西北工业大学

数字图像处理–论文翻译

原论文标题：Shadow Generation for Composite Image Using Diffusion Model

崔宇婷

计算机学院

计算机科学与技术

2024 年 11 月

学号：2022302610

Northwestern Polytechnical University

基于扩散模型的合成图像阴影生成

刘庆阳[1], 游俊琪[1], 王建廷[1], 陶欣浩[1], 张波[1], 李牛[1,2*]
[1] 上海交通大学 [2] miguo.ai
[1]{narumimaria,yjqsjtu2022,glory1299,taoxinhao,bo-zhang,ustcnewly}@sjtu.edu.cn

Abstract

在图像构图领域，为插入的前景生成逼真的阴影阴影仍然是一项艰巨的挑战。以前的工作开发了图像到图像翻译模型，这些模型是在配对训练数据的基础上进行训练的。然而，这些模型在生成具有准确形状和强度的阴影稀缺性和任务固有的复杂性。在本文中，我们借助具有丰富先验知识的基础模型自然阴影图像。具体来说，我们首先调整 *ControlNet* 以适应我们的任务，然后提出强度调制模块来改善阴影强度。此外，我们还利用新颖的数据采集管道将小规模 *DESOBA* 数据集扩展到 *DESOBAv2*。在 *DESOBA* 和 *DESOBAv2* 数据集以及真实合成图像上的实验结果表明我们的模型在阴影生成任务方面的卓越能力。数据集、代码和模型发布于 *https://github.com/bcmi/ObjectShadow-Generation-Dataset-DESOBAv2*。

1. 介绍

图像合成 [28] 的目的是将一幅图像的前景与另一幅背景图像合并，生成一幅合成图像，它在虚拟现实、艺术创作和电子商务等方面有着广泛的应用。简单地将前景粘贴到背景 [3], 上往往会造成视觉上的不一致，包括前景和背景之间的光照不协调、缺乏前景阴影/反射 [12, 34], 本文重点讨论阴影问题，即插入的前景在背景上没有可信的阴影，这会大大降低合成图像的真实感和质量。

如图 1 所示 1, 阴影生成是一项具有挑战性的任务，因为前景阴影由许多复杂因素（如照明信息和前景/背景的几何形状）确定。现有的阴影生成方法可以分为基于渲染的方法 [34–36] 和基于非渲染的方法 [12, 22, 53]. 基于渲染的方法通常对几何和光照进行限制性的假设，这在现实场景中很难得到满足。此外，[35, 36] 要求用户指定照明信息，这阻碍了其在我们的任务中的直接应用。基于非渲染的方法通常基于没有前景阴影的合成图像和具有前景阴影的真实的图像对来训练图像到图像转换网络。然而，由于训练数据的稀缺性和任务的艰巨性，这些方法难以生成具有合理形状和强度的阴影。

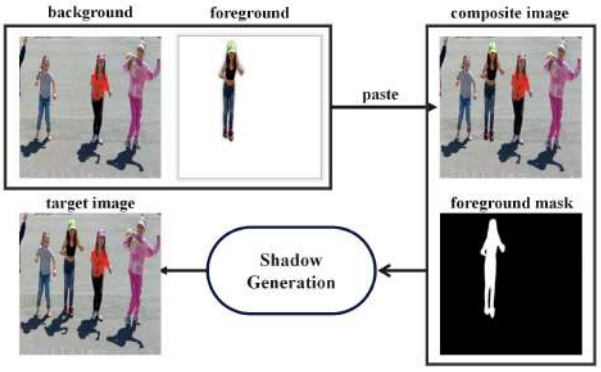

图 1. 将前景粘贴到背景上就能得到合成图像。阴影生成的目的是为合成图像中插入的前景生成可信的阴影，从而生成更逼真的图像。

最近，基金会模型（例如，稳定扩散 [32]）在大规模数据集上预训练，已经证明了图像生成和编辑的前所未有的潜力。在先前的关于对象引导的修补或合成的工作中 [44, 48] 他们表明，即使不考虑阴影问题，所生成的前景也伴随有阴影，这可能是因为基础模型中自然阴影图像的丰富先验知识。但是，它们只能在简单的情况下生成令人满意的阴影，并且对象外观可能会意外改变。

*Corresponding author.

我们的方法建立在条件基础模型 [52] 之上，并提出了几项关键创新。首先，我们修改控制编码器输入和噪声损失以适应我们的任务。我们观察到的是，生成的阴影强度（黑暗程度）并不令人满意。特别是当背景物体有阴影时，前景阴影和背景阴影之间的强度不一致会使整个图像变得不真实。因此，我们引入另一个强度编码器来调制前景阴影强度。具体来说，去噪 U-Net 被修改为输出噪声图和前景阴影掩模。强度编码器接收合成图像和背景阴影掩模，产生比例/偏差以调制前景阴影区域内的预测噪声。最后，我们设计了一个后处理网络来纠正颜色偏移和背景变化。

模型训练需要大量的无前景阴影的合成图像和有前景阴影的真实的图像对。现有的真实世界阴影生成数据集 DESOBA [12] 受到规模的限制（即，、1,012 张真实的图像和 3,623 对），这是由于手动阴影去除的成本很高，不足以训练我们的模型。为了确保充分的监督，我们设计了一个新的数据构造管道，它将 DESOBA 扩展到 DESOBAv 2（即，，21，575 个真实的图像和 28，573 对）。具体来说，我们首先收集大量的现实世界的图像与一个或多个对象阴影对。然后，我们使用预训练的对象-阴影检测模型 [41] 来预测对象-阴影对的对象和阴影掩模。接下来，我们应用预训练的修复模型 [32] 来修复检测到的阴影区域以获得去阴影图像。最后，基于真实的图像和去阴影图像，我们构造成对的合成合成图像和地面实况目标图像。

我们在 DESOBAv2 和 DESOBA 数据集上进行实验。结果表明，在利用大规模数据和基础模型的好处后，阴影生成任务有了显着的改善。我们的主要贡献可以总结如下：1）我们贡献了 DESOBAv2，一个大规模的真实世界的阴影生成数据集，这可以大大促进阴影生成任务。2) 我们提出了一个尖端的扩散模型，专门设计用于产生复合前景的阴影。3) 通过综合实验，我们验证了我们的数据集构建管道的有效性和我们提出的模型的优越性。

2. 相关工作

2.1. 图像合成

图像合成的目的是在背景图像上覆盖前景对象，以产生合成结果 [20, 22, 42, 46, 47]。以前的研究工作已经解决了可能损害合成图像质量的不同问题。例如，图像混合方法 [31, 42, 49, 51] 的目标是无缝地组合前景和背景。图像协调方法 [3–6, 40] 旨在纠正前景和背景之间的照明差异。尽管如此，上述方法在很大程度上忽略了前景投射到背景上的阴影。近来，生成式图像合成方法 [38, 44, 48] 可以将前景对象插入到背景中的边界框中，并且所插入的对象可能具有阴影效果。但是，它们只能在简单的情况下生成令人满意的阴影，并且对象外观可能会意外改变。

2.2. 阴影生成

在本文中，阴影生成任务的目标是生成逼真的阴影的复合前景。现有的方法可以被广泛地分类为基于渲染的方法和基于非渲染的方法。基于渲染的方法需要全面了解各种因素，如照明、反射、材质属性和场景几何体，以便为插入的对象生成阴影。然而，这种详细的知识依赖于用户输入 [15, 16, 21, 35, 36] 或模型预测 [1, 7, 19, 50],，这是劳动密集型的或不可靠的 [53]。例如，[35, 36] 可以通过用户控制产生令人信服的结果。然而，在合成图像中，照明信息应该从背景中自动推断出来，而不是由用户请求。

基于非渲染的方法 [12, 22, 25, 53] 旨在将没有前景阴影的输入合成图像转换为具有前景阴影的输出，从而绕过对上述因素的显式知识的需要。例如，ShadowGAN[53] 利用全局和局部条件反射来增强生成阴影的真实感。ARShadowGAN [22] 强调了背景阴影的重要性，并使用它来指导前景阴影的生成。SGRNet[12] 鼓励前景和背景之间的信息交换，并采用经典的光照模型以获得更好的阴影效果。工作 [25] 产生多个曝光不足的图像，并将它们融合以获得最终的阴影区域。DMASNet [39] 将阴影掩模预测分解为框预测和形状预测，实现了更好的跨域可转移性。

据我们所知，我们是第一个专注于阴影生成的基于扩散的方法。

2.3. 扩散模型

近年来，扩散模型已经成为图像生成和图像编辑的一个强大工具。这些模型将图像生成作为一系列随机过渡，从基本分布移动到所需的数据分布 [11].。扩散模型可以分为无条件扩散模型 [11, 37] [27, 32, 52].。无条件扩散模型的重点是通过捕捉自然图像的分布来生成逼真的图像，而不需要任何特定的输入条件。条

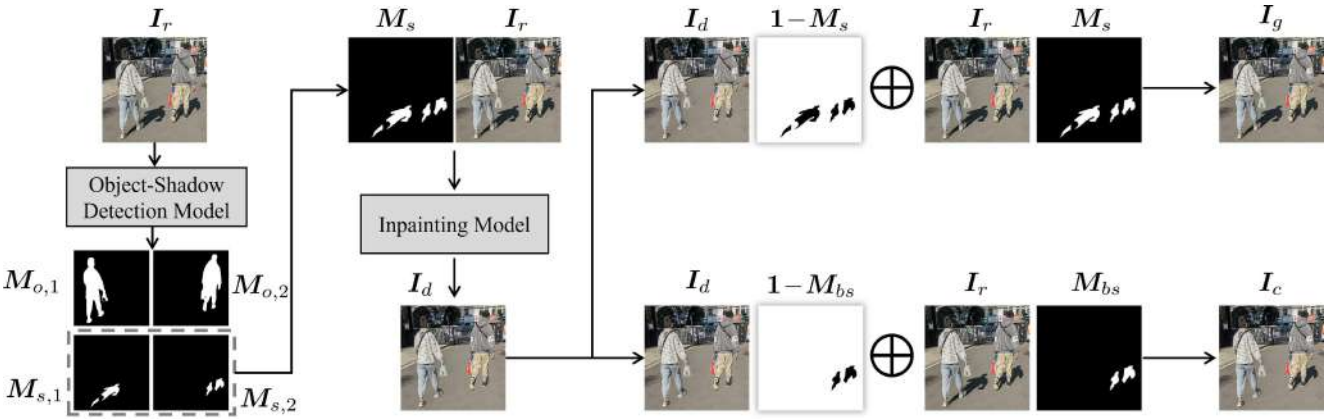

图 2. 数据集构建的管道。我们使用对象阴影检测模型 [41] 来预测真实的图像中的对象和阴影掩模对 $\boldsymbol{I}_r$. . 然后我们就能得到联盟 $\boldsymbol{M}_s$ 所有阴影蒙版中的一个作为修复蒙版，并应用修复模型 [32] 来获得去阴影图像 $\boldsymbol{I}_d$. 指定前景对象后，我们替换背景阴影区域 $\boldsymbol{M}_{bs}$ 在 $\boldsymbol{I}_d$ 与中国的同行 $\boldsymbol{I}_r$ 合成一个合成图像 $\boldsymbol{I}_c$ 并替换所有阴影区域 $\boldsymbol{M}_s$ 在 $\boldsymbol{I}_d$ 与中国的同行 $\boldsymbol{I}_r$ 获得地面真实目标图像 $\boldsymbol{I}_g$.

件扩散模型被设计成在特定条件输入（例如文本描述、语义掩码等）的指导下产生图像。ControlNet [52] 是一种流行的条件扩散模型，它为大型预训练的文本到图像扩散模型配备了空间感知和特定于任务的条件。我们建立我们的模型上 ControlNet，并提出了几个创新，以满足特定的要求，阴影生成。

3. Dataset Construction

我们的数据集构建管道如图 2所示，接下来将详细介绍。

3.1. 阴影图像采集

我们从两个来源收集了大量的真实世界的户外图像，这些图像在自然光的照射下跨越了各种场景。一方面，我们从公共网站上抓取在线图像，这些网站拥有重复使用的许可证。另一方面，我们聘请摄影师拍摄符合我们要求的户外场景照片。我们只保留至少有一个物体-阴影对的图像，达到 44，044 个图像。

3.2. 阴影消除

给定具有对象阴影对的真实的图像 $\boldsymbol{I}_r$ ，我们使用预训练的对象阴影检测模型 [41] 来预测 K 对对象和阴影掩模。我们使用 $\boldsymbol{M}_{o,k}$ (*resp.*, $\boldsymbol{M}_{s,k}$) 来表示对象（分别为阴影）掩模。我们将一个检测到的对象-阴影对称为一个检测到的实例。我们排除了没有检测到任何实例的图像。

随后，我们尝试删除所有检测到的阴影。我们已经尝试了一些最先进的阴影去除模型 [8, 9]，但由于泛化能力差，在野外的性能低于我们的预期。考虑到最近图像修复技术的快速发展 [14, 23, 29, 32, 45, 56] ，我们采用图像修复来去除阴影。虽然图像修复不能精确地保留背景信息 [10]，但我们观察到阴影区域的背景纹理通常非常简单，并且修复后的结果与原始背景具有相似的纹理。因此，我们粗略地将修复的结果视为去阴影的结果。

我们获得所有检测到的阴影掩模 $\boldsymbol{M}_s = \boldsymbol{M}_{s,1} \cup \boldsymbol{M}_{s,2} \cup \cdots \cup \boldsymbol{M}_{s,K}$[32] 来获得去阴影图像 $\boldsymbol{I}_d$。在实践中，我们观察到修复模型在某些情况下容易在修复区域中生成低质量的阴影。为了防止修复模型在修复区域产生不需要的阴影，我们采用了一些技巧，如扩大修复掩模和垂直翻转图像，可以有效地阻止修复过程中产生不需要的阴影。然而，在修补区域中可能仍然存在不期望的阴影或明显的伪影。

修复后，我们根据以下规则手动过滤对象-阴影对：1）我们使用低质量的对象遮罩或阴影遮罩移除对象-阴影对。2）我们移除那些在修复区域中具有生成阴影或明显伪影的对象-阴影对。手动过滤后，我们将剩余的对象-阴影对作为有效实例。我们有 21，575 张图片，28，573 个有效实例。

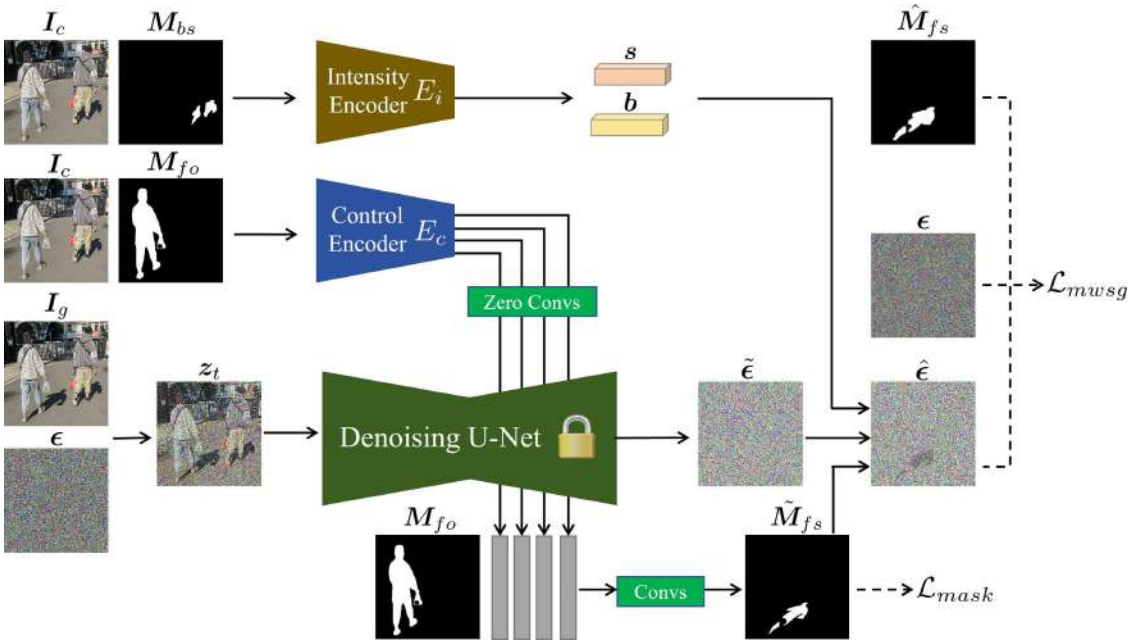

图 3. 我们的 SGDiffusion 的框架。我们适应 ControlNet（控制编码器和去噪 U-Net）的阴影生成任务。我们还引入了一个强度编码器来调制噪声图中的前景阴影区域，使其，从而导致在 $\hat{\epsilon}$ 为 0. 上的输出噪声 $\hat{\epsilon}$ is supervised by weighted noise loss $\mathcal{L}_{mwsg}$ 由基于扩展前景阴影遮罩 $\hat{M}_{fs}$

3.3. 图像合成

给定一对真实的图像 $\boldsymbol{I}_r$ 和去阴影图像 $\boldsymbol{I}_d$，我们从有效实例中随机选择第 k 个前景对象并合成合成图像。$\boldsymbol{M}_{o,k}$ (*resp.*, $\boldsymbol{M}_{s,k}$) 被称为前景对象 (*resp.*, shadow) 掩模 $\boldsymbol{M}_{fo}$ (*resp.*, $\boldsymbol{M}_{fs}$). 一种策略是用 $\boldsymbol{M}_{fs}$ 中的对应物替换 $\boldsymbol{I}_r$ 。但是，此策略可能会沿着阴影边界留下痕迹，在这种情况下，模型可能会找到生成阴影的捷径。另一种策略是用 $\boldsymbol{I}_r$ 中的对应物替换 $\boldsymbol{I}_d$ 中的其他对象的阴影区域 $\boldsymbol{M}_{bs} = \boldsymbol{M}_{s,1} \cup \cdots \cup \boldsymbol{M}_{s,k-1} \cup \boldsymbol{M}_{s,k+1} \cup \cdots \cup \boldsymbol{M}_{s,K}$ 以合成合成图像 $\boldsymbol{I}_d$ 其中只有所选前景对象不具有阴影，而所有其他对象具有阴影。我们采用第二种策略。

修复后，背景可能会发生轻微变化，因此 $\boldsymbol{I}_c$ 的背景可能与 $\boldsymbol{I}_r$ 的背景略有不同。为了确保一致的背景，我们通过用 $\boldsymbol{I}_r$ 中的对应物替换 $\boldsymbol{I}_g$ 中的所有对象的阴影。到目前为止，我们获得了 $\{\boldsymbol{I}_c, \boldsymbol{M}_{fo}, \boldsymbol{M}_{fs}, \boldsymbol{M}_{bs}, \boldsymbol{I}_g\}$ 形式的元组，它将用于模型训练。我们数据集的示例图像和更多统计数据可以在补充资料中找到。

4. Background

稳定扩散 [32] 是在潜在空间中运行的潜在扩散模型。首先，首先，利用编码器 [18] E_r 和解码器 D_r 使用 VAE 将 512×512 图像转换为 64×64。使用 E_r 将图像空间投影到潜在空间，并使用 D_r 将图像空间投影回图像空间。然后，在潜空间中执行前向扩散过程和后向去噪过程。去噪 U-Net [33] 由具有 12 块的编码器、中间块和具有 12 块的跳过连接解码器组成。

在训练期间，在去噪步骤 t 中将随机高斯噪声 $\boldsymbol{\epsilon}$ 添加到潜像 $\boldsymbol{z}_0$，产生噪声潜像 $\boldsymbol{z}_t$。给定时间步长 t 和文本提示 $\boldsymbol{c}_{txt}$，训练具有模型参数 $\epsilon_{\boldsymbol{\theta}}$ 的去噪 U-网以预测添加的声 $\boldsymbol{\epsilon}$。

为了支持空间条件信息（例如，, 边缘, 姿势, 深度), ControlNet[52] 将控制编码器 E_c 与预训练的稳定扩散集成在一起。具体地，控制编码器包含其 12 个编码块和跨四个分辨率的中间块（64×64,32×32,16×16,8×8 ）的可训练副本。它接受一个 512×512 条件图像作为输入。

从控制编码器输出的条件特征图 $\boldsymbol{c}_{img}$ 用于经由零卷积层增强去噪 U-Net[26] 中的 12 个跳过连接和中间块。虽然原始的稳定扩散是固定的以保留先验知识，但

是控制编码器可以结合附加条件来指导图像生成。

目标可以改写为

$$\mathcal{L}_{ctrl} = \mathbb{E}_{t,\boldsymbol{\epsilon}\sim\mathcal{N}(0,1)}\left[\|\boldsymbol{\epsilon} - \epsilon_{\boldsymbol{\theta}}(\boldsymbol{z}_t, t, \boldsymbol{c}_{txt}, \boldsymbol{c}_{img})\|_2^2\right]. \quad (1)$$

5. 方法

给定没有前景阴影的合成图像 $\boldsymbol{I}_c$ 以及前景对象掩模 $\boldsymbol{M}_{fo}$，我们的阴影生成扩散（SGDiffusion）模型旨在产生具有合理前景阴影的 $\tilde{\boldsymbol{I}}_g$。我们将在 5.1 节中调整 ControlNet[52] 以适应阴影生成任务，并在 5.1节中提出改进阴影强度的新模块。最后，我们将在 **??**节中简要介绍用于增强图像质量的后处理技术。

5.1. 使 ControlNet 适应阴影生成

对于阴影生成任务，有用的条件信息是输入合成图像 $\boldsymbol{I}_c$ 和前景对象遮罩 $\boldsymbol{M}_{fo}$，其中前景对象遮罩指示我们需要为其生成阴影的目标对象。我们将 $\boldsymbol{I}_c$ 与 $\boldsymbol{M}_{fo}$ 连接作为控制编码器 E_c 的输入。控制编码器输出条件特征图 $\boldsymbol{c}_{sg}$，其被注入去噪解码器以提供指导。对于文本提示符，我们尝试了几种变体，如 “the [object category] with shadow”，但它们对生成的阴影没有显著影响。因此，默认情况下我们使用空文本提示。

给定包括时间步长 t 和条件特征图 $\boldsymbol{c}_{sg}$ 的一组条件，具有模型参数 $\epsilon_{\boldsymbol{\theta}}$ 的去噪 U-Net 预测添加到噪声潜像 $\boldsymbol{z}_t$ 的噪声 $\boldsymbol{\epsilon}$：

$$\mathcal{L}_{sg} = \mathbb{E}_{t,\boldsymbol{\epsilon}\sim\mathcal{N}(0,1)}\left[\|\boldsymbol{\epsilon} - \epsilon_{\boldsymbol{\theta}}(\boldsymbol{z}_t, t, \boldsymbol{c}_{sg}))\|_2^2\right]. \quad (2)$$

为了使模型更加强调前景阴影区域，我们引入了加权噪声损失，为前景阴影区域分配更高的权重。我们通过一个扩张的核来扩展前景阴影蒙版，以获得扩展的蒙版 $\tilde{\boldsymbol{M}}_{fs}$。扩展的前景阴影区域中的权重是 w，而其他权重是 1，导致权重图 $\boldsymbol{W}_{fs}$。如果不扩大前景阴影区域，模型会被误导生成较大的阴影忽略阴影形状和边界的细节。通过将权重图 $\boldsymbol{W}_{fs}$ 应用于噪声损失，我们可以得到

$$\mathcal{L}_{wsg} = \mathbb{E}_{t,\boldsymbol{\epsilon}\sim\mathcal{N}(0,1)}\left[\|\boldsymbol{W}_{fs} \circ (\boldsymbol{\epsilon} - \epsilon_{\boldsymbol{\theta}}(\boldsymbol{z}_t, t, \boldsymbol{c}_{sg}))\|_2^2\right], \quad (3)$$

where denotes element-wise multiplication. 其中 $\circ$ 表示逐元素乘法。

$\boldsymbol{I}_c$ 的更多信息，我们通过向 $\boldsymbol{I}_c$ 的潜像添加噪声来获得 $\boldsymbol{z}_T$，而不是直接从高斯分布 $\mathcal{N}(0,1)$ 中采样。

5.2. 阴影强度调制

通过使用第5.1节中的自适应 ControlNet，我们观察到生成的前景阴影的强度不令人满意。特别是当背景具有对象阴影对时，生成的前景阴影通常比背景阴影明显更暗或更亮。前景阴影强度和背景阴影强度之间的这种不一致使得整个图像不真实。

因此，我们引入另一个强度编码器来调制前景阴影强度。具体来说，我们使用编码器 E_i 来提取强度相关信息。直观地说，通过观察背景阴影及其周围的非阴影区域，我们可以估计前景阴影的强度。因此，强度编码器 E_i 的输入应该包括合成图像 $\boldsymbol{I}_c$ 和背景阴影掩模 $\boldsymbol{M}_{bs}$。当没有背景阴影时，蒙版全黑。我们将 $\boldsymbol{I}_c$ 与背景阴影掩模 $\boldsymbol{M}_{bs}$ 连接作为强度编码器的输入。

强度编码器输出比例和偏差以调整前景阴影区域内的噪声图的强度。经调制的噪声图导致经调制的潜像，并且进一步导致经调制的前景阴影。因此，噪声图的强度调整最终体现在所生成的前景阴影的强度变化上。具体地，当噪声图具有 c 个通道时，E_i 输出 c-暗淡尺度矢量 $\boldsymbol{s}$ 和 c-暗淡偏置矢量 $\boldsymbol{b}$，其包含通道级尺度和偏置。$\boldsymbol{s}$ 和 $\boldsymbol{b}$ 用于调制前景阴影区域内的预测噪声图。

一个问题是在测试阶段前景阴影区域是未知的，因此我们需要预测前景阴影掩模。为了避免大量额外的计算成本，我们利用去噪 U-Net 中的特征图来预测前景阴影掩模。以前的作品通常在去噪 U-Net 中结合联合收割机不同层的特征图进行掩码预测。我们尝试了不同层的特征图，发现解码器特征图在阴影掩模预测中更有效。我们还使用前景对象遮罩，它可以提供有用的提示前景阴影的位置。我们将所有解码器特征映射和前景对象遮罩调整为相同大小，并按通道方式连接它们。该级联通过几个卷积层来预测前景阴影遮罩 $\tilde{\boldsymbol{M}}_{fs}$。$\tilde{\boldsymbol{M}}_{fs}$ 通过二进制交叉熵（BCE）损失和骰子损失 [24, 43] 使用地面实况前景阴影掩模 $\boldsymbol{M}_{fs}$ 进行监督：

$$\mathcal{L}_{mask} = \mathcal{L}_{bce}(\tilde{\boldsymbol{M}}_{fs}, \boldsymbol{M}_{fs}) + \mathcal{L}_{dice}(\tilde{\boldsymbol{M}}_{fs}, \boldsymbol{M}_{fs}). \quad (4)$$

当 t 较大时，$\boldsymbol{z}_t$ 接近随机噪声，因此解码器特征图不能提供预测阴影掩模的信息。因此，我们仅在时间步长 t 小时采用损失 $\mathcal{L}_{mask}$。我们将 t 的阈值设置为 σT，其中 T 是总步数。因此，仅当 t 小于阈值 σT 时才应用阴影强度调制。

在提供了预测的前景阴影掩模 $\tilde{\boldsymbol{M}}_{fs}$ 的情况下，我

们可以调制前景阴影区域内的噪声图。给定预测的噪声图 $\tilde{\boldsymbol{\epsilon}} = \epsilon_{\boldsymbol{\theta}}(\boldsymbol{z}_t, t, \boldsymbol{c}_{sg})$ ，我们将 $\tilde{\boldsymbol{\epsilon}}'$ 乘以通道尺度 $\boldsymbol{s}$，并加上通道偏差 $\boldsymbol{b}$ 以得到 $\tilde{\boldsymbol{\epsilon}}'$ 。然后，基于 $\tilde{\boldsymbol{M}}_{fs}$，我们将调制噪声图和原始噪声图联合收割机组合以得到最终噪声图：$\hat{\boldsymbol{\epsilon}} = \tilde{\boldsymbol{\epsilon}}' \circ \tilde{\boldsymbol{M}}_{fs} + \tilde{\boldsymbol{\epsilon}} \circ (1 - \tilde{\boldsymbol{M}}_{fs})$。

我们将等式 (3) 中的预测噪声图替换为（3）使用最终噪声图 $\hat{\boldsymbol{\epsilon}}$，并得到

$$\mathcal{L}_{mwsg} = \mathbb{E}_{t,\boldsymbol{\epsilon}\sim\mathcal{N}(0,1)}\left[\|\boldsymbol{W}_{fs} \circ (\boldsymbol{\epsilon} - \hat{\boldsymbol{\epsilon}})\|_2^2\right]. \quad (5)$$

我们总结了公式 1 中的掩模预测损失。 (4) 以及等式 (5) 中的加权噪声损失。（5）作为

$$\mathcal{L}_{all} = \mathcal{L}_{mask} + \lambda\mathcal{L}_{mwsg}, \quad (6)$$

其中 λ 是权衡参数。

5.3. 后处理

我们观察到生成的图像可能存在颜色偏移和背景变化问题。色移意味着整体色调偏离输入合成图像。背景变化意味着某些背景细节发生了变化。为了解决这些问题，我们创建了一个多任务后处理网络，该网络可以生成校正后的图像以及前景阴影掩模。然后，我们结合联合收割机输入的合成图像和校正图像的基础上预测的前景阴影掩模，以产生最终的图像。技术细节留待补充。

6. 实验

6.1. 数据集和评估指标

我们在 DESOBA[12] 和我们贡献的 DESOBAv 2 数据集上进行实验。我们将 DESOBAv 2 分为 21，088 个训练图像和 27，718 个元组，以及 487 个测试图像和 855 个元组。在 [12] 之后，测试集包含 BOS 图像（具有背景对象-阴影对）和无 BOS 图像。由于以下两个问题，我们的大部分实验都是基于 DESOBAv 2 数据集：1）DESOBAv 2 具有更大的测试集，支持更全面的评估。2)DESOBA 具有手动阴影去除和现有方法（*e.g.*，SGRNet）倾向于过拟合这样的伪像。

对于生成的结果，我们评估图像质量和掩模质量。对于图像评估，在 [12] 之后，我们采用 RMSE 和 SSIM，它们是基于地面真实目标图像和生成的图像计算的。全局 RMSE（GR）和全局 SSIM（GS）是在整个图像上计算的，而局部 RMSE（LR）和局部 SSIM（LS）是在地面实况前景阴影区域上计算的。对于掩模评估，在 [12] 之后，我们采用平衡错误率（BER），其基于地面真实二进制前景阴影掩模和通过阈值 0.5 获得的预测前景阴影掩模来计算。全局 BER（GB）是在整个图像上计算的，而局部 BER（LB）是在地面实况前景阴影区域上计算的。注意，扩散模型具有随机特性，并且阴影生成是一个多模态任务，即一个输入具有多个似然输出。类似于多模态修复评估 [54, 55]，我们为一个具有不同随机种子的测试图像生成 5 个结果，并选择最接近地面实况的结果（最高的本地 SSIM）来计算评估指标。

6.2. 实施细节

我们使用 PyTorch 1.12.1[30] 开发了我们的方法。我们的模型使用 Adam 优化器 [52] 在四个 NVIDIA RTX A6000 GPU 上以 50 个 epoch 的恒定学习率 $1e^{-5}$ 进行训练。我们的方法建立在 ControlNet 上 [52]。我们采用 ResNet18[17] 作为强度编码器。掩码预测器通过四个卷积层传递解码器特征映射和前景对象掩码的级联，ReLU 激活在前三层之后，Sigmoid 激活在最后一层之后。我们将超参数 w，σ 和 λ 分别设置为 10, 0.7, 和 1。

6.3. 与基线的比较

之后 [12]，我们与 ShadowGAN 进行比较 [53]、蒙版-ShadowGAN [13]、ARShadowGAN [22] 和 SGRNet [22][12]. 我们在 DESOBAv2 数据集上训练和测试所有方法。定量结果总结于表 1. 中。我们观察到，我们的 SGDiffusion 实现了最低的 GRMSE （LRMSE）和最高的 GSSIM （LSSIM），这表明我们的方法可以生成更接近真实阴影图像的阴影图像。最好的 GB 和 LB 结果表明，我们生成的阴影的形状和位置更准确。

为了进行定性比较，我们在图4中显示了几个示例结果。与基线方法相比，该模型产生的阴影具有更合理的形状和强度。此外，如行 1 所示，我们的方法可以考虑对象的自遮挡以生成不连续的阴影。如行 4 所示，我们的方法还可以考虑对象的材料，产生具有半透明效果的阴影。我们在补充资料中提供更多例子。

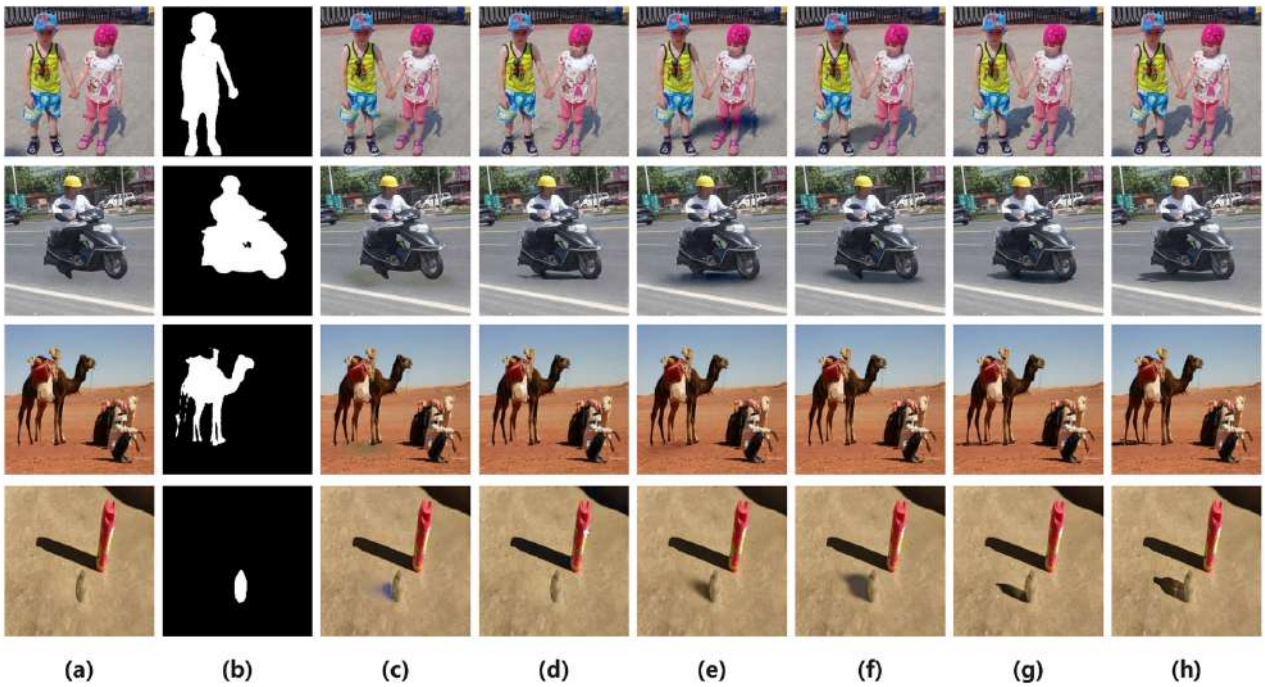

图 4. 不同方法在 DESOBAv 2 数据集上的视觉比较。从左到右是输入合成图像（a），前景对象掩模（B），ShadowGAN[53]（c），MaskshadowGAN[13]（d），ARShadowGAN[12]（e），SGRNet [22]（f），我们的 SGDiffusion（g），地面实况（h）的结果。

Method	BOS Test Images						BOS-free Test Images					
	GR ↓	LR ↓	GS ↑	LS ↑	GB↓	LB↓	GR ↓	LR ↓	GS ↑	LS ↑	GB↓	LB↓
ShadowGAN [53]	7.511	67.464	0.961	0.197	0.446	0.890	17.325	76.508	0.901	0.060	0.425	0.842
MaskshadowGAN [13]	8.997	79.418	0.951	0.180	0.500	1.000	19.338	94.327	0.906	0.044	0.500	1.000
ARShadowGAN [22]	7.335	58.037	0.961	0.241	0.383	0.761	16.067	63.713	0.908	0.104	0.349	0.682
SGRNet [12]	7.184	68.255	0.964	0.206	0.301	0.596	15.596	60.350	0.909	0.100	0.271	0.534
SGDiffusion	**6.098**	**53.611**	**0.971**	**0.370**	**0.245**	**0.487**	**15.110**	**55.874**	**0.913**	**0.117**	**0.233**	**0.452**

表 1. 不同方法在 DESOBAv2 数据集上的结果。最佳结果以粗体突出显示。

6.4. 消融研究

我们研究了加权噪声损失（WL）、强度调制（IM）和 SGDiffusion 后处理（PP）对 DESOBAv2 中 BOS 测试图像的影响。定量结果总结于表 2中。

在第 1 行中，我们报告了没有加权噪声损失的基本 ControlNet 的结果。对于 WL，行 3 和行 1 之间的比较强调了更加关注前景阴影区域的重要性。我们还报告了第 2 行中的特殊情况 †，其中在构建权重图时前景阴影遮罩未扩展。第 2 行中的结果与第 1 行中的结果相当，甚至更差，因为模型倾向于生成更大的阴影大小，而忽略形状和边缘细节。对于 IM，行 1 和行 5 之间的比较表明，强度调制可以通过调节阴影强度来显著改善阴影质量。我们还报告了第 4 行中的特殊情况 ∘，其中强度编码器输入不包含背景阴影掩模。第 4 行和第 5 行之间的比较表明，背景阴影掩模是有帮助的，因为背景阴影区域及其周围区域可以提供有用的线索来推断阴影强度。对于 PP，行 6 和行 7 之间的比较表明，后处理有效地校正了色移和背景变化，大大降低了全局 RMSE。我们还在补充资料中提供了消融版本的视觉结果。

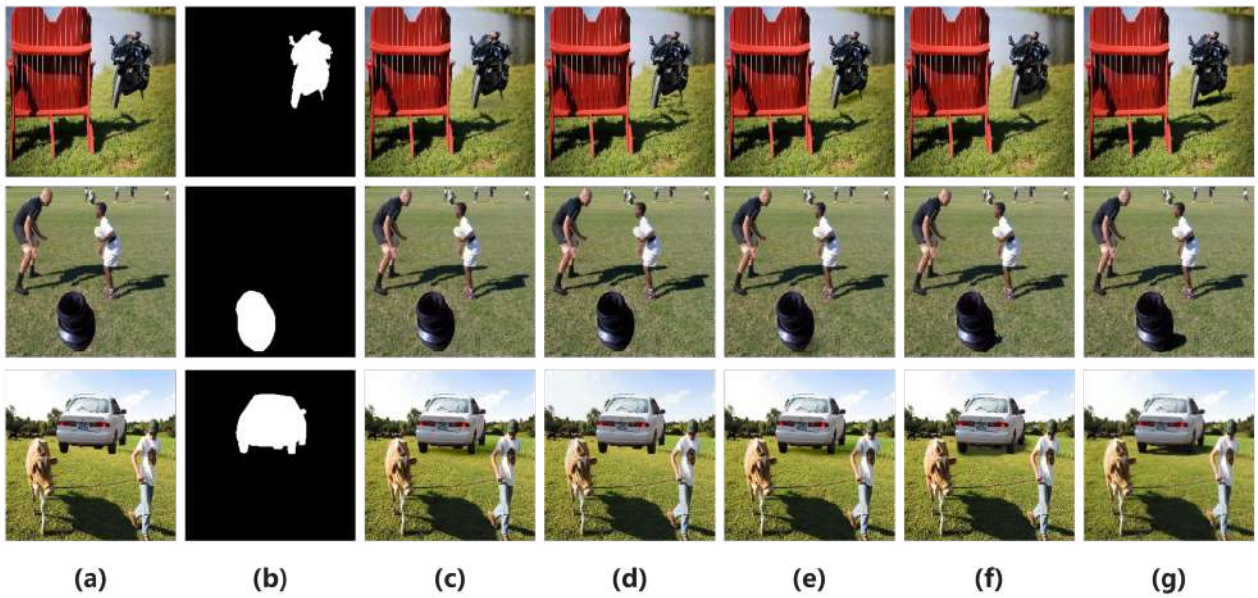

图 5. 不同方法在真实的合成图像上的视觉比较。从左到右是输入合成图像（a）、前景对象掩模（B）、ShadowGAN[53]（c）、MaskshadowGAN[13]（d）、ARShadowGAN[22]（e）、SGRNet[12]（f）、SGDiffusion（g）的结果。

Row	WL	IM	PP	GR ↓	LR ↓	GB↓	LB↓
1	-	-	+	8.285	59.753	0.271	0.534
2	†	-	+	8.319	59.491	0.282	0.563
3	+	-	+	7.041	53.829	0.249	0.492
4	-	○	+	7.410	56.121	0.269	0.536
5	-	+	+	7.357	54.159	0.262	0.526
6	+	+	-	13.447	55.231	0.245	0.487
7	+	+	+	6.098	53.611	0.245	0.487

表 2. 我们的方法在 DESOBAv 2 数据集的 BOS 测试图像上的消融研究。WL 是加权损失的缩写，† 表示没有扩展荫罩。IM 是强度调制的缩写，而 ○ 是指不使用背景荫罩。PP 是后处理的简称。

6.5. 真实的合成图像

我们在 [12] 提供的真实的合成图像上比较了不同的方法，其中背景图像和前景对象来自 DESOBA [12] 测试集。我们在 DESOBAv2 上训练所有方法，并在 DESOBA 上对其进行微调。不同方法的可视化结果如图5所示。这些结果证实了 SGDiffusion 能够熟练地合成具有精确轮廓、位置和方向的逼真阴影，这些阴影与背景对象-阴影对和前景对象信息兼容。相比之下，以前的方法往往产生模糊和错误的阴影。我们在补充资料中提供更多例子。

鉴于真实的合成图像缺乏地面实况图像，遵循 [12]，我们选择主观评估，在用户研究中使用 50 人类评分员。向每个参与者呈现由 5 方法生成的结果的图像对，并要求其选择具有更真实的前景阴影的图像。使用 Bradley-Terry 模型 [2]，我们在补充中报告了 B-T 评分，这再次证明了我们方法的优势。

7. 结论

在本文中，我们贡献了一个大规模的阴影生成数据集 DESOBAv2。我们还设计了一种新的基于扩散的阴影生成方法。大量的实验结果表明，我们的方法是能够产生合理的阴影复合前景，显着超过以前的方法。

8. 致谢

本工作得到了国家自然科学基金（批准号：62076162）、上海市科技重大专项（批准号：2021SHZDZX0102）的资助。

References

[1] Ibrahim Arief, Simon McCallum, and Jon Yngve Hardeberg. Realtime estimation of illumination direction for augmented reality on mobile devices. In *CIC*, 2012. 2

[2] Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952. 8

[3] Wenyan Cong, Jianfu Zhang, Li Niu, Liu Liu, Zhixin Ling, Weiyuan Li, and Liqing Zhang. Dovenet: Deep image harmonization via domain verification. In *CVPR*, 2020. 1, 2

[4] Wenyan Cong, Li Niu, Jianfu Zhang, Jing Liang, and Liqing Zhang. Bargainnet: Background-guided domain translation for image harmonization. In *ICME*, 2021.

[5] Wenyan Cong, Xinhao Tao, Li Niu, Jing Liang, Xuesong Gao, Qihao Sun, and Liqing Zhang. High-resolution image harmonization via collaborative dual transformations. In *CVPR*, 2022.

[6] Xiaodong Cun and Chi-Man Pun. Improving the harmony of the composite image by spatial-separated attention module. *TIP*, 2020. 2

[7] Marc-André Gardner, Yannick Hold-Geoffroy, Kalyan Sunkavalli, Christian Gagné, and Jean-Francois Lalonde. Deep parametric indoor lighting estimation. In *ICCV*, 2019. 2

[8] Lanqing Guo, Siyu Huang, Ding Liu, Hao Cheng, and Bihan Wen. Shadowformer: Global context helps image shadow removal. In *AAAI*, 2023. 3

[9] Lanqing Guo, Chong Wang, Wenhan Yang, Siyu Huang, Yufei Wang, Hanspeter Pfister, and Bihan Wen. Shadowdiffusion: When degradation prior meets diffusion model for shadow removal. In *CVPR*, 2023. 3

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 3

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

[12] Yan Hong, Li Niu, and Jianfu Zhang. Shadow generation for composite image in real-world scenes. *AAAI*, 2022. 1, 2, 6, 7, 8

[13] Xiaowei Hu, Yitong Jiang, Chi-Wing Fu, and Pheng-Ann Heng. Mask-shadowgan: Learning to remove shadows from unpaired data. In *ICCV*, 2019. 6, 7, 8

[14] Shaozong Huang and Lan Hong. Diffusion model for mural image inpainting. In *ITOEC*, 2023. 3

[15] Kevin Karsch, Kalyan Sunkavalli, Sunil Hadap, Nathan Carr, Hailin Jin, Rafael Fonte, Michael Sittig, and David Forsyth. Automatic scene inference for 3d object compositing. *ACM TOG*, 2014. 2

[16] Eric Kee, James F. O'Brien, and Hany Samir Farid. Exposing photo manipulation from shading and shadows. *ACM TOG*, 2014. 2

[17] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014. 6

[18] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *CoRR*, abs/1312.6114, 2013. 4

[19] Bin Liao, Yao Zhu, Chao Liang, Fei Luo, and Chunxia Xiao. Illumination animating and editing in a single picture using scene structure estimation. *Computers & Graphics*, 82:53–64, 2019. 2

[20] Chen-Hsuan Lin, Ersin Yumer, Oliver Wang, Eli Shechtman, and Simon Lucey. St-gan: Spatial transformer generative adversarial networks for image compositing. In *CVPR*, 2018. 2

[21] Bin Liu, Kun Xu, and Ralph R Martin. Static scene illumination estimation from videos with applications. *JCST*, 32(3):430–442, 2017. 2

[22] Daquan Liu, Chengjiang Long, Hongpan Zhang, Hanning Yu, Xinzhi Dong, and Chunxia Xiao. Arshadowgan: Shadow generative adversarial network for augmented reality in single light scenes. In *CVPR*, 2020. 1, 2, 6, 7, 8

[23] Guilin Liu, Fitsum A. Reda, Kevin J. Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. Image inpainting for irregular holes using partial convolutions. In *ECCV*, 2018. 3

[24] Jian Ma, Mingjun Zhao, Chen Chen, Ruichen Wang, Di Niu, Haonan Lu, and Xiaodong Lin. Glyphdraw: Learning to draw chinese characters in image synthesis models coherently. *CoRR*, abs/2303.17870, 2023. 5

[25] Quanling Meng, Shengping Zhang, Zonglin Li, Chenyang Wang, Weigang Zhang, and Qingming Huang. Automatic shadow generation via exposure fusion. *IEEE Transactions on Multimedia*, 2023. 2

[26] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *3DV*, 2016. 4

[27] Chong Mou, Xintao Wang, Liangbin Xie, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *CoRR*, abs/2302.08453, 2023. 2

[28] Li Niu, Wenyan Cong, Liu Liu, Yan Hong, Bo Zhang, Jing Liang, and Liqing Zhang. Making images real again: A comprehensive survey on deep image composition. *CoRR*, abs/2106.14490, 2021. 1

[29] Sibam Parida, Vignesh Srinivas, Bhavishya Jain, Rajesh Naik, and Neeraj Rao. Survey on diverse image inpainting using diffusion models. In *PCEMS*, 2023. 3

[30] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *NIPS*, 32, 2019. 6

[31] Patrick Pérez, Michel Gangnet, and Andrew Blake. Poisson image editing. In *SIGGRAPH*. 2003. 2

[32] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 1, 2, 3, 4

[33] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015. 4

[34] Yichen Sheng, Jianming Zhang, and Bedrich Benes. Ssn: Soft shadow network for image compositing. In *CVPR*, 2021. 1

[35] Yichen Sheng, Yifan Liu, Jianming Zhang, Wei Yin, A Cengiz Oztireli, He Zhang, Zhe Lin, Eli Shechtman, and Bedrich Benes. Controllable shadow generation using pixel height maps. In *ECCV*, 2022. 1, 2

[36] Yichen Sheng, Jianming Zhang, Julien Philip, Yannick Hold-Geoffroy, Xin Sun, He Zhang, Lu Ling, and Bedrich Benes. Pixht-lab: Pixel height based light effect generation for image compositing. In *CVPR*, 2023. 1, 2

[37] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *CoRR*, abs/2010.02502, 2020. 2

[38] Yi-Zhe Song, Zhifei Zhang, Zhe L. Lin, Scott D. Cohen, Brian L. Price, Jianming Zhang, Soo Ye Kim, and Daniel G. Aliaga. Objectstitch: Generative object compositing. In *CVPR*, 2023. 2

[39] Xinhao Tao, Junyan Cao, Yan Hong, and Li Niu. Shadow generation with decomposed mask prediction and attentive shadow filling. In *AAAI*, 2024. 2

[40] Yi-Hsuan Tsai, Xiaohui Shen, Zhe Lin, Kalyan Sunkavalli, Xin Lu, and Ming-Hsuan Yang. Deep image harmonization. In *CVPR*, 2017. 2

[41] Tianyu Wang, Xiaowei Hu, Pheng-Ann Heng, and Chi-Wing Fu. Instance shadow detection with a single-stage detector. *TPAMI*, 2022. 2, 3

[42] Huikai Wu, Shuai Zheng, Junge Zhang, and Kaiqi Huang. Gp-gan: Towards realistic high-resolution image blending. In *ACM MM*, 2019. 2

[43] Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *CVPR*, 2023. 5

[44] Binxin Yang, Shuyang Gu, Bo Zhang, Ting Zhang, Xuejin Chen, Xiaoyan Sun, Dong Chen, and Fang Wen. Paint by example: Exemplar-based image editing with diffusion models. In *CVPR*, 2023. 1, 2

[45] Shiyuan Yang, Xiaodong Chen, and Jing Liao. Uni-paint: A unified framework for multimodal image inpainting with pretrained diffusion model. In *ACM MM*, 2023. 3

[46] Fangneng Zhan, Jiaxing Huang, and Shijian Lu. Adaptive composition gan towards realistic image synthesis. *CoRR*, abs/1905.04693, 2019. 2

[47] Fangneng Zhan, Shijian Lu, Changgong Zhang, Feiying Ma, and Xuansong Xie. Towards realistic 3d embedding via view alignment. *CoRR*, abs/2007.07066, 2020. 2

[48] Bo Zhang, Yuxuan Duan, Jun Lan, Yan Hong, Huijia Zhu, Weiqiang Wang, and Li Niu. Controlcom: Controllable image composition using diffusion model. *arXiv preprint arXiv:2308.10040*, 2023. 1, 2

[49] He Zhang, Jianming Zhang, Federico Perazzi, Zhe Lin, and Vishal M Patel. Deep image compositing. In *WACV*, 2021. 2

[50] Jinsong Zhang, Kalyan Sunkavalli, Yannick Hold-Geoffroy, Sunil Hadap, Jonathan Eisenman, and Jean-François Lalonde. All-weather deep outdoor lighting estimation. In *CVPR*, 2019. 2

[51] Lingzhi Zhang, Tarmily Wen, and Jianbo Shi. Deep image blending. In *WACV*, 2020. 2

[52] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023. 2, 3, 4, 5, 6

[53] Shuyang Zhang, Runze Liang, and Miao Wang. Shadowgan: Shadow synthesis for virtual objects with conditional adversarial networks. *Computational Visual Media*, 5:105–115, 2019. 1, 2, 6, 7, 8

[54] Lei Zhao, Qihang Mo, Sihuan Lin, Zhizhong Wang, Zhiwen Zuo, Haibo Chen, Wei Xing, and Dongming Lu. Uctgan: Diverse image inpainting based on unsupervised cross-space translation. In *CVPR*, 2020. 6

[55] Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai. Pluralistic image completion. In *CVPR*, 2019. 6

[56] Haitian Zheng, Zhe Lin, Jingwan Lu, Scott Cohen, Eli Shechtman, Connelly Barnes, Jianming Zhang, Ning Xu, Sohrab Amirghodsi, and Jiebo Luo. Image inpainting with cascaded modulation gan and object-aware training. In *ECCV*, 2022. 3

Shadow Generation for Composite Image Using Diffusion Model

Qingyang Liu[1], Junqi You[1], Jianting Wang[1], Xinhao Tao[1], Bo Zhang[1], Li Niu[1,2*]
[1] Shanghai Jiao Tong University [2] miguo.ai
[1]{narumimaria,yjqsjtu2022,glory1299,taoxinhao,bo-zhang,ustcnewly}@sjtu.edu.cn

Abstract

In the realm of image composition, generating realistic shadow for the inserted foreground remains a formidable challenge. Previous works have developed image-to-image translation models which are trained on paired training data. However, they are struggling to generate shadows with accurate shapes and intensities, hindered by data scarcity and inherent task complexity. In this paper, we resort to foundation model with rich prior knowledge of natural shadow images. Specifically, we first adapt ControlNet to our task and then propose intensity modulation modules to improve the shadow intensity. Moreover, we extend the small-scale DESOBA dataset to DESOBAv2 using a novel data acquisition pipeline. Experimental results on both DESOBA and DESOBAv2 datasets as well as real composite images demonstrate the superior capability of our model for shadow generation task. The dataset, code, and model are released at https://github.com/bcmi/Object-Shadow-Generation-Dataset-DESOBAv2.

1. Introduction

Image composition [28] aims to merge the foreground of one image with another background image to produce a composite image, which has a wide range of applications like virtual reality, artistic creation, and E-commerce. Simply pasting the foreground onto the background often results in visual inconsistencies, including the incompatible illumination between foreground and background [3], lack of foreground shadow/reflection [12, 34], and so on. In this paper, we focus on the shadow issue, *i.e.*, the inserted foreground does not have plausible shadow on the background, which could significantly degrade the realism and quality of composite image.

As illustrated in Figure 1, shadow generation is a challenging task because the foreground shadow is determined by many complicated factors like the lighting information and the geometry of foreground/background. The existing shadow generation methods can be divided into rendering based methods [34–36] and non-rendering based methods [12, 22, 53]. Rendering based methods usually impose restrict assumptions on the geometry and lighting, which could hardly be satisfied in real-world scenarios. Besides, [35, 36] require users to specify the lighting information, which hinders its direct application in our task. Non-rendering based methods usually train an image-to-image translation network, based on pairs of composite images without foreground shadows and real images with foreground shadows. However, due to the training data scarcity and task difficulty, these methods are struggling to generate shadows with reasonable shapes and intensities.

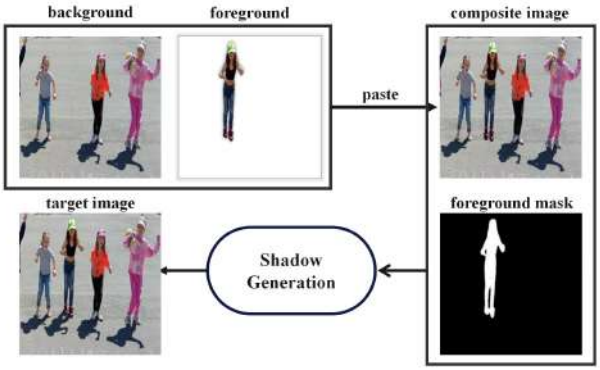

Figure 1. A composite image can be obtained by pasting the foreground on the background. Shadow generation aims to generate plausible shadow for the inserted foreground in the composite image to produce a more realistic image.

Recently, foundation model (*e.g.*, stable diffusion [32]) pretrained on large-scale dataset has demonstrated unprecedented potential for image generation and editing. In previous works [44, 48] on object-guided inpainting or composition, they show that the generated foregrounds are accompanied by shadows even without considering the shadow issue, probably because of the rich prior knowledge of natural shadow images in foundation model. However, they could only generate satisfactory shadows in simple cases and the object appearance could be altered unexpectedly.

*Corresponding author.

We build our method upon conditional foundation model [52] and propose several key innovations. First, we modify the control encoder input and the noise loss to fit our task. Then, we observe that the generated shadow intensity (the level of darkness) is unsatisfactory. Especially when the background objects has shadows, the intensity inconsistency between foreground shadow and background shadows make the whole image unrealistic. Therefore, we introduce another intensity encoder to modulate the foreground shadow intensity. Specifically, the denoising U-Net is modified to output both noise map and foreground shadow mask. The intensity encoder takes in the composite image and background shadow mask, producing the scale/bias to modulate the predicted noise within the foreground shadow region. Finally, we devise a post-processing network to rectify the color shift and background variation.

The model training requires abundant pairs of composite images without foreground shadows and real images with foreground shadows. The existing real-world shadow generation dataset DESOBA [12] is limited by scale (*i.e.*, 1,012 real images and 3,623 pairs) due to the high cost of manual shadow removal, which is insufficient to train our model. To ensure sufficient supervision, we design a novel data construction pipeline, which extends DESOBA to DESOBAv2 (*i.e.*, 21,575 real images and 28,573 pairs) using object-shadow detection and inpainting techniques. Specifically, we first collect a large number of real-world images with one or more object-shadow pairs. Then, we use pretrained object-shadow detection model [41] to predict object and shadow masks for object-shadow pairs. Next, we apply pretrained inpainting model [32] to inpaint the detected shadow regions to get deshadowed images. Finally, based on real images and deshadowed images, we construct pairs of synthetic composite images and ground-truth target images.

We conduct experiments on both DESOBAv2 and DESOBA datasets. The results reveal remarkable improvement in shadow generation task, after leveraging the benefits of large-scale data and foundation model. Our main contributions can be summarized as follows: 1) We contribute DESOBAv2, a large-scale real-world shadow generation dataset, which could greatly facilitate the shadow generation task. 2) We propose a cutting-edge diffusion model specifically designed to produce shadows for the composite foregrounds. 3) Through comprehensive experiments, we validate the efficacy of our dataset construction pipeline and the superiority of our proposed model.

2. Related Work

2.1. Image Composition

Image composition aims to overlay a foreground object on a background image to yield a composite result [20, 22, 42, 46, 47]. Previous research works have tackled different issues that can compromise the quality of composite images. For instance, image blending methods [31, 42, 49, 51] target at combining the foreground and background seamlessly. Image harmonization methods [3–6, 40] aim to rectify the illumination disparity between foreground and background. Nonetheless, the above methods largely overlook the shadow cast by the foreground onto the background. Recently, generative image composition methods [38, 44, 48] can insert a foreground object into a bounding box in the background and the inserted object is likely to have shadow effect. However, they could only generate satisfactory shadows in simple cases and the object appearance could be altered unexpectedly.

2.2. Shadow Generation

In this paper, the goal of shadow generation task is generating plausible shadow for the composite foreground. Existing methods can be broadly categorized into rendering based methods and non-rendering based methods. The rendering based methods necessitate a comprehensive understanding of factors like illumination, reflectance, material properties, and scene geometry to produce shadows for the inserted objects. However, such detailed knowledge relies on user input [15, 16, 21, 35, 36] or model prediction [1, 7, 19, 50], which is either labor-intensive or unreliable [53]. For example, [35, 36] could produce compelling results with user control. However, in the composite image, the lighting information should be inferred automatically from background instead of requested by users.

Non-rendering based methods [12, 22, 25, 53] aim to translate an input composite image without foreground shadow to an output with foreground shadow, bypassing the need for explicit knowledge of the aforementioned factors. For instance, ShadowGAN [53] utilizes both global and local conditional discriminator to enhance the realism of generated shadows. ARShadowGAN [22] emphasizes the importance of background shadow and uses it to guide foreground shadow generation. SGRNet [12] encourages the information exchange between foreground and background, and employs a classic illumination model for better shadow effect. The work [25] produces multiple underexposure images and fuses them to get the final shadow region. DMASNet [39] decomposes shadow mask prediction into box prediction and shape prediction, achieving better cross-domain transferability.

To the best of our knowledge, we are the first diffusion-based method focusing on shadow generation.

2.3. Diffusion Models

In recent years, diffusion models have emerged as a powerful tool in image generation and image editing. These models approach image generation as a series of stochastic transitions, moving from a basic distribution to the desired

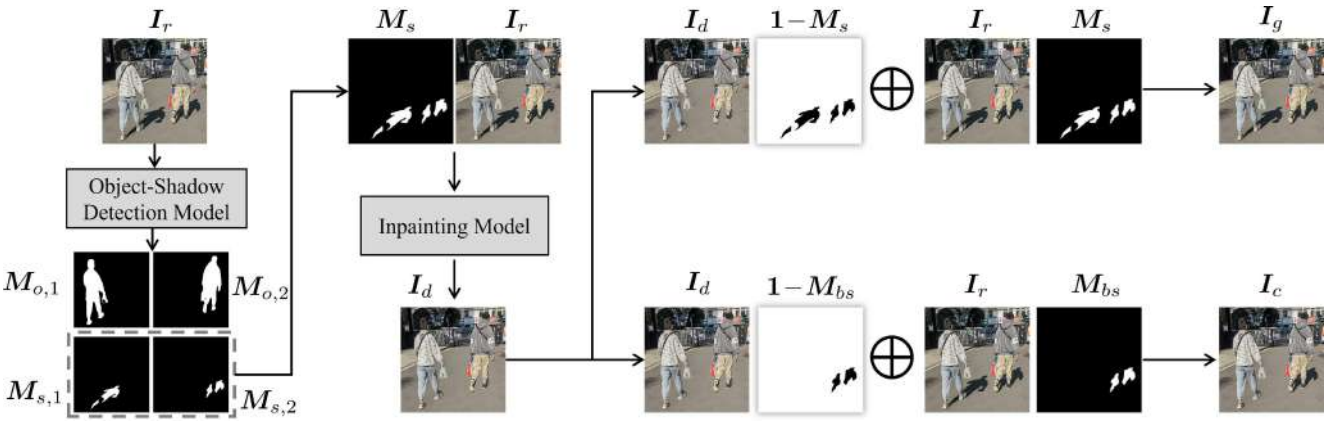

Figure 2. The pipeline of dataset construction. We use object-shadow detection model [41] to predict pairs of object and shadow masks in the real image $\boldsymbol{I}_r$. Then we obtain the union $\boldsymbol{M}_s$ of all shadow masks as the inpainting mask and apply inpainting model [32] to get a deshadowed image $\boldsymbol{I}_d$. After designating a foreground object, we replace the background shadow regions $\boldsymbol{M}_{bs}$ in $\boldsymbol{I}_d$ with the counterparts in $\boldsymbol{I}_r$ to synthesize a composite image $\boldsymbol{I}_c$, and replace all the shadow regions $\boldsymbol{M}_s$ in $\boldsymbol{I}_d$ with the counterparts in $\boldsymbol{I}_r$ to obtain the ground-truth target image $\boldsymbol{I}_g$.

data distribution [11]. Diffusion models can be divided into unconditional diffusion models [11, 37] and conditional diffusion models [27, 32, 52]. Unconditional diffusion models focus on generating realistic images by capturing the distribution of natural images, without the need of any specific input conditions. Conditional diffusion models are designed to produce images under the guidance of specific conditional inputs, such as text descriptions, semantic masks, and so on. ControlNet [52] is a popular conditional diffusion model, which equips large pretrained text-to-image diffusion models with spatial-aware and task-specific conditions. We build our model upon ControlNet and propose several innovations to meet the specific requirements of shadow generation.

3. Dataset Construction

The pipeline of our dataset construction is illustrated in Figure 2, which will be detailed next.

3.1. Shadow Image Collection

We harvest an extensive collection of real-world outdoor images with natural lighting across various scenes from two sources. On one hand, we crawl online images from public websites that have licenses for reuse. On the other hand, we hire photographers to capture photos in the outdoor scenes that satisfy our requirements. We only preserve the images with at least one object-shadow pair, arriving at 44,044 images.

3.2. Shadow Removal

Given a real image $\boldsymbol{I}_r$ with object-shadow pairs, we use the pretrained object-shadow detection model [41] to predict K pairs of object and shadow masks. We use $\boldsymbol{M}_{o,k}$ (*resp.*, $\boldsymbol{M}_{s,k}$) to denote the object (*resp.*, shadow) mask of the k-th object. We refer to one detected object-shadow pair as one detected instance. We eliminate the images without any detected instance.

Subsequently, we attempt to erase all the detected shadows. We have tried some state-of-the-art shadow removal models [8, 9], but the performance in the wild is below our expectation due to poor generalization ability. Considering the recent rapid advance of image inpainting [14, 23, 29, 32, 45, 56] techniques, we resort to image inpainting to remove the shadows. Although image inpainting cannot preserve the background information precisely, we observe that the background textures in the shadow region are usually very simple, and the inpainted result has similar textures with the original background. Thus, we roughly treat the inpainted results as deshadowed results.

We obtain the union of all detected shadow masks $\boldsymbol{M}_s = \boldsymbol{M}_{s,1} \cup \boldsymbol{M}_{s,2} \cup \cdots \cup \boldsymbol{M}_{s,K}$ as the inpainting mask and apply the pretrained inpainting model [32] to get a deshadowed image $\boldsymbol{I}_d$. In practice, we observe that the inpainting model is prone to generate low-quality shadow in the inpainted region in some cases. To prevent the inpainting model from generating undesirable shadows in the inpainted region, we adopt some tricks like dilating the inpainting mask and flipping images vertically, which can effectively obstruct undesirable shadow generation during inpainting. However, there may still exist undesirable shadows or noticeable artifacts in the inpainted region.

After inpainting, we manually filter the object-shadow pairs according to the following rules: 1) We remove the object-shadow pairs with low-quality object masks or shadow masks. 2) We remove those object-shadow pairs

with generated shadows or noticeable artifacts in the inpainted region. After manual filtering, we refer to the remaining object-shadow pairs as valid instances. We have 21,575 images with 28,573 valid instances.

3.3. Composite Image Synthesis

Given a pair of a real image $\boldsymbol{I}_r$ and a deshadowed image $\boldsymbol{I}_d$, we randomly select the k-th foreground object from valid instances and synthesize the composite image. $\boldsymbol{M}_{o,k}$ (*resp.*, $\boldsymbol{M}_{s,k}$) is referred to as the foreground object (*resp.*, shadow) mask $\boldsymbol{M}_{fo}$ (*resp.*, $\boldsymbol{M}_{fs}$). One strategy is replacing the shadow region $\boldsymbol{M}_{fs}$ of this foreground object in $\boldsymbol{I}_r$ with the counterpart in $\boldsymbol{I}_d$ to erase the foreground shadow. However, this strategy may leave traces along the shadow boundary, in which case the model may find a shortcut to generate the shadow. Another strategy is replacing the shadow regions $\boldsymbol{M}_{bs} = \boldsymbol{M}_{s,1} \cup \cdots \cup \boldsymbol{M}_{s,k-1} \cup \boldsymbol{M}_{s,k+1} \cup \cdots \cup \boldsymbol{M}_{s,K}$ of the other objects in $\boldsymbol{I}_d$ with the counterparts in $\boldsymbol{I}_r$ to synthesize a composite image $\boldsymbol{I}_c$, in which only the selected foreground object does not have shadow while all the other objects have shadows. We adopt the second strategy.

After inpainting, the background may undergo slight changes, so the background of $\boldsymbol{I}_c$ may be slightly different from that of $\boldsymbol{I}_r$. To ensure consistent background, we obtain the ground-truth target image $\boldsymbol{I}_g$ by replacing the shadow regions $\boldsymbol{M}_s$ of all objects in $\boldsymbol{I}_d$ with the counterparts in $\boldsymbol{I}_r$. Then, $\boldsymbol{I}_c$ and $\boldsymbol{I}_g$ form a pair of input composite image and ground-truth target image. So far, we obtain tuples in the form of $\{\boldsymbol{I}_c, \boldsymbol{M}_{fo}, \boldsymbol{M}_{fs}, \boldsymbol{M}_{bs}, \boldsymbol{I}_g\}$, which will be used for model training. Example images and more statistics of our dataset can be found in the supplementary.

4. Background

Stable Diffusion [32] is latent diffusion model operating in a latent space. First, 512×512 images are converted to 64×64 latent images using VAE [18] with encoder E_r and decoder D_r. The image space is projected to the latent space using E_r, and back to the image space using D_r. Then, the forward diffusion process and backward denoising process are performed in the latent space. The denoising U-Net [33] consists of an encoder with 12 blocks, a middle block, and a skip-connected decoder with 12 blocks.

During training, random Gaussian noise $\boldsymbol{\epsilon}$ is added to the latent image $\boldsymbol{z}_0$ in the denoising step t, producing a noisy latent image $\boldsymbol{z}_t$. Given time step t and text prompt $\boldsymbol{c}_{txt}$, the denoising U-Net with model parameters $\epsilon_{\boldsymbol{\theta}}$ is trained to predict the added noise $\boldsymbol{\epsilon}$.

To support spatial conditional information (*e.g.*, edge, pose, depth), ControlNet [52] integrates a control encoder E_c with pre-trained Stable Diffusion. Specifically, the control encoder contains trainable replicas of its 12 encoding blocks and middle block across four resolutions $(64 \times 64, 32 \times 32, 16 \times 16, 8 \times 8)$. It takes a 512×512 conditional image as input.

The conditional feature maps $\boldsymbol{c}_{img}$ output from control encoder are used to enhance the 12 skip-connections and middle block in denoising U-Net via zero convolution layers. While the original Stable Diffusion is fixed to retain prior knowledge, control encoder could incorporate additional conditions to guide image generation. The objective could be rewritten as

$$\mathcal{L}_{ctrl} = \mathbb{E}_{t,\boldsymbol{\epsilon}\sim\mathcal{N}(0,1)}\Big[\|\boldsymbol{\epsilon} - \epsilon_{\boldsymbol{\theta}}(\boldsymbol{z}_t, t, \boldsymbol{c}_{txt}, \boldsymbol{c}_{img})\|_2^2\Big]. \quad (1)$$

5. Method

Given a composite image $\boldsymbol{I}_c$ without foreground shadow as well as the foreground object mask $\boldsymbol{M}_{fo}$, our Shadow Generation Diffusion (SGDiffusion) model aims to produce $\tilde{\boldsymbol{I}}_g$ with plausible foreground shadow. We will adapt ControlNet [52] to shadow generation task in Section 5.1, and propose novel modules to improve the shadow intensity in Section 5.2. Finally, we will briefly introduce post-processing techniques to enhance the image quality in Section 5.3.

5.1. Adapting ControlNet to Shadow Generation

For shadow generation task, the useful conditional information is input composite image $\boldsymbol{I}_c$ and foreground object mask $\boldsymbol{M}_{fo}$, in which the foreground object mask indicates the target object we need to generate shadow for. We concatenate $\boldsymbol{I}_c$ with $\boldsymbol{M}_{fo}$ as the input of control encoder E_c. The control encoder outputs the conditional feature maps $\boldsymbol{c}_{sg}$, which are injected into the denoising decoder to provide guidance. For the text prompt, we have tried several variants like "the [object category] with shadow", but they have no significant impact on the generated shadows. Therefore, we use null text prompt by default.

Given a set of conditions including time step t and conditional feature maps $\boldsymbol{c}_{sg}$, the denoising U-Net with model parameters $\epsilon_{\boldsymbol{\theta}}$ predicts the noise $\boldsymbol{\epsilon}$ added to the noisy latent image $\boldsymbol{z}_t$:

$$\mathcal{L}_{sg} = \mathbb{E}_{t,\boldsymbol{\epsilon}\sim\mathcal{N}(0,1)}\Big[\|\boldsymbol{\epsilon} - \epsilon_{\boldsymbol{\theta}}(\boldsymbol{z}_t, t, \boldsymbol{c}_{sg}))\|_2^2\Big]. \quad (2)$$

To enforce the model to place more emphasis on the foreground shadow region, we introduce weighted noise loss, which assigns higher weights to the foreground shadow region. We expand the foreground shadow mask by a dilated kernel to get the expanded mask $\hat{\boldsymbol{M}}_{fs}$. The weights in the expanded foreground shadow region are w while the other weights are 1, leading to the weight map $\boldsymbol{W}_{fs}$. If we do not expand the foreground shadow region, the model will be misled to generate large shadows, overlooking the details of shadow shapes and boundaries. By applying weight map $\boldsymbol{W}_{fs}$ to the noise loss, we can arrive at

$$\mathcal{L}_{wsg} = \mathbb{E}_{t,\boldsymbol{\epsilon}\sim\mathcal{N}(0,1)}\Big[\|\boldsymbol{W}_{fs} \circ (\boldsymbol{\epsilon} - \epsilon_{\boldsymbol{\theta}}(\boldsymbol{z}_t, t, \boldsymbol{c}_{sg}))\|_2^2\Big], \quad (3)$$

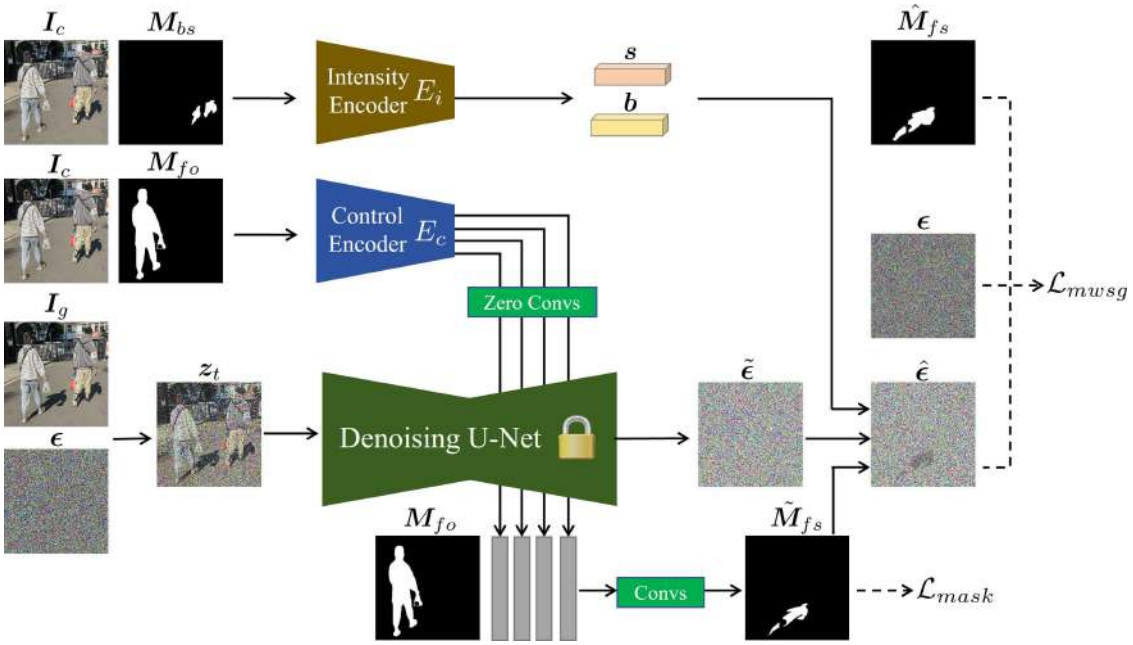

Figure 3. The framework of our SGDiffusion. We adapt ControlNet (Control Encoder and Denoising U-Net) to shadow generation task. We also introduce an intensity encoder to modulate the foreground shadow region in the noise map $\tilde{\boldsymbol{\epsilon}}$, leading to $\hat{\boldsymbol{\epsilon}}$. The output noise $\hat{\boldsymbol{\epsilon}}$ is supervised by weighted noise loss $\mathcal{L}_{mwsg}$ based on the expanded foreground shadow mask $\hat{\boldsymbol{M}}_{fs}$

where $\circ$ denotes element-wise multiplication.

During inference, to retain more information of input composite image $\boldsymbol{I}_c$ in the initial noise, we obtain $\boldsymbol{z}_T$ by adding noise to the latent image of $\boldsymbol{I}_c$, rather than directly sampling from the Gaussian distribution $\mathcal{N}(0,1)$.

5.2. Shadow Intensity Modulation

By using the adapted ControlNet in Section 5.1, we observe that the intensity of generated foreground shadow is unsatisfactory. Especially when the background has object-shadow pairs, the generated foreground shadow is often notably darker or brighter than background shadows. Such inconsistency between foreground shadow intensity and background shadow intensity makes the whole image unrealistic.

Therefore, we introduce another intensity encoder to modulate the foreground shadow intensity. Specifically, we use encoder E_i to extract intensity-relevant information. Intuitively, by observing background shadows and its surrounding unshadowed areas, we can estimate the intensity of foreground shadows. Thus, the input of intensity encoder E_i should include the composite image $\boldsymbol{I}_c$ and background shadow mask $\boldsymbol{M}_{bs}$. When there is no background shadow, the mask is all black. We concatenate $\boldsymbol{I}_c$ with background shadow mask $\boldsymbol{M}_{bs}$ as the input of intensity encoder.

The intensity encoder outputs scales and biases to adjust the intensity of noise map within the foreground shadow region. The modulated noise map results in the modulated latent image, and further results in the modulated foreground shadow. Therefore, the intensity adjustment of noise map is finally embodied in the intensity variation of generated foreground shadow. Specifically, when the noise map has c channels, E_i outputs the c-dim scale vector $\boldsymbol{s}$ and c-dim bias vector $\boldsymbol{b}$, containing channel-wise scales and biases. $\boldsymbol{s}$ and $\boldsymbol{b}$ are used to modulate the predicted noise map within the foreground shadow region.

One problem is that the foreground shadow region is unknown in the testing stage, so we need to predict the foreground shadow mask. To avoid much extra computational cost, we take advantage of the feature maps in the denoising U-Net to predict the foreground shadow mask. Previous works usually combine different layers of feature maps in denoising U-Net for mask prediction [24, 43]. We try different layers of feature maps and find that decoder feature maps are more effective in shadow mask prediction. We also use foreground object mask, which could provide useful hints for the location of foreground shadow. We resize all decoder feature maps and foreground object mask to the same size, and concatenate them channel-wisely. The concatenation passes through several convolutional layers to predict the foreground shadow mask $\tilde{\boldsymbol{M}}_{fs}$. $\tilde{\boldsymbol{M}}_{fs}$ is supervised with ground-truth foreground shadow mask $\boldsymbol{M}_{fs}$ by Binary Cross-Entropy (BCE) loss and Dice loss [26]:

$$\mathcal{L}_{mask} = \mathcal{L}_{bce}(\tilde{\boldsymbol{M}}_{fs}, \boldsymbol{M}_{fs}) + \mathcal{L}_{dice}(\tilde{\boldsymbol{M}}_{fs}, \boldsymbol{M}_{fs}). \quad (4)$$

When t is large, $\boldsymbol{z}_t$ is close to random noise and thus the decoder feature maps are not informative to predict shadow

mask. Hence, we only employ the loss $\mathcal{L}_{mask}$ when the time step t is small. We set the threshold of t as σT, in which T is the total number of steps. Accordingly, shadow intensity modulation is only applied when t is smaller than the threshold σT.

Provided with the predicted foreground shadow mask $\tilde{\boldsymbol{M}}_{fs}$, we can modulate the noise map within the foreground shadow region. Given the predicted noise map $\tilde{\boldsymbol{\epsilon}} = \epsilon_{\boldsymbol{\theta}}(\boldsymbol{z}_t, t, \boldsymbol{c}_{sg})$, we multiply $\tilde{\boldsymbol{\epsilon}}$ by channel-wise scales $\boldsymbol{s}$ and add channel-wise biases $\boldsymbol{b}$ to get $\tilde{\boldsymbol{\epsilon}}'$. Then, based on $\tilde{\boldsymbol{M}}_{fs}$, we combine the modulated noise map and original noise map to get the final noise map: $\hat{\boldsymbol{\epsilon}} = \tilde{\boldsymbol{\epsilon}}' \circ \tilde{\boldsymbol{M}}_{fs} + \tilde{\boldsymbol{\epsilon}} \circ (1 - \tilde{\boldsymbol{M}}_{fs})$.

We replace the predicted noise map in Eqn. (3) with the final noise map $\hat{\boldsymbol{\epsilon}}$ and get

$$\mathcal{L}_{mwsg} = \mathbb{E}_{t, \boldsymbol{\epsilon} \sim \mathcal{N}(0,1)} \left[\| \boldsymbol{W}_{fs} \circ (\boldsymbol{\epsilon} - \hat{\boldsymbol{\epsilon}}) \|_2^2 \right]. \tag{5}$$

We summarize the mask prediction loss in Eqn. (4) and weighted noise loss in Eqn. (5) as

$$\mathcal{L}_{all} = \mathcal{L}_{mask} + \lambda \mathcal{L}_{mwsg}, \tag{6}$$

where λ is a trade-off parameter.

5.3. Post-processing

We observe that the generated images could have color shift and background variation issues. Color shift means that the overall color tone deviates from the input composite image. Background variation means that some background details are changed. To solve these issues, we create a multi-task post-processing network which yields the rectified image together with the foreground shadow mask. Then, we combine input composite image and rectified image based on the predicted foreground shadow mask to produce the final image. The technical details are left to supplementary.

6. Experiments

6.1. Datasets and Evaluation Metrics

We conduct experiments on both DESOBA [12] and our contributed DESOBAv2 dataset. We split DESOBAv2 into 21,088 training images with 27,718 tuples and 487 test images with 855 tuples. Following [12], the test set contains BOS images (with background object-shadow pairs) and BOS-free images. Most of our experiments are based on DESOBAv2 dataset due to the following two concerns: 1) DESOBAv2 has larger test set which supports more comprehensive evaluation. 2) DESOBA has the artifacts caused by manual shadow removal and the existing methods (*e.g.*, SGRNet) tend to overfit such artifacts.

For the generated results, we evaluate both image quality and mask quality. For image evaluation, following [12], we adopt RMSE and SSIM, which are calculated based on the ground-truth target image and the generated image. Global RMSE (GR) and Global SSIM (GS) are calculated over the whole image, while Local RMSE (LR) and Local SSIM (LS) are calculated over the ground-truth foreground shadow region. For the mask evaluation, following [12], we adopt Balanced Error Rate (BER), which is calculated based on the ground-truth binary foreground shadow mask and the predicted foreground shadow mask obtained by threshold 0.5. Global BER (GB) is calculated over the whole image, while Local BER (LB) is calculated over the ground-truth foreground shadow region. Note that diffusion model has stochastic property and shadow generation is a multi-modal task, that is, one input has multiple plausible outputs. Similar to multi-modal inpainting evaluation [54, 55], we generate 5 results for one test image with different random seeds and select the one closest to the ground-truth (the highest Local SSIM) to calculate evaluation metrics.

6.2. Implementation Details

We develop our method with PyTorch 1.12.1 [30]. Our model is trained using the Adam optimizer [17] with a constant learning rate of $1e^{-5}$ over 50 epochs, on four NVIDIA RTX A6000 GPUs. Our method is built upon ControlNet [52]. We employ ResNet18 [10] as the intensity encoder. The mask predictor passes the concatenation of decoder feature maps and foreground object mask through four convolutional layers, with ReLU activation following the first three layers and Sigmoid activation following the last layer. We set the hyper-parameters w, σ, and λ as 10, 0.7, and 1, respectively.

6.3. Comparison with Baselines

Following [12], we compare with ShadowGAN [53], Mask-ShadowGAN [13], ARShadowGAN [22], and SGRNet [12]. We train and test all methods on DESOBAv2 dataset. The quantitative results are summarized in Table 1. We observe that our SGDiffusion achieves the lowest GRMSE, LRMSE and the highest GSSIM, LSSIM, which demonstrates that our method could generate shadow images that are closer to the ground-truth shadow images. The best GB and LB results demonstrate that the shapes and locations of our generated shadows are more accurate.

For qualitative comparison, we show several example results in Figure 4. Compared with the baseline methods, the shadows produced by our model have more reasonable shapes and intensities. Moreover, as shown in row 1, our method can take into account the self-occlusion of objects to generate discontinuous shadows. As shown in row 4, our method can also consider the material of the objects, producing shadows with translucency effects. We provide more examples in the supplementary.

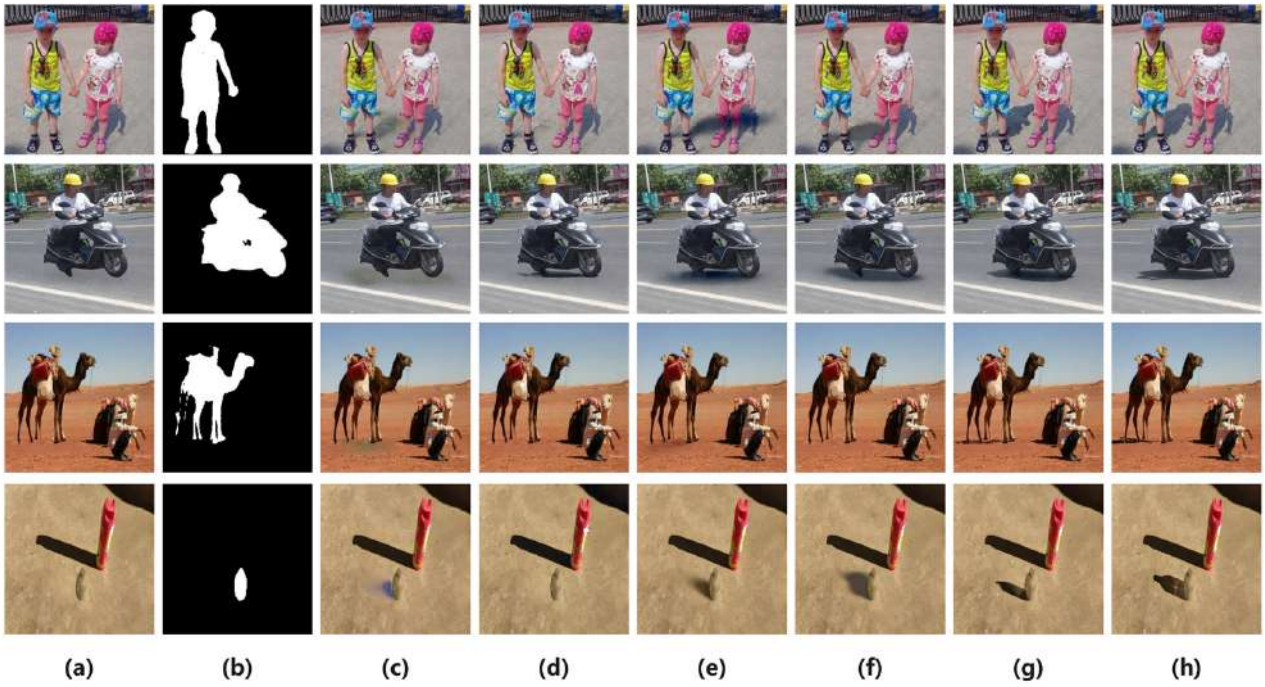

Figure 4. Visual comparison of different methods on DESOBAv2 dataset. From left to right are input composite image (a), foreground object mask (b), results of ShadowGAN [53] (c), MaskshadowGAN [13] (d), ARShadowGAN [22] (e), SGRNet [12] (f), our SGDiffusion (g), ground-truth (h).

Method	BOS Test Images						BOS-free Test Images					
	GR ↓	LR ↓	GS ↑	LS ↑	GB↓	LB↓	GR ↓	LR ↓	GS ↑	LS ↑	GB↓	LB↓
ShadowGAN [53]	7.511	67.464	0.961	0.197	0.446	0.890	17.325	76.508	0.901	0.060	0.425	0.842
MaskshadowGAN [13]	8.997	79.418	0.951	0.180	0.500	1.000	19.338	94.327	0.906	0.044	0.500	1.000
ARShadowGAN [22]	7.335	58.037	0.961	0.241	0.383	0.761	16.067	63.713	0.908	0.104	0.349	0.682
SGRNet [12]	7.184	68.255	0.964	0.206	0.301	0.596	15.596	60.350	0.909	0.100	0.271	0.534
SGDiffusion	**6.098**	**53.611**	**0.971**	**0.370**	**0.245**	**0.487**	**15.110**	**55.874**	**0.913**	**0.117**	**0.233**	**0.452**

Table 1. The results of different methods on DESOBAv2 dataset. The best results are highlighted in boldface.

6.4. Ablation Studies

We study the impact of weighted noise loss (WL), intensity modulation (IM), and post-processing (PP) of our SGDiffusion on BOS test images from DESOBAv2. The quantitative results are summarized in Table 2.

In row 1, we report the results of basic ControlNet without weighted noise loss. For WL, the comparison between row 3 and row 1 emphasizes the importance of paying more attention to the foreground shadow region. We also report a special case † in row 2, where the foreground shadow mask is not expanded when constructing the weight map. The results in row 2 are comparable or even worse than those in row 1, as the model tends to generate larger shadow size while ignoring shape and edge details. For IM, the comparison between row 1 and row 5 shows that the intensity modulation can significantly improve the shadow quality by adjusting the shadow intensity. We also report a special case ∘ in row 4, where the intensity encoder input does not contain background shadow mask. The comparison between row 4 and row 5 shows that background shadow mask is helpful, because the background shadow regions and their surrounding regions could provide useful clues to infer shadow intensity. For PP, the comparison between row 6 and row 7 demonstrates that post-processing effectively corrects color shift and background variations, substantially reducing the

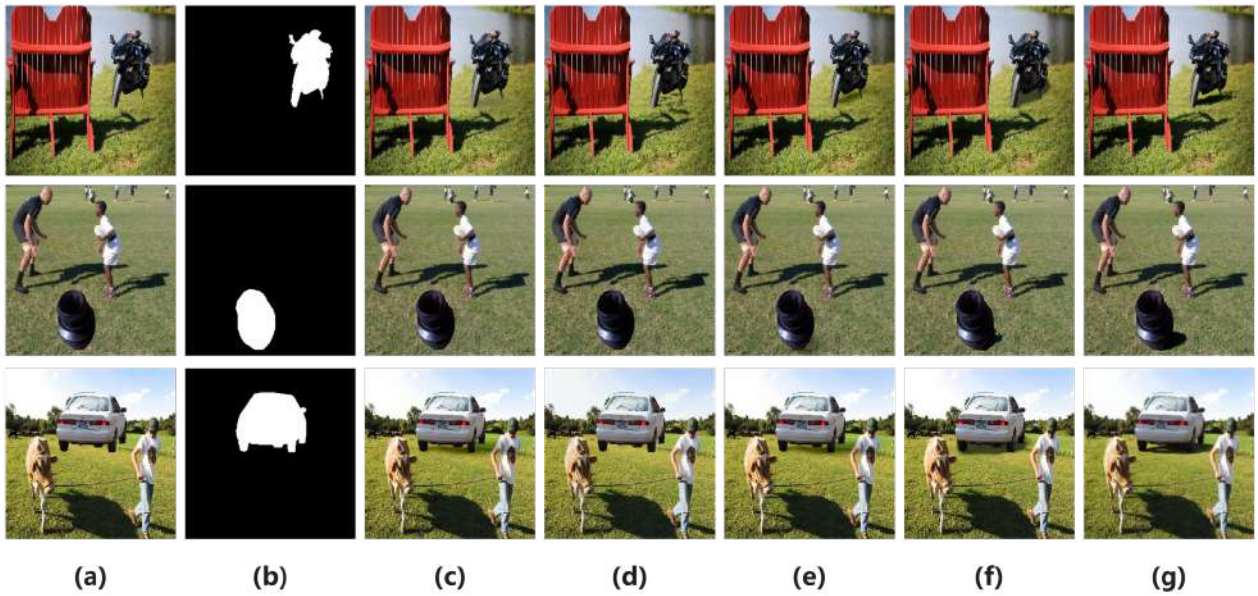

Figure 5. Visual comparison of different methods on real composite images. From left to right are input composite image (a), foreground object mask (b), results of ShadowGAN [53] (c), MaskshadowGAN [13] (d), ARShadowGAN [22] (e), SGRNet [12] (f), SGDiffusion (g).

Row	WL	IM	PP	GR ↓	LR ↓	GB↓	LB↓
1	-	-	+	8.285	59.753	0.271	0.534
2	†	-	+	8.319	59.491	0.282	0.563
3	+	-	+	7.041	53.829	0.249	0.492
4	-	○	+	7.410	56.121	0.269	0.536
5	-	+	+	7.357	54.159	0.262	0.526
6	+	+	-	13.447	55.231	0.245	0.487
7	+	+	+	6.098	53.611	0.245	0.487

Table 2. Ablation studies of our method on BOS test images from DESOBAv2 dataset. WL is short for weighted loss and † means without expanding shadow mask. IM is short for intensity modulation and ○ means without using background shadow mask. PP is short for post-processing.

global RMSE. We also provide the visual results of ablated versions in the supplementary.

6.5. Real Composite Images

We compare different methods on real composite images provided by [12], where background images and foreground objects are from the DESOBA [12] test set. We train all methods on DESOBAv2 and finetune them on DESOBA. The visual results of different methods are showcased in Figure 5. These results confirm that SGDiffusion adeptly synthesizes lifelike shadows with precise contours, locations, and directions, which are compatible with the background object-shadow pairs and foreground object information. In contrast, previous methods often produce vague and misdirected shadows. We provide more examples in the supplementary.

Given the absence of ground-truth images for real composite images, following [12], we opt for subjective evaluation, engaging 50 human raters in the user study. Each participant is presented with image pairs from the results generated by 5 methods, and asked to choose the image with more realistic foreground shadow. Using the Bradley-Terry model [2], we report the B-T scores in the supplementary, which again proves the advantage of our method.

7. Conclusion

In this paper, we have contributed a large-scale shadow generation dataset DESOBAv2. We have also designed a novel diffusion-based shadow generation method. Extensive experimental results show that our method is able to generate plausible shadows for composite foregrounds, significantly surpassing previous methods.

8. Acknowledgement

The work was supported by the National Natural Science Foundation of China (Grant No. 62076162), the Shanghai Municipal Science and Technology Major/Key Project, China (Grant No. 2021SHZDZX0102).

References

[1] Ibrahim Arief, Simon McCallum, and Jon Yngve Hardeberg. Realtime estimation of illumination direction for augmented reality on mobile devices. In *CIC*, 2012. 2

[2] Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952. 8

[3] Wenyan Cong, Jianfu Zhang, Li Niu, Liu Liu, Zhixin Ling, Weiyuan Li, and Liqing Zhang. Dovenet: Deep image harmonization via domain verification. In *CVPR*, 2020. 1, 2

[4] Wenyan Cong, Li Niu, Jianfu Zhang, Jing Liang, and Liqing Zhang. Bargainnet: Background-guided domain translation for image harmonization. In *ICME*, 2021.

[5] Wenyan Cong, Xinhao Tao, Li Niu, Jing Liang, Xuesong Gao, Qihao Sun, and Liqing Zhang. High-resolution image harmonization via collaborative dual transformations. In *CVPR*, 2022.

[6] Xiaodong Cun and Chi-Man Pun. Improving the harmony of the composite image by spatial-separated attention module. *TIP*, 2020. 2

[7] Marc-André Gardner, Yannick Hold-Geoffroy, Kalyan Sunkavalli, Christian Gagné, and Jean-Francois Lalonde. Deep parametric indoor lighting estimation. In *ICCV*, 2019. 2

[8] Lanqing Guo, Siyu Huang, Ding Liu, Hao Cheng, and Bihan Wen. Shadowformer: Global context helps image shadow removal. In *AAAI*, 2023. 3

[9] Lanqing Guo, Chong Wang, Wenhan Yang, Siyu Huang, Yufei Wang, Hanspeter Pfister, and Bihan Wen. Shadowdiffusion: When degradation prior meets diffusion model for shadow removal. In *CVPR*, 2023. 3

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 6

[11] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 3

[12] Yan Hong, Li Niu, and Jianfu Zhang. Shadow generation for composite image in real-world scenes. *AAAI*, 2022. 1, 2, 6, 7, 8

[13] Xiaowei Hu, Yitong Jiang, Chi-Wing Fu, and Pheng-Ann Heng. Mask-shadowgan: Learning to remove shadows from unpaired data. In *ICCV*, 2019. 6, 7, 8

[14] Shaozong Huang and Lan Hong. Diffusion model for mural image inpainting. In *ITOEC*, 2023. 3

[15] Kevin Karsch, Kalyan Sunkavalli, Sunil Hadap, Nathan Carr, Hailin Jin, Rafael Fonte, Michael Sittig, and David Forsyth. Automatic scene inference for 3d object compositing. *ACM TOG*, 2014. 2

[16] Eric Kee, James F. O'Brien, and Hany Samir Farid. Exposing photo manipulation from shading and shadows. *ACM TOG*, 2014. 2

[17] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014. 6

[18] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *CoRR*, abs/1312.6114, 2013. 4

[19] Bin Liao, Yao Zhu, Chao Liang, Fei Luo, and Chunxia Xiao. Illumination animating and editing in a single picture using scene structure estimation. *Computers & Graphics*, 82:53–64, 2019. 2

[20] Chen-Hsuan Lin, Ersin Yumer, Oliver Wang, Eli Shechtman, and Simon Lucey. St-gan: Spatial transformer generative adversarial networks for image compositing. In *CVPR*, 2018. 2

[21] Bin Liu, Kun Xu, and Ralph R Martin. Static scene illumination estimation from videos with applications. *JCST*, 32 (3):430–442, 2017. 2

[22] Daquan Liu, Chengjiang Long, Hongpan Zhang, Hanning Yu, Xinzhi Dong, and Chunxia Xiao. Arshadowgan: Shadow generative adversarial network for augmented reality in single light scenes. In *CVPR*, 2020. 1, 2, 6, 7, 8

[23] Guilin Liu, Fitsum A. Reda, Kevin J. Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. Image inpainting for irregular holes using partial convolutions. In *ECCV*, 2018. 3

[24] Jian Ma, Mingjun Zhao, Chen Chen, Ruichen Wang, Di Niu, Haonan Lu, and Xiaodong Lin. Glyphdraw: Learning to draw chinese characters in image synthesis models coherently. *CoRR*, abs/2303.17870, 2023. 5

[25] Quanling Meng, Shengping Zhang, Zonglin Li, Chenyang Wang, Weigang Zhang, and Qingming Huang. Automatic shadow generation via exposure fusion. *IEEE Transactions on Multimedia*, 2023. 2

[26] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *3DV*, 2016. 5

[27] Chong Mou, Xintao Wang, Liangbin Xie, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *CoRR*, abs/2302.08453, 2023. 3

[28] Li Niu, Wenyan Cong, Liu Liu, Yan Hong, Bo Zhang, Jing Liang, and Liqing Zhang. Making images real again: A comprehensive survey on deep image composition. *CoRR*, abs/2106.14490, 2021. 1

[29] Sibam Parida, Vignesh Srinivas, Bhavishya Jain, Rajesh Naik, and Neeraj Rao. Survey on diverse image inpainting using diffusion models. In *PCEMS*, 2023. 3

[30] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *NIPS*, 32, 2019. 6

[31] Patrick Pérez, Michel Gangnet, and Andrew Blake. Poisson image editing. In *SIGGRAPH*. 2003. 2

[32] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 1, 2, 3, 4

[33] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015. 4

[34] Yichen Sheng, Jianming Zhang, and Bedrich Benes. Ssn: Soft shadow network for image compositing. In *CVPR*, 2021. 1

[35] Yichen Sheng, Yifan Liu, Jianming Zhang, Wei Yin, A Cengiz Oztireli, He Zhang, Zhe Lin, Eli Shechtman, and Bedrich Benes. Controllable shadow generation using pixel height maps. In *ECCV*, 2022. 1, 2

[36] Yichen Sheng, Jianming Zhang, Julien Philip, Yannick Hold-Geoffroy, Xin Sun, He Zhang, Lu Ling, and Bedrich Benes. Pixht-lab: Pixel height based light effect generation for image compositing. In *CVPR*, 2023. 1, 2

[37] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *CoRR*, abs/2010.02502, 2020. 3

[38] Yi-Zhe Song, Zhifei Zhang, Zhe L. Lin, Scott D. Cohen, Brian L. Price, Jianming Zhang, Soo Ye Kim, and Daniel G. Aliaga. Objectstitch: Generative object compositing. In *CVPR*, 2023. 2

[39] Xinhao Tao, Junyan Cao, Yan Hong, and Li Niu. Shadow generation with decomposed mask prediction and attentive shadow filling. In *AAAI*, 2024. 2

[40] Yi-Hsuan Tsai, Xiaohui Shen, Zhe Lin, Kalyan Sunkavalli, Xin Lu, and Ming-Hsuan Yang. Deep image harmonization. In *CVPR*, 2017. 2

[41] Tianyu Wang, Xiaowei Hu, Pheng-Ann Heng, and Chi-Wing Fu. Instance shadow detection with a single-stage detector. *TPAMI*, 2022. 2, 3

[42] Huikai Wu, Shuai Zheng, Junge Zhang, and Kaiqi Huang. Gp-gan: Towards realistic high-resolution image blending. In *ACM MM*, 2019. 2

[43] Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *CVPR*, 2023. 5

[44] Binxin Yang, Shuyang Gu, Bo Zhang, Ting Zhang, Xuejin Chen, Xiaoyan Sun, Dong Chen, and Fang Wen. Paint by example: Exemplar-based image editing with diffusion models. In *CVPR*, 2023. 1, 2

[45] Shiyuan Yang, Xiaodong Chen, and Jing Liao. Uni-paint: A unified framework for multimodal image inpainting with pretrained diffusion model. In *ACM MM*, 2023. 3

[46] Fangneng Zhan, Jiaxing Huang, and Shijian Lu. Adaptive composition gan towards realistic image synthesis. *CoRR*, abs/1905.04693, 2019. 2

[47] Fangneng Zhan, Shijian Lu, Changgong Zhang, Feiying Ma, and Xuansong Xie. Towards realistic 3d embedding via view alignment. *CoRR*, abs/2007.07066, 2020. 2

[48] Bo Zhang, Yuxuan Duan, Jun Lan, Yan Hong, Huijia Zhu, Weiqiang Wang, and Li Niu. Controlcom: Controllable image composition using diffusion model. *arXiv preprint arXiv:2308.10040*, 2023. 1, 2

[49] He Zhang, Jianming Zhang, Federico Perazzi, Zhe Lin, and Vishal M Patel. Deep image compositing. In *WACV*, 2021. 2

[50] Jinsong Zhang, Kalyan Sunkavalli, Yannick Hold-Geoffroy, Sunil Hadap, Jonathan Eisenman, and Jean-François Lalonde. All-weather deep outdoor lighting estimation. In *CVPR*, 2019. 2

[51] Lingzhi Zhang, Tarmily Wen, and Jianbo Shi. Deep image blending. In *WACV*, 2020. 2

[52] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023. 2, 3, 4, 6

[53] Shuyang Zhang, Runze Liang, and Miao Wang. Shadowgan: Shadow synthesis for virtual objects with conditional adversarial networks. *Computational Visual Media*, 5:105–115, 2019. 1, 2, 6, 7, 8

[54] Lei Zhao, Qihang Mo, Sihuan Lin, Zhizhong Wang, Zhiwen Zuo, Haibo Chen, Wei Xing, and Dongming Lu. Uctgan: Diverse image inpainting based on unsupervised cross-space translation. In *CVPR*, 2020. 6

[55] Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai. Pluralistic image completion. In *CVPR*, 2019. 6

[56] Haitian Zheng, Zhe Lin, Jingwan Lu, Scott Cohen, Eli Shechtman, Connelly Barnes, Jianming Zhang, Ning Xu, Sohrab Amirghodsi, and Jiebo Luo. Image inpainting with cascaded modulation gan and object-aware training. In *ECCV*, 2022. 3