专题综述

基于任意粒度语言描述的图像着色方法

北京邮电大学 常征 张沛瑄 李思
北京大学 翁书晨 施柏鑫

本文是北京邮电大学与北京大学团队合作研究的成果，发表在NeurIPS 2023并获得Spotlight的工作L-CAD[1]（Language-based Colorization with Any-level Descriptions）。论文研究的问题是基于语言描述的图像着色。该任务要求模型能够在用户友好的自然语言描述指导下为灰度图像添加结构合理且主观视觉效果满意的颜色。先前的方法假设用户为图像中的大多数物体提供全面的颜色描述，而忽略了只描述主要物体、甚至完全缺少描述的情况。论文提出了一个统一的模型，用于解决基于任意粒度语言描述的图像着色问题。该方法利用预训练的跨模态生成模型，凭借其强大的语言理解能力和丰富的颜色先验知识来处理描述粒度的不确定性。该方法进一步设计了语义对齐模块，以使着色结果保持和输入灰度图一致的局部空间结构并防止鬼影效果。通过提出的新型采样策略，所提出的模型能够在多样且复杂的场景中实现实例感知的着色效果。广泛的实验结果显示了论文提出方法的优势，包括有效处理任意粒度的描述，以及在基于文本条件和自动着色方面超越其他模型的表现。

一、研究背景

图像着色是一项具有挑战性的任务，其目的在于将灰度图像转换为合理且视觉上令人愉悦的彩色图像。基于语言的着色方法[2, 3, 4]利用自然语言描述作为指导，以产生更可控的彩色图像。这些方法能够满足用户的特定需求，使他们能够提供更具体和细腻的颜色偏好。由于自动着色[5,6,7]在为常见对象（例如，花的颜色）确定颜色时经常遇到歧义，基于语言的着色在生成高质量和可定制的彩色图像方面显示出了令人满意的结果，且具有对用户友好的交互方式。尽管基于语言描述的着色方法通过特征融合[4, 8]、解耦颜色-对象空间[2, 9]和聚合相似区块[3]改进了语言描述与彩色化结果之间的一致性，但它们隐式地假设用户为图像中的大多数物体提供全面的颜色描述。这种假设通常导致次优的性能，尤其是对于没有相应颜色描述的对象。此外，本文观察到用户通常只为他们感兴趣的物体分配颜色。

图 1　基于任意粒度语言描述的图像着色

为了用具有不同粒度的语言描述给图像着色，本文提出了一个统一的模型，它能够自适应地理解任何粒度的描述，并按如下方式着色：(1) 对于包括所有物体的“完全”粒度描述，模型会根据用户的要求精确着色 (图 1 第一行，对四杯鸡尾酒的详尽描述)；(2) 对于只关注感兴趣物体的“部分”粒度描述，模型会根据图像语义来着色未提及的物体（图 1 第二行，仅对罐子和花朵的选择性描述)；(3) 对于缺乏有意义的颜色信息的极简粒度描述，它会切换到自动上色模型（图 1 第三行，遗漏了对披萨和餐厅的描述)。

为了实现上述目标，论文提出了 L-CAD，用于完成

基于任意粒度语言描述的图像着色。鉴于任意粒度描述中提及的物体存在固有的歧义，该方法利用了预训练的跨模态生成模型（即，Stable Diffusion[10]），通过其强大的语言理解能力以及丰富颜色先验知识，来帮助完成基于任意粒度描述的着色。然而，由于生成模型并不是专门为上色而设计的，它在与输入灰度图的空间对齐方面面临挑战。为了解决这个问题，该方法设计了一个灰度图指导的隐变量解码模块，它使着色结果在像素空间中与灰度图像对齐，保留了灰度图中的局部空间结构。此外，该方法通过通道扩展的卷积运算，使隐空间中的特征图与语言描述对齐，以防止鬼影的出现。另外，为了处理具有不同粒度和复杂场境下的描述，论文提出了一种实例感知的采样策略，它在隐空间中粗略估计物体轮廓并将颜色特征分配给它们对应的区域。由于这些改进，该方法可以不受描述中提及物体数量的限制，能够有效处理任意粒度的描述。

二、L-CAD方法介绍

2.1 前言：扩散模型

扩散模型[11, 12]作为一种生成模型，在图像生成方面取得了显著的成就。在前向过程中，随机采样一个高斯噪声对原图进行加噪：

$$x_t = \sqrt{\alpha_t}x_0 + \sqrt{1-\alpha_t}\epsilon_t$$

在反向过程中，它训练一个神经网络ϵ_θ来预测噪声。通过加入额外的条件，扩散模型能够生成与给定条件一致的结果，以完成条件生成任务。条件扩散模型的训练过程中，仍然使用均方误差作为损失函数：

$$\mathcal{L}_{dm} = \mathbb{E}_{t,x_0,\epsilon\sim\mathcal{N}(0,1)}[\| \epsilon_t - \epsilon_\theta(x_t, t, y) \|^2]$$

为了缓解高分辨率图像生成时的资源消耗，Stable Diffusion[10] 引入了感知压缩模型，将像素空间的图像压缩到隐空间，这使得扩散过程可以在尺度更小的隐空间中进行。具体来说，它采用了一个压缩编码器$\mathcal{E}$来将给定的图像x映射到隐空间$z = \mathcal{E}(x)$，并使用一个压缩解码器$\mathcal{D}$来重建图像$\tilde{x} = \mathcal{D}(z)$。这样，噪声预测网络的训练目标变成了学习隐变量$z$的分布，而不是图像$x$。

该方法的目标是利用 Stable Diffusion 的强大语言理解能力和丰富的颜色先验知识，实现基于任何级别描述的语言驱动的着色。

2.2 灰度图指导的隐变量解码

尽管 Stable Diffusion 在文本生成图像任务中展现出卓越性能，但它缺乏在着色任务中保留输入灰度图像局部空间结构的能力。为了解决这一问题，该方法提出在像素空间中加入额外的灰度图编码器。如图 2 所示，彩色图像被压缩编码器映射成隐变量，而灰度编码器从灰度图像中提取多尺度特征，保留局部结构语义。之后，该方法将这些特征使用跳跃链接的方式直接加到压缩解码器的相应尺度中，引导从隐空间到像素空间的解码过程。灰度编码器采用与压缩编码器相同的架构，其压缩编码器和解码器的权重固定，以保留来自预训练模型的先验知识。

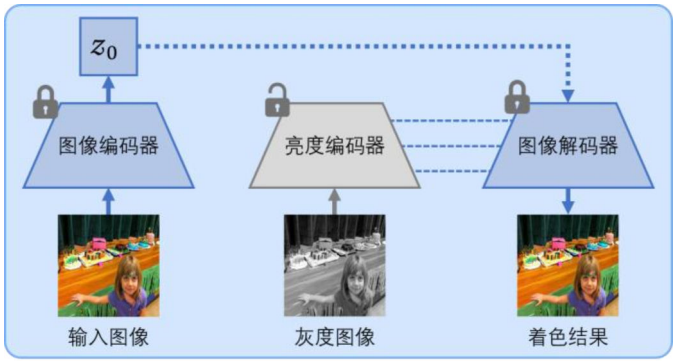


图 2　灰度图引导的隐变量解码

因为观察到在具有规则形状和锐利结构的区域中的错误像素显著损害了视觉感知，该方法估计了一个局部误差图$M^{art}_{h,w}$，并将其用于指示在解码结果中特定空间位置(h,w)遇到错误像素的概率：

$$M^{\text{art}}_{h,w} = \sum_{p\in\Omega_{(h,w)}}\left(\frac{\delta_p-\mu_p}{N_{win}}\right)^2,\ \mu_p = \sum_{p\in\Omega_{(h,w)}}\left(\frac{\delta_p-\mu_p}{N^2_{\text{win}}}\right)^2$$

其中，p表示以(h,w)为中心的局部窗口Ω的位置索引，N_{win}是局部窗口的大小。之后，该方法将局部误差图作为一个权重应用于图像重建损失中：

$$\mathcal{L}_{rec} = \| M^{\text{art}} \odot (x - \tilde{x}) \|_1$$

该方法在像素空间训练模型的总损失如下：

$$\mathcal{L}_{pix} = \mathcal{L}_{\text{rec}} + \alpha\mathcal{L}_{\text{dis}} + \beta\mathcal{L}_{\text{per}}$$

其中$\mathcal{L}_{dis}$为判别器损失，$\mathcal{L}_{per}$为感知损失。

2.3 隐空间语义对齐

在 Stable Diffusion[10]中，语言描述通过交叉注意力机制被注入隐空间的特征图中。然而在图像着色任务中，颜色特征被分配到预期区域之外会使生成的结果中产生鬼影。因此，需要在隐空间内保持语言描述和灰度图像之间的语义对齐。

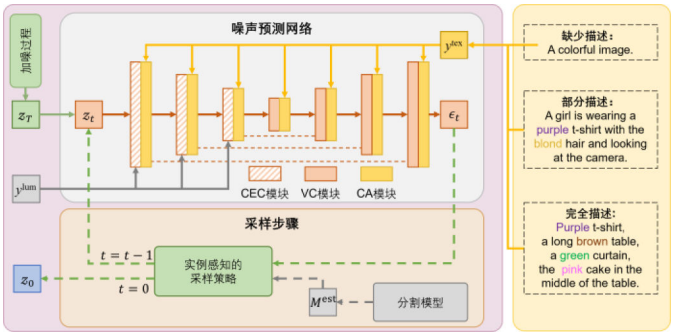


图 3　隐空间语义对齐

如图 3 所示，该方法用通道扩展卷积（CEC）模块替换了噪声预测网络中的普通卷积(VC)模块。这些 CEC 模块位于下采样网络内，它们接收来自像素空间的灰度编码器的亮度特征y^{lum}作为额外的引导。通过利用扩展的通道，CEC 模块可以有效地捕获隐空间中灰度的局部结构语义。VC 模块和 CEC 模块的区别如图 4 所示：

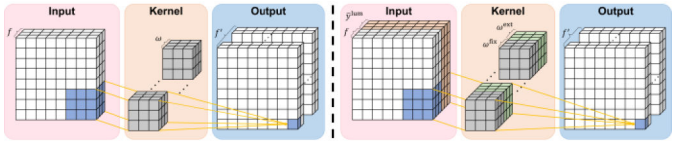


图 4　VC 模块和 CEC 模块

从数学角度来看，CEC 模块等同于使用卷积操作来提取特征，并将它们的输出加到下采样网络中。为了保留预训练生成模型的强大语言理解能力和丰富的颜色先验知识，该方法在训练过程中保持原始通道的参数w^{fix}不变。此外，扩展通道的权重w^{ext}被初始化为零，以确保在训练之前的模型保持与预训练的生成模型的功能等效性。该方法在下采样和上采样模块之间使用跳跃连接，以引导上采样过程中的亮度特征，训练损失为：

$$\mathcal{L}_{lat} = \mathbb{E}_{t,z_0,\epsilon\sim\mathcal{N}(0,1)}[\| \epsilon_t - \varepsilon_\theta(z_t, t, y^{\text{tex}}, y^{\text{lum}}) \|^2]$$

2.4 实例感知的采样策略

为确保语言描述中的颜色能准确地体现在图像里指定的物体上，本文提出了一种实例感知的采样策略。该策略采用了一个外部的分割模型（针对文本描述中提到的物体或区域进行分割，例如 SAM[13]）来估计描述中提到的物体的轮廓M^{est}。考虑到在第l个 CA 模块中的注意力图M_l^{att}控制每个颜色词的着色区域，该方法使用 Sigmoid 函数对注意力图归一化，并通过迭代优化来使它与缩放后的物体的轮廓$\hat{M}^{\text{est}}$对齐。之后，该方法应用 DDIM[12]完成去噪过程，如算法 1 所示。

算法 1：实例感知的采样策略

输入：粗略估计的物体轮廓 M^{est}
输出：着色过的隐变量 z_0
For $t = T \cdots 1$ do:
　　$\_, M_*^{\text{att}} = \epsilon_\theta(z_t, t, y^{\text{lum}}, y^{\text{tex}})$
　　For $l = 1 \cdots L$ do:
　　　　$\hat{M}_l^{\text{est}} \leftarrow \text{Downsampling}(M^{\text{est}}, l)$
　　　　$\mathcal{M} \leftarrow \text{Sigmoid}(M_l^{\text{att}})$
　　　　$\hat{M}_l^{\text{att}} \leftarrow M_l^{\text{att}} - \lambda\nabla_{\mathcal{M}}\mathcal{L}_{\text{BCE}}(\mathcal{M}, \hat{M}_l^{\text{est}})$
　　End
　　$\hat{\epsilon}_t, \_ = \epsilon_\theta(z_t, t, y^{\text{lum}}, y^{\text{tex}})\{M_*^{\text{att}} \leftarrow \hat{M}_*^{\text{att}}\}$
　　$z_{t-1} = \text{DDIM}(z_t, \hat{\epsilon}_t, t)$
End

三、实验结果

该方法在基于语言的图像着色数据集上进行实验：（1）扩展的COCO-Stuff数据集[9]，该数据集是在COCO-Stuff数据集[14]的基础上构建的，包括59K训练图像和2.4K评估图像；（2）多实例数据集[3]，它提供了在单一图像内有多个不同实例的样本，包括65K训练图像和7K评估图像。对于这两个数据集，每个图像都有相应的语言描述。

3.1 与基于语言的着色模型对比

L-CAD 与基于语言的图像着色方法进行了比较，例如 ML2018 [4]、L-CoDe [9]、L-CoDer [2]以及 L-CoIns [6]。先前基于语言的图像着色方法假设用户提供全面的颜色描述，因而在处理部分描述的样本时性能降低。相比之下，该方法利用了 Stable Diffusion[10]的先验知识以及新提出的实例感知采样策略，即使在描述程度不的情况下也能展现出生动的着色结果，如图 5 所示。

图 5　提出方法与基于语言的着色模型对比

3.2 与自动着色模型对比

L-CAD 还与全自动的图像着色方法进行了比较，例如 CIC[7]，InstColor[5]，ChromaGAN[15]，BigColor[16]，DISCO[17]以及 CT² [6]。如果用户有特殊要求，自动着色方法无法按照用户的需求改变图像中物体的颜色，只能按照数据集中物体的颜色分布着色，如图 6 所示。

图 6　提出方法与自动着色模型对比

3.3 与基于扩散模型的图像编辑方法对比

一些图像编辑方法例如 ControlNet[18]、Pix2PixZero[19]和 SDEdit[20]也能够基于预训练的扩散模型，使用描述来编辑图像。然而，这些方法并不是专门为着色任务设计的，这导致它们在保持局部空间结构、利用颜色先验以及学习物体与颜色词之间的对应关系

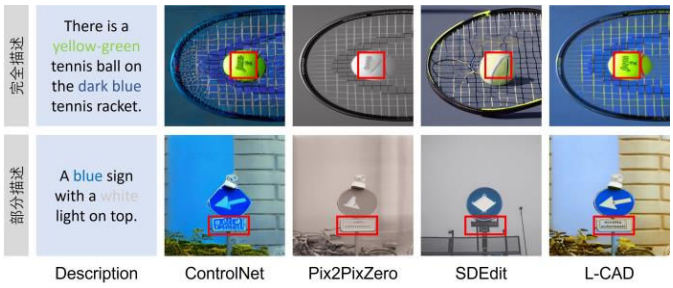


图 7　提出方法与基于扩散模型的图像编辑方法对比

方面存在挑战。这些限制使得它们难以完成任意粒度描述指导的实例感知着色。图 7 展示了它们的不适用性。

3.4 消融实验

为了研究该方法提出的模块和采样策略的影响，论文创建了三个基准模型进行消融实验。我们在图 8 中展示了消融实验的定性结果。

灰度引导的图像压缩（LIC）。该实验禁用了像素空间中的亮度编码器，导致模型缺少多尺度灰度特征进行引导，进一步使得模型无法正确保持局部空间结构。

语义对齐的隐变量（SLR）。在移除了指导隐空间语义对齐的灰度特征，并将扩展通道的卷积块替换为普通的卷积模块以后，着色图像中出现了明显的鬼影。

实例感知采样策略（ISS）。用标准的 DDIM[12]替换了实例感知采样策略以后，模型正确地根据文本描述的为物体分配颜色的性能显著降低了。

图 8　消融实验的定性结果展示

四、总结

该论文提出了一种基于任意粒度语言描述的图像着色模型。利用预训练模型的先验知识，其设计了新颖

的模块，以保持局部空间结构并防止出现鬼影，并进一步提出了实例感知采样策略，在复杂场景下实现实例感知的图像着色。与基于语言的着色方法进行比较的结果展示了该模型可以处理任何粒度的语言描述，这些粒度包括完全描述和部分描述，以及缺少描述。定性和定量结果都证明了该模型的优越性能。

五、前期相关工作

在L-CAD发表之前，北京邮电大学与北京大学的团队围绕语言引导的图像着色算法已经发表一系列工作，包括：L-CoDe[9]（Language-based Colorization using Color-object Decoupled Conditions）提出了颜色与对象解耦条件的着色网络，解决了颜色与对象耦合与不匹配的问题；L-CoDer[3]（Language-based colorization with Color-object Decoupling transformer）首次将transformer引入了到基于语言的图像着色任务中，统一了颜色描述与图像实例的特征表示；L-CoIns[4]（Language-based Colorization with Instance awareness）进一步设计了自动聚合的分组机制，强化了着色模型里实例感知的能力。

5.1 L-CoDe

基于语言描述的图像着色方法普遍面临颜色与物体的耦合和不匹配的问题：颜色与物体的耦合会导致难以将香蕉着色为红色，因为模型未曾见过红色的香蕉；而颜色与物体不匹配的问题会导致未被描述的物体会被错误地着色为其他物体的颜色。这些物体导致现有的模型难以准确地将句子中描述颜色的形容词映射为图像中指定物体的颜色。

为解决以上问题，L-CoDe 提出了颜色与对象解耦条件的着色网络。为了解决颜色与对象耦合的问题，其引入了颜色物体对应矩阵的预测器和新颖的注意力转移模块，这确保了颜色描述的特征被注入到对应物体的图像区域上；同时，为了解决颜色与物体不匹配问题，其采用了一个软门控的注入模块来有效地过滤掉不对应的颜色引导。进一步，它还提出了一个包含标注的“颜色与物体对”新数据集，以提供监督信号解决耦合问题。

5.2 L-CoDer

尽管 L-CoDe 在基于语言的着色任务中取得了一定的进展，但其性能受到三个关键问题的限制。（1）图像和语言属于不同的模态，因而提取的特征之间存在较大差距，这增加了理解语言描述的难度。（2）图像特征表示的语义在网络中是由局部到全局逐层演化的，单一的语言特征难以与多尺度的图像特征匹配，这降低了颜色表示的准确性。（3）基于卷积神经网络设计的着色模型往往是基于局部感知的，因此在着色局部亮度变化强烈的区域时容易出现伪影。

L-CoDer 首次将 transformer 引入了到基于语言的图像着色任务中，同时保持语言条件中颜色物体的解耦，以解决上述问题。L-CoDer 使用的 transformer 结构将图像与语言的特征统一表示为 tokens，并进一步支持语言描述中的颜色条件根据图像特征的变化的自适应调整。此外，由于 transformer 具备全局感受野，L-CoDer 对局部强烈的亮度变化具有鲁棒性。

5.3 L-CoIns

尽管 L-CoDe 和 L-CoDer 引入额外的标注来防止颜色-对象耦合和不匹配问题，但它们仍然难以处理包括多个不同实例的场景。例如，一张多个人物的合照，且每个人的衣着被语言描述指定为不同的颜色。

CoIns引入了多个可学习的分组向量，并提出了一个自动聚合具有相似颜色图像块的分组机制，促使分组向量能够自适应地表示图像中的多个实例，并最终能够在没有任何外部先验引导的情况下实现了实例感知的图像着色。此外，L-CoIns还提出了的亮度增强和颜色对比损失，打破了亮度和颜色词之间的统计相关性，驱动模型合成与语言描述更加一致的颜色。该工作进一步收集了一个多实例数据集，为同一图像中的多个实例提供了详细的语言描述。

责任编委 王金甲

参考文献

[1] Z. Chang, S. Weng, P. Zhang, Y. Li, S. Li, and B. Shi. L-CAD: Language-based Colorization with Any-level Descriptions using Diffusion Priors. In NeuIPS 2023.

[2] Z. Chang, S. Weng, Y. Li, S. Li, and B. Shi. L-CoDer: Language-based colorization with color-object decoupling transformer. In ECCV, 2022.

[3] Z. Chang, S. Weng, P. Zhang, Y. Li, S. Li, and B. Shi. L-CoIns: Language-based colorization with instance awareness. In CVPR, 2023.

[4] V. Manjunatha, M. Iyyer, J. Boyd-Graber, and L. Davis. Learning to color from language. In NAACL, 2018.

[5] J.-W. Su, H.-K. Chu, and J.-B. Huang. Instance-aware image colorization. In CVPR, 2020.

[6] S. Weng, J. Sun, Y. Li, S. Li, and B. Shi. CT2 : Colorization transformer via color tokens. In ECCV, 2022.

[7] J. Chen, Y. Shen, J. Gao, J. Liu, and X. Liu. Language-based image editing with recurrent attentive models. In CVPR, 2018.

[8] Zhigang Li, Gu Wang, Xiangyang Ji. CDPN: Coordinates-Based Disentangled Pose Network for Real-Time RGB-Based 6-DoF Object Pose Estimation. In ICCV, 2019.

[9] S. Weng, H. Wu, Z. C. Chang, J. Tang, S. Li, and B. Shi. L-CoDe: Language-based colorization using color-object decoupled conditions. In AAAI, 2022.

[10] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In CVPR, 2022.

[11] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. In NeurIPS, 2020.

[12] J. Song, C. Meng, and S. Ermon. Denoising diffusion implicit models. In ICLR, 2021.

[13] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, et al. Segment anything. arXiv preprint arXiv:2304.02643, 2023.

[14] H. Caesar, J. Uijlings, and V. Ferrari. COCO-Stuff: Thing and stuff classes in context. In CVPR, 2018.

[15] P. Vitoria, L. Raad, and C. Ballester. ChromaGAN: Adversarial picture colorization with semantic class distribution. In WACV, 2020.

[16] G. Kim, K. Kang, S. Kim, H. Lee, S. Kim, J. Kim, S.-H. Baek, and S. Cho. BigColor: Colorization using a generative color prior for natural images. In ECCV, 2022.

[17] M. Xia, W. Hu, T.-T. Wong, and J. Wang. Disentangled image colorization via global anchors. TOG, 2022

[18] L. Zhang and M. Agrawala. Adding conditional control to text-to-image diffusion models, 2023.

[19] G. Parmar, K. K. Singh, R. Zhang, Y. Li, J. Lu, and J.-Y. Zhu. Zero-shot image-to-image translation. arXiv preprint arXiv:2302.03027, 2023.

[20] C. Meng, Y. He, Y. Song, J. Song, J. Wu, J.-Y. Zhu, and S. Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations. In ICLR, 2021.

常征

北京邮电大学人工智能学院 2021 级硕士研究生，导师为李思副教授，主要研究方向为图像合成。
Email:zhengchang98@bupt.edu.cn

张沛瑄

北京邮电大学人工智能学院 2023 级硕士研究生，导师为李思副教授，主要研究方向为图像合成。
Email: pxzhang@bupt.edu.cn

李思

北京邮电大学人工智能学院副教授，博士生导师。北京邮电大学博士，美国布兰迪斯大学博士后，曾在新加坡国立大学、日本国立信息学研究所从事科研工作。研究方向为多模态智能信息处理，在 TPAMI、CVPR、ICCV、NeurIPS 以及 ACL 等期刊和会议上发表论文多篇。主持和参与国家自然科学基金、科技创新 2030—“新一代人工智能”重大项目课题、北京市自然科学基金等多个项目。
Email:lisi@bupt.edu.cn

翁书晨

北京大学计算机学院 2019 级博士生，导师为施柏鑫研究员，主要研究方向为跨模态图像编辑。
Email:shuchenweng@pku.edu.cn

施柏鑫

北京大学计算机学院多媒体信息处理全国重点实验室、视频与视觉技术国家工程研究中心研究员、博士生导师（“博雅青年学者”）；北京智源人工智能研究院青年科学家。2013 年博士毕业于日本东京大学，曾先后在麻省理工学院媒体实验室、新加坡科技设计大学、南洋理工大学、日本国立产业技术综合研究所从事研究工作。研究方向为计算摄像学与计算机视觉，发表论文 180 余篇（包括 TPAMI 论文 23 篇，计算机视觉三大顶级会议论文 69 篇）。论文获评国际计算摄像会议（ICCP）2015 年 Best Paper - Runner Up、入选 IJCV 专刊 Best Papers from ICCV 2015，2021 年获得日本大川研究助成奖。主持科技创新 2030—“新一代人工智能”重大项目、国家自然科学基金重点、国家级青年人才等多个项目。担任国际顶级期刊 TPAMI、IJCV 编委，顶级会议 CVPR、ICCV 领域主席。IEEE、CCF、CSIG 高级会员，APSIPA 杰出讲者。
Email:shiboxin@pku.edu.cn

热点追踪

动态场景新视角合成

西北工业大学 郭相 戴玉超

一、引言

新视角合成（NVS）是计算机视觉和图形学中一个长期且具有挑战性的问题，在虚拟现实、增强现实、数据增强、图像编辑等领域有很多应用。最近，可微神经渲染技术[1][2][3]特别是神经辐射场（NeRF）[1]的引入在短时间内极大地推动了这一领域的快速发展，并引起了广泛关注。NeRF[1]通过多层感知器（MLP）表示三维世界，将输入的三维坐标和视角方向映射到对应的不透明度和颜色，从而通过渲染生成逼真的图像。

最初的 NeRF 只能对静态场景建模，一系列工作将基于 NeRF 的框架从静态场景扩展到了动态场景[4][5][6][7][8][9][10][11]。其中一个很有前景的思路是使用规范空间表示法[8][9][12]。这种表示法将一个时刻设置为规范时刻，并用神经辐射场对规范时刻的静态场景进行建模。为了渲染其他时刻的图像，需要使用一个变形场来估计三维点从当前时刻移动到规范时刻的后向流。虽然基于后向流的规范表示法很容易实现，但后向流场是非光滑的。如图 1(b)所示，对于时间轴上的一个固定三维位置 p，会有不同类型的物体点覆盖位置 p，这就需要不连续的后向流将它们映射回规范空间（图 1(d)）。因此，常用的平滑运动模型（如 MLP）无法很好地拟合后向流。此外，由于运动模型的失效，规范空间的静态场景模型也会发生扭曲变形。

为了解决后向流的问题，提出使用前向流（Forward Flow）作为变形模型。通过使用前向变形流，将整个规范空间的辐射场从规范时刻翘曲到其他时刻，并在相应的时刻进行渲染。这样，对于时间轴上的同一个位置，变形模型估计的前向流将是平滑和连续的（图 1(c)和(e)）。需要注意的是，SNARF[13]使用了基于皮肤模型的前向翘曲，但它是为动态人体建模而设计的，不能用于一般场景。我们的目标是实现一般场景的动态建模，这意味着我们必须对整个空间进行翘曲。

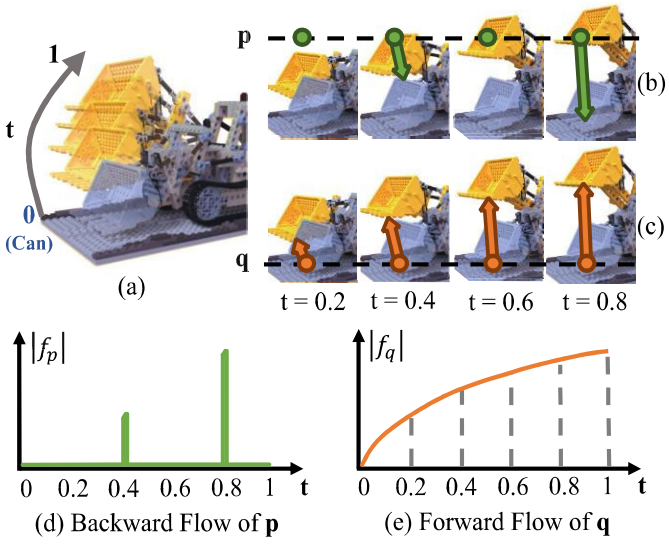


图 1 后向流 vs 前向流。本图展示了后向流和前向流变化的示例。(a) 动态场景示例。(b) 随着铲斗升起，不同类型的点覆盖了绿色的位置 p，这就需要非常不同的后向流来将这个点映射回规范空间。(d) 位置 p 后向流的模长随时间的变化不平滑。(c) 位置 q 的前向流将一个特定物体点从规范空间映射到其他时间，它是平滑和连续的。(e) 显示了位置 p 的前向流的模长变化是平滑的。

然而，将前向翘曲引入基于规范空间的动态 NeRF 方法仍有三个主要问题亟待解决。首先，现有方法中的传统辐射场无法进行显式的翘曲，这是因为辐射场是由 MLP 参数化的连续函数表示的。为了解决这个问题，我

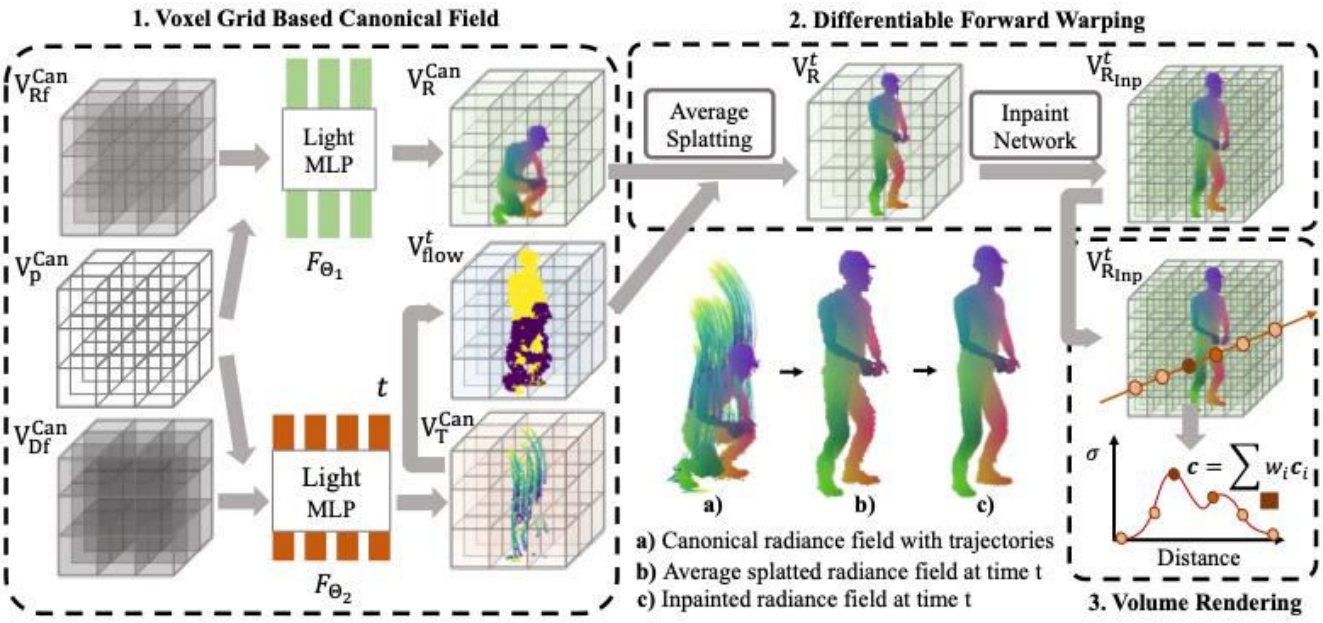


图 2　方法框架。a) 我们用基于体素网格的辐射场来表示规范空间静态场景的不透明度和颜色，基于体素网格的轨迹场来表示变形；b) 我们提出使用平均拼接的前向流对规范空间的静态辐射场进行前向翘曲；c) 使用填充网络对翘曲后的辐射场进行填充。 具体来说： 1. 基于体素网格的规范空间中包含两个模型，包括了一个基于体素网格的静态辐射场和一个基于体素网格的轨迹场。静态辐射场V_R^{Can}是由一个 MLP 估计的，它将辐射特征 V_{Rf}^{Can} 和相应的三维坐标 V_p^{Can}作为输入，估计辐射场三维点的颜色和不透明度。轨迹场 V_T^{Can} 由另一个 MLP 估计，它将形变特征 V_{Df}^{Can}和坐标V_p^{Can}作为输入，估计每个点的轨迹，然后就可以得到从规范时刻到时间 t 的前向流V_{flow}^t。 2. 可微分的前向翘曲操作首先将规范空间的辐射场V_R^{Can}根据V_{flow}^t翘曲，得到时间 t 的辐射场 V_R^t。 然后通过填充网络得到填充后的辐射场$V_{R_{Inp}}^t$； 3. 基于辐射场$V_{R_{Inp}}^t$，使用体积渲染（Volume Rendering）来渲染 t 时刻的图像。

们提出使用体素网格来表示规范空间辐射场，因为它是离散的并且有限的。基于体素的方法[14][15][16]已经证明了这种表示方法的有效性。另外两个问题是前向翘曲操作的固有特性带来的多对一和一对多映射问题。针对这两个问题，我们提出了一种由平均拼接操作和填充网络组成的可微分前向翘曲方法，分别解决了多对一和一对多的问题。

二、动态场景新视角合成相关工作

将 NeRF 从静态场景扩展到具有非刚性可变形物体的动态场景是一个热门的研究领域。一种可行的方法是构建 4D 时空表示。例如，Yoon 等人[17]结合了单视角深度和多视角立体深度来渲染具有三维变形的新视角图像。Gao 等人[5]使用时间不变模型（静态）和时间变化模型（动态）来表示场景，并通过场景流估计对动态模型进行正则化。NeRFlow[4]从一组 RGB 图像中学习动态场景的 4D 时空表示。Xian 等人[11]建立了一个 4D 时空辐射场，将时空位置映射到点的颜色和不透明度。同样，NSFF[6]将动态场景建模为关于外观、几何结构和三维场景运动的连续函数。DCT-NeRF[10]使用离散余弦变换(DCT)捕捉动态运动，即学习空间中每个点随时间变化的平滑稳定轨迹。

另一方面，D-NeRF[8]、Nerfies[7]、HyperNeRF[18]和 NR-NeRF[9]使用静态规范辐射场捕捉场景的几何和外观，然后在每个时刻学习规范空间的变形/位移场。具体来说，要渲染不同时刻的图像，就需要使用变形场来估算后向场景流，将三维点从当前时刻移动回规范辐射场。然而，对于时间轴上的同一 3D 位置，后向流场并不能保证平滑和连续。因此，规范辐射场通常会出现扭曲，类似于移动物体的平均形状。本文将重点解决后向流的不平滑问题。

除了这两个主要方向之外，目前还有一种趋势是加速基于体素网格表示的动态 NeRF 的训练。TiNeuVox[19]

使用轻量化的 MLP 对变形进行建模，并使用多分辨率为辐射网络获取特征，从而估算不透明度和颜色。V4D[20]使用三维特征体素对四维辐射场进行建模，并将额外的时间维度串联起来，同时提出了用于像素级的查找表。然而 V4D 主要侧重于提高图像质量，与 TiNeuVox 相比，其训练速度并不明显。DeVRF[21]也是以体素网格表示法为基础，它提出使用多视角数据来克服单目设置带来的奇异性问题。与其他使用单目设置的方法相比，多视角数据简化了运动和几何的学习。

三、动态场景新视角合成方法

我们使用一个基于体素网格的静态辐射场和一个基于体素网格的轨迹场来对规范空间中的场景进行建模。为了合成动态图像，我们提出将规范空间中的辐射场向前翘曲到相应的时刻，并根据翘曲得到的辐射场使用体渲染技术渲染图像。图 2 显示了提出方法的框架。该方法有三个主要组成部分：基于体素网格的规范空间模型、可微的前向翘曲方法和体渲染方法。

3.1 基于体素网格的规范空间模型

基于体素网格的规范空间中包含两个模型，包括了一个基于体素网格的静态辐射场和一个基于体素网格的轨迹场。静态辐射场包含了一个可学习的辐射特征V_{Rf}^{Can}和一个轻量化 MLP 网络F_{θ_1}，静态辐射场定义如下：

$$V_R^{Can} = F_{\theta_1}(V_{Rf}^{Can}, V_p^{Can}) \tag{1}$$

其中V_p^{Can}是体素网格在世界坐标中的坐标。

我们提出使用离散余弦变换（DCT）[10]来表示三维点的运动轨迹，以确保运动的平滑性。与静态辐射场类似，我们也使用了可学习的变形特征V_{Df}^{Can}和轻量化 MLP 网络F_{θ_2}，来估计轨迹场，其定义为：

$$V_T^{Can} = F_{\theta_2}(V_{Df}^{Can}, V_p^{Can}) \tag{2}$$

其中V_T^{Can}包含了每个体素的 DCT 轨迹系数。在给定时刻 t 的情况下，我们可以通过以下公式得到这些体素从规范空间到时刻 t 的前向流：

$$V_{flow}^t = f_{DCT^{-1}}(V_T^{Can}, t) - f_{DCT^{-1}}(V_T^{Can}, Can) \tag{3}$$

其中$f_{DCT^{-1}}$是 DCT 变换的逆变换，详细公式可以参考 DCT-NeRF[10]。

3.2 可微的前向翘曲方法

为了根据计算得到的前向流，将规范空间的静态辐射场从规范空间的时刻前向翘曲到对应的时刻，我们提出了一个可微的前向翘曲方法。该方法包含了两步：平均拼接和填充网络。对于平均拼接，受 Softmax-Splatting[22]的启发，我们提出将源网格中可能存在的多个值使用平均拼接融合到对应的目标网格。具体来说，我们提出一种简单而有效的方法：用三线性核计算这些值的"平均值"。形式上，假设我们需要通过流$f_{S\to T}$将源网格V^S翘曲到目标网格V^T，而 $\boldsymbol{p}$、$\boldsymbol{q}$ 是体素网格的索引。我们将$V^T = F_{warp}(V^S, f_{S\to T})$定义如下：

$$V^T[\boldsymbol{p}] = \frac{\sum_{\forall \boldsymbol{q}\in V^S} b[\boldsymbol{u}]\cdot V^S[\boldsymbol{q}]}{\sum_{\forall \boldsymbol{q}\in V^S} b[\boldsymbol{u}]} \tag{4}$$

$$b[\boldsymbol{u}] = \prod \max(0, 1 - |\boldsymbol{u}_i|), i \in \{x, y, z\} \tag{5}$$

$$\boldsymbol{u} = (\boldsymbol{q} + f_{S\to T}[\boldsymbol{q}]) - \boldsymbol{p} \tag{6}$$

其中x, y, z代表了体素网格的三个坐标轴。

通过以上的翘曲公式，我们可以将静态辐射场V_R^{Can}翘曲时刻 t：

$$V_R^t = F_{warp}(V_R^{Can}, V_{flow}^t) \tag{7}$$

由于一对多的问题存在，通过平均拼接后的辐射场存在空洞。为了解决这个问题，我们提出一个填充网络F_{θ_3}来填充V_R^t可能存在的空洞：

$$V_{R_{Inp}}^t = F_{\theta_3}(V_R^t) \tag{8}$$

填充网络是一个基于三维卷积的 UNet 网络结构，它可以通过学习，利用领域信息，填补存在的空洞。

3.3 体渲染方法

在得到时间 t 的辐射场$V_{R_{Inp}}^t$后，就可以使用体渲染技术[23]渲染图像射线的像素颜色。给定一条射线 $r(w) = \boldsymbol{o} + w\boldsymbol{d}$ 从像机中心 $\boldsymbol{o}$ 出发，以 $\boldsymbol{d}$ 为视角方向穿过图像平面上给定的像素，我们通过体渲染方法渲染出对应像素的颜色$C_{Inp}(r) = F_{render}(V_{R_{Inp}}^t, r)$。为此，我们获取射线$r$和体素网格相交的所有三维点$p$。然后，应用三线插值法获得每个三维点的密度σ和颜色$\mathbf{c}$，

$$(\sigma, \mathbf{c}) = F_{inter}(V_{R_{Inp}}^t, p) \tag{9}$$

最后，像素的颜色可以通过如下公式渲染得到：

$$C(r) = \sum_{k=1}^{K} T(w_k)\alpha(\sigma(w_k)\delta_k)\boldsymbol{c}(w_k) \quad (10)$$

$$T(w_k) = \exp(-\sum_{j=1}^{k-1}\sigma(w_j)\delta_j) \quad (11)$$

$$\alpha(\sigma(w_k)\delta_k) = 1 - \exp(-\sigma(w_k)\delta_k) \quad (12)$$

其中δ_k是射线上，相邻的两个采样点之间的距离。

三、动态场景新视角合成实验结果

在 D-NeRF 数据集上进行了相关测试，量化结果如表 1 。可以看到提出的方法在性能上与其他方法比较，有明显并且一致性的优势。

表 1　在 D-NeRF 数据集上的测试结果

方法	PSNR↑	SSIM↓	LPIPS↓
T-NeRF[8]	29.51	0.95	0.08
TiNeuVox-S[19]	30.75	0.96	0.07
TiNeuVox-B[19]	32.67	0.97	0.04
D-NeRF[8]	30.50	0.95	0.07
NDVG[12]	30.54	0.96	0.05
Ours	32.68	0.97	0.04

图 3 提供了一些可视化展示。可以渲染出准确而具有细节的图像，例如顶部场景中的头盔和手臂，也可以生成更清晰的边界，例如底部场景中的手和脚。

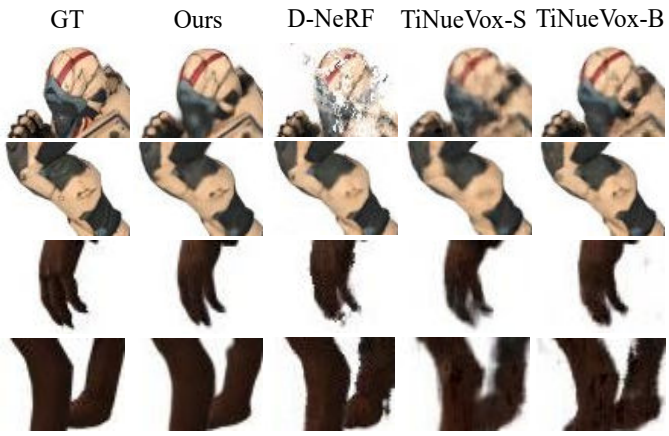


图 3　可视化结果

图 4 比较了提出方法与 D-NeRF[8]重建的规范空间场景模型。提出的方法可以恢复规范空间中的正确几何结构。例如，D-NeRF[8]所生成的球和机械臂位于整个轨迹的"平均"位置，而提出的方法则位于正确的位置。这表明提出的运动模型可以估计出更加精度的运动，从而减少了规范空间中场景模型的误差。

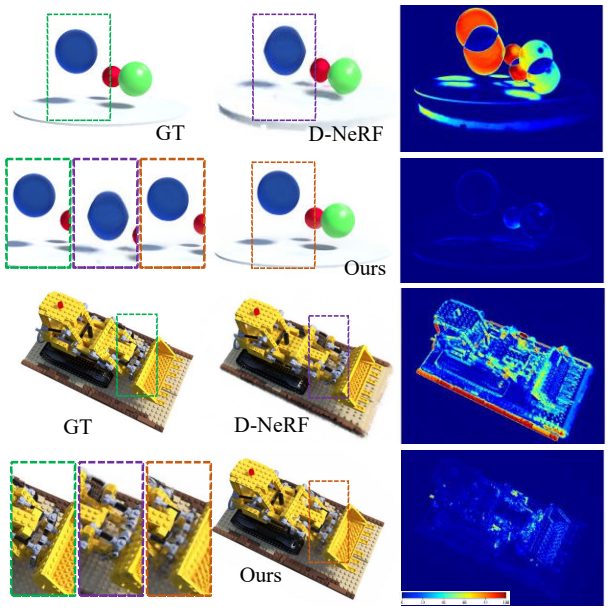


图 4　规范空间场景模型重建比较

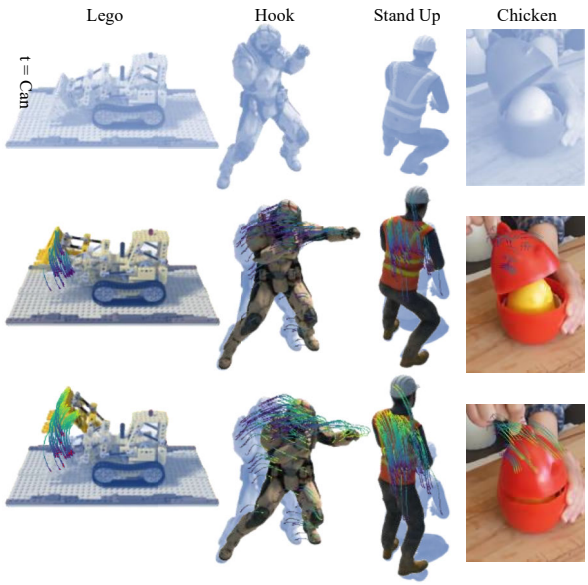


图 5　轨迹可视化

图 5 展示了由轨迹场学习到的轨迹。可以看到提出的方法可以学习到合理的运动轨迹。利用这一特点，在未来的工作中可以考虑引入几何约束、运动模型和先验知识等，来帮助模型提高轨迹的估计质量。

四、总结与展望

本文介绍了一种基于规范空间表示法的前向翘曲方法，用于动态场景的新视角合成。提出的方法在规范空间中对静态场景进行建模，并将整个场向前翘曲到其他时刻，以进行动态场景的渲染。为了解决多对一和一对多映射的问题，提出了一种由平均拼接和填充网络组

成的可微前向翘曲方案。提出的前向翘曲方法在公开数据集上实现了最优的性能。

方法的局限性和未来发展方向：我们目前实现的方法消耗显存相对较大，尤其是在真实场景中。此外，训练速度也相对较慢（每个场景一天）。由于我们的方法采用了前向翘曲技术，该方法可以获得平滑的轨迹场。因此在未来的工作中，我们还可以引入额外的约束条件和运动模型，来帮助模型学习更加精确的轨迹。

责任编委 储珺

参考文献

[1] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In Proceedings of the European Conference on Computer Vision (ECCV), 2020.

[2] Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[3] Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Ronen Basri, and Yaron Lipman. Multiview neural surface reconstruction by disentangling geometry and appearance. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), 2020.

[4] Yilun Du, Yinan Zhang, Hong-Xing Yu, Joshua B Tenenbaum, and Jiajun Wu. Neural radiance flow for 4d view synthesis and video processing. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2021.

[5] Chen Gao, Ayush Saraf, Johannes Kopf, and Jia-Bin Huang. Dynamic view synthesis from dynamic monocular video. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2021.

[6] Zhengqi Li, Simon Niklaus, Noah Snavely, and Oliver Wang. Neural scene flow fields for space-time view synthesis of dynamic scenes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2021.

[7] Keunhong Park, Utkarsh Sinha, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Steven M Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. Proceedings of the IEEE International Conference on Com- puter Vision (ICCV), 2021.

[8] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2021.

[9] Edgar Tretschk, Ayush Tewari, Vladislav Golyanik, Michael Zollho ̈fer, Christoph Lassner, and Christian Theobalt. Non-rigid neural radiance fields: Reconstruction and novel view synthesis of a dynamic scene from monocular video. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2021.

[10] Chaoyang Wang, Ben Eckart, Simon Lucey, and Orazio Gallo. Neural trajectory fields for dynamic novel view synthesis. arXiv preprint arXiv:2105.05994, 2021.

[11] Wenqi Xian, Jia-Bin Huang, Johannes Kopf, and Changil Kim. Space-time neural irradiance fields for free-viewpoint video. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2021.

[12] Xiang Guo, Guanying Chen, Yuchao Dai, Xiaoqing Ye, Jiadai Sun, Xiao Tan, and Errui Ding. Neural deformable voxel grid for fast optimization of dynamic view synthesis. In Proceedings of the Asian Conference on Computer Vision (ACCV), 2022.

[13] Xu Chen, Yufeng Zheng, Michael J Black, Otmar Hilliges, and Andreas Geiger. Snarf: Differentiable forward skinning for animating non-rigid neural implicit shapes. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2021.

[14] Thomas Mu¨ller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. ACM Transactions on Graphics (TOG), 2022.

[15] Cheng Sun, Min Sun, and Hwann-Tzong Chen. Direct Voxel Grid Optimization: Super-fast convergence for radiance fields reconstruction. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2022.

[16] Alex Yu, Ruilong Li, Matthew Tancik, Hao Li, Ren Ng, and Angjoo Kanazawa. Plenoctrees for real-time rendering of neural radiance fields. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2021.

[17] Jae Shin Yoon, Kihwan Kim, Orazio Gallo, Hyun Soo Park, and Jan Kautz. Novel view synthesis of dynamic scenes with globally coherent depths from a monocular camera. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[18] Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Ricardo Martin-Brualla, and Steven M Seitz. HyperNeRF: A higher- dimensional representation for topologically varying neural radiance fields. ACM Transactions on Graphics (TOG), 2021.

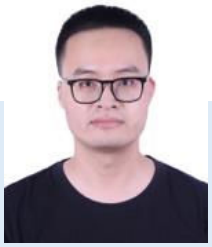
[19] Jiemin Fang, Taoran Yi, Xinggang Wang, Lingxi Xie, Xiaopeng Zhang, Wenyu Liu, Matthias Nießner, and Qi Tian. Fast dynamic radiance fields with time-aware neural voxels. ACM SIGGRAPH Asia, 2022.

[20] Wanshui Gan, Hongbin Xu, Yi Huang, Shifeng Chen, and Naoto Yokoya. V4d: Voxel for 4d novel view synthesis. arXiv preprint arXiv:2205.14332, 2022.

[21] Jia-Wei Liu, Yan-Pei Cao, Weijia Mao, Wenqiao Zhang, David Junhao Zhang, Jussi Keppo, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Devrf: Fast deformable voxel radiance fields for dynamic scenes. arXiv preprint arXiv:2205.15723, 2022.

[22] Simon Niklaus and Feng Liu. Softmax splatting for video frame interpolation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[23] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view syn- thesis. In Proceedings of the European Conference on Computer Vision (ECCV), 2020.



郭相

西北工业大学电子信息学院博士生，研究方向：新视角合成，视觉定位。
Email: guoxiang@mail.nwpu.edu.cn

戴玉超

西北工业大学电子信息学院教授，研究方向：机器视觉与人工智能
Email: daiyuchao@nwpu.edu.cn

热点追踪

RGBD1K：一个用于 RGB-D 目标跟踪的大规模数据集和基准

江南大学 朱学峰 徐天阳 吴小俊

一、摘要

RGB-D 目标跟踪近年来引起了广泛的关注，这主要得益于视觉和深度通道之间的信息合作，使其在性能上取得了令人瞩目的成果。然而，由于有限数量的带标注的 RGB-D 跟踪数据，大多数最先进的 RGB-D 目标跟踪器只是高性能 RGB 跟踪器的简单扩展，在离线训练阶段未充分挖掘深度通道的潜力。为缓解训练数据集不足的问题，我们发布了一个名为 RGBD1K 的新 RGB-D 数据集。RGBD1K 包含 1,050 个序列，总共约 250 万 RGB-D 图像对。为了展示在更大的 RGB-D 数据集上训练模型的好处，我们开发了一种基于 Transformer 的 RGB-D 跟踪算法，可作为未来使用新数据集 RGBD1K 进行视觉目标跟踪研究的基线方法。通过进行的大量实验表明，RGBD1K 数据集作为训练集可以显著提升 RGB-D 目标跟踪的性能，这为未来有效的跟踪器设计开辟了更多可能性。有关数据集和代码的详细信息在项目主页上提供：https://github.com/xuefeng-zhu5/SPT。

二、引言

视觉目标跟踪（Visual Object Tracking, VOT）旨在根据给定的目标初始状态，在视频的后续每一帧中预测给定目标对象的位置和尺度。目标跟踪在计算机视觉和模式识别领域扮演着重要角色，视觉目标跟踪技术的发展已经持续了几十年。特别是近年来，随着大规模标注数据集 (例如 GOT10K[1]、TrackingNet[2]和 LaSOT[3] 等）的发布，深度学习进一步加速了高性能视觉目标跟踪算法的发展。使用数百万帧标记的图像进行离线训练，跟踪网络能够学习强大的特征表示，相比传统的在线学习方法，取得了显著的性能提升[4]。

最近，随着低成本 RGB-D 传感器的广泛普及，视觉目标跟踪的研究已经从单模态 RGB 数据扩展到了多模态 RGB-D 视频数据。RGB-D 图像由三通道 RGB 图像和单通道距离深度（Depth）图组成。与传统的 RGB 跟踪相比，RGB-D 数据的附加深度图像提供了额外的空间信息，有助于在复杂场景中实现稳定的目标跟踪[5,6]。然而，现有的大多数 RGB-D 跟踪方法是建立在高性能的 RGB 跟踪器基础上，仅在在线跟踪阶段采用深度信息以支持部分遮挡物体的推理和重新检测消失的目标[7,8]。与 RGB 目标跟踪算法相比，多模态 RGB-D 跟踪的发展速度不如人意。其中主要原因是 RGB-D 跟踪的训练数据不足。公开可用的带有标注的 RGB-D 视频无法支撑 RGB-D 跟踪网络的离线训练。具体而言，现有的 RGB 跟踪数据集包含数千个视频序列，其中有数百万帧带标注图像，但现有的 RGB-D 跟踪标注数据数量要少很多，远远不足以推动 RGB-D 目标跟踪算法的快速发展。

为了进一步激发对 RGB-D 跟踪及其应用的研究，我们采集了一个名为 RGBD1K 的新 RGB-D 数据集。RGBD1K 数据集共包含 1050 个序列，其中，有 1,000 个视频用于训练，50 个视频用于测试。考虑到训练视频的标注成本以及长时视频的视频片段也能包含目标代表性的视觉和深度外观变化，足以支持跟踪模型的学习，因此，我们只对每个视频的前 600 帧进行标注。这样，RGBD1K 包含 60 万个标注图像帧可用于端到端基于深度网络的 RGB-D 跟踪方法的监督学习。对于测试集，所有图像帧都被标注，总共包含约 11.8 万帧。此外，我

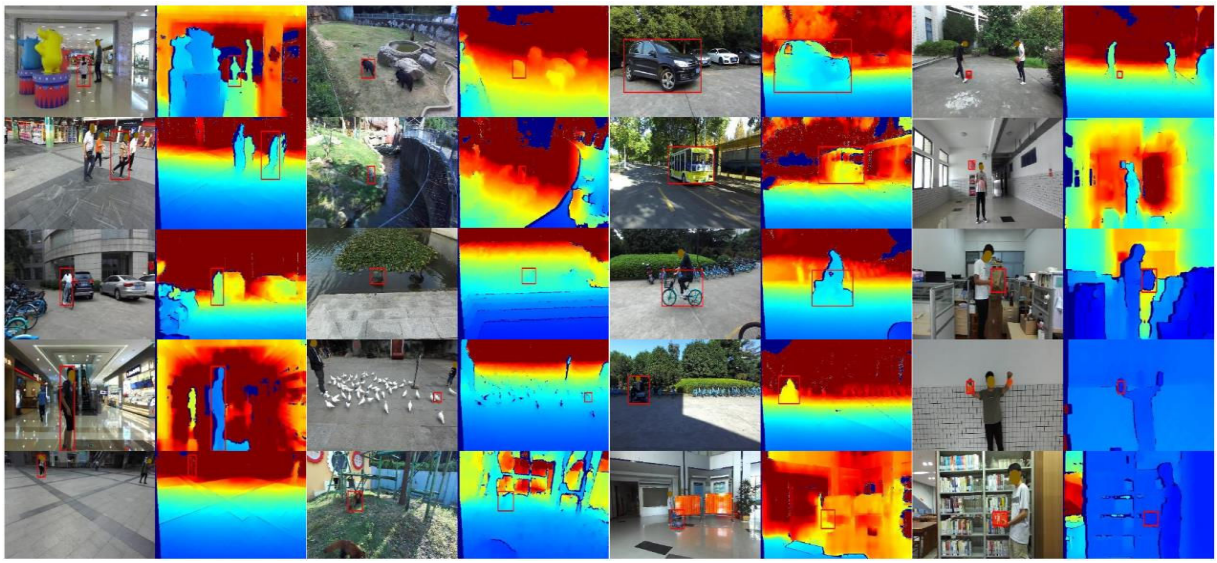


图 1　所提出多模态数据集 RGBD1K 的部分场景示例

们对测试集的每一帧都使用 15 个不同场景挑战属性进行标注。数据的挑战属性标注有助于对跟踪器性能与缺陷进行分析。表 1 对比了现有的 RGB-D 数据集，包括 PTB[9]、STC[10]、CDTB[11]、DepthTrack[12]以及所提出的 RGBD1K 数据集。从表 1 中可以看出，提出的 RGBD1K 具有最多的序列数、RGB-D 图像对数、标注量以及序列的平均长度。另外，为了展示新数据集对 RGB-D 跟踪性能的影响，我们提出了一种基于 Transformer 的跟踪算法，具体而言，我们将基于 RGB 的目标跟踪网络 STARK[13]扩展为 RGB-D 版本，并设计了一个用于融合两种模态特征的模块。该方法使用 RGBD1K 数据集的 1000 个训练序列进行离线训练。在 RGBD1K、DepthTrack 和 CDTB 数据集上进行的大量验证评估和相应的结果表明了 RGBD1K 数据集的有效性以及所提出的 RGB-D 跟踪方法的性能优势。

表 1　RGBD1K 和现有 RGB-D 跟踪数据集的对比

数据集	视频数	帧数	平均帧数	标注数	挑战属性
PTB	100	21,542	215	21,542	5
STC	36	9,009	250	9,009	12
CDTB	80	101,956	1,274	101,956	13
DepthTrack	200	294,591	1,473	294,591	15
RGBD1K	1,050	2,503,400	2,384	717,900	15

三、RGBD1K 数据集

3.1 视频序列

RGBD1K 包含 1,000 个训练序列和 50 个测试序列。总体而言，训练集共包含 2,385,500 个 RGB-D 图像对，测试集包含 117,900 图像对。RGBD1K 数据集的所有 1,050 个序列都是使用立体摄像头 ZED 在室内或室外采集的。ZED 相机提供了时间同步和像素对齐的 RGB 和 Depth 图像对。每个视频的帧率为每秒 25 帧。其中，RGB 图像以 24 位（每通道 8 位）的 JPEG 格式存储，深度图以 16 位 PNG 格式存储。

RGBD1K 涵盖了大量的对象类别，包括涉及人类、动物、交通工具和日常用品的 100 多种不同类型的目标物体，图 1 展示了不同目标类别的一些示例。我们选择了数十种不同的场景来拍摄这些序列，例如办公楼、购物中心、动物园、体育场等。此外，一些视频是从第一人称视角和俯视视角捕获的，用以模拟移动机器人、无人机和监控摄像机的视角。

3.2 数据标注

对于每个视频，我们使用目标边界框对图像进行标注。众所周知，数据标注对于科学研究至关重要但非常耗时。考虑到视频序列的片段可以包含足够的目标视觉和深度外观变化，同时为了减少标注的时间成本，对于训练集，我们只标注了每个视频中一个片段的图像帧。具体来说，对于训练集的每个序列，我们只标注了前 600 帧图像。尽管平均每个视频只标注了其长度的 1/4，但我们认为标注的片段中的外观变化足以让模型学习到目标和场景的时空变化[14]。此外，未标记的部分与标记的部分紧密相关，这样部分标记的数据可以直接用于监

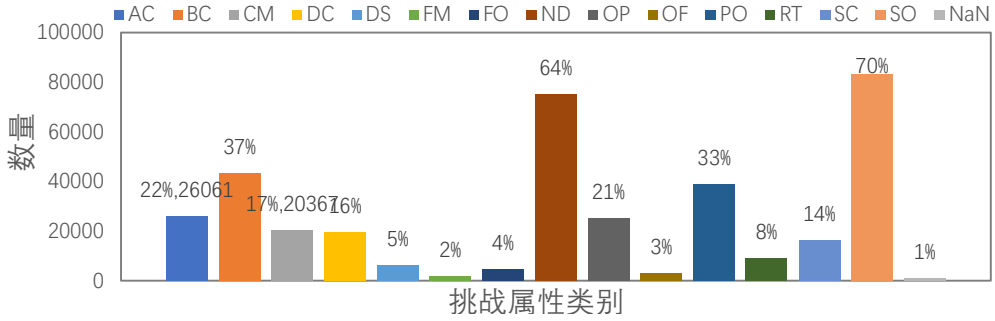


图 2　测试集各种挑战属性的数据分布

督学习同时也存着用于半监督学习的可能。对于测试集，每个序列的所有图像都进行了标注。

为了进一步分析跟踪算法的性能，我们使用了CDTB[11]和 DepthTrack[12]提出的 15 种场景挑战属性为测试集中的每一帧进行了标注。这些挑战属性包括纵横比变化（Aspect-ratio Change，AC），背景杂乱（Background Clutter，BC），摄像机运动（Camera Motion，CM），深度变化（Depth Change，DC），暗场景（Dark Scene，DS），快速运动（Fast Motion，FM），完全遮挡（Full Occlusion，FO），非刚性形变（Non-rigid Deformation，ND），平面外旋转（Out-of-plane Rotation，OP），超出视野（Out of Frame，OF），部分遮挡（Partial Occlusion，PO），反光目标（Reflective Target，RT），尺度变化（Size Change，SC），相似对象（Similar Objects，SO）和未分配（Unassigned, NaN）。其中 AC、DC、FM、SC 和 NaN 这些属性是根据 RGB-D 图像和目标边界框标注计算得到，其余 10 种挑战属性是人工手动标注的。这些场景的挑战属性有助于分析跟踪器在特定挑战下的优劣势。

此外，RGBD1K 测试集中每种挑战属性类别的帧数分布如图 2 所示。从图中可以看出，RGBD1K 数据集只有 1%的图像被标记为没有任何挑战属性，这表明 RGBD1K 测试集对于目标跟踪来说极具挑战性。在所有序列中，大约 64%的图像中的目标属于非刚性可形变物体。通常，可形变物体意味着极端外观变化出现的概率较高，这对于稳定跟踪来说更加困难。此外，70%的图像被标记为相似对象的挑战属性。背景中相似对象的干扰是实现鲁棒单目标跟踪的一个值得研究的重要问题。此外，背景嘈杂和局部遮挡也是 RGBD1K 测试集中重要的挑战因素。尽管某些属性包含的帧数较少，例如 FO 和 OF 只占 4%和 3%，但它们仍然对实际应用非常有价值。具有 FO 或 OF 属性的图像意味着目标在当前图像中是不可见的。尽管总体上只有 7%的图像中的目标是不可见的，但这意味着测试集中的每个视频平均有大约 165 帧的目标消失。目标频繁的长时间消失和重新出现使跟踪问题变得复杂，需要 RGB-D 跟踪器具备较强的感知能力。

3.3 性能评价指标

虽然对于 RGBD1K 的使用没有明确的限制，但在测试集上评估跟踪算法性能时，我们提倡使用长时跟踪评估协议[15]。这个长时跟踪评估协议被主要用于 VOT 竞赛的长时跟踪赛道和多模态 RGB-D 跟踪赛道。在 RGBD1K 数据集中有一定比例的图像中目标是不可见的，即目标可能在一个视频中多次消失和重新出现。因此，在 RGBD1K 上评估跟踪算法时，算法定位目标以及预测目标消失并再次捕获重新出现的目标的能力对于稳健的跟踪系统至关重要。因此，长时 VOT 评估协议非常适用于在我们的数据集上评估跟踪算法。其中主要的评价指标有跟踪精度（Precision，Pr）和召回率（Recall, Re）。具体而言，精度定义为在检测到目标的图像帧上，计算预测目标框和实际目标框的平均重叠比率。召回率表示在目标可见的图像帧上，计算预测目标边界框与标签边界框的平均重叠比率。最后最主要的性能指标是通过计算结合了跟踪精度和召回率的跟踪 F-分数（F-score）。此外，可以使用 VOT 竞赛的评测工具包[]非常方便地在 RGBD1K 上评估跟踪器。具体的三个评价指标计算公式如下：

$$\mathrm{Pr}(\tau_\theta)=\frac{1}{N_p}\sum_{t\in\{t:A_t(\tau_\theta)\neq\emptyset\}}\Omega\left(\mathrm{A}_{\mathrm{t}(\tau_\theta)},\mathrm{G}_\mathrm{t}\right),$$

$$\mathrm{Re}(\tau_\theta)=\frac{1}{N_g}\sum_{t\in\{t:G_t\neq\emptyset\}}\Omega\left(\mathrm{A}_{\mathrm{t}(\tau_\theta)},\mathrm{G}_\mathrm{t}\right),$$

$$F(\tau_\theta)=\frac{2\mathrm{Pr}(\tau_\theta)\mathrm{Re}(\tau_\theta)}{\mathrm{Pr}(\tau_\theta)+\mathrm{Re}(\tau_\theta)}.$$

其中，$\mathrm{G_t}$表示实际边界框，$A_{t(\tau_\theta)}$表示在第 t 帧的预测边界框。$\Omega\left(A_{t(\tau_\theta)},G_t\right)$表示实际边界框和跟踪预测之间的交并比（Intersection-over-Union, IoU）。τ_θ是一个置信度阈值。评估协议要求跟踪器一并报告预测边界框和置信度分数。如果在第 t 帧的预测置信度分数θ_t低于τ_θ，则$A_t(\tau_\theta)=\emptyset$。$N_p$是跟踪算法给出预测的帧数，即$A_t(\tau_\theta)\neq\emptyset$的帧数，$\mathrm{N_g}$是目标在视野里可见的帧数，即$G_t\neq\emptyset$的帧数。

四、基线 RGB-D 跟踪器

为了展示 RGBD1K 数据集的重要性，并激发新的 RGB-D 跟踪算法设计，我们提出了一种名为 SPT 的新的 RGB-D 跟踪基线方法。SPT 是从最近的基于 Transformer 的跟踪器 STARK[13]发展而来的。STARK 是一种高性能的基于 RGB 数据的目标跟踪器。SPT 是通过将 STARK-S（STARK 没有使用时序结构的版本）扩展为具有专用特征融合模块的 RGB-D 版本而构造的。SPT 的架构在图 3 中给出。首先，将两个模态的搜索区域和初始模板输入到骨干网络中以分别提取深度 CNN 特征。这里使用的骨干网络是 ResNet-50[16]网络。每个模态的搜索区域和模板的特征都是$\mathrm{H}\times\mathrm{W}\times\mathrm{C}$和$h\times w\times\mathrm{C}$大小的张量。然后，我们将每个模态的特征进行展平并级联，然后通过一个 6 层编码器层堆叠成的 Transformer 编码器来融合每个模态的模板-搜索区域特征信息。最后，两个模态特定编码器的输出通过设计的特征融合模块进行融合。

关于所提出的特征融合模块，首先，深度模态的 Transformer 编码器的输出和 RGB 模态的编码器的输出在通道维度上进行级联。然后采用一维卷积层来减少级联特征的通道数，从2C通道减少到C通道。最后，我们引入一个 2 层编码层组成的 Transformer 编码器，以进一步融合和增强两个模态的特征。每个编码器层包含多头自注意模块和前馈网络[13]。

网络框架的其余部分包括目标 query、Transformer 解码器和目标边界框预测头都保持和 STARK 算法一致。其中 Transformer 解码器由 6 层解码层堆叠而成，通过将可学习的目标 query 和融合特征作为输入生成输出，以学习目标通用的鉴别信息。每个

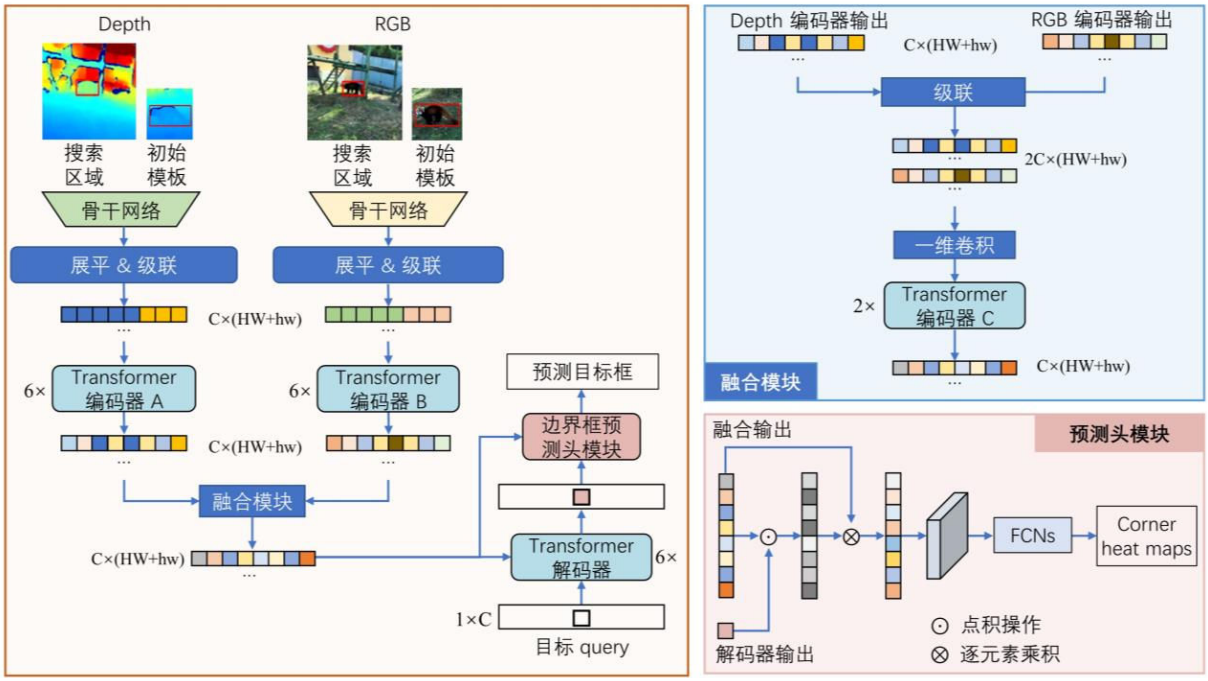


图 3　所提出的 RGB-D 目标跟踪基线算法的网络框架图

表 2　在 RGBD1K、DepthTrack 和 CDTB 上的消融对比实验结果

数据集	RGBD1K			DepthTrack			CDTB		
评价指标	Pr	Re	F-score	Pr	Re	F-score	Pr	Re	F-score
STARK-S	0.480	0.510	0.495	0.490	0.511	0.500	0.630	0.701	0.664
STARK-S-FT	0.509	0.537	0.522	0.497	0.517	0.507	0.638	0.706	0.670
SPT	0.545	0.578	0.561	0.527	0.549	0.538	0.654	0.726	0.688

解码器层由自注意模块、编码器-解码器注意模块和前馈网络组成。然后，Transformer 解码器的输出和融合特征一起输入到边界框预测头，以预测目标框坐标。

在边界框预测模块中，首先计算解码器输出与融合特征的相似度，并使用相似度来增强融合特征。然后，增强后的特征张量被调整形状并通过全卷积网络传递，生成左上角热图和右下角热图。通过左上角和右下角的点，目标边界框就可以被确定。最后，和 STARK 算法一样，SPT 网络训练损失函数是L_1损失和IoU损失组合[13]。

五、实验与分析

我们在 RGBD1K、DepthTrack 和 CDTB 数据集上进行了实验。在本节中，我们描述了所提出算法的实现细节，包括参数设置和实验平台。然后，本节呈现了消融研究的结果，以展示我们数据集以及提出的特征融合模块的有效性。最后，我们提供了比较实验的结果和相应的分析。

5.1 实验环境与设置

我们提出的 SPT 跟踪器是在有一块 Intel i9-CPU 和一块 NVIDIA GeForce RTX 3090 GPU 的计算机平台上进行训练和评估的。训练和测试参数设置与 STARK 算法相同，除了学习率和训练周期的设置。SPT 训练的学习率设置为10^{-5}，训练周期数为 250。对于 SPT 的骨干网络、Transformer 编码器 A、B、Transformer 解码器和边界框预测头网络，我们在训练时使用官方发布的 STARK-S 模型相应组件的权重进行初始化。SPT 的训练数据集为 RGBD1K 的训练集。

5.2 消融研究

为了展示提出的 RGBD1K 数据集对提升 RGB-D 跟踪性能的有效性，首先我们构建了三个跟踪器，包括 STARK-S[13]、STARK-S-FT 和我们的 SPT。STARK-S 是 ResNet-50 作为骨干网络（没有时间分支）的 STARK 跟踪器，是 SPT 的基线方法。对此，我们使用官方发布的 STARK-S 训练模型。STARK-S-FT 是在 RGBD1K 上仅使用训练集的所有 RGB 图像对 STARK-S 进行微调训练后的跟踪器。SPT 使用 RGBD1K 的 RGB-D 图像进行训练。

在 RGBD1K 测试集上的结果如表 2 所示。在使用 RGBD1K 训练集的 RGB 图像进行微调后，STARK-S-FT 在精度、召回率和 F-score 方面的性能从 0.480、0.510 和 0.495 提高到 0.509、0.537 和 0.522。同样使用 RGBD1K 训练集的 RGB-D 图像进行训练后，SPT 在精度、召回率和 F-score 方面进一步提高到 0.545、0.578 和 0.561。这个提高可以使我们得出结论，RGBD1K 的带标注的 RGB 图像和深度图像都有助于提高 RGB-D 跟踪性能。为了进一步确认 RGBD1K 的有效性，我们在另外两个数据集 DepthTrack 和 CDTB 上进行相同的实验，以探索 STARK-S、STARK-S-FT 和 SPT 之间的性能。值得注意的是，这些跟踪器仅使用 RGBD1K 进行训练，没有使用 DepthTrack 或 CDTB 的序列来微调跟踪网络或相应的超参数。在 DepthTrack 和 CDTB 上的结果也在表 2 中展示。

从表中可见，在 RGBD1K 数据集上训练后，STARK-S-FT 和 SPT 在 DepthTrack 和 CDTB 数据集上取得显著的性能提升。在使用 RGBD1K 的 RGB-D 数据训练，SPT 将 STARK-S 在两个数据集上的精度、召回率和 F-score 分别从 0.490、0.511 和 0.500 提高到 0.527、0.549 和 0.538，从 0.630、0.701 和 0.664 提高到 0.654、0.726 和 0.688。在 F-score 度量方面，SPT 分别提高了 STARK-S 在 DepthTrack 和 CDTB 上的性能 7.6%和 3.6%。显然，表 2 结果证明了提出的 RGBD1K 数据集在训练端到端的 RGB-D 跟踪器算法方面有着广泛优势。

表 3　在 RGBD1K、DepthTrack 和 CDTB 上的对比实验结果

跟踪算法	RGBD1K			DepthTrack			CDTB		
	Pr	Re	F-score	Pr	Re	F-score	Pr	Re	F-score
ATCAIS	0.511	0.451	0.479	0.500	0.455	0.476	0.709	0.696	0.702
DDiMP	0.557	0.534	0.545	0.503	0.469	0.485	0.703	0.689	0.696
TALGD	0.511	0.451	0.479	0.494	0.424	0.456	**0.728**	0.717	**0.722**
DAL	**0.562**	0.407	0.472	0.512	0.369	0.429	0.647	0.571	0.607
DeT	0.438	0.419	0.428	**0.560**	0.506	0.532	0.674	0.642	0.657
SPT	0.545	**0.578**	**0.561**	0.527	**0.549**	**0.538**	0.654	**0.726**	0.688

5.3 对比研究

我们将提出的 SPT 与最近的一些 RGB-D 跟踪算法进行了比较，这些跟踪器包括 VOT-RGBD 竞赛[18,19,4]上的算法 ATCAIS、DDiMP、TALGD 和 Siam_LTD，以及 DAL[17]和 DeT[12]。表 3 中呈现了这些算法在 RGBD1K, DepthTrack 和 CDTB 三个数据集上的跟踪结果。一般来说，F-score 是 VOT 协议中最重要的性能度量，跟踪器按照 F-score 值进行排名。从表中可以看出，在 RGBD1K 测试集上，SPT 实现了最佳的 F-score，而 RGB-D 跟踪器 DDiMP 是第二优秀的跟踪器。与 DDiMP 跟踪器相比，我们的 SPT 跟踪器在 F-score 方面取得了 2.9%的提高。此外，与 ATCAIS、TALGD、DAL 和 DeT 等 RGB-D 跟踪器相比，提出的 SPT 跟踪器也有显著优势。SPT 的跟踪性能提升表明使用提出的 RGBD1K 进行训练模型有助于实现更稳健的 RGB-D 目标跟踪。

为了更好地反映我们的 RGBD1K 对于跟踪性能提升的普遍优势，我们同样将 SPT 跟踪器与最先进的跟踪器在 DepthTrack 和 CDTB 数据集上进行了比较。值得注意的是，我们的 SPT 跟踪器是使用 RGBD1K 训练然后直接在 DepthTrack 和 CDTB 数据集上测试的，没有微调任何参数。在表 3 中可以看出，SPT 在 DepthTrack 数据集上实现了最佳的 F-score（0.538）和召回率 Recall（0.549）。在 CDTB 数据集上，SPT 在 Recall 方面显著优于其他最先进的跟踪器。尽管在 Precision 和 F-score 方面，SPT 相对于 DDiMP、ATCAIS 和 TALGD 较差，但这主要是因为这些算法使用了多个跟踪器的组合，且在 CDTB 数据集上过拟合。因此，这些算法的效率较低，而且在另外两个数据集 RGBD1K 和 DepthTrack 上性能大打折扣。以上结果证实，使用大量标注的 RGB-D 数据离线训练的 SPT 跟踪器可以实现较为鲁棒精准的 RGB-D 目标跟踪。另一方面，这些结果也证实了所提出的 RGBD1K 数据集对于 RGB-D 目标跟踪发展的重要性。

六、总结

在这项工作中，我们提出了一个用于 RGB-D 目标跟踪的大规模数据集，以及一个基于端到端深度网络的基线跟踪器。这项工作的动机是现有带标注 RGB-D 视频的稀缺阻碍了 RGB-D 跟踪的发展。所提出的 RGBD1K 数据集大大提升了现有 RGB-D 目标跟踪的标注数据量。为了展示 RGBD1K 数据集的实用性，我们设计了一种新的 RGB-D 跟踪基线方法，并使用 RGBD1K 训练集的所有 RGB-D 数据进行离线训练。使用 RGBD1K、DepthTrack 和 CDTB 数据集测试得到的广泛实验结果展示了在 RGBD1K 上进行训练算法的好处以及其推动未来 RGB-D 跟踪发展的潜力。

责任编委　崔海楠

参考文献

[1] Lianghua Huang, Xin Zhao, and Kaiqi Huang. Got-10k: A large high-diversity benchmark for generic object tracking in the wild [J]. IEEE TPAMI, 2019, 43(5): 1562-1577.

[2] Matthias Muller, Adel Bibi, Silvio Giancola, Salman Alsubaihi, Bernard Ghanem; "Trackingnet: A large-scale dataset and benchmark for object tracking in the wild." ECCV. 2018, pp. 300-317.

[3] Heng Fan, Liting Lin, Fan Yang, Peng Chu, Ge Deng, Sijia Yu, Hexin Bai, Yong Xu, Chunyuan Liao, Haibin Ling; "Lasot: A high-quality benchmark for large-scale single object tracking." CVPR, 2019, pp. 5374-5383.

[4] Matej Kristan, et al. "The ninth visual object tracking vot2021 challenge results." ICCV. 2021.

[5] Meshgi, Kourosh, et al. "An occlusion-aware particle filter tracker to handle complex and persistent occlusions." CVIU, 150 (2016): 81-94.

[6] Timur Bagautdinov, Francois Fleuret, and Pascal Fua. "Probability occupancy maps for occluded depth images." CVPR. 2015, pp. 2829-2837.

[7] Sion Hannuna, et al. "DS-KCF: a real-time tracker for RGB-D data." Journal of Real-Time Image Processing 16 (2019): 1439-1458.

[8] Ugur Kart, Joni-Kristian Kamarainen, and Jiri Matas. "How to make an rgbd tracker?." ECCVW. 2018.

[9] Shuran Song, and Jianxiong Xiao. "Tracking revisited using RGBD camera: Unified benchmark and baselines." ICCV. 2013, pp. 233-240.

[10] Jingjing Xiao, Rustam Stolkin, Yuqing Gao, Aleš Leonardis. "Robust fusion of color and depth data for RGB-D target tracking using adaptive range-invariant depth models and spatio-temporal consistency constraints." IEEE TCYB 48.8 (2017): 2485-2499.

[11] Alan Lukezic, Ugur Kart, Jani Kapyla, Ahmed Durmush, Joni-Kristian Kamarainen, Jiri Matas, Matej Kristan. "Cdtb: A color and depth visual object tracking dataset and benchmark." ICCV. 2019, pp. 10013-10022.

[12] Song Yan, Jinyu Yang, Jani Käpylä, Feng Zheng, Aleš Leonardis, Joni-Kristian Kämäräinen. "Depthtrack: Unveiling the power of rgbd tracking."ICCV. 2021, pp. 10725-10733.

[13] Bin Yan, Houwen Peng, Jianlong Fu, Dong Wang, Huchuan Lu. "Learning spatio-temporal transformer for visual tracking." ICCV, 2021, pp. 10448-10457.

[14] Jack Valmadre, Luca Bertinetto, Joao F. Henriques, Ran Tao, Andrea Vedaldi, Arnold W.M. Smeulders, Philip H.S. Torr, Efstratios Gavves. "Long-term tracking in the wild: A benchmark." ECCV. 2018, pp. 670-685.

[15] Alan Lukežič, Luka Čehovin Zajc, Tomáš Vojíř, Jiří Matas, Matej Kristan. "Now you see me: evaluating performance in long-term visual tracking." arXiv preprint arXiv:1804.07056 (2018).

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun. "Deep residual learning for image recognition." Proceedings of the CVPR. 2016, pp. 770-778.

[17] Yanlin Qian, Song Yan, Alan Lukežič, Matej Kristan, Joni-Kristian Kämäräinen, Jiří Matas. "DAL: A deep depth-aware long-term tracker." ICPR., 2021, pp. 7825-7832.

[18] Matej Kristan, et al. “The eighth visual object tracking VOT2020 challenge results.” ECCVW 2020: 547-601.

[19] Matej Kristan, et al. "The seventh visual object tracking VOT2019 challenge results." ICCV, 2019.

朱学峰

江南大学博士生，导师为江南大学吴小俊教授，主要研究方向为计算机视觉、目标跟踪。
Email: xuefeng.zhu@stu.jiangnan.edu.cn

徐天阳

博士，江南大学副教授，主要研究方向为模式识别、计算机视觉和人工智能。
Email: tianyang.xu@jiangnan.edu.cn

吴小俊

博士，江南大学二级教授、至善教授、研究生院院长，IAPR Fellow，AAIA Fellow，主要研究方向为模式识别与人工智能。
Email: wu_xiaojun@jiangnan.edu.cn