arXiv:2303.03651v2 [cs.CV] 1 Aug 2023

F2BEV：基于环视鱼眼相机图像的自动驾驶鸟瞰图生成

Ekta U. Samani[1], Feng Tao[2], Harshavardhan R. Dasari[3],Sihao Ding[2],和 Ashis G. Banerjee[4]

摘要—鸟瞰图（BEV）表示对于自动驾驶感知任务极为有用。然而，由于广角镜头会引入严重畸变，从环视鱼眼相机图像生成鸟瞰图颇具挑战。我们首次着手应对这一挑战，并提出基线方法 F2BEV，用于从鱼眼图像生成离散化鸟瞰图高度图和鸟瞰图语义分割图。F2BEV 由畸变感知空间交叉注意力模块和任务特定头组成；前者采用 Transformer 式架构，从鱼眼图像特征中查询并整合空间信息，后者则用于执行具体任务。我们在合成 FB-SSEM 数据集上评估了 F2BEV 的单任务和多任务变体，结果表明，与一种基于去畸变鱼眼图像运行的最先进 BEV生成方法相比，所有变体都能生成更优的 BEV 高度图和分割图（以交并比衡量）。我们还展示了使用 F2BEV 从真实世界鱼眼图像生成离散化高度图的结果。我们的数据集已公开发布于 https: // github.com/volvo-cars/FB-SSEM-dataset

一、引言

自动驾驶环视系统通常使用多个广角鱼眼镜头相机，捕捉汽车周围的近场区域。与针孔摄像头相比，这类相机的视场更大，因此只需较少的摄像头即可完整覆盖汽车的近场区域。然而，尽管鱼眼相机在汽车行业中十分普遍，针对鱼眼相机图像的视觉感知方法却相对较少 [1], [2]。鱼眼相机会产生严重的径向畸变，若不缩小视场或引入重采样伪影，就无法校正[3]。空间变化畸变还会使靠近摄像头的物体产生显著的外观变化。因此，目标检测和跟踪等单目感知任务会变得更加复杂 [4]。在环视系统的多视角设置中，复杂性会进一步增加，因为融合各个摄像头的预测结果需要复杂的后处理步骤 [5]。通过在由多个鱼眼相机图像生成的鸟瞰图（BEV）表示中执行感知任务，可以避免这种复杂性。

本研究部分由 Volvo Car Technology USA LLC 资助。[1]E. U. Samani 任职于美国华盛顿州西雅图市华盛顿大学机械工程系，邮编 98195。ektas@uw.edu[2]在本研究期间，F. Tao 和 S. Ding 任职于美国加利福尼亚州桑尼维尔市沃尔沃汽车公司，邮编 94085。{feng.tao, sihao.ding}@volvocars.com [3]H. R. Dasari 任职于美国加利福尼亚州桑尼维尔市沃尔沃汽车公司，邮编 94085。harshavardhan.reddy.dasari@volvocars.com[4]A. G. Banerjee 任职于美国华盛顿州西雅图市华盛顿大学工业与系统工程系和机械工程系，邮编 98195。ashisb@uw.edu

鸟瞰图表示包含丰富的语义信息，例如汽车周边环境中物体的位置和尺度，因此非常适合下游现实世界任务。例如，鸟瞰图高度图和分割图可用于碰撞规避和运动规划。此外，BEV地图提供了一种融合不同模态和不同时间信息的方式。即使使用针孔摄像头，通过环视摄像头（其安装方向与地面平行并朝向外侧）生成鸟瞰图表示仍是一项具有挑战性的问题，因为从透视视图（PV）图像到鸟瞰图的变换是病态的。已有多项研究尝试开发基于学习的方法，从透视视图图像生成复杂真实场景的鸟瞰图 [6]。然而，如何从鱼眼图像生成鸟瞰图表示仍是一个尚未解决的问题 [2]。

据我们所知，本研究是首个尝试直接从环视鱼眼相机图像中获取 BEV 表示的工作。具体而言，我们的贡献如下：

- 我们提出了一种有效的基线方法F2BEV（即鱼眼图像到BEV），该方法将鱼眼投影和畸变模型纳入一种最初为透视视图图像设计的基于Transformer的BEV生成方法中。我们提出了F2BEV的单任务和多任务变体，分别用于从鱼眼图像中单独或同时生成BEV分割和离散化高度图。
- 我们提出了一个名为FB-SSEM的合成数据集，它由来自自车运动模拟序列的环视鱼眼相机图像和BEV地图组成。
- 我们报告称，与一种在去畸变鱼眼图像上运行的最先进的BEV生成方法[7] 相比，使用F2BEV能够从鱼眼图像中获得更优的BEV分割和离散化高度图生成结果。

II. 相关工作

A. 基于视觉的*BEV*生成

基于相机图像的BEV生成通常在目标检测/跟踪以及车道线/地图分割等下游感知任务的背景下进行研究。近期文献研究了从单个透视视图（即前视图）或多个透视视图生成BEV。在这两种情况下，BEV生成方法要么显式依赖于几何信息（即相机投影过程），要么利用相机几何隐式学习从透视视图到BEV的变换。

一些方法通过逆透视映射（IPM）明确地纳入投影关系 [8]

进行预处理 [9], 后处理 [10], 或特征变换 [11]。IPM 假设所有点都位于水平面上，从而解决了将图像像素映射到三维空间中各点的病态问题。已有研究探索了生成对抗网络，以减少高度信息丢失所导致的失真 [12], [13]。此外，也有人提出基于深度预测的方法，通过将二维特征图提升到三维空间来保留高度信息 [14]–[16] ，或获取伪激光雷达点 [17]。

另一类方法利用神经网络隐式地纳入相机几何信息，学习透视视图与对应鸟瞰视图之间的映射。例如， [18]–[20] 中的方法采用多层感知机来学习这种变换。近年来，研究者提出了基于Transformer的方法 [6] ，以提升神经网络中的特征提取过程，其性能优于先前的方法。具体而言，这些基于Transformer的方法借助鸟瞰视图查询与输入图像特征之间的交叉注意力来执行视图变换。对于密集感知任务（例如分割），通常采用基于密集查询的方法，其中每个查询都与鸟瞰视图空间中的一个预定义空间位置相关联。这类方法通常利用几何线索来降低内存复杂度。例如，已有方法使用相机标定矩阵，为可变形注意力模块 [7], [21]获取参考点；也有方法通过将垂直图像扫描线预先分配给鸟瞰视图射线 [22]–[24], 来简化全局交叉注意力。

B. 鱼眼图像视觉感知方法

迄今为止，由于畸变带来的挑战，针对鱼眼图像提出的视觉感知方法相对有限。即使对于目标检测和语义分割等基本感知任务，标准视觉流程（即为针孔相机图像设计的流程）也无法直接应用于鱼眼图像，必须进行适配。例如，典型卷积神经网络在成像任务中取得卓越表现所依赖的平移不变性，由于鱼眼图像存在空间变化畸变，因此并不适用。因此，基于卷积神经网络的鱼眼图像语义分割方法会采用重叠金字塔池化模块 [25]或受限可变形卷积 [26]等专用模块。对于立体匹配和深度估计等基于几何的任务，一些方法会将特定的相机投影模型纳入其框架 [27], [28] ，或将相机参数作为条件输入提供给网络 [29]。

III. 方法

给定一组环视鱼眼图像，我们的目标是获得场景的离散化鸟瞰视图高度图和语义分割鸟瞰视图地图。为实现这一目标，我们提出了一种名为 F2BEV 的基于密集查询的方法。以下小节将介绍 F2BEV 的主要组成部分，图1展示了其整体架构。

A. 鱼眼图像到*BEV*特征

首先，我们将一个带有可学习位置嵌入的查询 $q_p \in \mathbb{R}^d$关联到尺寸为 h 和 w的 BEV平面中的每个网格单元 p 。自车位于该 BEV平面的中心，每个网格单元代表一个边长为 l 米的现实世界单元。我们在时序自注意力模块中使用这些查询，如 [7], 所述，以从前一时间步的 BEV特征中获取时序信息。具体而言，利用自运动将与前一时间戳 $t-1$ 对应的 BEV特征与当前时间戳 t 的查询对齐，并将其输入时序自注意力模块。

接下来，我们使用特征提取骨干网络，从与 n_{cam}环视鱼眼相机对应的图像中提取多尺度特征图。令 $\mathcal{F}_t = \{f_t^i, i \in \{1, \ldots, n_{\text{cam}}\}\}$ 表示提取的特征图，其中f_t^i是在时间 t从第i个视图输入的鱼眼图像中获得的多尺度特征图。随后，我们使用畸变感知空间交叉注意力模块，从这些多尺度特征中获取空间信息。

我们的畸变感知空间交叉注意力模块基于 [30] 中的相机模型和 [7]中的空间交叉注意力。首先，我们计算每个位置为 p =(x, y)的查询 q_p 所对应的真实世界位置（x_r, y_r），如下所示。

$$\left[(x_r = (x - \tfrac{w}{2})l \quad y_r = (y - \tfrac{h}{2})l\right] \tag{1}$$

然后，我们考虑 n_a 预定义的高度，以获得一组与位置 p 对应的三维参考点， $\mathcal{R}_p = \{(x_r, y_r, z_j)\ j \in \{1, \ldots, n_a\}\}$。

与 [7],类似，我们在畸变感知空间交叉注意力中使用资源高效可变形注意力 [31] ，并确保每个查询 q_p 仅在所有相机视图中的相关区域内查询空间信息。具体而言，查询 q_p 仅与具有对应于 $\mathcal{R}_p$中各点的有效二维投影的相机视图进行交互。需要注意的是，有效二维投影是指其坐标对应于未被摄像头或自车遮挡的图像区域（即图1中鱼眼图像的蓝色区域）。我们使用 $\mathcal{V}_{\text{valid}}$ 表示这些视图。随后，将 $\mathcal{R}_p$ 在 $\mathcal{V}_{\text{valid}}$ 中的二维投影用作查询 q_p 的参考点，并在其周围对与 $\mathcal{V}_{\text{valid}}$ 对应的多尺度特征图进行采样。然后，按照下式计算采样特征的加权和，以获得畸变感知空间交叉注意力（DA-SCA）输出。

$$\text{DA-SCA}(q_p, \mathcal{F}_t) = \frac{1}{\mathcal{V}_{\text{valid}}} \sum_{i \in \mathcal{V}_{\text{valid}}} \sum_{j=1}^{n_a} \Delta(q_p, (u_{ij}, v_{ij}), f_t^i), \tag{2}$$

其中 Δ 表示[31], 中所述的可变形注意力，而（u_{ij}， v_{ij}）表示从 $\mathcal{R}_p$中的第 j个三维点投影到第 i个摄像头视图上的二维点的像素坐标。

我们使用来自 [30] 的相机模型来获得（u_{ij}， v_{ij}）。该模型包含一个统一投影模型，可近似

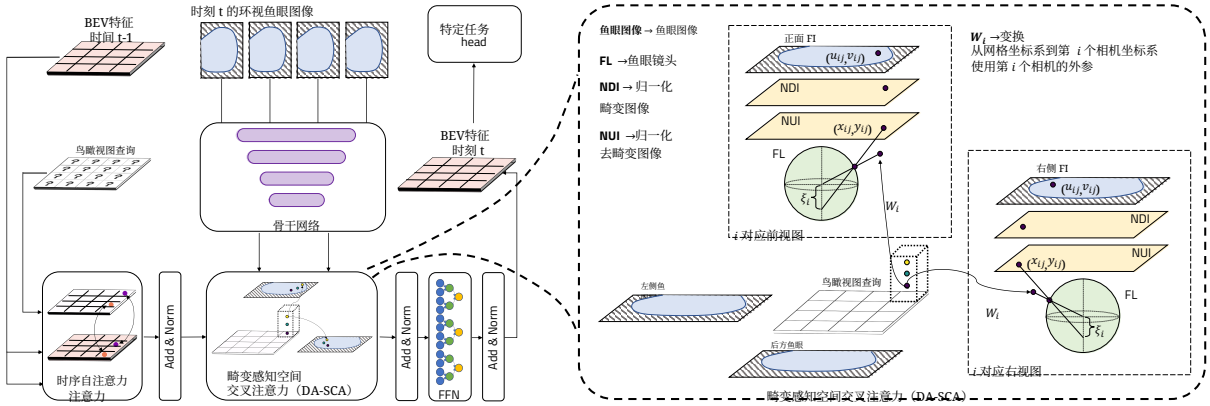


图1。 A F2BEV 从环视鱼眼相机图像生成鸟瞰图（BEV）的示意图 s

大范围的鱼眼镜头，以及一个径向-切向畸变模型。首先，我们使用相应的外参（旋转 R_i 和平移 T_i），将 $\mathcal{R}_p$中的三维参考点从BEV网格坐标系转换到第 i个相机坐标系。随后，将相机坐标系中的点投影到单位球上，并通过沿 z方向、大小为 ξ_i 的平移，将其转换到另一个参考坐标系。这里， ξ_i 表示第 i个相机的统一投影模型的镜面参数。然后，将这些平移后的点投影到归一化无畸变图像平面上。设（x_{ij}， y_{ij}）表示归一化无畸变图像平面中第j个平移点的坐标（对应于第 i个相机）。最后，我们分别使用 $[k_1, k_2]$ 和 $[k_3, k_4]$对径向畸变和切向畸变进行建模，并应用投影矩阵 K_i ，以获得（u_{ij}， v_{ij}），如下所示。

$$\begin{bmatrix} u_{ij} \\ v_{ij} \\ 1 \end{bmatrix} = K_i \begin{bmatrix} x_{ij} + \underbrace{x_{ij}\mathcal{L}(\rho_{ij})}_{\text{radial distortion}} + \underbrace{2k_3 x_{ij} y_{ij} + k_4(\rho_{ij}^2 + 2x_{ij}^2)}_{\text{tangential distortion}} \\ y_{ij} + \underbrace{y_{ij}\mathcal{L}(\rho_{ij})}_{\text{radial distortion}} + \underbrace{k_3(\rho^2 + 2y_{ij}^2) + 2k_4 x_{ij} y_{ij}}_{\text{tangential distortion}} \\ 1 \end{bmatrix} \tag{3}$$

其中 $\rho_{ij} = \sqrt{x_{ij}^2 + y_{ij}^2}$， $\mathcal{L}(\rho) = (k_1\rho^2 + k_2\rho^4)$，且 K_i 由公式（4）得到。

$$K_i = \begin{bmatrix} \gamma_{1i} & \alpha_i\gamma_{1i} & c_{1i} \\ 0 & \gamma_{2i} & c_2 i \\ 0 & 0 & 1 \end{bmatrix}, \tag{4}$$

其中， α_i 为斜切， $[\gamma_{1i}, \gamma_{2i}]$ 为焦距，$[c_{1i}, c_{2i}]$ 为第 i 个相机的主点坐标。

遵循标准Transformer架构，随后将鸟瞰视图查询输入前馈网络，以获取时间 t的BEV特征。然后将这些特征传递给任务专用头，以计算BEV地图。

B. 从*BEV*特征生成*BEV*地图

可以设计不同的任务专用头，从BEV特征中获取BEV地图。我们考虑两类头：单任务头和多任务头。

为生成BEV地图，我们考虑了两种单任务头。第一种是基于注意力的头，采用了 [32]中的掩码解码器。它使用一个可学习查询来表示每个高度类别（在分割图生成任务中则表示每个语义类别）。最终的BEV地图通过深度监督，基于多头注意力模块生成的注意力图得到。对每一层的注意力掩码进行深度监督，使多头注意力模块能够聚焦于有意义的语义区域。第二种单任务头是基于卷积的单任务头。它由多个级联上采样块组成，后接一个预测块，其中包含一个3×3 卷积层和一个激活层。每个块依次包含一次 $2\times$ 上采样操作（使用双线性插值）、一个丢弃层、一个 3×3 卷积层和一个ReLU激活层。上采样块无需增加查询数量（即降低 l），便可生成分辨率更高的BEV地图；相比之下，基于注意力的头生成的BEV地图与BEV特征具有相同的分辨率。

BEV高度图生成和分割图生成任务本质上彼此相关。此外，自动驾驶系统中的下游感知任务都能从这两类地图中获益，而同时生成它们在计算效率上也高于分别生成。因此，我们将单任务头扩展为基于注意力的多任务头和基于卷积的多任务头。在基于注意力的多任务头中，每个高度类别使用一个可学习查询，每个语义类别也使用一个可学习查询。与高度类别关联的查询会输入多头注意力模块，并使用相应的真实标注进行深度监督。同样，与语义类别关联的查询会输入多头注意力模块，并使用真实分割图进行深度监督。与单任务头一样，最终的 BEV 地图由相应的注意力图生成

另一方面，基于卷积的多任务头由两个分支组成，每个任务对应一个分支。两个分支共享级联上采样块，然后再分别进入各自的预测块。

IV. FB-SSEM 数据集

适用数据集的匮乏，是关于鱼眼图像感知方法的文献相对较少的原因之一。研究人员提出了 7自由度 [33] 和缩放增强 [25], [26] 等增强技术，用于将透视图像转换为鱼眼图像。近年来，研究人员提出了大规模合成 [34], [35]和真实世界数据集 [36]–[38] ，用于语义分割和目标检测任务。除了 Syn-WoodScape 之外，这些数据集都不包含真实值BEV高度图和分割标注。SynWoodScape 中的每个样本都代表自车的一次随机生成，并且仅关联一个先前时刻。该数据集未能捕捉与自车连续运动相关的时序信息，而已有研究表明，这类信息有助于 BEV生成 [7]。因此，我们提出了一种新的合成数据集 FB-SSEM，其中包含自车连续运动过程中生成的环视鱼眼相机图像和 BEV地图，以及相应的自车运动信息。

我们使用 Unity游戏引擎 [39] 模拟停车场环境，以构建我们的数据集。停车场中有停放的汽车/卡车、公交车、尺寸各异的电动汽车充电桩，以及高度各异的大型集装箱（位于边界处）。除自车外，停车场中的所有车辆均为静止状态。对于自车，我们使用前视广角相机来模拟其四个环视鱼眼相机 [4]。我们的数据集包含 20 段自车在白天以 0.35 m/s 的速度穿过停车场的运动序列。每段序列都代表一种不同的停车场布局，即停车场内所有车辆的摆放位置和地面纹理均有所不同。每段序列包含以 2 Hz 频率录制的 1000 个样本；每个样本包含来自四个车载鱼眼相机（即前、左、后、右相机）和 BEV相机的 RGB图像。此外，还生成了对应于全部五个视角的语义分割图，以及 BEV 的归一化高度图。此外，还获取了与每个样本对应的自车运动信息（三维旋转和平移）。我们为 BEV分割图设定了五个类别：汽车（自车和停放的汽车/卡车）、公交车、电动汽车充电桩、地面和不可行驶区域。对于本研究，我们将高度图离散为三个主要类别：高于汽车高度、低于汽车高度和与汽车同高。

V. 实验

A. 实现细节

我们在环视系统中使用四个摄像头，生成尺寸为 16.67m $\times$16.67m 的 BEV平面。我们选择 $l = 0.33$m，并将 2,500 个查询与 BEV平面关联。我们使用 PyTorch，并采用 [7] 中的时序自注意力实现，以及 [32]中的基于注意力的任务头实现。对于畸变感知空间交叉注意力，

表 I F2BEV 变体与 BEVFormer（使用未畸变图像）在离散化 BEV 高度图生成方面的性能比较

方法	按类别划分的性能（交并比）			总体性能	
	汽车下方级别	汽车处级别	汽车上方层级	Mean IoU	频率加权 IoU*
F2BEV-Attn-ST	0.9786	0.8470	0.8976	0.9078	0.8640
F2BEV-Conv-ST	0.9732	**0.8496**	0.8459	0.8896	0.8526
F2BEV-Attn-MT	**0.9787**	0.8485	**0.9044**	**0.9106**	**0.8673**
F2BEV-Conv-MT	0.9743	0.8478	0.8617	0.8946	0.8484
去畸变 →BEVFormer	0.9629	0.7656	0.8553	0.8612	0.7957

我们使用三个高度锚点来获得 R_p ，并分别将它们设置在高于地面 0m、0.25m 和 1.8m 的位置。畸变感知空间交叉注意力中使用的相机内参和外参通过 OpenCV 的 omnidir 相机标定模块获得。关于上述选择的讨论，请参阅第 V-C.1 节。

对于基于注意力的任务头，使用带有特征金字塔网络的 ResNet34 从鱼眼图像中提取特征；对于基于卷积的任务头，则使用双向特征金字塔网络。由时序自注意力、畸变感知空间交叉注意力和前馈网络组成的 Transformer 模块重复三次，以获得经过细化的 BEV 特征。同样，基于注意力的任务头为每项任务配置三个多头注意力模块。基于卷积的任务头由三个上采样块副本组成。

我们使用配备 NVIDIA A10G Tensor Core 图形处理器的 Ubuntu 20.04 AWS EC2 实例进行训练和测试。我们评估单任务和多任务 F2BEV 变体的性能，并将其与最先进的 BEVFormer [7]进行比较。由于 BEVFormer 需要以针孔相机图像作为输入，因此在将鱼眼图像输入 BEVFormer 之前，我们先对其进行去畸变。去畸变所用的焦距根据 OpenCV 对透视校正的建议进行选择 [40]。我们按照 (70-15-15)% 的训练集、验证集和测试集划分方式对每个序列进行划分，并先使用类别交叉熵损失训练模型，再使用多类别焦点损失进行微调。我们始终采用 BEVFormer 使用的随机采样策略，以增加自运动的多样性。测试时，我们依次为每个样本生成预测；也就是说，我们保存前一时间戳的 BEV特征，并将其用于当前时间戳的预测。

B. 结果

1) 模拟数据集上的结果： 表 I 和表 II 分别展示了离散高度图和分割图生成的性能。无论任务特定头部的类型如何，F2BEV 在这两项任务上的表现均优于 BEVFormer（其输入为未畸变图像）。

从表 I 可以看出，在离散化高度图生成任务中，基于注意力机制的单任务和多任务变体[1] （分别为 F2BEV-Attn-ST 和 F2BEV-Attn-MT）的总体性能略优于

[1]由于基于注意力机制的变体生成的预测分辨率较低（但覆盖相同的物理区域），我们对真实标注进行缩放以计算交并比。

表 II F2BEV各变体与BEVFORMER（使用未畸变图像）在鸟瞰视角分割图生成方面的性能比较

方法	按类别划分的性能（交并比）					总体性能	
	地面	电动汽车充电桩	不可行驶区域	Bus	Car	平均交并比	频率加权交并比*
F2BEV-Attn-ST	**0.9779**	**0.6971**	0.7293	0.8482	0.8987	**0.8303**	0.8587
F2BEV-Conv-ST	0.9734	0.6399	0.6851	0.8876	0.8848	0.8142	**0.8624**
F2BEV-Attn-MT	0.9776	0.6244	0.7597	0.8531	**0.9011**	0.8232	0.8591
F2BEV-Conv-MT	0.9743	0.6308	0.7105	**0.9066**	0.8847	0.8214	0.8560
去畸变 → BEVFormer	0.9679	0.6385	**0.8222**	0.8015	0.8408	0.8142	0.8149

基于卷积的单任务和多任务变体（分别为 F2BEV-Conv-ST 和 F2BEV-Conv-MT）。此外，F2BEV-Attn-MT 的总体性能最佳，按频率加权 IoU 指标[2]计算，较 BEV-Former 提升约 9%。还值得注意的是，两个多任务变体在生成鸟瞰图分割图的同时，总体性能与对应的单任务变体相当。我们将在本小节后文讨论多任务变体的分割性能。

图 2 展示了一些离散化高度图生成结果。请注意，为了便于观察，我们在所有图中都统一纳入了使用 F2BEV-Conv-ST 变体获得的相对高分辨率结果。第一行和第二行展示了 F2BEV 整合来自不同摄像头的信息的能力。具体而言，即使物体在一个摄像头视图中完全被遮挡、在另一个视图中仅部分可见（例如第一行中红色汽车后方的电动汽车充电桩），F2BEV 仍能在鸟瞰视图中正确放置这些物体。我们还看到，对于在一个视图中部分可见、在另一个视图中完全可见的长形物体（例如第二行中的公交车），模型也能正确放置。此外，该方法还展现出对物体与自车相对距离的隐式理解：在生成鸟瞰视图时，它会忽略鱼眼图像中所有相对较远的物体。结果还进一步体现了我们提出的畸变感知方法在鸟瞰视图生成过程中推断并保留高度信息的优势。例如，第一行和第二行展示了对因靠近自车而发生严重畸变的车辆进行正确高度分类的结果。

在具有五个语义类别的分割图生成任务中，我们从表 II 注意到，F2BEV 的单任务变体总体性能略优于其多任务变体。此外，F2BEV-Conv-ST 的总体性能最佳，在频率加权交并比指标上比 BEVFormer 提高了 5.8%。与离散化高度图生成的情况一样，所有 F2BEV 变体的表现都优于处理未畸变图像的 BEVFormer。我们还看到，生成的鸟瞰图分割图成功融合了来自多个摄像头的信息。例如，图 3 显示了电动汽车充电桩被正确放置在第二排白色汽车后方、第三排公交车后方。关于逐类别分割性能的进一步讨论，特别是电动汽车充电桩和不可行驶区域类别，请参阅第 V-C.2 节。

2) 真实世界图像上的结果：我们还在真实世界鱼眼图像上测试了 F2BEV 的性能。这些图像由安装在沃尔沃 XC90 汽车上的四台视场为 210° 的摄像头采集。我们使用自车上的 IMU 记录自车运动信息，并进行实时噪声滤波。由于无法获得真实标注，我们仅报告定性性能评估。图 4 展示了几张真实世界图像样本。

图 2 中的模拟样本与图 4 中的真实世界样本之间存在相当大的差异。为缩小这一差异，我们使用经过风格迁移、看起来像真实世界数据的模拟数据，训练 F2BEV-Conv-ST 模型以生成离散化高度图。我们使用真实世界图像样本作为风格图像，对模拟 FB-SSEM 数据集中的图像进行神经风格迁移 [41] 。此外，我们注意到，模拟鱼眼相机与安装在真实车辆上的相机在图像有效区域方面有所不同（即未被相机或自车覆盖的区域）。训练时，我们通过预先设定训练数据中每台相机的有效区域，使其与对应真实世界相机的有效区域一致，来处理这一差异。随后，我们使用训练好的模型为真实世界数据预测离散化鸟瞰图高度图。

图4展示了若干样本图像[3]生成的高度图。第一行和第二行展示了汽车在街道上行驶的实例。我们可以看到，送货卡车和路边停放汽车均被成功地放置在相应的高度图中。第三行展示了汽车在停满车辆的停车场中行驶的实例。从鱼眼图像中可以注意到，自车与其左侧停放汽车的距离比与其右侧大多数汽车的距离更近，这一点也反映在高度图中。补充材料中包含了该方法在真实世界图像和模拟图像上的性能视频。

C. 讨论

1) 设计选择：在所有实验中，我们使用与BEV平面相关联的2,500个查询。这一选择直接

[2]频率加权交并比的计算忽略了背景类别（即低于车辆高度类别和地面类别）

[3]我们仅报告高度图生成性能，因为训练和测试期间所考虑的输出类别相同，这与分割图生成不同。

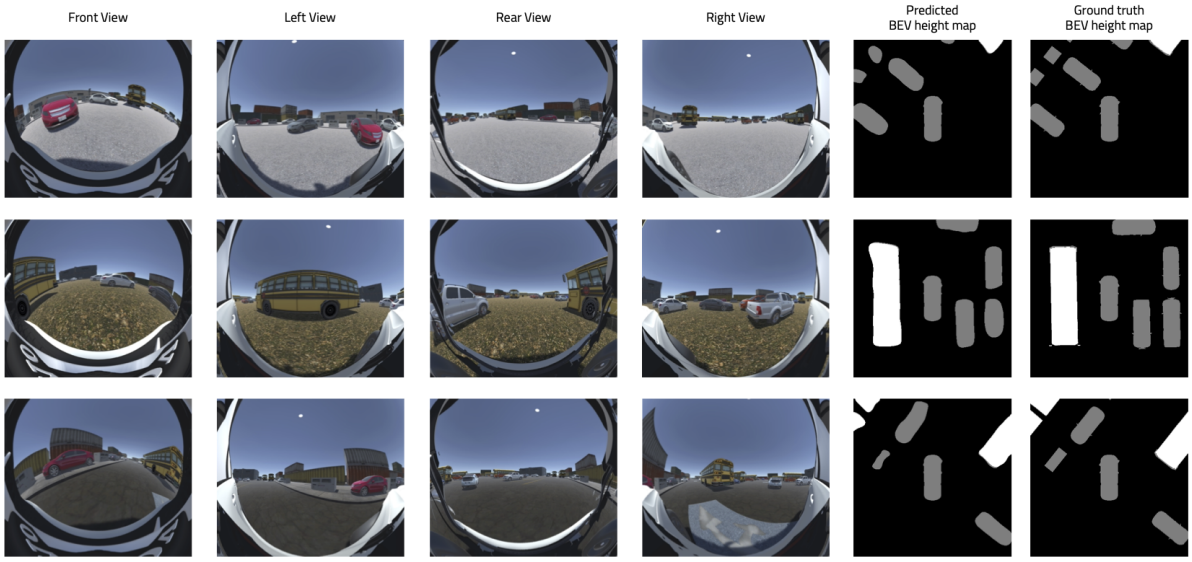


图 2. 使用 F2BEV-Conv-ST 从模拟鱼眼图像中获得的离散化鸟瞰图高度图示例。在这些 BEV 中，黑色表示“低于汽车高度”类别，灰色表示“与汽车同高”类别，白色表示“高于汽车高度”类别。

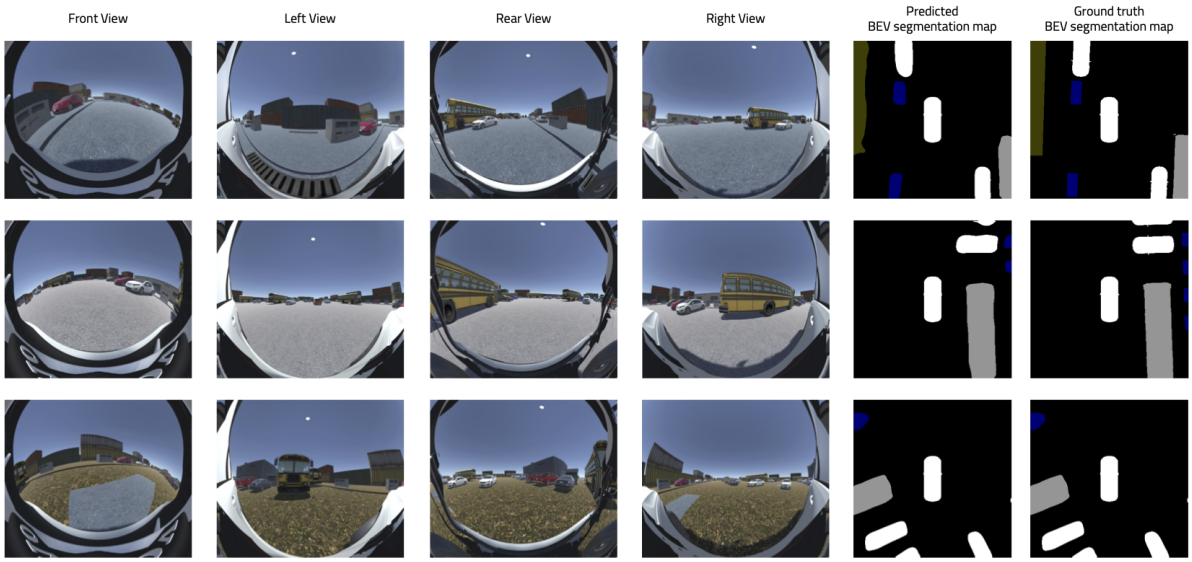


图 3. 使用 F2BEV-Conv-ST 从模拟鱼眼图像中获得的 BEV分割图预测示例。在这些 BEV 中，黑色表示地面类别，白色表示汽车类别，灰色表示公交车类别，蓝色表示电动汽车充电桩类别，橄榄绿表示不可行驶区域。

源于使用单个图形处理器进行训练时的内存限制。使用多图形处理器训练时，可以采用更多查询。如果降低 l，增加查询数量有助于利用基于注意力的头生成分辨率更高的特征图。它还有助于分割在BEV地图中显得狭窄的物体，例如路灯和车道线。另一种做法是在不改变 l 的情况下增加查询数量，从而为物理范围更大的区域生成BEV。在我们的实验中，每个查询 q_p 都关联三个高度锚点，以获得$\mathcal{R}_p$。这些高度锚点根据自车的大致离地间隙和高度设定。也可以考虑使用更多高度锚点，以获得包含三个以上高度类别的离散化高度图。

我们的畸变感知空间交叉注意力模块使用统一投影来近似鱼眼镜头，并基于径向切向畸变模型进行畸变校正，如 [30], 中所示，以从 $\mathcal{R}_p$ 获得二维参考点。我们的结果验证了将畸变纳入网络以生成 BEV地图，而不是对图像本身进行去畸变的可行性。我们认为，基于其他投影模型[37], [42]–[44] 的畸变感知注意力模块，也可以纳入为透视视图图像设计的标准 BEV生成方法中，从而比基于图像去畸变的方法取得更好的性能。

2）电动汽车充电桩和不可行驶区域类别的分割：在图 3 的第二行中，我们注意到预测的鸟瞰图中缺少两个电动汽车充电桩。这

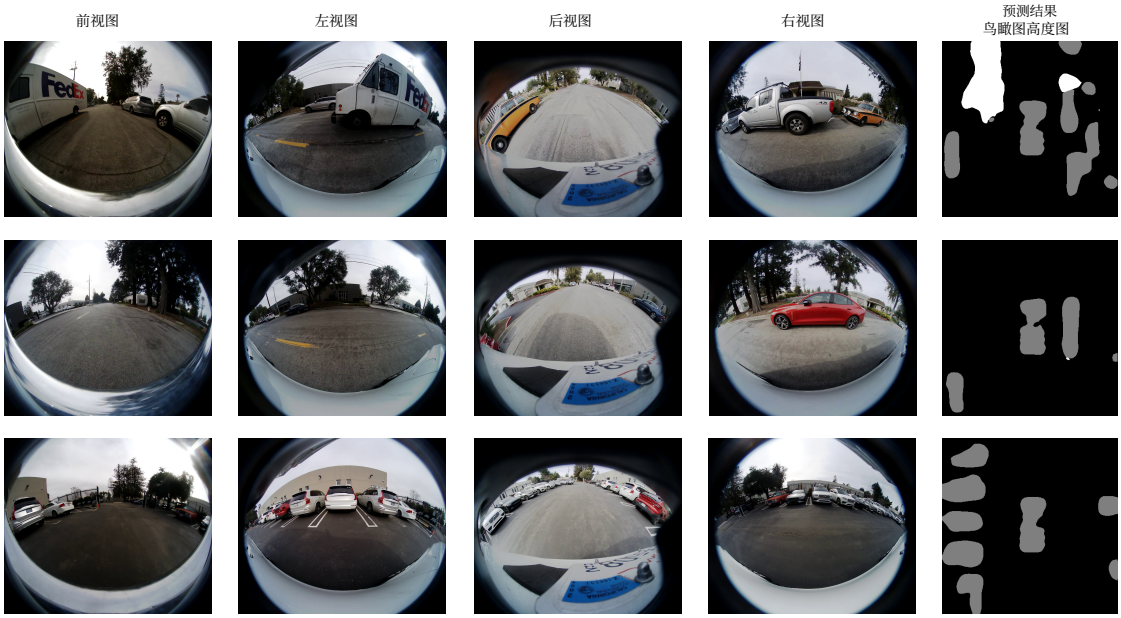


图 4. 使用 F2BEV-Conv-ST 从真实世界鱼眼图像中生成的离散化鸟瞰图高度图示例。在鸟瞰视图中，黑色表示“低于车身高度”类别，灰色表示“与车身等高”类别，白色表示“高于车身高度”类别。

一观察结果符合我们的预期；由于公交车完全遮挡了这两个电动汽车充电桩，它们并未出现在对应的相机视图中。真实鸟瞰图无法捕捉此类遮挡，因为它是由位于地面上方的透视相机生成的。我们当前基于交并比的定量评估未考虑此类情况，这是电动汽车充电桩类别分割性能相对较低的原因之一。

对于不可行驶区域，BEVFormer 的交并比高于 F2BEV。我们认为，这一趋势是由于与未畸变图像特征相比，鱼眼图像特征的采样相对稀疏。具体而言，不可行驶区域的某些部分远高于最高的高度锚点（1.8 米，约为汽车高度）。因此，对应不可行驶区域的图像特征采样较为稀疏。由于不可行驶区域在鱼眼图像中所占像素少于在未畸变图像中的占比（这是视场不同所致），采样不足对 F2BEV 的影响大于对 BEVFormer 的影响。我们认为，可以根据环境中的语义类别设置更具代表性的高度锚点，以解决这一问题。

3）畸变感知方法与输入图像去畸变的比较：我们在表 I 和表 II 中证明，即使在我们数据集相对简单的停车场环境中，在 F2BEV 中使用畸变感知空间交叉注意力，也比在 BEV-Former 中使用未畸变图像取得更好的性能。此外，将鱼眼图像去畸变为透视图像会因视场缩小而造成信息损失。当生成覆盖更大区域的 BEV地图时，这种信息损失会导致性能下降。采用更复杂的去畸变方法，从单张鱼眼图像中获取多张未畸变图像，也会使通过交叉注意力模块将信息整合到单张 BEV 中更具挑战性。

VI. 结论

在本文中，我们提出了 F2BEV：一种首创的方法，可直接根据环视鱼眼相机图像生成 BEV地图。该方法将鱼眼投影和畸变模型融入空间交叉注意力模块，以查询鱼眼图像特征中的空间信息，并将其整合为鸟瞰图表示。借鉴一种针对针孔相机图像的最先进BEV生成方法，我们将畸变感知空间交叉注意力模块置于Transformer式架构中，并结合时序自注意力模块和前馈网络，以获得BEV特征。我们采用单任务头和多任务头，从生成的BEV特征中生成离散化高度图和语义分割图。在合成FB-SSEM数据集上，无论选择哪种任务头，F2BEV在交并比方面均优于一种 [7] 基于去畸变鱼眼图像运行的最先进BEV生成方法。我们还展示了F2BEV根据从真实车辆获取的真实世界鱼眼图像生成BEV地图的性能。此外，我们公开了FB-SSEM 数据集，以促进该领域的进一步研究，例如开发独立于相机标定参数 [45] 的畸变感知模块，从而为自动驾驶任务获取不同类型的BEV地图（例如连续高度图、类图像RGB图）。

致谢

我们感谢 Zhipeng Liu、John Nousiainen、Chihiro Suga 和 Jeremy White 为构建用于生成 FB-SSEM 数据集的模拟环境所作的贡献。

参考文献

[1] J. Horgan、C. Hughes、J. McDonald 和 S. Yogamani，“基于视觉的驾驶辅助系统：综述、分类与进展”，载于 IEEE 国际智能交通系统会议，2015年，第 2032–2039 页。[2] V. R. Kumar、C. Eising、C. Witt 和 S. Yogamani，“自动驾驶环视鱼眼相机感知：概述、综述与挑战”，《IEEE 智能交通系统汇刊》，2023年。[3] V. R. Kumar、S. Yogamani、M. Bach、C. Witt、S. Milz 和 P. M¨ader，“UnRectDepthNet：采用通用框架处理常见相机畸变模型的自监督单目深度估计”，载于 IEEE 国际智能机器人与系统会议，2020年，第 8177–8183 页。[4] H. Rashed、E. Mohamed、G. Sistu、V. R. Kumar、C. Eising、A. El-Sallab 和 S. Yogamani，“面向自动驾驶的鱼眼相机通用目标检测：数据集、表示与基线”，载于 IEEE 冬季计算机视觉应用会议，2021年，第 2272–2280 页。[5] T. Fischer、Y.-H. Yang、S. Kumar、M. Sun 和 F. Yu，“CC-3DT：通过跨相机融合实现全景三维目标跟踪”，arXiv 预印本 arXiv:2212.01247，2022年。[6] Y. Ma、T. Wang、X. Bai、H. Yang、Y. Hou、Y. Wang、Y. Qiao、R. Yang、D. Manocha 和 X. Zhu，“以视觉为中心的BEV感知：综述”，arXiv 预印本 arXiv:2208.02797，2022年。[7] Z. Li、W. Wang、H. Li、E. Xie、C. Sima、T. Lu、Y. Qiao 和 J. Dai，“BEVFormer：通过时空变换器从多摄像头图像中学习鸟瞰视角表示”，载于欧洲计算机视觉会议，2022年，第 1–18 页。[8] H. A. Mallot、H. H. B¨ulthoff、J. Little 和 S. Bohrer，“逆透视映射简化了光流计算和障碍物检测”，《生物控制论》，第 64 卷，第 3 期，第 177–185 页，1991年。[9] Y. Kim 和 D. Kum，“基于深度学习，通过逆透视映射图像估计车辆位置和姿态”，载于 IEEE 智能车辆研讨会，2019年，第 317–323 页。[10] S. Sengupta、P. Sturgess、L. Ladick`y 和 P. H. Torr，“根据街景图像自动生成稠密视觉语义映射”，载于 IEEE 国际智能机器人与系统会议，2012年，第 857–862 页。[11] L. Reiher、B. Lampe 和 L. Eckstein，“一种仿真到现实的深度学习方法：将多台车载摄像头拍摄的图像变换为鸟瞰视角下的语义分割图像”，载于 IEEE 国际智能交通系统会议，2020年，第 1–7 页。[12] X. Zhu 、 Z. Yin 、 J. Shi 、 H. Li 和 D. Lin，“ 生成对抗式正视图到鸟瞰图合成”，载于 IEEE 国际三维视觉会议，2018年，第 454–463 页。[13] T. Bruls、H. Porav、L. Kunze 和 P. Newman，“正确（倾斜）的视角：利用增强型逆透视映射提升对道路场景的理解”，载于 IEEE 智能车辆研讨会，2019年，第 302–309 页。[14] J. Philion 和 S. Fidler，“Lift, Splat, Shoot：通过隐式反投影至三维，对任意相机配置拍摄的图像进行编码”，载于欧洲计算机视觉会议，2020年，第 194–210 页。[15] A. Hu、Z. Murez、N. Mohan、S. Dudas、J. Hawke、V. Badrinarayanan、R. Cipolla 和 A. Kendall，“FIERY：利用环视单目相机，从鸟瞰视角预测未来实例”，载于 IEEE 国际计算机视觉会议，2021年，第 15 273–15 282 页。[16] J. Huang、G. Huang、Z. Zhu 和 D. Du，“BEVDet：鸟瞰视角下的高性能多摄像头三维目标检测”，arXiv 预印本 arXiv:2112.11790，2021年。[17] Y. You、Y. Wang、W.-L. Chao、D. Garg、G. Pleiss、B. Hariharan、M. Campbell 和 K. Q. Weinberger，“Pseudo-LiDAR++：用于自动驾驶三维目标检测的精准深度”，载于国际学习表征会议，2020年。[18] C. Lu、M. J. G. van de Molengraft 和 G. Dubbelman，“基于卷积变分编码器—解码器网络的单目语义占据栅格映射”，《IEEE 机器人与自动化快报》，第 4 卷，第 2 期，第 445–452 页，2019年。[19] B. Pan、J. Sun、H. Y. T. Leung、A. Andonian 和 B. Zhou，“用于感知周边环境的跨视角语义分割”，《IEEE 机器人与自动化快报》，第 5 卷，第 3 期，第 4867–4873 页，2020年。[20] T. Roddick 和 R. Cipolla，“利用金字塔占据网络从图像预测语义地图表示”，载于 IEEE 计算机视觉与模式识别会议，2020年，第 11 138–11 147 页。[21] J. Lu、Z. Zhou、X. Zhu、H. Xu 和 L. Zhang，“将自车三维表示学习为光线追踪”，载于欧洲计算机视觉会议，2022年，第 129–144。

[22] A. Saha、O. Mendez、C. Russell 和 R. Bowden，“将图像转换为地图：无需 IPM 即可实现端到端 BEV 感知”，载于 IEEE 机器人与自动化国际会议，2022年，第 9200–9206 页。[23] S. Gong、X. Ye、X. Tan、J. Wang、E. Ding、Y. Zhou 和 X. Bai，“GitNet：基于几何先验的鸟瞰视图分割变换”，载于欧洲计算机视觉会议，2022年，第 396–411 页。[24] Y. Jiang、L. Zhang、Z. Miao、X. Zhu、J. Gao、W. Hu 和 Y.-G. Jiang，“PolarFormer：利用极坐标变换器进行多相机三维目标检测”，arXiv 预印本 arXiv:2206.15398，2022年。[25] L. Deng、M. Yang、Y. Qian、C. Wang 和 B. Wang，“使用鱼眼相机对城市交通场景进行基于卷积神经网络的语义分割”，载于 IEEE智能车辆研讨会，2017年，第 231–236 页。[26] L. Deng、M. Yang、H. Li、T. Li、B. Hu 和 C. Wang，“基于受限可变形卷积的使用环视摄像头进行道路场景语义分割”，IEEE智能交通系统汇刊，第 21 卷，第 10 期，第 4350–4362 页，2019年。[27] C. H¨ane、L. Heng、G. H. Lee、A. Sizov 和 M. Pollefeys，“使用平面扫描立体视觉在鱼眼图像上进行实时直接稠密匹配”，载于 IEEE三维视觉国际会议，第 1 卷，2014年，第 57–64 页。[28] C. Won、J. Ryu 和 J. Lim，“OmniMVS：面向全向立体匹配的端到端学习”，载于 IEEE计算机视觉国际会议，2019年，第 8987–8996 页。[29] V. R. Kumar、M. Klingner、S. Yogamani、M. Bach、S. Milz、T. Fin- gscheidt 和 P. M¨ader，“SVDistNet：在环视鱼眼相机上进行自监督近场距离估计”，IEEE智能交通系统汇刊，第 23 卷，第 8 期，第 10 252–10 261 页，2021年。[30] C. Mei 和 P. Rives，“从平面网格进行单视点全向相机标定”，载于 IEEE机器人与自动化国际会议，2007年，第 3945–3950 页。[31] X. Zhu、W. Su、L. Lu、B. Li、X. Wang 和 J. Dai，“可变形DETR：用于端到端目标检测的可变形变换器”，arXiv 预印本 arXiv:2010.04159，2020年。[32] Z. Li、W. Wang、E. Xie、Z. Yu、A. Anandkumar、J. M. Alvarez、P. Luo 和 T. Lu，“全景SegFormer：深入研究使用变换器的全景分割”，载于 IEEE计算机视觉与模式识别会议，2022年，第 1280–1289 页。[33] Y. Ye、K. Yang、K. Xiang、J. Wang 和 K. Wang，“面向鱼眼城市驾驶图像的通用语义分割”，载于 IEEE系统、人和控制论国际会议，2020年，第 648–655 页。[34] A. R. Sekkat、Y. Dupuis、P. Vasseur 和 P. Honeine，“OmniScape数据集”，载于 2020年 IEEE机器人与自动化国际会议（ICRA），2020年，第 1603–1608 页。[35] A. R. Sekkat、Y. Dupuis、V. R. Kumar、H. Rashed、S. Yogamani、P. Vasseur 和 P. Honeine，“SynWoodScape：用于自动驾驶的合成环视鱼眼相机数据集”，IEEE机器人与自动化快报，第 7 卷，第 3 期，第 8502–8509 页，2022年。[36] W. Maddern、G. Pascoe、C. Linegar 和 P. Newman，“1 年，1000 千米：牛津RobotCar数据集”，《国际机器人研究杂志》，第 36 卷，第 1 期，第 3–15 页，2017年。[37] S. Yogamani、C. Hughes、J. Horgan、G. Sistu、P. Varley、D. O’Dea、M. Uric´ar、S. Milz、M. Simon、K. Amende 等，“WoodScape：用于自动驾驶的多任务、多相机鱼眼数据集”，载于 IEEE计算机视觉国际会议，2019年，第 9308–9318 页。[38] Y. Liao、J. Xie 和 A. Geiger，“KITTI-360：用于二维和三维城市场景理解的新型数据集与基准”，IEEE模式分析与机器智能汇刊，2022年。[39] “Unity。” [在线]。可访问：https://unity.com/ [40] “全向相机标定。” [在线]。可访问：https: //docs.opencv.org/4.x/dd/d12/tutorial omnidir calib main.html [41] L. A. Gatys、A. S. Ecker 和 M. Bethge，“艺术风格的神经算法”，arXiv 预印本 arXiv:1508.06576，2015年。[42] D. Caruso、J. Engel 和 D. Cremers，“面向全向相机的大规模直接同步定位与建图”，载于 IEEE智能机器人与系统国际会议，2015年，第 141–148 页。[43] B. Khomutenko、G. Garcia 和 P. Martinet，“增强型统一相机模型”，IEEE机器人与自动化快报，第 1 卷，144，2015 1，137第。页 – 。[44] V. Usenko、N. Demmel 和 D. Cremers，“双球面相机模型”，载于 IEEE三维视觉国际会议，2018年，第 552–560 页。[45] L. Peng、Z. Chen、Z. Fu、P. Liang 和 E. Cheng，“BEVSegFormer：从任意相机配置进行鸟瞰视图语义分割”，载于 IEEE冬季计算机视觉应用会议，2023年，第 5935–5943 页。