arXiv:2304.06116v1 [cs.CV] 12 Apr 2023

AutoShot：一个短视频数据集和最先进的镜头边界检测

Wentao Zhu[1],Yufang Huang[3], Xiufeng Xie[1], Wenxian Liu[1],Jincan Deng[1], Debing Zhang[1], Zhangyang Wang[2], Ji Liu[1]1快手科技 [2]德克萨斯大学奥斯汀分校 [3]康奈尔大学

摘要

短视频的迅猛发展使其广受欢迎，并主导了新的社交媒体趋势。主流短视频平台，例如快手（*Kwai*）、*TikTok*、*In- stagram Reels* 和 *YouTube Shorts*，已经改变了我们消费和创作内容的方式。对于视频内容的创作和理解而言，镜头边界检测（*SBD*）是各种场景中最重要的组成部分之一。在这项工作中，我们发布了一个新的公开短视频镜头边界检测数据集，名为 *SHOT*。该数据集包含 *853* 个完整短视频和 *11,606* 个镜头标注，其中 *200* 个测试视频包含 *2,716* 个高质量镜头边界标注。借助这一丰富的新数据，我们通过在包含多种先进 *3D* 卷积网络和 *Transformer* 的搜索空间中进行神经架构搜索，优化视频镜头边界检测的模型设计。我们提出的方法名为 *AutoShot*，在新构建的 *SHOT* 数据集上进行生成和评估时，其 *F1* 分数高于以往最先进方法，例如比 *TransNetV2* 高出 *4.2%*。此外，为了验证 *AutoShot* 架构的泛化能力，我们直接在另外三个公开数据集 *ClipShots*、*BBC* 和 *RAI* 上对其进行评估，*AutoShot* 的 *F1* 分数分别比以往最先进方法高出 *1.1%*、*0.9%* 和 *1.2%*。*SHOT* 数据集和代码可在 *https://github.com/wentaozhu/AutoShot.git*。

1. 引言

短视频已在全球各年龄群体中广泛普及。在 5G 时代，短视频和视频广告的占比呈爆炸式增长，这得益于短视频相比图像和文本形式内容更丰富、传播效果更好且更具说服力 [36]。这一强劲趋势使得除了简单的视频分类之外，人们对时间精度高且全面的视频分析产生了显著而迫切的需求 [39,42]。镜头边界检测是实现全面时序视频分析的基础组成部分分析，也可作为多种任务的基本模块，例如，场景边界检测 [6,26], 视频结构化 [36], 和事件分割 [29]。例如，借助智能视频创作中精准的镜头边界检测，可以自动制作适用于不同平台且时长符合要求的奖励视频。

为加快视频时间边界检测的发展，研究人员通过费力的人工标注收集了多个数据集。传统的镜头边界检测数据集，例如BBC《地球脉动》纪录片系列 [1]和 RAI [2]，仅包含纪录片或访谈节目，其中的场景相对静态。Tang等人。 [32] 进一步构建了一个大规模视频镜头数据库 ClipShots，其中包含从 YouTube 和微博收集的不同类型的视频，涵盖体育、电视节目、动物等 20 多个类别。等。Shou等人 [29] 构建了一个通用事件边界检测（GEBD）数据集 Kinetics-GEBD，该数据集将片段定义为人类自然感知到某个事件发生的时刻。由于短视频与传统视频的时长差异很大，即 90% 的短视频时长不到一分钟，而其他数据集中的视频时长为 2 至 60 分钟，如表所示1以及图1右图，这显著造成了内容、呈现方式、时间动态和镜头转场方面的差异，如图1 左图所示。短视频数据集对于加快短视频镜头边界检测的发展并对其进行恰当评估至关重要。

另一方面，已有多项工作致力于提高视频镜头边界检 [11] 测（SBD）的准确率。DeepSBD 首先将深度时空 ConvNet 应用于视频SBD。深度结构化模型（DSM） [32] 设计了一种级联框架，以加快 SBD 的速度。TransNet [21] 使用扩张卷积单元 [38] 来处理调整大小后的帧序列。TransNetV2 [30] 采用了多种技术，例如，卷积核分解 [37], 批归一化 [14], 跳跃连接 [12], ，并进一步提高了片段镜头上的 F1 分数 [32] 以及 BBC [1]。

在这项工作中，我们首先收集了一个名为 SHOT 的短视频数据集，其中包含 853 个短视频和 11,606 条

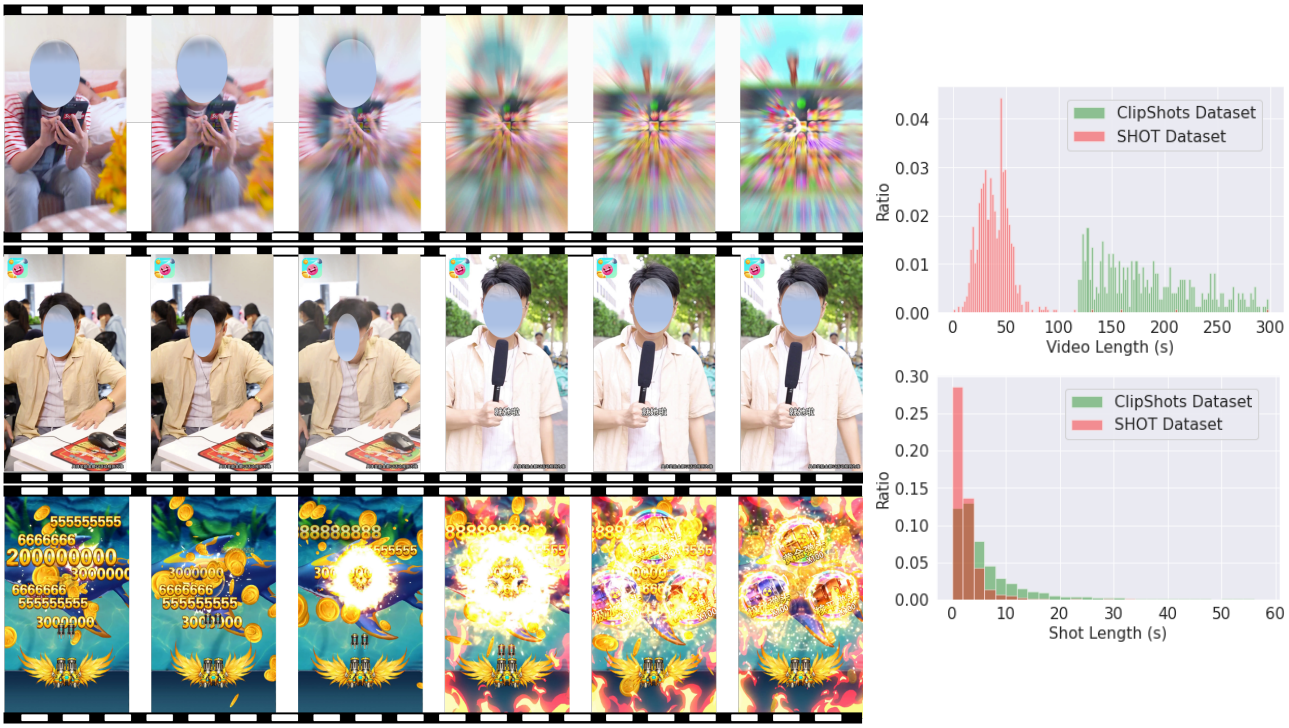


图1。左：在短视频中检测镜头边界可能是一项颇具挑战的任务。镜头转场可能由多个复杂的渐变转场组合而成（第一行），也可能是两个镜头中主体的快速转场（第二行）。游戏视频中，镜头内的视觉效果可能差异很大（第三行）。右：ClipShots 和我们收集的 SHOT 数据集测试集的视频时长和镜头时长（秒）对比。SHOT 中短视频的视频时长范围与 ClipShots 中测试视频的时长范围几乎没有重叠（上）。短视频的镜头时长均不超过 6 秒，而 ClipShots 中的镜头时长范围为 2 至 30 秒（下）。

人工镜头边界精细标注。200 个测试视频的 2,716 条镜头边界标注由专家经过两轮标注完成。利用这些丰富的新数据，我们旨在通过在涵盖多种先进 3D 卷积网络和 Transformer 的搜索空间中进行神经架构搜索，提高视频镜头边界检测的准确率。我们采用了单路径单次超网络策略和贝叶斯优化。搜索得到的模型名为 AutoShot，在 SHOT 数据集上的 F1分数比 TransNetV2 高 4.2%；在召回率与 TransNetV2 相同的情况下，其精确率指标高出 3.5%。我们还在 ClipShots、BBC 和 RAI 数据集上评估了搜索得到的 AutoShot 架构，其 F1分数分别比此前最先进方法高出 1.1%、0.9% 和 1.2%。我们的贡献总结如下： [41] [25, 40, 43, 45] [35, 39] [9] [28]

- 我们收集了一个短视频镜头边界检测数据集（SHOT），其中包含 853 个短视频和 11,606 条镜头边界标注。SHOT 将会公开发布，可用于推动各类短视频理解任务的发展。
- 我们设计了一个视频镜头边界检测搜索空间，其中涵盖了各种先进的 3D 卷积网络和 Transformer，并构建了用于镜头边界检测的神经架构搜索流程。
- 搜索得到的模型名为 AutoShot。事实证明，它是一种极具竞争力的镜头边界检测架构，不仅在其衍生的 SHOT 数据集上显著优于以往最先进的方法，在其他公开基准上也同样如此。

据我们所知，收集的 SHOT 数据集是首个用于短视频镜头边界检测的数据集，而 AutoShot 是首个专为镜头边界检测设计的神经架构搜索方法。

2. 相关工作

为收集视频镜头边界检测数据集，人们付出了大量努力 [1, 2, 4, 16, 27, 32],，这显著推动了先进镜头边界检测方法的发展。一类特定的视频边界检测旨在将视频解析为由人体动作构成的片段，通常使用展示人们执行各种动作的教学视频 [18, 31, 33]。另一类得到广泛研究的视频边界检测

是场景边界检测，即将视频切分为若干语义上相互独立的片段。电影和电视剧集常用于场景边界检测，MovieScenes [26] 和 AdCuepoints [6]是两个用于场景边界检测的大型电影数据集。此外，Shou 等人 [29] 收集了一个用于通用事件边界检测的新基准数据集 Kinectics-GEBD。对于视频广告场景边界检测，Wang 等人 [36] 最近收集了一个多模态视频广告理解数据集，其中包括场景边界检测和多模态场景分类。然而，场景边界检测通常基于预先提取的镜头，且场景分割的评估以镜头级而非帧级进行。

电视剧和脱口秀也被用于镜头边界检测，其中 BBC《地球脉动》纪录片系列 [1] 和 RAI [2] 是两个常用数据集，由数十个时长从半小时到一小时的视频组成。ClipShots [32] 通过从各种媒体平台收集多样化视频，改进了传统的镜头边界检测数据集，例如 YouTube 和微博上的视频，它是最具挑战性的大规模镜头边界检测数据集之一。如图 1 右图和表 1所示，短视频中的镜头转换与电影中的镜头转换截然不同，因此，收集短视频数据集对于推动短视频镜头边界检测的发展极为必要。据我们所知，目前唯一公开可用的短视频数据集是 SVD 数据集 [15]；但它是为近重复视频检索这一完全不同的任务而开发的。

镜头 镜头 边界检测的准确率得益于前述高质量数据集和深度学习而得到提升。在深度学习时代之前，PySceneDetect [3] 是一个流行的镜头 边界检测库，它依赖于传统特征，例如HSV色彩空间中帧之间的变化。近期关于视频边界检测的进展可分为两类：场景边界检测和镜头 边界检测。Rao 等. [26]提出了一种从局部到全局的场景分割框架，融合了跨片段、段落和影片的多模态信息。Chen等. [6] 进一步提出镜头对比自监督学习 [44] ，通过学习镜头表征，使相邻镜头相比随机选择的镜头具有更高的相似度，然后将学习到的镜头表征用于场景 边界检测。

场景边界检测高度依赖准确的镜头边界检测。DeepSBD [11]通过 C3D 网络 [34]预测 16 帧片段中出现过渡的可能性。DSM [32] 采用级联框架来加快镜头边界检测的速度。Gygli [10] 构建了一种无需任何后处理的快速镜头边界检测方法。TransNet [21] 采用膨胀卷积块 [38] ，并取得了与 DeepSBD 不进行后处理时相当的准确率。TransNetV2 [30] 凭借先进组件超越了以往最先进方法，例如跳跃连接 [12], 、批归一化 [14], 、核分解 [37], 、将帧相似度作为特征，以及多个分类头。借助 3D 卷积网络 [25], 、Transformer [35] 和神经架构搜索 [9], 的进展，AutoShot 能够从设计的搜索空间中自动确定最优的镜头边界检测架构；在收集的 SHOT、ClipShots、BBC 和 RAI 数据集上，其准确率优于以往方法。

3. SHOT：短视频镜头边界检测数据集

短视频凭借比传统媒体（即纯文本和静态图像媒体）更丰富的内容和更生动的效果，成为当今最流行的媒体之一。5G时代便捷且经济实惠的高速移动网络接入，加速了短视频平台（例如 Instagram Reels、YouTube Shorts 和 TikTok）的广泛普及。鉴于这些主流短视频平台拥有大量用户和视频广告，推进当前视频时序分割技术的发展至关重要，尤其是短视频镜头边界检测任务，因为它是许多后续语义理解任务的基础。

3.1. 短视频镜头边界检测面临的挑战

短视频通常定义为时长少于两分钟的视频。较短的视频时长使这些视频更容易传播，也使短视频比时长数小时的传统电影更受欢迎。另一方面，短视频时长较短，迫使整个事件在较短的时间段内发生，从而导致事件节奏快得多。这反过来又使短视频中的镜头时长大幅缩短，如图1 右图所示，这加剧了短视频镜头边界检测的难度。

收集的 SHOT 数据集与 ClipShots 测试集之间的视频时长差异巨大 [32] ，如图所示1 右图。几乎所有测试短视频的时长都不到 100 秒，而 ClipShots 中几乎所有测试视频的时长都超过 120 秒。总视频时长较短，直接导致 SHOT 数据集中的镜头转换十分频繁，如图所示1右图。SHOT 数据集测试集中的大多数镜头时长都在 5 秒以内，而 ClipShots 测试集中的镜头时长则介于 2 秒至 30 秒之间。由于视频时长和镜头时长分布差异巨大，传统的镜头边界检测数据集可能不适用于短视频镜头边界检测的开发。

如图所示，短视频镜头边界检测更具挑战性，也更为困难。1左图和图2，因为短视频中的场景比传统视频复杂得多。例如，为了达到更具说服力的效果，短视频中的镜头转场通常会组合使用多种复杂的渐变镜头转场（第一行）。第二个常见挑战是垂直三段式结构视频（图2第二行），其中只有视频的中间部分发生变化。最上方和最下方区域显示视频广告中的下载链接或品牌标识。由于压缩内容区域内发生变化的区域相对较小，垂直三段式结构视频大大增加了镜头边界检测的难度。虚拟场景中夸张的表情会导致游戏视频镜头边界检测出现误报（第三行）。实际上，游戏视频在视频广告和短视频中占有很大比例。因此，为面临挑战的短视频镜头边界检测收集短视频数据集并非易事。

3.2. SHOT 数据集

为加快短视频相关镜头边界检测的研究，我们从一个使用最广泛的短视频平台收集了 853 个短视频。数据集属性对比如表 1 所示。帧总数为 960,794，接近一百万帧。逐帧进行镜头边界标注是一项繁重的任务。数据标注和质量控制策略见附录。为去除个人隐私信息，我们采用最先进的人脸检测器 [7,24] 来检测并遮挡人脸区域。

受 TransNetV 中帧处理方式的启发2 [30],，我们开发了一种基于视频缩略图的标注方法，将每帧缩放至 48×27，如图 3 所示。每帧左上角会自适应显示帧编号，大幅减少了标注人员核对帧编号所需的工作量。如果帧编号位置的像素值较暗，就以浅色显示帧编号；否则，以深色显示帧编号。我们有三位专家和一个标注团队，共同完成了 960,794 帧的全部标注工作。标注团队对 459 个短视频进行标注，共获得 6,111 个镜头。标注完成后，我们对这 6,111 条标注进行了专家检查。我们随机抽取了 200 条镜头标注，发现这 6,111 条镜头标注中存在 2% 的错误率。考虑到镜头定义存在歧义，且标注人员之间的标注存在差异，我们认为 6,111 个镜头中 2% 的错误率是可以接受的。

测试集上的标注质量直接影响基准方法的准确评估，即短视频数据集的质量。为确保测试集上的镜头边界标注具有高质量

数据集	BBC	RAI	片段。	SHOT
复杂场景渐变变换	✗	✗	✗	✓
虚拟场景	✗	✗	✗	✓
三元视频	✗	✗	✗	✓
平均视频时长（秒）	2945	591	237	39.5
平均镜头时长（秒）	6.57	5.65	15.34	2.59

表1. 不同短边界检测数据集的比较即，BBC、RAI、ClipShots 和 SHOT，在复杂渐变转场、虚拟场景、三元视频、平均视频时长和平均镜头长度方面的比较。



图 2. SHOT 中短视频镜头边界检测的独特挑战，例如复杂的渐变镜头转场、垂直三元结构的视频，以及游戏视频虚拟场景中巨大的镜头内变化。

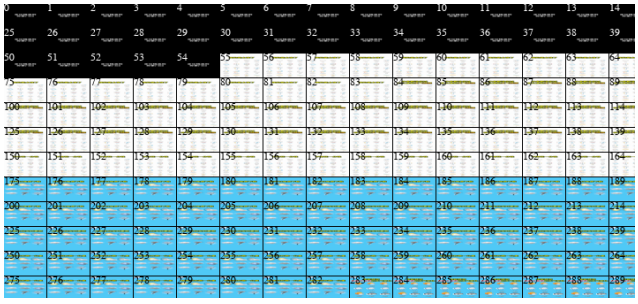


图 3. 我们收集的 SHOT 数据集中一个样例的缩略图。每帧的左上角显示帧编号。

在集合中，我们从剩余的 394 个由三位专家标注的短视频中随机选取 200 个作为测试集。

对于这 200 个测试短视频，我们进行了两轮标注，第一轮共标注出 2,616 个镜头。第二轮中，我们对标注结果进行了严格检查。除了修正少数无法通过我们手动设计的自动检查规则纠正的误报外，第二轮总共标注出 2,716 个镜头，即，从漏报中重新收集了 100 个镜头。

SHOT 数据集中的案例可见图 1左侧、图 2和图5。标注记录了每个镜头的起始帧和结束帧编号，在图粉色和浅色/白色区域中可视化呈现。5。所有用于训练和测试的短视频、镜头边界标注、评估指标脚本以及明确的视频级数据划分都将公开提供。

4. 自动镜头边界检测

4.1. 镜头边界检测搜索空间设计

我们基于先前最先进的方法之一 TransNetV 2 [30]，设计了一个 SuperNet，其中，特征表征学习网络可视为由六个因式分解膨胀深度 3D 卷积神经网络（DDCNNV2）模块依次构成。借助先进的 3D 卷积网络 [25] 和 Transformer [35]，，我们设计了一个镜头边界检测神经架构搜索空间。具体而言，AutoShot 在七个模块上进行架构搜索，并在第六个模块之后增加了自注意力层数量搜索。前六个模块中的搜索模块如图 4所示。我们首先在搜索空间中设计了四种因式分解 3D 卷积。设$\mathbf{x}$为当前模块的输入，则四种搜索模块可表示如下：

(1) DDCNNV2：我们对膨胀卷积分支的数量进行搜索 n_d，即，4 和 5；并对 2D 空间卷积的通道数进行搜索 n_c，即，分别为 TransNetV2 中原始 DD-CNNV2 模块输入通道数的 1、2 和 3 倍，如图4a 所示。DDCNNV2 可表示为

$$\begin{aligned}\mathbf{h} &= \mathrm{ReLU}(\mathrm{BN}(\mathrm{Concat}([\mathbf{h}_1, \mathbf{h}_2, \cdots, \mathbf{h}_{n_d}]))),\\ \mathbf{h}_i &= (\mathbf{T}_i \cdot \mathbf{S}_i) \cdot \mathbf{x} = \mathbf{T}_i(\mathbf{S}_i(\mathbf{x})),\ i = 1, \cdots, n_d,\end{aligned} \tag{1}$$

其中 $\mathbf{h}$ 是当前块的输出，$\mathbf{S}_i$ 是通道数为 $\lceil \frac{n_c}{n_d} \rceil$的 2D 空间卷积，$\mathbf{T}_i$ 是通道数为 $\lceil \frac{4F}{n_d} \rceil$、膨胀率为 2^{i-1}的 1 D 时间卷积，且 F 是图 4a中预定义的通道数。— DDCNNV2 的关键组件是膨胀时间 1D 卷积，以及由空间 2D 卷积和时间 1D 卷积构成的因式分解 3D 卷积。这种设计能够提取多样的时序上下文特征，并减少可学习参数的数量，从而有望降低过拟合。

(2) DDCNNV2A：为了统一空间 2D 卷积的特征提取器，我们可以采用共享的 2D 卷积，而不是多个空间 2D 卷积分支，如图 4b 所示。共享的空间 2D 卷积旨在提取统一的空间特征，以供后续多样化的时序特征提取使用。DDCNNV2A 可表示为

$$\begin{aligned}\mathbf{h} &= \mathrm{ReLU}(\mathrm{BN}(\mathrm{Concat}([\mathbf{h}_1, \mathbf{h}_2, \cdots, \mathbf{h}_{n_d}]))),\\ \mathbf{h}_i &= (\mathbf{T}_i \cdot \mathbf{S}) \cdot \mathbf{x} = \mathbf{T}_i(\mathbf{S}(\mathbf{x})),\ i = 1, \cdots, n_d,\end{aligned} \tag{2}$$

其中，$\mathbf{S}$ 是一个共享的 2D 空间卷积，其通道数为搜索得到的 n_c。

(3) DDCNNV2B：受伪 3D 网络设计的启发 [25]，，我们构建了另外两个搜索模块，以学习多种时空表示，如图4c 和4d所示。DDCNNV2B 可表示为

$$\begin{aligned}&\mathbf{h} = \mathrm{ReLU}(\mathrm{BN}((\mathbf{S} + \mathbf{T}) \cdot \mathbf{x})) = \mathrm{ReLU}(\mathrm{BN}(\mathbf{S}(\mathbf{x}) + \mathbf{T}(\mathbf{x}))),\\ &\mathbf{T}(\mathbf{x}) = \mathrm{Concat}([\mathbf{T}_1(\mathbf{x}), \mathbf{T}_2(\mathbf{x}), \cdots, \mathbf{T}_{n_d}(\mathbf{x})]).\end{aligned} \tag{3}$$

为确保空间特征和时间特征的通道数相等，2D 空间卷积的通道数固定为当前模块输入维度数的四倍 $4F$。

(4) DDCNNV2C：与 DDCNNV2B 不同，正如图 4d 所示，DDCNNV2C 的时间卷积仍可利用空间卷积的特征，其形式可表示为

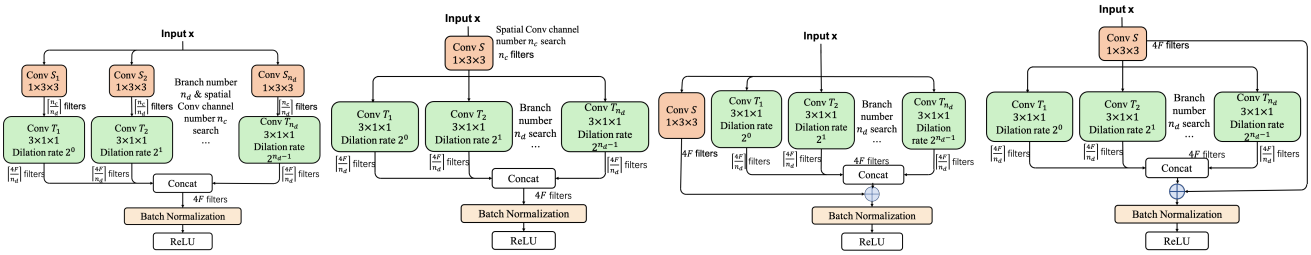
$$\begin{aligned}\mathbf{h} &= \mathrm{ReLU}(\mathrm{BN}((\mathbf{S} + \mathbf{T} \cdot \mathbf{S}) \cdot \mathbf{x}))\\ &= \mathrm{ReLU}(\mathrm{BN}(\mathbf{S}(\mathbf{x}) + \mathbf{T}(\mathbf{S}(\mathbf{x})))),\\ \mathbf{T}(\mathbf{S}(\mathbf{x})) &= \mathrm{Concat}([\mathbf{T}_1(\mathbf{S}(\mathbf{x})),\ \mathbf{T}_2(\mathbf{S}(\mathbf{x}))\ \cdots,\ \mathbf{T}_{n_d}(\mathbf{S}(\mathbf{x}))])。\end{aligned} \tag{4}$$

我们在六个因式分解卷积层之后进一步构建了一个一维时间 Transformer 模块，以增强时间建模能力。该模块的输入是逐帧卷积特征展平后的结果。我们在 Transformer 模块中搜索自注意力层的数量 $\{0, 1, 2, 3, 4\}$。

总而言之，AutoShot 包含七个搜索块。在前六个搜索块中，它针对 DDCNNV2 和 DDCNNV2A 搜索通道数以及分支数/膨胀率数。受逐元素加法中维度数一致性的限制，DDCNNV2B 和 DDCNNV2C 搜索分支数。因此，前六个搜索块中的每个搜索块都有 $3 \times 2 \times 2 + 2 \times 2 = 16$ 种选择。AutoShot 的搜索空间总共包含 $(16^6) \times 5 = 8.39 \times 10^7$ 个候选架构。

4.2. AutoShot 训练与搜索

构建用于表征学习的基础网络后，我们将基础网络的表征与原始输入帧的 RGB 直方图相似度，以及拼接块的可学习余弦相似度进行拼接



(a) DDCNNV2 (b) DDCNNV2A (c) DDCNNV2B (d) DDCNNV2C

图4. AutoShot 前六个块的搜索块示意图。它由四种因式分解的 3D 卷积块组成：(a) DDCNNV2，先进行 2D 空间卷积，再进行 1D 时间卷积；(b) DDCNNV2A，采用共享的 2D 空间卷积和 1D 时间卷积；(c) DDCNNV2B，累积 2D 空间卷积和 1D 时间卷积；(d) DDCNNV2C，兼顾 DDCNNV2A 和 DDCNNV2B。

特征 [5, 30] 以构成一个 4,864 维特征向量。随后添加一个包含 1,024 个神经元并采用 ReLU 激活的全连接层 [23] 。接下来，在最终的两个逐帧分类头之前使用一个 Dropout 层 [17]，其丢弃率为 0.5，用于对过渡中的单个中间帧进行分类 y 以及对所有过渡进行分类 z。

AutoShot 的 SuperNet 训练采用高效的权重共享策略。权重共享策略 [9, 20]将搜索空间编码到 SuperNet 中，所有候选架构共享 SuperNet 的权重。单次神经架构搜索 [9] 将 SuperNet 训练与架构搜索解耦，从而获得更高的准确率。我们还依次执行 SuperNet 训练和架构搜索两个步骤。我们采用单路径和均匀采样策略，以减少节点权重 [9]之间的共同适应。

我们为 AutoShot 实现了一种基于贝叶斯优化的架构搜索方法。贝叶斯优化通过拟合概率代理模型，并通过最大化采集函数来确定下一步要评估的配置，从而不断迭代。我们采用带有汉明核的高斯过程作为代理函数。在初始化阶段，我们进行随机探索，以获得良好的高斯过程模型。对于采集函数，我们使用可行概率 [8]。

$$\boldsymbol{p}|\boldsymbol{a} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{K}), \quad \text{Acc}|\boldsymbol{p}, \sigma^2 \sim \mathcal{N}(\boldsymbol{p}, \sigma^2 \boldsymbol{I}), \tag{5}$$

其中，变量 $\boldsymbol{p}$ 服从联合高斯分布， $\boldsymbol{a}$ 是一组已观测到的架构，汉明核 $K_{ij} = k(a_i, \ a_j)$，Acc 表示这些采用权重共享的架构经评估得到的准确率指标，即 F1值或在召回率固定时的精确率。GP 的参数 $\boldsymbol{\mu}$ 和 σ可通过最大化边际对数似然进行估计。

对于候选架构的重新训练，我们首先采用相同的两种分类交叉熵损失：即，过渡的单个中间帧 y和整个过渡 z，分别作为

SuperNet 训练

$$\mathcal{L}_{retrain} = -\sum_{i=1}^{N}\sum_{j=1}^{N_F}\left[\lambda_1 y_{i,j}\log\hat{y}_{i,j} + \lambda_2 z_{i,j}\log\hat{z}_{i,j}\right], \tag{6}$$

其中， N 是随机梯度下降（SGD）中的批量大小， N_F 是每个训练批次处理的每个训练样本所包含的预定义帧数，λ_1 和 λ_2 是两个分类头之间的权衡系数，而 $\hat{}y_{i,j}$ 和 $\hat{}z_{i,j}$ 分别是对过渡的单个中间帧和整个过渡的两种逐帧预测。

使用式（6）中的普通多头交叉熵分类损失进行重新训练后，我们进一步采用表现最佳的候选网络作为教师网络进行知识蒸馏 [13],，并利用权重嫁接 [22] 进一步提升最佳准确率。知识蒸馏用于使候选网络达到目标准确率，而权重嫁接则在对齐后的网络之间自适应地平衡嫁接信息，从而提升表征能力并提高准确率。基于熵的权重嫁接可表示为

$$\begin{aligned}
&\mathcal{L}(W) = -\sum_{i=1}^{N}\sum_{j=1}^{N_F}\left[\lambda_1\tilde{y}_{i,j}\log\hat{y}_{i,j} + \lambda_2\tilde{z}_{i,j}\log\hat{z}_{i,j}\right], \\
&\alpha = A \times (\arctan(c \times (H(W_l^{M_2}) - H(W_l^{M_1})))) + 0.5, \\
&W_l^{M_2} = \alpha W_l^{M_2} + (1-\alpha)W_l^{M_1},
\end{aligned} \tag{7}$$

其中，$\tilde{y}_{i,\ j}$ 和 $\tilde{z}_{i,\ j}$ 分别是教师网络对过渡中单个中间帧和所有过渡的两个逐帧预测， A 和 c是固定超参数，α 是基于熵的系数，用于平衡网络， H（·）是基于网络权重 $W_l^{M_1}$或 $W_l^{M_2}$的分箱计算的熵， l 是网络中的层， M_1 和 M_2 是两个网络，其中网络 M_2 接收来自网络 M_1的信息。

5. 实验

我们根据收集的 SHOT 上的评估指标进行神经架构搜索，并在 SHOT、ClipShots [32], BBC [1], 和 RAI [2]上验证搜索得到的最优架构的有效性。在 SuperNet 训练和候选网络再训练中，我们通过随机拼接两个镜头来构造每个训练样本，每个训练样本中的帧数 N_F 设为 60。超参数 λ_1， λ_2， A， c、熵计算中的分箱数量以及式（7）中的嫁接网络数量分别设为 5 、 0.1 、 0.4 、 1.0 、 10 和 3，这一设置通常沿用先前工作的做法 [22,30]。我们采用随机梯度下降，学习率为 0.1、动量为 0.9、批量大小为 16，训练轮次为 12。搜索过程中，每轮的种群数量为 48，总训练轮次为 100，其中初始化占 20 轮。我们使用 Open-Box 库 [19] 实现搜索中的贝叶斯优化。SuperNet 和候选网络训练均使用一块 32 GB NVIDIA Tesla V100 GPU，并使用八块 32 GB NVIDIA Tesla V100 GPU 加速搜索。所有代码、数据和训练好的模型都将公开，以确保可复现性。

在 SHOT 上进行神经架构搜索 我们在 SHOT 和 Clip- Shots 的合并训练集上训练 SuperNet，并根据收集到的 SHOT 数据集上的评估指标进行搜索。为确保公平比较，我们严格遵循以往工作的数据集协议和指标计算方法 [2,30]。此外，我们还在固定召回率 0.71（与 TransNetV2 相同）的条件下，根据精确率指标进行搜索，因为在实际应用中，较高的 F1 指标无法保证精确率和召回率同时更高。根据表 2 左侧，AutoShot 根据 F1 和精确率指标分别比 TransNetV2 高出 4.2% 和 3.5%。值得注意的是，PySceneDe- tect [3]在 SHOT 上的 F1分数低于 0.6，远逊于 AutoShot，而且它几乎无法处理具有挑战性的渐变转场。对于 AutoShot，我们发现，54% 的漏检镜头属于渐变转场，而渐变转场占 SHOT 中镜头的 30%。渐变转场的帧间差异不大，因此难以被完全检测出来。下文中，为保持一致性，我们仅根据 F1 指标将 AutoShot 与其他方法进行比较。

为了直观比较预测结果，我们使用视频缩略图来清晰展示差异。真实边界显示为粉色或浅色/白色，如图所示。 5。AutoShot 检测到的边界标记为粉色、青色或浅色，而 TransNetV2 检测到的边界以青色或浅色显示。AutoShot 能够成功检测细微镜头转场，如粉色所示，而 TransNetV2 未能检测到。对于明显转场，TransNetV2 和 AutoShot 均能成功检测，如浅色所示。



图5. 真实标注（粉色、浅色/白色）、TransNetV2（青色、浅色/白色）和 AutoShot（粉色、青色、浅色/白色）在 SHOT 数据集的四个片段上的镜头边界可视化比较。AutoShot 检测到了微小转场，以粉色表示。浅色/白色表示 TransNetV2 和 AutoShot 均成功检测到镜头边界。青色表示 TransNetV2 和 AutoShot 均产生了假阳性，这可能需要模型自适应地理解视频中的上下文语义。

TransNetV2 和AutoShot 均以 青色显示，它们是困难负镜头。减少假阳性可能需要自适应地理解整个视频中的上下文语义，而有些假阳性即使对人工标注者来说也很模糊。

其他数据集上的泛化能力 从 SHOT 数据集获得最优网络架构后，我们在另外三个公开且广泛使用的数据集 ClipShots [32], BBC [1], 和 RAI [2]上验证网络的泛化能力。这三个现有数据集与我们的 SHOT 数据集有很大差异，如第 3 节所示。BBC《地球脉动》纪录片系列 [1] 由 BBC教育电视系列节目《地球脉动》的 11 集组成。每集约 50 分钟，整个数据集包含约 4900 个镜头和 670 个场景。RAI [2] 数据集基于从 Rai Scuola视频档案中随机选取的十个广播视频，每个视频时长约半小时。ClipShots 收录了数千个来自 YouTube 的普通在线视频，而非短视频；与 BBC 和 RAI 相比，这是一个更具挑战性的数据集。我们采用与先前工作相同的数据集划分和协议 [2, 30]。

从表 2 可见，只需将搜索得到的AutoShot架构应用于三个数据集，就能始终获得比以往最先进方法更高的F1 分数，这充分验证了其有效性以及

方法	TransNetV2	AutoShot @F1	AutoShot @精确率
F1	0.799	**0.841**	0.826
精确率	0.904	0.923	**0.939**

数据集	ClipShots	BBC	RAI
DSMs	0.761	0.893	0.928
ST卷积网络	0.759	0.926	0.939
TransNet	0.735	0.929	0.943
TransNetV2	0.776	0.962	0.939
AutoShot	**0.787**	**0.971**	**0.955**

表 2。左：在召回率固定为与 TransNetV2 相同的情况下，根据 F1 分数和精确率指标，AutoShot 分别以 4.2% 和 3.5% 的优势超过 TransNetV2。右：搜索得到的最优架构在三个广泛使用的镜头边界检测数据集上得到验证。AutoShot 相较于深度显著性图始终取得最佳 F1值 [32], ST卷积网络 [11], TransNet [21] 和 TransNetV2 [30]。最佳 F1 分数以**粗体**表示。

AutoShot具有良好的泛化能力。特别是，AutoShot在ClipShots、BBC和RAI上的表现分别比以往的最先进方法高出1.1%、0.9%和1.2%。需要注意的是，我们基于PyTorch复现了TransNetV2，并在ClipShots上获得了0.776的F1分数，而原论文 [30]报告基于TensorFlow获得了0.779。

搜索空间的影响 我们研究了三种不同搜索空间的影响，即，AutoShot-S、-M 和 -L，见表 3 （左）。AutoShot-S 仅在搜索空间中使用 DDCNNV2A 组件，每个块有六种搜索选项。AutoShot-M 在搜索空间中使用 DDCNNV2 和 DDCNNV2A 组件，每个块有 12 种搜索选项。AutoShot-L 表示第 4.1 节定义的搜索空间，它取得了最佳 F1分数。将多种 3D ConvNet 变体（即 DDCNNV2、DDCNNV2A、B 和 C）组合到每个搜索块中，可以提高重新训练和搜索后候选架构的 F1分数，因为不同层的最优块各不相同，而 AutoShot 中更多的搜索选项有助于找到最优组合。

搜索到的架构具体而言，基于 F1分数 得到的架构 AutoShot@F1 为 DDCNNV2{(n_c=4F, n_d=4), A(n_c=4F, n_d=5), A(n_c=4F,n_d=5), A(n_c=4F, n_d=5), (n_c=12F, n_d=5), (n_c=8F, n_d=5)}，其浮点运算量（FLOPs）为 37 GMACs。基于精确率指标得到的架构 AutoShot@Prec. 为 DDCNNV2{(n_c=12F, n_d=4), (n_c=8F, n_d=4), B(n_d=4), C(n_d=4), B(n_d=5), B(n_d=4)}，其 FLOPs 为 30 GMACs。我们发现，最优架构确实采用了多种不同的模块，且运算次数少于 FLOPs 为 41 GMACs 的 TransNetV2。在 SHOT 上进行搜索会选择更多的膨胀卷积分支 n_d 和多种不同的模块，因为更多的膨胀卷积分支 n_d 和多种不同的模块能够增强模型针对各种时间粒度镜头转换的表征学习能力，而这类转换在短视频中大量存在。这可能也是 AutoShot 在 SHOT 上取得的提升远大于其在传统视频数据集上的提升的另一个原因。尽管这两种最优架构均未使用自注意力层，但 SuperNet 的训练以及 F1分数或精确率分数接近的架构都会使用自注意力。

AutoShot-	S	M	L
不重新训练	0.816	0.822	0.831
重新训练	0.833	0.837	**0.841**

方法	不使用知识蒸馏	使用知识蒸馏	使用知识蒸馏+权重嫁接
F1	0.825-0.837	0.832-0.838	**0.841**

表 3. 在收集的 SHOT 数据集上，不同搜索空间（左）、知识蒸馏（KD）和权重嫁接（右）的 F1 分数。最佳 F1 分数以**粗体**标示。

知识蒸馏和权重嫁接的效果我们基于第3 节（右）构建的搜索空间，对知识蒸馏和权重嫁接进行了消融实验，详见第4.1节。去掉式（6）中的知识蒸馏后，重新训练得到的 F1 分数范围为 0.825-0.837。我们选取性能最佳的架构（即 F1 分数为 0.837）作为教师网络，而知识蒸馏使候选架构的 F1 分数达到 0.832-0.838。随后，式（7）中的权重嫁接进一步将最佳 F1 分数提高了 0.3%。

6. Conclusion

在这项工作中，我们收集了一个新的短视频镜头边界检测数据集，名为 SHOT，它与基于传统长视频的镜头边界检测属于截然不同的场景。SHOT 可以显著加快各种基于短视频的应用的开发和评估，例如.,智能创作，以及视频场景分割与理解。借助这一新资源，我们通过在涵盖各种先进3D 卷积网络和 Transformer 的搜索空间中进行神经架构搜索，针对视频镜头边界检测任务优化模型设计。AutoShot在F1值和精确率分数方面分别超越了此前最佳的镜头边界检测方法 4.2% 和 3.5%。我们进一步在ClipShots、BBC 和 RAI 上验证了所搜索最优架构的泛化能力。实验结果表明，AutoShot 具有良好的泛化能力，并且在现有的三个公开数据集上优于此前最先进方法。

参考文献

[1] Lorenzo Baraldi, Costantino Grana, and Rita Cucchiara。用于广播视频场景检测的深度孪生网络。载于第23届ACM国际多媒体会议论文集，第1199–1202页，2015年。1，2，3，7[2] Lorenzo Baraldi, Costantino Grana, and Rita Cucchiara。通过层次聚类进行镜头与场景检测，以便重复利用广播视频。载于国际图像与模式计算分析会议论文集，第801–811页。施普林格，2015年。1，2，3，7[3]Brandon Castellano。PySceneDetect。http:// www . bcastell.com/projects/PySceneDetect，2022年。3，7[4] Saptarshi Chakraborty, Alok Singh, and Dalton Meitei Thounaojam。一种新颖的双阶段镜头边界检测算法：对运动和光照具有不变性。The Visual Computer，38(2):445–456，2022年。2[5] Vasileios Chasanis, Aristidis Likas, and Nikolaos Galat- sanos。使用支持向量机同时检测视频中的突变切换和叠化。Pattern Recognition Letters，30(1):55–65，2009年。6[6] Shixing Chen, Xiaohan Nie, David Fan, Dongqing Zhang, Vimal Bhat, and Raffay Hamid。用于场景边界检测的镜头对比式自监督学习。载于IEEE/CVF计算机视觉与模式识别会议论文集，第9796–9805页，2021年。1，3[7] Yuantao Feng, Shiqi Yu, Hanyang Peng, Yan-Ran Li, and Jianguo Zhang。高效人脸检测：综述与评估。IEEE生物特征、行为与身份科学汇刊，4(1):1–18，2021年。4[8] Jacob R Gardner, Matt J Kusner, Zhixiang Eddie Xu, Kil- ian Q Weinberger, and John P Cunningham。带不等式约束的贝叶斯优化。载于国际机器学习会议论文集，第2014卷，第937–945页，2014年。6[9] Zichao Guo, Xiangyu Zhang, Haoyuan Mu, Wen Heng, Zechun Liu, Yichen Wei, and Jian Sun。采用均匀采样的单路径单次神经架构搜索。载于欧洲计算机视觉会议论文集，第544–560页。施普林格，2020年。2，3，6[10] Michael Gygli。使用全卷积神经网络实现极快的镜头边界检测。载于2018年基于内容的多媒体索引国际会议（CBMI）论文集，第1–4页。IEEE，2018年。3[11]Ahmed Hassanien, Mohamed Elgharib, Ahmed Selim, Sung-Ho Bae, Mohamed Hefeeda, and Wojciech Ma- tusik。通过时空卷积神经网络实现大规模、快速且准确的镜头边界检测。arXiv预印本arXiv:1705.03281，2017年。1，3，8[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun。用于图像识别的深度残差学习。载于IEEE计算机视觉与模式识别会议论文集，第770–778页，2016年。1，3[13] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean。提炼神经网络中的知识。载于第28届神经信息处理系统国际会议深度学习研讨会论文集，2014年。6

[14] Sergey Ioffe 和 Christian Szegedy。批归一化：通过减少内部协变量偏移加速深度网络训练。载于《国际机器学习会议论文集》，第448–456页。PMLR，2015年。1, 3[15] Qing-Yuan Jiang、Yi He、Gen Li、Jian Lin、Lei Li 和 Wu-Jun Li。Svd：用于近似重复视频检索的大规模短视频数据集。载于《IEEE/CVF国际计算机视觉会议论文集》，第5281–5289页，2019年。3[16] Xuekun Jiang、Libiao Jin、Anyi Rao、Linning Xu 和 Dahua Lin。联合学习镜头的属性与构成，以检测视频中的边界。IEEE多媒体汇刊，2021年。2[17] Alex Krizhevsky、Ilya Sutskever 和 Geoffrey E Hinton。使用深度卷积神经网络进行ImageNet分类。载于《神经信息处理系统进展论文集》，第25卷，第1097–1105页，2012年。6[18] Hilde Kuehne、Ali Arslan 和 Thomas Serre。动作的语言：恢复目标导向型人类活动的句法与语义。载于《IEEE计算机视觉与模式识别会议论文集》，第780–787页，2014年。2[19] Yang Li、Yu Shen、Wentao Zhang、Yuanwei Chen、Huaijun Jiang、Mingchao Liu、Jiawei Jiang、Jinyang Gao、Wentao Wu、Zhi Yang 等。Openbox：通用黑盒优化服务。载于第27届SIGKDD知识发现与数据挖掘会议（KDD-2021）论文集, 2021年。7[20] Hanxiao Liu、Karen Simonyan 和 Yiming Yang。Darts：可微分架构搜索。载于《国际学习表征会议论文集》，2018年。6[21] Jakub Loko˘c、Gregor Koval˘cik、Tom´a˘s Sou˘cek、Jaroslav Moravec 和 P˘remysl ˘Cech。用于视频中有效已知项目检索的框架。载于第27届ACM国际多媒体会议论文集，第1777–1785页，2019年。1, 3, 8 – 。[22] Fanxu Meng、Hao Cheng、Ke Li、Zhixin Xu、Rongrong Ji、Xing Sun 和 Guangming Lu。深度神经网络的滤波器嫁接。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第6599–6607页，2020年。6, 7[23] Vinod Nair 和 Geoffrey E Hinton。修正线性单元提升受限玻尔兹曼机的性能。载于《国际机器学习会议论文集》，2010年。6[24] Hanyang Peng 和 Shiqi Yu。一种系统性的IoU相关方法：超越简化回归，实现更佳定位。IEEE图像处理汇刊，30：5032–5044，2021年。4[25] Zhaofan Qiu、Ting Yao 和 Tao Mei。利用伪3D残差网络学习时空表征。载于《IEEE国际计算机视觉会议论文集》，第5533–5541页，2017年。2, 3, 5[26] Anyi Rao、Linning Xu、Yu Xiong、Guodong Xu、Qingqiu Huang、Bolei Zhou 和 Dahua Lin。一种用于多模态电影场景分割的由局部到全局方法。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第10146–10155页，2020年。1, 3

[27] BS Rashmi 和 HS Nagendraswamy。基于分块累积方法的视频镜头边界检测。多媒体工具与应用，80(1)：641–664，2021年。2[28] Bobak Shahriari、Kevin Swersky、Ziyu Wang、Ryan P Adams 和 Nando De Freitas。让人类退出循环：贝叶斯优化综述。IEEE会刊，104(1)：148–175，2015年。2, 6[29] Mike Zheng Shou、Stan W Lei、Weiyao Wang、Deepti Ghadi-yaram 和 Matt Feiszli。通用事件边界检测：事件分割基准。载于《IEEE/CVF国际计算机视觉会议论文集》，2021年。1, 3[30] Tom´aˇs Souˇcek 和 Jakub Lokoˇc。Transnet v2：用于快速镜头转场检测的有效深度网络架构。arXiv预印本，arXiv:2008.04838，2020年。1, 3, 4, 5, 6, 7, 8[31] Sebastian Stein 和 Stephen J McKenna。结合嵌入式加速度计与计算机视觉，以识别食物准备活动。载于2013年ACM国际普适与泛在计算联合会议论文集，第729–738页，2013年。2[32] Shitao Tang、Litong Feng、Zhanghui Kuang、Yimin Chen 和 Wei Zhang。利用深度结构化模型快速定位视频镜头转场。载于《亚洲计算机视觉会议论文集》，第577–592页。施普林格，2018年。1, 2, 3, 7, 8[33] Yansong Tang、Dajun Ding、Yongming Rao、Yu Zheng、Danyang Zhang、Lili Zhao、Jiwen Lu 和 Jie Zhou。Coin：用于全面分析教学视频的大规模数据集。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第1207–1216页，2019年。2[34] Du Tran、Lubomir Bourdev、Rob Fergus、Lorenzo Torresani 和 Manohar Paluri。利用3D卷积网络学习时空特征。载于《IEEE国际计算机视觉会议论文集》，第4489–4497页，2015年。3 – 。[35] Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszko-reit、Llion Jones、Aidan N Gomez、Łukasz Kaiser 和 Illia Polosukhin。注意力就是你所需要的一切。载于《神经信息处理系统进展论文集》，第5998–6008页，2017年。2, 3, 5[36] Zhenzhi Wang、Zhimin Li、Liyu Wu、Jiangfeng Xiong 和 Qinglin Lu。腾讯多模态广告视频理解概述。载于第29届ACM国际多媒体会议论文集，第4725–4729页，2021年。1, 3[37] Saining Xie、Chen Sun、Jonathan Huang、Zhuowen Tu 和 Kevin Murphy。重新思考时空特征学习：视频分类中的速度与准确率权衡。载于《欧洲计算机视觉会议（ECCV）论文集》，第305–321页，2018年。1, 3[38] Fisher Yu 和 Vladlen Koltun。通过空洞卷积进行多尺度上下文聚合。载于《国际学习表征会议论文集》，2016年。1, 3[39] Xuefan Zha、Wentao Zhu、Lv Xun、Sen Yang 和 Ji Liu。用于时空表征学习的移位块Transformer。《神经信息处理系统进展》，34：11384–11396，2021年。1, 2

[40] Wentao Zhu、Yufang Huang、Liang Zeng、Xuming Chen、Yong Liu、Zhen Qian、Nan Du、Wei Fan 和 Xiaohui Xie。Anatomynet：用于头颈部解剖结构快速全自动全容积分割的深度学习。*Medical Physics*，46(2):576–589，2019年。2[41] Wentao Zhu、Tianlong Kong、Shun Lu、Jixiang Li、Dawei Zhang、Feng Deng、Xiaorui Wang、Sen Yang 和 Ji Liu。Speechnas：在大规模说话人验证中实现延迟与准确率更佳权衡的方法。载于*2021*年 *IEEE* 自动语音识别与理解研讨会（*ASRU*），第1102–1109页。IEEE，2021年。2[42] Wentao Zhu、Cuiling Lan、Junliang Xing、Wenjun Zeng、Yanghao Li、Li Shen 和 Xiaohui Xie。利用正则化深度 LSTM 网络进行基于骨架的动作识别的共现特征学习。载于*AAAI*人工智能会议论文集，第30卷，2016年。1[43] Wentao Zhu、Chaochun Liu、Wei Fan 和 Xiaohui Xie。DeepLung：用于自动肺结节检测与分类的深度 3D 双路径网络。载于*2018*年 *IEEE* 冬季计算机视觉应用会议（*WACV*），第673–681页。IEEE，2018年。2[44] Wentao Zhu、Jingya Liu 和 Yufang Huang。Hnssl：基于困难负样本的自监督学习。载于*IEEE/CVF*计算机视觉与模式识别会议研讨会（*CVPRW*）论文集，2023年。3[45] Wentao Zhu、Yeeleng S Vang、Yufang Huang 和 Xiaohui Xie。DeepEM：用于弱监督肺结节检测的结合期望最大化的深度 3D 卷积网络。载于医学图像计算与计算机辅助手术——*MICCAI 2018*：第*21*届国际会议，西班牙格拉纳达，*2018*年*9*月*16*日至*20*日，论文集，第*II*部分*11*，第812–820页。施普林格，2018年。2

arXiv:2304.06116v1 [cs.CV] 12 Apr 2023

AutoShot：一个短视频数据集与最先进的镜头边界检测

Wentao Zhu[1],Yufang Huang[3], Xiufeng Xie[1], Wenxian Liu[1],Jincan Deng[1], Debing Zhang[1], ZhangyangWang[2], JiLiu[11]快手科技 [2]德克萨斯大学奥斯汀分校 [3]康奈尔大学

1. SHOT 数据集

为加速短视频相关镜头边界检测的研究，我们从最广泛使用的短视频平台之一收集了 853 个短视频。数据集属性对比如表 1 所示。帧总数为 960,794，接近一百万帧。逐帧镜头边界标注是一项繁重的任务。

受 TransNetV 中帧处理方式的启发2 [，], 我们通过将每帧调整大小为 48×27 如图1所示，开发了一种基于视频缩略图的标注方法。每帧的编号会自适应地显示在帧的左上角，从而显著减少标注人员核对帧编号的工作量。如果帧编号所在位置的像素值较暗，我们就以浅色显示帧编号；否则，则以深色显示帧编号。我们有三位专家和一个标注团队，共同完成了对960,794帧的全部标注。标注团队对 459 个短视频进行了标注，共获得 6,111个镜头。标注完成后，我们对这6,111条标注进行了专家检查。我们随机选取了200条镜头标注，发现这6,111条镜头标注中存在2%的错误率。考虑到镜头定义的模糊性以及标注人员之间的标注差异，我们认为6,111 个镜头的 2% 错误率是可以接受的。

测试集上的标注质量直接影响基准方法的准确评估，即，也影响短视频数据集的质量。为保证测试集上镜头边界标注的高质量，我们从其余由三位专家标注的394个短视频中随机选取200个短视频作为测试集。专家团队是负责定义镜头转换并编写标注指南的制作团队。随后，标注指南和数据被发送给承包商，由其使用缩略图进行标注，制作团队负责开展质量审核。由于测试集的重要性，全部测试集均由制作团队完成标注。对于这200个测试短视频，我们采用两轮标注，第一轮共得到2,616个镜头。在第二轮中，我们对标注进行了严格检查。除了修正少量的误报标注之外，这些标注可能-

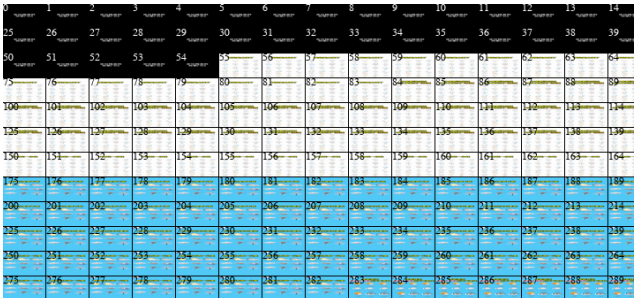


图1。我们收集的SHOT 数据集中的一个案例的缩略图。每帧的帧编号显示在帧的左上角。

数据集	BBC	RAI	ClipShots	SHOT
复杂场景渐变转场	✗	✗	✗	✓
虚拟场景	✗	✗	✗	✓
三元视频	✗	✗	✗	✓
平均视频长度（秒）	2944.8	590.5	236.5	39.5
平均镜头长度（秒）	6.57	5.65	15.34	2.59

表1. 不同镜头边界检测数据集的比较。

无法通过我们为自动检查设计的规则手动纠正，第二轮共得到 2,716 个镜头，即，从假阴性中重新收集 100 个镜头。

在该数据集中，每个短视频都有一个与视频同名的文本文件，其中包含镜头边界标注，即，也就是每行记录每个镜头的起始帧号和结束帧号，如下图2所示。若没有渐变转场（大多数镜头都是如此），则下一个镜头的起始帧号应为当前镜头结束帧号的下一帧。例如，第一个镜头的结束帧号为 72，第二个镜头的起始帧号为 73。若存在渐变转场，则起始帧号处为明确的转场，而上一镜头的结束帧号与当前镜头起始帧号之间的间隔就是渐变转场。例如，

```
0 72
73 102
109 127
128 135
136 175
176 206
207 229
230 252
253 264
265 301
302 384
385 422
423 443
444 542
543 640
641 716
```

图 2. 镜头边界标注文本文件。每行表示一个镜头的起始帧号和结束帧号。

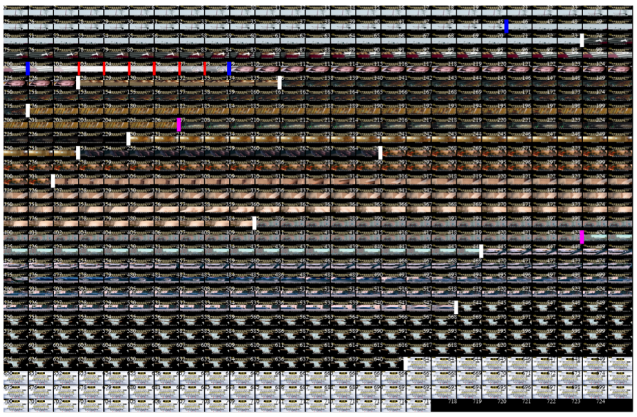


图 3. 与图 2所示相同视频的缩略图。浅色、红色和粉色标记真实标注的镜头边界标注。黄色圆圈表示从第 102 帧到第 108 帧的一次渐变转场。

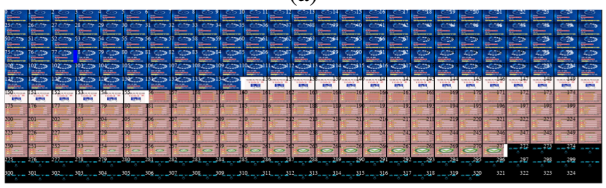
第 2 行的镜头结束帧号为 102，如图2所示；第 3 行的镜头起始帧号为 109。102 到 108 帧之间的帧属于渐变转场，如图3中的黄色圆圈所示。图3 是整个视频的缩略图，与图2相同，浅色、红色和粉色区域标示了真实标注的镜头边界。

为公平评估镜头边界检测，我们遵循 [?]中 ClipShots 数据集的评估方法。如果检测到的结束帧编号与标注相差不超过两帧，则该检测到的边界为真正例。对于渐变转场，如果检测到的结束帧编号位于渐变转场范围内，也视为真正例。指标评估脚本可在https://github.com/soCzech/TransNetV2/blob/master/training/metrics_utils.py#L26中找到。

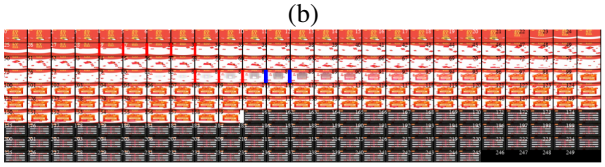
2. 更多视觉比较结果

我们在补充材料中添加了更多图示，用于比较图2 [?] 中的 AutoShot（蓝色、浅粉色、青色）、TransNetV4（绿色、黄色、浅青色）预测结果与真实标注（红色、浅粉色、黄色）。浅色表示 TransNetV2 和 AutoShot 在该帧都检测为真正例，如图4 （a）的视频缩略图所示。图4 （b）中的蓝色切点表明，AutoShot 成功检测到视频中段的轻微镜头转场（与真实标注相差不超过两帧）。图4 （c）展示了一个垂直三元结构的视频，其中场景变化轻微，因此轻微镜头转场很难检测。图4 （d）中的粉色切点表明，AutoShot 检测到了镜头转场，而 TransNetV2 未能检测到。图4（d）第一帧中的蓝色表示 AutoShot 出现了假阳性检测；从视觉上看，视频开头确实存在渐变转场。

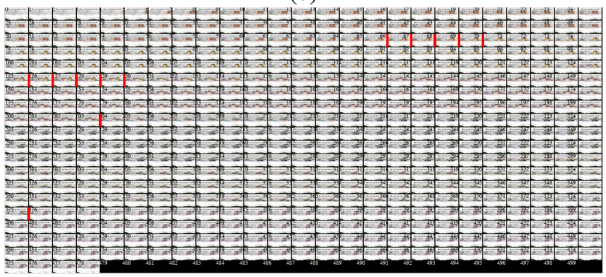
(a)



(b)



(c)



(d)

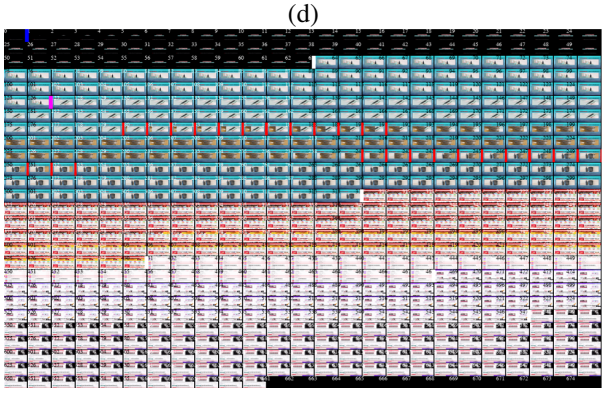


图4。在收集的 SHOT 数据集上，对真实标注（红色、浅色、粉色、黄色）、TransNetV2（绿色、黄色、浅色、青色）和 AutoShot（蓝色、粉色、浅色、青色）的镜头边界进行视觉比较。AutoShot 能够检测微小转场，以粉色显示。浅色表示 TransNetV2 和 AutoShot 均成功检测到镜头边界。