专题综述

跨模态医学影像合成研究与展望

上海科技大学 潘永生 西北工业大学 崔恒飞 夏勇

一、引言

医学影像能够显示身体部位和器官的信息，在疾病诊断、治疗和预后预测中发挥着重要的作用。常见医学影像包括磁共振影像(MRI)、计算机断层成像(CT)、正电子发射成像(PET)、X 光平片、光学成像等。由于成像原理的差异，各种影像获取的信息有所不同，例如 MRI 能提供有关软组织的信息，CT 主要用于成像高电子密度组织(如骨骼)，但也可以提供一定程度的软组织对比度，PET 则使用放射性示踪剂成像特定的生物学功能。同时，根据成像参数和所用试剂的差异，同一种医学影像又有不同的子类型，比如 MRI 包括 T1 加权序列、T2 加权序列等，PET 包括 FDG-PET、Aβ-PET 等。图 1 给出了一些常见的医学影像。

同时包含不同类型或者子类型的医学影像则被称为多模态影像[1]。由于不同影像模态存在一定互补性，多模态影像在临床应用中通常能够比单模态影像提供更多信息。然而，多模态医学影像的获取会遇到采集时间长、费用高、可能增加辐射剂量等困难。因此，人们期待能够使用图像处理技术进行跨模态医学影像合成，即使用某一种(或一些)模态的医学影像去生成另一种(或一些)模态的医学影像[2]。

跨模态医学影像合成虽然能为多模态影像诊断带来便利，但也存在一些技术挑战，例如临床失效问题，即合成影像和真实影像在诊断性能上具有明显的差异[3]。这是因为，各类影像模态的成像原理不同，目标模态能采集的某些信息在源模态影像中并不存在，导致合成的目标模态影像依然不具有这些信息。同时，合成模型受

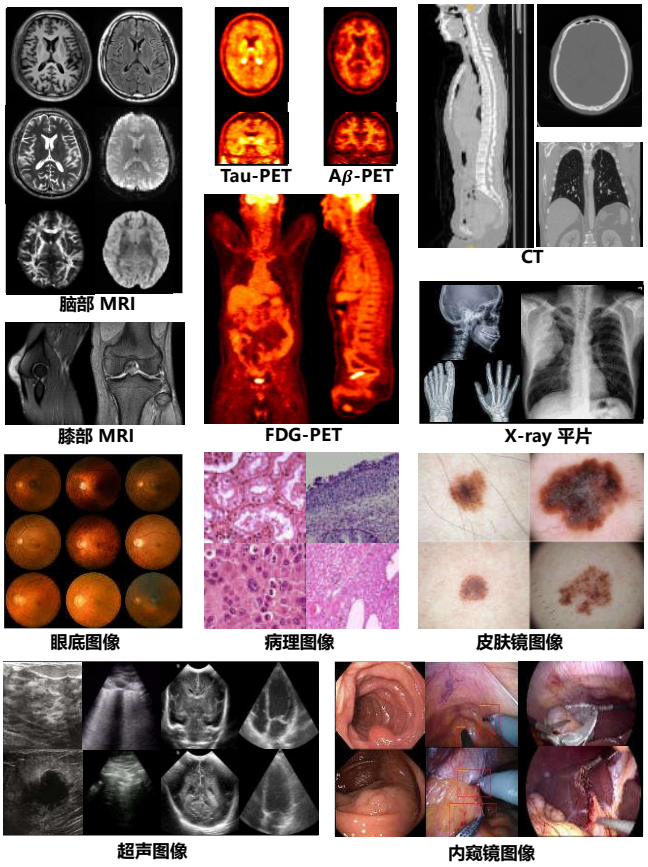

图 1　多种模态的医学影像

到自身表示能力的限制，会产生一定信息损失，特别当其训练过程受到约束条件的偏置化引导时，将在合成影像中产生与偏置相关的信息损失。当前，研究者们大多从模型本身入手，通过提高模型的表示能力或者设计针对具体任务的约束条件来提高合成影像的质量，所开发的跨模态医学影像合成技术已应用于影像采集、重建、配准、分割、检测、诊断等环节，给许多问题带来了新的解决思路和方法。

二、跨模态影像合成技术

2010 年以来，跨模态影像合成受到研究者越来越多的关注，产生了许多合成方法。本文将这些方法大致分为三类，包括传统合成方法、基于深度学习的合成方法和任务驱动的合成方法。

2.1 问题陈述

假设$\mathcal{X}=\{X_1,\ldots,X_s\}\sim\mathfrak{M}_X$ 是采集的S张源模态($\mathfrak{M}_X$)影像,$\mathcal{Y}=\ \{Y_1,\ldots,Y_T\}\sim\mathfrak{M}_Y$是采集的$T$张目标模态($\mathfrak{M}_Y$)影像。跨模态影像合成假设存在一种映射$\mathbb{G}:\mathcal{X}\to\mathcal{Y}$,

$$\forall\ X\sim\mathfrak{M}_X,\exists\ Y\sim\mathfrak{M}_Y\ \ s.t.\,\mathbb{G}(X)=Y.$$

并且，映射$\mathbb{G}$可以通过某种优化方法由给定的数据$\mathcal{X}$和$\mathcal{Y}$近似地估计出来，即

$$\hat{\mathbb{G}}=\ \arg\min_{\mathbb{G}}\mathfrak{D}(X,Y;\mathbb{G}),$$
$$s.t.X\in\mathcal{X},Y\in\mathcal{Y}.$$

其中，$\mathfrak{D}$是反映约束条件的优化目标。目前，影像合成技术均围绕设计映射模型$\mathbb{G}$和优化目标$\mathfrak{D}$而展开，两者相辅相成。典型的映射模型有字典学习、随机森林、卷积网络、编解码网络等，常见的优化目标有平均绝对误差、结构相似性、对抗损失，特征一致性等。

2.2. 传统跨模态影像合成方法

这类方法通常将影像划分成多个小块，并将每个块编码成一个表示向量，通过建立不同模态的配对的块表示向量之间的映射，再根据源模态块的编码产生对应的目标模态块。主要关注表示向量的设计和映射模型的建立，模型的求解过程以类似“数据检索”的方式进行，优化目标通常使用平均均方误差等容易计算的指标。这类方法包括字典学习随机森林等。

基于字典学习的方法[4]假设每个模态存在一个字典，每个图像块均可由字典中元素的稀疏表示得到，不同模态对应的图像块具有相同的字典编码。进行跨模态影像合成时，为不同的模态设置统一的编码和不同的字典，并根据稀疏表示原理通过最小化联合重建误差来求解字典和编码。

基于随机森林的方法[5]将影像合成视为回归问题，假设目标模态块(或其中心点/中心区域)的值是源模态块的因变量，并且这种关系可以通过回归模型得到。这类方法需要首先使用其他方法编码每个图像块，因此非常受编码方式的影响。为了提高表示能力，随机森林使用的表示向量通常是由多个尺度的多种简单特征组合而成的，如空间位置、离散傅里叶系数、类 haar 特征、平均亮度等。

2.3. 基于深度学习的跨模态影像合成方法

随着深度学习的发展，跨模态影像合成研究已逐渐转移到深度学习框架中。从简单卷积神经网络(CNN)，到变分自编码网络(VAE)、U-Net、生成对抗网络(GAN)等，深度学习的各种技术都在跨模态医学影像合成中有所应用。与传统方法相比，此类方法可以直接使用大规模的参数化模型以端到端的方式建立从源模态影像到目标模态影像的映射，并以数据驱动的方式自动提取图像(块)的表示特征，而不需要手工设计表示特征。由于其便于实现和性能优越，基于深度学习的跨模态影像合成技术目前已经占据了主导地位。

2.3.1 基于简单 CNN 的方法

基于简单 CNN 的方法早期常采用与传统方法类似的策略，首先将图像划分成一系列的小块，并使用每个源模态图像块预测目标模态图像块的中心值或中心区域的值。不同的是，这些方法通过堆叠多个卷积层来构建映射模型，并将特征提取和回归预测集成在一起，使用误差反向传播算法来优化模型参数。由于其数据驱动的学习方式，这些方法在训练样本充足的情况下能够达到比传统方法更好的性能。但是，这些方法的表示能力受到参数量和结构复杂性的限制(如使用的卷积层数量和卷积核大小)。因此，早期的 CNN 方法通常沿着增加模型复杂度的方向发展。Rongjian Li 等只使用了两层卷积来建立 MRI 和 PET 之间的映射[6]，Dong Nie 等使用四层卷积来建立 MRI 和 CT 之间的映射[7]，而 Lei Xiang 等则进一步串联了 3 个四层的卷积网络(共计 12 层)来建立 T1-MRI 和低剂量 PET 到高剂量 PET 的映射[8]。然而，增加卷积层的个数在增强表示能力的同时也会增加计算复杂度；同时，由于依然需要逐块甚至逐像素计算目标值，这些方法的计算效率并不比传统方法更

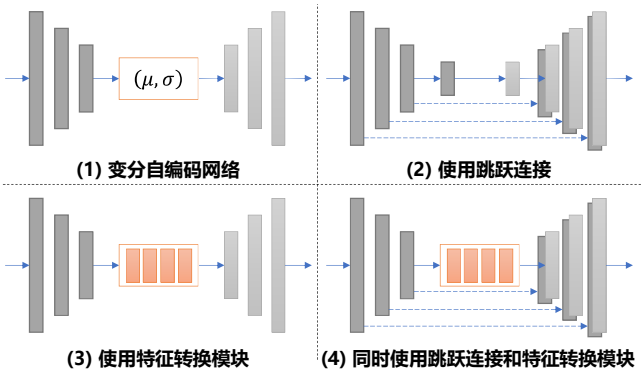

图 2　编解码网络及其变型

具优势。

2.3.2 基于编解码网络的方法

这类方法假设源模态和目标模态的影像在某一隐空间中存在共享的中间编码，因此其通常包含一个编码器和一个解码器，编码器将源模态图像(块)转换成中间编码，解码器将中间编码解码成目标模态图像(块)[9]。编码器通常采用多个带下采样的卷积网络来压缩源模态图像中的信息，解码器则采用多个带上采样的卷积网络来将压缩的信息恢复为目标模态图像。这类方法的优点是可以同时输出一个较大区域甚至整幅图像，不但避免了简单 CNN 方法中的逐像素计算，减少了计算代价，也有助于保留结构性信息，使合成图像具有更好的视觉效果。然而，这是一个“有损压缩”过程，存在信息丢失问题。例如，VAE 模型常采用均值和方差的组合作为中间编码，虽然保留了源模态图像中的统计信息，但同时抑制了个体化信息。有两种策略可以补偿损失的信息。一种是简化编码器和解码器的结构来减少信息损失，并在编解码之间增加额外的特征转换单元(如加入多个残差模块)来增强信息转换能力[2,10]。另一种是使用跳跃连接将编码器的中间特征同时作为解码器的输入来补偿缺失的信息[10,11]。需要主要的是，一般的编解码结构可以适用于配对或非配对的图像，而带跳跃连接的结构则只有在配对的图像上才有良好的性能[12]。图 2 给出了用于影像生成的编解码网络及其变型的结构示意图。

2.3.3 生成对抗网络方法

简单 CNN 和编解码网络一般使用确定性的简单优化目标，如平均绝对误差(MAE)、均方误差(MSE)等，从而会引入确定性偏置，即合成网络的优化方向始终朝向优化目标，导致与优化目标相背的信息传输被抑制，使得合成的图像在被抑制的信息方面呈现出平均的效果，虽然在 MSE、PSNR、SSIM 等指标上表现优异，但看起来却非常模糊。为了克服这种确定性偏置的不利影响，出现了基于 GAN 的影像合成技术[1]。GAN 使用一个判别网络(判别器，$\mathbb{D}$)来指导生成网络(生成器，$\mathbb{G}$)的学习过程：通过交替训练生成器和判别器，使两个网络以相互竞争的方式同步提高各自的能力，最终(在理想情况下)达到纳什均衡。这个过程中，生成器逐渐生成具有真实图像特征的合成图像，判别器不断提高对合成图像的鉴别能力。假设将真实样本和合成样本输入判别器$\mathbb{D}$时的判别标签分别为 1 和 0，那么 GAN 中的$\mathbb{D}$和$\mathbb{G}$通过不断迭代如下两个优化过程求解：

$$\max_{\mathbb{D}} \mathbb{E}_{Y\sim\mathcal{Y}}\left[\log\left(\mathbb{D}(Y)\right)\right] + \mathbb{E}_{X\sim\mathcal{X}}\left[\log\left(1-\mathbb{D}\left(\mathbb{G}(X)\right)\right)\right],$$

$$\min_{\mathbb{G}} \mathbb{E}_{X\sim\mathcal{X}}\left[\log\left(1-\mathbb{D}\left(\mathbb{G}(X)\right)\right)\right].$$

判别器$\mathbb{D}$的存在，使得 GAN 可以直接优化似然度本身，而不是 MAE 等确定性目标的对数似然的下界。$\mathbb{D}$的不断更新相当于不断变换生成网络$\mathbb{G}$的优化目标，因此避免了引入确定性偏置。由于需要训练两个网络，GAN 在训练时相比于其他方法需要更多的计算时间和空间，但在应用时只需要运行生成网络，其时间复杂度和其他深度学习方法是接近的。

作为一种学习策略，任何可微的模型都可以用于构建判别器和生成器，但常用于影像合成的 GAN 通常使用图 2 中的一种结构作为生成器，而判别器则主要使用图 3-(1) 所示的多层卷积结构。训练 GAN 的理想状态是达到纳什均衡，这有时候可以用梯度下降法或其衍生算法做到，但大部分时候是做不到的。目前还没有方法确保能达到纳什均衡，所以相比于使用确定性优化目标的方法，GAN 的训练是不稳定的。为此，确定性优化目标不得不被重新考虑进来，如加入 MAE、SSIM 等。此外，也可以使用感知损失(Perceptual loss)来提高 GAN 的稳定性。但感知损失依赖额外的网络来提取中间特征，并且只能在不同模态的配对图像上计算。如果在固定参数的预训练网络上计算，感知损失实际上相当于非常多但有限的确定性优化目标的组合，仍然会引入一定程度

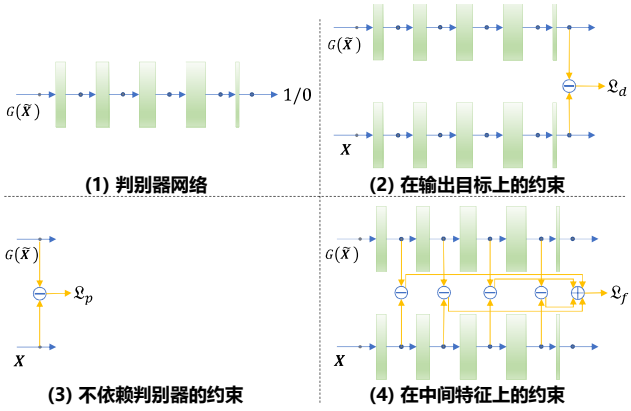

图 3　不同约束目标的对比示意图

的确定性偏置。此外，如果直接在判别网络上计算感知损失，也可以增强 GAN 的稳定性，此时多层感知损失也被称为特征匹配损失[13]。图 3 给出了不同约束目标的对比示意图，通过组合不同的约束目标可以得到 GAN 的不同变型。

三、面向任务的合成方法

许多跨模态影像合成技术并没有考虑具体问题的特点。因此，有的方法可能会产生“幻觉”，即倾向于合成训练数据中显著存在的图像模式，但该模式可能并不是下游任务所需要的[14]。为了解决该问题，需要在通用技术的基础上添加与任务相关的设计，形成对具体任务的偏置，从而使合成的图像保存更多有助于任务的信息，在具体任务上取得性能提升。前述感知损失就是一种这样的设计，其目的是为了平衡合成图像中的空间结构、形状与颜色、纹理等的保持与丢失。

在医学影像合成中，更多的是需要合成的影像对后续的诊断、分割、配准等任务有所帮助，这通常很难通过通用的合成方法实现，因为这些方法实际上主要包含与具体任务无关的偏置，这种无关偏置主导了模型的学习过程，使朝向具体任务的偏置被抑制了，进而产生“无用”的合成影像。这也是合成的医学影像在实际应用中很难得到认可的一个重要原因。为了提高合成的影像的“有用性”，针对具体问题设计专门的合成模型是一种有效的手段。因此，产生了一系列面向任务的影像合成方法[3, 15]。

3.1 面向任务的偏置

假设某种影像合成模型的驱动目标 $\mathfrak{D}$可以分解为与任务相关的部分$\mathfrak{D}_r$和与任务无关的部分$\mathfrak{D}_i$，即

$$\mathfrak{D}(X,Y;\mathbb{G}) = \mathfrak{D}_r(X,Y;\mathbb{G}) + \mathfrak{D}_i(X,Y;\mathbb{G})$$

其中，使用梯度下降算法时，对应的影像合成模型的优化方向为

$$\frac{\partial \mathfrak{D}}{\partial \mathbb{G}} = \frac{\partial \mathfrak{D}_r}{\partial \mathbb{G}} + \frac{\partial \mathfrak{D}_i}{\partial \mathbb{G}}$$

当$\left\|\frac{\partial \mathfrak{D}_i}{\partial \mathbb{G}}\right\| > 0$时，优化会偏向任务无关的方向，合成的影像相对于任务的价值降低；特别当$\left\|\frac{\partial \mathfrak{D}_i}{\partial \mathbb{G}}\right\| \gg \left\|\frac{\partial \mathfrak{D}_r}{\partial \mathbb{G}}\right\|$时，优化完全朝向与任务无关的方向，合成的影像对该任务完全没有价值。在 Cohen 等人给出的例子中，GAN 合成的影像虽然看起来更加真实，但在一些问题下甚至不如只使用 MAE 损失有用[14]，原因就在于判别器的优化方向在任务无关的方向上的分量大于在任务相关方向上的分量。因此，如果需要将合成的影像用于某一下游任务，必须突出任务相关的分量。

为了增加任务相关方向的优化分量，最直接的做法就是使用具体的任务模型作为影像合成目标，即让合成影像具有和真实影像在一个任务相关模型上有相近的输出。设$\mathbb{F}: \mathfrak{M}_Y \to \mathfrak{M}_T$是面向某一任务(如分割、分类、分割，配准等)的模型，其输入为合成模型$\mathbb{G}$的目标模态$Y \in \mathcal{Y} \sim \mathfrak{M}_Y$，输出为$Y$对应的任务标签$T \in \mathcal{T} \sim \mathfrak{M}_T$，该模型的求解目标为$\mathfrak{D}_s(Y,T;\mathbb{F}): T = \mathbb{F}(Y)$，那么将$\mathfrak{D}_t(X,Y;\mathbb{G}) = \mathfrak{D}_s(\mathbb{G}(X),T;\mathbb{F})$加入到合成模型的优化目标中，形成扩展的的优化目标

$$\tilde{\mathfrak{D}}(X,Y;\mathbb{G}) = \mathfrak{D}(X,Y;\mathbb{G}) + \mathfrak{D}_t(X,Y;\mathbb{G}),$$

即可产生一个面向任务的偏置$\frac{\partial \mathfrak{D}_t}{\partial \mathbb{G}}$。实验表明，这样的偏置可以显著提高合成模型对任务的适用性，将合成的影像用于训练对下游任务也有一定的帮助。比如在风格转换任务中，使用多尺度结构相似性约束作为优化目标，可以显著改善图像超分辨和去噪后的失真(即提高了结构相似性)[16]；在从其他模态合成 CT 影像的任务中，使用阈值分割约束即可提高合成的 CT 影像中不同组织(骨头、软组织、脂肪、气体等)间的差异性[17]。

3.2 通过网络模型形成偏置

对于很多实际问题，往往需要使用一个复杂的模型来得到近似的解决方案。例如，用一个卷积网络来进行病灶分割、生存期预测等。尽管这些模型很可能无法达到足够好的性能，但只要它们产生的结果与真实结果之间的误差在可接受范围，依然可以用它们来为影像合成模型产生面向任务的偏置。

同时，除了上述$\mathfrak{D}_t(X,Y;\mathbb{G})=\mathfrak{D}_s(\mathbb{G}(X),T;\mathbb{F})$的偏置目标外，还可以通过替换其中的$T$得到偏置目标的另一种形式：$\mathfrak{D}_t(X,Y;\mathbb{G})=\mathfrak{D}_s\big(Y,\mathbb{F}(\mathbb{G}(X));\mathbb{F}\big)$。与前一种形式相比，这种形式不需要引入训练样本的监督信息，故而在求解合成模型时可以保持和原来相同的数据量。此外，如果任务模型过于复杂，还可以使用图 3-(4)的方式，通过提取任务模型中间层的特征，并使用特征一致性约束来减少计算量。在之前的工作中，我们使用的大都是这样一种形式[3,10,15]。需要注意的是，根据网络的特点，浅层的特征对任务的表达能力通常弱于深层的特征，因此使用浅层的特征形成的对任务的偏置通常也要比深层的特征弱一些。

3.3 嵌入任务模型中的影像合成

虽然使用与任务相关的偏置能够使合成的影像更加适合该任务，但并不总是容易得到一个好的任务模型。因此，我们也希望合成的影像能帮助提高任务模型的性能[15]。此时，可以通过优化模型

$$\min_{\mathbb{F}}\mathfrak{D}_s(\mathbb{G}(X),T;\mathbb{F}),X\in\mathcal{X}$$

来利用从$\mathcal{X}$合成的影像提高任务模型$\mathbb{F}$；同时，$\mathcal{Y}$中的影像也可以通过

$$\min_{\mathbb{F}}\mathfrak{D}_s(Y,T;\mathbb{F}),Y\in\mathcal{Y}$$

同时加入到对$\mathbb{F}$的优化中。此时，合成模型$\mathbb{G}$通过优化

$$\min_{\mathbb{G}}\widetilde{\mathfrak{D}}(X,Y;\mathbb{G})$$

得到，以使其产生对任务的偏置。在优化过程中，$\mathbb{F}$和$\mathbb{G}$的求解要联合进行，其中$\mathfrak{D}_t$项可以根据目标模态影像和标签的数量选择。这个模型为许多任务提供了新的解决思路。以跨模态配准为例，同模态(如T1-MRI($\mathfrak{M}_X$)和T1-MRI($\mathfrak{M}_T$))之间的配准已经有许多成熟的工具，但跨模态(如 FDG-PET($\mathfrak{M}_Y$)和 T1-MRI($\mathfrak{M}_T$))的配准($\mathbb{F}$)依然是一个极具挑战的课题。在这个问题上，我们可以根据 FDG-PET 合成 T1-MRI[18]，进而用同模态配准方法得到配准参数，再应用于 FDG-PET；也可以根据 T1-MRI 合成 FDG-PET($\mathbb{G}$)，进而联合求解$\mathbb{F}$和$\mathbb{G}$得到 FDG-PET 和 T1-MRI 间的模型。

另外，有的任务模型期望同时使用多种模态来达到更好的性能[15,19]，但部分训练样本只有其中一种模态的影像($\mathfrak{M}_X$)，即另一种模态的影像($\mathfrak{M}_Y$)缺失了。此时，我们可以使用合成模型($\mathbb{G}$)来生成缺失的影像，同时提高多模态任务模型(记为$\widetilde{\mathfrak{D}}_s(X,Y,T;\mathbb{F}):T=\mathbb{F}(X,Y)$)的性能。这种情况下，依然可以通过联合优化$\mathbb{F}$和$\mathbb{G}$来求解，只需要将此时的偏差目标改为

$$\mathfrak{D}_t(X,Y;\mathbb{G})=\widetilde{\mathfrak{D}}_s(X,\mathbb{G}(X),T;\mathbb{F})$$

以阿尔茨海默病的影像诊断为例，同时使用 MRI($\mathfrak{M}_X$)和 PET($\mathfrak{M}_Y$)被认为是达到更好性能的有效手段，但许多样本没有采集 PET 影像。因此，多模态 MRI-PET 诊断模型面临数据缺失的问题。使用上述方法便可补全缺失数据，利用所有的样本训练多模态诊断模型。

四、跨模态影像合成的应用

跨模态影像合成技术在成像、重建、配准、分割、预测、诊断等医学影像智能计算的各个任务中都有所应用，同时也涉及到 MRI、PET、CT、X-光平片等影像模态。当前，跨模态医学影像合成的应用场景包括但不限于影像转换、数据扩充、数据统一、隐私保护、可信智能等。从合成图像的目的来说，这些应用大致可分为影像替代和影像补全两类。

4.1 影像的等效替代

某些模态的影像由于条件限制难以获取，但又在实际应用中必不可少。这时，可以尝试利用其他可获得模态的影像合成出该模态的影像。例如，PET 成像通常依赖 CT 影像进行辐射衰减校正，但在 PET/MRI 设备中无法采集 CT 影像，此时可以利用 MRI 影像或者未校正的 PET 影像合成 CT 影像，用以完成 PET 校正。同时，有些模态的影像采集过程需要依赖耗费高或者耗时长的成像方式，难以在临床中广泛的使用。例如，脑血容量(CBV)是评价颅内占位性病变最有用的参数，但 CBV 的

测量依赖于血流灌注成像技术，存在成像时间长、成本高、给患者带来极大不适等明显缺点。考虑到 CBV 影像中的信息可能也部分的存在于其他 MRI 序列中，可以尝试利用多个 MRI 序列来合成 CBV 影像，从而获得一个近似的结果[20]。

4.2 缺失影像的补全

对于有些任务，使用多种模态的影像相互配合,可以达到更好的精度。例如，同时使用 MRI 和 PET 影像可以建立更加准确的阿尔茨海默病诊断模型[3,15]，同时使用多种 MRI 序列可以提升胶质瘤分割精度，以便显现更加完整的病灶面貌[21]。但由于病人意愿或者条件限制等原因，数据不完整的情况非常普遍。这时，可以通过影像合成技术使用已有模态的影像合成缺失模态的影像，从而利用所有的样本来训练模型，并可以将该模型应用于可能存在缺失模态影像的样本上。

需要注意的是，影像替代和影像补全虽然技术上是相似的，但在目的上有着本质的区别。前者假设源模态包含目标模态的完整信息，希望挖掘源模态与目标模态的共有信息，并将这种信息以目标模态的模式呈现出来。后者则假设不同模态中的信息是互补的，希望合成的目标模态影像在共有信息的基础上呈现出与已有模态影像互补的信息，只有这样合成影像才能在下游任务中发挥正向的作用。例如，在 CT 影像中，不同组织通常对应不同的 HU 值(辐射衰减系数)，因此在使用 MRI 影像合成 CT 影像时，希望 MRI 影像所反映的组织分布信息以 HU 值的形式呈现。而在脑肿瘤的分割中，不同模态的影像都只有肿瘤区域的一部分信息，合成的影像只有提供本模态特有的与其他模态互补的信息，才能提高分割精度。这是很难做到的，即便使用面向任务的合成技术也很难达到期望，多模态模型性能的提升通常是得益于样本数量的增加或者数据多样性的提高。

4.3 在数据统一中的应用

医学影像通常呈现出多样化的特点，即使同一种模态的影像，也可能存在分辨率、噪声、数值分布等方面较大的差异。因此，将训练数据集上获得的模型应用于跨中心数据时，其性能会出现难以预测的下降。如果将这种富含多样性的影像视为广义的多模态影像，就可以使用跨模态影像合成技术将多中心的数据变成统一的风格，从而提高任务模型跨中心的迁移能力[22]。

五、总结与展望

本文介绍了跨模态医学影像合成方法及其应用，首先简介了该项研究的意义和技术难点，然后回顾了相关技术的发展和应用场景。目前，跨模态医学影像合成已得到广泛的研究，并在医学影像智能计算的诸多方向中有所应用，给许多问题的解决带来了新思路。但是，当前技术合成的影像和真实影像仍然存在较大的差异，面临明显的失效风险。展望未来，将与应用相关的目标和合成技术相结合，从而产生面向特定应用的合成技术是一条值得探索的可行技术路线。其中，如何融入目标任务相关的知识、如何设计与目标任务相关的模型、如何对影像合成模型进行专业化的改进等，都值得进一步深入研究。

责任编委 储珺

参考文献

[1] AS Fard, DC Reutens, and V Vegh. CNNs and GANs in MRI-based cross-modality medical image estimation. ArXiv Preprint, 2021, abs/2106.02198.

[2] Y Pan, M Liu, C Lian, Y Xia, and D. Shen. Spatially-constrained Fisher representation for brain disease identification with incomplete multi-modal neuroimages. IEEE Transactions on Medical Imaging, 2020, 39(9):2965-2975.

[3] Y Pan, M Liu, Y Xia, and D Shen. Disease-image-specific learning for diagnosis-oriented neuroimage synthesis with incomplete multi-modality data. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021.

[4] Y Huang, L Shao, and AF Frangi. Cross-modality image synthesis via weakly coupled and geometry co-regularized joint dictionary learning. IEEE Transactions on Medical Imaging, 2018, 37(3):815-827.

[5] A Jog, A Carass, S Roy, DL Pham, and JL Prince. Random forest regression for magnetic resonance image synthesis. Medical Image Analysis, 2017, 35:475–488.

[6] R Li, W Zhang, H Suk, L Wang, J Li, D Shen, and S Ji. Deep learning based imaging data completion for improved brain disease diagnosis. International Conference on Medical Image Computing and Computer-Assisted Intervention, 2014, 17 Pt 3: 305-12.

[7] D Nie, X Cao, Y Gao, W Li, and D Shen. Estimating CT image from MRI data using 3D fully convolutional networks. Proceedings of MICCAI workshop on Deep Learning and Data Labeling for Medical Applications (DLMIA), Athens, Greece, October 21, 2016, 170-178.

[8] L Xiang, Y Qiao, D Nie, L An, W Lin, Q Wang, and D Shen. Deep auto-context convolutional neural networks for standard-dose PET image estimation from low-dose PET/MRI. Neurocomputing, 2017, 267: 406-416.

[9] M Liu, T Breuel, and J Kautz. Unsupervised image-to-image translation networks. Advances in neural information processing systems, 2017.

[10] Y Pan and Y Xia. Ultimate Reconstruction: Understand your bones from orthogonal views. 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI), 2021, 1155-1158.

[11] F Isensee, PF Jaeger, SAA Kohl, J Petersen, and KH Maier-Hein. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods 2021, 18:203–211.

[12] H Yang, P Qian, and C Fan. An indirect multimodal image registration and completion method guided by image synthesis. Computational and Mathematical Methods in Medicine, 2020.

[13] T Wang, M Liu, J Zhu, A Tao, J Kautz, and B Catanzaro, High-resolution image synthesis and semantic manipulation with conditional GANs, IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, 8798-8807.

[14] JP Cohen, M Luck, and S Honari. Distribution matching losses can hallucinate features in medical image translation. Medical Image Computing and Computer Assisted Intervention (MICCAI), 2018, 529–536.

[15] Y Pan, Y Chen, D Shen, and Y Xia. Collaborative image synthesis and disease diagnosis for classification of neurodegenerative disorders with incomplete multi-modal neuroimages. Medical Image Computing and Computer Assisted Intervention (MICCAI), 2021, 480–489.

[16] R Malhotra, K Sharma, K Kumar, and N Rath. Integrating SSIM in GANs to generate high-quality brain MRI images. Data Engineering and Communication Technology. Springer, Singapore, 2021, 419-426.

[17] RR Colmeiro, C Verrastro, D Minsky, and T Grosges. Towards a whole body [18F] FDG positron emission tomography attenuation correction map synthesizing using deep neural networks. Journal of Computer Science and Technology, 2021.

[18] Q Yang, N Li, Z Zhao, X Fan, E Chang, and Y Xu. MRI cross-modality image-to-image translation. Scientific Reports. 2020, 10.

[19] J Wei, Y Pan, Y Xia, and D Shen. Learning to synthesize 7T MRI from 3T MRI with few data by deformable augmentation. MICCAI 2021 Workshop on Machine Learning in Medical Imaging (MLMI), 2021.

[20] Y Pan, J Huang, B Wang, P Zhao, Y Liu, and Y Xia. Cerebral blood volume prediction based on multi-modality magnetic resonance imaging. MICCAI 2021 Workshop on Simulation and Synthesis in Medical Imaging (SASHIMI), 2021.

[21] H Jia, Y Xia, W Cai, and H Huang. Learning High-Resolution and Efficient Non-local Features for Brain Glioma Segmentation in MR Images. Medical Image Computing and Computer Assisted Intervention (MICCAI), 2020, 480-490.

[22] H Lei, W Liu, H Xie, B Zhao, G Yue, and B. Lei, Unsupervised Domain Adaptation Based Image Synthesis and Feature Alignment for Joint Optic Disc and Cup Segmentation, IEEE Journal of Biomedical and Health Informatics, 2022 26(1): 90-102.

潘永生

上海科技大学生物医学工程学院博士。研究方向：图像合成、机器学习、疾病诊断。
Email: panysh@shanghaitech.edu.cn

崔恒飞

西北工业大学计算机学院副教授。研究方向：医学影像分析，模式识别。
Email: hfcui@nwpu.edu.cn

夏勇

西北工业大学计算机学院教授。研究方向：医学影像分析，图像处理，模式识别。
Email: yxia@nwpu.edu.cn

专题综述

用于物体位姿估计的端到端概率 PnP

同济大学 陈涵晟 田炜 熊璐

本文是同济大学团队解读其在CVPR 2022获得最佳学生论文奖的工作EPro-PnP[1]。论文研究的问题是基于单张图像估计物体在3D空间中的位姿。现有方法中，基于PnP几何优化的位姿估计方法往往通过深度网络提取2D-3D关联点，然而因为位姿最优解在反向传播时存在不可导的问题，难以实现以位姿误差作为损失对网络进行稳定的端到端训练，此时2D-3D关联点依赖其他代理损失的监督，这对于位姿估计而言不是最佳的训练目标。为解决这一问题，我们从理论出发，提出了EPro-PnP模块，其输出位姿的概率密度分布而非单一的位姿最优解，从而将不可导的最优位姿替换为了可导的概率密度，实现了稳定的端到端训练。EPro-PnP通用性强，适用于各类具体任务和数据，可以用于改进现有的基于PnP的位姿估计方法，也可以借助其灵活性训练全新的网络。从更一般的意义来说，EPro-PnP本质是将常见的分类softmax带入到了连续域，理论上可以推广至训练一般的嵌套了优化层的模型。

一、研究背景

我们研究的是 3D 视觉中的一个经典问题：基于单张 RGB 图像定位其中的 3D 物体。具体而言，给定一张含有 3D 物体投影的图像，我们的目标是确定物体坐标系到相机坐标系的刚体变换。这一刚体变换被称为物体的位姿，记作y，其包含两部分：1)位置(position)分量，可用 3x1 的位移向量t表示；2)朝向(orientation)分量，可用 3x3 的旋转矩阵R表示。

针对这一问题，现有方法可以分为显式和隐式两大类。显式方法也可称作直接位姿预测，即使用前馈网络(FFN)直接输出物体位姿的各个分量，通常是：1)预测物体的深度，2)找出物体中心点在图像上的2D投影位置，3)预测物体的朝向(朝向的具体处理方法可能比较复杂)。利用标有物体真实位姿的图像数据，可以设计损失函数直接监督位姿预测结果，轻松地实现网络的端到端训练，如图1所示。然而，这样的网络缺乏可解释性，在规模较小的数据集上易于过拟合。在3D目标检测任务中，显式方法占据主流[2, 3, 4]，尤其是对于规模较大的数据集，例如nuScenes[5]。

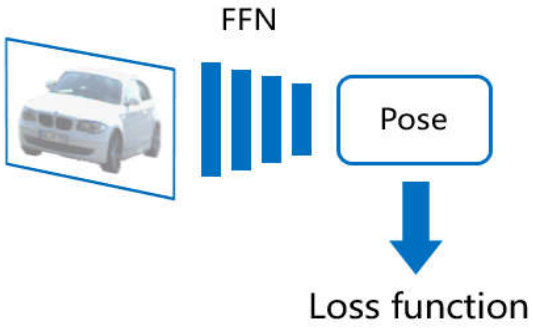

图 1 显示位姿估计网络结构示意图

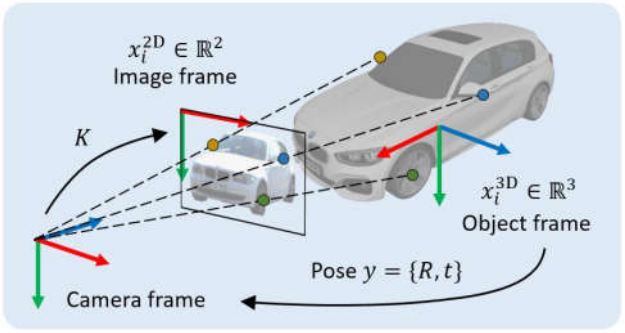

图 2 基于 PnP 的隐式位姿估计方法示意图

隐式方法则是基于几何优化的位姿估计方法，最典型的代表是基于 PnP 的位姿估计方法[6, 7, 8]。这类方法中，首先需要在图像坐标系中找出N个 2D 点(第i点 2D 坐标记作$x_i^{2D} \in \mathbb{R}^2$)，同时在物体坐标系中找出与之相关联的$N$个 3D 点(第$i$点 3D 坐标记作$x_i^{3D} \in \mathbb{R}^3$)，有时还需要获取各对点的关联权重(第$i$对点的关联权重记作$w_i^{2D} \in \mathbb{R}_+^2$)。如图 2 所示，根据透视投影约束，这$N$对 2D-3D 加权关联点隐式地定义了物体的最优位姿。具体而言，我们可以找出使重投影误差最小的物体位姿y^*：

$$y^* = \arg\min_y \frac{1}{2}\sum_i^N \|f_i(y)\|^2$$

其中$f_i(y) = w_i^{2D} \circ \left(\pi\left(Rx_i^{3D} + t\right) - x_i^{2D}\right)$，表示加权重投影误差，是位姿$y = \{R, t\}$的函数。$\pi(\cdot)$表示含有内参的相机投影函数，$\circ$表示元素乘积。PnP 方法常见于物体几何形状已知的 6 自由度位姿估计任务中。

如图 3 所示，基于 PnP 的方法也需要前馈网络去预测 2D-3D 关联点集$X := \{x_i^{3D}, x_i^{2D}, w_i^{2D} | i = 1 \cdots N\}$。相比于直接位姿预测，这一深度学习结合传统几何视觉算法的模型有非常好的可解释性，其泛化性能较为稳定，但在以往的工作中模型的训练方法存在缺陷。很多方法通过构建代理损失函数，去监督X这一中间结果，这对于位姿而言不是最优的目标。例如，已知物体形状的前提下，可以预先选取出物体的 3D 关键点，然后训练网络去找出对应的 2D 投影点位置[7]。这也意味着代理损失只能学习X中的部分变量，因此不够灵活。如果我们不知道训练集中物体的形状，需要从零开始学习X中的全部内容该怎么办?

显示和隐式方法的优势互补，如果能够通过监督 PnP 输出的位姿结果，端到端地训练网络去学习关联点集X，则可以将二者优势结合。为实现这一目标，一些近期研究利用隐函数求导实现了 PnP 层的反向传播[9, 10, 11]。然而，PnP 中的 argmin 函数在某些点是不连续不可导的，使得反向传播并不稳定，直接训练难以收敛。

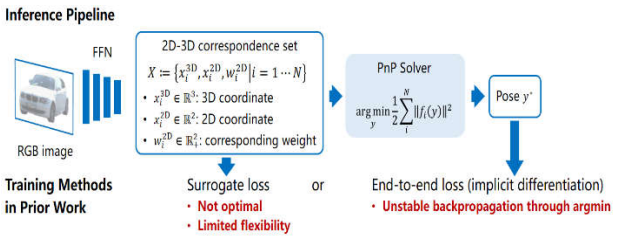

图 3　基于 PnP 的位姿估计网络结构及训练方法示意图

二、EPro-PnP方法介绍

1. EPro-PnP 模块

为了实现稳定的端到端训练，我们提出了端到端概率 PnP(end-to-end probabilistic perspective-n-point)，即 EPro-PnP，如图 4 所示。其基本思想是将隐式位姿视作一个概率分布，则其概率密度$p(y|X)$对于X是可导的。首先基于重投影误差定义位姿的似然函数：

$$p(X|y) = \exp -\frac{1}{2}\sum_i^N \|f_i(y)\|^2$$

若使用无信息先验，则位姿的后验概率密度为似然函数的归一化结果：

$$p(y|X) = \frac{p(X|y)}{\int p(X|y)\mathrm{d}y} = \frac{\exp -\frac{1}{2}\sum_i^N \|f_i(y)\|^2}{\int \exp -\frac{1}{2}\sum_i^N \|f_i(y)\|^2 \,\mathrm{d}y}$$

可以注意到，以上公式与常用的分类 softmax 公式($Softmax(a_i) = \exp a_i / \sum_j \exp a_j$)十分接近，其实 EPro-PnP 的本质就是将 softmax 从离散域搬到了连续域，把求和$\sum$换成了积分$\int$。

2. KL 散度损失

在训练模型的过程中，已知物体真实位姿y_{gt}，则可以定义目标位姿分布$t(y)$。此时可以计算KL散度$D_{KL}(t(y)|p(y|X))$作为训练网络所用的损失函数(因$t(y)$固定，实际上也就是交叉熵损失函数)。在目标$t(y)$趋近于Dirac函数的情况下，基于KL散度的损失函数可

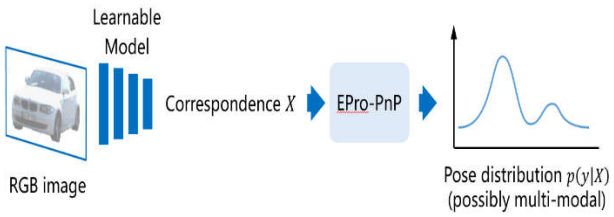

图 4　基于 EPro-PnP 的位姿估计网络结构示意图

以简化为以下形式：

$$L_{\mathrm{KL}} = \frac{1}{2}\sum_{i}^{N}\|f_i(y_{\mathrm{gt}})\|^2 + \log\int \exp -\frac{1}{2}\sum_{i=1}^{N}\|f_i(y)\|^2\,\mathrm{d}y + const$$

如对其求导则有：

$$\frac{\partial L_{\mathrm{KL}}}{\partial(\cdot)} = \frac{\partial}{\partial(\cdot)}\frac{1}{2}\sum_{i}^{N}\|f_i(y_{\mathrm{gt}})\|^2 - \mathop{\mathbb{E}}_{y\sim p(y|X)}\frac{\partial}{\partial(\cdot)}\frac{1}{2}\sum_{i=1}^{N}\|f_i(y)\|^2$$

可见，该损失函数由两项构成，第一项(记作L_{tgt})试图降低位姿真值y_{gt}的重投影误差，第二项(记作L_{pred})试图增大预测位姿$p(y|X)$各处的重投影误差。二者方向相反，效果如图 5(左)所示。作为类比，图 5(右)就是我们在训练分类网络是常用的分类交叉熵损失。

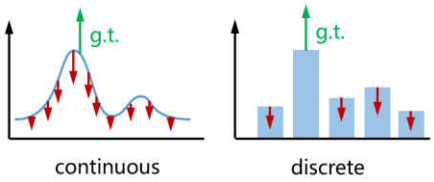

图 5　EPro-PnP 所用的连续损失与离散分类损失的类比

3. 蒙特卡洛位姿损失

需要注意到，KL 损失中的第二项$L_{\mathrm{pred}} = \log\int \exp -\frac{1}{2}\sum_{i=1}^{N}\|f_i(y)\|^2\,dy$中含有积分，这一积分没有解析解，因此必须通过数值方法进行近似。综合考虑通用性，精确度和计算效率，我们采用蒙特卡洛方法，通过采样来模拟位姿分布。具体而言，我们采用了一种重要性采样算法——Adaptive Multiple Importance Sampling(AMIS)[12]，计算出K个带有权重v_j的位姿样本y_j，我们将这一过程称作蒙特卡洛 PnP：

$$\{y_j, v_j | j = 1\cdots K\} = PnP_{\mathrm{MC}}(X)$$

据此，第二项L_{pred}可以近似为关于权重v_j的函数，且v_j可以反向传播：

$$L_{\mathrm{pred}} = \log\int p(X|y)\mathrm{d}y \approx \log\frac{1}{K}\sum_{j=1}^{K} v_j$$

位姿采样的可视化效果如图 6 所示。

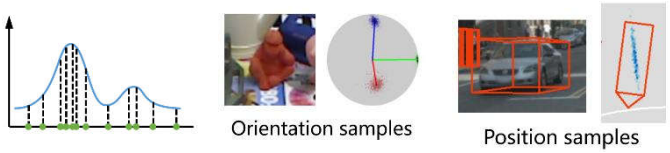

图 6　位姿采样示意图及可视化

4. 针对 PnP 求解器的导数正则化

尽管蒙特卡洛 PnP 损失可以用于训练网络得到高质量的位姿分布，但在推理阶段，还是需要通过 PnP 优化求解器来得到最优位姿解y^*。常用的高斯-牛顿及其衍生算法通过迭代优化求解y^*，其迭代增量是由代价函数$\frac{1}{2}\sum_{i=1}^{N}\|f_i(y)\|^2$的一阶和二阶导数决定的。为使 PnP 的解$y^*$更接近真值$y_{\mathrm{gt}}$，可以对代价函数的导数进行正则化。设计正则化损失函数如下：

$$L_{\mathrm{reg}} = l(y^* + \Delta y, y_{\mathrm{gt}})$$

其中，Δy为高斯-牛顿迭代增量，与代价函数的一阶和二阶导数有关，且可以反向传播，$l(\cdot,\cdot)$表示距离度量，对于位置使用 smooth L1，对于朝向使用 cosine similarity。在y^*与y_{gt}不一致时，该损失函数促使迭代增量Δy指向实际真值。

三、基于EPro-PnP的位姿估计网络

我们在 6 自由度位姿估计和 3D 目标检测两个子任务上分别使用了不同的网络。其中，对于 6 自由度位姿估计，在 ICCV 2019 的 CDPN[8]基础上稍加修改并用 EPro-PnP 训练，用来进行消融实验；对于 3D 目标检测，在 ICCVW 2021 的 FCOS3D[13]基础上设计了全新的变形关联(deformable correspondence)检测头，以证明 EPro-PnP 可以训练网络在没有物体形状知识的情况下直接学出所有 2D-3D 点和关联权重，从而展现 EPro-PnP 在应用方面的灵活性。

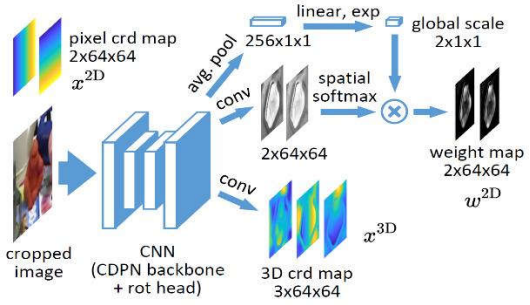

图 7　6 自由度位姿估计网络结构

1. 用于 6 自由度位姿估计的稠密关联网络

网络结构如图 7 所示，只是在原版 CDPN[8]的基础上修改了输出层。原版 CDPN 使用已经检测到的物体 2D 框裁剪出区域图像，输入到 ResNet34 backbone 中。原版 CDPN 将位置与朝向解耦为两个分支，位置分支使用直接预测的显式方法，而朝向分支使用稠密关联和 PnP 的隐式方法。为了研究 EPro-PnP，改动后的网络只保留了稠密关联分支，其输出为 3 通道的 3D 坐标图，以及 2 通道关联权重，其中关联权重经过了 spatial softmax 和 global weight scaling。增加 spatial softmax 目的是对权重w_i^{2D}进行归一化，使其具有类似 attention map 的性质，可以关注相对重要的区域，实验证明权重归一化也是稳定收敛的关键。Global weight scaling 反映了位姿分布$p(y|X)$的集中程度。该网络仅需 EPro-PnP 的蒙特卡洛位姿损失就可以训练，此外可以增加导数正则化，以及在物体形状已知的情况下增加额外的 3D 坐标回归损失。

2. 用于 3D 目标检测的变形关联网络

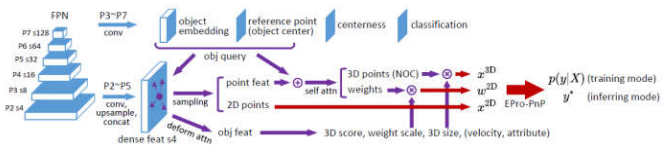

图 8　3D 目标检测网络结构

网络结构如图 8 所示。总体而言是基于 FCOS3D[13]检测器，参考 deformable DETR[14]设计的网络结构。在 FCOS3D 的基础上，保留其 centerless 和 classification 层，而将其原有的位姿预测层替换为 object embedding 和 reference point 层，用于生成 object query。参考 deformable DETR，我们通过预测相对于 reference point 的偏移量得到 2D 采样位置(也就得到了x_i^{2D})。采样后的 feature 经由 attention 操作聚合为 object feature，用于预测物体级别的结果(3D score，weight scale，3D box size 等)。此外，采样后各点的 feature 在加入 object embedding 并经由 self-attention 处理后输出各点所对应的的 3D 坐标x_i^{3D}和关联权重w_i^{2D}。所预测的$x_i^{3D}, x_i^{2D}, w_i^{2D}$全部可由 EPro-PnP 的蒙特卡洛位姿损失训练得到，不需要额外正则化就可以收敛并有较高的精度。在此基础上，可以增加导数正则化损失和辅助损失进一步提升精度(具体细节在我们论文的补充材料中给出)。

四、实验结果

1. 6 自由度位姿估计任务

表 1　6 自由度估计任务的实验结果

Method		ADD-0.1d
CDPN without translation head		74.54
+ Batch=32, LM solver (fair baseline)		79.96
Basic EPro-PnP Loss		**92.66 (+12.70)**
+ Tricks	+ Regularize derivatives	93.43
	+ Initialize from CDPN	95.76
	+ Long schedule (320 ep.)	95.80

使用 LineMOD[19]集进行实验，并严格与 CDPN baseline 进行比对，主要结果见表 1。可见，增加 EPro-PnP 损失进行端到端训练，精度显著提升(+12.70)。继续增加导数正则化损失，精度进一步提升。在此基础上，使用原版 CDPN 的训练结果初始化并增加 epoch(保持总 epoch 数与原版 CDPN 的完整三阶段训练一致)可以使精度进一步提升，其中预训练 CDPN 的优势部分来源于 CDPN 训练时有额外的 mask 监督。

表 2　与其它 6 自由度估计方法的比较

Method	Type	ADD-0.1d
CDPN	PnP + Explicit depth	89.86
HybridPose	Hybrid geometric constraints	91.3
GDRNet	PnP + Explicit depth	93.6
DPOD	PnP + Explicit refiner	95.15
EPro-PnP (ours)	PnP	95.80
PVNet-RePOSE	PnP + Implicit refiner	96.1

表 2 是 EPro-PnP 与各种领先方法[8, 15, 16, 17, 18]的比较。由较落后的 CDPN 改进而来的 EPro-PnP 在精度上接近 SOTA，并且 EPro-PnP 的架构简洁，完全基于 PnP 进行位姿估计，不需要额外进行显式深度估计或位姿精修，因此在效率上也有一定优势。

2. 3D 目标检测任务

使用 nuScenes[5]数据集进行实验，与其他方法对

表 3　3D 目标检测任务的实验结果

Method	Type	NDS↑	mAP↑	mATE↓	mAOE↓
MonoDIS	Explicit (direct prediction)	0.384	0.304	0.738	0.546
CenterNet	Explicit (direct prediction)	0.400	0.338	0.658	0.629
FCOS3D	Explicit (direct prediction)	0.428	0.358	0.690	0.452
PGD	Explicit + Ground constraint	0.448	**0.386**	0.626	0.451
EPro-PnP	PnP	**0.453**	0.373	**0.605**	**0.359**

比结果如表 3 所示。EPro-PnP 不仅相对 FCOS3D 有了明显提升，还超越了 FCOS3D 的另一个改进版本 PGD。更重要的是，EPro-PnP 目前是唯一在 nuScenes 数据集上使用几何优化方法估计位姿的。因 nuScenes 数据集规模较大，端到端训练的直接位姿估计网络已具有较好性能，而我们的结果说明了端到端地训练基于几何优化的模型能做到在大数据集上取得更加优异的性能。

3. 可视化分析

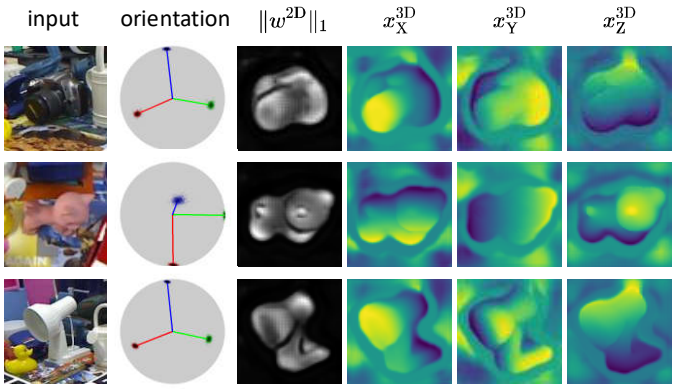

图 9　稠密关联网络的预测结果

图 9 显示了用 EPro-PnP 训练的稠密关联网络的预测结果。其中，关联权重图$\|w^{2D}\|_1$对图像中的重要区域进行了高光，类似于 attention 机制。由损失函数分析可知，高光区域对应的是重投影不确定性较低以及对位姿变动较为敏感的区域。

3D 目标检测的结果如图 10 所示。其中左上视图显示了变形关联网络采样出的 2D 点位置，红色表示w_i^{2D}水平 X 分量较高的点，绿色表示w_i^{2D}垂直 Y 分量较高的点。绿色点一般位于物体上下两端，其主要作用是通过物体高度来推算物体的距离，这一特性并非人为指定，完全是自由训练的结果。右图显示了俯视图上的检测结果，其中蓝色云图表示物体中心点位置的分布密度，反映了物体定位的不确定性。一般远处的物体定位不确定

图 10　3D 目标检测结果

性大于近处的物体。

EPro-PnP 的另一重要优势在于，能够通过预测复杂的多峰分布来表示朝向的模糊性。如图 11 所示，Barrier 由于物体本身旋转对称，朝向经常出现相差 180° 的两个峰值；Cone 本身没有特定的朝向，因此预测结果在各个方向均有分布；Pedestrian 虽不完全旋转对称，但因图像不清晰，不易判断正面和背面，有时也会出现两个峰值。这一概率特性使得 EPro-PnP 对于对称物体不需要在损失函数上做任何特殊处理。

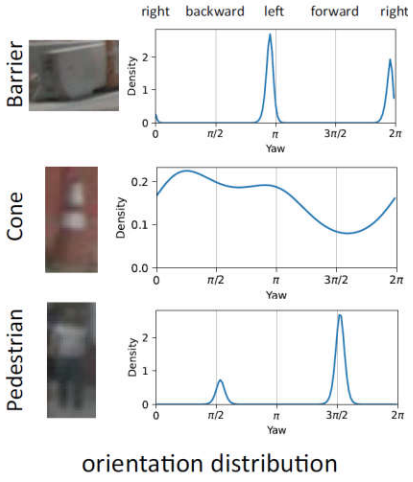

图 11　3D 目标检测网络预测的模糊朝向

四、总结

EPro-PnP 将原本不可导的最优位姿转变为可导的位姿概率密度，使得基于 PnP 几何优化的位姿估计网络可实现稳定且灵活的端到端训练。EPro-PnP 可应用于一般的 3D 物体位姿估计问题，即使在未知 3D 物体几何形状的情况下，也可以通过端到端训练学习得到物体

的 2D-3D 关联点。因此，EPro-PnP 拓宽了网络设计的可能性，例如我们提出的变形关联网络，这在以往是不可能训练的。此外，EPro-PnP 也可以直接被用于改进现有的基于 PnP 的位姿估计方法，通过端到端训练释放现有网络的潜力，提升位姿估计精度。从更一般的意义来说，EPro-PnP 本质是将常见的分类 softmax 带入到了连续域，不仅可用于其它基于几何优化的 3D 视觉问题，理论上还可以推广至训练一般的嵌套优化层的模型。

责任编委 王金甲

参考文献

[1] Hansheng Chen, Pichao Wang, Fan Wang, Wei Tian, Lu Xiong, Hao Li. EPro-PnP: Generalized End-to-End Probabilistic Perspective-n-Points for Monocular Object Pose Estimation. In CVPR 2022.

[2] Ze Dennis Park, Rares Ambrus, Vitor Guizilini, Jie Li, Adrien Gaidon. Is Pseudo-Lidar Needed for Monocular 3D Object Detection? In ICCV, 2021.

[3] Tai Wang, Xinge Zhu, Jiangmiao Pang, Dahua Lin. Probabilistic and Geometric Depth: Detecting Objects in Perspective. In Conference on Robot Learning (CoRL), 2021.

[4] Yue Wang, Vitor Guizilini, Tianyuan Zhang, Yilun Wang, Hang Zhao, Justin Solomon. DETR3D: 3D Object Detection from Multi-View Images via 3D-to-2D Queries. In Conference on Robot Learning (CoRL), 2021.

[5] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, Oscar Beijbom. nuScenes: A Multimodal Dataset for Autonomous Driving. In CVPR, 2020.

[6] Mahdi Rad, Vincent Lepetit. BB8: A Scalable, Accurate, Robust to Partial Occlusion Method for Predicting the 3D Poses of Challenging Objects without Using Depth. In ICCV, 2017.

[7] Sida Peng, Yuan Liu, Qixing Huang, Xiaowei Zhou, Hujun Bao. PVNet: Pixel-Wise Voting Network for 6DoF Pose Estimation. In CVPR, 2019.

[8] Zhigang Li, Gu Wang, Xiangyang Ji. CDPN: Coordinates-Based Disentangled Pose Network for Real-Time RGB-Based 6-DoF Object Pose Estimation. In ICCV, 2019.

[9] Dylan Campbell, Liu Liu, Stephen Gould. Solving the Blind Perspective-n-Point Problem End-to-End with Robust Differentiable Geometric Optimization. In ECCV, 2020.

[10] Bo Chen, Alvaro Parra, Jiewei Cao, Nan Li, Tat-Jun Chin. End-to-End Learnable Geometric Vision by Backpropagating PnP Optimization. In CVPR, 2020.

[11] Eric Brachmann, Carsten Rother. Learning Less is More - 6D Camera Localization via 3D Surface Regression. In CVPR, 2018.

[12] Jean-Marie Cornuet, Jean-Michel Marin, Antonietta Mira, Christian P. Robert. Adaptive Multiple Importance Sampling. Scandinavian Journal of Statistics, 39(4):798–812, 2012.

[13] Tai Wang, Xinge Zhu, Jiangmiao Pang, Dahua Lin. FCOS3D: Fully Convolutional One-Stage Monocular 3D Object Detection. In ICCV Workshops, 2021.

[14] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, Jifeng Dai. Deformable DETR: Deformable Transformers for End-to-End Object Detection. In ICLR, 2021.

[15] Chen Song, Jiaru Song, Qixing Huang. HybridPose: 6D Object Pose Estimation under Hybrid Representations. In CVPR, 2020.

[16] Gu Wang, Fabian Manhardt, Federico Tombari, Xiangyang Ji. GDR-Net: Geometry-Guided Direct Regression Network for Monocular 6D Object Pose Estimation. In CVPR, 2021.

[17] Sergey Zakharov, Ivan Shugurov, Slobodan Ilic. DPOD: 6D Pose Object Detector and Refiner. In ICCV, 2019.

[18] Shun Iwase, Xingyu Liu, Rawal Khirodkar, Rio Yokota, Kris M. Kitani. RePOSE: Fast 6D Object Pose Refinement via Deep Texture Rendering. In ICCV, 2021.

[19] Stefan Hinterstoisser, Stefan Holzer, Cedric Cagniart, Slobodan Ilic, Kurt Konolige, Nassir Navab, Vincent Lepetit. Multimodal Templates for Real-Time Detection of Texture-less Objects in Heavily Cluttered Scenes. In ICCV, 2011.

陈涵晟

同济大学汽车学院 2020 级硕士研究生，导师为熊璐教授，副导师为田炜助理教授，主要研究方向为 3D 计算机视觉。

Email:hanshengchen97@gmail.com

田炜

博士毕业于德国卡尔斯鲁厄理工学院，现任同济大学汽车学院助理教授、硕士生导师，主要研究方向为面向智能驾驶的环境目标感知技术和轨迹预测技术，上海市浦江人才计划入选者，主持国家自然科学基金青年项目、上海市自然科学基金面上项目等横纵向课题，曾参与德国联邦教育研究部、德国博世研究院智能驾驶项目开发，并担任国际会议 IEEE ITSC2015、FUSION2021、CVCI2021 分会场主席，发表智能驾驶领域 SCI/EI 论文近 40 篇，著有专著 2 部。

Email: tian_wei@tongji.edu.cn

熊璐

工学博士、教授、博士生导师。现任同济大学新能源汽车工程中心副主任。长期从事汽车底盘控制、分布式驱动电动汽车动力学控制、智能驾驶相关科研工作，主持和参与国家重点研发计划项目、国家自然科学基金项目、973 计划、863 计划和国家支撑计划等多项国家和省部级项目；发表 SCI/EI 论文 100 余篇，授权专利 40 余项，参撰英文著作 2 部；曾获 2011 年中国汽车工业科技进步三等奖、2013 年上海市科技进步一等奖、2019 年上海市科技进步一等奖等多项奖励。任《同济大学学报》编委和国内外多个期刊的评审专家、国家自然基金和科技部重点研发计划等项目评审专家，担任国际汽车工程师学会(SAE)智能网联汽车技术委员会联合主席、中国汽车工程学会汽车智能交通分会副秘书长、中国汽车工程学会青年委员会副主任委员、中国自动化学会车辆控制与智能化专委会委员。

Email: xiong_lu@tongji.edu.cn

热点追踪

DINE：基于黑盒模型的无监督领域自适应学习

中科院自动化研究所 梁坚 赫然 新加坡国立大学 胡大鹏 冯佳时

一、摘要

为了减轻对标注数据的依赖，无监督领域自适应学习旨在将已有相关标记数据集(源域)中的知识转移到新的未标记数据集(目标域)上。现有的方法需要访问原始的源域数据，并依赖于其中的信息以识别目标样本，这一设置在数据隐私愈发重要的场景下难以得到有效的部署。近两年来，有少量研究希望利用在源数据上学习得到的白盒源模型代替源域数据来进行目标域数据的自适应学习，但这一方式依旧存在模型遭受逆生成攻击而泄露数据的风险。本文探讨了一种有趣的无监督领域自适应的问题设置，即在目标域自适应期间只能接触黑盒源模型(即只有网络预测可见)。具体地，我们提出了一种称为 DINE 的两步知识自适应学习框架。相关成果被 CVPR 2022 录用为口头报告。

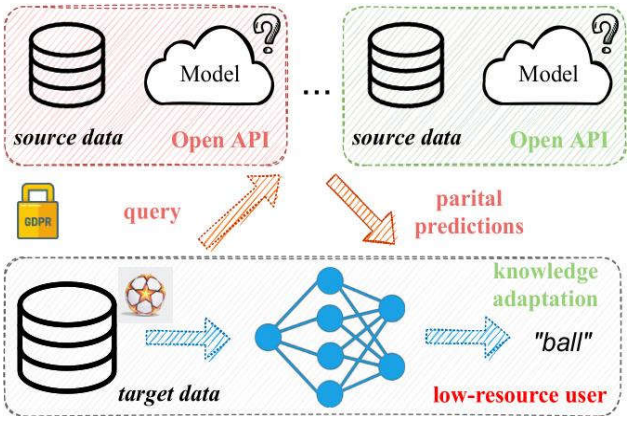

图 1 基于黑盒源模型的无监督领域自适应问题

二、引言

无监督领域自适应学习旨在将已有相关标记数据集(源域)中的知识转移到新的未标记数据集(目标域)上。现有领域自适应方法需要访问原始的源数据，它们通常使用领域对抗性训练[1]或最大平均差异最小化[2]等手段来对齐源域与目标域的特征分布。然而传统领域自

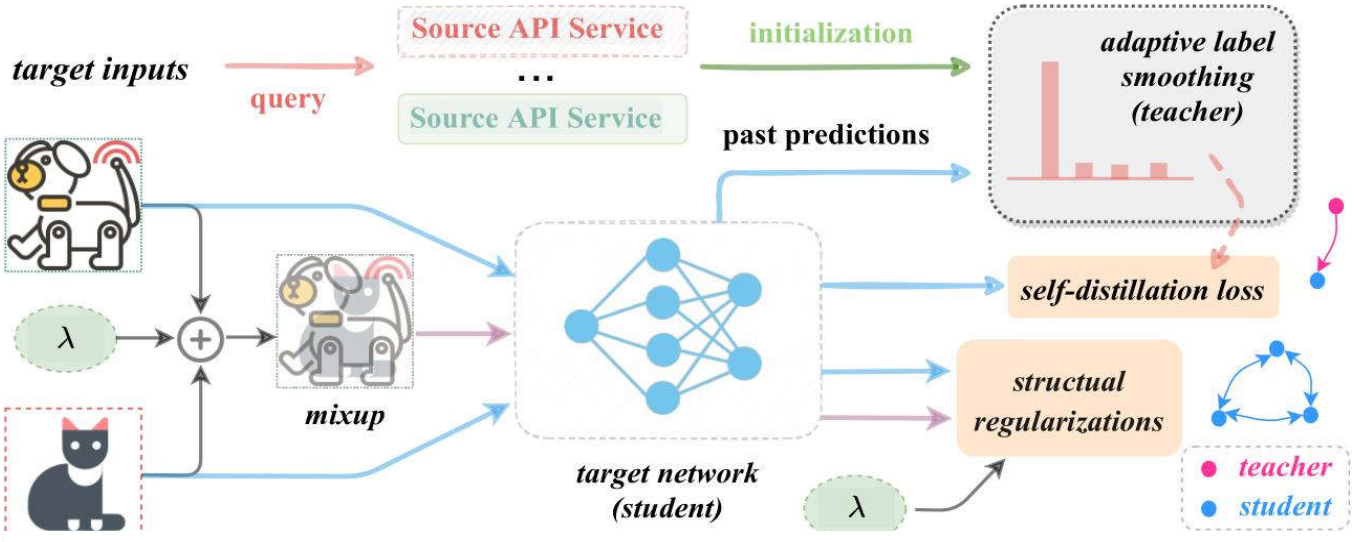

图 2 基于结构化知识蒸馏的无监督领域自适应方法整体框架

适应方法不能很好地应用于对数据安全比较重视的场景(如个人医疗信息、网络浏览历史等隐私数据)。近些年，有研究[3]提出使用在源数据上训练好的模型而非源数据来进行无监督目标域的领域自适应学习，取得了媲美数据依赖方法的识别性能。然而这种基于白盒模型进行知识迁移的方式依旧存在着被对抗生成学习等技术攻击导致数据泄露的风险。为了更好保护源域数据的数据安全，本文研究了基于黑盒模型的领域自适应学习问题(如图 1 所示)。在学习过程中，源数据和源模型信息均无法访问，用户仅能利用源模型的输出概率分布来进行知识迁移。同时，我们受知识蒸馏[4]框架的启发，提出了一种结构化知识蒸馏的新颖方法，在有效去除含噪教师信息的同时充分考虑了目前域数据的潜在结构化约束。此外，新方案不再需要目标域网络架构同源域架构一致，可以在资源受限的客户端采用轻量级网络架构进行知识抽取与整合。大量实验结果表明，我们提出的 DINE 框架在仅使用黑盒模型的条件下可以取得媲美依赖于源数据及基于白盒模型的迁移方法识别性能。

三、正文

本文所提出的 DINE 框架主要由两部分构成，即知识蒸馏阶段以及模型微调阶段，其中第一步为结构化知识蒸馏阶段，整体方案如图 2 所示。

在蒸馏阶段，我们采用适应性自知识蒸馏方法让目标(学生)模型来学习源(教师)模型的输出预测。我们将目标实例在多个源模型上的输出概率分布的平均作为指导分布，最小化目标模型与指导分布之间的 KL 散度损失。然而，由于目标域与源域的差异性，源模型的输出概率并不完全可信，因此我们提出一种自适应的标签平滑策略以调整来自源模型的输出分布。具体来说，我们保留源模型输出分布当中最大的 r 个值(r 默认为 1)，并且将其余类的概率修改为同一个值。通过这种方式，模型能够更加关注到最大的值并忽略部分噪声。不同于伪标签策略，我们不完全依赖于有噪声的伪标签，而是利用最大值作为置信度。为了进一步消除源模型预测中的噪声，我们用源模型的输出分布均值与目标模型预测分布之间的指数移动平均值作为指导分布，形成最终的自蒸馏损失。

此外，为了利用目标域中数据的结构化信息，我们对蒸馏过程进行结构正则化来约束知识蒸馏的过程。首先，我们通过 MixUp[5]来利用成对结构信息，通过最小化成对样本的混合输入在目标模型上的输出分布与一对样本输出分布的混合之间的交叉熵损失来优化网络。在此基础上，我们还考虑了蒸馏过程中目标域的全局结构信息。在蒸馏过程中，由于类别不平衡问题使得部分类相对容易学习，这可能会导致目标模型错误地将一些混淆的目标样本识别为此类。因此我们最大化模型输入与输出之间的互信息以鼓励样本总体的标签分布均匀。最后，在微调阶段，我们再一次应用互信息最大化来进一步精炼上一步蒸馏后的目标模型。

四、实验结果

表 1 DINE 及其他算法在 Office 数据库上的识别结果

Method	Type	A→D	A→W	D→A	D→W	W→A	W→D	Avg.
No Adapt.	Pred.	79.9	76.6	56.4	92.8	60.9	98.5	77.5
NLL-OT	Pred.	88.8	85.5	64.6	95.1	66.7	98.7	83.2
NLL-KL	Pred.	89.4	86.8	65.1	94.8	67.1	98.7	83.6
HD-SHOT	Pred.	86.5	83.1	66.1	95.1	68.9	98.1	83.0
SD-SHOT	Pred.	89.2	83.7	67.9	95.3	71.1	97.1	84.1
DINE	Pred.	91.6	86.8	72.2	96.2	73.3	98.6	86.4
DINE (full)	Pred.	91.7	87.5	72.9	96.3	73.7	98.5	86.7
ResNet-50↑↑, ViT↓↓ (source backbone) → ResNet-50 (target backbone)								
No Adapt.	Pred.	88.2	89.2	74.5	97.2	77.2	99.3	87.6
NLL-OT	Pred.	91.3	91.4	76.4	97.2	78.2	99.4	89.0
NLL-KL	Pred.	91.7	91.8	76.3	97.2	78.4	99.0	89.1
HD-SHOT	Pred.	88.8	90.9	75.3	97.7	77.7	99.5	88.3
SD-SHOT	Pred.	91.6	92.8	77.8	98.7	78.5	**99.7**	89.8
DINE	Pred.	94.2	94.6	80.7	**98.8**	81.5	99.5	91.6
DINE (full)	Pred.	**95.5**	**94.8**	**81.2**	98.5	**82.0**	**99.7**	**91.9**

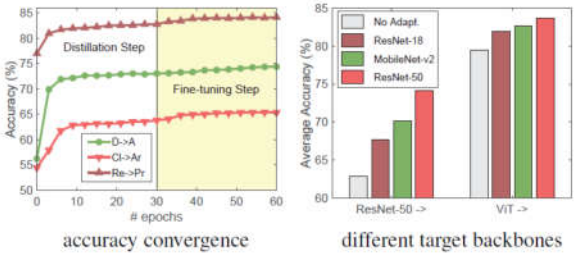

图 3 DINE 的收敛性分析及其对网络架构的敏感度分析

表 1 展示了我们的 DINE 和一些基准方法在 Office 数据库的识别准确率，其中上、下两部分分别为使用 ResNet50 和 ViT 作为源域基础架构进行迁移的结果。可以直观地发现，DINE 具有很大的性能优势。图 3 进一步展示了 DINE 在两个阶段的收敛性以及整体算法对目标域不同架构的敏感度。

责任编委 崔海楠

参考文献

[1] Ganin, Yaroslav, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. "Domain-adversarial training of neural networks." Journal of Machine Learning Research 17, no. 1 (2016): 2096-2030.

[2] Long, Mingsheng, Yue Cao, Jianmin Wang, and Michael Jordan. "Learning transferable features with deep adaptation networks." In International Conference on Machine Learning, pp. 97-105. PMLR, 2015.

[3] Liang, Jian, Dapeng Hu, and Jiashi Feng. "Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation." In International Conference on Machine Learning, pp. 6028-6039. PMLR, 2020.

[4] Hinton, Geoffrey, Oriol Vinyals, and Jeff Dean. "Distilling the knowledge in a neural network." arXiv preprint arXiv:1503.02531 2 (2015)..

[5] Zhang, Hongyi, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. "mixup: Beyond empirical risk minimization." arXiv preprint arXiv:1710.09412 (2017).

梁坚

中科院自动化研究所副研究员。主要研究方向为领域自适应、特征表示及迁移学习等。
Email: jian.liang@nlpr.ia.ac.cn

赫然

中科院自动化研究所研究员。主要研究方向为视觉内容生成、生物特征识别、机器学习等。
Email: rhe@nlpr.ia.ac.cn