贾小云, 刘颜荦, 翁佳顺. 基于 RV1126 的人脸检测轻量化算法的改进[J]. 智能计算机与应用,2025,15(9):192-197. DOI: 10.20169/j.issn.2095-2163.250929

基于 RV1126 的人脸检测轻量化算法的改进

贾小云, 刘颜荦, 翁佳顺
(陕西科技大学 电子信息与人工智能学院, 西安 710021)

摘 要: 采用 RV1126 芯片作为硬件平台,针对人脸检测模型较大、检测准确率较低的问题提出了一种改进 RetinaFace 的人脸检测算法,实现了模型部署在嵌入式设备下快速框人脸位置的准确输出。在 RetinaFace 模型的基础上,算法使用轻量化骨干结构 MobileNetV3 网络结构作为特征提取网络,并通过注意力机制 CBAM 加强特征提取和降低特征图通道数来保证模型检测准确率和提高泛化能力。仿真实验结果表明,模型在体积由 104.4 MB 减小到 8.29 MB 的同时,准确率仅由 94.1%降低到 93.0%,验证了算法的有效性。

关键词: RV1126 芯片; 模型部署; 人脸检测; CBAM; 轻量化
中图分类号: TP391 **文献标志码**: A **文章编号**: 2095-2163(2025)09-0192-06

Improvement of the lightweight algorithm of face detection based on RV1126

JIA Xiaoyun, LIU Yanluo, WENG Jiashun
(School of Electronic Information and Artificial Intelligence, Shaanxi University of Science & Technology, Xi'an 710021, China)

Abstract: Using the RV1126 chip as the hardware platform, an improved RetinaFace face detection algorithm is proposed to solve the problem of large face detection model and low detection accuracy, which realizes the accurate output of the face position of the fast frame when the model is deployed in the embedded device. On the basis of the RetinaFace model, the algorithm uses the lightweight backbone structure MobileNetV3 network structure as the feature extraction network, strengthens the feature extraction and reduces the number of feature map channels through the attention mechanism CBAM to ensure the detection accuracy of the model and improve the generalization ability. The experimental results show that while the volume of the model is reduced from 104.4 MB to 8.29 MB, the accuracy is only reduced from 94.1% to 93.5%, which verifies the effectiveness of the algorithm.

Key words: RV1126 chip; model deployment; face detection; CBAM; lightweighting

0 引 言

目前随着国内物联网的发展,其技术应用已逐渐融入了人们的生活,从食堂刷卡到出入门禁,而万物互联的应用场景将深刻改变人们的生活方式。其中,嵌入式人脸识别应用广泛。人脸识别作为公共空间环境中的生物特征识别技术,可被用于城市智慧安防、智慧校园的门禁或考勤人脸识别监测,具有高安全可靠性、便捷性的优势[1]。但是仍存在神经网络过于复杂,参数量过大等问题,导致部署到嵌入式设备上比较困难。RetinaFace 作为目前获得广泛认可的高性能人脸检测网络之一,采用了多分支的架构,同时执行人脸分类、边界框回归和关键点回归等任务[2-4]。因而基于改进 RetinaFace 深度学习算法,框选需识别的对象目标候选区域,可以对各种条件下的人脸进行像素级别定位,推动人员管理的智能化发展。同时,边缘计算设备的应用相较于服务器有着更低的价格、更低的能耗和更好的应用环境,因此一个好的边缘计算设备对于人脸识别的发展有着不可忽略的作用。RV1126 作为一款由国内瑞芯微(Rockchip)公司研发的边缘计算平台,不仅支持 NPU 可以提供高效的人工智能算力,而且支持多种深度学习框架和工具,对于各种物联设备提供了高效的嵌入式人工智能解决方案。

本文选择 RetinaFace 作为人脸检测的基础模型,同时使用轻量级主干网络 MobileNetV3 来提取

作者简介: 贾小云(1971—),女,副教授,主要研究方向:智能信息处理,图形图像处理及数据采集,软件体系结构与构建技术。Email: 1036875244@qq.com; 刘颜荦(1999—),男,硕士研究生,主要研究方向:智能信息处理,基于深度学习的图像与视频抠图;翁佳顺(1999—),男,硕士研究生,主要研究方向:智能信息处理,基于深度学习的图像与视频抠图。

收稿日期: 2024-01-06

图像特征。这一方法首先利用锚框提前定义多个参考框，然后对这些参考框进行分类以判断其中是否包含人脸。通过参考框回归，精确计算人脸边界框的位置和尺寸。其次，使用注意力机制 CBAM 和降低特征图通道数来保证模型检测准确率，并提高泛化能力。最后，通过模型转换、模型部署等步骤将其部署到国产开发板 RV1126 上完成总体系统设计。

1　系统总体方案设计

系统以 RV1126 开发板为主控核心，将摄像头、LCD 显示屏等连接到 RV1126 开发板上，通过在上位机训练好模型，再将模型部署到开发板中，从而构建了人脸检测系统。系统总体如图 1 所示。

图 1　系统总体方案设计图

Fig. 1　Overall scheme design of the system

1.1　特征提取模块

RV1126[5] 开发板是基于瑞芯微公司的 RV1126 处理器的硬件开发平台。处理器采用了双核 ARM Cortex-A7 架构提供高性能计算，集成了多种硬件加速器和接口，适用于多种嵌入式系统和应用，包括智能摄像头、机器视觉、工业自动化和物联网等领域。此外，RV1126 还支持 MIPI 接口，使其能够直连传感器，通过 MIPI 信号传输数据。这使得 RV1126 在图像处理和人工智能决策方面表现出色，为各种应用场景提供了高性能和灵活性。其硬件参数见表 1。

表 1　RV1126 硬件配置

Table 1　Hardware configurations of RV1126

硬件	RV1126 配置
CPU	四核 ARM Cortex-A7
NPU	2.0Tops，support INT8/ INT16
DDR	DDR3 1 GB/2 GB
Flash	eMMC 8 GB/16 GB、SD Card 接口
显示	MIPI-DSI 接口，1080P@60 fps
Camera	双 MIPI-CSI 接口
ISP	1 400 万 ISP2.0 with 3 帧 HDR
RTC	外部 RTC
网络	无线 WiFi、4 G 模块接口
声卡	RK809 集成音频解码器
外围接口	Debug、USB、模拟 Speaker 接口

1.2　人脸检测模块

人脸检测模块以摄像头 IMX415 与 LCD 显示屏组成。摄像头采用 IMX415，主要用于对活体人脸的输入，通过 IMX415 将人脸拍摄到显示屏中。摄像头模组由对角线 CMOS 有源像素型图像传感器组成，具有方形像素阵列和 8.46 M 有效像素及镜头 F/NO.2.0，视场角（水平）130 Diag，并且还具有优化的 TV 畸变（<15%）。此外，还包括 VCM 系统以实现自动对焦。该模块通过水平/垂直合并和二次采样实现 10-bit/12-bit 数字输出，用于拍摄高速高清运动画面。实现高灵敏度、低暗电流，并且该模块还具有可变存储时间的电子快门功能。LCD 显示屏采用 5.5 寸 MIPI 接口电容触摸屏模块，MIPI 接口采用 26 pin 和 18 pin 的 FPC 插座，该模块分辨率为 720×1 080 竖屏（60 帧）。

1.3　存储模块

采用外接 SD 卡进行系统、程序和数据的储存。将录入的人脸数据进行保存，当输入人脸时与库中人脸进行对比得出检测结果。SD 卡具有存储容量大、安全保密性强、功耗低、数据传输速度快、体积小等特点，是系统硬件不可或缺的一部分。

1.4　存储模块

供电采用 USB 接口供电，电压为 5 V，电流为 2.5 A。为了避免功率不足可能会导致的各种各样的问题，如宕机、USB 设备不工作等情况，因此需要稳定的电源适配器。

2　RetinaFace 人脸检测算法的轻量化改进

2.1　整体网络结构

为了在嵌入式上实现人脸检测，对 RetinaFace 进行了一定的改进，其改进后的结构如图 2 所示。改进点主要包括：

（1）将特征提取网络替换为轻量化网络 MobileNetV3，减少模型计算量。

（2）将 CBAM 嵌入 MobileNetV3 网络，提高模型运行精度。

（3）在 RetinaFace 中的 SSH 模块，通过减半 out_c 参数来降低特征图的通道数。此外，还引入深度可分离卷积替代 SSH 模块中的标准卷积。这一改变旨在提高卷积操作的效率，使其更轻量化，更适用于各种不同任务需求。

2.2　主干网络设计

2.2.1　MobileNetV3 网络

MobileNetV3 是一种轻量级神经网络，其特点是参数少、计算量小、推理时间短。网络结构见表 2。其中，包括二维卷积操作、逆残差卷积操作、SE 注意力机制、激活函数类 NL 和卷积操作步长 s 等组件[6]。

通过这样的设计,MobileNetV3 网络能够在不同尺度下提取有意义的特征,并以特征金字塔的形式传递给后续处理步骤。如此一来,将有助于网络更全面、准确地捕捉图像中的信息,从而提高模型性能。

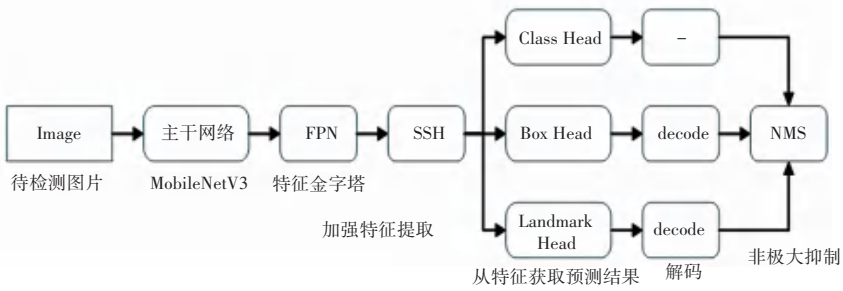

图 2 改进的 RetinaFace 算法网络框架结构

Fig. 2 Network framework structure of improved RetinaFace algorithm

表 2 MobileNetV3

Table 2 MobileNetV3

Input	Operator	CBAM	NL	s
$224^2\times3$	conv2d	-	HS	2
$112^2\times16$	bneck, 3×3	-	RE	1
$112^2\times16$	bneck, 3×3	-	RE	2
$56^2\times24$	bneck, 3×3	-	RE	1
$56^2\times24$	bneck, 3×3	√	RE	2
$28^2\times40$	bneck, 3×3	√	RE	1
$28^2\times40$	bneck, 3×3	√	RE	1
$28^2\times40$	bneck, 3×3	-	HS	2
$14^2\times80$	bneck, 3×3	-	HS	1
$14^2\times80$	bneck, 3×3	-	HS	1
$14^2\times80$	bneck, 3×3	-	HS	1
$14^2\times80$	bneck, 3×3	√	HS	1
$14^2\times112$	bneck, 3×3	√	HS	1
$14^2\times112$	bneck, 3×3	√	HS	2
$7^2\times160$	bneck, 3×3	√	HS	1
$7^2\times160$	bneck, 3×3	√	HS	1
$7^2\times160$	conv2d, 1×1	-	HS	1
$7^2\times960$	pool, 7×7	-	-	1
$1^2\times960$	conv2d 1×1, NBN	-	HS	1
$1^2\times1\ 280$	conv2d 1×1, NBN	-	-	1

2.2.2 CBAM 注意力机制

采用 CBAM 代替骨干网络中的 SENet,使模型在通道、空间维度上都能关注重要特征。CBAM[7]是一种用于深度卷积神经网络的注意力机制。与 SENet 相比,CBAM 不仅加强了 SENet 的特征提取能力,还加入了空间注意力模块。在输入特征图依次经过 CBAM 注意力模块的通道、空间注意力模块后,输入特征图中的通道、空间特征信息得到自适应特征细化。主要目的就是增强网络的感受野(Receptive Field)和对不同特征图的自适应关注,从而提高网络性能。

通道注意力模块用于对不同通道的特征图进行自适应的加权,以突出重要的特征通道并减少对无关信息的依赖。模块核心是计算每个通道的权重,然后将这些权重应用于特征图,从而让网络自动学习。该模块首先将输入的特征图分别经过全局最大池化和全局平均池化操作得到 2 个特征图。接着,将其送入多层感知器压缩其通道数,再将通道数进一步扩张至原通道数。最后,通过 Sigmoid 激活函数生成的通道注意力模块与输入的特征图做乘法操作,生成空间注意力模块需要的输入特征。

空间注意力模块用于对特征图的不同空间位置进行自适应的加权,以突出重要的空间区域。即通过计算每个位置的权重,并将这些权重应用于特征图,以使网络集中关注重要的区域。该模块将通道注意力模块的输出作为其输入,随后,经过一个标准卷积层、卷积核大小为 7×7,用于生成空间注意力模块所需的特征图。这一过程有助于将通道关注信息与空间特征图结合,以提高网络性能,进而加强对不同空间位置的自适应关注。

2.3 轻量化人脸检测模块

本文还对 RetinaFace 中的轻量化人脸检测(SSH)模块进行了一些优化。其中,将 SSH 模块中的输出通道数 out_c 减少为 out_c/2。这一调整有助于减少通道中的冗余信息,降低了网络的复杂性,并在一定程度上提高了网络的泛化能力。此外,为了进一步优化性能,本文还引入了深度可分离卷积[8]来替代原 SSH 模块中的标准卷积,即采用 Conv_dw 代替 Conv_bn。这一改进可以提高模型的效率和计算速度,使其更适合实时人脸检测应用。检测模块 SSH 结构如图 3 所示。

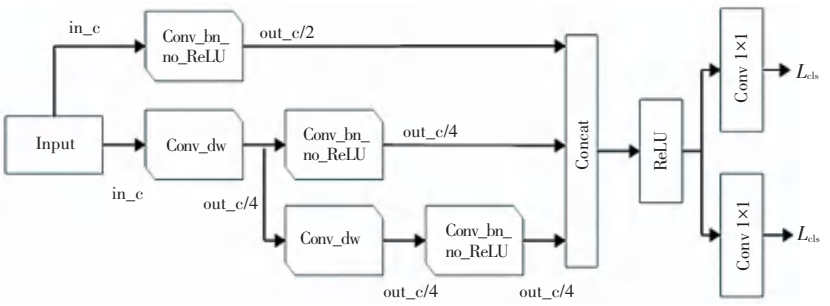

图 3　人脸检测模块 SSH 结构

Fig. 3　SSH structure of the face detection module

3　实验设计和结果分析

3.1　数据集

实验中采用 WIDER FACE[9] 官方数据集。该数据集包括 32 203 张图像和 393 703 个人脸框的大规模人脸检测数据集。这些图像涵盖了各种尺度、遮挡和光照变化，呈现了复杂多样的场景。数据集被划分为 3 个子集：训练集占 40%、验证集占 10% 和测试集占 50%。在 WIDER FACE 数据集中，人脸检测的难度通过基于 EdgeBox 的检测率进行评估，分为容易、中等和困难三个不同难度级别。这种多层次的难度分类有助于评估算法在面对不同复杂度场景时的性能表现。

3.2　模型训练

改进后的算法在 Windows 10 操作系统下，使用 Python 3.7 进行编程。实现过程中采用了 PyTorch 作为深度学习框架，并利用 CUDA 11.7 进行 GPU 加速。硬件方面，主要使用了 GeForce RTX 3060 GPU 型号，以充分利用其并行计算能力，提高算法的运算速度和性能。

3.3　模型部署

模型部署[10] 主要流程如图 4 所示。在 PC 端安装 Python 版本的工具链（rknn－toolkit），并将 PyTorch 模型转换为 ONNX 格式后，将. onnx 导出为. rknn 模型。该步骤主要是验证训练的模型是否可以成功导出、卷积算子是否支持、是否可以量化等功能。接着，在 PC 端配置 C 版本工具链（rknpu），将 Python 版本的工程改写为 C++推理工程，利用 arm－32 bit 的工具链将其编译为可执行文件。再将得到的 RKNN 模型和可执行文件，以及 RK 一些必要的依赖库放入芯片中运行。

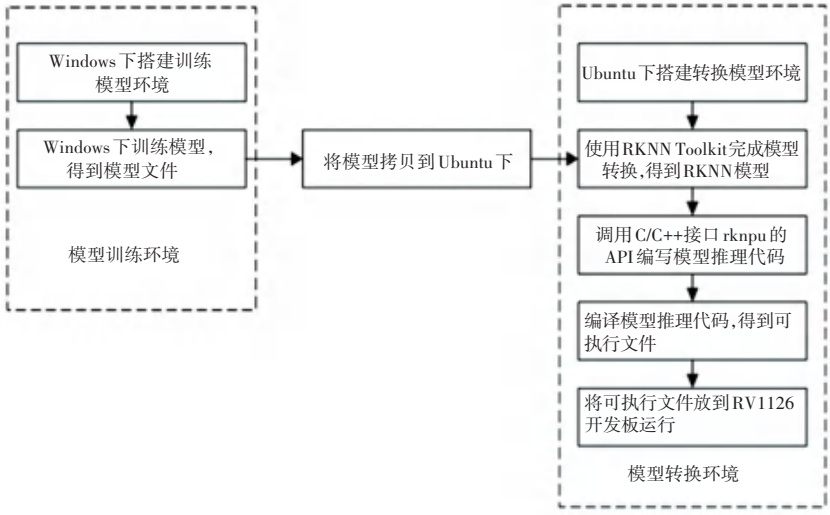

图 4　模型部署流程

Fig. 4　Deployment process of the model

3.4　实验结果分析

为了验证改进后的人脸检测算法的优化效果，本文从算法的检测性能以及模型大小两个方面进行评估。以 WIDER FACE 验证集作为实验测试集。采用平均精确率（AP）和模型大小（Size）来评价改进后算法的检测效果。其中，AP 的值反映单一目标的检测效果，其计算公式为：

$$AP = \int_0^1 p(r)\,dr \tag{1}$$

其中，$p(r)$ 表示真正率和召回率的映射关系。真正率(p) 和召回率(r) 的计算公式为：

$$p = \frac{TP}{TP + FP} \tag{2}$$

$$r = \frac{TP}{TP + FN} \tag{3}$$

其中，TP(True Positive)表示预测结果为正样本，实际也为正样本，即正样本被正确识别的数量；FP(False Positive)表示预测结果为正样本、实际为负样本，即误报的负样本数量；FN(False Negative)表示预测结果为负样本、实际为负样本，即负样本被正确识别的数量。改进后的算法与一些常见的人脸检测网络及 Retinaface 采用 ResNet50[11] 和 MobileNet0.25[12] 进行测试对比见表 3。

表 3 人脸检测性能对比

Table 3 Comparison of face detection performance

算法	AP/%	Size/MB
Faster R-CNN	91.0	1 126.40
SSD	92.5	114.20
RetinaFace(ResetNet50)	94.1	104.40
Retinaface(MobileNet0.25)	90.7	1.71
本文算法	93.0	8.29

由表 3 可见，当 ResNet50 为主干网络时，此时的模型准确率达到很高的水平，超过了一些主流的人脸检测网络，但是其模型大小为 104.40 MB，体积过于庞大不利于嵌入式部署。当 MobileNet0.25 为主干网络时，其模型大小仅为 1.71 MB，但准确率低于常见人脸检测网络。本文研发算法的模型准确率仅次于 ResNet50 为主干网络，但是模型大小却为 8.05 MB，能够满足在嵌入式设备部署的条件，证明了改进后的人脸检测方法的有效性。其实现效果如图 5 所示。由图 5 可见设备能很好地检测到人脸范围，并且将手机照片用设备检测时，检测失败。

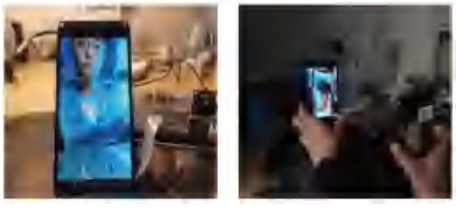

图 5 人脸检测模型部署效果

Fig. 5 Effect of face detection model deployment

3.5 消融实验

为进一步证明改进后的特征提取网络和改进的 SSH 模块对模型的性能提升的作用，进行消融实验：分别将 MobileNetV3、CBAM 和 SSH 进行替换，测试不同模型在测试集上的 AP 和 Size，实验结果见表 4。

表 4 消融实验结果对比

Table 4 Comparison of ablation experimental results

ID	MobileNetV3	CBAM	SSH	AP/%	Size/MB
1				94.1	104.40
2	√			90.7	8.05
3			√	94.6	102.00
4	√	√		92.1	8.29
5	√		√	91.4	8.23
6	√	√	√	93.0	8.17

由表 4 可见，改进前的模型准确率虽然不低、达到 94.1%，但是缺点是模型体积太大，对于嵌入式设备来说不适合部署。改进后的模型在加入 MobileNetV3 作为主干网络时模型准确度由 94.1% 下降到 90.7%，但模型体积大小从 104.40 MB 下降到 8.05 MB，对于模型来说进行了极大轻量化，适合嵌入式设备条件，证明加入 MobileNetV3 后的有效性。在加入 CBAM 模块后，模型体积由 8.05 MB 上升至 8.29 MB，但相比原模型还是下降很多，且模型准确率又提升至 92.1%。加入改进 SSH 模块后，模型复杂程度降低，并且泛化能力提升，因此模型的准确率和大小均有改善，满足模型部署。

4 结束语

针对嵌入式 RV1126 设备的人脸检测工作，本文改进了 RetinaFace 人脸检测算法，采用 MobileNetV3 作为特征提取网络降低模型的参数数量。同时，引入了 CBAM 注意力机制和轻量化 SSH 模块，以提高检测准确性并进一步提升模型的检测效率。根据实验结果与原模型分析，改进后的 RetinaFace 算法不仅满足了轻量化的需求，而且提高了检测准确率，使其更适应不同条件下的应用。未来研究将继续改进网络结构，采用更多的数据集进行训练，并将其部署到不同嵌入式设备中，提升在不同环境下人脸检测的性能和效率。

参考文献

[1] 赵锋. 基于 ZYNQ 的高精度轻量化人脸识别技术与应用研究[D]. 太原：中北大学，2023.

[2] 王鹏，尹勇，宋策. 基于改进 RetinaFace 和 YOLOv4 的船舶驾驶员吸烟和打电话行为检测[J]. 上海海事大学学报，2022，43(4)：44-50.

[3] DENG Jiakang, GUO Jia, VERVERAS E, et al. RetinaFace: Single-shot multi-level face localisation in the wild[C]//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2020: 5202-5211.

[4] SUN Miao, CAO Yingjie, CHIANG P Y. Energy-aware Retinaface: A power efficient edge-computing SOC for face detector in 40 nm [C]//Proceedings of 2021 IEEE 14th International Conference on ASIC (ASICON). Piscataway, NJ: IEEE, 2021: 1-4.

[5] ZHU Jiayi, ZHOU Yi, JIANG Daozhong, et al. Infrared image Target detection system based on RV1126[C]//Proceedings of Seventh Asia Pacific Conference on Optics Manufacture and 2021 International Forum of Young Scientists on Advanced Optical Manufacturing. San Francisco, CA: SPIE, 2022: 1216650.

[6] HOWARD A, SANDLER M, CHU C, et al. Searching for MobileNetV3[C]//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway, NJ: IEEE, 2019: 1314-1324.

[7] ZHAO Zhengxuan, CHEN Kaixu, YAMANE S. CBAM-Unet++: Easier to find the target with the attention module "CBAM"[C]//Proceedings of 2021 IEEE 10th Global Conference on Consumer Electronics (GCCE). Piscataway, NJ: IEEE, 2021: 655-657.

[8] 胡清礼，胡建强，余小燕，等. 基于深度可分离卷积的心音自动分类[J]. 计算机应用与软件，2023，40(6)：154-159.

[9] YANG Shuo, LUO Ping, LOY C C, et al. WIDER FACE: A face detection benchmark [C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2016: 5525-5533.

[10] 高宗斌，崔永杰，李凯. 基于 T-YOLO-LITE 树干检测的模型部署方法[J]. 计算机应用与软件，2021，38(2)：132-139.

[11] ZAKARIA N, MOHAMED F, ABDELGHANI R, et al. VGG16, ResNet-50, and GoogLeNet deep learning architecture for breathing sound classification: A comparative study[C]//Proceedings of 2021 International Conference on Artificial Intelligence for Cyber Security Systems and Privacy (AI-CSP). Piscataway, NJ: IEEE, 2021: 1-6.

[12] JAHNAVI K, SANDEEP N S, DEEPIKA R, et al. Detection of COVID-19 using ResNet50, VGG19, MobileNet, and Forecasting; using Logistic Regression, Prophet, and SEIRD model[C]//Proceedings of 2023 7th International Conference on Computing Methodologies and Communication (ICCMC). Erode, India: WikiCFP, 2023: 1538-1542.