对 DIFFUSEMIX:Label-Preserving Data Augmentation with Diffusion Models 的复现研究

摘要

本文研究了基于输入变换的对抗图像生成，旨在提升对抗攻击在计算机视觉中的有效性与可迁移性。首先，本文回顾了 FGSM、I-FGSM、MI-FGSM 等经典攻击方法，并介绍了 DiffuseMix、Block Shuffle and Rotation 等相关研究。受 DiffuseMix 启发，本文提出了一种新型方法：将 DiffuseMix 作为变换算子集成到 MI-FGSM 中，随后简化模型，移除扩散部分，仅保留分形图形生成机制，并结合分块旋转思想改进攻击的迭代过程。实验结果表明，单独引入分形图形相较于传统攻击方法能够有效提升对抗图像的攻击性能，引入分块旋转方法也能有效增强攻击效果。然而，与当前的最先进 SOTA 方法相比，本文方法仍存在一定差距。总体而言，分形图形与分块旋转的结合为生成优秀的对抗图像提供了新的思路，并为未来的研究方向提供了思考方向。

关键词：对抗攻击；图像增强；分形图形

1 引言

对抗攻击自提出以来，在计算机视觉领域的影响逐渐显现，且其历史演变推动了技术创新和理论发展。最初，2014 年 Goodfellow 等人提出了对抗样本的概念，并展示了通过微小扰动可以使深度神经网络产生错误分类。这一发现揭示了深度学习模型在面对看似无害的扰动时的脆弱性。此后，研究者们深入探讨了对抗攻击的不同形式，包括白盒攻击和黑盒攻击，并不断探索如何防御这些攻击。对抗攻击的出现和发展，虽然带来了一些挑战，但也成为推动计算机视觉技术进步的重要动力。对抗攻击虽然在计算机视觉任务中带来了挑战，但也促使了该领域的技术进步和创新，具有积极的推动作用。首先，对抗攻击的存在迫使研究者们更加关注视觉模型的鲁棒性。为了应对对抗攻击，研究者们开发了各种防御技术，如对抗训练、梯度遮蔽、输入数据的预处理等方法，这些技术增强了模型在面对恶意干扰时的稳定性和准确性。通过对抗训练，模型能够在训练过程中学习到如何区分正常数据与带有扰动的数据，从而提高其对真实世界中复杂环境的适应能力。这种鲁棒性提升不仅在防御对抗攻击方面有效，也使得视觉系统在实际应用中能够处理更多的噪声和干扰，增强了系统的可靠性和稳定性。

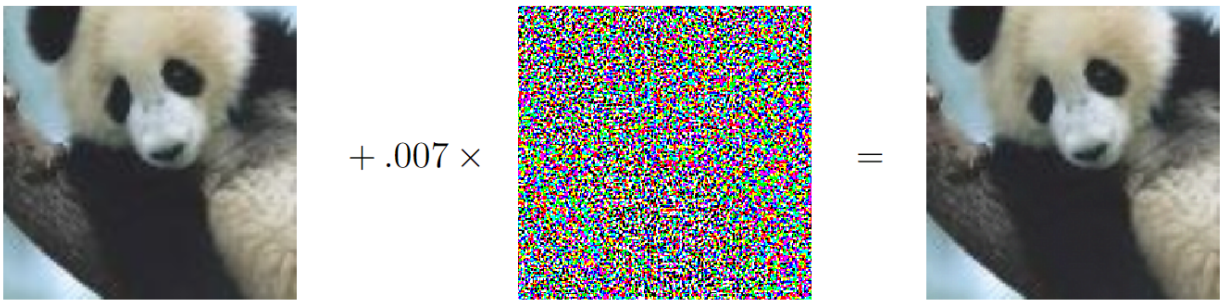

图 1. 对抗攻击示意图

其次，对抗攻击也推动了模型的透明度和可解释性研究。在应对对抗攻击时，研究者不仅注重提高模型的精度，还加大了对模型内部机制的分析，探索如何使模型决策过程更加透明。了解模型为何在特定情况下被对抗攻击误导，有助于揭示深度学习模型的潜在弱点，并帮助设计出更具可解释性的模型。这种解释性提升对于安全性、可审计性和可信赖性至关重要，尤其是在金融、医疗和公共安全等高风险领域。通过深入研究模型的工作原理和攻击方式，能够开发出更加可信、可验证的人工智能系统。

对抗攻击推动了跨学科的技术融合和新的创新应用。在计算机视觉领域，对抗攻击不仅促进了算法和防御技术的发展，还激发了硬件设计、加密技术、网络安全等领域的创新。研究人员在防御对抗攻击时，往往需要结合硬件加速、数据隐私保护等技术，推动了跨领域的协作和技术突破。这种跨学科的合作为计算机视觉的发展带来了更多的机会和解决方案，也促使了新兴应用场景的诞生。例如，结合区块链技术的防护机制可以增强数据的安全性，而在移动设备上实现更加高效的对抗防御也开辟了新的市场需求。因此，对抗攻击不仅对视觉任务本身产生了深远的影响，还推动了更广泛的技术进步和创新。

通过阅读一种新的图像增强方法:DIFFUSEMIX: Label-Preserving Data Augmentation with Diffusion Models，考虑到其引入的扩散模型和分形图形能为图像引入更多丰富的信息，本文提出了一种新的基于输入变换的对抗图像生成方法，并在此基础上进行了大量实验与改进。

2 相关工作

本次攻击方法在白盒模型上使用基于梯度的攻击方法生成对抗样本，生成了对抗样本后，再在不同的黑盒模型上进行测试。下面是一些相关的基于梯度的攻击方法：

2.1 FGSM

FGSM [2]（Fast Gradient Sign Method）是最早且最经典的对抗攻击方法之一，由 Ian Goodfellow 等人在 2014 年提出。FGSM 通过利用深度学习模型的梯度信息，生成能够使模型预测错误的对抗样本，它是一种非常高效、简单且具有强攻击性的攻击方法。其核心思想是利用模型的梯度信息对输入数据进行微小的扰动，从而生成对抗样本。这些扰动会迫使模型做出错误的预测。具体来说，FGSM 通过计算输入图像对损失函数的梯度，然后沿着梯度的方向添加扰动。为了确保扰动对图像的影响尽可能小，FGSM 仅沿着梯度的符号方向进行调整。

$$\mathbf{x}^{adv} = \mathbf{x} + \epsilon \cdot \text{sign}\left(\nabla_x J(\mathbf{x}, \mathbf{y}; \theta)\right)$$

- 通过反向传播计算输入图像 x 相对于损失函数的梯度 $\nabla_x J(\mathbf{x}, \mathbf{y}; \theta)$
- 然后，根据梯度信息，沿着梯度的符号方向生成扰动 $\epsilon \cdot \text{sign}\left(\nabla_x J(\mathbf{x}, \mathbf{y}; \theta)\right)$
- 将生成的扰动加到原始输入图像上，得到新的对抗样本 $\mathbf{x}^{adv}$

FGSM 攻击通过非常简单且计算高效的方式，在非常短的时间内生成对抗样本。尽管其产生的扰动通常较小（通过控制的大小来调整扰动幅度），但是这些对抗样本通常能使得深度神经网络产生错误的预测。由于扰动很小，人眼几乎难以察觉，因此对抗样本对人类视觉系统来说几乎是“无害”的，但对机器学习模型却能造成显著影响。

2.2 I-FGSM

I-FGSM [4] 是对原始 FGSM 的一种改进版本，旨在通过多次迭代更新扰动，从而生成更强、更有效的对抗样本。I-FGSM 在每次迭代时计算新的梯度并更新对抗样本，从而逐步增强攻击的强度。这个方法通过反复调整对抗样本，使得其在增加扰动幅度的同时，仍保持对人眼的不可察觉性，但却能够有效地误导深度学习模型。

其核心思想与 FGSM 类似，都是基于梯度信息生成对抗样本，但 I-FGSM 通过多次迭代优化对抗扰动，逐步增强攻击效果。每次迭代中，I-FGSM 通过计算输入图像相对于损失函数的梯度，沿着梯度方向调整输入图像，并生成新的对抗样本。通过重复这一过程，I-FGSM 能够在每次更新时逐步增强扰动的强度，从而使得最终的对抗样本能够更加有效地攻击目标模型。

$$\mathbf{x}_t^{adv} = \mathbf{x}_{t-1}^{adv} + \alpha \cdot \text{sign}\left(\nabla_{\mathbf{x}_{t-1}^{ad}} v\left(\mathbf{x}_{t-1}^{adv}, \mathbf{y}; \theta\right)\right)$$

其中 $\mathbf{x}_t^{adv}$ 是第 t 次迭代的对抗样本，α 是每次迭代的步长。与单次更新的 FGSM 相比，I-FGSM 通常能够生成更强大的对抗样本，从而在对抗训练、模型评估等场景中产生更大的攻击效果。

2.3 MI-FGSM

MI-FGSM [1]（Momentum Iterative Fast Gradient Sign Method）是对 I-FGSM 的进一步改进，旨在通过引入动量机制，增强对抗攻击的效果。MI-FGSM 是由 Dong et al. 在 2017 年

提出的，它的核心思想是通过动量的积累来缓解迭代过程中扰动的震荡，使得生成的对抗样本更加有效，并提高攻击的成功率。

在 I-FGSM 中，每次迭代时的梯度更新仅仅基于当前步骤的梯度，而 MI-FGSM 则利用一个动量变量（类似于动量梯度下降中的动量项），将历史梯度信息加权累积起来，从而使得扰动的更新方向更加平稳和一致。这种方法有助于提高攻击效果，特别是在对抗训练或者防御性模型下，生成的对抗样本往往能更有效地攻击目标模型。

$$g_t = \mu \cdot g_{t-1} + \frac{\nabla_{\mathbf{x}_{t-1}^{adv}} J\left(\mathbf{x}_{t-1}^{adv}, \mathbf{y}; \theta\right)}{\left\| \nabla_{\mathbf{x}_{t-1}^{adv}} J\left(\mathbf{x}_{t-1}^{adv}, \mathbf{y}; \theta\right) \right\|_1},$$
$$\mathbf{x}_t^{adv} = \mathbf{x}_{t-1}^{adv} + \alpha \cdot \operatorname{sign}\left(g_t\right), g_0 = 0$$

MI-FGSM 通过引入动量机制，显著提高了对抗攻击的效果，尤其是在面对高鲁棒性模型时，能够生成更强大的对抗样本。它通过减少梯度震荡、增强攻击方向的一致性，使得攻击更加稳定和高效，成为了生成对抗样本的重要方法之一。尽管增加了计算复杂度，但在攻击效果和效率之间做出了较好的平衡，是对抗攻击领域中的一种有效策略。此外，假设 T 是一个变换算子，现有的基于输入变换的攻击通常集成到 MI-FGSM 中以提高对抗可迁移性，即使用下列式子替换原式中的部分。

$$\nabla_{\mathbf{x}_{t-1}^{adv}} J\left(\mathcal{T}(\mathbf{x}_{t-1}^{adv}), \mathbf{y}; \theta\right)$$

2.4 DiffuseMix

DIFFUSEMIX: Label-Preserving Data Augmentation with Diffusion Models（CVPR2024）[3] 是一种利用扩散模型进行图像增强的方法。扩散模型（Diffusion Models）最近在图像生成和修复任务中表现出了优异的性能，它们通过逐步添加噪声并学习逆过程来生成清晰图像。传统的数据增强方法（如旋转、平移、裁剪等）在某些情况下可能无法充分生成多样化的训练数据，而扩散模型生成的图像可以有效地增强数据多样性。

DiffuseMix 面临的一个关键挑战是扩散模型生成图像时的不稳定性和不可控性。为了克服这个问题，作者提出了一种创新的方法，通过引入一组精心设计的 prompt 来控制扩散模型的生成过程。这些 prompt 能够引导扩散模型生成具有更好可控性的图像，确保生成的图像在保留原图像语义的基础上，还能引入新的信息，从而丰富图像的内容。

首先，为了提高生成图像的可控性，作者设计了一组 prompt 来指导扩散模型的生成过程。这些 prompt 帮助模型生成带有滤镜风格的图像，从而在保留原始图像的语义信息的同时，也能加入新的风格或内容。例如，通过引导模型生成某种特定风格（如油画风格、素描风格等）的图像，使得生成的图像在视觉上具有更多变种和多样性。

随后，生成图像与原图像拼接的过程是通过预定义的四种 mask 来完成的。具体来说，作者使用线性组合的方式将原图像和生成的图像在这些 mask 区域内进行拼接融合。这种方法能够有效地将生成图像与原图像结合，同时保持图像的语义一致性。通过这种拼接，既保留了原图像的核心信息，又引入了生成图像中新的风格或细节，极大地丰富了数据集的多样性。

最后，作者从提前收集的分形图形库中随机选取一张图像，将其与拼接后的图像进行进一步融合。这一过程在增加图像的复杂性和多样性的同时，保证了生成图像仍然具备原图像

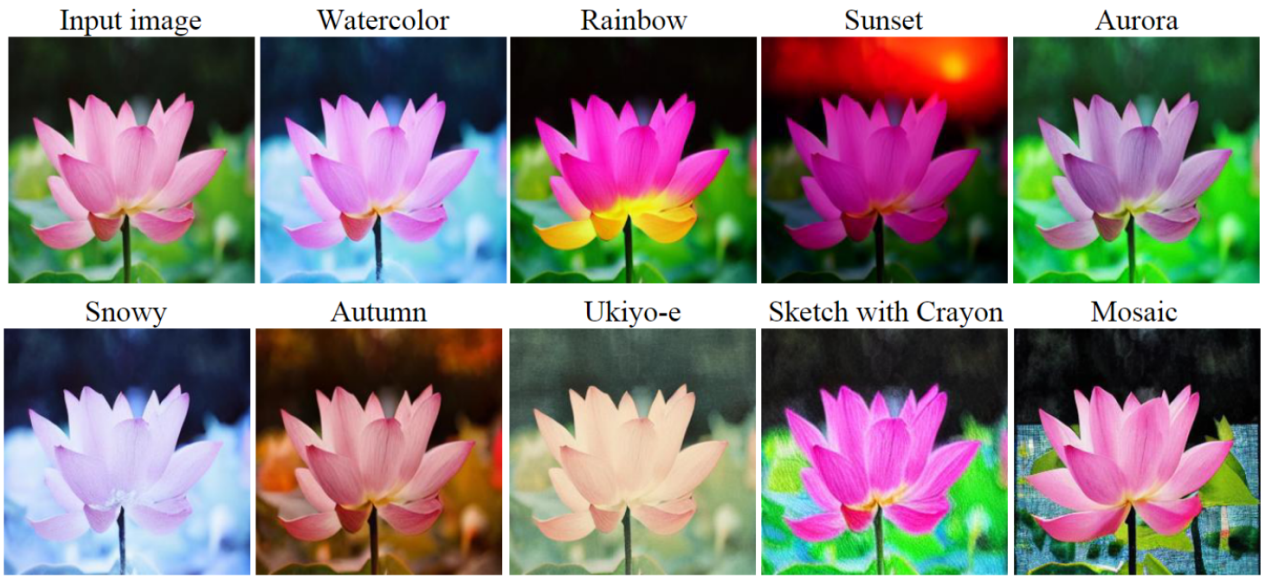

图 2. 经过扩散模型生成的图像

的基本结构一致性。我们推测，分形图形的引入能使得生成图像在视觉上更加丰富，并且能在不同的任务中提高模型的鲁棒性和泛化能力。

2.5 Block Shuffle and Rotation

《Boosting Adversarial Transferability by Block Shuffle and Rotation》[5] 是 2024 年 CVPR 接受发表的一篇基于输入变换的对抗攻击方法论文。作者在文中提出了一种新颖的对抗攻击方法，旨在通过打破图像内在的空间结构关系来提高攻击的可迁移性。与现有攻击方法相比，这种新方法在提升对抗攻击的迁移能力方面具有显著优势。

现有的对抗攻击方法虽然在白盒攻击场景中表现优异，但其可迁移性较差，往往难以在不同的模型之间有效迁移。不同的模型在架构和参数上存在差异，如何捕捉并利用这些模型之间的共享特征，成为提升攻击可迁移性的关键。作者提出，打破图像的空间一致性，重新排列图像中的局部块，可以帮助攻击方法在不同模型间迁移得更好。

此外，作者还深入分析了注意力热图的差异问题。研究发现，传统攻击方法在白盒模型上生成的对抗样本，其注意力热图与目标黑盒模型的热图存在显著差异，这种差异限制了攻击的迁移性。为了克服这一问题，作者设计了一种块重排和旋转的输入变换方法，使得对抗样本在不同模型之间的注意力热图更加一致，从而有效提升了攻击的可迁移性。

作者提出，人类在物体的某些部分被遮挡住时，可以心理重建出被遮挡的部分，得益于我们对物体固有的内在关系的认知。比如马腿位于马的身体下方，这些是对马的固有内在关系的认知。作者从这一思路出发，尝试使用混洗和旋转扰乱内在关系，影响注意力热图。

新的输入变换方法首先将图像随机拆分成 n×n 个小块，以打破图像中局部区域之间的空间结构关系。然后，作者对这些小块进行随机混洗（重新排列），进一步扰乱图像的内在顺序和上下文信息。为了增强对抗性并进一步破坏图像的结构性关系，每个小块还会在 $[-\tau,\tau]$ 的范围内，随机旋转一个角度 。这一过程不仅打乱了图像块之间的相对位置，还通过旋转操作增加了图像的多样性，进一步提高了对抗样本的强度和可迁移性。通过这种方式，生成的对抗样本在不同模型上表现出更高的迁移效果，克服了传统方法中由于图像内在关系未被有效

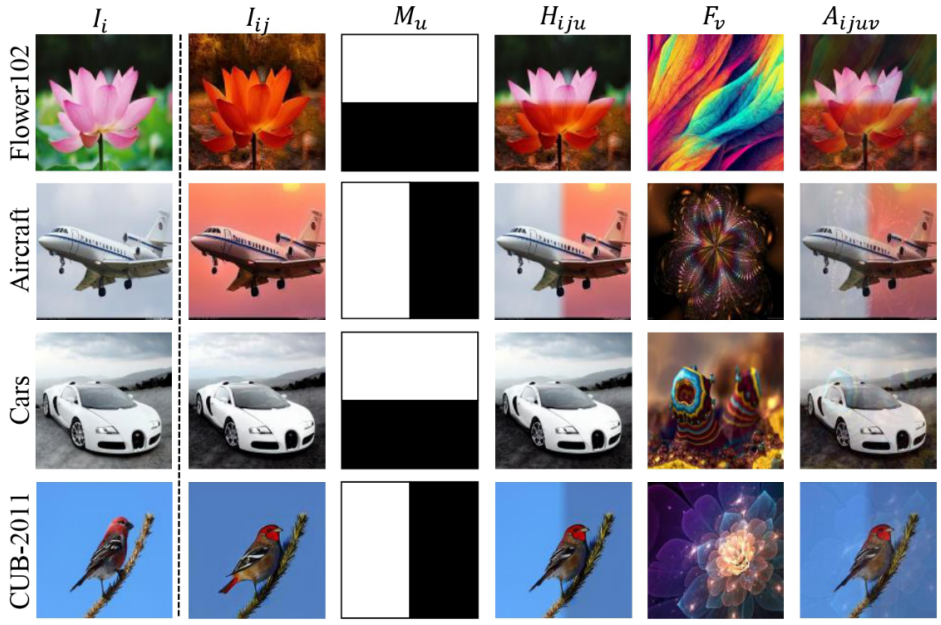

图 3. DiffuseMix 图像增强示例

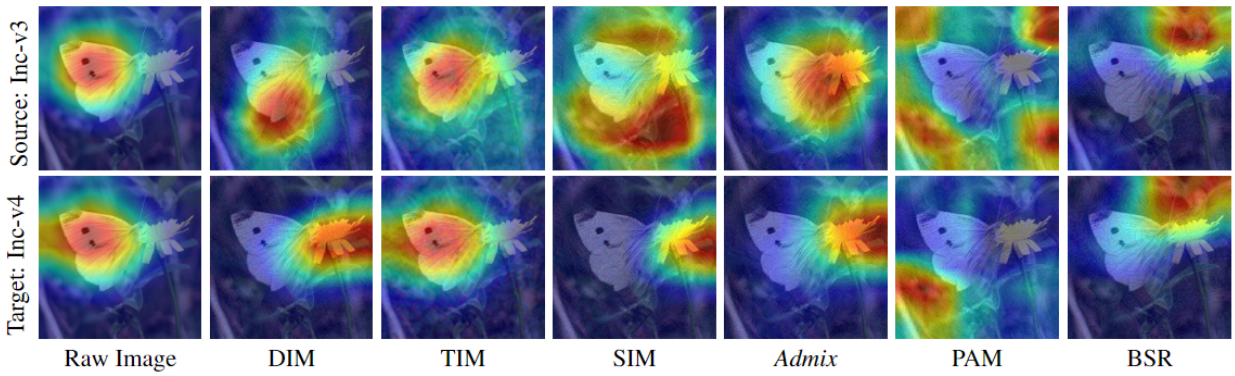

Figure 2. Attention heatmaps of adversarial examples generated by various input transformations using Grad-CAM.

图 4. 注意力热图示意

图 5. 注意力热图的打乱与重排

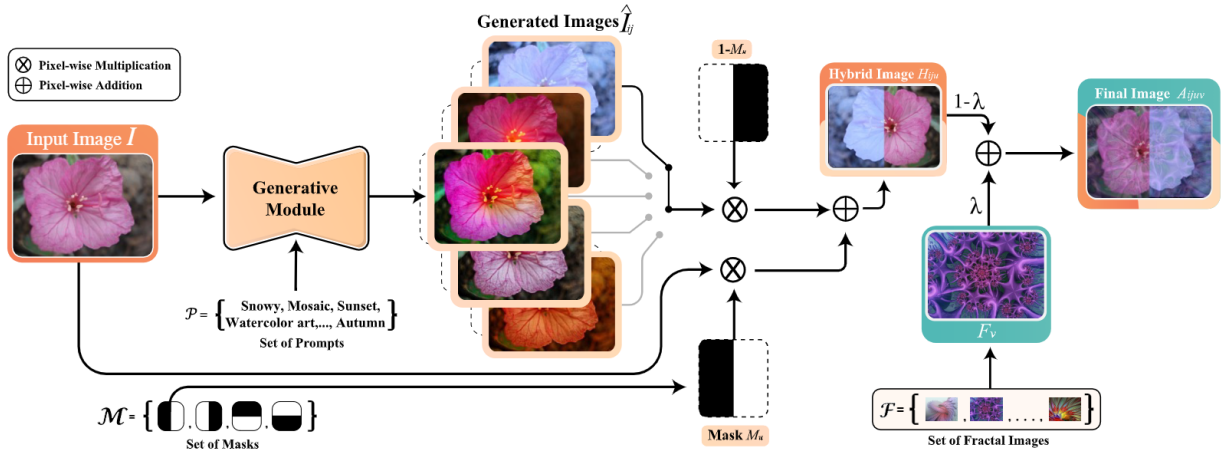

图 6. DiffuseMix 方法架构图

破坏而导致的迁移性能不足的问题。

3 本文方法

正是因为分形图形和扩散模型对提升图像风格细节存在巨大的潜力，在接下来的工作中，我们将这个新的图像增强方法（DiffuseMix）引入基于输入变换的对抗攻击中。

3.1 DiffuseMix

受到 DiffuseMix 方法的启发，在本篇工作中，我首先将 DiffuseMix 方法移植为一个变换算子 $\mathcal{T}$，并将其集成到 MI-FGSM 方法中，从而提出了一种全新的对抗图像生成方法。在实验开始之前，通过分析可知，生成模型能够引入更多的图像细节，这为生成更加复杂的对抗图像提供了巨大的潜力。同时，分形图形的复杂性也为提高对抗攻击方法的可迁移性提供了强大的支持。

3.2 分形迭代

分形图形（Fractal Graphics）是指通过自相似的几何结构生成的图形，其特点是无论从哪个尺度上观察，图形的结构都会展示出相似的形态。这种结构通常是通过数学公式或递归算法生成的。分形图形的一个显著特征是自相似性，即在不同的尺度下，图形的细节和整体形态会保持相似。

考虑到扩散模型在可控性和时间开销方面的挑战，我在方法 3.1 的基础上对其进行了简化，移除了扩散模型的生成部分，仅保留了分形图形的引入。预计这一调整能够在保持对抗性强度的同时，提升方法的效率，从而在性能上取得一定的改进。

3.3 块旋转与分形迭代

在方法 3.2 的基础上，通过融合 Block Shuffle and Rotation 的思想，进一步增强了对抗样本的生成过程。在进行分形图形融合之前，首先对输入图像进行分块处理，并对每个小块

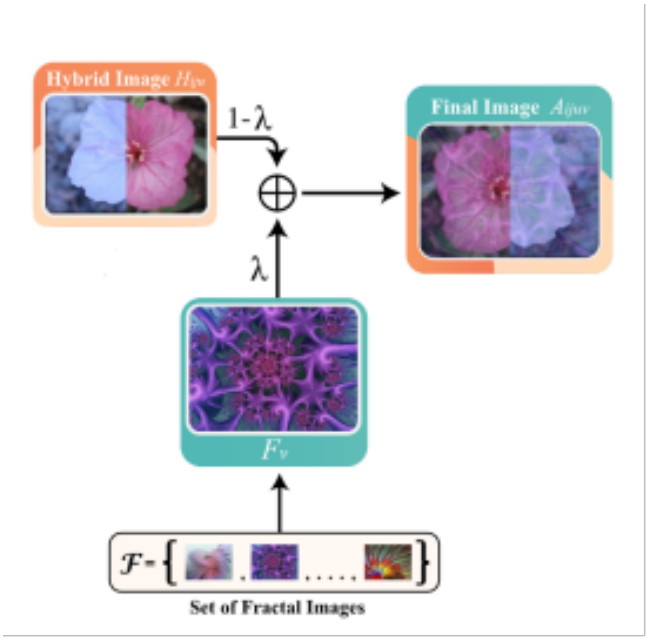

图 7. 分形融合

进行细微的随机旋转。通过这种方式，图像的原始结构被进一步扰乱，从而有效打破图像内部的空间关系和局部一致性。

3.4 多重块旋转与分形迭代

在方法 3.3 的基础上，我对每次迭代过程进行了改进。对于同一张输入图像 x，在每次迭代中生成 n 张不同的对抗样本，并将这些对抗样本的生成信息进行平均。这一策略通过引入更多的随机性和不确定性，使得生成的对抗样本在多个方向上展现出更多的变化。通过集成多次生成的结果，可以有效避免模型对特定模式的过拟合，同时提高对抗样本的普适性。

4 复现细节

4.1 与已有开源代码对比

开源代码实现了图像增强方法，相较于开源代码，我在一个对抗攻击检测框架上复现了其功能，并做了性能测试。DiffuseMix 开源代码：https://github.com/khawar-islam/diffuseMix。对抗攻击检测框架：https://github.com/Trustworthy-AI-Group/TransferAttack。在移植过程中，参考了 DiffuseMix 源代码中的分形图形库与图像增强方法，将其移植到 MI-FGSM 方法中，进行效果检测。随后，通过移除扩散模型模块，引入块旋转思想继续了后续的实验与改进。

4.2 实验环境搭建

TransferAttack 框架配置：

- Python >= 3.6
- PyTorch >= 1.12.1

Attack	CNNs				ViTs		
	Resnet-18	ResNet-101	ResNeXt-50	DenseNet-101	ViT	PiT	Visformer
MI-FGSM	100	42.7	46.8	74.6	17.4	23.4	33.6
DiffuseMix(分形迭代)	100	51.8	55	84.7	18.3	25.3	40.6
DiffuseMix	99.7	30.4	35	64.2	11.8	16.1	25.7
DeCoWA(SOTA)	100	85.1	87.7	98.4	55.3	65.5	80.4

表 1. 初步实验结果

- Torchvision >= 0.13.1
- timm >= 0.6.12

数据集：从 ImageNet 验证集中随机抽取了 1,000 幅图像，其中每幅图像属于一个类别，并且能够被所采用的模型正确分类。

4.3 创新点

- 引入扩散模型用于生成对抗样本，为对抗样本的生成引入更多丰富信息。
- 引入分形图形用于生成对抗样本，为对抗样本的生成引入更多复杂信息。
- 将分形图形和块旋转思想结合用于生成对抗样本。

5 实验结果分析

5.1 DiffuseMix and 分形迭代

表1是在代理模型 ResNet-18 上生成的对抗样本在其他模型上的攻击结果：

在实验过程中，我发现引入扩散模型会导致生成对抗图像的效率大幅下降，并且实验结果没有带来预期的进步，反而有所退步。具体而言，在其他模型上的攻击成功率甚至低于原始的 MI-FGSM 方法。因此，我决定将生成图像的部分 l 从方法中移除，重新测试仅引入分形图形的策略。

新的实验结果表明，在 ResNet-18 上生成的对抗样本在 ResNet-101 等其他模型上的攻击效果有所提高，表现出了更好的迁移性。然而，与当前的 SOTA 方法相比，仍然存在一定的差距。这表明，尽管本方法在某些模型上取得了一定进展，但在提升攻击性能和进一步优化可迁移性方面，仍然有提升空间。受上述实验结果的启发，随后实验中的 DiffuseMix 都将移除扩散模型部分。

5.2 BR_fractal and BR_fractal_multi

对于 BR_fractal 和 BR_fractal_multi 的性能分析，这里做了更详细的实验与对比。表2是分别在不同的代理模型：Inc-v3、Inc-v4、IncRes-v2、ResNet-101、ResNet-50 上，不同的攻击方法对不同的模型的攻击效果。

Model	Attack	Inc-v3	Inc-v4	IncRes-v2	Res-101	VGG-19	DN-121	Avg
Inc-v3	MI-FGSM	97.8	48.4	47.8	28.3	48.5	44.1	52.48
	diffusemix	98.6	67.7	63.4	38.7	60.5	61.8	**65.12**
	BR_fractal	99.0	66.0	65.5	39.1	62.6	62.6	**65.8**
	BR_fractal_multi	98.3	66.2	63.3	36.6	60.3	58.9	**63.93**
	decowa(sota)	95.4	82.2	79.4	51.0	74.2	75.4	76.27
Inc-v4	MI-FGSM	57.1	96.7	45.5	31.2	51.9	49.7	55.35
	diffusemix	69.6	98.3	63.6	42.4	66.7	64.2	**67.47**
	BR_fractal	73.7	98.1	65.6	44.0	68.4	67.2	**69.5**
	BR_fractal_multi	72.6	97.8	66.0	44.2	67.2	66.0	**68.97**
	decowa(sota)	87.7	97.6	83.6	60.8	85.0	84.9	83.27
IncRes-v2	MI-FGSM	62.7	57.6	98.3	35.8	55.4	50.4	60.03
	diffusemix	77.4	75.7	98.5	47.4	66.1	67.1	**72.03**
	BR_fractal	79.3	74.9	98.2	46.7	69.1	68.5	**72.78**
	BR_fractal_multi	78.9	76.1	**98.8**	49.8	69.4	68.9	**73.65**
	decowa(sota)	84.7	82.0	92.4	55.4	76.7	76.9	78.02
ResNet-101	MI-FGSM	42.3	36.9	29.8	87.8	50.4	51.7	49.82
	diffusemix	53.6	49.4	43.0	87.4	61.4	66.2	**60.17**
	BR_fractal	53.5	50.9	43.2	81.9	61.7	65.0	**59.37**
	BR_fractal_multi	56.7	53.3	49.3	83.5	63.6	68.0	**62.4**
	decowa(sota)	89.2	90.6	87.6	96.8	90.3	94.4	91.48
ResNet-50	MI-FGSM	32.3	26.8	21.2	30.0	41.1	40.5	31.98
	diffusemix	41.6	34.5	30.2	43.2	54.1	54.1	**42.95**
	BR_fractal	42.8	38.1	30.6	43.3	57.7	56.2	**44.78**
	BR_fractal_multi	45.6	41.1	35.7	46.7	57.9	58.0	**47.5**
	decowa(sota)	83.2	83.3	79.0	84.3	88.0	89.9	84.62

表 2. 在 CNNs 模型上的测试效果

Surrogate	Attack	ViT-B/16	LeViT-256	PiT-B	CaiT-S-24	ConViT-B	TNT-s	Visformer-S	Avg
Inc-v3	MI-FGSM	13.6	24.1	18.2	16.9	17.4	21.3	22.7	19.17
	diffusemix	16.5	34.6	22.6	24.0	21.7	28.5	32.5	**25.77**
	BR_fractal	17.9	33.6	23.1	25.1	24.0	29.5	32.0	**26.46**
	BR_fractal_multi	18.3	33.8	21.9	23.7	24.0	29.8	31.9	**26.2**
	decowa(sota)	27.2	48.5	36.1	36.0	34.6	43.2	45.7	38.76
Inc-v4	MI-FGSM	13.6	25.1	18.8	18.8	16.5	22.5	27.2	20.36
	diffusemix	19.5	36.6	27.1	26.5	23.7	30.9	38.2	**28.93**
	BR_fractal	20.7	37.5	25.2	27.9	26.1	32.8	36.8	**29.57**
	BR_fractal_multi	22.1	38.1	28.1	30.0	25.6	33.3	39.6	**30.97**
	decowa(sota)	33.8	57.8	44.7	45.2	41.8	51.3	56.6	47.31
IncRes-v2	MI-FGSM	16.5	30.2	24.0	22.8	22.4	27.1	30.9	24.84
	diffusemix	22.6	42.0	30.9	31.8	28.4	38.6	40.6	**33.56**
	BR_fractal	24.7	41.9	30.9	32.5	29.4	37.4	39.1	**33.7**
	BR_fractal_multi	26.3	44.5	31.4	34.2	30.1	37.6	41.5	**35.09**
	decowa(sota)	31.5	50.3	39.6	40.2	39.3	46.1	48.4	42.2
ResNet-101	MI-FGSM	14.7	25.6	23.1	19.2	17.5	22.3	30.2	21.8
	diffusemix	20.2	35.1	29.9	27.6	24.3	32.5	41.8	**30.2**
	BR_fractal	22.4	40.3	34.8	30.4	27.4	34.7	43.3	**33.33**
	BR_fractal_multi	25.9	42.6	36.5	32.3	30.5	35.7	45.0	**35.5**
	decowa(sota)	66.1	81.8	**75.5**	75.7	70.7	77.1	83.9	75.83
ResNet-50	MI-FGSM	8.6	17.8	14.6	11.1	10.1	16.1	18.9	13.89
	diffusemix	12.8	25.5	20.2	17.0	15.2	22.1	26.6	**19.91**
	BR_fractal	13.2	27.1	22.2	19.5	18.6	25.5	29.2	**22.19**
	BR_fractal_multi	14.8	29.4	25.4	19.8	18.6	28.1	32.6	**24.1**
	decowa(sota)	52.8	73.9	67.8	64.4	59.8	69.4	76.4	66.36

表 3. 在 Transformer 类模型上的测试效果

表3则是在不同的代理模型：Inc-v3、Inc-v4、IncRes-v2、ResNet-101、ResNet-50 上生成的对抗图像对 Transformer 类模型的攻击效果，这也是评价一个攻击方法的可迁移性的参考。

通过上述实验分析，可以得出以下结论：在卷积神经网络（CNNs）类模型中，仅引入分形图形相比于 MI-FGSM 方法，确实表现出一定的性能提升。这表明，分形图形作为一种独特的扰动形式，能够有效增强对抗样本的攻击能力，优于传统的攻击方法。此外，在进行分形融合之前，先对图像进行分块旋转处理，也能显著增强对抗图像的攻击性能。该处理方法通过引入更多的图像变换，使得生成的对抗样本更加多样化，从而提高了对抗效果。

在跨模型攻击的实验中，随着方法的逐步改进，对抗图像的攻击成功率也在不断提升。然而，尽管如此，与当前最先进的 SOTA 方法相比，仍然存在一定的差距。虽然改进后的方法在攻击性能上有了可观的提升，但仍有空间进一步优化，以更好地接近或超越现有的最佳攻击技术。因此，尽管当前方法展示了良好的潜力，但在实际应用中仍需进一步研究和完善，以实现更强大的跨模型攻击能力。

6 总结与展望

相比于完整的 DiffuseMix 方法，单独引入分形图形能够生成更有效、迁移性更好的对抗样本。具体而言，分形图形的引入能够增强对抗图像的鲁棒性和攻击成功率。这表明，分形图形作为一种有效的结构化扰动手段，能够更好地干扰模型的识别过程，从而生成更具挑战性的对抗样本。此外，实验还表明，在对图像进行处理之前，引入分块旋转同样能够提升对抗图像的攻击效果。这一策略通过增加图像的变换多样性，使得对抗样本更加难以预测和防御，从而进一步增强了攻击的有效性。综合来看，分形图形与分块旋转的结合，为生成更加强大且难以防范的对抗图像提供了有力的支持。

参考文献

[1] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun 2018.

[2] Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *Cornell University - arXiv,Cornell University - arXiv*, Dec 2014.

[3] Khawar Islam, Muhammad Zaigham Zaheer, Arif Mahmood, and Karthik Nandakumar. Diffusemix: Label-preserving data augmentation with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27621–27630, 2024.

[4] Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio. *Adversarial examples in the physical world*, page 99–112. Jul 2018.

[5] Kunyu Wang, Xuanran He, Wenxuan Wang, and Xiaosen Wang. Boosting adversarial transferability by block shuffle and rotation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24336–24346, 2024.