

MANGO: 基于多模态注意力的正则化流融合学习方法

Thanh-Dat Truong¹, Christophe Bobda², Nitin Agarwal³, Khoa Luu¹

¹CVIU Lab, University of Arkansas, USA ²University of Florida, USA

³University of Arkansas at Little Rock, USA

{ tt032, khoaluu } @uark.edu cbobda@ece.ufl.edu, nxagarwal@ualr.edu

<https://uark-cviu.github.io>

Abstract

近年来，多模态学习取得了很大成功。然而，目前的多模态融合方法采用 Transformer 的注意机制来隐式地学习多模态特征的潜在相关性。因此，多模态模型无法捕捉每种模态的基本特征，难以理解多模态输入的复杂结构和相关性。本文介绍了一种新颖的基于多模态注意的正态流 (MANGO) 方法，以开发显式、可解释和易处理的多模态融合学习。特别地，我们提出了一种新的可逆交叉注意 (ICA) 层，以开发基于正态流的多模态数据模型。为了在我们提出的可逆交叉注意层中有效捕捉多模态数据的复杂潜在相关性，我们提出了三种新的交叉注意机制：模态对模态交叉注意 (MMCA)、模态间交叉注意 (IMCA) 和可学习模态间交叉注意 (LICA)。最后，我们引入了一种新的基于多模态注意的正态流，以实现我们提出的方法在多维多模态数据上的可扩展性。我们在三种不同的多模态学习任务上，即语义分割、图像到图像翻译和电影类型分类的实验结果，展示了该方法的最新 (SoTA) 性能。

1 简介

人类的感知通过多种输入渠道（如视觉、文本或音频）以多模态方式解释周围的世界。基于深度学习的多模态融合方法极大地提高了各种问题的性能，例如，分类 [20, 38, 37, 58]，动作识别 [12, 54, 57]，语义分割 [65, 56, 59, 61, 60]，目标检测 [72]。最近推出的大型多模态模型，如 ChatGPT [1]、Gemini [52]、Chaemelon [51]、LLaMMA [53] 等，旨在通用辅助用途的应用中也表现出了令人印象深刻的性能。

多模态融合方法取得关键成功的关键在于对输入模态之间的互动和关联建模机制。最近的方法 [20, 65] 采用 Transformer 的注意力机制 [64] 来隐式建模跨模态的关联。通过在大规模数据上训练，注意力模型能够隐式学习数据中表现出的底层关联。例如，视觉-语言融合模型 [28, 27, 48] 使用早期融合，其中视觉标记和文本标记同时输入到 Transformer 模型中。然后，Transformers 将通过注意力机制的二阶关联学习来学习视觉和文本标记之间的关联与对齐。在这种形式下，这些多模态融合方法是不考虑对齐的，即对跨模态的对齐和关联没有充分利用 [65]。此外，隐式融合方法往往在不明确建模每个模态的独特特征和关联的情况下关联跨模态的信息。这样可能忽略特定模态的贡献，尤其是在一个模态包含更多数据或更强的信号时，会导致次优表现

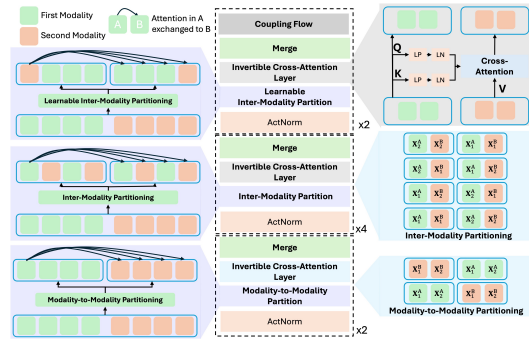


Figure 1: 通过具有可逆交叉注意的多模态归一化流的方法进行跨模态融合

[65, 20]。由于隐式方法无法单独建模每个模态的重要性, 这些方法可能无法捕捉多模态数据中复杂的结构和互补信息。隐式建模方法也缺乏可解释性, 因为难以理解或表示每个模态对输出的贡献。其他方法 [33, 11, 8] 采用后期融合, 即在每个模态被确定后再融合特征。然而, 后期融合忽略了跨模态的低层次互动。因此, 与单模态方法相比, 直接采用注意力的融合无法提高性能 [65, 20]。

尽管最近大多数的多模态方法采用注意力机制来隐式捕捉多模态相关性, 但对显式建模方法的研究较少。基于正态变换的模型是一种常见的显式建模方法。通过在数据和潜在空间之间的双射映射来建模数据的精确似然性, 基于正态变换的模型允许稳定和可靠的训练, 使得对基础多模态数据分布的模型表示有更好的理解。特别地, 通过堆叠一组双射变换, 显式模型可以构造复杂的分布, 使它们能够以直接控制参数的方式捕捉多模态数据分布。因此, 这种显式建模方法提高了解释性, 并且能更好地理解潜在空间中的多模态特征和相关性, 这在之前的方法中是具有挑战性的。与之前的方法相比, 通过正态变换进行显式建模提供了一种更好的多模态融合机制, 因为它可以捕捉多模态数据的基础结构和相关性, 而不会让任何单一模态占主导地位。因此, 显式建模能够实现更精确、灵活和稳健的多模态融合, 提高在需要理解和良好对齐多模态数据的任务中的表现。

多模态归一化流的挑战。虽然显式建模是多模态融合的一种潜在方法, 但开发多模态归一化流需要付出许多努力。事实上, 当前基于归一化流的模型存在两个重要限制。首先, 尽管仿射耦合层 [5, 21] 赋予了模型易处理性和可逆性的特性, 但这一层限制了模型的表达能力。与 Transformer 中的注意力机制不同 [64], 耦合层无法捕捉多模态数据 [48] 中的大范围数据依赖性和相关性。其次, 将基于归一化流的模型扩展到高维数据是一个具有挑战性的问题。这需要在模型中堆叠更多的双射层, 导致训练期间的高计算成本和难以收敛 [5]。虽然隐式建模方法通过使用潜在模型 (例如, 潜在扩散 [42]) 缓解了计算开销, 但对基于归一化流的方法中解决此开销问题的研究有限。因此, 迫切需要解决这些限制, 以开发一种高效的基于多模态归一化流的模型。

本工作的贡献。本文引入了一种新的多模态基于注意力的正规化流 (MANGO), 这是一种明确、可解释、易处理的方法, 用于解决多模态融合问题 (图 1)。据我们所知, 这是首批将正规化流方法应用于多模态融合学习的研究之一。我们的贡献可以总结如下。首先, 我们提出了一种新的可逆交叉注意 (ICA) 层, 用于基于正规化流的模型。所提出的 ICA 层可以有效解决标准正规化流中耦合层的局限性, 同时保持其可处理性和可逆性。其次, 为了捕捉模态之间的相关性和对齐性, 我们提出了三种新的分区交叉注意机制, 包括模态到模态交叉注意 (MMCA)、模态间交叉注意 (IMCA) 和可学习式模态间交叉注意 (LICA)。第三, 我们提出了一种新颖的基于多模态注意的正规化流方法, 结合潜在模型以实现其对高维多模态数据融合的可扩展性。我们的方法可以解决计算负担的限制, 同时有效地建模多模态数据中的复杂相关性。最后, 我们在三个多模态学习任务上进行了实验, 即语义分割、图像到图像的翻译和电影类型分类, 结果显示 MANGO 在不同方面的有效性, 与现有的多模态模型相比, 展现了其领先的性能。

2 相关工作和背景

2.1 相关工作

注意力模型。Transformer 中的注意力机制在单模态和多模态学习中表现出色 [64, 28, 65]。通过使用二阶相关性, 注意力机制能够捕捉跨输入模态的长期关系。Transformer 中有两种常见的注意力类型, 即自注意力和交叉注意力。自注意力专注于学习单一输入模态内的相关性 [64], 而交叉注意力则对跨模态的关系进行建模, 使得模型可以分析从一个模态到另一个模态的复杂相关性 [68]。Transformer 已成为一种主流方法, 并且在开发各种多模态任务中产生了深远的影响, 例如大型视觉-语言模型 [28, 27]、RGB-D 物体分割 [65]。

多模态融合。多模态融合学习在各种任务中相比单一模态表现出了显著优势, 例如语义分割 [65, 20]、图像到图像翻译 [18]、动作识别 [12]、目标检测 [72] 等。早期的多模态融合学习方法采用简单的特征拼接来融合来自多种模态的信息 [7, 76]。随后, 后来的工作通过使用神经网络进行深度融合进一步改进了不同模态间的融合, 例如使用 RNN [2]、LSTM [50]、注意力机制 [55, 35] 等。另一种方法采用神经架构搜索来寻找适用于多模态融合的合适网络结构 [25, 73, 10]。目前最先进的融合方法利用早期融合, 在数据层面通过 Transformers [65, 20, 28] 捕捉跨模态交互。通过在初始阶段结合所有模态作为输入标记, Transformers 将通过自注意力机制学习跨模态间的相关性 [64]。后续工作通过像素级融合 [20]、剪枝技术 [65] 或动态多模态融合 [70] 进一步改进了早期融合方法。然而, 需要注意的是, 这些当前的多模态融合方法是一种隐式建模的方法。

显式建模通过正则流。为了开发可逆网络，RealNVP [5] 首先引入了一个仿射耦合层，其中它的逆版本和雅可比矩阵的对数行列式可以被轻松计算。后续的工作 [4, 21] 通过引入非线性独立成分估计 [4]、可逆卷积 [21, 34]、激活正则化 [21]、自回归建模 [17]、多尺度架构 [5]、等变正则流 [9] 进一步改进了耦合层。另一种方法 [15, 48] 通过在缩放和平移网络中使用 Transformers 增强了耦合层的表现力。然而，它仍然无法解决数据中的长距离依赖和复杂的跨模态相关性的问题。最近的研究进一步开发了基于条件流的方法，例如，条件图像合成 [31, 32]，使用条件可逆网络 [47]，或两个可逆网络 [49]。

2.2 归一化流的局限性

典型的归一化流模型 [5, 4, 21] 是通过可逆仿射耦合层设计的为：

$$\begin{aligned} \mathbf{X}_1, \mathbf{X}_2 &= \text{partition}(\mathbf{X}) \\ \mathbf{Y}_1 &= \mathbf{X}_1, \quad \mathbf{Y}_2 = \mathbf{X}_2 \odot \exp(\mathcal{S}(\mathbf{X}_1)) + \mathcal{T}(\mathbf{X}_1) \\ \mathbf{Y} &= \text{merge}([\mathbf{Y}_1, \mathbf{Y}_2]) \end{aligned} \quad (1)$$

，其中 $\mathbf{X}$ 是输入， partition 是一种划分方法，例如，RealNVP [5] 采用棋盘划分方法， $\mathcal{S}$ 和 $\mathcal{T}$ 是深度神经网络， merge 是合并函数， $\odot$ 是按元素的矩阵乘法。

限制。流模型的成功依赖于可逆层的设计。然而，目前的仿射耦合层在建模复杂数据时效率仍然不高。首先，由于设计简单，耦合层的表达能力有限。通过残差网络 [5] 设计的 $\mathcal{S}$ 和 $\mathcal{T}$ 无法捕捉高维数据中表示的复杂关系。因此，仍然很难捕捉数据中高度复杂的依赖关系或相关性，尤其是在多模态数据中。其次，扩展到高维数据会增加流模型的复杂性，可能导致训练不稳定且效率低下。如果耦合层的数量较少，模型可能会无法捕捉多模态数据中的复杂关系和依赖性。这会导致在密度估计或融合建模等任务中的表现不佳。因此，高维数据还需要更多的层来捕捉所有令牌之间的必要相关性，从而增加计算成本。在本文中，我们将开发一种新的基于注意力的归一化流方法，以解决归一化流和多模态融合中的这些限制。

3 提出的基于多模态注意力的归一化流 (MANGO) 方法

最近的大多数多模态模型采用具有注意力机制的 Transformer 来学习跨模态相关性 [65, 20, 27]。然而，之前的研究表明这种融合方法效率不高 [35]。实际上，通过自监督或弱监督学习的相关性无法在模态间提供明确的注意力建模，并且当多模态输入的信息稀疏时会变得无效。此外，由于跨模态相关性通常是高维且复杂的，开发能够捕捉复杂相关性的多模态模型具有挑战性。

因此，为了解决这个问题，本文将跨模态关联建模为联合分布。然后，可以使用基于 Normalizing Flow 的模型进一步对联合分布进行建模，这是一种可处理但强大的用于建模具有双射映射函数的复杂分布的方法。图 2 展示了我们提出的基于多模态注意力的 Normalizing Flow 框架的概述。形式上，令 $\mathbf{X}$ 为多模态输入（例如，RGB 和深度图像）， G 为将输入映射到潜在空间的双射网络，即 $\mathbf{Z} = G(\mathbf{X})$ 。预测 $\hat{\mathbf{Y}}$ 可以通过投射头得到，表示为 $\hat{\mathbf{Y}} = \text{TaskHead}(\mathbf{Z})$ ，其中 TaskHead 为产生任务特定输出（例如，语义分割）的投射头。然后，多模态数据分布 $p(\mathbf{X})$ 可以通过基于 Normalizing Flow 的模型 G 形成，如方程 (2)。在方程

$$p(\mathbf{X}) = \pi(\mathbf{Z}) \left| \frac{\partial G(\mathbf{X})}{\partial \mathbf{X}} \right| \quad (2)$$

中， $\pi(\mathbf{Z})$ 是先验正态分布。在我们的方法中，我们假设输入 $\mathbf{X}$ 可以被标记为 $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_N]$ ，其中 N 是标记的数量。为简化起见，我们假设 $\mathbf{X}$ 包含两个输入模态（例如，RGB 和深度图像），即 $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_M, \mathbf{x}_{M+1}, \dots, \mathbf{x}_N]$ ，其中 $[\mathbf{x}_1, \dots, \mathbf{x}_M]$ 和 $[\mathbf{x}_{M+1}, \dots, \mathbf{x}_N]$ 分别属于第一和第二模态。

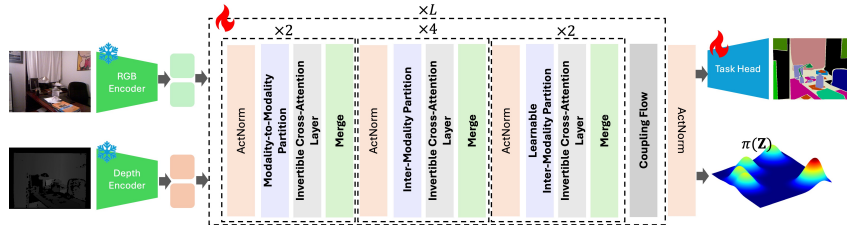


Figure 2: 我们提出的基于多模态注意力的归一化流 (MANGO) 方法来进行融合学习。

3.1 提出的可逆交叉注意力 (ICA)

我们引入了一种新颖的可逆交叉注意力来解决基于归一化流模型的先前限制。注意力机制的成功依赖于通过二阶相关性来探索特征之间关系的能力。特别地，注意力层的设计可以被表述为 $\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q} \times \mathbf{K}^T}{\sqrt{d}}\right) \mathbf{V}$ ，其中 $\mathbf{Q}$ 、 $\mathbf{K}$ 和 $\mathbf{V}$ 是通过将输入 $\mathbf{X}$ 应用线性投影获得的查询、键和数值特征， $\times$ 是缩放点积。查询和键用于通过缩放点积学习注意力权重。然后，这一注意力信息累积到数值向量中，这使得最终输出能够携带标记之间的相关性。受到这种注意力设计的启发，我们在耦合层中提出了 ICA，如方程 (3) 所示。

$$\begin{aligned} \mathbf{X}_1, \mathbf{X}_2 &= \text{partition}([\mathbf{x}_1, \dots, \mathbf{x}_N]) \\ \mathbf{Q} &= \text{LN}(\text{LP}(\mathbf{X}_1)), \quad \mathbf{K} = \text{LN}(\text{LP}(\mathbf{X}_1)), \quad \mathbf{V} = \mathbf{X}_2 \\ \mathbf{Y}_1 &= \mathbf{X}_1, \quad \mathbf{Y}_2 = \text{softmax}\left(\frac{\mathbf{Q} \times \mathbf{K}^T}{\sqrt{d}}\right) \mathbf{V} \\ \mathbf{Y} &= \text{merge}([\mathbf{Y}_1, \mathbf{Y}_2]) \end{aligned} \quad (3)$$

，其中 LN 是层归一化，LP 是线性投影， d 是特征维度。这种交叉注意力机制旨在通过注意力权重来建模标记之间的交互。输入的第二个补丁 ($\mathbf{X}_2$) 中的注意力信息被嵌入到输入的第二个补丁 ($\mathbf{X}_1$) 中。通过扩展到多个可逆交叉注意力层并交替标记分区，我们提出的方法可以有效捕捉输入之间的相关性，尤其是在多模态数据中，因为输入分区之间的注意力信息会交错交换。

可逆性。当前大规模生成模型（如大型语言模型 (LLM) [53]、大型视觉语言模型 (LVM) [28, 27]）的成功依赖于自回归建模。实际上，自回归形式自然地与数据的性质相匹配，其中每个输入符号依赖于前面的符号。这种建模方法能够模拟多模态数据中的高度复杂的依赖关系并保持一致性和连贯性。在我们的学习方法中，我们建议通过自回归形式来建模可逆注意力层。尤其是，我们的 ICA 层在式子 (3) 中可以按照式子 (4) 进行改造。

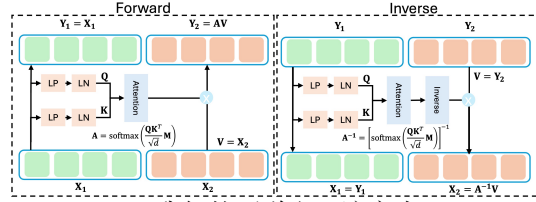


Figure 3: 我们的可逆交叉注意力 (ICA)。

$$\mathbf{Y}_2 = \text{softmax}\left(\frac{\mathbf{Q} \times \mathbf{K}^T}{\sqrt{d}} \mathbf{M}\right) \mathbf{V} \quad (4)$$

其中 $\mathbf{M}$ 是上三角矩阵，以确保自回归建模特性。在这种形式下，我们 ICA 的逆过程可以按照式子 (5) 进行构建。

$$\begin{aligned} \mathbf{Y}_1, \mathbf{Y}_2 &= \text{partition}([\mathbf{y}_1, \dots, \mathbf{y}_N]) \\ \mathbf{Q} &= \text{LN}(\text{LP}(\mathbf{Y}_1)), \quad \mathbf{K} = \text{LN}(\text{LP}(\mathbf{Y}_1)), \quad \mathbf{V} = \mathbf{Y}_2 \\ \mathbf{X}_1 &= \mathbf{X}_1, \quad \mathbf{X}_2 = \left[\text{softmax}\left(\frac{\mathbf{Q} \times \mathbf{K}^T}{\sqrt{d}} \mathbf{M}\right) \right]^{-1} \mathbf{V} \\ \mathbf{X} &= \text{merge}([\mathbf{X}_1, \mathbf{X}_2]) \end{aligned} \quad (5)$$

图 3 展示了 ICA 层的前向和反向过程。设 $\mathbf{A} = \text{softmax}\left(\frac{\mathbf{Q} \times \mathbf{K}^T}{\sqrt{d}} \mathbf{M}\right)$ 为交叉注意力矩阵。得益于自回归建模，逆矩阵 $\mathbf{A}$ 始终存在，因为 $\mathbf{A}$ 是上三角矩阵。需要注意的是，由于 softmax 的性质， $\mathbf{A}$ 的对角线始终大于 0。因此，我们的方法可以有效地确保交叉注意力层的可逆性。受 [48, 64] 启发， d 将是一个可学习的参数，用于捕捉一个通用的尺度。

可操作性。基于正常化流模型的一个关键属性是雅可比矩阵行列式的可操作性，即 $\det\left(\frac{\partial \mathbf{Y}}{\partial \mathbf{X}}\right)$ 。形式上，我们 ICA 的雅可比矩阵行列式可以如公式 (??) 中形成。由于 $\mathbf{A}$ 是一个上块三角矩阵，因为其自回归形式，行列式可以简单地计算为矩阵对角线上的乘积。

3.2 跨模态注意力的分区方法

如方程 (3) 所示，分区方法在学习不同模态之间的相关性中起着至关重要的作用，因为它将决定在可逆的交叉注意力层中哪些注意信息将被交换。为了支持不同模态间的相关性学习，我们提出设计三种不同的分区方法来捕捉不同类型的跨模态注意力（图 4）。

为简单起见，我们将多模态输入 $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_M, \mathbf{x}_{M+1}, \dots, \mathbf{x}_N]$ 重写为 $\mathbf{X} = [\mathbf{x}_1^A, \dots, \mathbf{x}_M^A, \mathbf{x}_1^B, \dots, \mathbf{x}_K^B]$ ，其中 K 是第二模态的 token 数，即 $N = M + K$ 。

模态到模态的交叉注意力 (MMCA)。为了捕捉从第一个模态到第二个模态 (或反之) 的注意力, 方程 (3) 中的分区函数可以形成如下:

$$\underbrace{\begin{cases} \mathbf{X}_1 = [\mathbf{x}_1^A, \dots, \mathbf{x}_M^A] \\ \mathbf{X}_2 = [\mathbf{x}_1^B, \dots, \mathbf{x}_K^B] \end{cases}}_{\text{partition}_{A \rightarrow B}^{MMCA}} \text{ or } \underbrace{\begin{cases} \mathbf{X}_1 = [\mathbf{x}_1^B, \dots, \mathbf{x}_K^B] \\ \mathbf{X}_2 = [\mathbf{x}_1^A, \dots, \mathbf{x}_M^A] \end{cases}}_{\text{partition}_{B \rightarrow A}^{MMCA}} \quad (6)$$

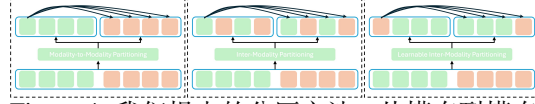


Figure 4: 我们提出的分区方法: 从模态到模态的交叉注意力 (左)。模态间交叉注意力 (中)。可学习的模态间交叉注意力 (右)。

其中 $\text{partition}_{A \rightarrow B}^{MMCA}$ 和 $\text{partition}_{B \rightarrow A}^{MMCA}$ 是模态到模态的分区方法。

虽然第一个分区方法 $\text{partition}_{A \rightarrow B}^{MMCA}$ 允许 ICA 层捕捉第一种模态与第二种模态的相关性, $\text{partition}_{B \rightarrow A}^{MMCA}$ 将以相反的方向建模注意力, 即从第二种模态到第一种模态。在这种方法下, 可以有效地跨模态交换内部注意力信息。然后, 相应分区函数的合并方法可以如方程 (7) 中所述进行表述。

$$\underbrace{\text{merge}([\mathbf{Y}_1 \mathbf{Y}_2]) = [\mathbf{Y}_1, \mathbf{Y}_2]}_{\text{partition}_{A \rightarrow B}^{MMCA}} \text{ or } \underbrace{\text{merge}([\mathbf{Y}_1 \mathbf{Y}_2]) = [\mathbf{Y}_2, \mathbf{Y}_1]}_{\text{partition}_{B \rightarrow A}^{MMCA}} \quad (7)$$

这种合并方法旨在通过重新组织输出标记对应于输入 $\mathbf{X}$ 中其原始位置的标记位置来保持标记位置的一致性。

跨模态交叉注意力 (IMCA)。为了对模型间的跨注意力进行建模, 我们的分区函数可以表述为如下形式,

$$\text{partition}^{IMCA} = \begin{cases} \mathbf{X}_1 = [\underbrace{\mathbf{x}_1^A, \dots, \mathbf{x}_{M/2}^A}_{\mathbf{x}_A^1}, \underbrace{\mathbf{x}_1^B, \dots, \mathbf{x}_{K/2}^B}_{\mathbf{x}_B^1}] \\ \mathbf{X}_2 = [\underbrace{\mathbf{x}_{M/2+1}^A, \dots, \mathbf{x}_M^A}_{\mathbf{x}_A^2}, \underbrace{\mathbf{x}_{K/2+1}^B, \dots, \mathbf{x}_K^B}_{\mathbf{x}_B^2}] \end{cases} \quad (8)$$

其中 partition^{IMCA} 是跨模态分区方法。我们的分区方法有四种不同的分区方式, 即 $(\mathbf{X}_1, \mathbf{X}_2) \in \{([\mathbf{X}_A^1, \mathbf{X}_B^1], [\mathbf{X}_A^2, \mathbf{X}_B^2]), ([\mathbf{X}_A^1, \mathbf{X}_B^2], [\mathbf{X}_A^2, \mathbf{X}_B^1]), ([\mathbf{X}_A^2, \mathbf{X}_B^1], [\mathbf{X}_A^1, \mathbf{X}_B^2]), ([\mathbf{X}_A^2, \mathbf{X}_B^2], [\mathbf{X}_A^1, \mathbf{X}_B^1])\}$ 。我们跨模态分区方法允许交叉注意力层有效捕捉模态之间的相关性。然后, 第一个分区 ($\mathbf{X}_1$) 中的跨模态注意力可以嵌入到第二个分区 ($\mathbf{X}_2$) 中。接着, 分区函数 partition^{IMCA} 的合并方法可以按照方程 (9) 进行表述。

$$\text{merge}(\mathbf{Y}_1, \mathbf{Y}_2) = [\mathbf{Y}_A^1, \mathbf{Y}_A^2, \mathbf{Y}_B^1, \mathbf{Y}_B^2] \quad (9)$$

其中 $\mathbf{Y}_A^1, \mathbf{Y}_A^2, \mathbf{Y}_B^1, \mathbf{Y}_B^2$ 是由 ICA 生成的 $\mathbf{X}_A^1, \mathbf{X}_A^2, \mathbf{X}_B^1, \mathbf{X}_B^2$ 的相应输出。

可学习的跨模态注意力 (LICA)。为了进一步提高 IMCA 的学习效果, 我们引入了一种新的可学习的跨模态注意力, 如下所示,

$$\mathbf{X}' = [\mathbf{x}'_1, \dots, \mathbf{x}'_N] = \mathbf{X} \mathbf{W}_{per} \quad \text{partition}^{LICA} = \begin{cases} \mathbf{X}_1 = [\mathbf{x}'_1, \dots, \mathbf{x}'_{N/2}] \\ \mathbf{X}_2 = [\mathbf{x}'_{N/2+1}, \dots, \mathbf{x}'_N] \end{cases} \quad (10)$$

其中 $\mathbf{W}_{per}$ 是可学习的排列矩阵。

为了保持矩阵 $\mathbf{W}_{per}$ 的排列性质, 我们采用将 LU 分解 [21] 作为 $\mathbf{W}_{per} = \mathbf{P} \mathbf{L} (\mathbf{U} + \text{diag}(\mathbf{s}))$, 其中 $\mathbf{P}$ 是固定的排列矩阵, $\mathbf{L}$ 和 $\mathbf{U}$ 分别是可学习的下三角和上三角矩阵, 它们在对角线上为 1 和 0, $\mathbf{s}$ 是可学习的向量。由于 $\mathbf{W}_{per}$ 是排列矩阵, 因此可以计算出逆排列矩阵 $\mathbf{W}_{per}^{-1}$, 并且可以通过向量 $\mathbf{s}$ 确定 $\frac{\partial \mathbf{X}'}{\partial \mathbf{X}}$ 的雅可比行列式, 即 $\log \det \left| \frac{\partial \mathbf{X}'}{\partial \mathbf{X}} \right| = \sum (\log |\mathbf{s}|)$ 。我们的方法可以使用提出的 LICA 方法有效地捕捉输入模态之间的底层交叉注意力。然后, 通过逆排列 $\mathbf{W}_{per}^{-1}$ 可以形成可学习分区函数的合并方法, 如下所示:

$$\text{merge}([\mathbf{Y}_1, \mathbf{Y}_2]) = [\mathbf{Y}_1, \mathbf{Y}_2] \mathbf{W}_{per}^{-1} \quad (11)$$

3.3 多模态潜在归一化流

典型的基于似然的模型有两个阶段。首先，感知压缩阶段侧重于去除高频细节，同时学习较少的语义信息。其次，语义压缩阶段将学习数据 [42] 中表示的语义和概念组成。因此，第二阶段起到更重要的作用，因为它是一个实质上的生成模型，学习多模态数据中表示的语义结构和跨模态关联。原始数据通常表示在高维空间中，例如高分辨率图像或长序列数据。然而，数据的语义信息可以在更低维空间中表示，因为输入空间具有冗余维度。因此，基于输入数据的固有维数，我们旨在找到一个在感知上等效但计算上高效的空间，以用于我们的基于多模态归一化流的方法。

知觉压缩。受前人成果 [42] 的启发，我们提出将数据投影到一个维度更低但在表示上含有更多有意义信息的特征空间中。令 $\mathcal{E}$ 为将输入 $\mathbf{X}$ 映射到潜在特征 $\mathbf{F}$ 的编码器，即 $\mathbf{F} = \mathcal{E}(\mathbf{X})$ 。然后，解码器 $\mathcal{D}$ 将特征映射回其原始数据空间，即 $\mathbf{X} = \mathcal{D}(\mathbf{F})$ 。编码器 $\mathcal{E}$ 和解码器 $\mathcal{D}$ 的设计可以多样化，例如 PCA，自动编码器 [41, 14]。然而，为了实现知觉压缩的最佳能力，我们采用自动编码器方法 [41, 14] 来开发编码器和解码器。该方法可以提供在一个在感知上等同于数据空间的新输入空间，同时保持低维空间。

潜在正则流模型。我们建议通过其潜在空间的多模态特征 $\mathbf{F}$ 来建模数据分布，而不是在其原始高维空间上建模多模态数据 $\mathbf{X}$ 。我们将这种方法命名为基于多模态注意力的潜在模型正则化流方法。通过我们的方法，基于流的模型不需要学习对高维数据进行感知压缩。相反，我们的正则流方法将专注于学习多模态数据的语义信息和关联性。因此，在使用高效计算成本的同时，我们的模型表现出更好的扩展特性。此外，通过可逆交叉注意力层设计的双射网络 G 提供了通过二阶相关学习进行更好的多模态建模。

基于注意力的归一化流网络。我们的双射网络 G （图 2）由 L 个块组成，其中每个块包含八个可逆的交叉注意力层和一个耦合层 [5]。前两个交叉注意力层采用 MMCA 分区。接下来的四个交叉注意力层执行不同的 IMCA 分区方法。接下来的两个层使用 LICA 交叉注意力层。然后，采用耦合层来增强双射块的内部表达能力。

学习 MANGO 以任务为导向。给定多模态输入 $\mathbf{X}$ 和特定任务的标签 $\mathbf{Y}$ ，MANGO 可以通过负对数似然和任务特定学习目标（如方程 (??) 中）联合优化。在此， $\mathbf{Z} = G(\mathcal{E}(\mathbf{X}))$ ， $\hat{\mathbf{Y}}$ 是对应任务的预测， $\mathcal{L}_{task}$ 是对应预测任务的损失， θ 是模型的参数。

4 实验结果

4.1 实现和基准测试

实现。我们的双射网络 G 由 $L = 12$ 交叉注意力块组成。对于感知压缩编码器 $\mathcal{E}$ ，我们采用 [14] 的视觉编码器用于 RGB 和深度图像。我们利用 [40] 中的文本编码器处理文本数据。为了公平比较，我们使用来自 [70, 65] 的语义分割、图像翻译和电影类型分类的任务头。我们的实验在 4 NVIDIA A100 GPU 上进行。我们的训练使用来自 [65] 的相同学习超参数，输入图像大小为 256×256 以确保公平比较。

语义分割。该任务使用 RGB 和深度图像的两个同质输入来预测分割图。我们在 NYUDv2 [36] 和 SUN RGB-D [45] 上进行了实验。NYUDv2 包含 795 张用于训练和 654 张用于测试的图像，而 SUN RGB-D 包含 5,285 个用于训练和 5,050 个用于测试的样本。

图像到图像翻译。按照标准协议 [65]，我们采用 Taskonomy [75] 进行多模态图像翻译任务。这个大规模的室内场景数据集提供了十多种多模态数据，例如 RGB、深度、法线、阴影、纹理、边缘等。我们使用 1,000 张高质量图像的子集进行训练，500 张用于验证。

MM-IMDB 电影类型分类。MM-IMDB 是一个用于电影类型分类的大规模多模态数据集。为了进行公平比较，我们采用了 [70] 的训练和测试划分。具体来说，我们实验中的数据包括 15,552 个训练数据和 2,608 个验证数据。在这个多模态学习任务中，我们使用来自图像和文本两种模态的输入。

Table 1: 与现有方法比较 NYUDv2 和 SUN RGB-D 上的 RGB-D 语义分割性能。我们的评价指标包括像素准确率 (Pixel Acc.) (%), 平均准确率 (mAcc.) (%), 平均交并比 (mIoU) (%)。

Method	Inputs	NYUDv2			SUN RGB-D		
		Pixel Acc.	mAcc.	mIoU	Pixel Acc.	mAcc.	mIoU
CNN-based models							
FCN-32s [30]	RGB	60.0	42.2	29.2	68.4	41.1	29.0
RefineNet [26]	RGB	74.4	59.6	47.6	81.1	57.7	47.0
FuseNet [13]	RGB+D	68.1	50.4	37.9	76.3	48.3	37.3
SSMA [62]	RGB+D	75.2	60.5	48.7	81.0	58.1	45.7
RDFNet [39]	RGB+D	76.0	62.8	50.1	81.5	60.1	47.7
AsymFusion [67]	RGB+D	77.0	64.0	51.2	-	-	-
CEN [66]	RGB+D	77.7	65.0	52.5	83.5	63.2	51.1
Transformer-based models							
DPLNet [6]	RGB+D	-	-	59.3	-	-	52.8
DFormer [71]	RGB+D	-	-	57.2	-	-	52.5
EMSA-Net [44]	RGB+D	-	-	59.0	-	-	50.9
W/O Fusion (Tiny) [65]	RGB	75.2	62.5	49.7	82.3	60.6	47.0
Feature Concat (Tiny) [65]	RGB+D	76.5	63.4	50.8	82.8	61.4	47.9
TokenFusion (Tiny) [65]	RGB+D	78.6	66.2	53.3	84.0	63.3	51.4
W/O fusion (Small) [65]	RGB	76.0	63.0	50.6	82.9	61.3	48.1
Feature Concat (Small) [65]	RGB+D	77.1	63.8	51.4	83.5	62.0	49.0
TokenFusion (Small) [65]	RGB+D	79.0	66.9	54.2	84.7	64.1	53.0
GeminiFusion (MIT-B5) [20]	RGB+D	80.3	70.4	57.7	83.8	65.3	53.3
MANGO	RGB+D	81.5	71.6	59.2	83.9	67.2	54.1

4.2 与现有最先进方法的比较

语义分割。表 1 展示了我们与之前的多模态方法在多模态语义分割上的对比结果。我们的结果表明，所提出的方法在 NYUDv2 和 SUN RGB-D 数据集上均达到了最先进的性能。我们的模型在所有评估指标和数据集上始终优于之前的方法。特别是，我们提出方法的 mIoU 结果在两个数据集上比 GeminiFusion 高 1.5 % 和 0.6 %。我们的结果表明，我们对多模态融合的显式建模相比于之前的融合方法显示出了明显的优势 [20, 65]。图 5 通过我们的归一化流可视化了我们的融合方法相对于之前的融合方法，即 TokenFusion [65] 的结果。

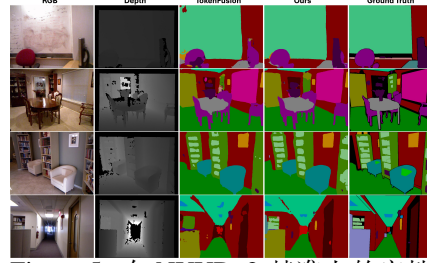


Figure 5: 在 NYUDv2 基准上的定性比较。

图像到图像的翻译。我们在表 2 所示的五种不同多模态图像到图像翻译学习设置中展示了我们的结果。在这五种不同的多模态学习设置中，我们的结果始终优于之前的方法。特别是，与之前的 GeminiFusion [20] 相比，我们的模型获得了更好的 FID 分数，即在 Shade+Texture → RGB 和 Depth+Normal → RGB 基准测试中，比 GeminiFusion 低 1.71 和 29.37。这些结果进一步验证了我们的方法在捕捉跨模态复杂相关性方面的出色能力。

Table 2: 在 Taskonomy 上的多模态图像翻译性能与先前的多模态方法的比较。我们使用 FID / KID ($\times 10^{-2}$) 对 RGB 目标进行评估, 并使用 MAE ($\times 10^{-1}$) / MSE ($\times 10^{-1}$) 对法线、阴影和深度目标进行评估。

Method	RGB (↓)	RGB (↓)	Normal (↓)	Shade (↓)	Depth (↓)
Ground Truth	78.82 / 2.83	90.08 / 4.58	1.34 / 2.85	1.26 / 7.02	0.51 / 0.78
Self-Attention [65]	73.87 / 2.86	86.79 / 3.91	1.26 / 2.76	1.19 / 7.36	0.39 / 0.70
Align [48]	82.30 / 4.20	105.00 / 4.91	1.52 / 3.25	1.41 / 7.21	0.45 / 0.90
C2F [65]	82.87 / 4.56	104.52 / 5.70	1.52 / 3.25	1.50 / 7.72	0.50 / 0.96

MM-IMDB 电影类型分类。表 3 展示了我们的方法在多模态分类基准测试上的结果。如表 3 所示，我们提出的方法在微平均和宏平均 F1 分数上均优于以前的方法，并达到了最先进的性能。特别是，我们的微平均和宏平均 F1 分数结果比之前的方法 [69] 分别高出 3.5 % 和 4.9 %。这些结果说明，我们的方法在同质和异质输入上表现更好。图 6 展示了我们在图像到图像翻译基准测试上的结果。

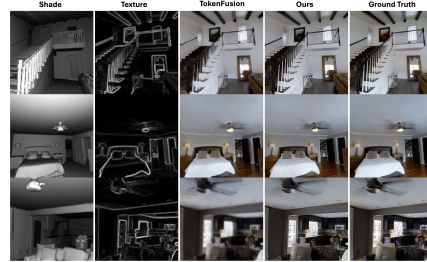


Figure 6: 图像到图像基准的定性比较。

4.3 消融研究

可逆交叉注意层的有效性。为了说明我们提出的可逆交叉注意层的影响，我们进行实验将我们提出的层与其他流模型进行比较，即 Affine Coupling Layer [5]、Glow [21]、Flow++ [15] 和 AttnFlow [48]。如表 4 所示，我们提出的可逆交叉注意层始终优于先前的耦合方法。特别是，我们的方法在 NYUDv2 和 SUN RGBD 基准测试中的 mIoU 结果分别达到了 59.2 % 和 54.1 %。这些结果清楚地说明了我们提出的方法在建模多模态数据中的相关性和复杂结构方面的优势。

不同分区方法的效果。表 5 展示了不同分区方法的实验结果。正如结果所示，使用模态到模态和模态间的交叉注意力，在 NYUDv2 和 SUN RGBD 基准上的 mIoU 结果分别达到了 58.0 % 和 53.7 %。此外，当采用可学习的模态间交叉注意力时，与不使用 LICA 相比，我们的 mIoU 结果进一步提高到了 59.2 % 和 54.1 %。实验结果证实了我们提出的方法在通过交叉注意力机制建模模态间关联方面的有效性。

潜在模型的有效性。这些实验研究了我们潜在模型方法的有效性。如表 6 所示，使用潜在模型后，我们的基于多模态归一化流的模型的表现持续在语义分割基准上得到改善。所提出的方法达到了最先进的结果，其中我们最佳模型的 mIoU 在 NYUDv2 和 SUN RGBD 基准

Table 3: 在 MM-IMDB 数据集上与先前多模态方法的电影类型分类性能比较。我们的指标包括微平均和宏平均 F1 分数。

Method	Modality	Micro F1 (%)	Macro F1 (%)
Image Network [70]	T	40.0	25.3
Text Network [70]	T	59.2	47.2
Late Fusion [24]	I+T	59.6	51.0
LRTF [29]	I+T	59.2	49.3
ML-Matrix [19]	I+T	58.5	48.4
DynMM [70]	I+T	60.4	51.6
COCA [74]	I+T	67.7	62.6
MFMM [3]	I+T	67.5	61.6
BLIP [22]	I+T	67.4	62.8
ReFNet [43]	I+T	68.0	58.7
BridgeTow [69]	I+T	68.2	63.3
MANGO	I+T	71.7	68.2

上的结果分别达到了 59.2 % 和 54.1 %。这些结果突出了使用感知压缩编码器产生较低但更有效的表示空间的优点。

Table 4: 可逆层的有效性。

Layer	NYUDv2			SUN RGBD		
	Pixel Acc.	mAcc.	mIoU	Pixel Acc.	mAcc.	mIoU
Coupling Layer [5]	76.0	63.4	50.8	79.8	59.9	48.5
Glow [21]	77.0	66.4	53.0	80.3	61.9	49.1
Flow++ [15]	77.5	68.1	54.2	81.5	62.0	50.5
AttFlow [48]	79.5	69.9	56.5	82.5	65.1	52.2
MANGO	81.5	71.6	59.2	83.9	67.2	54.1

交叉注意力块数量的有效性。表 6 展示了我们的方法在使用不同数量 (L) 的交叉注意力块时的结果。正如我们的结果所示, 使用更深网络的多模态分割模型表现出更好的性能。具体来说, 使用 $L = 12$ 个可逆交叉注意力块, NYUDv2 和 SUN RGBD 基准上的 mIoU 性能分别达到了 59.2 % 和 54.1 %。虽然较少的块可能降低计算成本, 但更深的模型可以更好地利用多模态数据中特征的相关性。

5 结论

我们的论文通过基于注意力的正规化流模型引入了一种新的显式建模方法, 用于多模态融合学习。我们提出的 ICA 层结合三种不同的交叉注意力机制, 能够有效捕捉多模态数据中的复杂结构和潜在关联。我们还引入了一种新的潜在方法来规范化流, 以增强我们在多模态数据上的可扩展性。我们在三个标准基准测试上的深入实验, 即语义分割、图像到图像转换和电影类型分类, 展示了我们方法的有效性。我们的研究证明了在选定的超参数和基准条件下, 可逆交叉注意力层在多模态学习中的有效性。然而, 在目标平衡和可扩展性方面仍存在限制。详细的限制在我们的附录中进行了讨论。

Table 5: 分区方法的有效性。

	MMCA	IMCA	LICA	NYUDv2			SUN RGBD		
				Pixel Acc.	mAcc.	mIoU	Pixel Acc.	mAcc.	mIoU
✓				79.3	68.8	56.4	82.4	64.6	51.3
✓	✓			80.2	70.8	58.0	83.3	66.2	53.7
✓		✓	✓	81.5	71.6	59.2	83.9	67.2	54.1

Table 6: 潜在模型的有效性

# Blocks	Latent Model	NYUDv2			SUN RGBD		
		Pixel Acc.	mAcc.	mIoU	Pixel Acc.	mAcc.	mIoU
6	✗	75.9	63.5	51.0	79.4	59.1	47.3
	✓	77.5	65.8	52.3	79.6	60.5	48.1
8	✗	78.0	65.5	53.1	80.8	60.8	49.4
	✓	78.1	65.3	54.1	84.4	60.0	51.4
12	✗	80.7	70.4	58.0	83.4	65.8	53.5
	✓	81.5	71.6	59.2	83.9	67.2	54.1

References

- [1] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
- [2] J. Arevalo, T. Solorio, M. Montes-y Gómez, and F. A. González. Gated multimodal units for information fusion. *arXiv preprint arXiv:1702.01992*, 2017.
- [3] L. Braz, V. Teixeira, H. Pedrini, and Z. Dias. Image-text integration using a multimodal fusion network module for movie genre classification. In *11th International Conference of Pattern Recognition Systems (ICPRS 2021)*, volume 2021, pages 200–205, 2021.
- [4] L. Dinh, D. Krueger, and Y. Bengio. Nice: Non-linear independent components estimation. *arXiv preprint arXiv:1410.8516*, 2014.
- [5] L. Dinh, J. Sohl-Dickstein, and S. Bengio. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*, 2016.
- [6] S. Dong, Y. Feng, Q. Yang, Y. Huang, D. Liu, and H. Fan. Efficient multimodal semantic segmentation via dual-prompt learning. *arXiv preprint arXiv:2312.00360*, 2023.
- [7] A. Ephrat, I. Mosseri, O. Lang, T. Dekel, K. Wilson, A. Hassidim, W. T. Freeman, and M. Rubinstein. Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. *arXiv preprint arXiv:1804.03619*, 2018.
- [8] G. Evangelopoulos, A. Zlatintsi, A. Potamianos, P. Maragos, K. Rapantzikos, G. Skoumas, and Y. Avrithis. Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention. *IEEE Transactions on Multimedia*, 15(7):1553–1568, 2013.
- [9] V. Garcia Satorras, E. Hoogeboom, F. Fuchs, I. Posner, and M. Welling. E (n) equivariant normalizing flows. *Advances in Neural Information Processing Systems*, 34:4181–4192, 2021.
- [10] M. I. E. Ghebriout, H. Bouzidi, S. Niar, and H. Ouarnoughi. Harmonic-nas: Hardware-aware multimodal neural architecture search on resource-constrained devices. In *Asian Conference on Machine Learning*, pages 374–389. PMLR, 2024.

- [11] M. Glodek, S. Tschechne, G. Layher, M. Schels, T. Brosch, S. Scherer, M. Kächele, M. Schmidt, H. Neumann, G. Palm, et al. Multiple classifier systems for the classification of audio-visual emotional states. In *Affective Computing and Intelligent Interaction: Fourth International Conference, ACHI 2011, Memphis, TN, USA, October 9–12, 2011, Proceedings, Part II*, pages 359–368. Springer, 2011.
- [12] X. Gong, S. Mohan, N. Dhingra, J.-C. Bazin, Y. Li, Z. Wang, and R. Ranjan. Mmg-ego4d: Multimodal generalization in egocentric action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6481–6491, 2023.
- [13] C. Hazirbas, L. Ma, C. Domokos, and D. Cremers. Fusetnet: incorporating depth into semantic segmentation via fusion-based cnn architecture. In *Asian Conference on Computer Vision*, November 2016.
- [14] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.
- [15] J. Ho, X. Chen, A. Srinivas, Y. Duan, and P. Abbeel. Flow++: Improving flow-based generative models with variational dequantization and architecture design. In *International conference on machine learning*, pages 2722–2730. PMLR, 2019.
- [16] C. Hu, B. Fu, P. Yu, L. Zhang, X. Shi, and Y. Chen. An explicit multi-modal fusion method for sign language translation. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3860–3864. IEEE, 2024.
- [17] C.-W. Huang, D. Krueger, A. Lacoste, and A. Courville. Neural autoregressive flows. In *International conference on machine learning*, pages 2078–2087. PMLR, 2018.
- [18] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 172–189, 2018.
- [19] S. M. Jayakumar, W. M. Czarnecki, J. Menick, J. Schwarz, J. Rae, S. Osindero, Y. W. Teh, T. Harley, and R. Pascanu. Multiplicative interactions and where to find them. In *International Conference on Learning Representations*, 2020.
- [20] D. Jia, J. Guo, K. Han, H. Wu, C. Zhang, C. Xu, and X. Chen. Geminifusion: Efficient pixel-wise multimodal fusion for vision transformer. *arXiv preprint arXiv:2406.01210*, 2024.
- [21] D. P. Kingma and P. Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. *Advances in neural information processing systems*, 31, 2018.
- [22] J. Li, D. Li, C. Xiong, and S. Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022.
- [23] M. Liang, J. Hu, C. Bao, H. Feng, F. Deng, and T. L. Lam. Explicit attention-enhanced fusion for rgb-thermal perception tasks. *IEEE Robotics and Automation Letters*, 8(7):4060–4067, 2023.
- [24] P. P. Liang, Y. Lyu, X. Fan, Z. Wu, Y. Cheng, J. Wu, L. Y. Chen, P. Wu, M. A. Lee, Y. Zhu, et al. Multibench: Multiscale benchmarks for multimodal representation learning. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021.
- [25] X. Liang, Q. Guo, Y. Qian, W. Ding, and Q. Zhang. Evolutionary deep fusion method and its application in chemical structure recognition. *IEEE Transactions on Evolutionary Computation*, 25(5):883–893, 2021.
- [26] G. Lin, A. Milan, C. Shen, and I. Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.
- [27] H. Liu, C. Li, Y. Li, and Y. J. Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024.
- [28] H. Liu, C. Li, Q. Wu, and Y. J. Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.

- [29] Z. Liu, Y. Shen, V. B. Lakshminarasimhan, P. P. Liang, A. Zadeh, and L.-P. Morency. Efficient low-rank multimodal fusion with modality-specific factors. In *Annual Meeting of the Association for Computational Linguistics*, 2018.
- [30] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.
- [31] Y. Lu and B. Huang. Structured output learning with conditional generative flows. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 5005–5012, 2020.
- [32] A. Lugmayr, M. Danelljan, L. Van Gool, and R. Timofte. SrfLOW: Learning the super-resolution space with normalizing flow. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pages 715–732. Springer, 2020.
- [33] E. Morvant, A. Habrard, and S. Ayache. Majority vote of diverse classifiers for late fusion. In *Structural, Syntactic, and Statistical Pattern Recognition: Joint IAPR International Workshop, S+ SSPPR 2014, Joensuu, Finland, August 20-22, 2014. Proceedings*, pages 153–162. Springer, 2014.
- [34] S. Nagar, M. Dufraisie, and G. Varma. Cinc flow: Characterizable invertible 3x3 convolution. *arXiv preprint arXiv:2107.01358*, 2021.
- [35] A. Nagrani, S. Yang, A. Arnab, A. Jansen, C. Schmid, and C. Sun. Attention bottlenecks for multimodal fusion. *Advances in neural information processing systems*, 34:14200–14213, 2021.
- [36] P. K. Nathan Silberman, Derek Hoiem and R. Fergus. Indoor segmentation and support inference from rgb-d images. In *ECCV*, 2012.
- [37] H.-Q. Nguyen, T.-D. Truong, X. B. Nguyen, A. Dowling, X. Li, and K. Luu. Insect-foundation: A foundation model and large-scale 1m dataset for visual insect understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21945–21955, 2024.
- [38] X.-B. Nguyen, C. N. Duong, X. Li, S. Gauch, H.-S. Seo, and K. Luu. Micron-bert: Bert-based facial micro-expression recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1482–1492, 2023.
- [39] S.-J. Park, K.-S. Hong, and S. Lee. Rdfnet: Rgb-d multi-level residual feature fusion for indoor semantic segmentation. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.
- [40] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.
- [41] A. Razavi, A. Van den Oord, and O. Vinyals. Generating diverse high-fidelity images with vq-vae-2. *Advances in neural information processing systems*, 32, 2019.
- [42] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.
- [43] S. Sankaran, D. Yang, and S.-N. Lim. Refining multimodal representations using a modality-centric self-supervised module, 2022.
- [44] D. Seichter, S. Fishedick, M. Köhler, and H.-M. Gross. Efficient multi-task rgb-d scene analysis for indoor environments. In *IEEE International Joint Conference on Neural Networks (IJCNN)*, pages 1–10, 2022.
- [45] S. Song, S. P. Lichtenberg, and J. Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 567–576, 2015.
- [46] S. Song, J. Liu, Y. Li, and Z. Guo. Modality compensation network: Cross-modal adaptation for action recognition. *IEEE Transactions on Image Processing*, 29:3957–3969, 2020.
- [47] M. Sorkhei, G. E. Henter, and H. Kjellström. Full-glow: Fully conditional glow for more realistic image generation. In *DAGM German Conference on Pattern Recognition*, pages 697–711. Springer, 2021.

- [48] R. S. Sukthanker, Z. Huang, S. Kumar, R. Timofte, and L. Van Gool. Generative flows with invertible attentions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11234–11243, 2022.
- [49] H. Sun, R. Mehta, H. H. Zhou, Z. Huang, S. C. Johnson, V. Prabhakaran, and V. Singh. Dual-glow: Conditional flow-based generative model for modality transfer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10611–10620, 2019.
- [50] Z.-X. Tan, A. Goel, T.-S. Nguyen, and D. C. Ong. A multimodal lstm for predicting listener empathic responses over time. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pages 1–4. IEEE, 2019.
- [51] C. Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024.
- [52] G. Team, R. Anil, S. Borgeaud, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, K. Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.
- [53] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.
- [54] T.-D. Truong, Q.-H. Bui, C. N. Duong, H.-S. Seo, S. L. Phung, X. Li, and K. Luu. Direcformer: A directed attention in transformer approach to robust action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20030–20040, 2022.
- [55] T.-D. Truong, C. N. Duong, T. De Vu, H. A. Pham, B. Raj, N. Le, and K. Luu. The right to talk: An audio-visual transformer approach. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1105–1114, October 2021.
- [56] T.-D. Truong, N. Le, B. Raj, J. Cothren, and K. Luu. Freedom: Fairness domain adaptation approach to semantic scene understanding. In *IEEE/CVF Computer Vision and Pattern Recognition (CVPR)*, 2023.
- [57] T.-D. Truong and K. Luu. Cross-view action recognition understanding from exocentric to egocentric perspective. *Neurocomputing*, 614:128731, 2025.
- [58] T.-D. Truong, H.-Q. Nguyen, X.-B. Nguyen, A. Dowling, X. Li, and K. Luu. Insect-foundation: A foundation model and large multimodal dataset for vision-language insect understanding. *International Journal of Computer Vision*, pages 1–26, 2025.
- [59] T.-D. Truong, H.-Q. Nguyen, B. Raj, and K. Luu. Fairness continual learning approach to semantic scene understanding in open-world environments. *Advances in Neural Information Processing Systems*, 36, 2024.
- [60] T.-D. Truong, U. Prabhu, B. Raj, J. Cothren, and K. Luu. Falcon: Fairness learning via contrastive attention approach to continual semantic scene understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 15065–15075, 2025.
- [61] T.-D. Truong, U. Prabhu, D. Wang, B. Raj, S. Gauch, J. Subbiah, and K. Luu. Eagle: Efficient adaptive geometry-based learning in cross-view understanding. *Advances in Neural Information Processing Systems*, 37:137309–137333, 2024.
- [62] A. Valada, R. Mohan, and W. Burgard. Self-supervised model adaptation for multimodal semantic segmentation. *International Journal of Computer Vision (IJCV)*, jul 2019. Special Issue: Deep Learning for Robotic Vision.
- [63] A. Valada, R. Mohan, and W. Burgard. Self-supervised model adaptation for multimodal semantic segmentation. *International Journal of Computer Vision*, 128(5):1239–1285, 2020.
- [64] A. Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.
- [65] Y. Wang, X. Chen, L. Cao, W. Huang, F. Sun, and Y. Wang. Multimodal token fusion for vision transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12186–12195, 2022.
- [66] Y. Wang, W. Huang, F. Sun, T. Xu, Y. Rong, and J. Huang. Deep multimodal fusion by channel exchanging. *Advances in neural information processing systems*, 33:4835–4845, 2020.

- [67] Y. Wang, F. Sun, M. Lu, and A. Yao. Learning deep multimodal feature representation with asymmetric multi-layer fusion. In *ACM International Conference on Multimedia (ACM MM)*, 2020.
- [68] X. Wei, T. Zhang, Y. Li, Y. Zhang, and F. Wu. Multi-modality cross attention network for image and sentence matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10941–10950, 2020.
- [69] X. Xu, C. Wu, S. Rosenman, V. Lal, W. Che, and N. Duan. Bridgetower: Building bridges between encoders in vision-language representation learning. *arXiv preprint arXiv:2206.08657*, 2022.
- [70] Z. Xue and R. Marculescu. Dynamic multimodal fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2575–2584, 2023.
- [71] B. Yin, X. Zhang, Z. Li, L. Liu, M.-M. Cheng, and Q. Hou. Dformer: Rethinking rgb-d representation learning for semantic segmentation. *arXiv preprint arXiv:2309.09668*, 2023.
- [72] J. Yin, J. Shen, R. Chen, W. Li, R. Yang, P. Frossard, and W. Wang. Is-fusion: Instance-scene collaborative fusion for multimodal 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14905–14915, 2024.
- [73] Y. Yin, S. Huang, and X. Zhang. Bm-nas: Bilevel multimodal neural architecture search. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 8901–8909, 2022.
- [74] J. Yu, Z. Wang, V. Vasudevan, L. Yeung, M. Seyedhosseini, and Y. Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022.
- [75] A. R. Zamir, A. Sax, W. Shen, L. J. Guibas, J. Malik, and S. Savarese. Taskonomy: Disentangling task transfer learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3712–3722, 2018.
- [76] H. Zhao, C. Gan, A. Rouditchenko, C. Vondrick, J. McDermott, and A. Torralba. The sound of pixels. In *Proceedings of the European conference on computer vision (ECCV)*, pages 570–586, 2018.

A 额外消融研究

交叉注意块数量的有效性。我们进行了一项使用 16 个交叉注意块的消融研究。如表 7 所示，尽管使用更多的交叉注意块会增加计算量，但它有助于提升模型性能。

Table 7: 交叉注意力模块数量的有效性。

# Blocks	NYUv2			SUN RGBD		
	Acc.	mAcc.	mIoU	Acc.	mAcc.	mIoU
6	77.5	65.8	52.3	79.6	60.5	48.1
8	78.1	65.3	54.1	84.4	60.0	51.4
12	81.5	71.6	59.2	83.9	67.2	54.1
16	83.1	75.1	61.7	85.4	68.7	55.6

计算成本。如表 8 所示，我们方法的参数、GFLOPs 和推理时间与之前的方法有竞争力。同时，我们在两个分割基准上取得了最先进的性能。

Table 8: 计算成本的比较。

Method	NYUdv2 mIoU	SUN RGB-D mIoU	PARAMS	GFLOPS	Inference Time
TokenFusion [65]	54.2	53.0	45.9M	108	126 ms
GeminiFusion [20]	57.7	53.3	75.8M	174	153 ms
MANGO	59.2	54.1	72.9M	152	144 ms

注意力可视化。如图 7 所示，我们的可逆交叉注意力层能够捕捉深度图像（红色框）区域与 RGB 图像之间的注意力交互。这个结果说明了我们提出的注意力层在捕捉跨模态关联上的有效性。

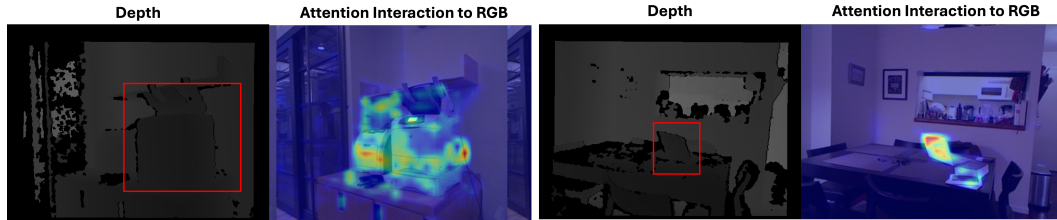


Figure 7: ICA 层的注意力可视化。

B 局限性讨论

我们的实验选择了一组学习超参数和基准来支持我们的假设。然而，我们的工作可能存在一些局限性。我们的研究考察了我们提出的可逆交叉注意层在多模态学习中的有效性。因此，尚未充分探索学习目标之间平衡权重的研究，我们将这一实验留作未来的工作。由于计算限制，我们的实验仅限于基准的标准规模。然而，我们假设根据论文中提出的基本理论，所提出的方法能够推广到更大规模的数据和基准设置。