构建一种模态　　基于语义重构的 L-平衡区块链　　N

Zhijie Tan[†], *Xiang Yuan*[†], *Shengwei Meng*[‡], *Yakun Huang*[‡], *Weiping Li*[†], *Zhonghai Wu*[†], *Tong Mo*[†]

[†]北京大学软件与微电子学院，中国北京
[‡]北京邮电大学计算机科学学院，中国北京

arXiv:2303.02428v1 [cs.MM] 4 Mar 2023

摘要

当前的大型区块链系统（BTC 闪电网络、以太坊等）普遍面临持久化率低、存储成本高的问题。因此，用户倾向于在现有区块链系统上存储单模态（文本）信息。受语义通信算法的启发，本文提出了一种新算法，以解决区块链上文本与视觉模态严重失衡的问题。对原始视觉图像进行语义采样后，生成的语义文本将存储在链上，终端用户则可以使用**相**对**最**优**语**义**同**位**体**选择算法重建语义相似的图像。在 DIV2K 数据集上的实验表明，采用我们算法的区块链在可接受的语义信息损失下，对原始视觉数据的存储容量可提升 43 万倍，持久化率可提升 55 万倍。

索引词— 多模态、区块链、语义重建

1. 引言

区块链被视为元宇宙的基础设施，具备永久文件存储和去中心化验证机制等特性。然而，这些特性也带来了文件大小和持久化率方面的限制，尤其是在将文件上链时。永久文件存储意味着区块链的总磁盘容量会不断增长，而去中心化验证则意味着文件持久性必须由超过 50% 的验证节点进行验证。更糟的是，这些限制会随着节点数量的增加而加剧。以太坊 [1], 目前拥有近 400,000 个活跃地址。以2021年9月15日至2022年9月15日为例，其平均区块生成时间为 12 至 14 秒，平均区块大小约为 2 MB[1]。对于视觉模态数据而言，这样的持久化带宽（约 10MB/分钟）极其有限，会导致区块链严重拥塞。此外，大规模视觉数据持久化还会给每个去中心化节点带来难以承受的存储压力。

为了解决这些限制，通常会使用星际文件系统（IPFS） [2]或其他云存储方法（如 Google 云端硬盘等），为相应的视觉数据生成唯一的哈希值或超链接。然而，外部文件链接的永久存储并不意味着数据能够永久保存：区块链本身无法保证星际文件系统或其他云盘中文件的可访问性和存在性，这给区块链系统的设计带来了新的不确定性，因此并非完美的解决方案。除了使用外部文件链接，传统图像压缩方法（每分钟压缩 500 张图像，并将压缩比设得极高，达到 1:100）仍然难以实现理想的上链持久化率，因为通信网络由近 400,000 名用户组成。因此，尽管当前的移动 Web2 世界中，社交媒体以视觉数据为主，相较于文本数据，这些数据更易于所有人理解，但面向未来 Web3 世界的区块链仍然极度缺乏视觉数据。

受当前流行的语义通信算法启发[3, 4], ，本文的核心思想是以链下计算资源换取链上持久化率和文件大小。对于拥有去中心化节点的区块链，链下计算任务可以并行完成，而链上数据持久化则需要在串行的区块链区块中进行验证。随着区块链节点数量的增加，链下计算任务的效果会减弱。得益于图像转文本[5, 6], 、文本转图像 [7, 8]及其他多模态技术[9], 领域的突破，强大的预训练模型能够提取视觉图像的高层语义信息，而另一个模型则可以重建**语义同位体**（由相同语义信息生成的不同重建图像）。需要强调的是，与图像压缩方法相比，我们的语义重建算法更关注语义相似度，而非像素级视觉相似度。人类看到一幅图像时，会理解该图像的高层语义信息。如果图像略微平移或旋转，其语义信息不会受到太大影响，但这会彻底破坏 PSNR（峰值信噪比） [10] 或 SSIM（结构相似性） [11]等常见的视觉相似度评价指标。为应对这些挑战，

[1]数据来源：https://www.ycharts.com/。

诅咒，本文提出了一种算法，使区块链只需将相应的自然语义文本（T）上链。然而，在语义重建过程中，从低维文本数据重建高维视觉数据会不可避免地带来一定的随机性。随着语义提示确定性的提高，随机性将会降低。本文设计了**相**对**最**优**语**义**同**位**体**选择（**ROSIS**）算法，以降低与模型无关的随机性。

我们将主要贡献总结如下：提出了一种语义重建方法，以便将更多视觉数据上链。设计了 ROSIS算法，以减少语义重建过程中的语义损失。实验表明，该方法具有巨大潜力。

2. 方法

我们算法的原型框架如图 1所示。整个算法由语义采样器（图像转文本模型）、语义重建器（文本到图像模型）和 ROSIS算法组成。本节将重点介绍算法的定义与设计。

2.1. 问题表述

基于学习的图像重建过程可定义为：

$$I_o \xrightarrow{f(\theta_1)} X \xrightarrow{g(\theta_2)} I_r \tag{1}$$

在（1）中，I_o 是原始图像，f(θ_1)是一个参数为 θ_1 的采样映射，X 是中间表示，g(θ_2)是一个参数为 θ_2 的重建映射，I_r是重建图像。从信息瓶颈理论[12, 13]的角度来看，f(θ_1)旨在最小化互信息 $MI(I_o,\ X;\ \theta_1)$，而 g(θ_2)旨在最大化 $MI(X,\ I_l;\ \theta_2)$。I_l 是从 I_o下采样得到的有损压缩图像。因此，图像有损重建问题可定义为：

$$\arg\min_{\theta_1,\theta_2} MI(I_o, X;\theta_1) - \gamma MI(X, I_l;\theta_2) \tag{2}$$

γ 用于平衡 f 和 g。语义重建问题也可以类似地定义为：

$$\arg\min_{\theta_1,\theta_2} MI(Inf(I_o), T;\theta_1) - \gamma MI(T, Inf(I_l);\theta_2) \tag{3}$$

在式（3）中，Inf（I）是 I的语义表示，而 T是中间语义表示。我们的算法选择文本（自然语言）作为语义表示。

2.2. 解耦的语义采样器与重建器

人们还会进一步质疑，为什么在（3）中，X 需要是文本。从压缩比的角度来看，X 可以被压缩成更轻量的表示形式。然而，这种表示形式暗含一个假设：g 能够将每个$x \in X$ 映射为可接受的 $i_r \in I_r$，这会带来一项限制：f和 g 紧密耦合（通常在同一数据集上训练，或针对特定任务进行训练）。这项限制要求所有采样和重建模型都必须上链。如果不上链，不含语义信息的 x 就无法解码为可接受的图像，因而无法满足永久文件存储的要求。因此，用户需要从区块链下载对应的重建器模型，这会给区块链带来沉重的通信和存储压力。如果 x 是语义文本，那么 f 和 g 可以是任意一对性能强大的图像转文本和文本转图像预训练模型。这样就不会给区块链带来额外的通信或存储压力。因此，我们的方法采用解耦的语义采样器和重建器，这意味着这些模型和图像对双方都是透明的。因此，f 和 g 的θ_1 和 θ_2 几乎无法微调。事实上，在我们提出的算法中，θ_1 和 θ_2 是固定的。

2.3. 相对最优语义同位素选择（ROSIS）算法

由于 θ_1 是固定的，T 也变为固定值。固定的 T表明，当前问题与式（3）略有不同，因此可以用余弦距离替代互信息。式（3）中的语义重建问题可以转换为：

$$\arg\min_{\hat{T}} Cosine_dist(T, Inf(\hat{I}_r)) \tag{4}$$

$\hat{}\ \hat{}\ \hat{}I_r$ 是由 T重建出的图像。T 是原始 T 与随机种子的组合。由于第2.2节所述的透明性，重建器无法看到原始图像和采样器，因此在式（4）中，文本 T 是唯一可靠的信号；而在式（2）和式（3）中，I_o 更为可靠。对于一般任务而言，当压缩比达到10000:1或更高时，从 $t \in T$重建 $i_r \in I_r$ 必然存在不确定性。对于文本转图像模型，即使输入文本 T相同，不同的随机种子也会生成不同的图像。因此，我们提出了ROSIS算法：$\hat{}$ 生成不同的随机种子以获得不同的 T，并选择语义最接近的图像作为最终的 I_r。对于数据集和模型事先已知的任务，ROSIS算法可以作为一种与模型和数据无关的方法，基于微调后的模型获取额外的信息增益。例如，ROSIS算法可以改进人脸到身份 [14] 和身份到人脸 [15] 任务。如果数据集中存在一些相似的面部，身份到人脸模型可能会生成一些相似的面部。随后即可选出最佳面部。

3. 实验

3.1. 实验环境、数据集与指标

实验环境：实验部分使用的操作系统为 Ubuntu 20.04，中央处理器为英特尔至强（R）

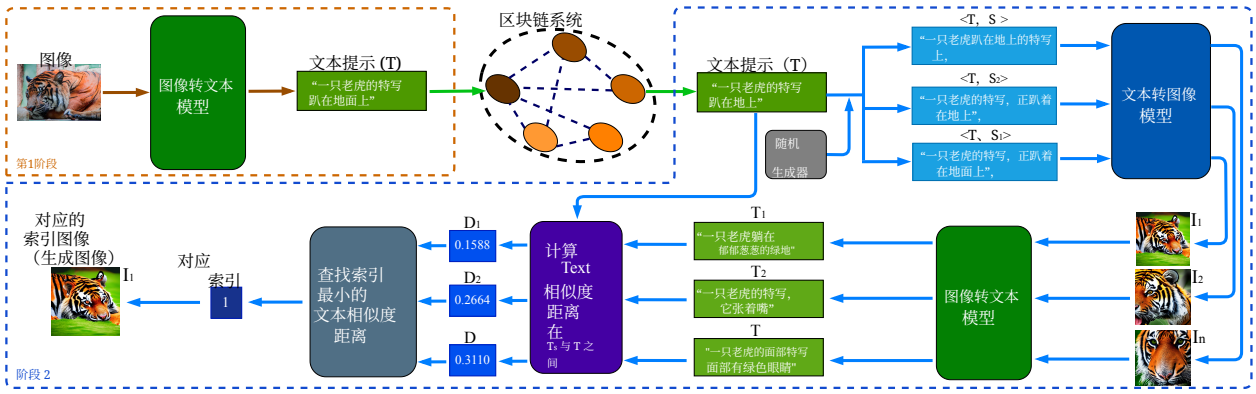


图 1。语义重建原型的整体架构。第1阶段：语义采样阶段。原始图像被转换为文本提示 T。阶段 2：语义重建阶段。从区块链中获取文本提示 T。我们通过文本到图像模型从 T 生成 n 个语义同位体（ I_1、……、 I_n ），然后分别将其转换为 T_1、……、 T_n 。计算 T_s 与 T 之间的语义距离，并将语义距离最小的对应索引所对应的 I_r 视为最终重建结果。

E5-2603 v4@1.7GHz，内存容量为 64GB。本文使用 PlatONE[2] 作为区块链系统进行模拟。燃气上限为 150m，难度为 1。区块链中有 10 个节点，ROSIS算法使用 4 个随机种子。我们使用 4 个 Nvidia V100 GPU 运行深度学习模型。图像首先转换为 Base64 格式的序列化数据，然后按每 100KB 分块。

数据集：本文使用的数据集为 DIV2K[16]。DIV2K 由 1000 张高分辨率（2K）图像组成，包含大量内容各异的图像（人物、老虎、建筑等），可有效验证不同对象的语义重建性能。

指标：**TPT（总持久化时间）**：所有文件传输到区块链并成功上链持久化后的总持久时间。开始时间是发送开始上传图像的命令时，结束时间是包含图像数据的最后一个区块的生成时间。**WT（等待时间）**：从发送开始上传的命令到包含图像数据的最后一个区块生成，一张图像所耗费的总时间。受区块链区块大小限制，一张图像通常需要拆分成不同区块，然后上传到区块链。等待时间反映了区块链的实时拥堵情况。**mWT（最短等待时间）/ MWT（最长等待时间）**：所有图像中最快/最慢完成上链所需的时间。它们反映了用户在最佳/最差情况下需要承受的等待时间。**F.S.（文件大小）**：所有文件占用的总存储空间。**Dis.（距离）**：Inf（I_ori）与 Inf（I_rec）之间的余弦距离。该值介于 0 和 2 之间。**TSS（语义采样时间）/ TRS（语义重建时间）**：对一张图像进行语义采样和语义重建所需的时间。以上所有指标均遵循越小越好的原则。

[2]http s://github.com/PlatONEnterprise/PlatONE- Go

3.2. 文件大小上的算法性能 P.T.

语义重建方法的原型如图 1 所示。该原型使用 $BLIP$ [5]作为图 1 中的图像转文本模型。 $BLIP$ 的视觉编码器使用 Transformer 块，其权重由 ViT-L/16[17]初始化。文本编码器则使用权重由 $BERT_{base}$ [18]初始化的 Transformer 块。对于给定的文本提示 T，使用 Stable Diffusion（图 1 中的文本转图像模型。）[7] 来还原对应的图像 I_o 。 Inf（I）由 all-$MiniLM$-$L6$-$v2$ [19]计算得出。

本实验结果如表1所示。我们将本框架与常用的 .jpg 和 .webp 方法、最先进的 RC [21] 和 SlimCAEs [20]方法（PSNR>25DB）进行了比较。表1中的 W.75/J.75 表示将 DIV2K 中的原始图像（.png）以默认质量 75 保存为 .webp/.jpg（PSNR>25DB）。表1中的 W.1/J.1 表示将 DIV2K 中的原始图像（.png）以质量 1 保存为 .webp/.jpg（PSNR>20DB）。

从表1中很容易看出，与图像压缩方法相比，本文提出的框架在持久化速度和文件大小方面具有无可比拟的优势。由于将 .txt 压缩为 .gz 文件不会造成信息损失，本文后文所述的方法均指压缩后的文本（.gz 文件）。即使与 .jpg/.webp 相比，我们的方法在文件大小比或总持久时间比方面仍可达到近 10000 倍的优势。mWT 和 MWT 为我们提供了两条线索。其一，对于分辨率相似的图像，$\frac{MWT}{mWT}$（表1中的 W.1）可以达到 6.82，这是因为这些图像的每像素比特数（bits per pixel）不同。相比之下，从图像中采样得到的文本提示文件大小相近。其二，我们的方法可以将海量视觉数据映射为足够小的文件，使其能够存储在一个区块链区块中，从而大幅减轻整个区块链系统的通信和同步压力。这解释了为何表1中与我们的方法对应的 mWT 和 MWT 指标均与总持久时间一致。

表1。语义重建与图像压缩方法的比较。J.75/J.1表示将压缩质量（1至95之间的整数）设为75/1。J./W.表示.jpg/.webp。我们的方法/我们的方法*分别表示原始文本提示（.txt）/压缩文本提示（.gz）。TPT（总持久化时间）、mWT/MWT（最短等待时间/最长等待时间）、F.S.（文件大小）、Dis.（原始图像与所选图像之间的语义距离）、TSS（每张图像的语义采样时间）/TSR（每张图像的语义重建时间）的定义见第3.1节。最佳结果以粗体显示。

	总持久时间（秒）	最短等待时间（毫秒）	最长等待时间（毫秒）	文件大小（MB）	语义距离$_{mean}$	距离$_{.min}$	距离$_{.max}$	每张图像的语义采样时间$_{mean}$(s)	每张图像的语义重建时间$_{mean}$(s)
Ori.	7209.580	3744.01	19804.14	3800.515	0.0	0.0	0.0	-	-
J.75	1020.774	533.61	2258.87	380.160	**0.253**	0.0548	0.763	-	-
J.1	169.245	89.48	373.03	51.007	0.375	0.115	0.894	-	-
W.75	780.385	457.10	1912.74	338.183	0.254	0.0548	0.700	-	-
W.1	178.157	63.99	436.63	52.191	0.287	0.0901	0.721	-	-
S[20]	720.143	396.13	1660.63	307.63	0.254	**0.0543**	**0.684**	-	-
R[21]	743.211	426.52	1820.86	324.12	0.254	0.0544	0.697	-	-
Ours	0.0440	0.0440	0.0440	0.0371	0.376	0.0876	0.712	12.51	10.11
我们的方法*	**0.0131**	**0.0131**	**0.0131**	**0.00879**	0.376	0.0876	0.712	12.51	10.11

表2。采用/不采用 ROSIS算法时的语义距离。

	Dis_{mean}	Dis_{min}	Dis_{max}
我们的方法*	0.376	0.0876	0.712
最差	0.652	0.205	0.967
随机	0.498	0.163	0.894

考虑两幅图像之间的语义距离，我们可以指出，本文提出的方法所达到的平均语义距离与质量为 1 的 .jpg 图像的 Dis_{mean} 相近，但持久化时间最多仅为其 1/10000，文件大小最多仅为其 1/6000。Dis_{max} 反映了最坏情况下的语义损失，这表明我们的方法优于 .jpg 系列中的所有方法。换言之，该框架在所有图像上的表现均可接受，即使在最差情况下也是如此（表1中的 0.712）。从 Dis_{min} 中可以看出，在某些情况下，该框架能够生成极其准确的语义表达（表1中的 0.0876）。

有人可能会进一步质疑，语义采样器/重构器处理单张图像所需的时间接近 10 秒，这似乎太长了。假设共享区块链上有 1000 个运行节点的主机，每个节点拥有一张来自 DIV2K 的不同图像，并且它们需要对这些图像进行持久化。采用我们的方法时，只会生成一个区块链区块，包括 TSS&TSR 在内的总耗时接近 30 秒。而采用 J.1方法时，总耗时将至少增加到 170 秒。此外，对于该区块链，存储在近 510 个区块链区块中的 J.1文件（约 51MB）必须在 1000 个分布式主机之间进行验证，而不是仅在我们的本地主机上验证；因此，在可能发生拥塞或受到其他因素影响的情况下，每台主机的实际总耗时和等待时间都会急剧增加。造成这一现象的主要原因是，链下计算可以并行执行，而链上持久化只能通过共享区块链上的串行区块进行验证。随着节点数量增加，TSS&TSR 的影响将会减小。

3.3. ROSIS算法的消融实验

表2列出了在最坏情况和随机情况下不使用ROSIS的实验结果。最坏情况是指选择与原始图像语义距离最远的图像作为语义同位体；随机情况则是指随机选择语义同位体。与最坏情况相比，ROSIS算法的平均距离、最小距离和最大距离分别提高了73.4%、134.0%和35.8%。与随机情况相比，ROSIS算法的平均距离、最小距离和最大距离分别提高了32.4%、86.0%和25.6%。虽然增加随机种子数量会带来更好的结果，但为节省计算资源，应谨慎考虑随机种子的数量。

4. 结论与未来工作

过去十年间，在 DIV2K 数据集上，每像素比特数的增幅不到 10%（相较于 .webp） [21, 20] ，而以太坊用户共享的网络总带宽为 10MB/min。因此，我们提出了一种语义重建算法，以在这种有限带宽下构建模态均衡区块链。ROSIS算法旨在减少从低维文本重建高维图像时由随机性导致的语义损失。实验表明，该方法实现了突破：在语义损失可接受的情况下，DIV2K 数据集上链所需的空间使用量不到原来的 1/430,000，持久时间不到原来的 1/550,000。未来，我们将探索可行的算法，以生成合适的 T，从而减少重建器选择造成的差异。此外，对于模型更新频率较低的专用任务，我们将投入更多时间来获得更好的表示 X。

5. 致谢

本研究得到中国辽宁省科学技术研究“$JieBangGuaShuai$”计划资助：基于联盟区块链的智能电子政务系统，项目编号为 2021JH1/10400010。

6. 参考文献

[1]Chris Dannen，以太坊与 *Solidity* 入门，第 1 卷，Springer，2017年。

[2] Erik Daniel 和 Florian Tschorsch，“IPFS 及其同类：新一代点对点数据网络的定性比较”，*IEEE* 通信调查与教程，第 24 卷，第 1 期，第 31–52 页，2022年。

[3] Yakun Huang、Yuanwei Zhu、Xiuquan Qiao、Zhijie Tan 和 Boyuan Bai，“AITransfer：面向实时点云视频流的渐进式人工智能驱动传输”，载于第 *29* 届 *ACM* 多媒体国际会议论文集，2021年，第 3989–3997 页。

[4] Huiqiang Xie、Zhijin Qin、Geoffrey Ye Li 和 Biing-Hwang Juang，“深度学习赋能的语义通信系统”，*IEEE* 信号处理汇刊，第 69 卷，第 2663–2675 页，2021年。[5] Junnan Li、Dongxu Li、Caiming Xiong 和 Steven Hoi，“BLIP：通过自举实现统一视觉-语言理解与生成的语言-图像预训练”，*arXiv* 预印本 *arXiv:2201.12086*，2022 年。[6] Matteo Stefanini、Marcella Cornia、Lorenzo Baraldi、Silvia Cascianelli、Giuseppe Fiameni 和 Rita Cucchiara，“从展示到讲述：基于深度学习的图像描述综述”，*IEEE* 模式分析与机器智能汇刊，2022年。[7] Robin Rombach、Andreas Blattmann、Dominik Lorenz、Patrick Esser 和 Bj¨orn Ommer，“利用潜在扩散模型进行高分辨率图像合成”，载于*IEEE/CVF* 计算机视觉与模式识别会议论文集，2022年，第 10684–10695 页。[8] Aditya Ramesh、Prafulla Dhariwal、Alex Nichol、Casey Chu 和 Mark Chen，“利用 CLIP 潜变量进行分层文本条件图像生成”，*arXiv* 预印本 *arXiv:2204.06125*，2022年。[9] Alec Radford、Jong Wook Kim、Chris Hallacy、Aditya Ramesh、Gabriel Goh、Sandhini Agarwal、Girish Sastry、Amanda Askell、Pamela Mishkin、Jack Clark 等，“从自然语言监督中学习可迁移视觉模型”，载于国际机器学习会议。PMLR，2021年，第 8748–8763 页。

[10] Hamid R Sheikh、Muhammad F Sabir 和 Alan C Bovik，《近期全参考图像质量评估算法的统计评估》，*IEEE* 图像处理汇刊，第15卷，第11期，第3440–3451页，2006年。

[11] Zhou Wang、Alan C Bovik、Hamid R Sheikh 和 Eero P Simoncelli，《图像质量评估：从误差可见性到结构相似性》，*IEEE* 图像处理汇刊，第13卷，第4期，第600–612页，2004年。[12] Naftali Tishby、Fernando C Pereira 和 William Bialek，《信息瓶颈方法》，*arXiv* 预印本 *physics/0004057*，2000年。[13] Ravid Shwartz-Ziv 和 Naftali Tishby，《通过信息揭开深度神经网络的黑箱》，*arXiv* 预印本 *arXiv:1703.00810*，2017年。[14] Wenyi Zhao、Rama Chellappa、P Jonathon Phillips 和 Azriel Rosenfeld，《人脸识别：文献综述》，*ACM* 计算调查（*CSUR*），第35卷，第4期，第399–458页，2003年。[15] Jianxin Sun、Qiyao Deng、Qi Li、Muyi Sun、Min Ren 和 Zhenan Sun，《Anyface：自由风格的文本到人脸合成与操控》，载于*IEEE/CVF* 计算机视觉与模式识别会议论文集，2022年，第18687–18696页。[16] Eirikur Agustsson 和 Radu Timofte，《NTIRE 2017年单幅图像超分辨率挑战赛：数据集与研究》，载于*IEEE* 计算机视觉与模式识别会议研讨会论文集，2017年，第126–135页。[17]Alexey Dosovitskiy、Lucas Beyer、Alexander Kolesnikov、Dirk Weissenborn、Xiaohua Zhai、Thomas Unterthiner、Mostafa Dehghani、Matthias Minderer、Georg Heigold、Sylvain Gelly 等，《一幅图像胜过16×16个词：用于大规模图像识别的Transformer》，*arXiv* 预印本 *arXiv:2010.11929*，2020年。[18]Jacob Devlin、Ming-Wei Chang、Kenton Lee 和 Kristina Toutanova，《BERT：用于语言理解的深度双向Transformer预训练》，*arXiv* 预印本 *arXiv:1810.04805*，2018年。[19] Wenhui Wang、Furu Wei、Li Dong、Hangbo Bao、Nan Yang 和 Ming Zhou，《MiniLM：面向任务无关压缩的预训练Transformer深度自注意力蒸馏》，神经信息处理系统进展，第33卷，第5776–5788页，2020年。[20] Fei Yang、Luis Herranz、Yongmei Cheng 和 Mikhail G Mozerov，《适用于实用神经图像压缩的可伸缩压缩自编码器》，载于*IEEE/CVF* 计算机视觉与模式识别会议论文集，2021年，第4998–5007页。[21] Fabian Mentzer、Luc Van Gool 和 Michael Tschannen，《利用有损压缩学习更好的无损压缩》，载于*IEEE/CVF* 计算机视觉与模式识别会议论文集，2020年，第6638–6647页。