基于流的抗噪微小目标定位

Huixin Sun[1],Linlin Yang[2],Ronyu Chen[3],
Kerui Gu[3], Baochang Zhang[1],Angela Yao[3],
Xianbin Cao[ll]北京航空航天大学，中国[2]中国传媒大学，中国[3]新加坡国立大学，新加坡

arXiv:2601.00617v1 [cs.CV] 2 Jan 2026

摘要

尽管通用目标检测取得了显著进展，但与常规尺度目标相比，微小目标的检测性能仍存在持续差距。我们证明，微小目标对标注噪声高度敏感，优化严格的定位目标可能导致对噪声过拟合。为此，我们提出基于流的微小目标定位（TOLF），这是一种抗噪定位框架，利用归一化流进行灵活的误差建模和不确定性引导优化。我们的方法通过基于流的误差建模捕捉复杂的非高斯预测分布，从而在噪声监督下实现鲁棒学习。具备不确定性感知能力的梯度调制机制进一步抑制模型从高不确定性、易受噪声影响的样本中学习，从而减轻过拟合并稳定训练。在三个数据集上开展的大量实验验证了我们方法的有效性。特别是，TOLF在AI-TOD数据集上使DINO基线的平均精度提升了1.2%。

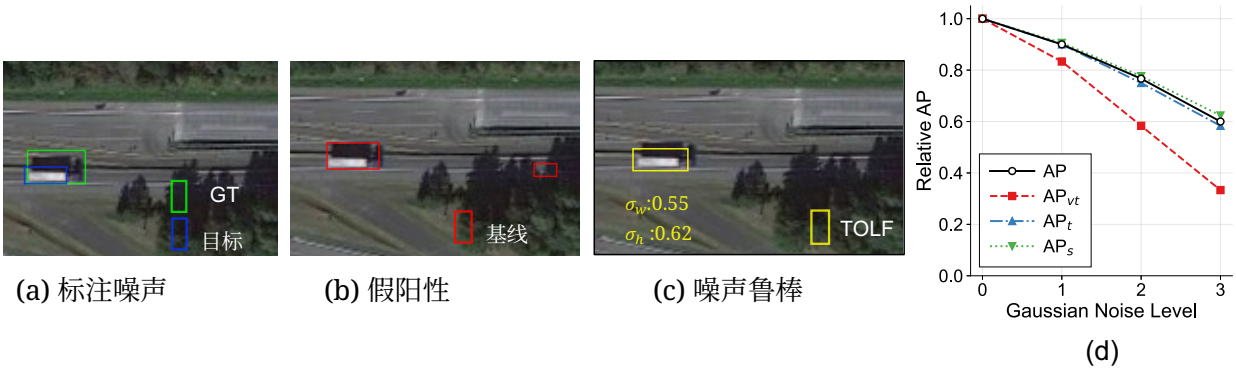
关键词：微小　目标检测、噪声鲁棒性、归一化　　流

1. 引言

深度神经网络（DNNs） [1] 的近期进展显著推动了目标检测领域的发展 [2]。尽管取得了这些进展，微小目标检测（TOD）仍然是一项极具挑战性的问题 [3]。微小目标的像素输入极为有限（少于 16×16 像素 [4]），与通用目标检测相比，它们在通用检测器上的性能会严重下降 [2]。例如，作为先进的基于查询的检测器，DINO [5], 在 AI-TOD 中对中等目标的平均精度为 37.6%，而对微小目标的平均精度仅为 9.9% [4]。如此低的性能远不能满足安全关键型现实应用的需求，例如交通管理 [6],、驾驶辅助 [7], 和异常检测 [8]。

微小目标本身有限的像素输入是 TOD 面临的主要挑战，阻碍了充分提取判别性前景特征 [9]。在杂乱环境中，这一挑战会进一步加剧 [9],：普遍存在的遮挡、复杂的背景噪声 [10], 以及极低的信噪比，会导致特征表示空间中的歧义 [11]。因此，通用检测器可能产生特征偏差，倾向于将前景与和其相似的背景区域区分开来，从而导致 TOD 中出现漏检和假阳性。近期研究通过上采样或专用架构提升特征分辨率 [12], ，利用上下文信息弥补像素输入有限的问题 [13], ，并采用辅助自重建模块 [9] 来改进目标判别。

在这项工作中，我们揭示了微小目标容易受到标注噪声影响，并面临过拟合风险。由于分辨率有限且存在视觉模糊性，微小目标的人工标注往往存在标注不一致，包括漏标目标、边界框不准确或类别错误 [4]。为了量化真实世界微小目标数据集中标注噪声的普遍程度，我们人工检查了图 2 中 AI-TOD 测试集里随机选取的 10 张图像中的 532 个边界框。结果显示，近 34.2% 的标注存在噪声。微小目标的交并比敏感性会加剧这些误差，即使轻微偏差也可能大幅改变定位质量。对于一个 10×10 目标，仅 2 像素的偏移就会导致交并比下降超过 20%；而同样的误差对一个 100×100 目标仅会造成 5% 的性能下降。在这种情况下，针对严格定位标准（例如，交并比为 1.0 ）进行优化，可能会导致模型对标注噪声过拟合，而非学到有效的回归。如图 1(b) 所示，这种过拟合



(a) 标注噪声　(b) 假阳性　(c) 噪声鲁棒　(d)

图 1：过拟合标签噪声导致的异常预测。(a) 不准确的真实标注覆盖了背景阴影。(b) 过拟合会在背景区域产生假阳性，这些区域与噪声标注相似。(c) TOLF 在不确定位置的置信度较低，且定位更准确。(d) 噪声敏感性分析：向训练标注注入高斯噪声，并测量不同目标尺度下的检测性能。结果表明，微小目标的性能退化最严重。通过训练 1× FCOS 检测器 [14]来评估检测性能。模型在 AI-TOD [4] trainval 上训练，并在 AI-TOD 测试集上验证。

可能导致背景区域中的假阳性增加。此外，我们还进行了噪声敏感性分析，以量化训练阶段的标签噪声对不同目标尺度的影响。我们向训练边界框的中心注入标准差为$\sigma \in \{1.0、2.0、3.0\}$ 像素的高斯噪声，并在干净数据上进行评估。如图 1(d) 所示，随着噪声水平升高，所有尺度目标的性能均有所下降，其中微小目标的性能下降幅度最大。当噪声为 $\sigma = 3.0$ 像素时，整体平均精度下降 40.0%，而极微小和微小目标的平均精度下降 66.7%。这种对标注噪声的高度敏感性凸显了微小目标检测中采用鲁棒定位目标函数的重要性。

鉴于上述分析，我们提出微小目标定位流（TOLF），这是一种利用归一化流进行灵活预测分布建模的鲁棒定位框架，能够处理不确定性和标签噪声。不同于受高斯假设或固定先验限制的传统不确定性方法，TOLF 使用可逆归一化流，显式学习预测与噪声标注之间的误差分布。因此，TOLF 能够捕捉复杂的噪声结构，包括厚尾、偏度和多模态性。此外，TOLF 的损失函数能够感知不确定性。通过基于不确定性的加权自适应降低噪声样本的权重，它抑制了对标注误差的过拟合，并在严重噪声条件下保持稳定训练，从而避免离群值主导损失景观。TOLF 将灵活的误差建模与不确定性感知优化融为一体，提供了一种有原则、数据驱动的解决方案，从源头上缓解过拟合，实现更强的定位鲁棒性和更高的准确率。

总而言之，我们的主要贡献有三点：

1. 我们证明，微小目标检测器极易受到标注噪声的影响，并指出严格的局部定位目标可能导致模型过拟合噪声标签。为解决这一问题，我们提出了**TOLF**，这是一种采用灵活分布建模的抗噪定位框架。2. TOLF 集成了基于归一化流的误差建模组件，以捕捉复杂的非高斯误差模式；还采用不确定性感知梯度调制机制，自适应地抑制来自高不确定性、易受噪声影响的样本的梯度。3. TOLF 显著提升了训练稳定性，并进一步提高了微小目标检测器的最先进准确率，为依赖固定先验或高斯假设的传统不确定性建模方法提供了一种有原则、由数据驱动的替代方案。

2. 相关工作

2.1. 微小目标检测

深度卷积神经网络（深度神经网络）的进展显著提升了目标检测任务的性能 [2]。尽管取得了这些进展，由于输入像素数量本身有限，微小目标检测仍然是一项具有挑战性的问题 [4]。

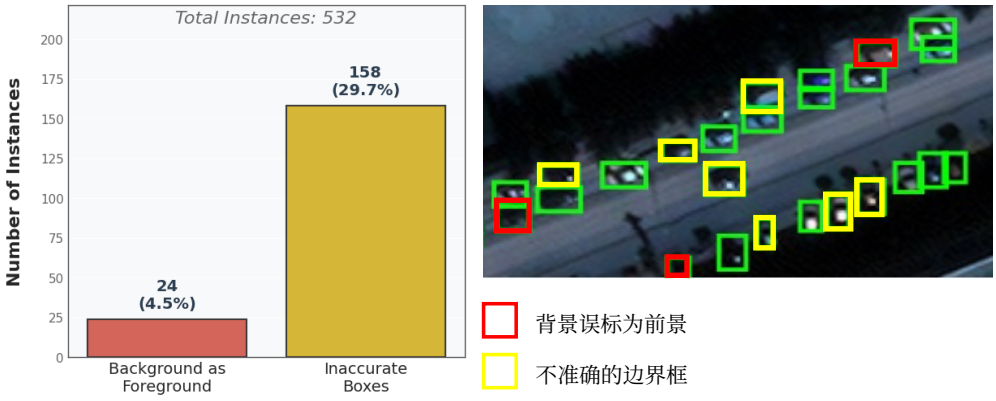


图2：基于人工检查 10 张 AI-TOD 测试图像中的 532 个边界框得出的标注质量统计。红色边界框/条形表示被误标为前景的背景区域，黄色边界框/条形表示不准确或过于宽松的边界框。结果表明，近 34.2% 的标注存在噪声。

主要难点包括特征表示能力较弱 [12]，下采样过程中的信息丢失 [15]，以及 IoU 计算敏感度提高导致正样本分配数量减少 [16]。针对这些问题的现有方法大致可分为四类：特征增强、数据增强、尺度感知训练和基于超分辨率的方法。

特征增强。 一个重要的研究方向是改进多尺度特征表示。SSD [17]利用不同分辨率的特征检测目标。FPN [12]引入自顶向下路径，并通过横向连接融合不同尺度的语义信息和空间信息。PANet [18] 和 Recursive-FPN [19]等方法进一步扩展了这一框架。TridentNet [20] 通过采用具有不同感受野的多个分支，针对不同目标尺寸进一步增强多尺度检测。SET [21]在异构架构中放大微小目标的频率特征。

数据增强。 除了常规的数据增强（例如翻转、旋转、缩放）之外，Kisantal等人 [15]通过在训练图像中过采样并复制粘贴微小目标来提升检测效果。少样本目标检测（FSOD）的最新进展也凸显了跨模态知识迁移在缓解微小目标类别数据稀疏问题中的作用。

尺度感知训练。 检测器通常难以在不同目标尺度下保持准确率。SNIP [22] 通过将训练限制在特定尺度区间内的目标上来解决这一问题。UGS [23] 将目标定位重新表述为分类任务，以稳定微小目标的梯度。

基于超分辨率的方法。 一些方法通过超分辨率技术增强微小目标特征。PGAN [24]将基于GAN的超分辨率集成到检测流程中。然而，这些方法通常会带来较高的训练和推理成本。近期策略着重改进标签分配和候选框精炼 [16]，，这对于提高微小目标的召回率和定位精度至关重要。

与现有的小目标检测方法不同，我们的方法通过基于流的不确定性建模来应对标注噪声和过拟合，为微小目标检测带来了全新的视角。

2.2. 使用噪声标签进行学习

噪声已成为现代机器学习范式中的关键组成部分。从注入结构随机性的丢弃层，到利用扰动实现鲁棒性的对抗训练，噪声驱动机制如今已成为提升深度神经网络（DNNs）泛化能力、稳定性和收敛性的核心手段 [25, 26, 27, 28,29, 30]。近期研究进一步指出，不确定性作为预测中的一种隐式噪声，可以通过揭示模型弱点来指导学习 [31]。这些进展与一种转向有益噪声学习的趋势相契合：通过受控地注入或利用噪声来提升模型性能。

医学中的噪声　学习。 标签噪声在医学诊断等任务中构成了重大挑战，因为不一致的

- 一致的标注 c　　误导模型。诸如 DAL [32] 这样的方法引入动态感知损失函数，以自适　　y

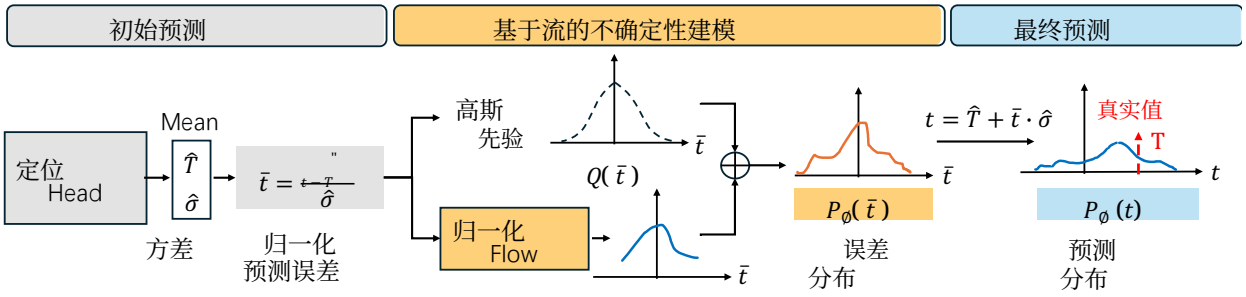


图3：噪声鲁棒定位框架 TOLF 概览。定位头为每个边界框预测均值 $\hat{T}$ 和不确定性 $\hat{\sigma}$。随后，使用归一化流 G_ϕ 对归一化预测误差进行建模，以捕捉复杂的非高斯误差分布。这使得能够稳健地估计预测分布 $P_\phi(t)$，并将不确定性和标签噪声纳入考量。

地平衡拟合能力与鲁棒性，而自步学习框架 [6] 借助医学指南检测并缓解标签噪声，从而提升多疾病诊断任务中的可解释性和性能。这些方法凸显了噪声感知学习对于提升标签敏感型应用中鲁棒性和可靠性的重要性。

多模态学习中的噪声。 多模态学习在处理含噪或不完整的数据流时面临重大挑战。传统方法（例如不完整多模态框架） [33], 通过选择性降低不可靠通道的权重，同时维持潜在跨模态一致性，来应对低质量或缺失信号。除了这些防御性策略外，近期研究还发现，经过精心设计的噪声能够主动提升多模态学习效果。在视觉语言模型中，有意注入噪声可通过促使模型具备抵御扰动的鲁棒性 [34]来增强跨模态对齐。同样，在对比学习框架中，常见的数据增强可以重新诠释为正向激励噪声，从而改善表征学习 [35]。这些方法共同表明，含噪或缺失的模态不应被视为纯粹有害因素。相反，通过将此类缺陷重新视为有益噪声注入的契机，多模态框架能够提升鲁棒性、改善对齐效果，并增强泛化能力。

目标检测中的噪声。 与图像分类相比，目标检测面临的标签噪声更加多样和复杂。这类噪声主要表现为四种类型：标签缺失、额外标签、类别偏移和不准确边界框。一些先前研究 [36, 37] 假设所有类型的噪声都会出现，并试图同时解决所有类型的噪声；另一些研究则 [38, 39]专注于处理特定类型的噪声（例如 不准确边界框） [39]。

总的来说，这些进展确立了抗噪学习在安全关键的真实噪声场景中的重要性。对于微小目标检测（TOD），这一点尤为关键，因为标注噪声会损害定位准确度和稳定性，因此促使我们开展这项研究。

3. 方法

本节介绍我们在噪声标注下实现鲁棒微小目标定位的方法。我们首先分析现有定位不确定性建模范式的局限性（第 3.1 节），然后介绍微小目标定位流（TOLF）框架，该框架联合学习灵活的预测分布和不确定性感知优化（第 3.2 节）。

3.1. 定位不确定性建模

传统定位。 沿用先前的检测器 [40, 41, 42], , 我们将定位目标和预测表示为：

$$\begin{aligned}\{T_x, T_y, T_w, T_h\} &= \{\frac{x - x_a}{w_a}, \frac{y - y_a}{h_a}, \log\frac{w}{w_a}, \log\frac{h}{h_a}\},\\ \{\hat{T}_x, \hat{T}_y, \hat{T}_w, \hat{T}_h\} &= \{\frac{\hat{x} - x_a}{w_a}, \frac{\hat{y} - y_a}{h_a}, \log\frac{\hat{w}}{w_a}, \log\frac{\hat{h}}{h_a}\},\end{aligned} \tag{1}$$

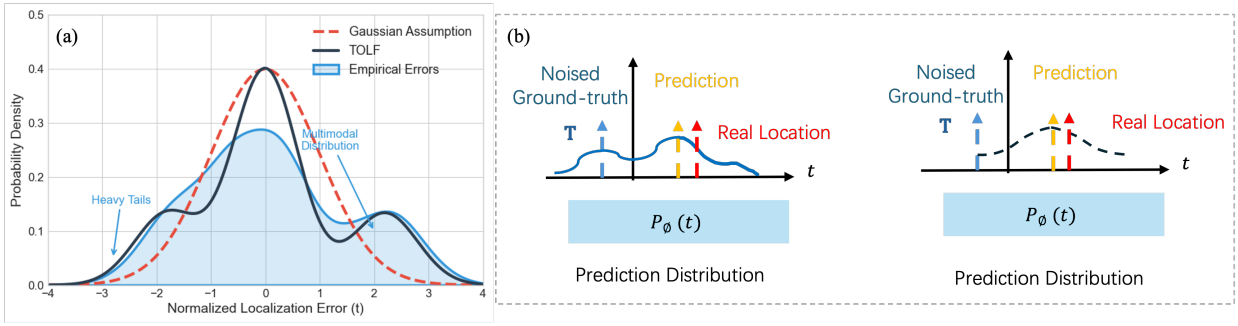


图 4: （a）TOLF 实现的灵活分布建模。与高斯假设相比，该方法能够更好地捕捉现实世界噪声。（b）抗噪定位示意图。模型预测以期望真实位置为中心的分布 $P_\theta(t)$，而不是回归到带噪的狄拉克真值。分布监督减少了对标签噪声的过拟合，并实现了不确定性感知定位。

ˆ其中（x_a，y_a，w_a，h_a）表示锚点坐标，（x，y，w，h）表示真值坐标，而（ˆx，ˆy，ˆw，h）分别表示预测坐标。x 的 $\mathcal{L}_2$ 损失可表示为：

$$\mathcal{L}_2(T_x, \hat{T}_x) = \|T_x - \hat{T}_x\|_2^2, \tag{2}$$

该方法也可应用于 y、w 和 h。为简化表示，我们使用 T 表示变换参数：$T = (T_x, T_y, T_w, T_h)$。从最大似然估计（MLE）的角度来看，$\mathcal{L}_2$ 损失假设定位误差服从同方差高斯分布：

$$P(T \mid \hat{T}) = \mathcal{N}(T; \hat{T}, \sigma^2), \tag{3}$$

其中，方差 σ^2 是固定的，并在所有训练样本之间共享。然而，目标定位通常表现出异方差性 [43], ，即定位不确定性会因目标尺寸、遮挡或模糊等因素而在不同样本之间显著变化。

基于高斯的建模。近期方法通过显式建模定位不确定性来处理该问题。为联合学习定位结果及其置信度，[44]将边界框回归表述为最小化狄拉克真值分布 P_D 与预测高斯分布P_Θ之间的 KL 散度 $D_{\mathrm{KL}}(\cdot)$。这两个分布可表示为：

$$P_\Theta(t) = \mathcal{N}(t \mid \hat{T}, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(t-\hat{T})^2}{2\sigma^2}\right), \quad P_D(t) = \delta(t - T), \tag{4}$$

其中，ˆT 表示预测均值，T 表示定位真值目标，而 σ表示学习得到的不确定性。最后，回归损失由 KL 散度推导得出：

$$\mathcal{L}_{\mathrm{reg}} = D_{\mathrm{KL}}(P_D, |, P_\Theta) = \frac{(T-\hat{T})^2}{2\sigma^2} + \frac{1}{2}\log\sigma^2 + C, \tag{5}$$

其中 $C = \frac{1}{2}\log(2\pi)$ 为常数。请注意，高斯形式是固定且对称的，因此其表示多模态或长尾标注误差的能力有限。

基于分类的监督。基于分类的监督可以减轻噪声标签的影响，这也可以从分布的角度来理解。广义焦点损失（GFL V1） [45] 通过将连续目标量化为离散软分布，引入了一种用于定位的分类范式。对于回归目标$T \in [-\alpha, \alpha]$，连续范围被划分为$n+1$ 个区间，网格点为$Y = \{\mathbf{y}_0, \mathbf{y}_1, \ldots\ldots \mathbf{y}_n\}$。定位头生成的类别分布隐式地对不确定性进行建模，其中网络预测逻辑值$\mathbf{l} \in \mathbb{R}^{n+1}$，再通过 softmax 转换为概率：

$$\mathbf{p}_i = \frac{\exp(\mathbf{l}_i)}{\sum_{k=0}^{n}\exp(\mathbf{l}_k)}, \quad \forall i \in \{0, \ldots, n\}. \tag{6}$$

真实值使用基于相邻网格i_l和i_r的双热目标进行编码：

$$\mathbf{p}_i^* = \begin{cases} |\mathbf{y}_i - T| \cdot \frac{n+1}{2\alpha}, & i = i_l, i_r \\ 0, & \text{otherwise}, \end{cases} \tag{7}$$

通过交叉熵进行优化：

$$\mathcal{L}_{\text{CE}} = -\mathbf{p}_{i_l}^* \log \mathbf{p}_{i_l} - \mathbf{p}_{i_r}^* \log \mathbf{p}_{i_r}. \tag{8}$$

虽然通过离散分布对非高斯不确定性进行建模，但二热目标编码（式 7）施加了分段线性结构，这种结构可能与定位中的真实不确定性不一致，从而限制了预测分布的灵活性和保真度。此外，固定数量的分箱也限制了尾部的表达能力。

3.2. TOLF：微小目标定位流

框架。为缓解微小目标检测中对含噪定位标签的过拟合，我们提出 TOLF，这是一种基于流的框架，可通过不确定性估计学习灵活的预测分布。整体概览如图 3 所示。

TOLF 引入了一个概率定位头，为每个边界框坐标同时输出预测均值 $\hat{}T_i$ 和不确定性 $\hat{\sigma}_i$ 。遵循 [46], ，我们对归一化预测误差的分布进行建模，而非对预测坐标进行建模，其定义如下：

$$\bar{t}_i = \frac{T_i - \hat{T}_i}{\hat{\sigma}_i}, \tag{9}$$

其中，T_i 是真实值，$\hat{}T_i$ 是预测均值，$\hat{\sigma}_i$ 是预测不确定性。

为捕捉标注噪声的复杂特性，我们使用归一化流对 $\bar{}t_i$ 的分布进行建模。该框架通过一系列可逆映射，将简单的基分布（例如高斯分布）转换为复杂的目标分布。

使用归一化流，我们通过灵活的误差分布 $P_\phi(\bar{}t_i)$ 对 $\bar{}t_i$ 进行建模。$P_\phi(\bar{}t_i)$ 是一种灵活的密度近似器，能够通过可逆变换突破参数化约束。遵循 [46], ，我们将该分布定义为：

$$P_\phi(\bar{t}_i) = Q(\bar{t}_i) \cdot G_\phi(\bar{t}_i) \cdot s, \tag{10}$$

其中，$Q(\bar{}t_i) = \mathcal{N}(0,1)$ 是标准高斯先验，$G_\phi(\bar{}t_i)$ 是由流模型学习得到的密度校正，而s 是归一化常数，用于确保 P_ϕ 的积分为 1：

$$s = \left(\int Q(\bar{t}) G_\phi(\bar{t}) d\bar{t} \right)^{-1}. \tag{11}$$

与传统参数化模型（例如高斯分布和拉普拉斯分布）相比，归一化流具有更强的表达能力，能够对分布的偏度、厚尾和多峰性进行建模。这些特性对于在真实场景中的标注噪声条件下实现稳健定位至关重要。

学习得到的似然 $P_\phi(\bar{}t_i)$ 通过负对数似然损失提供监督。对于每个回归目标 T_i，损失分量计算如下：

$$\begin{aligned} \mathcal{L}_{\text{nf}}^{(i)} &= -\log P_{\Theta,\phi}(T_i | \mathcal{I}) \\ &= -\log P_\phi(\bar{t}_i) + \log \hat{\sigma}_i \\ &= -\log Q(\bar{t}_i) - \log G_\phi(\bar{t}_i) - \log s + \log \hat{\sigma}_i. \end{aligned} \tag{12}$$

总回归损失通过对所有边界框参数求和计算（例如，{x、y、w、h}）：

$$\mathcal{L}_{\text{nf}} = \sum_{i \in \{x,y,w,h\}} \mathcal{L}_{\text{nf}}^{(i)}. \tag{13}$$

该公式使模型能够学习回归目标的均值和方差，同时还可通过流模型实现灵活的非高斯误差建模。

表 1：在 AI-TOD [4] 上使用各种框架的主要结果。模型在 AI-TOD 训练验证集上训练，并在 AI-TOD 测试集上验证。我们报告了不同 IoU 阈值下的平均精度（%），以及根据 AI-TOD 标准计算的不同尺寸目标的平均精度（%）。* 表示使用 P2~P6 FPN 特征。**加粗**表示最佳结果。

框架	AP	平均精度$_{0.5}$	平均精度$_{0.75}$	AP$_{vt}$	AP$_{t}$	AP$_{s}$
PAA [47]	10.0	26.5	6.7	3.5	10.5	13.1
ATSS [48]	11.6	28.5	7.6	2.5	11.9	15.9
CenterNet [49]	13.4	39.2	5.0	3.8	12.1	17.7
DetectoRS [19]	14.8	32.8	11.5	0.0	10.8	18.3
DotD [50]	16.1	39.2	10.6	8.3	17.6	18.1
NWD [51]	20.8	49.3	14.3	6.4	19.7	29.6
SR-TOD [9]	24.0	54.6	17.1	10.1	24.8	29.3
单阶段						
FCOS [14]	12.0	29.0	8.0	2.5	11.9	17.1
w/ TOLF	**13.2**	**32.1**	**9.0**	**3.2**	**13.5**	**19.4**
FCOS* [14]	15.1	35.8	10.2	5.9	16.6	18.8
w/ TOLF	**16.7**	**37.5**	**12.2**	**6.8**	**18.0**	**21.9**
多阶段						
Faster R-CNN [42]	11.1	26.3	8.1	0.0	7.2	23.3
w/ TOLF	**12.8**	**28.9**	**10.3**	**0.2**	**9.5**	**25.1**
Cascade R-CNN [52]	13.6	30.3	10.6	0.0	9.9	25.5
w/ TOLF	**15.4**	**33.1**	**11.8**	**0.5**	**11.3**	**27.5**
RFLA [16]	21.7	50.5	15.3	8.3	21.8	24.5
w/ TOLF	**23.0**	**53.2**	**17.6**	**10.1**	**23.7**	**27.9**
基于Transformer的						
DINO-5scale [5]	23.2	56.6	15.4	9.9	23.1	29.3
w/ TOLF	**24.4**	**57.0**	**17.2**	**11.0**	**24.4**	**31.0**

不确定性感知梯度调制。 TOLF 通过双重梯度调制提升噪声鲁棒性。

$$\frac{\partial \mathcal{L}_{\text{nf}}^{(i)}}{\partial \hat{T}_i} = \frac{\partial \mathcal{L}_{\text{nf}}^{(i)}}{\partial \bar{t}_i} \cdot \frac{\partial \bar{t}_i}{\partial \hat{T}_i} = \left(-\partial_{\bar{t}_i} \log Q(\bar{t}_i) - \partial_{\bar{t}_i} \log G_\phi(\bar{t}_i)\right) \cdot \left(-\tfrac{1}{\hat{\sigma}_i}\right)$$
$$= \frac{1}{\hat{\sigma}_i}\left(\partial_{\bar{t}_i} \log Q(\bar{t}_i) + \partial_{\bar{t}_i} \log G_\phi(\bar{t}_i)\right). \tag{14}$$

$1/\hat{\sigma}_i$ 项会对高不确定性预测的梯度进行自适应衰减，而 $\partial \log G_\phi$ 则引导更新朝着学习到的误差分布的高密度区域进行。这会抑制针对噪声标签的更新，同时保留针对干净且定位良好的目标的稳定更新。

$$\frac{\partial \mathcal{L}_{\text{nf}}^{(i)}}{\partial \hat{\sigma}_i} = \frac{\partial \mathcal{L}_{\text{nf}}^{(i)}}{\partial \bar{t}_i} \cdot \frac{\partial \bar{t}_i}{\partial \hat{\sigma}_i} + \frac{\partial}{\partial \hat{\sigma}_i}(\log \hat{\sigma}_i)$$
$$= \left(-\partial_{\bar{t}_i} \log Q(\bar{t}_i) - \partial_{\bar{t}_i} \log G_\phi(\bar{t}_i)\right) \cdot \left(-\tfrac{T_i - \hat{T}_i}{\hat{\sigma}_i^2}\right) + \tfrac{1}{\hat{\sigma}_i}$$
$$= \frac{T_i - \hat{T}_i}{\hat{\sigma}_i^2}\left(\partial_{\bar{t}_i} \log Q(\bar{t}_i) + \partial_{\bar{t}_i} \log G_\phi(\bar{t}_i)\right) + \tfrac{1}{\hat{\sigma}_i}. \tag{15}$$

公式 (15) 揭示了 TOLF 的鲁棒不确定性学习。当较大的误差 $|T_i - \hat{T}_i|$源于标注噪声而非模型误差时，梯度项 $\partial \log G_\phi$ 会减小更新幅度，从而防止不确定性过度膨胀。同时，$1/\hat{\sigma}_i$ 分量可防止不确定性坍缩为零，维持校准效果。这两种机制共同构成了一种平衡梯度调制方案，在降低噪声标注权重的同时，为定位良好的目标保留稳定更新，从而避免过拟合并确保可靠收敛。

表2：在 TinyPerson [54] 上使用不同框架的检测结果。所有模型均在训练集上训练，并在验证集上评估。我们按照 TinyPerson基准，在不同 IoU 阈值和目标尺寸下报告平均精度（%）。**粗体** 表示各基础检测器组中的最佳结果。

框架	平均精度$^{tiny}_{50}$	平均精度$^{tiny1}_{50}$	平均精度$^{tiny2}_{50}$	平均精度$^{tiny3}_{50}$	平均精度$^{small}_{50}$	平均精度$^{tiny}_{25}$	平均精度$^{tiny}_{75}$
FCOS [14]	16.9	3.9	12.4	29.3	36.8	40.5	1.5
Faster R-CNN [42]	43.6	48.3	53.5	43.6	56.7	64.1	5.4
RetinaNet [58]	15.5	3.0	14.4	29.1	46.8	48.4	1.3
RetinaNet/ TOLF	**17.2**	**3.7**	**15.8**	**29.6**	**47.3**	**51.6**	**1.5**
AutoAssign [59]	21.0	7.1	19.7	32.3	48.1	55.0	1.4
AutoAssign/TOLF	**22.1**	**7.5**	**20.8**	**33.2**	**50.3**	**57.3**	**2.0**

4. 实验

4.1. 数据集与实现细节

数据集。我们在三个基准测试数据集上评估了我们的方法：**AI-TOD** [4], **DOTA-v2**。**0** [53], 和 **TinyPerson** [54]。我们的主要实验基于 AI-TOD，该数据集具有挑战性，其目标平均尺寸仅为 12.8 像素，显著小于标准检测数据集（如 MS COCO 中的 99.5 像素） [2]。我们还将该方法应用于 DOTA-v2.0 和 TinyPerson，这两个数据集都包含高分辨率的航空或无人机影像，并且微小目标密度较高。

实现细节。我们在配备 NVIDIA RTX 3090 GPU 的计算机上进行了实验。所有基于 CNN 的模型均采用 ResNet-50 [55] 主干网络，使用随机梯度下降（SGD）优化器训练 12 个 epoch，动量为 0.9，权重衰减为 0.0001，批量大小为 2。初始学习率为 0.005，并在第 8 和第 11 个 epoch 时衰减。数据处理遵循各数据集的默认配置（例如，AI-TOD 的固定值为 800×800）。我们还训练了一个基于Transformer的检测器 DINO [5], , 采用 5 个尺度的特征图，训练 36 个 epoch，作为基线。训练使用 Adam优化器，权重衰减为 0.0001，并遵循 DETR [56] 的随机裁剪和尺度增强策略。

我们提出的定位范式不受归一化流具体设计的限制。在实验中，我们采用 RealNVP [57] ，因为其训练速度快且稳定。我们将可逆变换表示为一个全连接架构，其中包含 L_{fc} 层，每层有 N_{n} 个神经元，即 $L_{\mathrm{fc}} \times N_{\mathrm{n}}$。默认情况下，我们设定 $L_{\mathrm{fc}} = 3$ 和$N_{\mathrm{n}} = 64$。该流模型十分轻量，对整体训练过程的额外开销可以忽略不计。

4.2. AI-TOD 上的结果

我们在 AI-TOD基准 [60], 上使用多种检测器评估 TOLF，并与最先进的小目标检测方法进行比较。如表 1 所示，TOLF 始终能使所有基线的平均精度提升 ~2%。值得注意的是，它使单阶段 FCOS [14]检测器的平均精度分别提升 1.2% 和 1.6%。加入 P2~P6 FPN 特征——这是利用高分辨率 P2 检测微小目标的代表性 TOD 配置——后，TOLF 进一步使 FCOS 的平均精度提升 1.6%。TOLF 也能有效泛化至多阶段检测器，使 Faster R-CNN [42] 和 Cascade R-CNN [52]的平均精度分别提升 1.7% 和 1.8%。尤为重要的是，TOLF 与最先进的 RFLA [16]方法相辅相成，带来 1.3% 的平均精度增益。对于基于Transformer的检测器，TOLF 在 DINO-5scale [5]上达到 24.4% 的平均精度（提升 1.2%），优于包括 DotD [50],NWD [51], 和 SR-TOD [9]在内的竞争方法。

*4.3. DOTA-v2.0*和*TinyPerson*上的结果

我们在两个具有挑战性的微小目标检测基准上评估了 TOLF 的有效性：DOTA-v2.0 [53] 和 TinyPerson [54], , 这两个基准都包含密集分布的低分辨率目标。如表 3 所示，TOLF 在 DOTA-v2.0 上始终能提升多种检测器的性能。对于 FCOS，TOLF 将平均精度提高了 1.5%，其中 AP_{vt} 提高了 0.6%，AP_t提高了 0.7%。在 AutoAssign 的基础上，TOLF 带来了更显著的提升，平均精度提高了 1.7%，其中 AP_{vt} 和 AP_t分别提高了 0.3% 和 1.4%。与先前的方法 RFLA [16]和 DCFL [61], 相比，TOLF 优于二者；在与 FCOS 配合使用时，其平均精度分别比 RFLA 和 DCFL 高出 3.2% 和 1.5%。

表 2 报告了 TinyPerson 数据集上的结果。TOLF 在两种不同的基线上均带来了稳定的提升。对于 RetinaNet，TOLF 将平均精度$^{tiny}_{50}$ 提高了 1.7%，并将平均精度$^{tiny}_{25}$ 提高了 3.2%。应用于 AutoAssign 时，TOLF

表 3：检测性　　能在 DOTA-v2.0 [53] 上的表现。所有模型均在 DOTA-v2.0 训练集上训练，并在验证集上评估。TOLF 被　　s 应用于四种有代表性的 基础检测器。**粗体** 表示各组中的最佳结果。

框架	AP	AP_{vt}	AP_t	AP_s
ATSS [48]	32.7	0.7	6.9	23.4
ATSS/ TOLF	**34.1**	**0.8**	**7.3**	**24.3**
Faster R-CNN [42]	35.6	0.0	7.1	28.9
Faster R-CNN/TOLF	**36.5**	**0.4**	**7.5**	**29.5**
FCOS [14]	31.8	0.3	4.0	19.4
FCOS 使用 RFLA [16]	32.1	**0.7**	6.8	23.6
FCOS/ TOLF	**33.3**	0.6	**7.1**	**24.8**
AutoAssign [59]	33.8	0.9	7.3	22.4
AutoAssign w/TOLF	**35.5**	**1.2**	**8.7**	**23.9**

实现了显著的 平均精度$^{tiny}_{50}$ 提高了 1.1%，并提升了所有子类别的性能。这些结果凸
凸显了 TOLF 的泛 化能力和鲁棒性在不同架构及真实场景中的表现。

表 4：MS COCO [2] 上的检测性能。请注意，模型在 COCO train2017 上训练，并在 COCO val2017 上验证。

框架	AP	AP_{vt}	AP_t	AP_s	AP_m
FCOS [14]	36.4	7.9	19.6	27.2	43.6
FCOS 搭配 TOLF	37.2	8.9	20.8	28.3	44.5

4.4. COCO 上的结果

我们进一步在大规模基准测试 MS COCO [2], 上验证了 TOLF 的性能。如表 4 所示，TOLF 相较于 FCOS 基线带来了显著提升，整体平均精度达到 +0.8。根据 AI-TOD 指标，微小目标的提升尤为显著：超小目标的平均精度为 +1.0，微小目标的平均精度为 +1.2。这些结果证实，TOLF 不仅在微小目标数据集上有效，在通用目标检测基准测试上也同样有效。

如图 5 所示，学习得到的逐坐标残差分布呈非高斯且不对称的形态。对于左侧坐标，在 −0.9 和 0.9 附近出现了两个众数。0 附近的高密度表示标注一致，而厚尾则表明存在歧义或离群值。这些结果表明，正规化流模型能够捕捉超出高斯假设范围的标注噪声。

4.5. 可视化

如图 5 所示，学习得到的逐坐标残差分布呈非高斯且不对称。对于左坐标，在 −0.9 和 0.9 附近出现了两个峰值。0 附近的高密度表明标注一致，而厚尾则表明存在歧义或离群值。这些结果表明，归一化流模型能够捕捉超出高斯假设范围的标注噪声。红色方框区域突出显示了方差较高的区域，表明存在局部不确定性，这会调节分布锐度。这通过在高不确定性区域进行梯度衰减来指导噪声感知训练，从而增强对含噪标注的鲁棒性。

图 4 展示了 TOLF 设计在具有挑战性的场景中带来的显著改进。与基线相比，我们的方法实现了更准确的定位，并减少了杂乱环境中的假阳性，证实了 TOLF 在真实噪声条件下提升检测可靠性的有效性。

4.6. 消融研究

为评估我们框架中各组件的有效性，我们以 FCOS* 作为基线，在 AI-TOD 测试集上开展了广泛的消融研究。结果如表 5 所示。

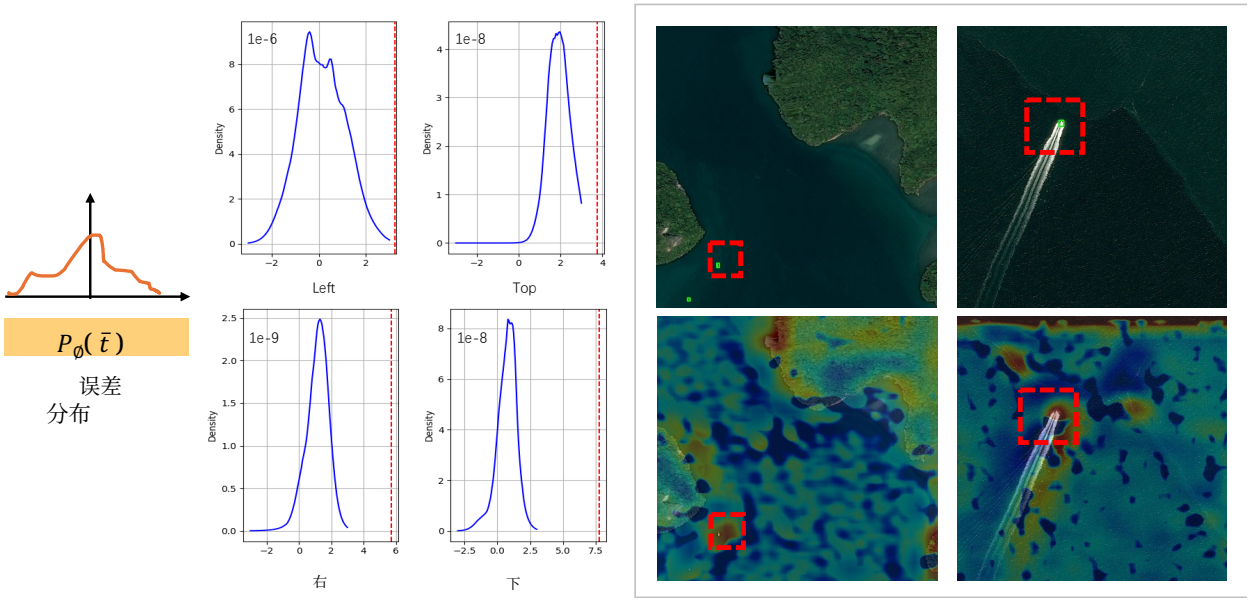


图5：**左**。 使用归一化流学习得到的边界框坐标（左、右、上、下）的一维误差分布。每个分布$P_\theta(t)$绘制了特定坐标（例如左边界）的残差误差，该误差以其他坐标的固定值为条件。**右**。 四个坐标的平均预测方差 σ 与输入叠加显示。

表5：TOLF 组件在 AI-TOD 测试集上的消融研究。所有实验均以 FCOS* 为基线。

变体（λ）	AP	平均精度$_{0.5}$	平均精度$_{0.75}$	AP$_{vt}$	AP$_t$	AP$_s$
FCOS*	15.1	35.8	10.2	5.9	16.6	18.8
核心组件						
+ 归一化流	15.9	36.7	10.8	6.3	17.0	19.7
+ 不确定性	15.7	36.2	10.6	6.1	16.8	19.3
+ TOLF（$\lambda = 0.1$）	**16.7**	**37.5**	**12.2**	**6.8**	**18.0**	**21.9**
定位不确定性比较						
使用 KL损失 [44]	15.6	36.2	10.9	5.7	17.2	20.5
使用 GFocal [43]	16.0	36.8	11.4	6.0	17.6	20.7
损失权重						
TOLF（$\lambda = 0.01$）	16.1	36.7	11.6	6.3	17.3	20.9
TOLF（$\lambda = 0.1$）	**16.7**	**37.5**	**12.2**	**6.8**	**18.0**	**21.9**
TOLF（$\lambda = 1.0$）	16.2	36.6	11.0	6.0	17.4	20.6

组件分析。 我们首先评估 TOLF 的核心组件各自的作用。使用对数似然目标（$\mathcal{L}_{\text{flow}} = -\log P_\phi(T_i - \hat{T}_i)$）的归一化流对残差误差进行建模后，平均精度从 15.1% 提升至 15.9%，表明超越固定高斯假设的灵活、数据驱动的误差建模能够增强模型在标注噪声下的鲁棒性。将不确定性感知加权引入最终损失函数（$\mathcal{L}_{\text{uncertainty}} = |T_i - \hat{T}_i|/\sigma_i$）后，平均精度取得了相近的提升，达到 15.7%，这表明根据预测不确定性自适应调节梯度能够减轻噪声标签或模糊标签的影响。结合这两个组件后得到完整的 TOLF框架，平均精度达到 16.7%，较基线绝对提升 +1.6%，并且在 AP$_{0.5}$、AP$_{0.75}$ 以及所有目标尺度（vt/t/s）上均取得了稳定增益。

与其他损失的比较。 我们将 TOLF 基于似然的损失与常用的不确定性感知损失进行比较：KL损失 [44] 和 GFocal [43]。虽然这两种替代方法都优于基线，但 TOLF 在所有指标上均取得了更高的准确率，表明通过流显式建模残差误差分布，比假设预定义误差形式能够实现更稳健的定位。

Effect of TOLF 的损失权重 λ。 我们进一步研究了不确定性调节权重 λ 的影响。如

表 6：TOLF 各组件在 AI-TOD 数据集上的效率分析。

方法	性能 (%)			Cost		
	AP	AP_{vt}	AP_t	时间（毫秒）	十亿次浮点运算	参数量（M）
FCOS 基线	12.0	2.5	11.9	17.2	126.1	37.0
+ TOLF（完整）	13.7	2.9	13.1	19.8	127.4	37.5
+ TOLF（简化版）	13.5	2.8	13.0	18.1	126.8	37.2

表 7：AI-TOD 数据集上不同图块尺寸下的性能对比

图块尺寸	方法	AP	AP_{vt}	AP_t	AP_s
8×8	DINO-5scale	22.8	9.2	22.7	29.0
	TOLF	**24.2**	**10.3**	**24.1**	**30.5**
16×16	DINO-5scale	22.3	8.7	22.2	28.5
	TOLF	**24.4**	**10.5**	**24.3**	**30.7**
32×32	DINO-5scale	20.1	6.5	20.0	26.8
	TOLF	**23.0**	**9.1**	**22.9**	**29.4**

表 5 所示，较小的值 $\lambda = 0.1$ 表现最佳，在保留有用梯度与抑制易受噪声影响的更新之间取得了平衡。

效率分析。我们展示了所提出的 TOLF 框架仅引入极少的计算开销，同时提升了性能。如表 6 所示，完整的 TOLF 实现仅使推理时间增加 15.1%（+2.6 毫秒），计算复杂度增加 1.0%，同时平均精度提高 1.7%。为进一步优化效率，我们探索了一种简化的 RealNVP [57] 配置，将耦合层数量从 6 层减少至 3 层，并在每个变换块中采用更浅的神经网络。该简化变体保留了 98.5% 的性能提升，同时将额外延迟降低至仅 5.2 %。性能方面的微小折衷表明，TOLF 有望部署于资源受限环境，同时不损害其核心效能。

对图块尺寸变化的鲁棒性。近期基于 Transformer 的网络已显示出对图块尺寸变化的敏感性 [62], ，这可能影响特征粒度和容量 [63]。我们研究了 TOD 中的图块尺寸变化，并在相同设置下评估 TOLF，以考察其鲁棒性。如表 7 所示，将图块尺寸减小至8×8 会使 DINO-5scale 获得微小提升，主要惠及微小目标。TOLF 在所有配置下均持续提升，并在 32×32 时保持稳定，超过 DINO-5scale 2.9 个平均精度。这表明，即使特征表示较为粗糙，TOLF 仍能保持检测质量。

5. 结论

本文致力于解决标注噪声条件下微小目标鲁棒定位这一挑战，这是一个至关重要却尚未得到充分研究的问题。我们发现，传统微小目标检测器对噪声标签高度敏感，尤其是在采用严格定位目标进行训练时，这种目标会无意中加剧过拟合。为此，我们提出 TOLF，一种抗噪定位框架，通过归一化流对残差误差进行建模，并借助不确定性引导优化抑制不可靠的监督信号。TOLF 通过可逆变换实现灵活的非高斯误差建模，并引入不确定性感知梯度调制，对方差较高、易受噪声影响的预测赋予较低权重。在三个具有挑战性的基准数据集上开展的大量实验表明，TOLF 始终能够提高微小目标的检测准确率。本研究凸显了灵活的标签噪声建模对于提升微小目标检测器可靠性的重要性。

参考文献

[1] K. He, J. Sun，受限时间成本下的卷积神经网络，载于：CVPR，2015，页码 5353–5360。[2] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C. L. Zitnick，微软 COCO：上下文中的常见物体，载于：ECCV，Springer，2014年，页码 740–755。[3] C. Yang, Z. Huang, N. Wang，QueryDet：用于加速高分辨率小目标检测的级联稀疏查询，载于：CVPR，2022，页码 13668–13677。[4] J. Wang, W. Yang, H. Guo, R. Zhang, G.-S. Xia，航拍图像中的微小目标检测，载于：国际模式识别会议，2021，页码 3791–3798。[5] H. Zhang, F. Li, S. Liu, L. Zhang, H. Su, J. Zhu, L. M. Ni, H.-Y. Shum，DINO：采用改进去噪锚框的端到端目标检测 DETR，国际学习表征会议（2023年）。[6] J. Long, J. Ren，基于匹配网络和自步学习的可解释多疾病诊断与标签噪声检测，《模式识别》148（2024）110178。[7] R. Liang, Y. Li, Y. Yi, J. Zhou, X. Li，一种用于驾驶视频无监督交通异常检测的记忆增强多任务协作框架，《模式识别》（2025年）111789。[8] L. Zhang, G. Wang, M. Chen, F. Ren, L. Shao，一种用于无人机目标检测的增强型抗噪哈希，《模式识别》143（2023年）109762。[9] B. Cao, H. Yao, P. Zhu, Q. Hu，可见且清晰：在差异图中寻找微小目标，载于：ECCV，2024年。[10] S. Wang, F. Nie, Z. Wang, R. Wang, X. Li，用于异常检测的模糊加权主成分分析，ACM《数据知识发现汇刊》19（3）（2025年）1–22。[11] J. Zhou, Z. He, D. Zhang, S. Liu, X. Fu, X. Li，用于水下目标检测的空间残差，《IEEE 模式分析与机器智能汇刊》（2025年）。[12] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, S. Belongie，用于目标检测的特征金字塔网络，载于：计算机视觉与模式识别会议，2017，页码 2117–2125。[13] B. Du, Y. Huang, J. Chen, D. Huang，融合全局上下文增强的自适应稀疏卷积网络，用于加速无人机图像上的目标检测，载于：计算机视觉与模式识别会议，2023年，页码 13435–13444。[14] Z. Tian, C. Shen, H. Chen, T. He，FCOS：全卷积单阶段目标检测，载于：ICCV，2019年，页码 9627–9636。[15] M. Kisantal, Z. Wojna, J. Murawski, J. Naruniec, K. Cho 等，面向小目标检测的数据增强，载于：计算机科学与信息技术会议论文集，第 9 卷，计算机科学与信息技术会议论文集，2019。[16] C. Xu, J. Wang, W. Yang, H. Yu, L. Yu, G.-S. Xia，RFLA：基于高斯感受野的微小目标检测标签分配，载于：ECCV，2022，页码 526–543。[17] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, A. C. Berg，SSD：单次多框检测器，载于：ECCV，2016，页码 21–37。[18] S. Liu, L. Qi, H. Qin, J. Shi, J. Jia，用于实例分割的路径聚合网络，载于：CVPR，2018，页码 8759–8768。[19] S. Qiao, L.-C. Chen, A. Yuille，检测器：通过递归特征金字塔和可切换空洞卷积检测目标，载于：CVPR，2021，页码 10213–10224。[20] Y. Li, Y. Chen, N. Wang, Z. Zhang，用于目标检测的尺度感知三叉戟网络，载于：ICCV，2019，页码 6054–6063。

[21] H. Sun, R. Wang, Y. Li, L. Yang, S. Lin, X. Cao, B. Zhang, 《SET：用于微小目标检测的频谱增强》，载于：计算机视觉与模式识别会议论文集，2025年，第4713–4723页。[22] B. Singh, L. S. Davis, 《目标检测中尺度不变性的分析：SNIP》，载于：计算机视觉与模式识别会议，2018年，第3578–3587页。[23] H. Sun, Y. Li, L. Yang, X. Cao, B. Zhang, 《面向小目标检测的不确定性感知梯度稳定化》（2025）。arXiv:2303.01803。网址：https://arxiv.org/abs/2303.01803[24] J. Li, X. Liang, Y. Wei, T. Xu, J. Feng, S. Yan, 《用于小目标检测的感知生成对抗网络》，载于：计算机视觉与模式识别会议，2017年，第1222–1230页。[25] I. Goodfellow, 《深度学习》（2016）。[26] X. Li, 《正向激励噪声》，IEEE神经网络与学习系统汇刊 35 (6) (2024) 8708–8714。[27] H. Zhang, S. Huang, Y. Guo, X. Li, 《变分正向激励噪声：噪声如何使模型受益》，IEEE《模式分析与机器智能汇刊》（2025）。[28] S. Huang, Y. Xu, H. Zhang, X. Li, 《将有益噪声作为图增强进行学习》，载于：第42届国际机器学习会议论文集（国际机器学习会议），2025年。[29] Z. Huang, X. Qiu, Y. Ma, Y. Zhou, J. Chen, H. Zhang, C. Zhang, X. Li, 《Nfig：基于下一频率预测的自回归图像生成》，arXiv预印本 arXiv:2503.07076 (2025)。[30] K. Jiang, Z. Shi, D. Zhang, H. Zhang, X. Li, 《用于基于预训练模型的类增量学习的噪声混合》，arXiv预印本 arXiv:2509.16738 (2025)。[31] A. Kendall, Y. Gal, 《计算机视觉的贝叶斯深度学习需要哪些不确定性？》，《神经信息处理系统进展》30 (2017)。[32] X.-C. Li, X. Xia, F. Zhu, T. Liu, X.-Y. Zhang, C.-L. Liu, 《面向标签噪声学习的动态感知损失》，《模式识别》144 (2023) 109835。[33] S. Kanwal, I. A. Taj, 《不完整RGB-D显著目标检测：隐藏、关联与融合》，《模式识别》（2024）110700。[34] S. Huang, H. Zhang, X. Li, 《利用噪声增强视觉语言对齐》，载于：AAAI人工智能会议论文集，第39卷，2025年，第17449–17457页。[35] H. Zhang, Y. Xu, S. Huang, X. Li, 《对比学习中的数据增强是在估计正向激励噪声》，arXiv预印本 arXiv:2408.09929 (2024)。[36] S. Chadwick, P. Newman, 《使用噪声数据训练目标检测器》，载于：2019年IEEE智能车辆研讨会（IV），IEEE，2019年，第1319–1325页。[37] J. Li, C. Xiong, R. Socher, S. Hoi, 《面向带噪声标注的抗噪目标检测》，arXiv预印本 arXiv:2003.01285 (2020)。[38] X. Zhang, Y. Yang, J. Feng, 《学习利用带噪标签的实例定位目标》，载于：AAAI人工智能会议论文集，第33卷，2019年，第9219–9226页。[39] C. Liu, K. Wang, H. Lu, Z. Cao, Z. Zhang, 《利用不准确边界框实现鲁棒目标检测》，载于：欧洲计算机视觉会议，Springer，2022年，第53–69页。[40] R. Girshick, J. Donahue, T. Darrell, J. Malik, 《用于精确目标检测和语义分割的丰富特征层次结构》，载于：计算机视觉与模式识别会议，2014年，第580–587页。

[41] R. Girshick, 《Fast R-CNN》，载于：ICCV，2015年，第1440–1448页。[42] S. Ren, K. He, R. Girshick, J. Sun, 《Faster R-CNN：利用区域提议网络实现实时目标检测》，载于：NeurIPS，第28卷，2015年。[43] X. Li, C. Lv, W. Wang, G. Li, L. Yang, J. Yang, 《广义焦点损失：面向密集目标检测的高效表征学习》，IEEE TPAMI（2022）。[44] Y. He, C. Zhu, J. Wang, M. Savvides, X. Zhang, 《利用不确定性进行边界框回归，实现精确目标检测》，载于：计算机视觉与模式识别会议，2019年，第2888–2897页。[45] X. Li, W. Wang, L. Wu, S. Chen, X. Hu, J. Li, J. Tang, J. Yang, 《广义焦点损失：学习密集目标检测中的合格分布式边界框》，载于：NeurIPS，2020年。[46] J. Li, S. Bian, A. Zeng, C. Wang, B. Pang, W. Liu, C. Lu, 《利用残差对数似然估计进行人体姿态回归》，载于：IEEE/CVF国际计算机视觉大会论文集，2021年，第11025–11034页。[47] K. Kim, H. S. Lee, 《用于目标检测的IoU预测概率锚框分配》，载于：ECCV，2020年，第355–371页。[48] S. Zhang, C. Chi, Y. Yao, Z. Lei, S. Z. Li, 《通过自适应训练样本选择弥合基于锚框与无锚框检测之间的差距》，载于：计算机视觉与模式识别会议，2020年，第9759–9768页。[49] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, Q. Tian, 《CenterNet：用于目标检测的关键点三元组》，载于：ICCV，2019年，第6569–6578页。[50] C. Xu, J. Wang, W. Yang, L. Yu, 《用于航拍图像微小目标检测的点距离》，载于：计算机视觉与模式识别会议，2021年，第1192–1201页。[51] J. Wang, C. Xu, W. Yang, L. Yu, 《用于微小目标检测的归一化高斯沃瑟斯坦距离》，arXiv预印本 arXiv:2110.13389 (2021)。[52] Z. Cai, N. Vasconcelos, 《Cascade R-CNN：深入探索高质量目标检测》，载于：计算机视觉与模式识别会议，2018年，第6154–6162页。[53] G.-S. Xia, X. Bai, J. Ding, Z. Zhu, S. Belongie, J. Luo, M. Datcu, M. Pelillo, L. Zhang, 《DOTA：用于航拍图像目标检测的大规模数据集》，载于：计算机视觉与模式识别会议，2018年，第3974–3983页。[54] X. Yu, Y. Gong, N. Jiang, Q. Ye, Z. Han, 《用于微小行人检测的尺度匹配》，载于：WACV，2020年，第1257–1265页。[55] K. He, X. Zhang, S. Ren, J. Sun, 《用于图像识别的深度残差学习》，载于：计算机视觉与模式识别会议，2016年，第770–778页。[56] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, S. Zagoruyko, 《基于Transformer的端到端目标检测》，载于：ECCV，2020年，第213–229页。[57] L. Dinh, J. Sohl-Dickstein, S. Bengio, 《使用Real NVP进行密度估计》，载于：国际学习表征会议，2017年。[58] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, 《用于密集目标检测的焦点损失》，载于：ICCV，2017年，第2980–2988页。[59] B. Zhu, J. Wang, Z. Jiang, F. Zong, S. Liu, Z. Li, J. Sun, 《AutoAssign：用于密集目标检测的可微标签分配》，arXiv预印本 arXiv:2007.03496 (2020)。[60] C. Xu, J. Wang, W. Yang, H. Yu, L. Yu, G.-S. Xia, 《航拍图像中的微小目标检测：归一化沃瑟斯坦距离与新基准》，《ISPRS摄影测量与遥感杂志》190 (2022) 79–93。

[61] C. Xu, J. Ding, J. Wang, W. Yang, H. Yu, L. Yu, G.-S. Xia, 《面向定向微小目标检测的动态由粗到细学习》，发表于：计算机视觉与模式识别会议，2023年，第 7318–7328 页。[62] C. Chen, Z. Huang, C. Zou, M. Zhu, K. Ji, J. Liu, J. Chen, H. Chen, C. Shen, 《Hieratok：多尺度视觉分词器改进图像重建与生成》，arXiv 预印本 arXiv:2509.23736（2025年）。[63] Y. Shi, X. Guo, W. Yin, M. Jia, Q. Zhang, X. Hu, W. Liu, X. Wang, 《二维高斯与视觉分词器相遇》，arXiv 预印本 arXiv:2508.13515（2025年）。