基于注意力的自闭症谱系障碍筛查与特权模式**摘要**

本文提出了一种自闭症谱系障碍（ASD）自动定量筛查的新框架。它旨在解决当前临床环境中的两个问题：1）ASD患病率缺乏临床资源（美国为1.7%），以及2）ASD筛查的主观性。这项工作具有三个独特的特征：首先，它提出了一种具有特权模式框架的ASD筛查，该框架在训练期间集成了来自两种行为模式的信息，并在测试时提高了每种单一模式的性能。拟议的框架不要求各模式之间在主题上有重叠。其次，它开发了第一个计算模型，使用拍照任务对ASD患者进行分类，其中受试者在更生态的环境中自由探索他们的环境。拍照揭示了受试者的注意力偏好，将ASD患者与健康人区分开来，并且也很容易在现实世界的临床环境中实施，而无需先进的诊断仪器。第三，这项研究首次利用了观看图像时眼球运动的时间信息，编码了ASD患者和健康对照者之间更详细的行为差异。实验表明，我们的ASD筛查模型可以实现卓越的性能，比以前最先进的方法有相当大的优势。此外，我们使用不同模式的框架展示了拍照和图像查看任务的性能改进，提供了一种通用范例，该范例采用多种行为数据源进行更准确的ASD筛查。该框架也适用于各种场景，其中难以跨不同模式获得一对一的成对关系。

**1.介绍**

自闭症谱系障碍（ASD）是一种可遗传的终身神经发育障碍（NDD），具有复杂的病因和原因。它在全球普遍存在，在美国，每59名儿童中就有1例受到影响[2]。虽然目前被认为是ASD治疗最有效的临床途径[3]，但早期诊断和干预依赖于一支医疗专业知识团队，这些诊断仪器既耗时又临床要求高。由于ASD的患病率和有限的临床资源，它们没有广泛应用。此外，人类的评估是主观的，往往不一致，也是间歇性的。因此，协助ASD筛查的自动和客观工具具有重要的临床和社会需求。

视觉注意网络在大脑中无处不在，许多NDD与对视觉刺激的非典型注意有关。例如，长期以来，ASD患者对面部或其他社会刺激有非典型的关注[6，7，27，28，29]。最近关于自然场景刺激的研究表明，ASD患者与健康人之间的差异更复杂或更细[36]。本文开发了新的计算机视觉技术，以解决ASD筛查中的现有挑战。它提出了一种新方法，可以记录和建模具有更高的生态有效性和实际可行性的注意力偏好。此外，由于问题的复杂性，例如ASD内部或多个NDD之间的相当大的异质性[24]），以及临床数据的稀缺性，它强调了重要性，并提出了利用多种行为模式以及时间信息来编码准确ASD筛查所需的更详细和全面信息的方法。

具体而言，我们建议将两种与人类视觉注意力相关的不同形式，即从拍照任务和图像查看任务中记录的注意力偏好，用于ASD筛查。在拍照任务中，受试者在环境中自由移动，并通过拍照来识别他们喜欢的感兴趣区域，而在图像查看任务中，受试者通过眼动追踪设备记录的眼球运动来查看不同的图像。我们没有在每个模式上独立筛选ASD，而是提出了一种具有特权模式框架的新型ASD筛查，该框架在培训期间整合了各种模式，并使每种模式在测试中受益。我们的框架由三个主要组成部分组成：为了利用视觉注意网络的普遍性和深度神经网络（DNN）的学习潜力，我们开发了两个DNN模型，每个模型编码受试者拍摄的照片的特征（拍照）或眼球运动的时间信息（图像查看），以根据注意力偏好对ASD进行分类。为了利用来自两种模态的丰富和互补的信息，我们提出了一种多模态蒸馏方法，该方法学习多种模态的共享嵌入空间（即始终可用的主模态和仅在训练期间适用的特权模态），并将多模态知识从共享空间提炼到每个模态。与现有方法相比，我们的框架具有以下优势：1）与以前的方法[17，22]不同，以前的方法较少关注时间信息并独立训练特征编码器和ASD分类器，我们的图像查看模型以端到端的方式开发，并获取眼球运动的时间信息;2）与多模态方法[21，34]不同，多模态方法要求在测试期间所有模态的可用性，这在临床场景中很难实现，我们的框架可以部署在每个独立的模态上;3）所提出的多模态蒸馏方法不是依赖于模式之间的一对一成对对关系，类似于使用特权模态方法学习的其他模式[10，16，20，23]，而是可以在不同模态之间传递丰富的信息，而不会在受试者中重叠。

综上所述，本文有三大贡献：

我们超越了一种模式，并提出了一种利用多种行为数据来源的特权模式框架的ASD筛查。在我们的背景下，模式之间没有主题重叠。

我们开发了第一个基于拍照任务筛选ASD的计算模型。尽管任务具有挑战性，但我们的模型优于人类专家，并实现了合理的性能。

通过结合眼球运动的时间信息，我们的图像查看任务模型能够实现新的最先进的性能。

**2.相关工作**

**自动 ASD 筛查。**有几种计算模型可以自动识别ASD患者。Anzulewicz等人[1]使用智能平板电脑设备记录儿童的运动模式，并提出三种基于决策树的模型，用于基于这些模式识别ASD。受到ASD患者难以识别面部和解释面部情绪的研究结果的启发[29]，Liu等人[22]评估了儿童的面部扫描模式并检测ASD患者。为了在图像查看过程中捕获ASD和对照组之间注视模式的差异，Wang等人[36]建议训练具有预定义特征的支持向量机（SVM）模型，以根据ASD的个体的注视模式对其进行分类。Jiang和Zhao[17]后来通过引入一种新的深度神经网络方法扩展了[36]的思想，该方法突出了凝视模式的差异，从而为准确的ASD筛查带来了更多的判别特征。人们还探索了用于分类ASD的不同类型的神经影像学技术[21，34]。虽然这些方法取得了合理的结果，但它们要么只考虑单一的模式，而较少关注时间信息的利用[1，17，22，36]，要么依赖于在临床场景中难以部署的资源要求苛刻的仪器获得的多模态数据[21，34]。

**学习特权信息。**学习特权信息（LUPI）是[31，32]中提出的一种范例，它指定了某些特权信息在训练期间可用但在测试中不适用的场景。在本文中，我们专注于特权信息对应于与所有可用方式不同的模式的情况，即使用特权模式学习。Hoffman等人[16]提出了一种多流幻觉架构，该架构学习不同模态之间的映射，并在使用单模态进行测试时模拟多模态场景。在[20]中，Lambert等人利用特权模态的特征来学习辍学单元的超参数。Garcia等人[10]建议将模态幻觉[16]与知识蒸馏[14]结合起来，以特权模态进行行动识别。在[23]中，Luo等人通过在训练期间在各种模态之间构建图结构，将多模态知识提炼成单模态网络。这些方法依赖于不同模式之间的一对一成对关系，这在临床环境中通常很难实现患者数据，例如，在我们的ASD筛查实验中，两种模式的数据是从两组受试者中单独收集的，没有重叠。

在这项工作中，我们提出了一种具有特权模式框架的ASD筛查，该框架结合了用于ASD筛查的多种行为模式。我们的框架不依赖于不同模式的受试者重叠或先进的诊断仪器来收集数据，因此适用于常规临床环境中的ASD筛查。通过结合眼球运动的时间信息和来自两种模式的丰富知识，所提出的模型能够实现卓越的性能，并优于以前最先进的方法。

**3.有特权模态的ASD筛查**

神经发育障碍（如ASD）通常以多种症状为特征，其中单一模式可能无法提供足够的信息用于诊断目的。在这项工作中，我们建议使用更多的行为方式来筛选ASD，这些行为方式将提供补充和丰富的信息。多模态数据对于病症的异质性和临床人群数据的稀缺性尤为重要。

具体而言，我们提出了一种具有特权模式框架的ASD筛查，该框架利用两种不同的行为模式，即拍照和图像观看。它在训练期间结合了来自两种模式的信息，并且只需要一种模式进行测试，即我们将一种模式视为有利于学习另一种主要模式的特权模式。与现有的多模态[21，34]或使用特权模态学习[10，16，20，23]方法不同，我们的框架不依赖于部署期间所有模态的可用性或跨模态的一对一成对关系（例如主题重叠），使其在现实世界的临床场景中更实用。

在本节中，我们阐述了所提出框架的三个主要组成部分，包括分别用于拍照和图像查看任务的ASD筛查的两个DNN模型，以及一种多模态蒸馏方法，该方法将多模态知识从共享空间提炼到每个独立模态。

**3.1照片拍摄中的ASD筛查**

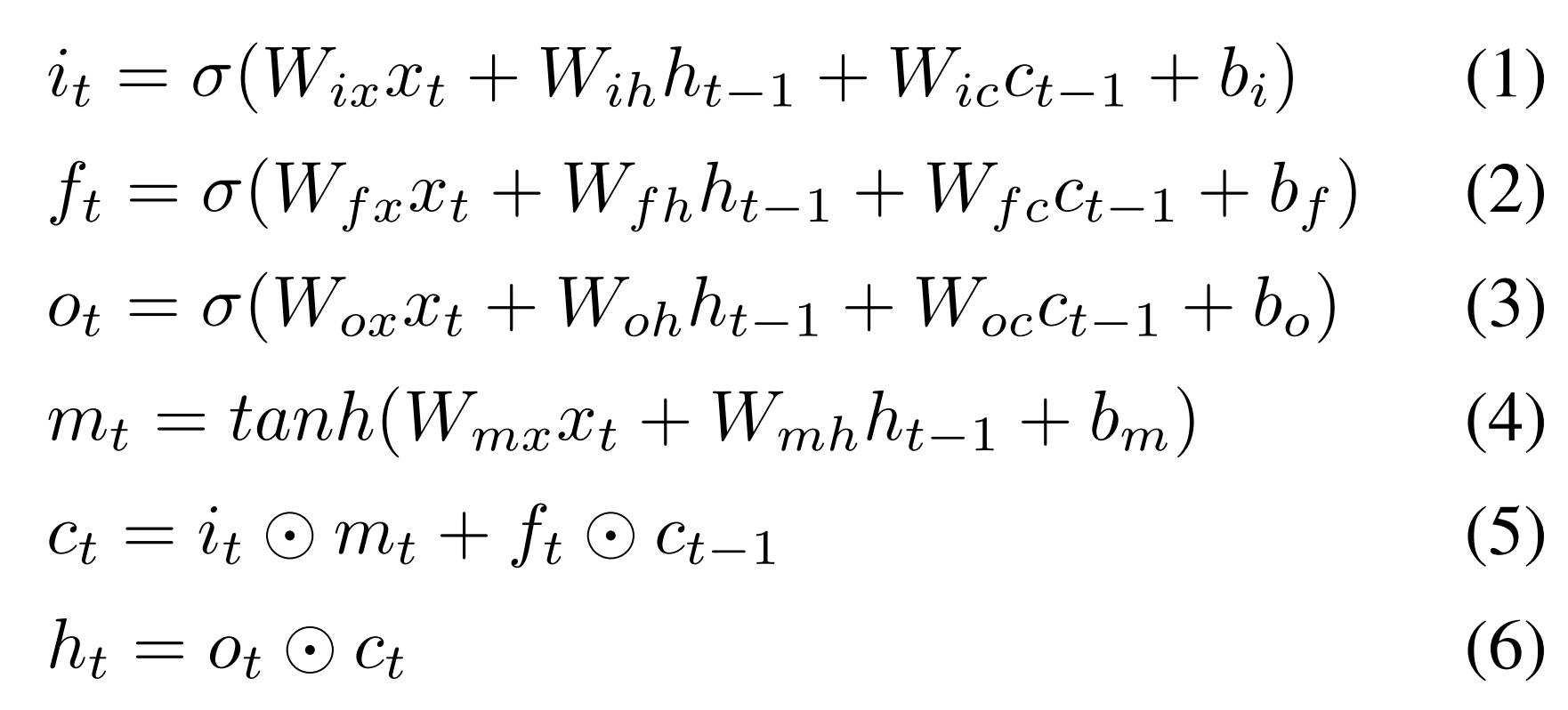
与以前遵循被动图像查看程序的基于视觉注意的ASD筛查方法[17，22，36]不同，我们的拍照任务允许受试者自由地与各种场景互动，并在第一人称设置中识别感兴趣的区域或物体，从而为揭示一个人的注意偏好提供了一个更加生态的范式。此外，人们拍摄的照片提供了额外的信息，显示了他们在社交互动中的行为。例如，由于社交互动减少，患有ASD的人可能不会要求人们调整姿势，从而导致人们不摆姿势或看着相机的低质量照片。受ASD患者拍摄的照片往往与健康人拍摄的照片具有不同的特征的发现[35]的启发，例如注意力偏好和照片质量的差异，在本文中，我们的目标是通过用受试者拍摄的照片池来描述这些差异来筛选ASD。

为了实现上述目标，我们建议利用CNN来学习有意义的特征，并利用递归神经网络（RNN）来捕获一系列照片的特征。如图1（上图）所示，所提出的模型由两个主要组件组成：1）一个编码器模块，该模块首先使用最先进的ResNet-50 [13]将原始图像数据投影到高级视觉特征，然后使用长短期记忆（LSTM）[15]网络按顺序遍历照片序列中不同图像的特征，以及2）一个分类器模块，该模块接收LSTM的最终隐藏状态并进行预测（即ASD）或控制）。给定特定受试者拍摄的照片序列，我们首先通过ResNet-50计算序列中不同图像的视觉特征，然后应用全局平均池（GAP）将空间特征转换为描述相应图像抽象信息的向量。然后将这些矢量按顺序转发到LSTM，通过反复更新隐藏状态来捕获照片的特征。在获得编码整个序列信息的最终隐藏状态后，我们直接将其馈送到我们的分类器（即单个全连接层）以识别ASD患者。

**3.2图像查看中的ASD筛查**

使用自然主义场景的ASD研究的最新进展已经为ASD患者如何与健康受试者看起来不同产生了一些新的见解。例如，对于包含丰富社会和语义内容的复杂刺激，Wang等人[36]观察到，受试者群体之间对语义对象的注视顺序和延迟显着不同，这表明时间信息在筛选任务中的作用。图 2 突出显示了使用时间信息的重要性，该信息揭示了 ASD 和对照组之间凝视模式的显著差异，即使聚合的注视点图相似。此外，由于临床数据的稀缺性阻止了过于复杂的模型设计，这些方法通常分别训练特征编码器和ASD分类器，而没有明确地将学习的视觉特征与ASD筛查相关联，这使得它们难以实现令人满意的性能。在本节中，我们将介绍一种 DNN 模型来解决这些问题，以便通过图像查看实现更准确的 ASD 筛查。我们的模型以端到端的方式进行了优化，它会自动将视觉特征与ASD筛查联系起来，并接收眼球运动的时间信息，以破译在图像观看过程中记录的视觉注意力的更多辨别性特征。

如图1（底部）所示，所提出的图像查看模型与我们的拍照模式模型具有相似的设计。然而，与目标基于一系列照片对ASD进行分类的照片拍摄不同，对于图像查看，我们的目标是根据每个特定图像上捕获的注意力模式（即眼球运动）来区分ASD患者。因此，给定特定受试者的图像及其相应的视觉扫描路径，而不是将CNN（卷积神经网络）与GAP一起用于特征提取，我们首先从CNN获得有用的视觉特征，然后在每个眼睛固定物附近提取特征（即最接近每个固定点的2048维特征向量）。然后将提取的特征按顺序馈送到我们的LSTM（请注意，对于图像查看，我们使用类似于[11]的LSTM变体以获得更好的性能，表示为LSTM∗），基于扫描路径内的注视顺序，捕获眼球运动的时间信息。在每次眼睛固定时，该过程可以表示如下：



其中xt是第一只眼睛固定附近提取的视觉特征，W和b是LSTM中的可训练参数，σ sigmoid函数，ht−1和ct−1表示包含先前眼球运动的时间信息的隐藏状态和记忆向量。i、f 和 o 是输入门，忘记 LSTM 的门和输出门，m 进一步编码基于 xt 和 ht−1 的特征。在视觉扫描路径结束时计算的隐藏状态h被输入到分类器中，用于预测ASD患者。

**3.3. 通过共享空间进行多模态蒸馏**

通过上述每种模式的ASD筛查模型，这里的关键是有效地整合来自两种不同模式的信息，以进一步提高ASD筛查对每种模式的性能。为此，我们提出了一种多模态蒸馏方法，使模型能够通过共享空间从不同类型的行为数据中学习。我们的方法受到交叉模式检索和匹配方法[5，12，25]的启发，但是，它在目标和方法上都与它们显着不同：1）与[5，12，25]的目标是用源模态的数据检索目标模态中的样本不同，我们的目标是创建一个共享空间，以便在不同模态之间转移知识，以提高每种模态的性能。2）在跨模式匹配中，例如[5]，不同模式的模块在多任务学习框架[4]（不同模式的联合培训）下进行优化，其中很难在两种模式下实现令人满意的表现。相反，我们提出了一种新颖的方法，通过共同训练共享空间来提炼多模态知识，但在学习共享空间后，通过分离不同模态的模型来克服上述困难，因此每个模型都可以专注于自己的模态以最好地优化它。图3所示方法的实施过程。

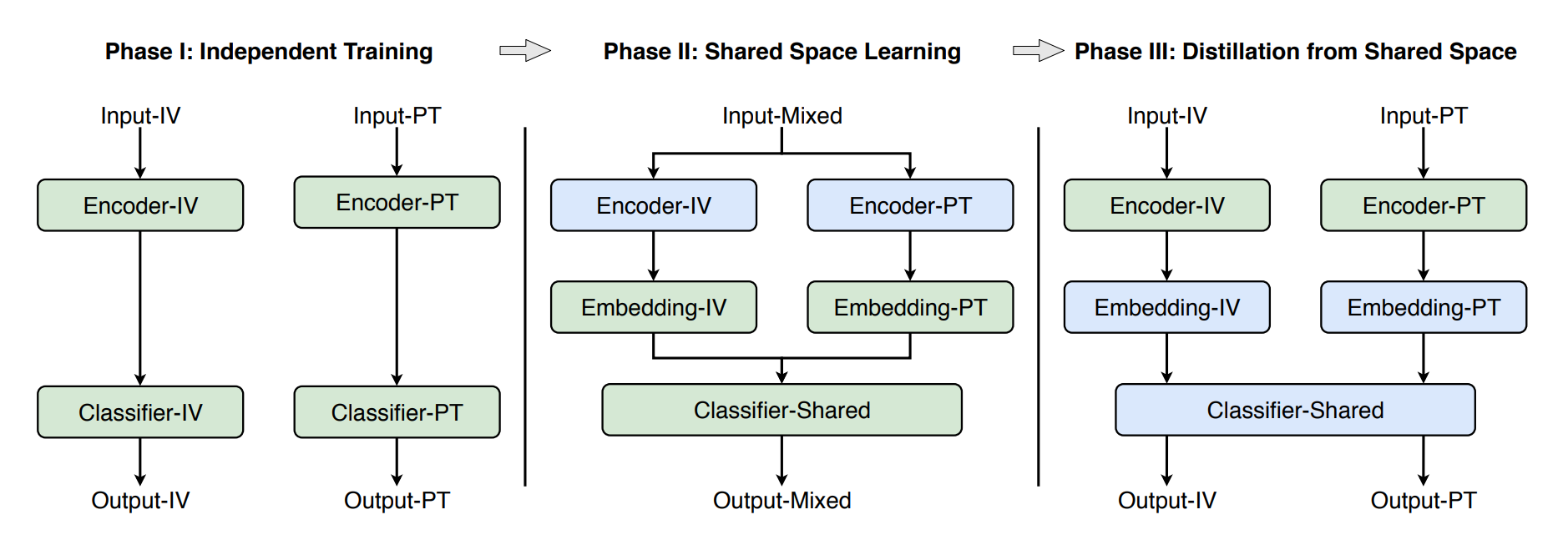
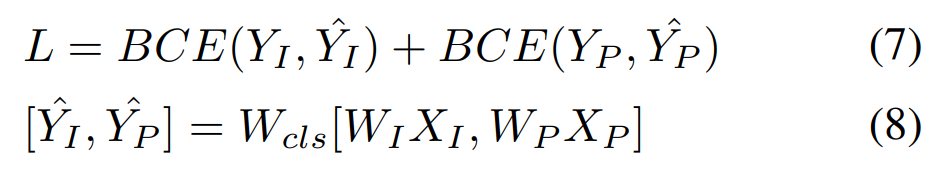


图 3使用特权模式框架进行拟议的ASD筛查的过程。不同的训练阶段在顶部以粗体文本突出显示。Encoder-IV、Encoder-PT、Classifier-IV 和 Classifier-PT 与图 1 所示相同。蓝色模块在训练阶段是固定的，而绿色模块则是优化的。

我们的方法遵循的直觉是，首先构建一个编码多模态知识的共享空间，然后鼓励每个模态的模块从这样的空间中学习。具体来说，为了开发一个共享的嵌入空间，并充分理解每个模态，我们的方法首先独立地优化模型的相应模态（独立训练）。然后，利用所学的模态特定知识，我们整合了每种模态的模型（共享空间学习），并通过在具有以下损失函数L的两种模态上进行联合训练来构建共享空间：



其中 YI 和 YP 是图像查看和拍照模态的实况注释，Yˆ I 和 Yˆ P 是各自的模型预测，BCE 表示二元交叉熵损失。XI和XP是由两种模态（即图3中的Encoder-IV和CocoderPT）的编码器提取的特征，WI和WP表示两种模态的嵌入层（即Embeding-IV，EmbeddingPT），Wcls对应于共享分类器（即分类器共享）。通过固定模态特定的模块（即编码器-IV和编码器-PT），并且仅使用上述方程优化嵌入层以及共享分类器（即嵌入IV，嵌入PT和分类器共享），我们通过学习来自两种模态的基本知识来构建共享空间。

为了将多模态知识从共享空间提炼到每个模态，同时减轻多任务学习中的训练困难，而不是继续进行联合训练[4，5]，我们建议解开不同模态的模型，并分别根据它们自己的模态进行优化。具体而言，在共享空间蒸馏阶段，嵌入层和共享分类器（即嵌入-IV，嵌入-PT和分类器共享）是固定的，并且仅优化每个独立模态的模块（即编码器-IV和编码器-PT），鼓励每个模态的模块通过多模态知识和学习的对齐特征表示来适应共享空间。

上述程序通过学习共享空间将不同的模态连接起来，并鼓励模型从中提炼出多模态知识，以提高每个模态的性能。在我们的上下文中，使用在两种模态之间共享相同的分类器，我们的方法在两种行为模态中学习一致的特征表示，用于ASD筛查，允许它们相互补充，并通过共享空间中编码的多模态知识相互增强其特征表示。我们注意到，所提出的方法适用于在测试中只有部分模式（在我们的例子中只有一种模式）可用并且不存在跨模式的一对一成对关系的情况，这在临床环境中很常见。第4.2节表明，所提出的具有特权模式的方法可以提高ASD筛查模型在两种不同模式上的性能，而第4.3节则分析了在共享空间中学到的知识。

**4.实验**

在本节中，我们将报告实施细节并对所建议的方法进行全面评估。

**4.1实施**

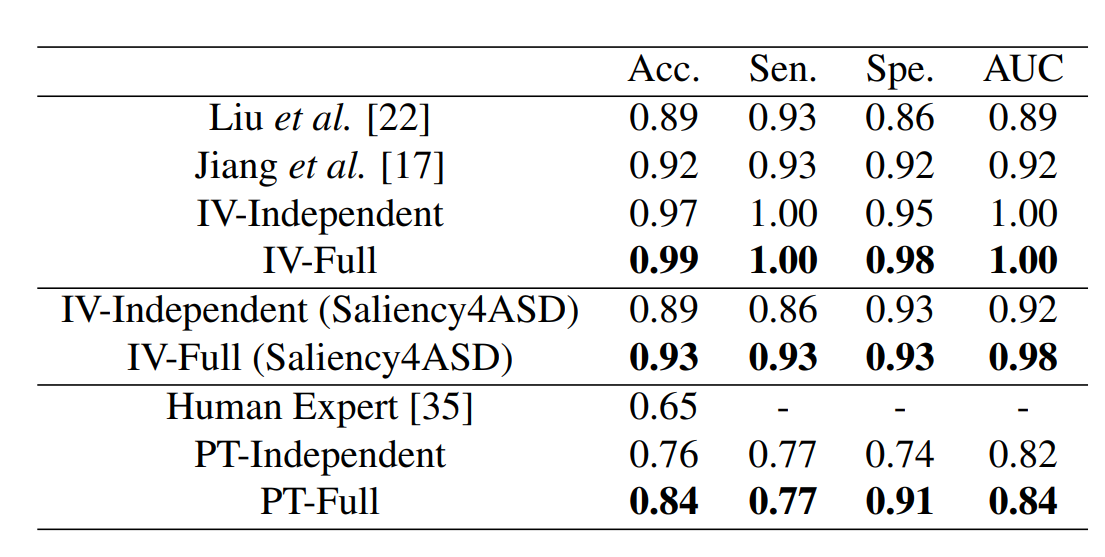
**数据集和评估。**我们首先介绍这项工作中使用的数据。对于照片拍摄，22名患有ASD的人和23名对照组（即年龄，性别和智商匹配的健康人）参加了我们的实验。他们被指示在室内和室外场景中拍照，每人平均拍摄40张照片。对于图像查看，我们使用来自20个ASD和19个控件的眼动追踪数据。在查看来自OSEE[37]眼动追踪数据集的700张图像时记录了双目眼球运动，我们在评估中将患者的不同眼睛视为两个受试者，类似于[17]。使用簇固定法提取眼注[19]。请注意，两种方式之间没有受试者重叠。为了证明我们方法的普遍性，我们还在最近的Saliency4ASD [9]数据集上进行了实验。

对于评估，我们采用[17，22]中广泛使用的留一主题对象交叉验证，它能够返回对错误概率的几乎无偏估计[33]。请注意，由于 Saliency4ASD [9] 不提供受试者 ID，因此我们在训练数据中使用来自同一组（即 ASD 或对照组）的第i个注视序列来构造第i轮交叉验证的验证数据。与[17，22]类似，我们通过准确性，灵敏度（即真阳性率），特异性（即真阴性率）和ROC曲线下面积（AUC）来评估我们的模型。

**型号规格。**我们模型中使用的ResNet-50 [13]首先在ImageNet数据集[8]上进行预训练，然后与两个模型中的其他模块联合优化。LSTM（LSTM\* 和 LSTM）的嵌入大小为 512，而在多模态蒸馏中，对于 Embedding-IV 和 Embedding-PT（见图 3），我们将它们的大小设置为 1024。对于图像查看，我们使用原始图像及其所有相应的眼睛注视作为输入，而对于拍照，我们从特定主题的照片库中随机抽取12张照片（根据经验结果设置的照片数量）作为单个输入序列。

**训练。**为了训练我们基于注意力的ASD筛查模型，对于照片拍摄，我们遍历了所有受试者（不包括用于验证的样本），每个受试者都与从他的照片池中随机抽样的图片序列相关联。对于图像查看，我们使用与[17]相同的图像选择技术来选择前100个判别性图像，这些图像最能区分ASD和对照组之间的凝视模式，并在每轮验证中训练受试者。我们利用具有二进制交叉熵损失的 Adam [18] 优化器来训练使用权重衰减 10−5 和梯度削波 10 来训练我们所有的模型。两个任务的批大小都设置为 12。在独立训练期间，模型被训练为10和180个epoch以进行图像查看和拍照，学习速率初始化为10-4，每2和30个epoch除以2。为了学习共享空间，我们共同优化了学习速率为5×10-6和单个epoch的两种模态的模型（由于两种模态的数据集具有不同的大小，我们不断训练模型，直到处理两种模态的数据）。在成功学习共享空间后，我们分别对模型进行图像查看和拍照（图3中的共享空间蒸馏阶段）进行3个和60个纪元的训练。

表 1 ASD筛查的模型间比较。我们的图像查看数据集Saliency4ASD [9]和我们的照片拍摄数据集的结果被水平线除以，并从上到下列出。IV-独立和PT-独立是我们关于图像查看和拍照的单模态模型。对于这两种模式，我们采用多模态蒸馏的完整型号均表示为 IV-Full 和 PT-Full。使用四个评估指标，包括准确性（ACC.），敏感性（Sen.），特异性（Spe.）和AUC。最佳结果以粗体文本突出显示。



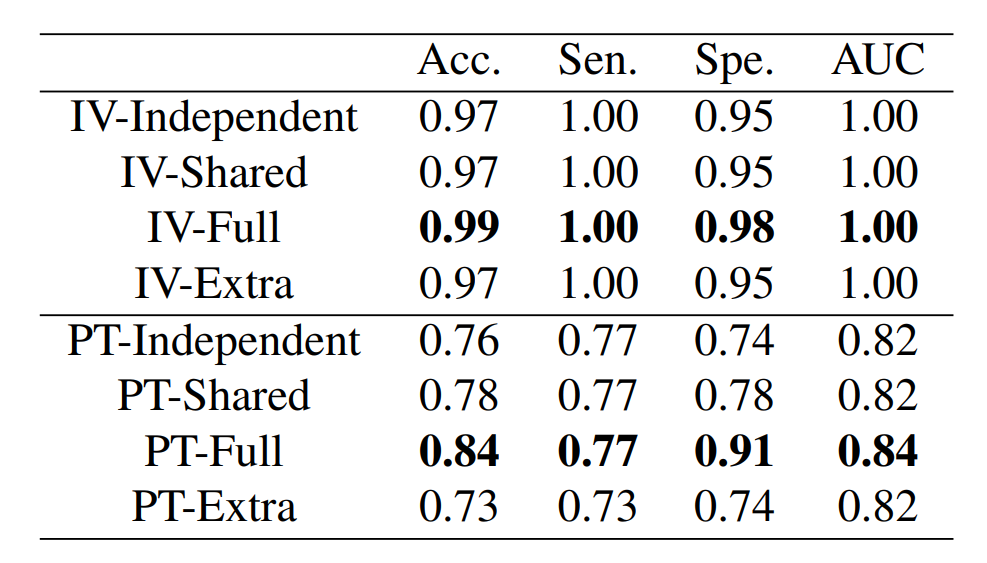
**按主题分类。**由于ASD筛查的评估是在受试者的基础上进行的，因此将我们的样本预测（对不同图像或照片序列的预测）转换为受试者的预测，我们平均了受试者所有样本的置信度（用于图像查看的前100个判别图像和用于拍照的5个随机采样序列），并利用预定义的阈值，即0.5， 识别泛自闭症障碍。

**4.2结果**

在本节中，我们报告了实验结果，以证明我们的ASD筛查的有效性与特权模式框架。我们首先在所提出的模型和相关的最新技术之间进行模型间比较。具体来说，对于图像查看，我们将我们的模型与[17，22]进行比较，后者也根据眼睛注视来筛选ASD，而对于照片拍摄，我们使用人类专家的表现[35]（三个人类专家同意相同标签的百分比，即ASD或对照）作为参考，因为这是这项任务的第一个计算模型。然后，我们在所提出的多模态蒸馏方法的不同阶段对模型进行模型内比较。表1和表2显示了模型间和模型内比较的定量结果。

如表1所示，所提出的具有时间信息的影像查看模态模型在所有评估指标上都能显著优于目前最先进的ASD筛查模型。通过使用循环模块以时间顺序扫描不同眼睛固定的视觉特征，我们的模型在识别ASD个体（敏感性）方面实现了100%的准确率，在区分健康人（特异性）方面实现了95%的准确率。在拍照时识别ASD患者更具挑战性，人类专家的表现为65%。我们的模型在仅使用自己的模式进行训练时显示出合理的性能（总体准确率为76%）。

表 2所提出的多模态蒸馏的模型内比较.在特定模态的每个部分中，前三个结果对应于不同训练阶段后的模型（参见图3，结果按与训练阶段相同的顺序排列），而最后一行（即 - Extra）显示了具有附加层（图3中的嵌入IV或嵌入PT）的单模态性能，并且训练周期的数量与-Full相同。



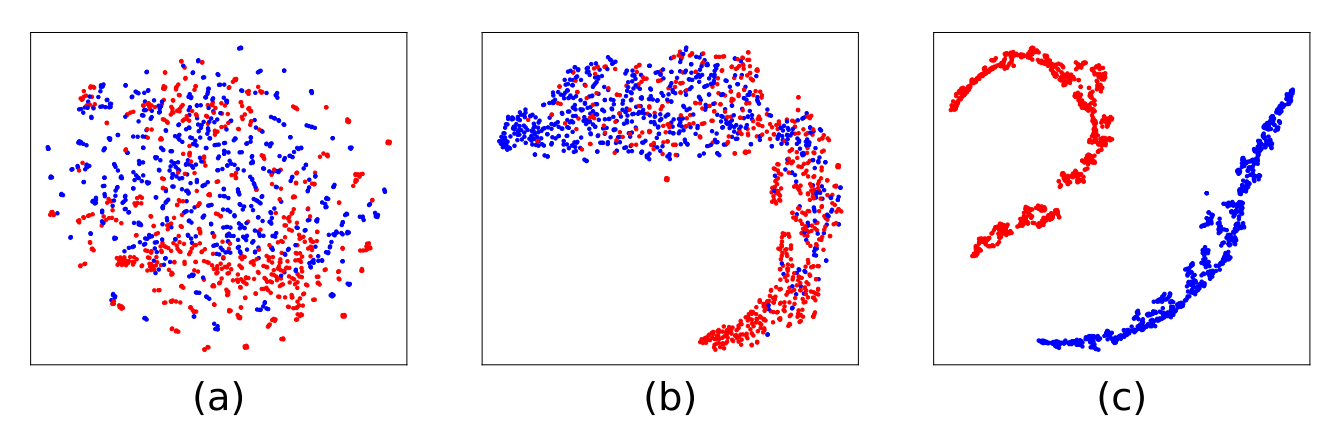


图 4在所提出的多模态蒸馏方法的三个不同训练阶段提取的特征的t-SNE可视化。从（a）到（c）是独立培训，共享空间学习和从共享空间阶段提炼的结果。红点表示 ASD 的样本，而蓝点对应于对照组的样本。

此外，通过将这两种模式与拟议的多模态蒸馏方法相结合，我们能够在这两种模式上实现相当大的改进。具体而言，我们进一步将图像查看的整体精度从97%（单模态性能）提高到99%，并将照片的整体精度从76%显着提高到84%。

根据表2中报告的模型内比较，ASD筛查的性能在三个阶段中单调地增加。特别是，在共享空间学习阶段构建共享空间，我们的多模态蒸馏方法通过在共享空间蒸馏阶段将多模态知识提炼到每个独立模态的模块中显着提高性能。我们进一步研究了从多模态知识中学到的对齐特征，并使用t-SNE [30]可视化比较在三个阶段（拍照模式）中学到的特征。如图4所示，仅从单一模态学习的特征没有足够的区分性，因此具有不同标签（ASD或Control）的数据点混合在一起。在共享空间学习之后，不同标签的样本开始向不同的方向移动。最后，通过将多模态知识从共享空间转移到独立模态的模块，对齐的特征变得更加具有鉴别性，并被很好地分成两个集群。

为了验证所提出的多模态蒸馏方法的贡献，我们在每个单一模态上训练我们的模型，但具有额外的层（图3中的嵌入IV或嵌入PT）和与多模态蒸馏相同数量的训练周期。我们在表 2 中将此方法表示为 -Extra。结果表明，在单个模态上附加层和训练具有不显着（图像查看）甚至负面（拍照）的效果。结果证实，使用所提出的方法在提高基于注意力的ASD筛查性能方面，跨不同模式的提炼知识起着至关重要的作用，我们的改进不仅仅是由于模型修改或额外训练周期的优势。

**4.3共享空间学到了什么？**

到目前为止，我们已经证明，通过学习两种模态共享的空间，并将多模态知识从空间提炼到每种模态的模块，我们能够大大提高ASD筛查的准确性。为了更多地了解我们的多模态蒸馏方法的有效性，在本节中，我们重点通过定性和定量评估来分析在共享空间中学到的知识。更具体地说，我们研究了共享空间在关联这两种模式方面学到了什么，以及为什么它能够在每种独立模式上有益于ASD筛查。

**定性评估**。为了理解共享空间如何关联这两种模态，我们首先提取由独立空间中不同模态的模块计算的特征（在独立训练阶段之后在编码器-IV和编码器-PT中计算的特征）和共享空间中的特征（在共享空间学习阶段之后在嵌入IV和嵌入PT中计算的特征），然后根据其相应的特征匹配两种模态之间的最近输入。我们使用余弦相似度作为匹配最近输入的距离度量，这在自然语言处理中被广泛用于匹配不同有意义的单词[26]。通过比较独立空间和共享空间之间最近的输入，我们能够揭示共享空间中模式的排列。图5显示了匹配示例的定性结果，即照片拍摄（整个照片序列，每个序列有12张由一个对象拍摄的照片）和图像观看（突出显示固定区域的图像）之间的最近输入。每行代表一对匹配项。请注意，在图像查看中，仅比较固定区域。我们有以下三个主要观察结果：

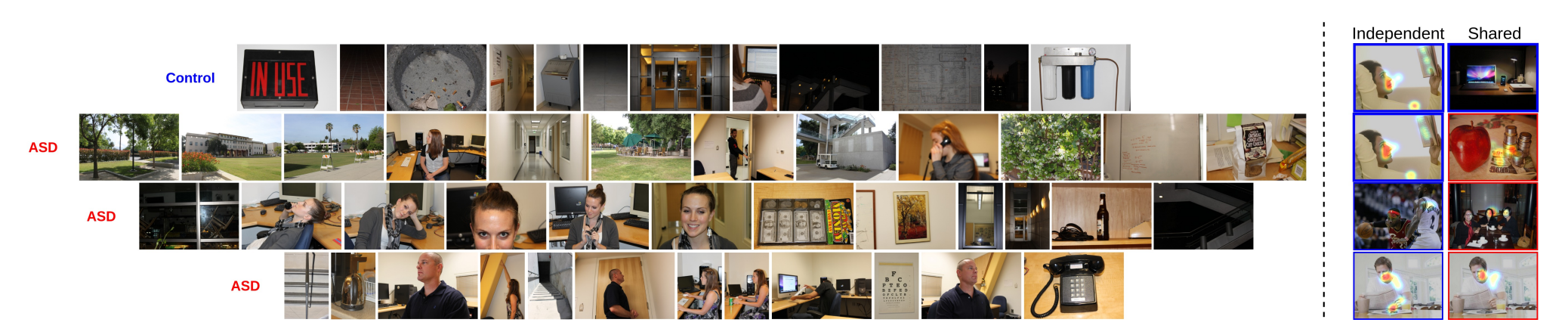


图 5照片拍摄（照片序列，左）和图像查看（具有固定区域的图像，右）之间最近的输入，最好以数字形式进行缩放查看。对于每行，从左到右是整个照片序列（每个序列总共有12张图像，显示在第1-12列上），来自独立（第13列）和共享空间（第14列）中图像查看方式的匹配对应物。图像观察模式中的眼睛固定被可视化为具有喷射色彩映射的高斯模糊显著性图。用于拍照的标签显示在左侧，而用于图像查看的标签则显示为彩色框（ASD为红色，控件为蓝色）。

观察一：独立空间中的匹配示例通常具有不一致的语义含义。例如，在第 1-2 行中，在照片拍摄中有许多非人类对象的照片序列与图像查看中对人脸的注视相匹配。

观察二：通过学习多模态，共享空间中的匹配示例在高级语义意义上表现出一致性。在第1-2行中，与非人类物体的照片序列与对非人类物体（包括笔记本电脑和硬币）的注视相匹配，而在第3-4行中，在照片拍摄中具有许多人脸的照片序列与在图像查看中高度关注人脸的注视相匹配。

观察三：在共享空间中匹配的最近的例子不仅具有相似的语义含义，而且还具有一致的ASD筛查标签。在独立空间中，四行示例中有三行在拍照和图像查看之间具有不一致的标签，而在共享空间中，两种模式的标签是相同的。

这些观察表明，与独立空间不同，通过我们的方法学习的共享空间能够用高级语义概念来桥接这两种模态。此外，将样本与一致的标签（即观察III）精确匹配的能力表明，我们的方法不仅可以将两种模态与其视觉外观（例如语义概念）相关联，还可以将其预测标签相关联。因此，我们能够在两种模态中对齐具有相似语义含义和相同标签（即我们上下文中的ASD或Control）的特征，允许不同的模态通过转移其特定于模态的知识来相互补充，从而提高各自的模型性能。

**定量评估。**为了进一步支持我们在定性评估中的发现，特别是观察III，我们进行了跨模态匹配和检索实验，以定量评估共享空间对一致标签的关联实例的有效性。具体而言，给定来自一种模式（拍照，总共使用45个样本）的源输入以及特定的预测标签，我们计算将它们与具有相同标签的另一种模式（图像查看）中的输入相匹配的准确性。结果表明，使用我们的多模态蒸馏方法，Recall@5（在另一模态中至少5个匹配输入中至少有1个具有一致标签的源输入百分比）从62.2%（独立空间）提高到95.6%（共享空间），证实了我们在定性实验中的观察结果。

**5.结论**

在本文中，我们提出了一种具有特权模式框架的ASD筛查，该框架集成了来自两种不同模式（即拍照和图像查看）的大量信息，并相互提高每种模式的性能。我们的框架具有三个主要的新颖性，包括两个用于在两种模态上进行ASD筛选的DNN模型，以及一种多模态蒸馏方法，该方法将多模态知识从共享空间提取到每种模态。它不需要模态之间的一对一成对关系，也不要求所有模态在测试时的可用性，从而提供了在实际临床环境中利用多个数据源的通用范例。实验结果表明，所提出的模型能够获得新的最新成果，并且跨两种模态的知识提炼进一步提高了它们的性能。