arXiv:2303.02968v2 [cs.CV] 7 Mar 2023

DwinFormer：用于端到端单目深度估计的双窗口 Transformer

Md Awsafur Rahman，*IEEE* 学生会员，以及 Shaikh Anowarul Fattah，*IEEE* 高级会员

摘要—从单张图像中估计深度在计算机视觉领域至关重要，并具有广泛的应用。传统方法受局部感受野的限制，难以兼顾一致性与精细细节，因此其实用性受到影响。这种长程依赖的缺失，根源在于架构中的卷积神经网络部分。本文提出一种双窗口 Transformer 网络 DwinFormer，用于端到端单目深度估计，同时利用局部特征和全局特征。DwinFormer 由双窗口自注意力 Transformer 和交叉注意力 Transformer 构成，分别称为 Dwin-SAT 和 Dwin-CAT。Dwin-SAT 能够无缝提取精细且具备局部感知能力的特征，同时捕获全局上下文。它利用局部和全局窗口注意力，有效捕捉短程与长程依赖，无需采用注意力掩码或窗口移动等复杂且计算开销巨大的操作。此外，Dwin-SAT 引入了归纳偏置，从而具备平移等变性等理想特性，并降低了对大规模数据的依赖。此外，传统解码方法通常依赖跳跃连接，这在融合编码器与解码器特征时可能造成语义差异，并导致全局上下文缺失。相比之下，Dwin-CAT 同时采用局部和全局窗口交叉注意力，无缝融合编码器与解码器特征，兼顾精细的局部信息和具备上下文感知能力的全局信息，从而有效弥合语义鸿沟。在 NYU-Depth-V2 和 KITTI 数据集上开展的大量实验所获得的实证结果表明，所提出的方法优于现有方法，并在室内和室外环境中始终取得更佳表现。

索引词—**注意力机制、计算机视觉、深度估计、Transformer**

I. 引言

深PETH度信息在许多不同领域都有广泛的应用，[1]—[4]，例如重新对焦/散景效果、自动驾驶、机器人运动、增强现实等。深度传感器（如激光雷达和飞行时间传感器）价格昂贵，且只能采集低分辨率的稀疏深度信息，因此亟需开发单图像（单目）深度估计方法，以获取高分辨率、稠密的深度信息。然而，仅依赖单张图像进行深度估计是一个病态问题 [5]，因为场景中往往没有足够的信息来准确确定物体的深度，因此这一任务极具挑战性。近年来，基于卷积神经网络（CNN）的深度学习方法已

M. A. Rahman 和 S. A. Fattah 任职于孟加拉国达卡 1000 号孟加拉国工程与技术大学电气与电子工程系，电子邮箱：(awsaf49@gmail.com 和 fattah@eee.buet.ac.bd)。

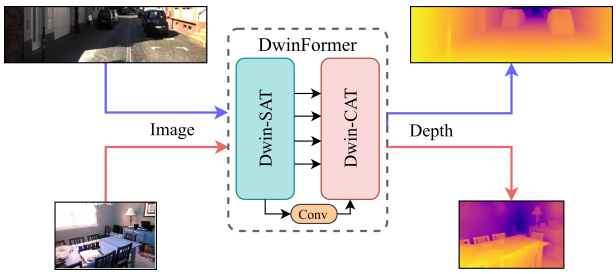

图 1：所提出方法的图示摘要。

主导了单目深度估计 [6]。但是，卷积神经网络受其固有局部性的限制，无法实现全局上下文感知的逐像素预测，因此在细节丰富且一致的深度图之间存在权衡 [7]。尽管有多种方法试图缓解这些局限 [8]，上述问题依然存在。

近年来，Transformer [9]–[11] 在单目深度估计方面展现出令人期待的成果。它们以能够捕获数据中的长程依赖而著称。几年前，视觉Transformer [12] 将纯Transformer应用于图像，取得了超越卷积神经网络的表现，验证了Transformer在计算机视觉中的实用性。Transformer已应用于逐像素预测，解决了局部感受野受限的问题。然而，它们也带来了新的挑战，包括缺乏多尺度特征、对图像分辨率的要求高于文本，以及自注意力的计算复杂度随图像尺寸呈二次增长。此后，研究人员尝试通过分层Transformer [13] 解决这些问题，但其对全局感受野（跨窗口连接）的覆盖仍然有限 [14]。近期研究，如 [14]和 [15]，在解决这些问题方面取得了进展。尽管如此，这些研究也各自面临挑战。 [15]中采用全局稀疏注意力，导致图像尺寸增大时全局特征的质量下降。与此同时， [14] 所述研究存在全局注意力效率低下的问题，造成关键信息丢失和性能不佳。

尽管上述近期基于 Transformer 的方法在使深度边缘与物体边界对齐方面取得了成功，但由于难以有效融合编码器和解码器特征，它们通常难以为像素准确分配深度标签。通常使用跳跃连接来融合编码器和解码器特征，即在拼接特征后应用卷积。由于卷积固有的局部性，语义信息的流动受到限制，无法跨越

较长距离，进而影响模型为像素预测正确深度标签的能力。为缓解这一问题， [16] 文献提出了跳跃注意力模块，该模块使用基于局部窗口的交叉注意力，以考虑上下文的方式整合编码器-解码器特征。尽管基于 Transformer 的编码器-解码器特征融合具有优势，跳跃注意力模块仍受限于局部窗口注意力有限的感受野，因此只能有效利用有限范围内输入像素的信息，而未能利用全局信息，这可能会限制其整体性能。本文提出了一种基于 Transformer 的架构，用于端到端深度估计，同时解决上述现有方法的问题。所提方法的主要贡献可概括如下：

1）提出了一种双窗口 Transformer 网络 DwinFormer，用于端到端单目深度估计。本文引入双窗口自注意力（Dwin-SAT）和交叉注意力（ Dwin-SAT）Transformer，以有效捕获长程依赖和局部细粒度细节。2）提出的 Dwin-SAT 为基于 Transformer 的骨干网络引入了一种有效设计，通过同时利用局部和全局窗口注意力，捕获局部和全局上下文信息，从而缓解深度图中细节与一致性之间的权衡。3）为弥合编码特征与解码特征之间的语义鸿沟，提出的 Dwin-CAT解码器无缝融合编码器和解码器特征，同时兼顾细粒度局部信息和具备上下文感知能力的全局信息。

II. 相关工作

Eigen 等人 [6] 首先提出了一种由粗到精的深度估计方法。Zhou 等人 [17] 首次提出了一种从单目视频中预测相机自运动和深度图的方法。随后，Laina 等人 [18] 提出了一种完全基于卷积神经网络的残差网络，并引入上投影模块以提升性能。Ranjan 等人 [19] 将深度估计与光流估计及运动分割联合训练，以实现协同效果。Yourun 等人 [20] 开发了一种新颖的多尺度方法，通过对预测深度图进行下采样，并在多个分辨率下执行图像合成，以提升模型性能。

近年来，Transformer 因能够捕获长程依赖而成为编码器的热门选择。TransDepth [10], 和 DPT [9] 最早采用视觉 Transformer（ViT）编码器和卷积解码器进行深度估计。随后，SwinDepth [21] 采用名为 Swin Transformer [13] 的分层视觉 Transformer 作为编码器，并使用多尺度卷积块作为解码器。另一方面，AdaBin [22] 使用一个额外的基于 Transformer 的模块，通过将可能的深度范围划分为多个区间，并将区间中心进行线性组合，来自适应地估计图像的深度值。该领域的研究还很少。不过，最近 PixelFormer [16], 采用基于窗口注意力的 Transformer，有效融合了由编码器和解码器生成的特征。这种方法避免了简单拼接这些特征所带来的局限，因为这种做法可能会引发语义鸿沟 [23]。尽管该方法采用基于 Transformer 的编码器-解码器特征融合，但由于局部窗口注意力的限制，其跳跃注意力模块受到输入像素范围有限的严重制约，因而未能充分发挥潜力。

本质上，卷积神经网络在深度图的细节与一致性之间难以兼顾；而 Transformer 虽有望捕获全局信息，却受限于较高的计算复杂度、缺乏多尺度特征以及归纳偏置。近期研究尝试通过不同类型的窗口注意力克服这些局限，但也面临新的挑战，例如全局覆盖有限、全局上下文受损以及全局上下文丢失。同时，由于存在语义鸿沟，编码器与解码器特征的融合仍然是一项挑战；尽管跳跃注意力模块 [16] 改进了融合过程，但仍受其局部感受野的限制。

III. 方法

A. 问题定义

所提出的方法将 RGB 图像的深度图建模为所有像素处最大深度的概率。给定图像 $I \in \mathbb{R}^{H\times W\times 3}$，我们构建模型 $\mathcal{F}(\Theta)$，以预测概率掩码 $\hat{}p \in [0,1]^{H\times W\times 1}$，表示图像中所有像素处最大深度的可能性。该掩码被转换为深度图 $\hat{}y = \hat{}p \odot$ 最大深度。通过相对于真实深度图 y，最小化所选损失函数 $\mathcal{L}(\Theta,\ y,\ \hat{}y)$，优化模型参数 Θ，即通过 $\Theta^* = \arg\min_\Theta \mathcal{L}(\Theta,\ y,\ \hat{}y)$。

B. 所提出的 *DwinFormer* 架构概述

如图2所示，所提出的架构引入了双窗口自注意力 Transformer（Dwin-SAT）主干网络来处理输入图像 I。Dwin-SAT 主干网络采用多个双窗口自注意力（ Dwin-SA）层，提取表示图像在不同分辨率尺度下信息的特征图。这些分辨率尺度由 i 表示，定义为 $\frac{1}{2^{i+1}}$ 相对于输入图像 I，其中 $i \in \{1,2,3,4\}$ 和 $i = 5$ 的分辨率与 $i = 4$相同。Dwin-SA 利用局部窗口自注意力（ Lwin-SA）和全局窗口自注意力（Gwin-SA），从输入图像中提取具有局部和全局上下文信息的特征。随后，所提出的双窗口交叉注意力（Dwin-CA）会进一步优化这些特征，并在不同阶段用于分层融合编码器-解码器特征图。Dwin-CA 分别利用局部窗口交叉注意力（Lwin-CA）和全局窗口交叉注意力（Gwin-CA），将编码器-解码器特征与局部和全局上下文融合。最后，Dwin-CA 层堆叠后得到的解码特征由DepthHead 模块处理，以估计最大深度的逐像素概率。下文将介绍 DwinFormer 的各个模块。

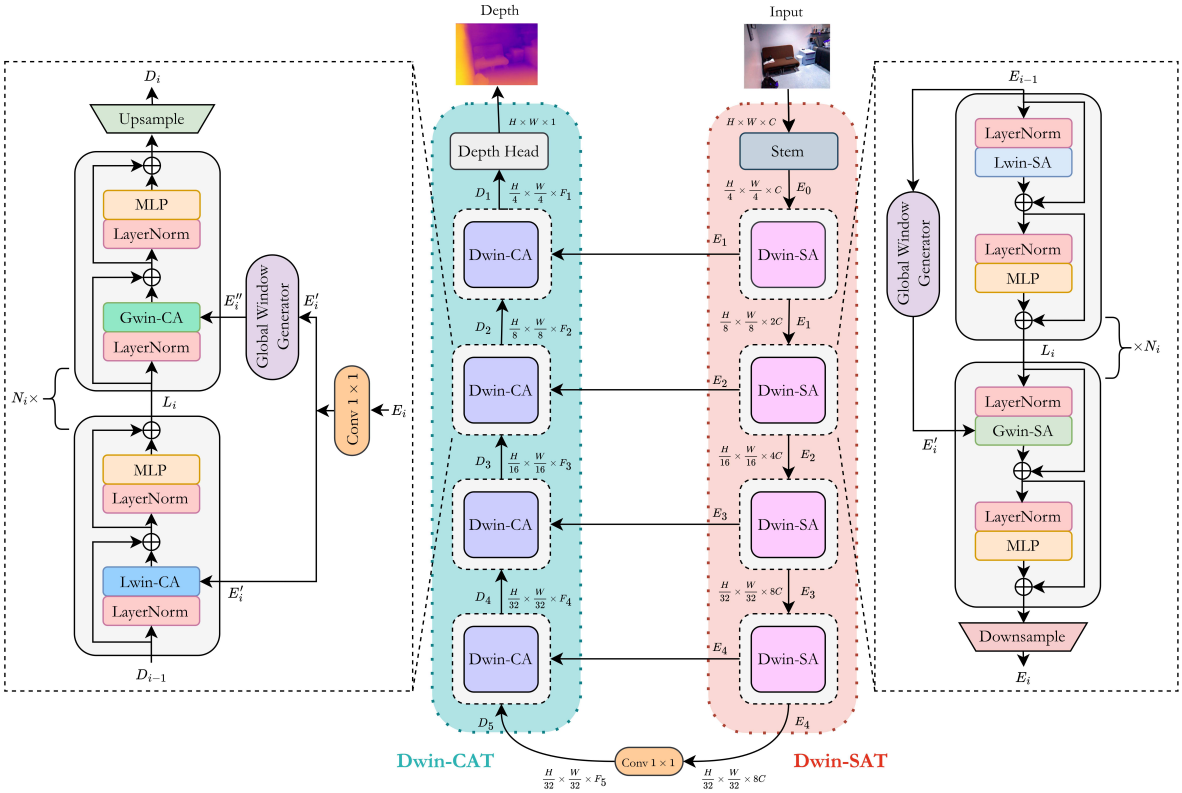

图2：DwinFormer 架构示意图。输入图像通过 Dwin-SAT 中的一系列 Dwin-SA 模块编码为多尺度特征图。随后，Dwin-CAT 将编码特征重建为深度图；该模块融合了同一层级的 Dwin-SA 特征和前一层级的 Dwin-CA 特征。最后，深度头对解码特征进行处理，生成最终深度图。

C. 所提出的编码器架构：*Dwin-SAT*

如图2所示，所提出的 Dwin-SAT 主干网络通过迭代地将空间维度缩小、将通道维度扩大 2 倍，有效提取多个分辨率下的特征。该过程始于 Stem 模块，它从输入图像 I 中生成重叠图块，以便应用 Transformer。随后，这些图块通过一个步幅为 C的 3×3 卷积层投影到 2×2 维嵌入空间中，并采用 1×1 填充。接着，使用步幅卷积将图像块数量减少一半，使 Dwin-SAT 主干网络能够持续提取不同细节层次的特征。最后，Dwin-SAT 提取的特征经由 1×1 卷积处理，以精确控制解码器的通道数。具体而言，Dwin-SAT 使用 Dwin-SA 模块，通过交替使用具有局部和全局上下文的 Lwin-SA 与 Gwin-SA 模块来提取空间特征。与 Gwin-SA 不同，Gwin-SA 使用全局窗口和局部窗口执行全局注意力，而 Lwin-SA 则按照 Swin Transformer [13]中所述的方式，仅使用局部窗口执行局部注意力。Gwin-SA 中的全局窗口由全局窗口生成器（GWG）根据整个输入特征图生成，并由所有局部窗口共享，因此减少了参数量和 FLOPs。它将信息限制在整个输入特征用于与局部查询特征进行交互的映射。具体而言，如图3所示，生成器由一系列 Fused-MBConv [14] 模块和一个最大池化层组成。然后，在每个阶段的末尾使用下采样模块，降低所得特征的空间维度，并将通道维度增加2倍。下采样模块采用与 [14] 相同的架构，包括一个改进的 Fused-MBConv 模块，后接一个 2×2 卷积核大小为 3×3 的步幅卷积和一个层归一化层；该结构利用卷积神经网络中的空间特征收缩，从而引入归纳偏置并促进通道间交互。在 Dwin-SAT 的末端，使用一个 1×1 卷积来控制各尺度下每个解码器层的输出通道数。对于给定尺度 i，编码器可表示为

$$
\begin{aligned}
{}_1Q_l^e &= LayerNorm(E_{i-1}) \\
K_l^e &= MLP(E_{i-1}); V_l^e = MLP(E_{i-1}) \\
L_i &= \text{Lwin-SA}({}_1Q_l^e, K_l^e, V_l^e) + E_{i-1} \\
L_i &= L_i + MLP(LayerNorm(L_i)) \\
{}_2Q_l^e &= LayerNorm(L_i); E_i' = GWG(E_{i-1}) \\
K_g^e &= MLP(E_i'); V_g^e = MLP(E_i') \\
E_i &= \text{Gwin-SA}({}_2Q_l^e, K_g^e, V_g^e) + L_i \\
E_i &= E_i + MLP(LayerNorm(E_i)) \\
E_i &= \text{Downsample(E}_\text{i}\text{)}
\end{aligned}
\tag{1}
$$

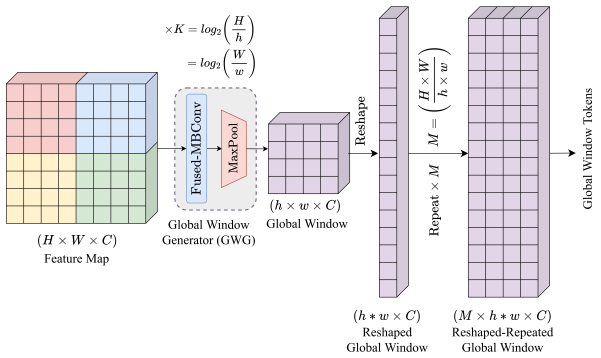

图3：输入特征图的维度为 H，W，随后被送入全局窗口生成器（GWG）。GWG 反复使用融合 MBConv 和最大池化层 K 次，生成维度为 h，w 的全局窗口，再将其重塑并重复 M 次。

D. 双窗口自注意力（*Dwin-SA*）

Dwin-SAT 主干网络的主要计算算子是 Dwin-SA。一般而言，它通过上下文化过程进行运算，即计算每个查询位置与所有键位置之间的成对关系，以形成注意力图。该图充当加权函数，用于对值进行线性组合，从而生成每个查询位置的输出。简言之，每个查询位置的输出表示为值的加权和，其中权重通过注意力机制利用查询和键计算得到。Dwin-SA 模块采用两种子机制进行特征提取。首先，Lwin-SA 机制将图像划分为较小的窗口（称为局部窗口），并在这些窗口内的图像块之间执行自注意力，以提取细粒度局部特征。其次，Gwin-SA 是一种新颖的注意力机制，它利用全局窗口与图像中任意位置的图像块进行交互，从而能够考虑局部窗口之外的信息。Gwin-SA 通过由 GWG 生成的全局键和值获取图像中的全局信息；GWG 表示整幅图像。这使得 Gwin-SA 在进行注意力决策时能够考虑全局上下文。图4 直观展示了这些注意力机制，它们也可表示如下：

$$\text{Lwin-SA}\left(Q_l^e, K_l^e, V_l^e\right) = S\left(Q_l^e K_l^{eT}/\sqrt{d} + B\right) V_l^e \quad (2)$$

$$\text{Gwin-SA}\left(Q_l^e, K_g^e, V_g^e\right) = S\left(Q_l^e K_g^{eT}/\sqrt{d} + B\right) V_g^e \quad (3)$$

其中，Q_l^e, K_l^e, V_l^e分别表示编码器局部窗口中的查询、键和值，而 K_g^e, V_g^e表示由 GWG 生成的编码器全局窗口中的键和值。对于编码器中的局部窗口和全局窗口，$Q, K, V \in \mathbb{R}^{M^2 \times d}$；$d$表示查询/键的维度，$M^2$ 表示窗口中的图像块数量，S 表示 Softmax函数。根据 [13]，[14],假设相对位置沿各轴的取值范围为 $[-M+1, M-1]$，可学习的相对位置偏置 B 从偏置矩阵 $\hat{}B \in \mathbb{R}^{(2M-1)\times(2M-1)}$中采样得到。

所提出的 Gwin-SA 机制受到 GCViT [14] 的启发，但采用了一种独特的注意力计算方法，如图 4所示。GCViT 中的全局注意力使用全局查询和局部键值，根据局部窗口令牌对全局窗口令牌进行上下文建模。即便如此，这一过程仍可能导致全局窗口令牌失去非局部性，因为全局查询仅影响注意力权重，而图像特征则由局部值提供。在这种方法中，每个查询位置的输出都通过对值进行加权平均来计算，权重则由查询和键经由注意力机制得出。相比之下，Gwin-SA 机制采用相反的方法，使用局部查询和全局键值，根据全局窗口令牌对局部窗口令牌进行上下文建模。借助全局值所提供的图像特征，局部窗口令牌能够在注意力中获得非局部性。Gwin-SA 机制的另一项优势是能够对每个较小区域进行更深入的处理，并最大限度地利用全局信息，从而实现更深入的处理。GCViT 方法与 Gwin-SA 机制的主要区别总结如下：

- GCViT 中的全局注意力：⇒ 全局查询 +局部键值 ⇒ 依据局部窗口对全局窗口进行上下文化
- DwinFormer 中的 Gwin-SA：⇒ 局部查询 +全局键值⇒ 依据全局窗口对局部窗口进行上下文化

E. 所提出的解码器架构：*Dwin-CAT*

所提出的双窗口交叉注意力变换器（Dwin-CAT）如图 2 所示。此处，每个分辨率尺度下各解码器层的输出通道数 i 表示为 $F_i = 2^{i+4}$，其中 $i \in 1$，2、3、4、5。解码器接收来自编码器的最低分辨率特征图，并逐步增强该特征图，以生成最终的深度表示。在每个层级，Dwin-CAT 通过跳跃连接和上采样整合编码器特征图中的细节。然而，包含丰富信息的编码特征有时会与包含更多语义信息的解码特征不匹配，使优化过程变得困难。为弥合这一语义鸿沟，Dwin-CAT 使用 Dwin-CA 组件整合局部和全局上下文，从而生成具有更丰富上下文信息的特征。编码特征首先经过 1×1 卷积，以使其通道与先前的解码特征相匹配，并便于进行注意力计算。Dwin-CAT 中的 Dwin-CA 组件使用局部窗口交叉注意力（Lwin-CA）和全局窗口交叉注意力（Gwin-CA）融合编码特征与解码特征。随后，融合得到的特征经过上采样层，同时提升空间维度并降低通道维度。解码器的运行方式与 Dwin-SAT 类似，但它在编码特征之间使用交叉注意力

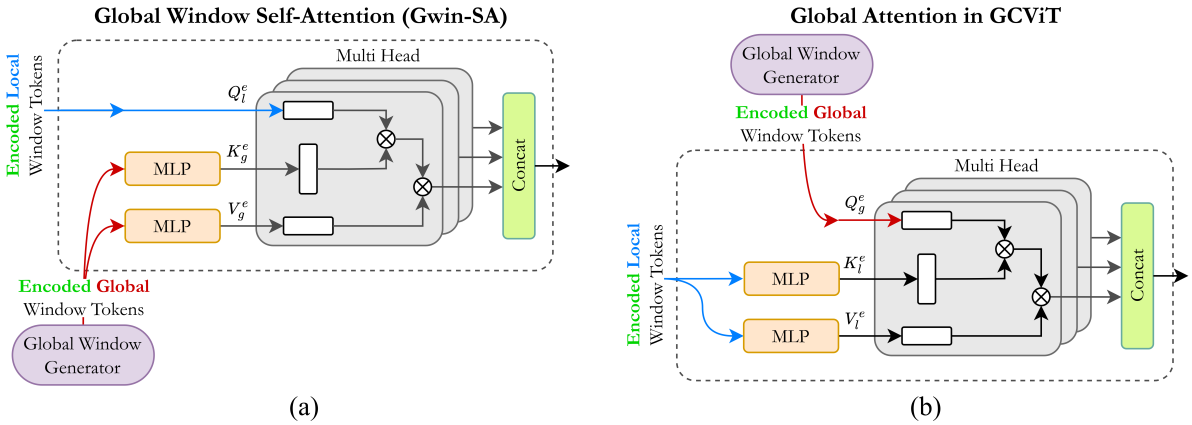

图4：所提出的 Dwin-SAT 中的全局窗口自注意力与 GCViT 中的全局注意力的比较。虽然这两种方法都对编码窗口标记应用注意力，但 Gwin-SA 使用全局键值和局部查询，而 GCViT 使用全局查询和局部键值。此处，使用全局窗口生成器来生成全局查询键值。

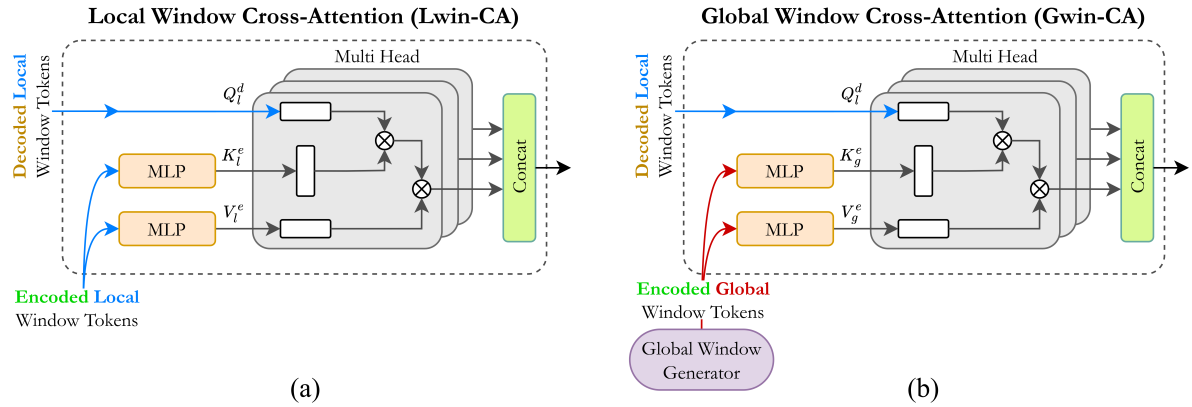

图5：所提出的 Dwin-CAT 中的局部与全局窗口交叉注意力（Lwin-CA 和 Gwin-CA）模块利用编码特征与解码特征之间的交叉注意力。两个模块都将解码特征用作局部查询，并将编码特征用作键值。不过，Lwin-CA 使用编码特征作为局部键值，而 Gwin-CA 则通过全局窗口生成器将其用作全局键值。

和解码特征之间使用交叉注意力，其中查询来自编码器，键和值来自解码器。用数学形式表示，解码器层可表示如下：

$$
\begin{aligned}
{}_1Q_l^d &= LayerNorm(D_{i-1}); E_i' = Conv_{1\times1}(E_i) \\
K_l^e &= MLP(E_i'); V_l^e = MLP(E_i') \\
L_i &= \text{Lwin-CA}({}_1Q_l^d, K_l^e, V_l^e) + D_{i-1} \\
L_i &= L_i + MLP(LayerNorm(L_i)) \\
{}_2Q_l^d &= LayerNorm(L_i); E_i'' = GWG(E_i') \\
K_g^e &= MLP(E_i''); V_g^e = MLP(E_i'') \\
D_i &= \text{Gwin-CA}({}_2Q_l^d, K_g^e, V_g^e) + L_i \\
D_i &= D_i + MLP(LayerNorm(D_i)) \\
D_i &= \text{Upsample(D}_\text{i}\text{)}
\end{aligned}
\tag{4}
$$

F. 双窗口交叉注意力（*Dwin-CA*）

双窗口交叉注意力（Dwin-CA）是一种有效的技术，能够无缝整合编码器和解码器特征，同时解决二者之间的语义差异它们。与双窗口自注意力（Dwin-SA）类似，它采用局部与全局窗口注意力机制。不过，它并非在编码特征内进行自注意力，而是在编码特征与解码器特征之间进行交叉注意力。这使 Dwin-CA 能够弥合丰富且细致的编码特征图与解码特征图之间的语义鸿沟，后者包含了关于最终深度表示的大量信息。该方法分别通过局部窗口交叉注意力（Lwin-CA）和全局窗口交叉注意力（Gwin-CA）融入局部与全局上下文。Lwin-CA 使用局部窗口在编码器与解码器之间执行交叉注意力，从而提取细粒度局部特征；Gwin-CA 则使用全局窗口在编码器与解码器之间执行交叉注意力，以捕获长程依赖。图 5展示了所提出的交叉注意力机制示意图，其数学表示如下

$$
\text{Lwin-CA}\left(Q_l^d, K_l^e, V_l^e\right) = S\left(Q_l^d K_l^{eT}/\sqrt{d} + B\right) V_l^e \tag{5}
$$

表 I：NYUv2 数据集上的结果。最佳结果以**粗体**标示，次佳结果带下划线，符号 ↑或 ↓表示值越高/越低越好

方法	绝对相对误差 ↓	均方根误差 ↓	$\log_{10}$ ↓	δ_1 ↑
Eigen 等人 [6]	0.158	0.641	-	0.769
DORN [24]	0.115	0.509	0.051	0.828
Yin 等人 [25]	0.108	0.416	0.048	0.872
BTS [26]	0.110	0.392	0.047	0.885
TransDepth [10]	0.106	0.365	0.045	0.900
DPT [9]	0.110	0.367	0.045	0.904
Adabins [22]	0.103	0.364	0.044	0.903
P3Depth [27]	0.104	0.356	0.043	0.898
SwinDepth [21]	0.100	0.354	0.042	0.909
Depthformer [11]	0.100	0.345	-	0.911
NeWCRFs [28]	0.095	0.334	0.041	0.922
PixelFormer [16]	0.090	0.322	0.039	0.929
DwinFormer（本文提出）	**0.081**	**0.280**	**0.034**	**0.951**

$$\text{Gwin-CA}\left(Q_l^d, K_g^e, V_g^e\right) = S\left(Q_l^d K_g^{eT}/\sqrt{d} + B\right) V_g^e \quad (6)$$

这些公式中使用的符号，例如 Q_l^d、 K_l^e、 V_l^e、K_g^e、 V_g^e、d、 B和 S，均与公式 2 和 3中先前定义的符号一致，但 Q_l^d除外，它表示解码器局部窗口中的查询。

G. 深度头

所提出架构中的深度头使用由 Dwin-CAT 生成的深度特征图，其尺寸为 ($\frac{H}{4}$， $\frac{W}{4}$， C)。随后，深度特征图经过一个 3×3卷积层，转换为每个像素对应的最大深度概率图。接着，对该概率图应用 Sigmoid函数，得到深度图，从而连续地估计每个像素的深度。最后，将生成的深度图上采样 $4\times$ 至与真实值相同的分辨率。

H. 损失函数

本文采用经过缩放处理的尺度不变损失（SI）。该损失的数学表达式如下：

$$\mathcal{L}_{\text{pixel}} = \alpha\sqrt{\frac{1}{T}\sum_i g_i^2 - \frac{\lambda}{T^2}\left(\sum_i g_i\right)^2} \quad (7)$$

ˆ 其中， d_i 为真实深度， d_i 为估计深度， T 表示具有有效真实值ˆ的像素数量， g_i 的计算方式为 $g_i = \log_e(d_i) - \log_e(d_i)$。实验中使用的参数 λ 和 α 分别设为 0.85 和 10。

IV. 实验

使用 KITTI [29] 和 NYU Depth V2 [30] 数据集开展了大量实验，以验证所提出新方法的有效性。通过对定量和定性标准进行全面分析，对所提出的方法进行了充分评估，并与当前最先进技术进行了比较。此外，还开展了消融实验，以清晰展示各个组件的重要性。

表 II：KITTI 数据上的结果。最佳结果以**粗体**标示，次佳结果以下划线标示，符号 ↑或 ↓ 表示值越高/越低越好

方法	绝对相对误差 ↓	均方根误差 ↓	对数$_{10}$ ↓	δ_1 ↑
Eigen 等人 [6]	0.203	6.307	0.282	0.702
DORN。 [24]	0.072	2.727	0.120	0.932
Yin 等人 [25]	0.072	3.258	0.117	0.938
BTS [26]	0.059	2.756	0.096	0.956
TransDepth [10]	0.064	2.755	0.098	0.956
Adabins [22]	0.058	2.360	0.088	0.964
DPT [9]	0.060	2.573	0.092	0.959
SwinDepth [21]	0.064	2.643	-	0.957
Depthformer [11]	0.058	2.285	-	0.967
NeWCRFs [28]	0.052	2.129	0.079	0.974
PixelFormer [16]	0.051	2.081	0.077	0.976
DwinFormer（本文提出）	**0.047**	**1.959**	**0.073**	**0.980**

A. 数据集

NYU Depth V2 是一个室内数据集，包含来自 464 个场景的 12 万组 RGB-深度对，官方训练集/测试集划分为 5 万/654 张图像，深度上限为 10 米。所提出的方法以 480×640 的分辨率进行训练。

KITTI 是一个由 61 个场景的立体图像构成的室外数据集，训练集/测试集划分为 26K/697 张图像，由 [6]，定义，深度上限为 80 米。所提方法采用 704×352 的分辨率进行训练。

B. 实现细节

本研究采用 TensorFlow 框架实现所提方法。训练期间使用 Adam 优化器 [31], ，批量大小为 8。KITTI 和 NYUV2 数据集均训练 30 个周期，初始学习率为 3×10^{-6}，随后线性逐渐增至 2.5×10^{-5}，再通过余弦调度逐渐降低。所提方法使用 8 块 NVIDIA V100 GPU 进行训练。为防止过拟合，采用了多种数据增强技术，例如随机旋转、水平翻转、亮度、对比度和色调调整、灰度化以及噪声。所提出的架构设计支持使用 GCViT 的预训练权重初始化编码器骨干网络，从而无需从头开始在 ImageNet 上训练骨干网络。编码器和解码器各层的输出通道数取决于架构的规模；对于 XXTiny、XTiny、Tiny、Small、Base 和 Large，C 分别设为 64、64、64、96、128 和 192。其余架构参数与 GCViT [14] 保持相似。

C. 评估指标

为了评估并比较所提出方法与现有方法的性能，我们采用五项指标 [6]，分别为平均相对误差（Abs Rel）：$\frac{1}{n}\sum_{i\in n}\left|d_i - \hat{d}_i\right| \hat{d}_i$；均方根误差（RMS）：$\sqrt{\frac{1}{n}\sum_{i\in n}\left\|d_i - \hat{d}_i\right\|^2}$；平均（$\log_{10}$）误差：$\frac{1}{n}\sum_{i\in n}\left|\log_{10}(d_i) - \log_{10}\left(\hat{d}_i\right)\right|$；阈值准确率（$\delta_i$）：

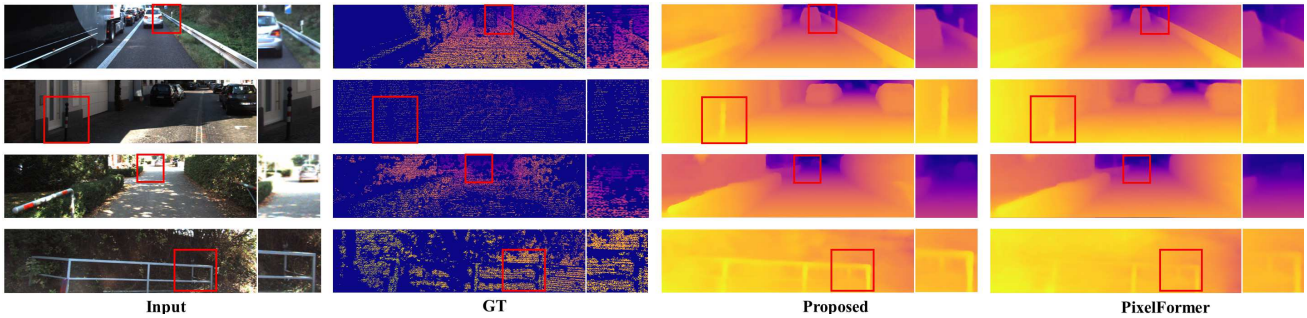

图 6：KITTI 数据集上以往最先进方法与本文所提方法的定性比较。

d_i s.t. $\max\left(\frac{d_i}{\hat{d}_i}, \frac{\hat{d}_i}{d_i}\right) = \delta < thr$ 时的百分比；平方相对差（Sq Rel）：$\frac{1}{n}\sum_{i\in n} \frac{\|d_i - \hat{d}_i\|^2}{\hat{d}_i}$。其中，$n$ 是每张深度图中的像素总数，d_i，而 $\hat{d}_i$ 分别表示第 i 个像素的真实值和预测深度。

D. 与现有方法的性能比较

1) 定量分析：表I和II分别对室外数据集 KITTI 和室内数据集 NYUv2 进行了定量分析，并将本文所提方法与现有技术进行了比较。比较结果表明，本文所提方法表现出卓越的性能，在多项指标上均大幅优于其他方法。唯一的例外是指标 δ_3，该指标的结果已高度饱和，但即便如此，本文所提方法的结果仍具有竞争力。这些结果有力地证明了本文所提方法在单目深度估计领域的出色性能，以及其生成高精度深度图的能力。

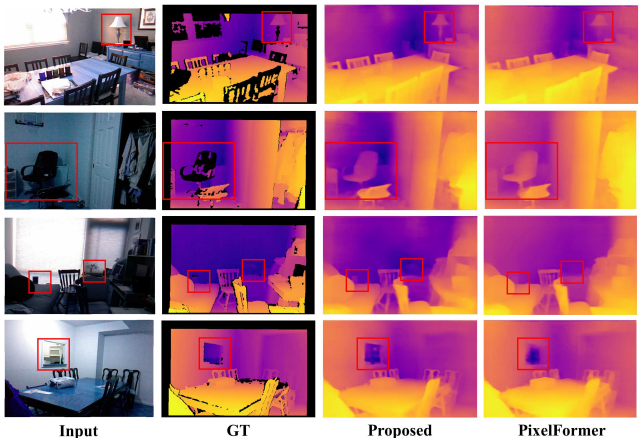

图 7：NYUv2 数据集上以往最先进方法与本文所提方法的定性比较。

2) 定性分析：图 6 和 7 分别展示了所提出的方法与现有技术在室外数据集 KITTI 和室内数据集 NYUv2 上的对比，提供了深入的见解。从图中显而易见，所提出的方法能够有效识别物体边界和锐利边缘等深度线索，同时生成一致且细致的深度图，即使对于 RGB 图像中信息缺失的物体也不例外。这凸显了所提出方法的优越性。

表 III：Dwin-SAT 在 KITTI Eigen 划分上的消融实验。结果用 ↑ 表示较好，用 ↓ 表示较差。最佳结果以**粗体**显示

方法	绝对相对误差 ↓	平方相对差 ↓	δ_1 ↑
EfficientNet-B7 [32]	0.056	0.165	0.968
SwinT-Large [13]	0.053	0.153	0.972
ConvNeXt-Large [33]	0.052	0.154	0.974
MaxViT-Large [15]	0.051	0.140	0.975
GCViT-Large [14]	0.050	0.138	0.977
Dwin-SAT（本文提出）	**0.047**	**0.136**	**0.980**

E. 消融实验

1）Dwin-SAT 的有效性：表 III 全面比较了本文提出的 Dwin-SAT 与其他成熟骨干网络的性能。在这项消融实验中，为了进行比较，所有编码器均使用 Dwin-CAT 作为解码器。显然，本文提出的 Dwin-SAT 显著优于其他方法。这归功于其独特的能力：能够有效提取兼具局部和全局上下文的特征，从而有力验证了所提出架构的有效性。

2）Dwin-CAT 的有效性：表 IV 展示了我们提出的 Dwin-CAT 模块与现有技术相比的性能。在此次消融实验中，为了进行比较，所有解码器均采用 Dwin-SAT 作为编码器。结果清楚地表明，所提出的 Dwin-CAT 模块表现最为出色，显著优于其他方法，从而验证了该模块的有效性。

V. 结论

本文提出了一种基于Transformer的设计——DwinFormer，用于端到端单目深度估计。所提出的编码器和解码器同时采用自注意力和交叉注意力机制，并利用

表 IV：Dwin-CAT 在 KITTI Eigen 划分上的消融实验。结果以 ↑ 表示更优，以 ↓ 表示更差。最佳结果以粗体标出。

方法	绝对相对误差 ↓	平方相对差 ↓	δ_1 ↑
Add-Conv	0.060	0.176	0.961
Concat-Conv	0.058	0.174	0.963
Nested-Skip [34]	0.056	0.168	0.965
SAM [16]	0.052	0.160	0.969
Dwin-CAT（提出的方法）	**0.047**	**0.136**	**0.980**

局部和全局感受野。该方法有效应对了深度图中一致性与细节之间的权衡，以及弥合编码特征与解码特征之间语义差异等挑战。因此，该方法实现了高效的特征提取和更好的深度图重建，从而生成更准确的深度图。实验结果有力地证明了所提方法优于现有方法，其在NYUv2和KITTI数据集上的显著性能提升便是明证。

参考文献

[1] B. Busam、M. Hog、S. McDonagh 和 G. Slabaugh，《Sterefo：利用立体视觉实现高效图像重新聚焦》，载于*IEEE/CVF* 计算机视觉国际会议研讨会论文集，2019 年，第 0–0 页。[2] W. Chuah、R. Tennakoon、R. Hoseinnezhad、D. Suter 和 A. Bab- Hadiashar，《面向更安全自动驾驶车辆应用的语义引导长距离立体深度估计》，*IEEE* 智能交通系统汇刊，第 23 卷，第 10 期，第 18 916–18 926 页，2022 年。[3] M. Mancini、G. Costante、P. Valigi 和 T. A. Ciarfuglia，《利用全卷积网络进行障碍物检测的快速稳健单目深度估计》，载于*2016 IEEE/RSJ* 智能机器人与系统国际会议（*IROS*）。IEEE，2016 年，第 4296–4303 页。[4] J. Ping、Y. Liu 和 D. Weng，《虚拟现实与增强现实系统中的深度感知比较》，载于*2019 IEEE* 虚拟现实与三维用户界面会议（*VR*）。IEEE，2019 年，第 1124–1125 页。[5] V. Hedau、D. Hoiem 和 D. Forsyth，《思考盒子内部：利用基于房间几何结构的外观模型与上下文》，载于欧洲计算机视觉会议。Springer，2010 年，第 224–237 页。[6] D. Eigen、C. Puhrsch 和 R. Fergus，《利用多尺度深度网络从单幅图像预测深度图》，神经信息处理系统进展，第 27 卷，2014 年。[7] S. M. H. Miangoleh、S. Dille、L. Mai、S. Paris 和 Y. Aksoy，《通过内容自适应多分辨率融合将单目深度估计模型提升至高分辨率》，载于*IEEE/CVF* 计算机视觉与模式识别会议论文集，2021 年，第 9685–9694 页。[8] L.-C. Chen、G. Papandreou、I. Kokkinos、K. Murphy 和 A. L. Yuille，《Deeplab：利用深度卷积网络、空洞卷积和全连接条件随机场进行语义图像分割》，*IEEE* 模式分析与机器智能汇刊，第 40 卷，第 4 期，第 834–848 页，2017 年。[9] R. Ranftl、A. Bochkovskiy 和 V. Koltun，《用于密集预测的视觉Transformer》，载于*IEEE/CVF* 计算机视觉国际会议论文集，2021 年，第 12 179–12 188 页。[10] G. Yang、H. Tang、M. Ding、N. Sebe 和 E. Ricci，《用于连续逐像素预测的基于Transformer的注意力网络》，载于*IEEE/CVF* 计算机视觉国际会议论文集，2021 年，第 16 269–16 279 页。[11] A. Agarwal 和 C. Arora，《Depthformer：用于单目深度估计的多尺度视觉Transformer与全局-局部信息融合》，载于*2022 IEEE* 图像处理国际会议（*ICIP*）。IEEE，2022 年，第 3873–3877 页。[12] A. Dosovitskiy、L. Beyer、A. Kolesnikov、D. Weissenborn、X. Zhai、T. Unterthiner、M. Dehghani、M. Minderer、G. Heigold、S. Gelly 等，《一幅图像胜过 16×16 个词：用于大规模图像识别的Transformer》，*arXiv* 预印本 *arXiv:2010.11929*，2020 年。

[13] Z. Liu、Y. Lin、Y. Cao、H. Hu、Y. Wei、Z. Zhang、S. Lin 和 B. Guo，《Swin Transformer：使用移位窗口的层次化视觉Transformer》，载于 IEEE/CVF 计算机视觉国际会议论文集，2021 年，第 10 012–10 022 页。[14] A. Hatamizadeh、H. Yin、J. Kautz 和 P. Molchanov，《全局上下文视觉Transformer》，arXiv 预印本 arXiv:2206.09959，2022 年。[15] Z. Tu、H. Talebi、H. Zhang、F. Yang、P. Milanfar、A. Bovik 和 Y. Li，《MaxViT：多轴视觉Transformer》，arXiv 预印本 arXiv:2204.01697，2022 年。[16] A. Agarwal 和 C. Arora，《注意力无处不在：利用跳跃注意力进行单目深度预测》，载于 IEEE/CVF 计算机视觉应用冬季会议论文集，2023 年，第 5861–5870 页。[17] T. Zhou、M. Brown、N. Snavely 和 D. G. Lowe，《从视频中无监督学习深度与自运动》，载于 IEEE 计算机视觉与模式识别会议论文集，2017 年，第 1851–1858 页。[18] I. Laina、C. Rupprecht、V. Belagiannis、F. Tombari 和 N. Navab，《利用全卷积残差网络进行更深层次的深度预测》，载于 2016 年第四届国际三维视觉会议（3DV）。IEEE，2016 年，第 239–248 页。[19] A. Ranjan、V. Jampani、L. Balles、K. Kim、D. Sun、J. Wulff 和 M. J. Black，《竞争性协作：深度、相机运动、光流和运动分割的联合无监督学习》，载于 IEEE 计算机视觉与模式识别会议论文集，2019 年，第 12 240–12 249 页。[20] Y. Zhang、M. Gong、J. Li、M. Zhang、F. Jiang 和 H. Zhao，《利用多尺度感知进行自监督单目深度估计》，IEEE 图像处理汇刊，第 31 卷，第 3251–3266 页，2022 年。[21] Z. Cheng、Y. Zhang 和 C. Tang，《Swin-Depth：利用Transformer和多尺度融合进行基于单目的深度估计》，IEEE 传感器期刊，第 21 卷，第 23 期，第 26 912–26 920 页，2021 年。[22] S. F. Bhat、I. Alhashim 和 P. Wonka，《Adabins：利用自适应分箱进行深度估计》，载于 IEEE/CVF 计算机视觉与模式识别会议论文集，2021 年，第 4009–4018 页。[23] N. Zioulis、G. Albanis、P. Drakoulis、F. Alvarez、D. Zarpalas 和 P. Daras，《Hybrid Skip：一种受生物学启发的 U-Net架构跳跃连接》，IEEE Access，2022 年。[24] H. Fu、M. Gong、C. Wang、K. Batmanghelich 和 D. Tao，《用于单目深度估计的深度序数回归网络》，载于 IEEE 计算机视觉与模式识别会议论文集，2018 年，第 2002–2011 页。[25] W. Yin、Y. Liu、C. Shen 和 Y. Yan，《对深度预测中的虚拟法线施加几何约束》，载于 IEEE/CVF 计算机视觉国际会议论文集，2019 年，第 5684–5693 页。[26] J. H. Lee、M.-K. Han、D. W. Ko 和 I. H. Suh，《从大到小：用于单目深度估计的多尺度局部平面引导》，arXiv 预印本 arXiv:1907.10326，2019 年。[27] V. Patil、C. Sakaridis、A. Liniger 和 L. Van Gool，《P3Depth：具有分段平面先验的单目深度估计》，载于 IEEE/CVF 计算机视觉与模式识别会议论文集，2022 年，第 1610–1621 页。[28] W. Yuan、X. Gu、Z. Dai、S. Zhu 和 P. Tan，《用于单目深度估计的神经窗口全连接条件随机场》，载于 IEEE/CVF 计算机视觉与模式识别会议论文集，2022 年，第 3916–3925 页。[29] A. Geiger、P. Lenz、C. Stiller 和 R. Urtasun，《视觉遇见机器人学：KITTI 数据集》，《国际机器人研究期刊》，第 32 卷，第 11 期，第 1231–1237 页，2013 年。[30] N. Silberman、D. Hoiem、P. Kohli 和 R. Fergus，《从RGB-D图像进行室内分割与支撑推断》，载于欧洲计算机视觉会议。Springer，2012 年，第 746–760 页。[31] D. P. Kingma 和 J. Ba，《Adam：一种随机优化方法》，arXiv 预印本 arXiv:1412.6980，2014 年。[32] M. Tan 和 Q. Le，《EfficientNet：重新思考卷积神经网络的模型缩放》，载于机器学习国际会议。PMLR，2019 年，第 6105–6114 页。[33] Z. Liu、H. Mao、C.-Y. Wu、C. Feichtenhofer、T. Darrell 和 S. Xie，《面向 2020 年代的卷积网络》，载于 IEEE/CVF 计算机视觉与模式识别会议论文集，2022 年，第 11 976–11 986 页。[34] Z. Zhou、M. M. R. Siddiquee、N. Tajbakhsh 和 J. Liang，《Unet++：重新设计跳跃连接以利用图像分割中的多尺度特征》，IEEE 医学成像汇刊，第 39 卷，第 6 期，第 1856–1867 页，2019 年。