arXiv:2303.03105v2 [cs.MM] 7 Mar 2023

基于置信度的事件中心在线视频问答：在新构建的 ATBS 数据集上

Weikai Kong[⋆‡]、*Shuhong Ye*[⋆‡]、*Chenglin Yao*[⋆]、*Jianfeng Ren*[⋆†]

[⋆]宁波诺丁汉大学计算机科学学院
[†]宁波诺丁汉大学卓越研究与创新灯塔研究院，宁波诺丁汉大学，中国

摘要

深度神经网络推动了视频问答（VideoQA）的发展，但闭路电视和直播等视频流的实际应用对求解器提出了更高要求。为应对时长未知的长视频所带来的视频问答挑战，我们定义了一类新问题，称为在线开放式视频问答（O[2]VQA）。该问题要求求解器具备在线状态更新机制，以判断已收集的信息是否足以得出答案。随后，我们提出基于置信度的事件中心在线视频问答（CEO-VQA）模型来解决这一问题。此外，我们构建了一个名为背景流中的答案目标（ATBS）的数据集，用于评估这一新开发的在线视频问答应用。与需要观看完整视频的基线视频问答方法相比，实验结果表明，所提方法取得了显著的性能提升。

索引词— 在线视频问答、视频理解、开放式视频问答、视频问答

1. 引言

视频问答（VideoQA）有益于许多现实应用，例如智能对话 [1], 多媒体推荐 [2], 以及网络视频摘要[3]。在开放式视频问答任务中，给定一段视频和一个相关问题，模型需要给出正确答案。近期方法 [4–7] 通常采用注意力机制来建模视频与问题之间的交互。

大多数现有模型都专注于回答有关固定长度短视频的问题 [4–11]。但在现实场景中，视频可能来自时长难以预料的闭路电视或直播。在线视频问答面临三项挑战：1）信息量过大，导致关键特征被稀释；2）干扰事件可能使求解器困惑；3）缺少一种机制来确认所收集的信息是否足以回答问题。我们将其建模为 **在**线**开**放式**视**频**问**题**解**答（[2]视频问答）问题，如图1所示。给定一段长度难以预料的视频和一个相关问题，求解器应在线处理视频，在收集到足够信息时终止，并用便于人类理解的句子回答问题。

为解决这些问题，我们提出了一种基于置信度的事件中心在线视频问答（CEO-VQA）模型，用于开放式[2]视频问答任务。该模型由两个主要模块组成：1）基于置信度的目标事件定位器，用于定位与回答问题最相关的目标事件；2）目标事件问答模型，利用视觉和语言线索回答问题。具体而言，模型采用BridgeFormer [12] 来计算语言特征与视觉特征之间的相似度，并据此更新置信度分数。随后，基于置信度的目标事件定位器会过滤无关事件并定位目标事件。最后，将视觉特征和语言特征整合到目标事件问答模型中，并输入答案解码器，以得出正确答案。

为便于评估这一新开发的开放式视频问答任务，我们构建了一个大型视频数据集。我们采用背景 + 目标方法生成新数据集**答**案**目**标**背**景**流**（ATBS），以模拟复杂的多事件视频流。具体而言，我们将 MSRVTT 数据集中的 1 万个目标 视频片段[13] 分别插入 DiDeMo 数据集的 10,464 个背景 视频中的一个 [14]，并将目- 标视频的问答对作为标注。因此，构建的数据集共包含 1 万个视频和 24.4 万个已标注问答对。实验结果表明，所提出的数据集比现有视频问答数据集更具挑战性 [13, 15]。

我们的贡献有三点：1）为推动视频问答研究，我们提出了一项新颖且具有挑战性的视频问答任务——在线开放式视频问答（O[2]VQA）；2）我们构建了一个大型 ATBS 数据集，以促进 O[2]VQA 任务的研究；3）我们提出了 CEO-VQA 模型，以更有效地解决 O[2]VQA 问题。

‡ 作者贡献相同。本研究部分得到了中国国家自然科学基金项目（批准号：72071116）的资助，部分得到了宁波市科学技术局项目（批准号：2019B10026和2022Z173）的资助。通讯作者：jianfeng.ren@nottingham.edu.cn。

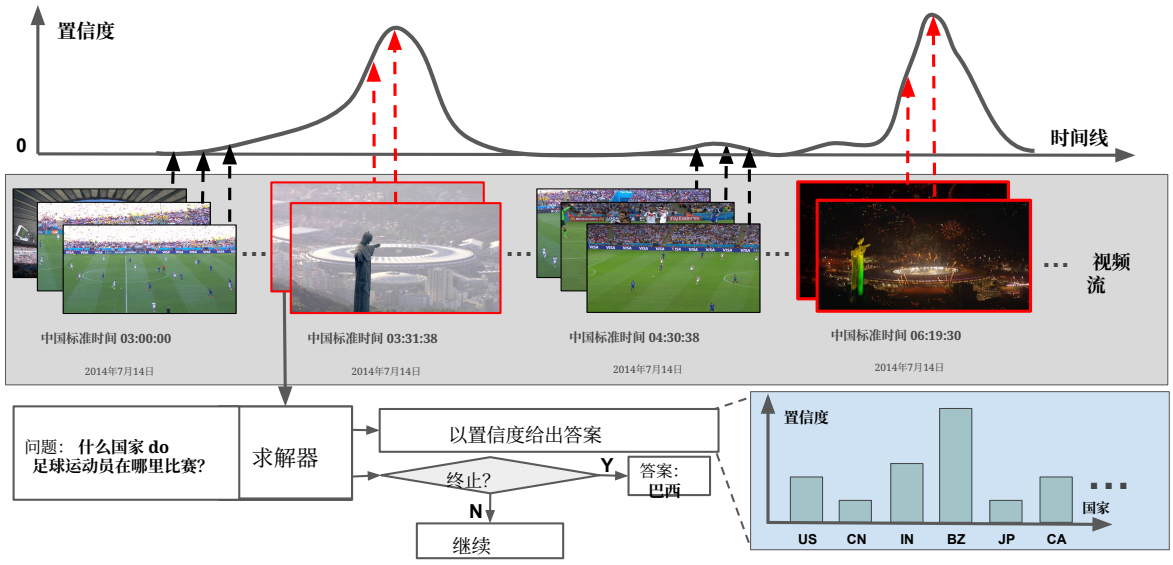

图 1。在线开放式视频问答（O²VQA）任务示意图。对于每个视频帧，求解器会为每个可行的候选答案计算置信度分数，并判断证据是否充分，足以给出有把握的答案。

2. 所提方法

2.1. 所提方法概述

为应对 O²VQA 面临的挑战，我们提出了一种基于置信度的事件中心在线视频问答（CEO-VQA）模型。主要流程如算法1所示。所提出的 CEO-VQA 包含两个主要模块：1）基于置信度的目标事件定位器。给定一个在线视频流和一个问题，通过使用 BridgeFormer [12] 评估从问题中提取的文本特征与从视频流中提取的视觉特征之间的注意力信息，从而得出答案置信度。随后，基于置信度的目标事件定位器会过滤无关事件并定位目标事件。为降低时间复杂度，我们以斐波那契方式双向遍历视频。2）目标事件问答。在定位目标事件后，使用 TimeSformer [16]从视频关键帧中提取视觉特征，并使用 BERT-base 模型 [17, 18] 提取语言特征。随后，将视觉特征和语言特征拼接起来，并输入多模态编码器进行跨模态学习。最后，由答案解码器生成可靠的答案 [8, 10]。

2.2. 基于置信度的 TargetEvent 定位器

与现有的会对整个视频进行编码的视频问答模型不同，体育赛事直播等在线视频流需要根据不断输入的视频流，低延迟地回答问题。此外，无关事件常常会干扰对视觉信息的理解，尤其是在长视频中。因此，我们提出一种基于置信度的目标事件定位器（CTEL），用于过滤无关事件并找到目标事件。与事件定位相比，问题提供的语言监督远弱于事件定位中给定的文本描述。

回答问题的置信度是基于当前视频帧与给定问题之间的注意力信息，使用 BridgeFormer [12]计算得到的。许多深度神经网络已被开发用于视觉识别 [22–24]。本文使用 12 层视觉变换器 [25–27]提取视觉特征。该模型将每个视频帧划分为N_u 图像块，并将这些图像块线性投影为一系列视觉标记。随后，将分类标记 [CLS] 添加到视觉标记中。在向这些标记添加可学习的位置嵌入后，执行空间自注意力，并生成一系列视觉特征 $\boldsymbol{U} = [\boldsymbol{u}_{cls}, \boldsymbol{u}_1, ..., \boldsymbol{u}_{N_u}]$。类似地，使用一个采用 DistilBERT 架构的 12 层文本变换器来编码问题 [28], 。问题被标记化为 N_p 个标记，与一个 [CLS] 标记拼接，并添加可学习的位置嵌入。然后，在文本标记上执行自注意力，并生成一系列特征 $\boldsymbol{P} = [\boldsymbol{p}_{cls}, \boldsymbol{p}_1, ..., \boldsymbol{p}_{N_p}]$。最后，对视觉特征 [CLS]标记和问题特征 u_{cls} p_{cls} 进行归一化，并通过点积相乘，以获得视频帧的置信度。

为了去除干扰事件并定位目标事件，我们设计了一种滞后阈值处理方案，其中 c_{min}表示预期的最低置信度，c_{max} 表示置信度足够高时的阈值。一旦当前帧的置信度足够高，即 $c \geq c_{max}$，我们便从当前帧开始，按照斐波那契方式双向遍历视频，直至 $c < c_{min}$。遍历得到的视频帧 $\mathcal{T}_B$ 和 $\mathcal{T}_F$ 构成目标事件 $\mathcal{T}$。斐波那契采样

主要是为了减少计算量。

算法 1： 基于置信度的事件中心式在线视频问答。

输入： 在线视频流 $\mathcal{O}$、问题 $\boldsymbol{q}$、答案空间 $\mathcal{A}$

输出： 针对问题的答案 $\boldsymbol{\alpha}$

初始化置信度阈值 c_{min}, c_{max};
对于每一帧 $\boldsymbol{f} \in \mathcal{O}$ **执行**
  $c \leftarrow$ 视频文本编码器$(\boldsymbol{q}, \boldsymbol{f})$;
  如果 $c \geq c_{max}$ **then**
    $\mathcal{T}_B$:=反向遍历帧，直到 $c < c_{min}$;
    $\mathcal{T}_F$: =前向遍历帧，直到 $c < c_{min}$;
    $\mathcal{T}$: $= \mathcal{T}_B \cup \mathcal{T}_F$;
    中断;
  end
结束
视频特征 $\boldsymbol{V}$: =视频编码器$(\mathcal{T})$;
问题特征 $\boldsymbol{Q}$: =问题编码器$(\boldsymbol{q})$;
$\boldsymbol{M}$:=多模态编码器$(\boldsymbol{Q}, \boldsymbol{V})$;
$\boldsymbol{\alpha} \leftarrow$答案解码器$(\boldsymbol{M}, \mathcal{A})$;

2.3. 目标事件问答

定位到目标事件 $\mathcal{T}$ 后，我们设计了一个在线目标事件问答模型来生成正确答案。我们使用 12 层的 TimeSformer [16]和 BERT-base 模型 [17, 18] 分别对目标事件中的帧和给定问题进行编码，以提取视觉和语言特征。每帧被划分为图块，并线性投影为图块标记。随后，将这些标记与可学习的位置嵌入和时间嵌入相加，并输入 TimeSformer 注意力块，以执行分解的时空注意力。编码后各帧的特征向量经过平均池化，得到一个特征向量 $\boldsymbol{V} = [\boldsymbol{v}_{cls}, \boldsymbol{v}_1, ..., \boldsymbol{v}_{N_v}]$，其中 $\boldsymbol{v}_{cls}$ 是用于分类的 [CLS]标记。对于问题编码，问题被分词为一个包含 N_q 个标记的序列。然后对输入的问题标记执行自注意力，并得到特征序列 $\boldsymbol{Q} = [\boldsymbol{q}_{cls}, \boldsymbol{q}_1, ..., \boldsymbol{q}_{N_q}]$ 作为输出。

将提取的文本特征 $\boldsymbol{Q}$ 和视觉特征 $\boldsymbol{V}$ 拼接成一个序列，并输入多模态Transformer [19]，在不添加其他嵌入的情况下，对各项特征执行自注意力。输出序列表示为 $\boldsymbol{M} = [\boldsymbol{m}_{cls}, \boldsymbol{m}_1, ..., \boldsymbol{m}_{N_q+N_v}]$。最后，使用多层感知机融合其 $\boldsymbol{m}_{cls}$[CLS] 标记，并采用答案解码器[8, 10] 来生成正确答案。

3. ATBS数据集的构建

为了评估针对新开发的 O²VQA 任务所提出的方法，我们构建了一个**答案目标背景流**（ATBS）数据集[1]。对于 O²VQA 任务，要正确回答问题所需的视觉线索可能在任意时间多次出现，且持续时长未知。为模拟现实场景，我们设计了一种背景 +目标方法，其中背景是一段相对较长的背景视频，用于模拟在线视频流，而

目标 是一个包含回答给定问题所需信息的目标短视频片段。目标 视频会被插入背景 视频中的随机位置，以构成一个视频样本。

背景视频：为了模拟视频流中出现多个事件的现实场景，我们选用 Distinct Describable Moments (DiDeMo) [14] 数据集作为背景视频的来源。该数据集中的每个视频都包含多个不同的事件。它包含 10,464 个视频，视频平均时长约为 48 秒。

目标视频：我们使用 MSRVTT 数据集 [13]中的视频片段作为目标视频，其中每个片段对应一个不同的事件。该数据集包含 1 万个视频片段，平均时长约为 14秒。

数据集生成： MSRVTT 数据集 [13] 中的每个视频片段都会与从 DiDeMo 数据集 [14]中随机选取的一个背景视频进行唯一配对。目标视频片段的帧和背景视频的帧都会调整为224×224。随后，将目标视频片段插入背景视频中随机选取的视频帧之后，从而生成我们数据集中的最终视频，模拟目标事件在视频流中随机出现的情况。接着，将目标 视频片段对应的问答对选作生成视频的标注。需要注意的是，对相似视频进行配对可能会产生多个“正确”答案，而背景视频也可能包含具有干扰性的“正确”答案。尽管随机配对可能会将目标视频与无关的背景视频合并，但它能够确保问题具有唯一答案。最终数据集包含 1 万个视频，共计 24.4 万个问答对。

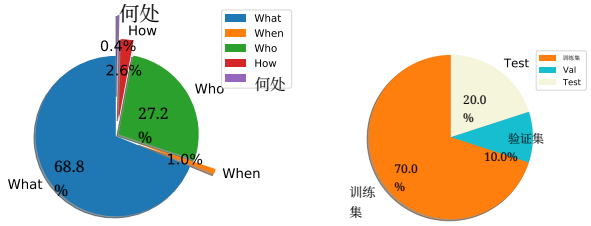

图 2。ATBS数据集中问题类型（左）和训练测试划分策略（右）的分布。

ATBS数据集是首个用于 O²VQA 任务的数据集。我们将其与现有视频问答数据集进行了比较，结果汇总于表1。所构建的数据集平均时长最长，总时长位居第二，问答对数量最多。ATBS数据集包含 5 种问题类型。问题类型的分布汇总如下

[1]访问地址：https://kenn-san.github.io/ATB S

表1。与现有视频问答数据集的比较。

数据集	Task	视频构成	总时长	平均时长	视频数量	# 问答对
MSRVTT-QA [15]	视频问答	仅目标	38.89小时	14s	10,000	243,680
MSVD-QA [15]	视频问答	仅目标	5.47小时	10s	1,970	50,505
TGIF-QA [29]	视频问答	仅目标	47.27小时	3s	56,720	103,919
视频问答 [30]	视频问答	仅目标	226.25小时	45s	18,100	174,775
ATBS	O²VQA	背景 +目标	175h	63s	10,000	243,680

见图 3。答案空间中的标签数量为 1290。对于数据集划分，我们选取 7k 个视频作为训练集，1k 个视频用于验证，2k 个视频用于测试集。

4. 实验结果

4.1. 实验设置

由于 O²VQA 是一项新颖的视频问答任务，目前没有现成的基准模型。因此，我们采用 ALPRO [19], 这一最先进的视频问答模型来处理 O²VQA 任务。与仅处理部分视频的所提方法相比，ALPRO 可以对完整视频进行编码。CTEL 中的视频-文本编码器 BridgeFormer [12], 使用官方发布的权重进行初始化[2]。我们通过经验设定置信度阈值，即，$c_{max} = 0.4$ 和 $c_{min} = 0.3$。初始学习率设为 5e-5，随后各轮训练中线性衰减。训练批大小为 48。优化器采用 AdamW，权重衰减为 1e-3。实验在 NVIDIA A100 GPU 上进行。所提方法和 ALPRO 各训练 20 轮约需 3 GPU小时。

4.2. 对比结果

我们首先比较基线模型 AL- PRO [19] 在不同数据集上的性能，以展示我们新构建的 ATBS 数据集和新开发的 O² VQA 任务的难度。如表 2所示，ALPRO [19] 在 ATBS 数据集上的准确率远低于在其他数据集上的准确率，这表明 ATBS 更具挑战性，因为有用的视觉线索在长视频中被稀释了。

表2。ALPRO [19] 在视频问答数据集上的准确率。我们的 ATBS 数据集比其他数据集更具挑战性。

数据集	what	who	how	何处	何时	总体
MSVD-QA	37.05	58.61	81.62	46.43	72.41	45.94
MSRVTT-QA	36.05	52.24	85.67	42.80	78.88	42.12
ATBS	19.16	33.58	77.27	25.00	54.55	24.80

[2]获取地址：https://github.com/tencentarc/mcq

随后，我们在 ATBS 数据集上将所提模型与 ALPRO [19]和 **TargetEvent** 进行比较，其中**TargetEvent**在进行视频问答时直接使用真实标注的目标事件。如表 3所示，尽管我们的方法只处理部分视频，但它将 ALPRO [19]的整体性能从 24.8% 提升到了 27.1%。结果证明了所提方法的有效性。我们的方法在“谁”和“如何”这两类问题上表现更优。具体而言，在“谁”类问题上，它显著优于基线模型 8.1%。这表明我们的方法能够更精准地定位与人物相关的目标事件，并给出更准确的答案。对于“如何”“何处”和“何时”这几类问题，我们的方法取得了与可直接获取目标事件的 **TargetEvent** 相同的性能。我们的方法能够快速定位目标事件，而 ALPRO [19] 需要利用整段视频来回答问题。因此，我们的方法运行速度快于 ALPRO，例如，在 Tesla V100 GPU 上，使用一段视频回答问题时，我们的方法耗时 28.5 秒，而 ALPRO 耗时 63.6 秒。

表 3 TargetEvent、ALPRO [19]与所提出的**CEO-VQA**在 top-1 准确率（%）上的比较。

方法	什么	谁	如何	何处	何时	总体
TargetEvent	24.38	43.25	79.55	25.0	54.55	31.1
ALPRO [19]	19.16	33.58	77.27	25.0	54.55	24.8
CEO-VQA	19.20	41.68	79.55	25.0	54.55	27.1

5. 结论

本文定义了一项新颖且具有挑战性的视频问答任务，称为在线开放式视频问答（O²VQA）。为解决这一问题，我们提出了基于置信度的事件中心在线视频问答（CEO-VQA）模型。具体而言，我们设计了置信度引导的目标事件定位器，用于定位目标事件，并判断收集到的信息是否足以给出有把握的答案。随后，我们开发了目标事件问答模型来得出最终答案。为评估所提方法，我们构建了一个大型 ATBS 数据集。在 ATBS 数据集上的实验结果表明，O²VQA 是一项更具挑战性的任务，且我们的方法优于当前最先进的视频问答模型。

参考文献

[1] H. Wang、D. Guo、X. Hua 和 M. Wang，《用于视频问答的成对 VLAD 交互网络》，载于 *ACM MM*，2021 年，第 5119–5127 页。1

[2] A. Theodoridis、C. Kotropoulos 和 Y. Panagakis，“使用超图和组稀疏性的音乐推荐”，载于 *ICASSP*，2013 年，第 56–60 页。1

[3] M. Wang、R. Hong、G. Li、Z. Zha、S. Yan 和 T. Chua，“通过标签定位和关键镜头识别实现事件驱动的网络视频摘要”，*IEEE 多媒体汇刊*，第 14 卷，第 4 期，第 975–985 页，2012 年。1

[4] X. Li、L. Gao、X. Wang、W. Liu、X. Xu、H. Shen 和 J. Song，“采用多样性学习的可学习聚合网络用于视频问答”，载于 *ACM MM*，2019 年，第 1166–1174 页。1

[5] L. Gao、P. Zeng、J. Song、Y. Li、W. Liu、T. Mei 和 H. S hen，“用于视频问答的结构化双流注意力网络”，载于*AAAI*，2019 年，第 33 卷，第 6391–6398 页。

[6] Z. Zhang, Z. Liu, T. Zhang, “面向视频问答的时空文本协同注意力网络”，*ACM TOMM*，第 15 卷，第 2s 期，第 1–18 页，2019 年。

[7] W. Jin, Z. Zhao, M. Gu, J. Yu, J. Xiao, Y. Zhuang, “用于视频问题解答的对象关系多交互网络”，载于*ACM MM*，2019年，第1193–1201页。1

[8] T. Le、V. Le、S. Venkatesh 和 T. Tran，“用于视频问答的分层条件关系网络”，发表于 *CVPR*，2020年，第 9972–9981 页。2，3

[9] J. Xiao, X. Shang, X. Yang, S. Tang, T.-S. Chua, “将视频视为对象：用于多粒度问题回答的条件图层级结构”，载于*AAAI*，2022 年，第 36 卷，第 2804–2812 页。

[10] C. Fan, X. Zhang, S. Zhang, W. Wang, C. Zhang, H. Huang, 《用于视频问答的异构记忆增强多模态注意力模型》，载于*CVPR*，2019，第 1999–2007 页。2，3

[11] J. Gao、R. Ge、K. Chen 和 R. Nevatia，《用于视频问答的运动-外观协同记忆网络》，载于*CVPR*，2018，第 6576–6585 页。1

[12] Y. Ge、Y. Ge、X. Liu、D. Li、Y. Shan、X. Qie 和 P. Luo，《通过选择题连接视频-文本检索》，载于 *CVPR*，2022，第 16167–16176 页。1，2，4

[13] J. Xu, T. Mei, T. Yao, and Y. Rui, “MSR-VTT：用于连接视频与语言的大型视频描述数据集”，发表于 *CVPR*，2016 年，第 5288–5296 页。1，3

[14] LA. Hendricks, O. Wang, E. Shechtman, J. Sivic, T. Darrell, 和 B. Russell，《利用自然语言定位视频中的片段》，发表于*ICCV*，2017，第 5803–5812 页。1，3

[15] D. Xu、Z. Zhao、J. Xiao、F. Wu、H. Zhang、X. He 和 Y. Zhuang，《通过对外观和运动进行逐步细化的注意力实现视频问答》，发表于 *ACM MM*，2017 年，第 1645–1653 页。1，2，4

[16] 你需要的全部注意力，载于 Vaswani 等人，*神经信息处理系统进展*，2017 年。2，3

[17] J. Devlin、M. Chang、K. Lee 和 K. Toutanova，“BERT：用于语言理解的深度双向 Transformer 预训练”，发表于 *北美计算语言学协会人类语言技术分会年会*，2019 年，第 4171–4186 页。2，3

[18] J. Liu、J. Ren、Z. Lu、W. He、Cui M、Z. Zhang 和 R. Bai，“用于自动化医学执照考试的基于跨文档注意力的门控融合网络”，《专家系统与应用》，第 205 卷，117588，2022 年2，3页码。。[19] D. Li、J. Li、H. Li、J. Niebles 和 Steven C. Hoi，“对齐与提示：结合实体提示的视频与语言预训练”，发表于*CVPR*，2022 年，第 4953–4963 页。2，3，4

[20] S. Zhang、J. Su 和 J. Luo，“利用自然语言中的时间关系进行视频片段定位”，载于 *ACM MM*，2019，第 1230–1238 页。2

[21] S. Zhang, H. Peng, J. Fu, Y. Lu, J. Luo, “学习用于视频片段定位的二维时间邻接网络”，载于*AAAI*，2020年，第 34 卷，第 12870–12877 页。2

[22] C. Yao, S. Wang, J. Zhang, W. He, H. Du, J. Ren, R. Bai, and J. Liu, 《基于加权时空表示的 EfficientNet 人脸面具攻击基于远程光电容积脉搏波的欺骗检测》，载于国际图像处理会议，2021 年，第 3872–3876 页。2

[23] X. Song, J. Chen, Y. Wang, R. Bai, Y. Jiang, X. Wang, “用于解决具有大面积侵蚀缺口的拼图的孪生判别深度强化学习”，发表于*AAAI*，2023年。

[24] X. Song, Y. Yang, Y. Jiang, R. Bai, 《使用拼图判别网络解决大型侵蚀缺口拼图》，发表于*ICASSP*，2023 年。2

[25] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit 和 N. Houlsby，《一张图像胜过 16×16 个词：面向大规模图像识别的 Transformer》，发表于*ICLR*，2021 年。2

[26] S. Chen、W. He、J. Ren 和 X. Jiang，《基于注意力的双流视觉 Transformer：用于雷达步态识别》，发表于 *ICASSP*，2022 年，第 3668–3672 页。

[27] W. He、J. Zhang、J. Ren、R. Bai 和 X. Jiang，《用于视觉类比推理的基于注意力的关系推理器与层次化 ConViT》，发表于 *AAAI*，2023 年。2

[28] V. Sanh, L. Debut, J. Chaumond, T. Wolf, 《DistilBERT，BERT 的蒸馏版本：更小、更快、更便宜、更轻量，》*arXiv* 预印本 *arXiv:1910.01108*，2019。2

[29] Y. Jang, Y. Song, C. D. Kim, Y. Yu, Y. Kim, G. Kim, “TGIF-QA：迈向视觉问答中的时空推理”，发表于*CVPR*，2017，第 2758–2766 页。4

[30] K. Zeng、T. Chen、C. Chuang、Y. Liao、J. Niebles 和 M. Sun，“利用视频描述学习视频问答”，发表于*AAAI*，2017 年，第 31 卷，第 4334–4340 页。4