进行了定量评估，并将其与最先进的（SOTA）方法进行了比较。评估涉及从伪装图像、显著目标图像和通用图像中采样的特定前景物体。本文提出的方法表现出卓越的性能。

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 [65]$_{22}$	$\mathcal{F}$	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**

• **定量分析**　为评估伪装图像生成的质量，本文构建了一个大规模测试集，其中包括三种不同来源的前景物体，以评估模型对不同图像领域的适应性。显著性物体子集和通用物体子集分别从显著目标检测和图像分割领域的数据集中抽取，图像数量与 COD 测试集保持一致。使用 FID 和 KID 测量生成结果与实际 COD 数据集之间的距离，结果显示在 表1 中。

三个子集的结果呈阶梯状分布，表明模型性能受图像域差距的影响很大，一般物体比显著物体更难转换。基于图像混合的方法会产生较大的结果，因为它们会机械地将前景特征转向与背景特征一致，从而导致生成结果的视觉特征主要由背景决定。当背景图像来自于随机取样时，相关指标也会表现出一定程度的随机性。另一方面，基于图像修复的方案倾向于为物体生成合适的背景，通常表现出更好的性能。

此外，本文还发现 LCGNet 在一般对象子集上的验证结果存在异常值，这是以下原因共同造成的。首先，模型在三个子集上的合成难度逐渐增加。伪装物体本身来自隐蔽的场景，易于隐藏。显著物体的大小和位置适中，通常具有完整的结构。通用物体种类丰富，大小不一，要为其找到合适的伪装环境具有挑战性。随着复杂性的增加，这些方法逐渐难以完美地隐藏通用目标，导致在特定子集内的性能下降。在这种情况下，LCGNet 最大限度地舍弃了前景特征，结果主要取决于随机采样的背景（图4）。它受前景的负面影响最小，同时受背景的随机性影响最大，因此会产生异常结果。不过，本文的方法在整个测试集上取得了最佳性能。

• **用户调研**　由于图像生成质量和伪装效果都需要基于人类感知进行评价，本文进行了用户研究，以获得人对生成结果的主观判断。为此，本文沿用了前人在伪装图像生成方面的研究方法，随机选取了 20 组前景图像，并采用不同的方法生成结果。对于基于风格迁移的方法，本文使用了从 Places365 中随机抽取的额外图片作为背景输入，且所有方法都保持一致的背景图像。本文邀请 20 位参与者根据三个问题对结果进行评分：

-Q#1: 哪个结果中的目标最难找到？

-Q#2: 哪个结果伪装最自然？

-Q#3: 哪个结果最接近真实的伪装图像数据集？

对每个问题，参与者需要选择排名前 3 的选项，其中 1 为最高。用户调查结果显示在 图5 中。虽然 LCGNet 在 **Q#1** 中获得了更多的选票，但这是因为其生成的结果中几乎失去前景特征。本方法被认为在视觉呈现方面产生了更自然、更接近真实数据集的结果。

表 2. **定量消融实验**。通过逐步将每个模块添加到基础模型中，比较它们对生成结果的质量和成本的影响。结果表明，本文提出的方法是有效的，而且几乎不需要成本。

Module			Prams	MAC	FPS	Overall	
BKRM	RCEM	LMP	(M)↓	(G)↓	(Hz)↑	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

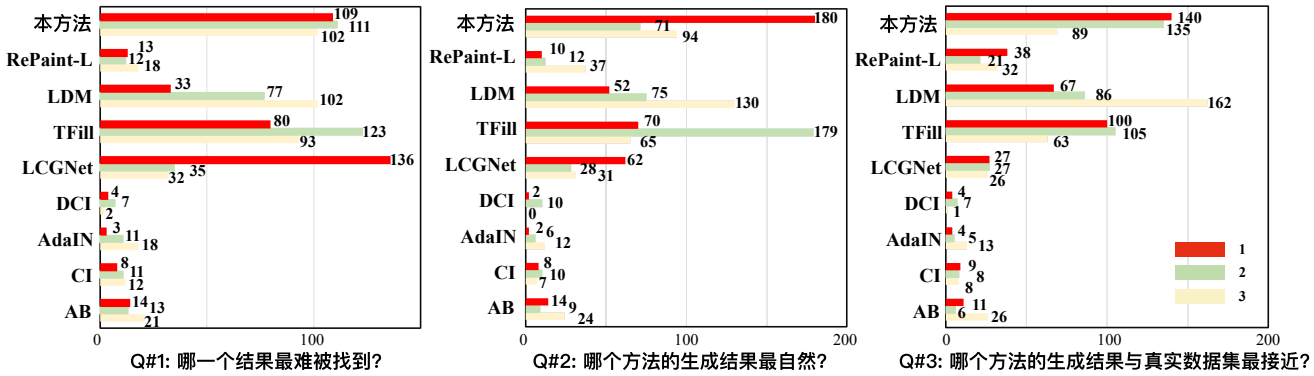


图 5. **对 9 种不同方法生成的伪装图像的用户调研**。本文的方法被认为能产生最自然、视觉效果最接近真实伪装图像的结果。

图 6. **消融实验的可视化分析**。本文对消融实验过程中的样本进行了可视化展示，以直观显示所提出的模块的有效性。

4.3. 消融实验

本文通过向基础的 LDM 逐步添加模块来进行消融实验，以评估所提出的方法中每个组件的有效性。如表2 所示，随着本文提出的模块的引入，生成的伪装图像质量逐渐提高，证明了其有效性。当同时应用本文提出的三个模块时，模型性能达到顶峰，在 FID 和 KID 指标上分别提高了 33.14% 和 28.71%。此时，引入这三个模块仅为模型增加了约 0.01M 个参数和 0.02G 的计算量，推理速度仅降低了 0.04Hz。这些结果清楚地表明，本文的方法是有效的，而且几乎不增加任何成本。

本文进一步对消融实验过程中的样本进行可视化展示，以直观展示本文提出的模块的有效性。图6 中显示了其中两组结果。在从目标到背景的图像生成过程中，LDM 在关注伪装特性方面面临挑战。由于任务的复杂性，它在特定区域生成背景时失效，导致出现黑色色块。通过结合隐空间背景知识检索模块（BKRM），模型显式学习前景与背景之间的相似性，从而使生成的背景与前景更接近。此外，推理驱动的条件增强模块（RCEM）通过加入背景重建损失来增强场景的真实感，迫使模型进行推理并准确重建背景特征。最后，局部遮蔽池化（LMP）的引入将模型的注意力从全局特征转移到局部前景特征，增强了生成背景的纹理多样性。

5. 结论

本文提出了一种用于伪装图像生成的潜在背景知识检索和增强的扩散模型（LAKE-RED）。与现有方法不同，本文的生成范式无需人为指定背景。通过知识检索和推理增强，本文从前景中获得了强大的背景条件，从而能够生成高质量的伪装图像，超越其他 SOTA 的伪装图像生成方法。本文的方法不局限于特定的前景目标或人类选择的背景。这使其能够大规模生成伪装图像，并为将来将伪装视觉感知扩展到更多领域提供了可能性。

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. Coco-stuff: 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, 6

[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] Pancheng Zhao, Peng Xu, Pengda Qin, Deng-Ping Fan, Zhicheng Zhang, Guoli Jia, Bowen Zhou, and Jufeng Yang. Lake-red: Camouflaged images generation by latent background knowledge retrieval-augmented diffusion. In *CVPR*, 2024. 1

[63] 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

[64] 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

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

[66] 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

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

[68] 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.

[69] 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