arXiv:2303.03171v1 [cs.CV] 6 Mar 2023

用于变化描述的邻域对比 Transformer

Yunbin Tu、Liang Li、Li Su、Ke Lu 和 Qingming Huang，*IEEE* 会士

摘要—**变化描述旨在用自然语言描述一对相似图像之间的语义变化。与一般的图像描述相比，这项任务更具挑战性，因为它既要求捕捉细粒度变化信息，同时避免受到无关视角变化的干扰，还需要解决变化描述中的句法歧义。本文提出一种邻域对比变换器，以提升模型在不同场景下感知各种变化的能力，以及理解复杂句法结构的能力。具体而言，我们首先设计邻域特征聚合，将邻近上下文融入每个特征，从而在显著参照物的指引下快速定位不显著变化。随后，我们设计公共特征提炼，在邻域层面对两幅图像进行比较，并从每幅图像中提取共同属性，以学习二者之间有效的对比信息。最后，我们引入词语间的显式依赖关系来校准 Transformer 解码器，从而帮助模型在训练过程中更好地理解复杂句法结构。大量实验结果表明，所提出的方法在三个具有不同变化场景的公开数据集上取得了最先进的性能。代码可在 https://github.com/tuyunbin/NCT 获取。**

索引词—**变化描述、邻域对比变换器、句法依存关系。**

I. 引言

变化描述旨在描述两幅语义相似的图像之间发生的变化，这是视觉与语言领域的一项新任务 [1]–[3]。它进一步拓展了传统的图像描述 [4]，[5]，即，需要同时处理两幅图像并描述它们之间的差异。这推动了探索图像对之间关系与差异的研究。此外，它还有广泛的应用，例如为普通用户或视障用户解释复杂图像编辑效果、输出有关监控区域的日志，以及生成有关病理变化的报告 [6]–[8]。



图 1. 变化描述示例。第一个示例来自图像编辑场景，其中被移除的物体并不显眼。第二个示例展示了物体移动和适度视角变化同时存在的情况，且变化对象被部分遮挡。最后一个示例展示了物体移动和极端视角变化同时存在的情况，此时真实移动被伪移动所淹没。变化对象和指称对象分别用红框和白框标出。

关键挑战主要体现在两个方面。首先，模型应具备细粒度语义理解能力，因为变化信息通常难以准确定位。例如，在图 1 (a) 中，被移除的女孩因位置不显眼且形状模糊而容易被忽略。在图 1 (b) 中，变化难以定位，因为移动的方块被部分遮挡。其次，模型应避免受到无关干扰因素的影响，只描述真实语义变化。在动态环境中，由于相机抖动、拍摄时间不同等多种因素，几乎不可能在相同视角下获取两幅图像。在图 1 (c) 中，视角发生极端变化，导致未变化物体出现明显的伪移动，这可能会掩盖真实变化，并误导模型生成不准确的句子。

针对上述挑战，已有一些研究尝试加以解决。尽管取得了一些进展，这些方法仍难以学习有效的变化表征。具体而言，它们比较

本研究得到中国国家重点研发计划项目（编号：2018AAA0102000）的资助，并部分得到中国国家自然科学基金项目（编号：U21B2038、61931008）以及中国科学院青年创新促进会项目（编号：2020108）的资助。（通讯作者：*Liang Li*、*Li Su*。）

Yunbin Tu就职于中国科学院大学计算机科学与技术学院，中国北京101408（电子邮箱：tuyunbin22@mails.ucas.ac.cn）。

Liang Li 就职于中国中国科学院计算技术研究所，北京 100190，中国（电子邮箱：liang.li@ict.ac.cn）

Li Su 就职于中国科学院大学计算机科学与技术学院，北京 101408，中国（电子邮箱：Suli@ucas.ac.cn）。

Ke Lu 就职于中国科学院大学工程科学学院，北京，中国；鹏城实验室，南山，深圳，广东，中国。（电子邮箱：luk@ucas.ac.cn）。

Qingming Huang 任职于中国科学院大学计算机科学与技术学院（中国北京 101408），以及中国科学院计算技术研究所（中国北京 100190）（电子邮箱：qmhuang@ucas.ac.cn）。

本文的可下载补充材料由作者提供，网址为 http://ieeexplore.ieee.org.。材料包含实现细节以及三个数据集的更多定性示例。该材料大小为 0.98MB。

两幅图像时主要采用全局或局部层面的比较方式：全局比较指直接相减 [9], [10] ，局部比较则基于单独的特征匹配来计算二者的相似度 [8], [11],[12]。前者过于粗略，无法捕捉不显著或被遮挡的变化。后者更为合理，但容易受到极端视角变化的影响，例如，图 1 (c) 中的每个宾语似乎都在移动。在这种情况下，这种单独匹配无法学习变化的稳定特征。我们认为，要学习能够适应视角变化的有效特征，模型应在邻域层面比较细节。原因在于，图像中的空间邻近对象高度相关 [13], ，若某个宾语发生变化，它与邻近对象之间的关系也会随之改变。这种关系变化有助于挖掘不显著变化。此外，伪变化实际上是对象尺度和位置的失真，因此这些邻近对象之间的关系不会受到影响。基于此，我们尝试在邻域层面动态整合特征，从而帮助模型抵抗视角变化并定位真实变化。

此外，我们观察到，变化描述通常由两部分组成：语义变化和一个指称对象，因此包含复杂句法结构。如图1(c)所示，该句的主句是“棕色小方块移动了”。然而，主语“方块”和谓语“移动了”被一个描述指称对象“球”的从句分隔开。在这种情况下，“移动了”一词比起“方块”更靠近“球”一词。在训练过程中，如果模型不理解词语之间的句法关系，就可能从真实标注描述中学到错误信息。据我们所知，现有方法忽略了这个问题。事实上，如果模型注意到“方块”和“移动了”之间的直接依存关系，就可以避免上述误解。因此，有必要在训练过程中引入显式的依存关系，以帮助模型理解描述文本的句法结构。

本文提出一种邻域对比变换器，用于在不同变化场景下准确定位变化，并赋予模型依存关系的句法知识，以解决结构歧义问题。具体而言，给定一个图像对，首先设计邻域特征聚合，将邻域上下文融入每幅图像的特征中。这有助于模型抵抗视角变化，并在邻近指称对象的引导下感知细粒度变化。随后，基于相似度匹配，定制公共特征提炼，以建立上述两幅图像特征之间的对应关系，从而总结其共同特征。接着，通过移除共同特征，计算每幅图像中变化的稳定特征，并将其融合以学习图像对之间的对比特征。随后，将这些对比特征输入 Transformer 解码器以生成描述。在训练过程中，我们向解码器提供词语之间依存关系的先验知识，这有助于理解真实标注描述中的复杂句法结构。

本文的贡献总结如下：（1）提出一种邻域对比变换器，通过对图像对进行邻域对比来精确定位变化，并设计了邻域特征聚合以探索细粒度变化并抵御视角变化；设计了一种公共特征提炼方法，用于捕捉每幅图像的判别属性，并构建其对比特征以生成句子。（2）本研究首次尝试通过引入词语间的显式依赖关系来解决这一任务中的句法结构歧义。（3）大量实验表明，我们的方法在三个公开数据集上的表现优于最先进方法。

II. 相关工作

图像/视频描述。 在介绍变化描述相关研究之前，我们首先回顾近期发表的传统图像/视频描述研究。TTA [14] 从视频中检测视觉标签，以弥合视觉文本鸿沟，并提出一种文本时间注意力模型，用于建立词语与帧之间的对齐。LSRT [15] 提出长短期关系变换器，以充分挖掘宾语之间的关系，进而生成描述。I²Transformer [16]从不同模态中学习嵌入了内部与跨模态关系的表示，并将其输入标准Transformer以生成描述。P+D注意力 [17]在金字塔图像特征图上提出双重注意力模块，可探索视觉与语义之间的相关性，并优化生成的描述文本。MSA [18]提出多分支自注意力，并将其重复多次，以增强通用自注意力模型在描述生成过程中的表达能力。HTG+HMG [19]通过设计两类图，即语言到视觉异构图和视觉到视觉同构图，提出了一种关系感知注意力。

变化描述。 这是视觉描述中的一项新任务，同时也更具挑战性。这是因为它需要理解两幅图像的内容，并进一步描述它们之间的差异。先驱工作 [6] 基于监控场景描述变化。工作 [7] 阐述了两幅图像之间的编辑变换，如图1（a）所示。这两项工作的共同点是，它们都检测并描述两幅良好对齐的图像之间的变化。事实上，在拍摄图像时会存在视角偏移，这给区分真实变化与伪变化带来了挑战。鉴于此，Park等 [9] 和 Kim等 [11]分别发布了两个数据集，涵盖中等程度的视角变化（图1（b））和极端视角变化（图1（c））。为了描述视角变化下的语义变化，Park等提出了一个用于定位和描述变化的DUDA模型，其中通过对两幅未对齐的图像进行相减来建模差异，这可能会计算出带有噪声的差异特征 [20]。

为缓解这一问题，Hosseinzadeh等人 [10] 利用 TIRG [21] 的检索模型对 DUDA 进行正则化。Tu 等人 [20] 通过衡量相减得到的变化与图像对之间的关系，判断变化是否确实发生。与直接相减不同，一方面，相关工作 [8], [11], [22] 首先基于特征相似度提取两幅图像之间的共同特征，然后移除这些特征，以明确捕捉变化特征。另一方面，相关工作 [12], [23]匹配两幅图像之间的相似特征，以隐式推断变化特征。为增强视觉文本对齐，Kim 等人

[11]引入循环一致性模块，以优化生成句子。Yao 等人 [12] 通过从预训练到微调的范式，对细粒度跨模态对齐进行建模。

此外，Liao等人 [24]引入物体深度的三维信息，以应对视角变化。他们首先将图像输入预训练的深度估计模型，获得深度图。同时，他们利用预训练的 Yolov4 获取物体的边界框。随后，根据这些深度图和物体边界框，计算出物体深度。由于物体深度的准确性在很大程度上取决于这两个预训练模型的性能，因此计算得到的深度信息并不可靠。此外，引入三维信息会增加模型的复杂度。即便如此，利用三维知识仍是克服视角变化影响的另一种思路。这启发我们在未来进一步探索这一任务。

然而，上述方法主要基于全局层面（直接相减）或局部层面（单独进行特征匹配）来捕捉两幅图像之间变化的特征，却没有尝试在邻域层面学习变化特征。此外，这些方法忽略了真实描述中的语法歧义问题。为此，我们提出一种邻域对比变换器。它在邻域层面对比两幅图像，首先捕捉每幅图像的差异属性，然后学习它们之间的对比信息。此外，该方法利用依存关系解决变化描述中的结构歧义问题。

描述生成中的对比特征学习。学习对比特征旨在根据相似/不相似的图像对 [25]，对相似/不相似的图像表示进行建模。近期研究已在群组描述生成 [26]和胸部X光报告生成 [25]中尝试了这一思路。一方面，给定两组图像，Li 等人 [26] 提出使用自注意力机制从每个图像组中提取共同属性，然后捕捉两组之间的对比信息。另一方面，给定一张胸部X光图像和一组正常图像，Liu 等人 [25]提出了一种对比注意力模型，用于学习输入图像与正常图像之间的对比特征。我们的方法与这些方法有两个主要区别。首先，我们的任务中存在无关干扰因素（例如，视角变化）这带来了区分真实变化与伪变化的额外挑战。其次，与这些方法逐个匹配特征不同，我们的方法先聚合邻近特征，再在邻域层面进行特征匹配，以构建对比特征，旨在识别细粒度变化，同时不受视角变化的影响。

描述生成中使用的句法知识。已有一些研究尝试在描述生成中利用词性（PoS）知识以及词语之间的句法依存关系。一方面，Hou 等人 [27] 提出通过学习 PoS 序列与词语的联合概率来建模句法结构并利用语义基元。Wang 等人 [28] 提出词性生成器，用于预测句子的全局句法词性信息。Zhang等人 [29]和 Deng 等人 [30] 提出根据每个词语的 PoS 自适应地生成该词语，从而改善跨模态对齐。另一方面，Zheng 等人 [31]提出解码句法成分（主语、宾语和谓语），以定位视频片段中的动作。Zhao 等人 [32] 设计了一种多模态依存树构建方法，用于捕捉较长且复杂的视频描述中的句法和语义依存关系。

在变化描述中，大多数工作都专注于学习准确的变化表征以进行描述生成，却忽视了句法知识的利用。与 Zhang 等人和 Deng 等人类似，Tu 等人 [20]引入词性信息，并提出一种基于注意力的视觉切换机制，以动态利用视觉信息。不同于这项工作，我们旨在利用词语之间的显式句法依存关系来消除变化描述句法结构的歧义，这有助于模型在训练过程中区分真实描述中的变化对象及其指称对象。

III. 方法

如图 2 所示，我们的方法架构由四个部分组成：(1) 邻域特征聚合模块用于识别细粒度变化，并抵抗无关的视角变化；(2) 公共特征提炼模块从每幅图像中提取差异信息，并学习两幅图像之间的对比信息；(3) 对比变化定位器用于定位两幅图像中的具体变化特征；(4) 句法感知 Transformer 解码器将学习到的变化特征转换为自然语言句子，并预测词语之间的句法依存关系。

A. 邻域特征聚合

形式上，给定“之前” I_{bef} 和“之后”I_{aft}的两幅图像，我们利用现成的卷积神经网络提取它们的网格特征，分别记为 X_{bef} 和 X_{aft}，其中 $X \in \mathbb{R}^{C\times H\times W}$。C、H、W分别表示通道数、高度和宽度。尽管卷积神经网络能够捕捉局部空间上下文，但这些相关性是基于单幅图像、在没有视角变化的情况下建模的，因而无法直接迁移到变化描述任务中。此外，语义对应领域的最新研究 [33]表明，局部自注意力在捕捉相邻元素之间的关系方面表现出色。受此启发，我们设计了一个邻域特征聚合模块，通过整合两幅图像中同一邻域的空间上下文，动态更新每个特征。

具体而言，对于 $X_{bef(aft)} = \{x_1, \ldots, x_N\}$（$N = HW$），其中 $x_i \in \mathbb{R}^C$，我们首先通过共享线性变换，将第 i个网格单元 x_i 的特征投影到维度为 $\mathbb{R}^D$ 的低维嵌入空间：

$$x_i' = M_v x_i + b_v + pos(\tilde{x}, \tilde{y}), \tag{1}$$

其中 $M_v \in \mathbb{R}^{D\times C}$ 和 $b_v \in \mathbb{R}^D$ 是可训练参数。pos($\tilde{x}$, $\tilde{y}$) $\in \mathbb{R}^D$ 是第 i个网格特征的可学习位置嵌入。这里，$\tilde{x}$ 和 $\tilde{y}$ 分别表示图像高度和宽度维度上每个特征的顺序。位置嵌入层是两个查找表，大小分别为 (H, D/2) 和 (W, D/2)。根据每个特征的顺序，我们可以从高度和宽度维度学习其位置嵌入，并将它们拼接为该特征的位置嵌入。然后，对于每个网格特征 x_i'，我们选取其 $r \times r$ 个邻近特征

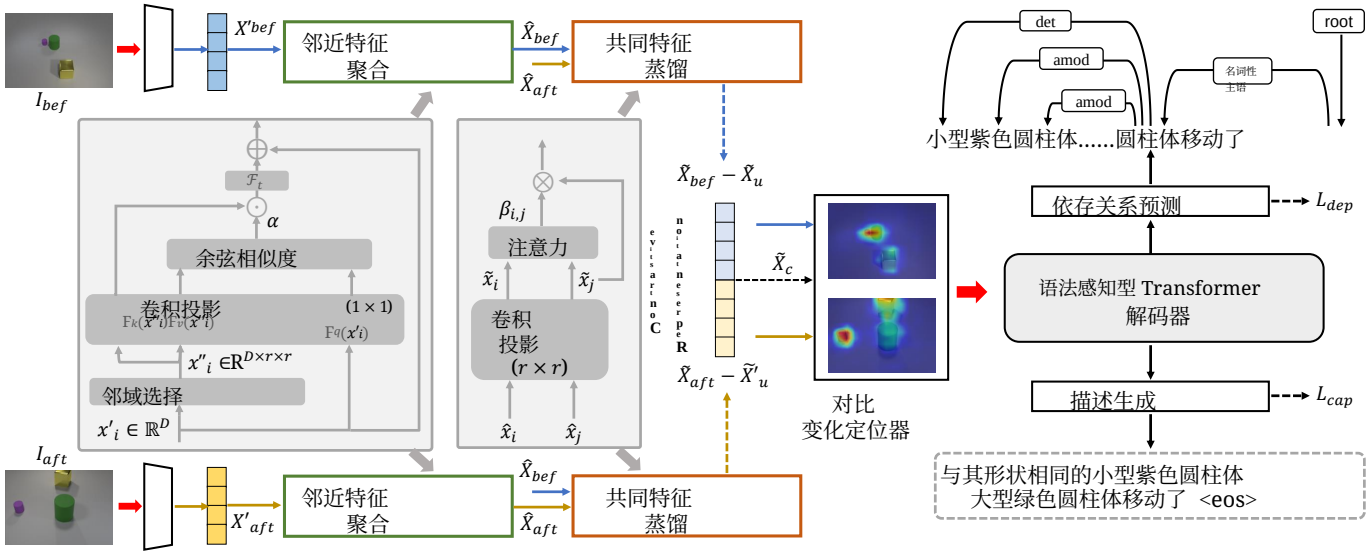


图2. 所提出的邻域对比变换器架构，包括邻域特征聚合、公共特征提炼、对比变化定位器和语法感知型 Transformer 解码器。

并获取邻域特征表示 $X''_{bef(aft)} \in \mathbb{R}^{C\times H\times W\times r\times r}$。接下来，我们计算 x'_i 与其 $r\times r$ 个邻域特征之间的余弦相似度：

$$e = \Phi\left[\mathcal{F}_q\left(x'_i\right), \mathcal{F}_k\left(x''_i\right)\right], \quad \alpha \sim \mathrm{Softmax}\left(e\right), \tag{2}$$

其中，$\alpha \in \mathbf{R}^{r\times r}$ 是关系系数，表示应从邻近特征中获取多少信息；Φ 是余弦相似度函数。$\mathcal{F}_q \in \mathbb{R}^D$ 和 $\mathcal{F}_k \in \mathbb{R}^{D\times r^2}$是两个卷积层。最后，通过聚合来自邻近特征的相关信息，将 x'_i 更新为$\hat{x}_i$：

$$\hat{x}_i = x'_i + \mathcal{F}_t(\sum_{r,r} \alpha \odot \mathcal{F}_v\left(x''_i\right)), \hat{x}_i \in \mathbb{R}^D, \tag{3}$$

其中，$\odot$ 表示逐元素乘法。上述$\hat{}$ $\hat{}$操作将原始特征更新为 X_{bef} 和 X_{aft}，使模型能够识别不明显和被遮挡的变化，同时不受视角变化的影响。

B. 共同特征蒸馏

如图1所示，与微小变化相比，图像对之间的大多数属性都是相同的。因此，从两幅图像中找出并移除共同部分是很自然的做法，剩余信息则可视为对比特征。受此启发，我们设计了一个公共特征提炼模块，用于比较两幅图像并学习有效的对比表示。

$\hat{}$ 在此，我们学习 X_{bef} 相较于$\hat{}X_{aft}$的变化特征。具体而言，我们首先利用带有深度可分离卷积的共享变换层$\hat{}$，将 X_{bef} 和$\hat{}X_{aft}$ 投影到公共语义空间中。这进一步捕获了同一邻域内的空间相关性：

$$\tilde{X}_{bef(aft)} = \mathcal{F}_{depth}\left(\hat{X}_{bef(aft)}, s\right), \tilde{X} \in \mathbb{R}^{D\times H\times W}, \tag{4}$$

其中，s 的核大小为 $r\times r$，我们将$\tilde{X}_{bef(aft)}$ 重塑为 $\tilde{X}_{bef(aft)} \in \mathbb{R}^{N\times D}$。然后，我们通过点积注意力衡量 $\tilde{X}_{bef}$ 中的每个特征 $\tilde{x}_i^b$ 与 X_{aft} 中的每个特征$\tilde{x}_j^a$ 之间的相似度：

$$\beta_{i,j} = \frac{\exp\left(\beta'_{i,j}\right)}{\sum_j \exp\left(\beta'_{i,j}\right)}, \quad \beta'_{i,j} = \tilde{x}_i^{bT}\tilde{x}_j^a, \tag{5}$$

其中 $\beta_{i,\ j} \in B$ 是一组相似度分数，用于指示哪些˜ ˜特征是 X_{bef} 和 X_{aft}之间的共同属性。˜然后，在学习到的相似度分数矩阵 B 的指导下，从 X_{aft} 中提取共同特征：

$$\tilde{X}_u = B\cdot\tilde{X}_{aft}. \tag{6}$$

˜ ˜接下来，我们从 X_{bef} 中移除共同特征 X_u，以提炼变化特征：

$$\tilde{X}_c^{bef} = \tilde{X}_{bef} - \tilde{X}_u. \tag{7}$$

基于这一类比，我们参照 X_{bef}提炼其中的变化特征 $\tilde{X}_c^{aft}$ $\tilde{X}_{aft}$。最后，我们通˜过融合上述变化特征来构建两幅图像之间的对比表征，该操作由一个带有 ReLU 激活函数的全连接层实现：

$$\tilde{X}_c = \mathrm{ReLU}\left(\left[\tilde{X}_c^{bef}; \tilde{X}_c^{aft}\right] W_h + b_h\right). \tag{8}$$

C. 对比变化定位器

˜学习到对比表征 X_c后，我们引入一种基于空间注意力机制的对比变化定位器，用于精确定位两幅图像中的变化。具体而言，它首先通过˜使用 X_c 分别查询每幅图像表征，从而生成两张注意力图：

$$\gamma_{bef} = \sigma\left(MLP\left(\left[\tilde{X}_c; \tilde{X}_{bef}\right]\right)\right), \quad \gamma_{aft} = \sigma\left(MLP\left(\left[\tilde{X}_c; \tilde{X}_{aft}\right]\right)\right), \tag{9}$$

其中， MLP 是一个两层多层感知机，中间采用 ReLU 激活函数。 [、]和 σ 分别表示拼接操作和 Sigmoid 激活函数。此外，我们分别在空间维度上对每幅图像表征进行加权求和池化，从而定位变化的具体特征：

$$\begin{aligned} l_{bef} &= \textstyle\sum_{H,W} \gamma_{\text{bef}} \odot \tilde{X}_{bef}, l_{bef} \in \mathbb{R}^D, \\ l_{aft} &= \textstyle\sum_{H,W} \gamma_{\text{aft}} \odot \tilde{X}_{aft}, l_{\text{aft}} \in \mathbb{R}^D. \end{aligned} \tag{10}$$

D. 语法感知语言解码器

借助池化后的变化特征 l_{bef}， l_{aft}及其差异特征 l_{diff}，我们首先将它们拼接为 $V \in \mathbb{R}^{3\times D}$。然后，Transformer 解码器学习词嵌入特征 $E[W] = \{E[w_1], ..., E[w_m]\}$ 与视觉特征 V之间的跨模态对齐。最后，解码器利用变化的注意力特征生成句子，在此过程中，我们引入词语间依存关系的句法知识来校准解码器。这旨在解决变化描述中的句法歧义问题。

1）背景知识：我们首先简要回顾标准 Transformer 解码器的框架。其关键模块是缩放点积注意力。给定查询矩阵$Q \in \mathbb{R}^{T_q\times d_k}$，键矩阵 $K \in \mathbb{R}^{T_v\times d_k}$ 和值矩阵$V \in \mathbb{R}^{T_v\times d_v}$，注意力结果计算如下：

$$\text{Attention}\ (Q, K, V) = \text{Softmax}\left(\frac{QK^\top}{\sqrt{d_k}}, \text{dim} = 1\right) V. \tag{11}$$

多头注意力基于缩放点积注意力。它由 h 个不同的“头”组成。对于每个头，注意力结果的计算方式如下：

$$\text{head}\ _i = \text{Attention}\left(QW_i^Q, KW_i^K, VW_i^V\right). \tag{12}$$

随后，多头注意力操作会将所有头拼接起来，其定义如下：

$$\text{MultiHead}\ (Q, K, V) = \ \text{Concat}\ _{i=1...h}\left(\ \text{head}\ _i\right) W^O. \tag{13}$$

此外，每个注意力层的输出 x 会输入一个基于非线性变换的前馈网络（FFN）：

$$\text{FFN}(\boldsymbol{x}) = \text{GELU}\left(\boldsymbol{x}W_{f1} + \boldsymbol{b}_{f1}\right) W_{f2} + \boldsymbol{b}_{f2}. \tag{14}$$

接下来，我们介绍词语之间的依存关系。在自然语言处理领域，依存句法分析是指考察句子中语言单位（例如，词语）之间的依存关系，以确定句子的语法结构。也就是说，句法依存关系指词语通过有向链接彼此相连的关系。动词被视为分句结构的中心，并标记为“根节点”。其他所有句法词语都通过有向链接直接或间接地连接到“根节点”。在图 2 中，我们主要展示了这一变化描述中的主要词语。依存标签表示两个词语之间的关系。例如，词语“移动”改变了名词“圆柱体”的含义。因此，我们可以发现一条从“移动”指向“圆柱体”的依存关系，其中“移动”是中心词，“圆柱体”是子节点或依存词。该依存关系的标签是“nsubj”，表示该句的名词性主语。动词“移动”是这一依存结构中的根节点。此外，我们注意到另一个“圆柱体”和“移动”之间没有有向链接。基于这些有向链接，模型能够更好地理解句子中的复杂结构，从而识别出哪个宾语发生了变化。

2) 解码阶段：解码器包含由 N 个相同层组成的堆栈。在第 l 个解码器层中，掩码自注意力层首先将词嵌入特征 $E[W] = \{E[w_1], ..., E[w_m]\}$ 作为输入，并对它们之间的关系进行建模；该层可防止模型看到未来的词语。其操作定义如下：

$$\hat{E}[W] = \ \text{LN}\ (E[W] + \ \text{MultiHead}\ (E[W], E[W], E[W])), \tag{15}$$

其中，LN 是层归一化 [34] 的缩写。然后，ˆ 解码器利用注意力特征 $E[W]$ ，基于交叉注意力层从 V 中查询最相关的特征：

$$\hat{H} = \ \text{LN}\ (E[\hat{W}] + \ \text{MultiHead}\ (E[\hat{W}], V, V)). \tag{16}$$

ˆ 随后，将 H 传入前馈层：

$$\tilde{H} = \ \text{LN}(\hat{H} + \text{FFN}(\hat{H})). \tag{17}$$

最后，通过两个独立的单隐藏层计算目标词和依存关系的概率分布：

$$\begin{aligned} W &= \text{Softmax}\left(\tilde{H}W_c + b_c\right), \\ D &= \text{Softmax}\left(\tilde{H}W_d + b_d\right), \end{aligned} \tag{18}$$

其中， $W_c \in \mathbb{R}^{D\times U}$、 $W_d \in \mathbb{R}^{D\times n}$、 $b_c \in \mathbb{R}^U$和 $b_d \in \mathbb{R}^n$ 是待学习的参数。 U 是词汇表大小； n 是依存关系的数量。

E. 联合训练

我们通过最大化观测到的词语序列和依存关系的似然，以端到端方式联合训练描述生成器和依存关系预测器。给定目标真实描述词（w_1^c、...、 w_m^c）和依存关系（w_1^d、...、 w_m^d），我们分别最小化描述生成器和依存关系预测器的负对数似然损失：

$$\begin{aligned} L_{cap}(\theta_c) &= -\sum_{t=1}^{m} \log p\left(w_t^c \mid w_{<t}^c; \theta_c\right), \\ L_{dep}(\theta_d) &= -\sum_{t=1}^{m} \log p\left(w_t^d \mid w_{<t}^d; \theta_d\right), \end{aligned} \tag{19}$$

其中， θ_c 和 θ_p 分别是描述生成器和依存关系预测器的参数。 m 是描述和依存关系的长度。最终的损失函数按如下方式优化：

$$L(\theta) = L_{cap} + \lambda L_{dep}, \tag{20}$$

其中， λ 是权衡参数，用于平衡描述生成器和依存关系预测器的贡献。

IV. 实验

A. 数据集

图像编辑请求数据集 [7]由 3,939 对真实图像和 5,695 条编辑指令组成。训练集中的每对图像对应一条指令，验证集和测试集中的每对图像则对应三条指令。该数据集中的变化物体通常不明显且难以辨认。我们采用官方划分，其中 3,061 对图像用于训练，383 对用于验证，495 对用于测试。

CLEVR-Change 是一个大规模合成数据集 [9]，具有适度的视角变化。它包含79,606对图像和493,735条描述文本，涵盖五种变化类型，即“颜色”“纹理”“添加”“删除”和“移动”。它包含两种变化设定：场景变化与伪变化并存，以及仅有伪变化。我们采用官方划分，其中67,660对用于训练，3,976对用于验证，7,970对用于测试。

CLEVR-DC 是一个大规模合成数据集 [11]，用于模拟极端视角偏移。它包含48,000对图像，其变化类型与CLEVR-Change相同。我们采用官方划分，分别将85%用于训练、5%用于验证、10%用于测试。

B. 评估指标

继最先进方法之后， [8], [10], [35], 我们使用五种指标来评估生成句子，即 BLEU-4 (B) [36], METEOR (M) [37],ROUGE-L (R) [38],CIDEr (C) [39], 和 SPICE (S) [40]。BLEU-4 用于语料库级别的四元组匹配比较，并已广泛应用于机器翻译任务。METEOR 旨在通过精确的词元匹配来衡量候选句子与参考句子之间的关系。ROUGE-L 基于最长公共子序列（LCS），计算两个句子中按相同顺序共存的词语相关性。CIDEr 是近期提出的、专为图像描述任务设计的指标，用于捕捉人类对共识的判断。SPICE 也是一种新指标，专为图像描述任务设计，用于比较候选句子与参考句子之间的语义命题内容。我们基于 Microsoft COCO 评估服务器 [41]计算结果。

C. 实现细节

为了公平比较，我们采用最先进方法，使用在 ImageNet 数据集上预训练的 ResNet-101 模型 [42] 提取图像对的网格特征，特征维度为 $1024 \times 14 \times 14$。我们首先将这些特征投影到 512 维。整体模型中的隐藏层大小和解码器中的词嵌入维度分别设为 512 和 300。为了获得真实依存关系，我们利用预训练的双仿射解析器 [44] 提取训练集每个句子的显式依存关系。邻域特征聚合层数设为 1，依存标签数量设为 49，邻域范围 r 设为 3，解码器层数设为 2，注意力头数量设为 8。

在训练期间，对于 CLEVR-Change 和 CLEVR-DC，我们将批次大小和学习率分别设置为 128 和 2×10^{-4}。在图像编辑请求任务中，批次大小和学习率分别设置为 32 和 2×10^{-4}。我们使用 Adam 优化器 [45]来最小化公式（20）中的负对数似然损失。在推理阶段，采用贪心解码策略生成目标描述。训练和推理均使用 PyTorch [46]在 RTX 3090 GPU 上实现。

表一 CLEVR-Change 上与最先进方法的总体性能比较。“*”表示该模型使用三个预训练任务进行训练。

方法	总计				
	B	M	R	C	S
DUDA (ICCV'19)	47.3	33.9	-	112.3	24.5
M-VAM (ECCV'20)	50.3	37.0	69.7	114.9	30.5
DUDA+TIRG (CVPR'21)	51.2	37.7	70.5	115.4	31.1
IFDC (TMM'21)	49.2	32.5	69.1	118.7	-
R^3Net+SSP（EMNLP'21）	54.7	39.8	73.1	123.0	32.6
VACC（ICCV'21）	52.4	37.5	-	114.2	31.0
SRDRL+AVS（ACL'21）	54.9	40.2	73.3	122.2	**32.9**
SGCC（ACM MM'21）	51.1	**40.6**	**73.9**	121.8	32.2
MCCFormers-D（ICCV'21）	52.4	38.3	-	121.6	26.8
PCL（不含 PT）（AAAI'22）	32.7	27.7	57.2	89.8	-
PCL（含 PT）（AAAI'22）*	51.2	36.2	71.7	**128.9**	-
NCT（我们的方法）	**55.1**	40.2	73.8	**124.1**	**32.9**

表二 CLEVR-CHANGE 上场景变化与最先进方法的比较。

方法	场景变化				
	B	M	R	C	S
DUDA（ICCV'19）	42.9	29.7	-	94.6	19.9
DUDA+TIRG（CVPR'21）	49.9	34.3	65.4	101.3	27.9
IFDC（TMM'21）	47.2	29.3	63.7	105.4	-
R^3Net+SSP（EMNLP'21）	52.7	36.2	69.8	116.6	30.3
SRDRL+AVS（ACL'21）	52.7	36.4	69.7	114.2	30.8
NCT（本文方法）	**53.1**	**36.5**	**70.7**	**118.4**	**30.9**

D. 性能比较

1）CLEVR-Change 数据集上的结果：我们从以下方面将所提出的方法与最先进方法进行比较：1）总体性能，评估在场景变化和伪变化下的整体性能；2）场景变化；3）不同的变化类型。十种对比方法分别为 DUDA [9], M- VAM [8], IFDC [22],DUDA+TIRG [10], R^3Net+SSP [35],VACC [11], SRDRL+AVS [20],MCCFormers-D [23], SGCC[24],, 以及使用和不使用 PT（预训练）的 PCL [12]。其中，PCL 设计了三项预训练任务，以增强图像差异与描述文本之间的细粒度对齐。PCL 的作者使用 8K 个预热步数对模型进行预训练，总迭代次数为 250K。与 PCL 不同，其他对比方法均以端到端方式进行训练。因此，为了进行公平比较，我们分别比较使用和不使用预训练的 PCL。结果如表一至表五所示。

在表一中，我们可以观察到：1）NCT 在大多数指标上都取得了更优的结果；2）请注意，MCCFormers-D 同样基于 Transformer，并根据特征相似度识别变化。它与我们的方法有两个主要区别。首先，它对两组特征进行逐一匹配。相比之下，我们的 NCT 旨在在邻域层面对两幅图像进行比较，以捕捉它们之间的对比特性，这有助于感知细粒度变化，同时

表III 不同变化类型的 CIDEr 评估详细结果：“(C) 颜色”、“(T) 纹理”、“(A) 添加”、“(D) 删除”和“(M) 移动”。

方法	CIDEr				
	C	T	A	D	M
DUDA (ICCV'19)	120.4	86.7	108.3	103.4	56.4
M-VAM (ECCV'20)	122.1	98.7	126.3	115.8	82.0
DUDA+TIRG (CVPR'21)	120.8	89.9	119.8	123.4	62.1
IFDC (TMM'21)	133.2	99.1	128.2	118.5	82.1
R^{3}Net+SSP（EMNLP'21）	139.2	123.5	122.7	121.9	**88.1**
SRDRL+AVS（ACL'21）	136.1	122.7	121.0	126.0	78.9
SGCC (ACM MM'21)	128.0	122.9	117.1	116.9	77.1
PCT（带 PT）（AAAI'22）	131.2	101.1	**133.3**	116.5	81.7
NCT（我们的方法）	**140.2**	**128.8**	128.4	**129.0**	86.0

表四 SPICE 评估的详细分类。

方法	SPICE				
	C	T	A	D	M
DUDA（ICCV'19）	21.2	18.3	22.4	22.2	15.4
M-VAM（ECCV'20）	28.0	26.7	30.8	32.3	22.5
DUDA+TIRG（CVPR'21）	29.7	27.4	31.4	30.8	23.5
R^{3}Net+SSP（EMNLP'21）	31.6	30.8	32.3	31.7	25.4
SRDRL+AVS（ACL'21）	**32.4**	30.9	**33.0**	32.4	25.4
SGCC（ACM MM'21）	30.0	31.1	30.8	30.1	25.3
NCT（我们的方法）	**32.4**	**31.8**	32.3	**32.6**	**25.5**

不受视角变化的影响。其次，与基于标准Transformer进行描述生成的方法不同，我们利用句法依存关系来校准解码器，从而更好地理解变化描述中复杂的句法结构。3）与 SGCC 相比，所提出的 NCT 在 METEOR 和 ROUGE-L 指标上的表现略低。我们推测，SGCC 利用了比我们更多的视觉模态，例如通过 Yolov4 [47] 提取的语义属性，以及由 Monodepth2 [48]计算的图像深度图。相比之下，NCT 在其他指标上大幅超越了 SGCC。4）与 PCL 相比，NCT 即使不经预训练也大幅超越了它。对于经过预训练的 PCL，NCT 在三项指标上也表现更优。在 CIDEr 上，NCT 略低。我们推测，PCL 利用了三项预训练任务（总计 8K 个预热步数和 250K 次迭代）来增强模型。

如表II所示，NCT在每项指标上均优于最先进方法，尤其是在CIDEr得分上实现了大幅提升。在表III-V中，我们使用CIDEr、SPICE和METEOR指标，在特定变化类型下将NCT与最先进方法进行比较。特别是，CIDEr和SPICE专门用于评估描述生成性能。结果表明，我们的NCT在几乎所有类别中均优于最先进方法。这表明我们的方法在不同变化类型下具有良好的泛化能力。

总之，与不同情况下的最先进方法相比，所提出的 NCT 取得了令人鼓舞的性能。这一优势源于：1）邻域特征聚合和公共特征提炼有助于学习可靠的对比特征，并抵抗无关的视角变化；2）句法依存关系可以解决变化描述中的结构歧义问题。

表 V METEOR 评估的详细分析。

方法	METEOR				
	C	T	A	D	M
DUDA（ICCV'19）	32.8	27.3	33.4	31.4	23.5
M-VAM+RAF（ECCV'20）	35.8	32.3	37.8	36.2	27.9
DUDA+TIRG（CVPR'21）	36.1	30.4	37.8	36.7	27.0
IFDC (TMM'21)	33.1	27.9	36.2	31.4	31.2
R^{3}Net+SSP (EMNLP'21)	38.9	35.5	38.0	37.5	30.9
SRDRL+AVS（ACL'21）	39.0	35.6	38.9	**38.0**	30.1
SGCC (ACM MM'21)	37.8	36.1	38.9	36.7	**32.8**
NCT（我们的方法）	**39.1**	**36.3**	**39.0**	37.2	30.5

表 VI CLEVR-DC 上与最先进方法的比较。

方法	B	M	R	C	S
DUDA	40.3	27.1	-	56.7	16.1
DUDA + CC	41.7	27.5	-	62.0	16.4
M-VAM	40.9	27.1	-	60.1	15.8
M-VAM+CC	41.0	27.2	-	62.0	16.4
VA	44.5	29.2	-	70.0	17.1
VACC	45.0	29.3	-	71.7	**17.6**
NCT	**47.5**	**32.5**	**65.1**	**76.9**	15.6

2）CLEVR-DC 数据集上的结果：实验还在一个新近发布的、具有极端视角变化的合成数据集（ICCV'21）上进行。我们将其与六种最先进方法进行比较：DUDA/DUDA+CC [9], M-VAM/M-VAM+CC [8], 以及 VA/VACC [11]。

比较结果如表 VI 所示。我们发现，NCT 在大多数指标上均大幅优于最先进方法。这验证了我们的方法在任何视角变化下都具有良好的鲁棒性。这主要得益于其能够在邻域层面捕获图像对之间的对比信息，因为在视角变化下，未发生变化的局部邻域内部及局部邻域之间的物体关系是稳定的。

表 VII 在图像编辑请求数据集上与最先进方法的比较。

方法	B	M	R	C	S
多头注意力	6.1	11.8	35.1	22.8	-
静态关系注意力	5.8	12.6	35.5	20.7	-
动态关系注意力	6.7	12.8	37.5	26.4	-
NCT	**8.1**	**15.0**	**38.8**	**34.2**	**12.7**

3) 图像编辑请求数据集上的结果：我们在另一个具有挑战性的数据集——图像编辑请求数据集上进行了实验。该数据集中的变化物体通常模糊且不显眼。我们将 NCT 与 Tan 等人 [7] 报告的三种最先进方法进行比较：多头注意力、静态关系注意力和动态关系注意力。

表 VII 显示，NCT 大幅优于最先进方法。这表明，所提出的方法能够通过捕捉邻域对比特征，并基于词语间的显式依赖关系实现句法消歧，准确描述“源”图像中的哪些部分经过了编辑。

简言之，在上述三个数据集上进行的实验表明，我们的方法能够很好地泛化到不同变化场景中的变化定位与描述任务。

表 VIII 基于 CLEVR-CHANGE 总体性能的消融研究

方法	B	M	R	C	S
Diff-sub	53.3	38.8	72.1	119.7	31.8
NFA	54.3	39.7	73.1	121.9	32.0
CFD	54.1	39.6	73.1	122.8	32.2
ST	53.7	39.4	72.7	120.7	31.9
不含 S 的 NCT	54.6	40.0	73.6	123.4	32.5
NCT	**55.1**	**40.2**	**73.8**	**124.1**	**32.9**

E. 消融研究

我们进行了消融研究，以验证所提出的各个模块及完整模型的有效性。(1) Diff-sub 是一种基于 Transformer 的基线模型，通过直接相减计算差异特征。具体而言，它首先直接对两幅图像的图像特征进行相减，以获得差异表示。然后，根据相减后的表示，使用空间注意力机制从每幅图像中选出最相关的特征。最后，将特定变化特征输入标准 Transformer 解码器，以进行描述生成。(2) NFA 在相减之前进行邻域特征聚合。(3) CFD 提炼共同特征以构建对比特征，而非采用直接相减。(4) ST 指使用语法依存关系增强基线模型的语法感知 Transformer 解码器。(5) 不含 S 的 NCT 是不包含语法依存关系的邻域对比变换器。(6) NCT 是所提出的完整模型：邻域对比变换器与语法依存关系。

消融研究基于 CLEVR-Change 上的总体性能进行，结果如表 VIII 所示。我们观察到：1）与基线模型相比，各模块和完整模型均取得了稳定的性能提升；2）NFA 和 CFD 的性能相近，而二者结合后性能更佳；3）仅在基线模型中加入句法依存关系，带来的提升有限，与 NCT 结合后则取得最佳性能。上述观察表明：1）所提出的各模块和完整模型均有效；2）各模块不仅发挥各自的独特作用，还能相互补充；3）只有当模型通过邻域特征聚合和公共特征提炼学到有效的对比特征时，语法感知解码器才能利用这些特征生成正确的句子。

F. 评估模型在不同程度视角变化下的鲁棒性

为评估 NCT 在不同程度视角变化下的鲁棒性，遵循先驱工作 [9]，我们计算两幅图像中各个对象（变化对象除外）边界框的交并比，其中交并比越低，难度越高。在此，我们使用 SPICE 评估 Diff-sub（基线模型）和 NCT 生成的句子在对象边界框交并比不同时的表现。结果如图3左侧子图所示。值得注意的是，NCT 始终大幅优于基线模型。这表明所提出的方法

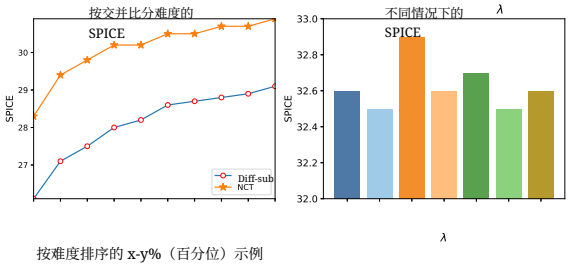


图3. 左侧子图展示了按视角变化（通过交并比衡量）细分的描述性能（SPICE）可视化结果；右侧子图展示了权衡参数 λ 对 CLEVR-Change 的影响。

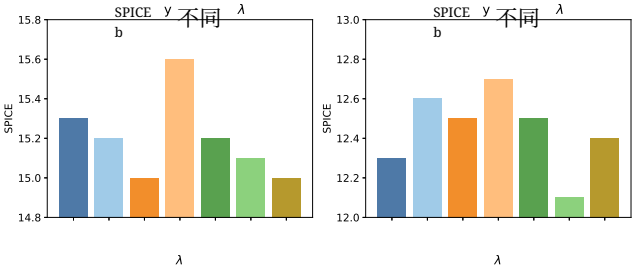


图4. 权衡参数 λ 对 CLEVR-DC（左侧子图）和图像编辑请求（右侧子图）的影响。

能够识别可靠的变化，并应对不同程度的视角变化。

G. 权衡参数 λ 的研究

本节讨论式（20）中的权衡参数 λ 对 CLEVR-Change、CLEVR-DC 和图像编辑请求的影响。该参数用于平衡描述生成器和依存关系预测器的贡献。在 CLEVR-Change 上，采用不同取值时得到的 SPICE 分数如图3右侧子图所示。我们发现，随着 λ 增大或减小，NCT 的性能会发生变化。这主要是因为整个模型会过度关注其中一部分，而忽略来自另一部分的监督信号。基于此，我们通过经验将 λ 设为 0.01。此外，另外两个数据集 CLEVR-DC 和图像编辑请求的结果如图4所示。采用不同取值时，CLEVR-DC 上得到的 SPICE 分数见左侧子图，图像编辑请求上的分数见右侧子图。值得注意的是，与 CLEVR-Change 上的实验结果类似，随着 λ 增大或减小，NCT 的性能会发生变化。在这两个数据集上，较好的取值为 0.02。以上分析表明，不同数据集上的权衡参数 λ 取值相近，这验证了所提出方法在不同变化场景下具有良好的鲁棒性。

H. 邻域范围参数研究 r

本节将分析邻域范围 r 的影响。这里，我们分别将其设为 3 和 5。需要注意的是，当该值大于 5 时，计算成本会急剧增加，所需资源超过一块 RTX 3090 GPU。对于不同的

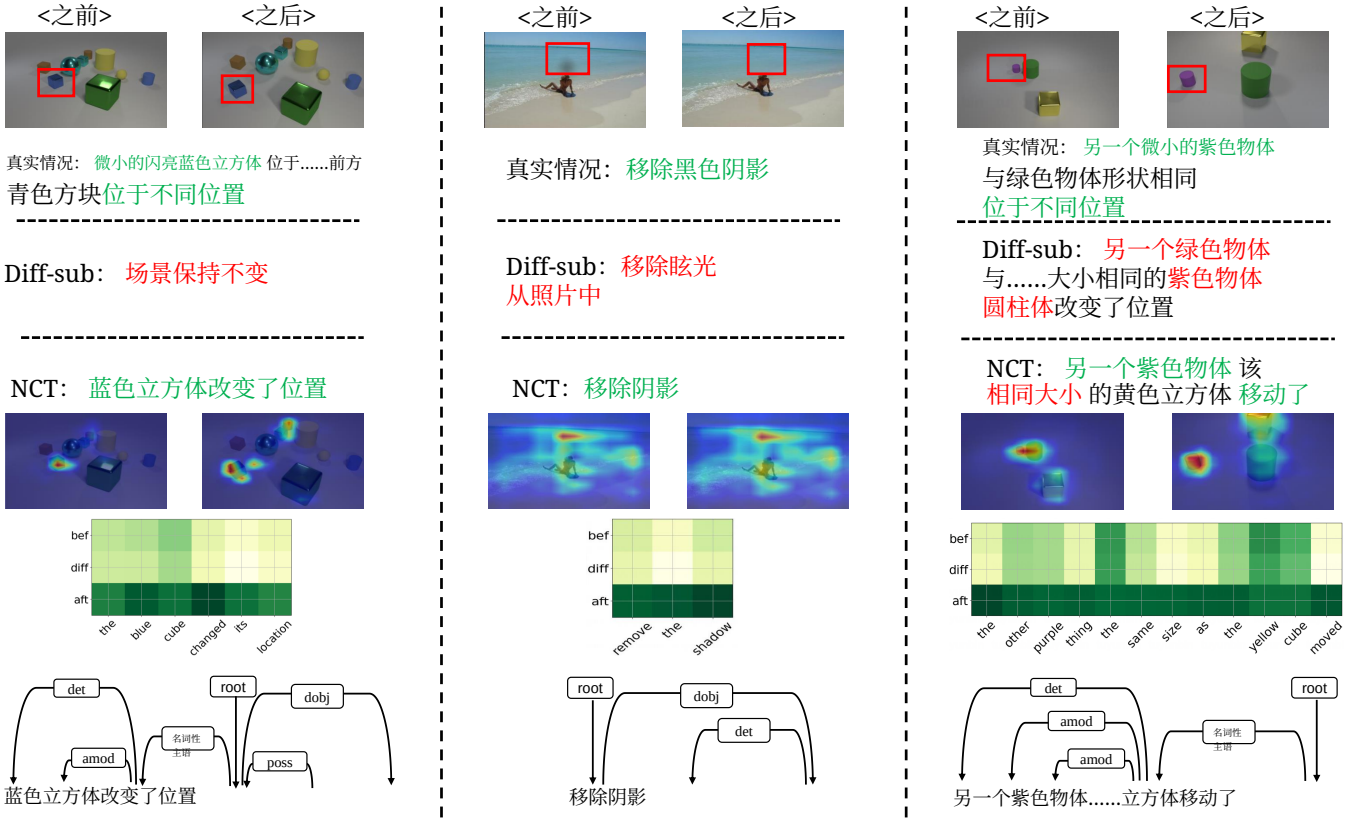


图5. CLEVR-Change、图像编辑请求和 CLEVR-DC 上的定性示例。对于每个示例，我们报告了 Diff-sub 和 NCT 生成的描述文本，以及真实情况（GT）描述文本。描述文本中正确和错误的部分分别用绿色和红色标出。我们在“之前”和“之后”的图像上展示变化定位结果，并通过热力图可视化展示变化特征与对应词语之间的语义对齐。我们展示了每个示例的预测依赖关系。真实变化以红框标出。

取值，描述生成性能和参数数量如表IX所示。

我们发现，随着邻域范围增大，性能并没有明显提升，而且大多数指标上的结果甚至有所下降。我们推测，模型只需要最近的指称对象来引导其定位变化对象，因此邻域范围 3×3 是合适的。基于此，我们通过经验将三个数据集上的 r 设为3。

表IX 研究邻域范围 r参数在三个数据集上的影响，其中 CC、CD 和 IER 分别是 CLEVR-Change、CLEVR-DC 和图像编辑请求的简称。

r	Set	参数	B	M	C	S
3 × 3	CC	26.65M	**55.1**	**40.2**	124.1	**32.9**
5 × 5	CC	26.74M	54.9	39.9	**124.8**	32.7
3 × 3	CD	26.70M	**47.5**	**32.5**	**76.9**	**15.6**
5 × 5	CD	26.79M	44.4	31.3	71.1	14.5
3 × 3	IER	34.07 M	8.1	**15.0**	34.2	**12.7**
5 × 5	IER	34.17M	**9.5**	14.7	**36.9**	11.9

I. 定性分析

为了评估 NCT 在变化定位和描述生成方面的整体性能，我们在 CLEVR-Change、图像编辑请求和 CLEVR-DC，如图 5 所示。对于每个图像对，我们报告 Diff-sub 的基线模型和我们的 NCT 生成的描述文本，以及真实标注（绿色词语）。为评估变化物体的准确性，我们还根据变化检测的注意力权重将变化结果可视化。对于第一个示例，物体移动幅度很小，导致 Diff-sub 误判为没有发生任何变化。对于第二个示例，被移除的物体过于模糊，难以察觉。在这种情况下，Diff-sub 直接对两幅图像做差以计算变化特征，错误地将“阴影”判断为“眩光”，因而无法生成准确的句子。对于第三个示例，极端的视角变化导致所有物体都出现伪移动，使 Diff-sub 错误识别真实变化的物体。此外，这一失败的另一个可能原因是，真实标注句子的句法结构较为复杂。也就是说，相比“紫色物体”，“绿色物体”距离变化类型更近，这可能使 Diff-sub 将变化物体误判为“绿色物体”。与 Diff-sub 相比，所提出的 NCT 能够准确定位并描述变化物体。这主要得益于以下几点：1）邻域特征聚合帮助模型识别真实变化，同时不受视角变化影响；2）公共特征提炼能够有效概括图像对的共同属性，并从每幅图像中提取具有区分性的特征，从而构建它们之间具有对比性的

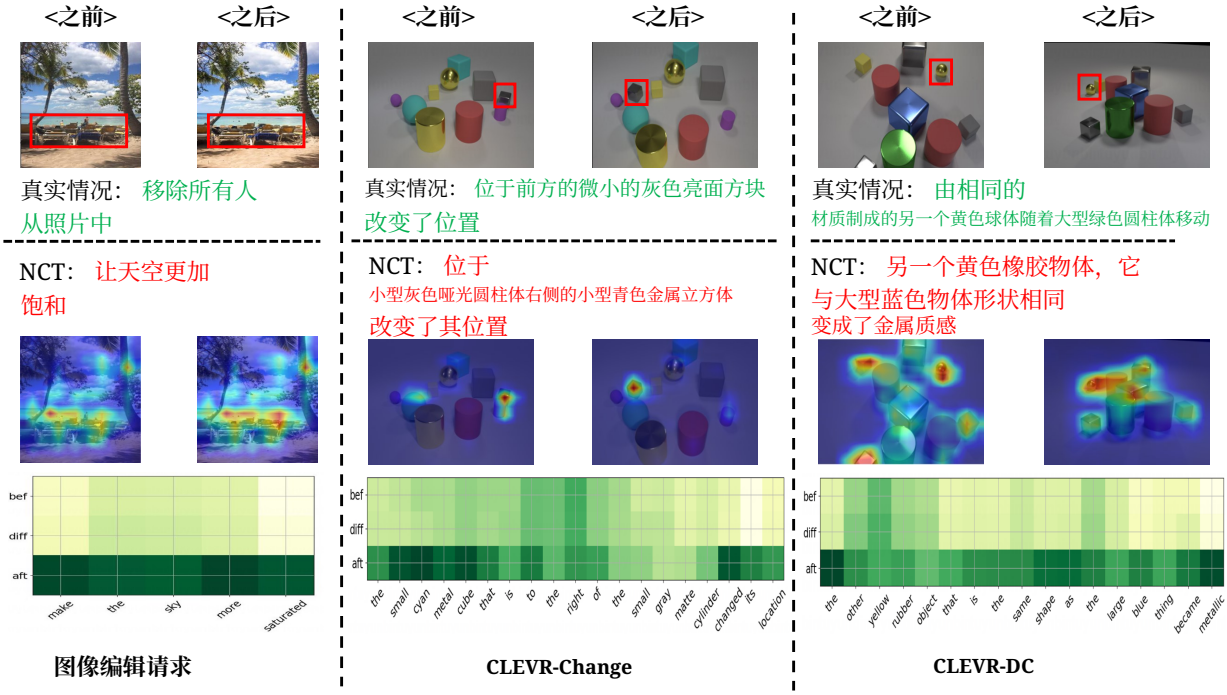


图6. NCT 在图像编辑请求、CLEVR-Change 和 CLEVR-DC 测试集划分上的失败示例。

特征；3）引入词语之间的依存关系有助于解决句子中的句法歧义，并理解其复杂句法结构。例如，在第三个示例中，我们的 NCT 能够预测真实变化的物体“紫色物体”与变化类型“移动”之间的有向链接，从而识别真实变化的物体。

此外，我们发现，在第三个示例中，NCT 预测微小的紫色物体与黄色立方体大小相同。这种误解可能有以下原因：1）NCT 主要根据“之后”图像比较它们的尺寸。2）该数据集的变化类型不包括尺寸变化。在这种情况下，模型难以准确比较两个物体的尺寸。因此，未来研究有必要进一步探索如何识别物体尺寸的变化。补充材料中展示了更多定性示例。

J. 讨论

图6展示了NCT在三个具有不同变化场景的数据集上产生的失败案例。对于所有变化场景，我们可以观察到，所提出的NCT能够成功定位变化物体。然而，它无法用准确的句子描述这些物体。对于这些失败案例，我们推测，每个示例中的变化视觉信号都出现在不显眼的区域，且特征较弱，因此容易被大多数未变化物体的信号淹没。因此，解码器无法获得足够的视觉信息来生成描述。在我们看来，有两种可能的解决方案。其一是利用其他视觉模态来增强网格特征，例如语义分割特征，它能够捕捉更细粒度的视觉信息 [49], ，从而增强对变化信号较弱的物体的特征表示。另一种解决方案是采用从预训练到微调的范式，例如在变化描述数据集上对视觉特征进行微调。具体而言，我们可以使用与我们的框架相配合的预训练特征提取器（例如，视觉Transformer [50]）。随后，我们不冻结其参数，而是将其与所提出的NCT联合训练。这样，训练损失便可以反向传播到特征提取器，从而增强图像特征的表征能力。我们将在后续工作中尝试这些策略。此外，我们注意到，在CLEVR-DC上，可视化的变化定位注意力权重存在噪声。未来，我们将尝试从利用三维知识的角度解决极端视角变化问题。

V. 结论

本文提出邻域对比变换器（NCT），用于在不同变化场景下定位并描述变化。在 NCT 中，邻域特征聚合模块有助于克服视角变化的影响，并在周围显著参照物的引导下快速发现不显著的变化。共同特征蒸馏模块能够从每幅图像中提取共同属性，并学习图像对之间的对比表征。此外，我们引入词语之间的显式依赖关系，以校准 Transformer 的解码器，从而帮助其在训练过程中理解变化描述中的复杂句法结构。大量实验表明，在具有不同变化场景的三个公开数据集上，NCT 的表现大幅优于最先进方法，这也表明它具备良好的泛化能力，能够应对各种变化设定。

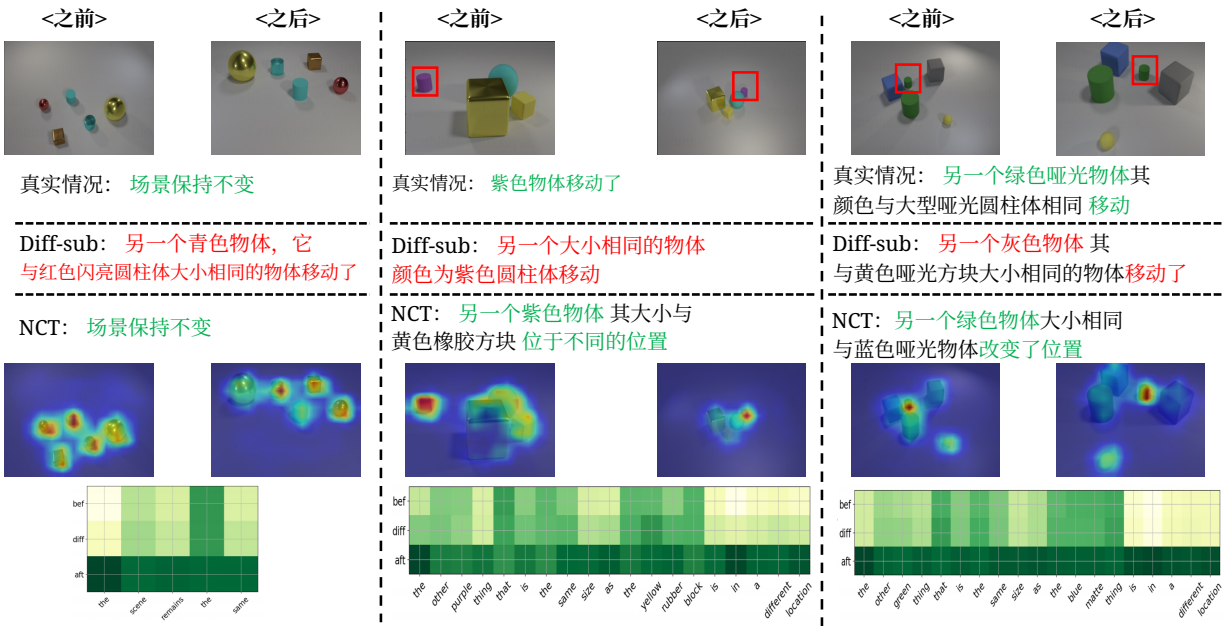
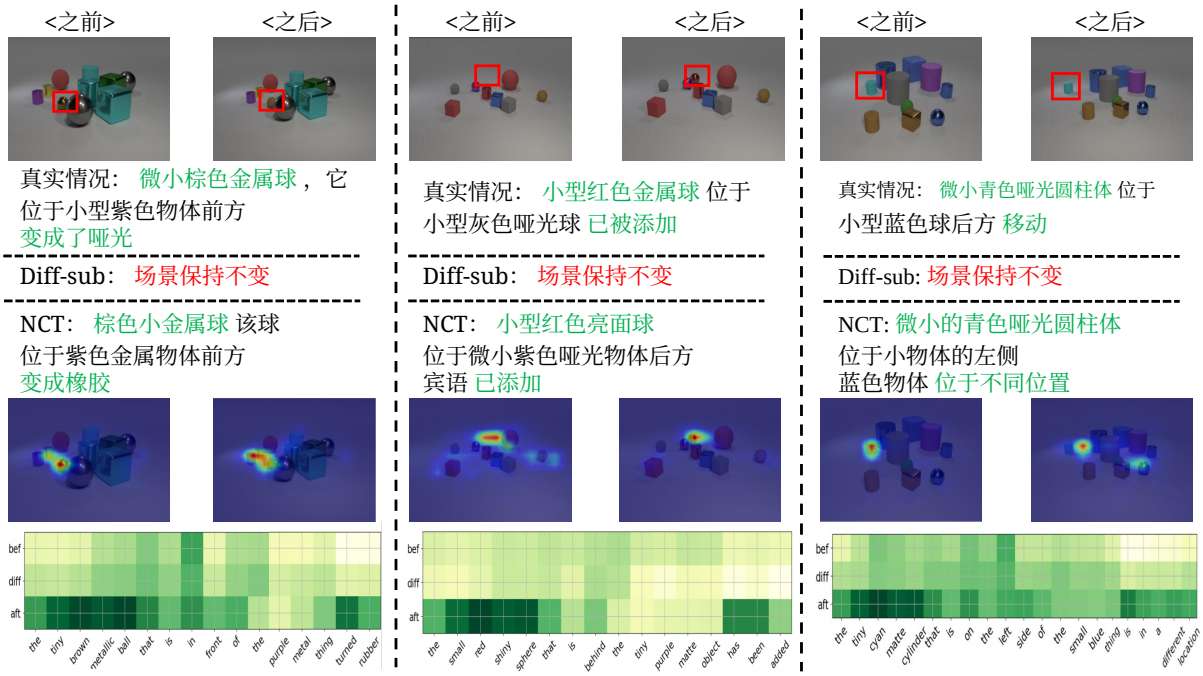


图7. CLEVR-DC测试集划分中的定性示例。



Fig。8. CLEVR-Ch 测试集划分上的定性示例 ange。

附录：实现细节及三个数据集上的更多定性示例

在附录中，我们首先提供了所提方法的更多实现细节。在三个数据集上，我们通过总计 10K 次迭代将模型训练至收敛。训练和推理均使用 PyTorch 在 RTX 3090 GPU 上实现。训练阶段在三个数据集上使用的资源如表 X 所示。可以看出，我们的方法不需要太多资源和训练时间，因此它可以便于其他研究人员轻松复现。

表 X 三个数据集上的训练时间和 GPU内存使用情况

	训练时间	GPU内存
CLEVR-Change	4 小时	13G
CLEVR-DC	2 小时	8G
图像编辑请求	30 分钟	4G

接下来，我们展示三个数据集上有关变化定位和描述生成的更多定性示例，这些示例

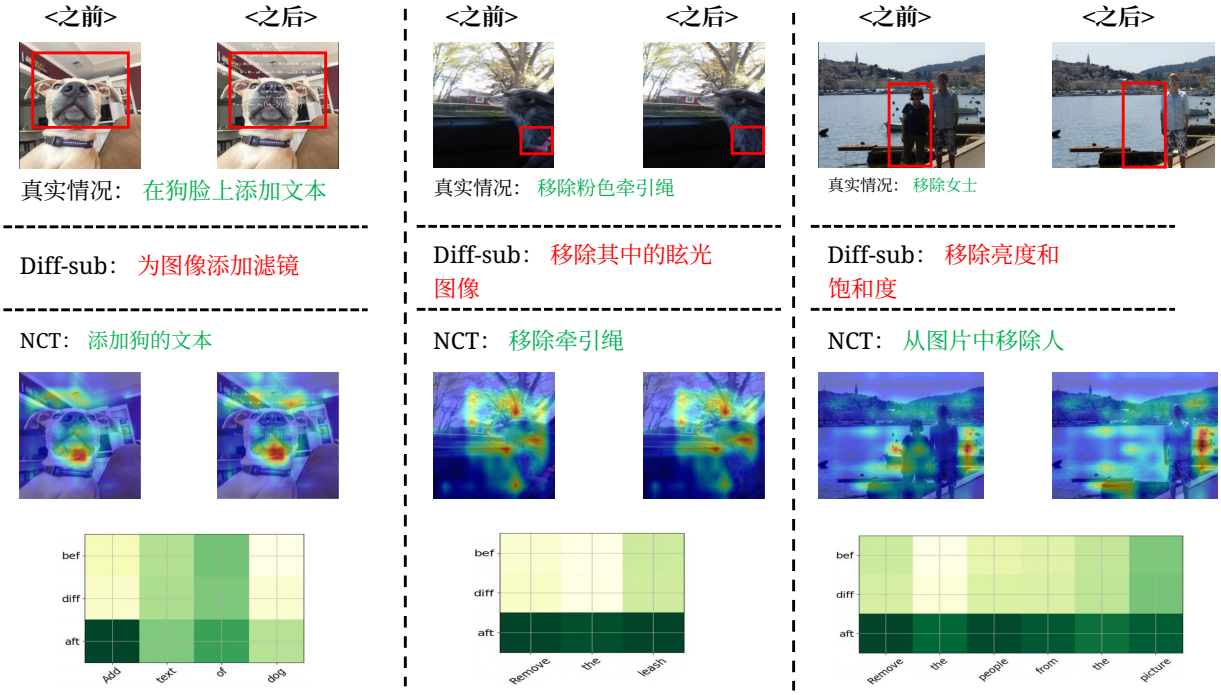


图 9. 图像编辑请求测试集划分上的定性示例。

如图 7–9 所示。对于旨在模拟极端视角变化的 CLEVR-DC 数据集，场景中的所有宾语都存在明显的伪移动，如图 7 所示。这会误导 Diff-sub 基线模型，使其得出错误结果。CLEVR-Change 上的定性示例如图 8 所示。由于变化物体部分遮挡或不明显，基线模型无法定位这些变化，并误判为没有任何变化。相比之下，所提出的 NCT 能准确区分这些细粒度变化与伪变化，并生成相关描述。值得注意的是，在图 7 左侧的示例中，热图突出显示了全部五个宾语。我们推测，热图是根据对比变化定位器（第 III-C 节）的注意力权重生成的。当没有任何变化时，图像对的已学习对比表征不会包含变化对象的信息。背景特征远弱于物体特征。在这种情况下，定位器会关注物体特征，并为每个物体特征分配相似的注意力权重，因此可视化的热图会突出显示全部五个宾语。对于图 9 中的图像编辑请求，我们可以看到，每个示例中的变化信息都非常模糊，难以发现，但与基线模型相比，我们的模型仍能定位变化对象，从而生成理想的描述。我们方法的优越结果主要得益于以下几点：1）邻域特征聚合帮助模型处理无关的视角变化并定位细粒度变化；2）公共特征提炼能够捕获图像对的联合信息，并从每幅图像中提取差异属性，从而构建它们之间的对比特征；3）在词语之间引入显式的依存关系，有助于消除复杂句法结构的歧义训练期间变化描述句中的结构。

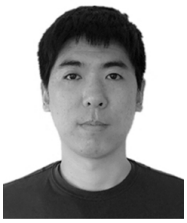
参考文献

[1] J. Wang、B.-K. Bao 和 C. Xu，《Dualvgr：用于视频问答的双视觉图推理单元》，《*IEEE*多媒体汇刊》，第24卷，第3369–3380页，2022年。[2] S. Liu、A. Li、J. Wang 和 Y. Wang，《结合词共现的视频描述双向最大熵训练》，《*IEEE*多媒体汇刊》，第1–1页，2022年。[3] X. Liu、L. Li、S. Wang、Z.-J. Zha、Z. Li、Q. Tian 和 Q. Huang，《用于弱监督指代表达定位的实体增强自适应重建网络》，*IEEETransactions on PatternAnalysis and Machine Intelligence*，第45卷，第3期，第3003–3018页，2023年。[4] J. Wu、T. Chen、H. Wu、Z. Yang、G. Luo 和 L. Lin，《采用全局-局部判别目标的细粒度图像描述生成》，《*IEEE*多媒体汇刊》，第23卷，第2413–2427页，2021年。[5] Z. Zhang、Q. Wu、Y. Wang 和 F. Chen，《在图像描述生成中自适应地从语言上下文探索成对关系》，《*IEEE*多媒体汇刊》，第24卷，第3101–3113页，2022年。[6] H. Jhamtani 和 T. Berg-Kirkpatrick，《学习描述相似图像对之间的差异》，载于*2018*年自然语言处理实证方法会议论文集，2018年，第4024–4034页。[7] H. Tan、F. Dernoncourt、Z. Lin、T. Bui 和 M. Bansal，《通过语言表达视觉关系》，载于计算语言学协会第*57*届年会论文集，2019年，第1873–1883页。[8] X. Shi、X. Yang、J. Gu、S. Joty 和 J. Cai，《发现另一侧的目标：用于变化描述生成的视角适配匹配编码器》，载于欧洲计算机视觉会议。施普林格，2020年，第574–590页。[9] D. H. Park、T. Darrell 和 A. Rohrbach，《鲁棒变化描述生成》，载于*IEEE/CVF*国际计算机视觉会议论文集，2019年，第4624–4633页。[10] M. Hosseinzadeh 和 Y. Wang，《通过辅助任务学习实现图像变化描述生成》，载于*IEEE/CVF*计算机视觉与模式识别会议论文集，2021年，第2725–2734页。[11] H. Kim、J. Kim、H. Lee、H. Park 和 G. Kim，《基于循环一致性的无关变化描述生成》，载于*IEEE/CVF*国际计算机视觉会议论文集，2021年，第2095–2104页。

[12] L. Yao、W. Wang 和 Q. Jin，《结合预训练与对比学习的图像差异描述》，载于AAAI人工智能大会论文集，2022年，第3108–3116页。[13] H. Wu、B. Xiao、N. Codella、M. Liu、X. Dai、L. Yuan 和 L. Zhang，《Cvt：将卷积引入视觉Transformer》，载于IEEE/CVF国际计算机视觉会议论文集，2021年，第22–31页。[14] Y. Tu、C. Zhou、J. Guo、S. Gao 和 Z. Yu，《增强视频描述生成中目标词与对应帧之间的对齐》，《模式识别》，第111卷，文章编号107702，2021年。[15] L. Li、X. Gao、J. Deng、Y. Tu、Z.-J. Zha 和 Q. Huang，《用于视频描述生成的全局门控长短期关系Transformer》，《IEEE图像处理汇刊》，第31卷，第2726–2738页，2022年。[16] Y. Tu、L. Li、L. Su、S. Gao、C. Yan、Z.-J. Zha、Z. Yu 和 Q. Huang，《I2Transformer：用于电视节目描述的内外关系嵌入变换器》，《IEEE图像处理汇刊》，第31卷，第3565–3577页，2022年。[17] L. Yu、J. Zhang 和 Q. Wu，《用于图像描述生成的金字塔特征图双重注意力》，《IEEE多媒体汇刊》，第24卷，第1775–1786页，2022年。[18] J. Ji、X. Huang、X. Sun、Y. Zhou、G. Luo、L. Cao、J. Liu、L. Shao 和 R. Ji，《用于图像描述生成的多分支距离敏感自注意力网络》，《IEEE多媒体汇刊》，第1–1页，2022年。[19] Y. Tu、C. Zhou、J. Guo、H. Li、S. Gao 和 Z. Yu，《通过图学习实现视频描述生成的关系感知注意力》，《模式识别》，第136卷，文章编号109204，2023年。[20] Y. Tu、T. Yao、L. Li、J. Lou、S. Gao、Z. Yu 和 C. Yan，《用于变化描述生成的语义关系感知差异表征学习》，载于计算语言学协会研究成果：ACL-IJCNLP 2021，2021年，第63–73页。[21] N. Vo、L. Jiang、C. Sun、K. Murphy、L.-J. Li、L. Fei-Fei 和 J. Hays，《组合文本与图像以实现图像检索：一次实证探索》，载于CVPR，2019年，第6439–6448页。[22] Q. Huang、Y. Liang、J. Wei、Y. Cai、H. Liang、H.-f. Leung 和 Q. Li，《采用实例级细粒度特征表示的图像差异描述》，《IEEE多媒体汇刊》，第24卷，第2004–2017页，2021年。[23] Y. Qiu、S. Yamamoto、K. Nakashima、R. Suzuki、K. Iwata、H. Kataoka 和 Y. Satoh，《使用Transformer描述并定位多个变化》，载于IEEE/CVF国际计算机视觉会议论文集，2021年，第1971–1980页。[24] Z. Liao、Q. Huang、Y. Liang、M. Fu、Y. Cai 和 Q. Li，《用于变化描述生成的含三维信息场景图》，载于第29届ACM国际多媒体会议论文集，2021年，第5074–5082页。[25] F. Liu、C. Yin、X. Wu、S. Ge、P. Zhang 和 X. Sun，《用于自动生成胸部X光报告的对比注意力》，载于计算语言学协会研究成果：ACL-IJCNLP 2021，2021年，第269–280页。[26] Z. Li、Q. Tran、L. Mai、Z. Lin 和 A. L. Yuille，《通过自注意力和对比特征实现上下文感知群组描述》，载于IEEE/CVF计算机视觉与模式识别会议论文集，2020年，第3440–3450页。[27] J. Hou、X. Wu、W. Zhao、J. Luo 和 Y. Jia，《面向视频描述生成的联合句法表征学习与视觉线索转换》，载于IEEE/CVF国际计算机视觉会议（ICCV）论文集，2019年10月。[28] B. Wang、L. Ma、W. Zhang、W. Jiang、J. Wang 和 W. Liu，《基于门控融合网络和词性序列引导的可控视频描述生成》，载于IEEE/CVF国际计算机视觉会议论文集，2019年，第2641–2650页。[29] J. Zhang、K. Mei、Y. Zheng 和 J. Fan，《将词性引导融入图像描述生成》，《IEEE多媒体汇刊》，第23卷，第92–104页，2021年。[30] J. Deng、L. Li、B. Zhang、S. Wang、Z. Zha 和 Q. Huang，《用于视频描述生成的句法引导层次注意力网络》，IEEE Transactions on Circuits and Systems for Video Technology，第32卷，第2期，第880–892页，2022年。[31] Q. Zheng、C. Wang 和 D. Tao，《用于视频描述生成的句法感知动作定位》，载于IEEE/CVF计算机视觉与模式识别会议论文集，2020年，第13 096–13 105页。[32] W. Zhao、X. Wu 和 J. Luo，《用于视频描述生成的多模态依存树》，《神经信息处理系统进展》，第34卷，第6634–6645页，2021年。[33] D. Zhao、Z. Song、Z. Ji、G. Zhao、W. Ge 和 Y. Yu，《用于语义对应的多尺度匹配网络》，载于IEEE/CVF国际计算机视觉会议论文集，2021年，第3354–3364页。[34] J. L. Ba、J. R. Kiros 和 G. E. Hinton，《层归一化》，*arXiv* 预印本 *arXiv:1607.06450*，2016年。[35] Y. Tu、L. Li、C. Yan、S. Gao 和 Z. Yu，《R^3Net：用于变化描述的关系嵌入式表示重建网络》，载于*2021*年自然语言处理实证方法会议论文集，2021年，第9319–9329页。[36] K. Papineni、S. Roukos、T. Ward 和 W.-J. Zhu，《BLEU：一种用于机器翻译自动评估的方法》，载于第*40*届计算语言学协会年会论文集，2002年，第311–318页。[37] S. Banerjee 和 A. Lavie，《Meteor：一种与人工判断相关性更高的机器翻译评估自动指标》，载于机器翻译和／或摘要的内在与外在评估指标*ACL*研讨会论文集，2005年，第65–72页。[38] C.-Y. Lin，《ROUGE：一种用于摘要自动评估的工具包》，载于文本摘要新方向，2004年，第74–81页。[39] R. Vedantam、C. Lawrence Zitnick 和 D. Parikh，《CIDEr：基于共识的图像描述评估》，载于*IEEE*计算机视觉与模式识别会议论文集，2015年，第4566–4575页。[40] P. Anderson、B. Fernando、M. Johnson 和 S. Gould，《SPICE：语义命题式图像描述评估》，载于欧洲计算机视觉会议。施普林格，2016年，第382–398页。[41] X. Chen、H. Fang、T.-Y. Lin、R. Vedantam、S. Gupta、P. Doll´ar 和 C. L. Zitnick，《Microsoft COCO Captions：数据收集与评估服务器》，*arXiv* 预印本 *arXiv:1504.00325*，2015年。[42] K. He、X. Zhang、S. Ren 和 J. 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，2016年，第770–778页。[43] O. Russakovsky、J. Deng、H. Su、J. Krause、S. Satheesh、S. Ma、Z. Huang、A. Karpathy、A. Khosla、M. Bernstein 等，《ImageNet大规模视觉识别挑战赛》，国际计算机视觉期刊，第115卷第3期，第211–252页，2015年。[44] T. Dozat 和 C. D. Manning，《用于神经依存句法分析的深度双仿射注意力》，*arXiv* 预印本 *arXiv:1611.01734*，2016年。[45] D. P. Kingma 和 J. Ba，《Adam：一种随机优化方法》，*arXiv* 预印本 *arXiv:1412.6980*，2014年。[46] A. Paszke、S. Gross、F. Massa、A. Lerer、J. Bradbury、G. Chanan、T. Killeen、Z. Lin、N. Gimelshein、L. Antiga 等，《PyTorch：一种命令式风格的高性能深度学习库》，神经信息处理系统进展，第32卷，2019年。[47] A. Bochkovskiy、C.-Y. Wang 和 H.-Y. M. Liao，《Yolov4：实现目标检测速度与精度的最佳平衡》，*arXiv* 预印本 *arXiv:2004.10934*，2020年。[48] C. Godard、O. Mac Aodha、M. Firman 和 G. J. Brostow，《深入探讨自监督单目深度估计》，载于*IEEE/CVF*国际计算机视觉会议论文集，2019年，第3828–3838页。[49] M. Wu、X. Zhang、X. Sun、Y. Zhou、C. Chen、J. Gu、X. Sun 和 R. Ji，《DifNet：增强图像描述的视觉信息流》，载于*IEEE/CVF*计算机视觉与模式识别会议论文集，2022年，第18 020–18 029页。[50] 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：用于任意尺度图像识别的Transformer》，载于第*9*届国际学习表征会议，*ICLR*，2021年。



Yunbin Tu 获得杭州电子科技大学自动化专业理学学士学位，以及昆明理工大学模式识别与智能系统专业理学硕士学位。目前，他正在中国科学院大学计算机科学与技术学院攻读博士学位。他的研究兴趣包括多媒体内容分析，尤其是视频和变化描述。



Liang Li 于2008年获得西安交通大学理学学士学位，并于2013年获得中国科学院北京计算技术研究所博士学位。2013年至2015年，他在中国科学院大学计算机与控制工程系从事博士后研究。目前，他担任中国科学院计算技术研究所副教授。他还曾在多个国际期刊的委员会任职，并

会议论文。Li博士已在经同行评审的期刊和会议上发表了60余篇论文。他的研究兴趣包括多媒体内容分析、计算机视觉和模式识别。



Li Su 于2009年获得中国科学院研究生院计算机科学博士学位。她目前是中国科学院大学计算机科学与技术学院教授。她的研究兴趣包括媒体计算和内容分析。



Ke Lu （IEEE高级会员）1971年3月13日出生于宁夏。他分别于1998年7月和2003年7月获得中国西安西北大学数学系硕士学位和计算机科学系博士学位。2003年7月至2005年4月，他在中国北京中国科学院自动化研究所从事博士后研究。目前，他是中国科学院大学的杰出教授。他还兼任双重

受聘于中国广东深圳的鹏城实验室。他目前的研究领域主要包括计算机视觉、三维图像重建和计算机图形学。



Qingming Huang 于1988年和1994年分别获得哈尔滨工业大学计算机科学理学学士学位和博士学位。他现任中国科学院大学计算机科学与技术学院讲席教授、副院长。他合著了400余篇学术论文，发表于IEEE TPAMI、TIP、TKDE、TMM和TCSVT等国际期刊，以及NeurIPS、ACM Multimedia、

ICCV、CVPR、ECCV、VLDB、AAAI和IJCAI等国际顶级会议。他是IEEE会士。目前的研究兴趣包括多媒体计算、图像/视频处理、模式识别和计算机视觉。