基于通道方差引导特征融合的AI生成图像检测方法

罗世淦,唐雨行

(广西警察学院 信息技术学院,广西 南宁 530028)

摘 要: 针对扩散模型与生成对抗网络产生的图像在底层像素和高层语义上具有显著差异,导致现有检测方法在面对跨模型、高保真生成图像时泛化性能不足、检测率较低的问题,提出一种基于通道方差引导的语义—像素特征融合检测方法(CVGF)。该方法构建双分支特征提取网络:像素级分支利用高通滤波机制捕获底层伪造指纹,语义级分支引入预训练CLIP模型提取高层语义异常。其核心创新在于设计了通道方差引导融合模块,通过对异构特征通道的统计方差进行建模,引入极轻量化的可学习基础权重(W_{base}),实现对存在关键伪造线索通道的自适应加权与特征高效整合。在ProGAN、CycleGAN、GauGAN、VQDM和DALL·E 2这5类主流生成模型上的实验表明,CVGF展现出优异的跨域检测性能。其中,其在CycleGAN上的准确率达到88.57%,相较于基线方法提升了36.71个百分点;在VQDM上达到79.17%,提升了16.52个百分点;在DALL·E 2上的虚假图像检测率达到80.30%,显著优于所对比的基线方法。CVGF通过统计先验引导特征权重分配,有效整合了像素层级与语义层级的判别证据,增强了检测器对未见生成模型的识别能力与鲁棒性。该方法验证了语义证据与像素证据的强互补性,为应对日益复杂的AI生成图像检测领域提供了有效的技术路径。

关键词: AI生成图像检测;特征融合;通道方差;语义—像素双分支;跨模型泛化

DOI: 10. 11907/rjdk. 261223 扫描二维码阅读全文:

中图分类号: TP391.41 **文献标识码:** A **文章编号:** 1672-7800(2026)007-0084-10

AI-Generated Image Detection via Channel Variance Guided Feature Fusion

LUO Shigan, TANG Yuxing

(*School of Information Technology*, *Guangxi Police College*, *Nanning* 530028, *China*)

Abstract: Cross-model generalization remains a critical bottleneck in synthetic image detection, primarily because images produced by diffusion models and GANs diverge significantly at both pixel-level statistics and high-level semantics. To bridge this gap, we introduce the Channel Variance-Guided Semantic-Pixel Feature Fusion Detection Method (CVGF). Our framework adopts a dual-branch architecture for feature extraction. The pixel branch isolates low-level forgery fingerprints using a high-pass filtering strategy, while the semantic branch leverages a frozen CLIP model to capture inconsistencies in high-level content. The central novelty of our work is a channel variance-guided fusion module. Rather than simply concatenating features, this module models the statistical dispersion across heterogeneous feature channels and employs an extremely lightweight learnable base weight (W_{base}) to adaptively amplify channels that carry critical tampering evidence. We benchmarked CVGF against five representative generative paradigms: ProGAN, CycleGAN, GauGAN, VQDM, and DALL·E 2. The method demonstrates strong cross-domain resilience. Notably, CVGF achieves 88.57% accuracy on CycleGAN (a gain of 36.71 percentage points over baselines), 79.17% on VQDM (up 16.52 points), and a detection rate of 80.30% on DALL·E 2. These results underscore that allocating feature weights based on statistical priors successfully unifies pixel-domain artifacts with semantic inconsistencies, bolstering detector robustness against previously unseen generative architectures.This method verifies the strong complementarity between semantic evidence and pixel evidence, providing an effective technical path for addressing the increasingly complex field of AI-generated image detection.

Key Words: AI-generated image detection; feature fusion; channel variance; semantic-pixel dual branch; cross-model generalization

收稿日期: 2026-05-11

基金项目: 广西重点研发计划项目(桂科AB25069434);广西高校中青年教师科研基础能力提升项目(2025KY0874)

作者简介: 罗世淦(2006-),男,广西警察学院信息技术学院学生,研究方向为信息安全;唐雨行(1992-),女,硕士,广西警察学院信息技术学院高级工程师,研究方向为人工智能与数字取证。本文通讯作者:唐雨行。

0　引言

近年来，以扩散模型（DALL·E 2[1]、VQDM[2]）和生成对抗网络（ProGAN[3]、CycleGAN[4]、GauGAN[5]）为代表的图像生成技术取得了显著进展，其合成图像的视觉质量已达到人眼难以分辨的水平。与此同时，此类技术被滥用于虚假新闻传播、身份伪造、司法证据篡改等场景的风险日益加剧，对公共舆论安全、个人隐私保护乃至司法公正构成了现实威胁，使得AI生成图像的自动化检测成为一项兼具学术价值与社会意义的迫切需求。

现有检测方法大致可分为两类。以PatchCraft[6]为代表的像素特征方法通过分析底层统计差异捕获生成指纹，在主流GAN上效果显著，但面对像素高保真的生成模型时性能急剧下滑，在CycleGAN上的准确率仅为51.86%。另一类以UnivFD[7]为代表，依赖CLIP预训练模型的语义特征进行判别，对GauGAN、CycleGAN等语义合成型模型表现良好，却在扩散模型上几乎完全失效，其在DALL·E 2上的虚假图像检测率仅1.70%。

两类方法的失效模式恰好互补，说明伪造痕迹同时分布在像素细节与语义结构两个层面，单一层级的特征难以应对跨域检测需求。本文目标是设计一种能够同时利用两个层级证据且在不同生成模型间保持泛化能力的轻量化检测方法。受此启发，本文构建了语义—像素双分支检测架构，在保留像素分支底层指纹捕获能力的同时，引入语义分支弥补其在高层判别上的不足。初步实验也印证了这一思路的可行性，即仅将两类特征直接相加，可将ProGAN上的准确率从95.03%提升至99.80%。

然而，直接相加并非最优融合方式。语义特征与像素特征处于不同的高维空间，768个通道中并非每一个都承载有效的判别信息——部分通道可能仅包含噪声，直接相加时这些通道会稀释有用信号。因此，如何识别并聚焦于真正具有判别力的通道，是提升融合效果的关键。常见的注意力机制或可学习加权策略往往引入较多参数，在小数据集上容易过拟合。对此，本文提出了通道方差引导融合方法（Channel Variance Guided Fusion，CVGF）。其基本思路是：在一个训练批次中，语义—像素差异特征的某个通道若呈现较高的方差，则意味着该通道上两类证据的分歧较大，更可能反映出生成模型在像素保真与语义一致之间的顾此失彼，因而应赋予更高的融合权重。从信息论的角度看，方差较大的通道携带了更多关于样本类别的信息——如果某通道的值对所有样本几乎相同，则它对分类几乎无贡献；波动越大，该通道与类别标签之间的潜在互信息越高。整个机制仅需$2C$个可学习参数（C=768），保持了较低的参数开销。

本文主要贡献包括3个方面：①提出语义—像素双分支检测架构，引入CLIP ViT-L/14图像编码器作为语义分支，实验验证了像素证据与语义证据在AI生成图像检测中的互补性；②设计了基于通道方差统计先验的轻量化融合模块CVGF，以批次通道方差引导权重分配，在参数量仅为$2C$的条件下取得了优于可学习加权融合的效果；③在ProGAN、CycleGAN、GauGAN、VQDM与DALL·E 2这5类生成模型上开展了系统实验，其中在CycleGAN上较PatchCraft提升了36.71个百分点，在VQDM上提升了16.52个百分点，这得益于双分支骨干冻结与极低的可训练参数量（30.72 K），模型训练效率较高，约5轮即可基本稳定。

1　相关工作

1.1　底层特征检测方法

这类方法的出发点是：生成模型在卷积上采样过程中会在像素层面留下可追溯的“指纹”，具体表现为空间域的纹理异常或频域中的周期性模式，可作为区分真伪图像的依据。

早期工作主要聚焦于频域分析。Frank等[8]发现，GAN生成图像在频域存在可检测伪影，使用简单分类器即可区分真伪；Wang等[9]进一步指出不同GAN架构的生成结果在频域上共享某些伪影模式，具备作为通用检测依据的潜力。Zhang等[10]和Durall等[11]分别从高频分量异常的角度提出基于频谱的检测框架。Chen等[12]则尝试将RGB信息与频域特征融合用于人脸伪造检测。Zhou等[13]进一步提出像素级映射预处理，主动破坏检测器依赖的语义捷径，迫使模型关注可泛化的高频伪影。然而，Ricker等[14]系统评估了现有GAN检测方法对扩散模型的适用性，发现扩散模型生成的图像缺乏GAN特有的网格状频域伪影，且倾向于低估高频信息，从而产生更少的可检测伪影，导致传统基于GAN的频域特征难以直接迁移。

PatchCraft在这类方法中与本文关系最为密切，它通过图像块重组破坏语义结构，并比较纹理丰富区与贫乏区之间的像素相关性差异[6]。其假设是生成图像因上采样引入的周期性伪影，会在两类区域之间呈现异常的相关性模式。该方法在主流GAN上效果优异，但由于在预处理阶段丢弃了语义信息，故在像素高保真生成模型（如CycleGAN、GauGAN）上可用的判别线索十分有限。

1.2　语义特征检测方法

与底层特征方法不同，这类方法利用预训练视觉—语言模型的高层表示以判别真伪，认为生成图像在对象属性、空间关系或场景一致性等语义层面可能偏离真实图像分布。

Ojha等[7]提出的UnivFD是其中最具代表性的工作，它以CLIP ViT-L/14图像编码器为特征提取器，借助CLIP在大规模图文对上习得的跨模态语义表达进行判别。由于ViT-L/14的自注意力机制能够建模全局结构关系，该方法在GauGAN、CycleGAN等语义合成型生成模型上取得了

不错的效果。NS-Net 提示 CLIP 视觉特征中的高层语义信息会干扰判别，通过 NULL-Space 投影解耦语义信息，在 40 种生成模型上检测准确率提升 7.4 个百分点[15]。GCS-Net 通过 LoRA 轻量微调 CLIP 并融合 LAB/HSV 颜色空间特征，在 20 个数据集上平均准确率达 84.57%，验证了增强语义特征的有效性[16]。Cozzolino 等[17]优化了基于 CLIP 的特征提取策略，提升了多个测试集上的检测准确率。Corvi 等[18]专门考察了扩散模型的检测问题，发现其在传统频域特征上与真实图像差异甚小，需要引入多层级特征才能有效判别。

然而，语义特征方法在扩散模型场景下普遍表现欠佳。究其原因在于，扩散模型采用渐进去噪策略，生成结果在语义层面与真实图像高度一致，二者的差异主要体现在像素级的高频纹理上——而这恰恰是 CLIP 等以语义对齐为训练目标的模型所不敏感的。

1.3 多特征融合方法

鉴于单一类型特征难以覆盖所有生成模型，有研究者尝试融合多种证据以提升跨域泛化能力。Yao 等[19]提出的 MSAFNet 在空间金字塔池化基础上引入通道注意力，对不同尺度特征的权重进行动态分配[20-21]。Lin 等[22]改进了 Xception 网络，通过 CBAM 模块与自注意力机制在通道和空间维度上分别对特征进行增强。Zhai 等[23]将局部注意力引入空间—频率联合建模，提供了一种频域与语义信息协同利用的思路。此外，Wang 等[24]从轻量化 VIT 与 CNN 的结构关联出发，通过将轻量级 ViT 的高效架构设计逐步融入 MobileNetV3，提出了纯 CNN 模型 RepViT，在移动端实现了高精度与低延迟的平衡。这些研究验证了多元特征整合的价值，但其融合机制大多依赖额外的可学习权重或交互矩阵，在小数据集场景下有一定的过拟合风险。

综合来看，底层特征方法擅长检测具有明显频域/纹理伪影的 GAN 生成图像，面对像素高保真模型和扩散模型时则力不从心；语义特征方法在语义合成型生成模型上表现突出，但对扩散模型普遍容易漏检；融合方法虽然拓宽了特征覆盖面，但代价往往是引入更多可学习参数。本文做法是保留像素分支的底层指纹捕获能力，同时引入语义分支，并以通道方差统计先验引导二者的轻量化融合，在不显著增加参数开销的前提下实现跨域检测性能提升。

2 本文方法

2.1 整体框架

AI 生成图像检测可形式化为二分类问题：给定图像 I，模型输出判定 $y \in \{0, 1\}$，其中 0 为真实图像，1 为生成图像。本文目标是学习映射 $f: I \to y$，使其能够泛化到多种生成模型（包括 GAN 与扩散模型）。

本文模型整体架构如图 1 所示，采用语义—像素双分支设计。像素级分支 $\mathcal{B}_p$ 负责从图像底层统计中提取生成指纹特征 $F_p \in \mathbb{R}^C$，具体通过纹理分区采样、高通滤波与差分增强实现；语义级分支 $\mathcal{B}_s$ 则利用预训练的 CLIP ViT-L/14 图像编码器提取高层语义特征 $F_s \in \mathbb{R}^C$，其中 $C = 768$。两组特征送入 CVGF 融合模块（见图 2）进行整合，输出融合特征 $F_{fusion} \in \mathbb{R}^C$，最后经分类头映射为判定概率 $P \in [0, 1]$。其中，像素级分支的特征提取部分沿用 PatchCraft[6] 的已有设计，语义级分支直接采用预训练的 CLIP ViT-L/14 图像编码器，二者均不涉及架构改动。本文主要贡献集中在特征对齐策略（线性投影与 LayerNorm）和 CVGF 融合模块设计上。

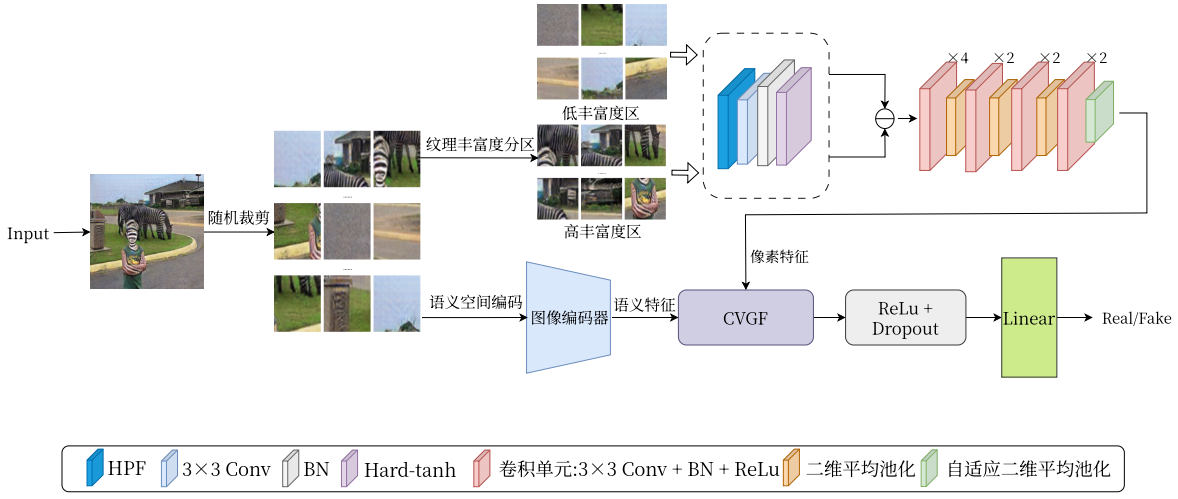


Fig. 1 Overall architecture of the model

图 1 模型整体架构

2.2 像素级分支

像素级分支沿用 PatchCraft 的设计思路，从图像底层统计中提取生成指纹[6]。该思路具体步骤如下：

首先，对输入图像 I 随机裁剪 $N = 192$ 个尺寸为 $M \times M$（$M = 32$）的图像块作为候选集。为量化各块的纹理复杂度，引入局部总变分描述子 l_{div}，定义为相邻像素在水平、

垂直及两个对角方向上灰度差的绝对值之和，如式(1)所示。

$$l_{div}=\sum_{i=1}^{M}\sum_{j=1}^{M-1}\left(\left|x_{i,j}-x_{i,j+1}\right|\right)+\sum_{i=1}^{M-1}\sum_{j=1}^{M}\left(\left|x_{i,j}-x_{i+1,j}\right|\right)+\sum_{i=1}^{M-1}\sum_{j=1}^{M-1}\left(\left|x_{i,j}-x_{i+1,j+1}\right|\right)+\sum_{i=1}^{M-1}\sum_{j=1}^{M-1}\left(\left|x_{i+1,j}-x_{i,j+1}\right|\right) \tag{1}$$

其中，$x_{i,j}$表示图像块(i,j)位置的灰度值。将192个候选块按l_{div}降序排列后，取前64块拼接为纹理丰富区R，后64块拼接为纹理贫乏区P，中间64块舍弃。之所以只取两端，是为了拉大富纹理区与贫纹理区之间的统计反差，使后续差分操作更容易捕获异常。

随后，R与P分别送入特征提取算子$H(\cdot)$。该算子由预设参数的高通滤波器、3 × 3卷积层、批归一化层与Hard-tanh激活函数级联组成，用于滤除低频语义成分、放大高频指纹响应。通过计算差分$D_{pix}=H(R)-H(P)$，可抵消图像内容本身的干扰，锁定两类区域在底层统计上的分布偏差。

D_{pix}随后进入特征增强模块，其由多个3 × 3卷积单元(含批归一化与ReLU)和自适应池化层堆叠而成，汇聚为32维的原始像素特征$\hat{F}_p\in\mathbb{R}^{32}$。至此，基于PatchCraft的像素特征提取完毕。

由于语义分支的输出维度为768，这里通过一层线性投影将$\hat{F}_p$升维至$C=768$，再经LayerNorm归一化，得到最终参与融合的像素特征$F_p\in\mathbb{R}^C$。训练时，该分支的骨干网络(沿用PatchCraft的部分)使用预训练权重并保持冻结，只有线性投影层和LayerNorm参与梯度更新。

2.3　语义级分支

语义级分支采用CLIP ViT-L/14的图像编码器作为特征提取器。选择ViT-L/14而非更小的ViT-B/16，是因为前者提供了768维的特征空间，语义容量更大，能学习到更丰富的语义特征信息。

给定输入图像I，编码器输出[CLS]token最后一层的表示，即原始语义特征$\hat{F}_s\in\mathbb{R}^C$。为了让语义特征与像素特征处于可比的分布空间，$\hat{F}_s$经LayerNorm进行归一化后得到$F_s\in\mathbb{R}^C$。这里不引入额外的线性投影，目的是尽量保留CLIP在预训练阶段形成的语义结构。编码器全部参数在训练过程中保持冻结，该分支仅有LayerNorm参与梯度更新。

2.4　CVGF融合模块

CVGF是本文提出的核心模块，其作用是将像素分支与语义分支的输出进行自适应融合。CVGF架构如图2所示。

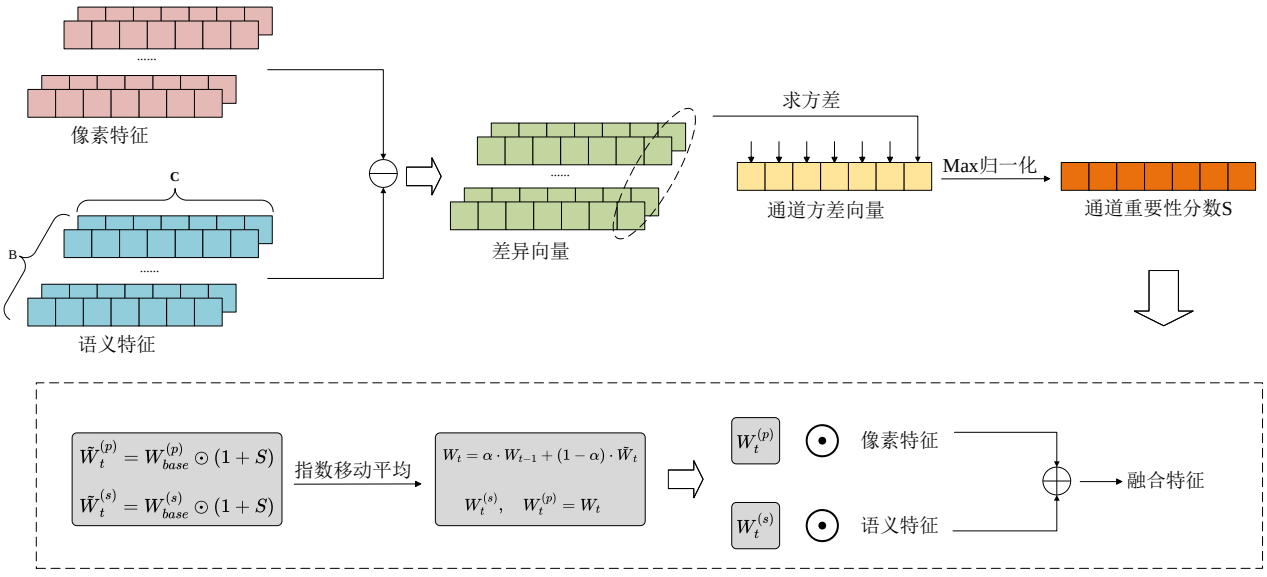


Fig. 2　CVGF architecture

图2　CVGF架构

2.4.1　设计动机

融合的关键问题在于：两个分支的C维特征中，并非每个通道都同等重要，不同通道承载的判别信息类型和强度各有差异。直接相加时，那些不含有效判别信号的通道会稀释有用信息，因此需要一种机制以识别并聚焦于真正具有区分力的通道。CVGF的做法是利用通道方差完成这一筛选。具体而言，定义差异向量$D=F_s-F_p$，在一个训练批次的B个样本上观察D的每个通道：如果某通道的方差较大，说明语义证据与像素证据在该维度上的分歧在不同样本间波动剧烈，这往往对应着生成模型在像素保真与语义一致之间顾此失彼的维度，判别价值较高；反之，方差小的通道意味着两类证据的差异在样本间比较稳定，区分力有限。因此，CVGF对高方差通道赋予更高的融合权重。从特征选择的角度看，这与方差准则的思路一致；从信息论的角度看，方差较大的通道在差异向量D上携带了更多关于样本类别的信息——如果某通道对所有样本几乎取同一值，它对分类几乎无贡献，波动越大，与类别标签之间的潜在互信息越高。

为了直观验证上述设计思路，本文统计了训练过程中20个批次的通道方差分布。如图3所示，在768个通道中，方差落在[0，0.1)区间的通道平均占比超过75%，方差≥0.3的通道通常不超过50个，方差≥0.5的通道仅为个位数。这一分布表明，判别性信息确实集中在少数高方差通道上，大部分通道对真伪区分的贡献有限。这从实验层面验证了CVGF通过方差加权聚焦高判别力通道的合理性，也说明了直接相加融合会引入大量低方差通道的噪声从而稀释有效信号。

2.4.2 权重计算

像素分支与语义分支的输出$F_p, F_s \in \mathbb{R}^C$已处于同一特征空间。定义差异向量$D = F_s - F_p \in \mathbb{R}^C$，对于包含$B$个样本的训练批次，第$j$个样本的差异向量记为$D^{(j)}$，其第$i$个通道的取值为$D_i^{(j)}$。该通道在批次内的均值和方差分别如式(2)、式(3)所示。

$$\mu_i = \frac{1}{B}\sum_{j=1}^{B} D_i^{(j)} \tag{2}$$

$$\sigma_i^2 = \frac{1}{B}\sum_{j=1}^{B}(D_i^{(j)} - \mu_i)^2 \tag{3}$$

由此得到通道方差向量$\sigma^2 \in \mathbb{R}^C$。对其做最大值归一化，得到各通道的重要性分数$S \in [0, 1]^C$，如式(4)所示。

$$S_i = \frac{\sigma_i^2}{\max\limits_j \sigma_j^2 + \epsilon} \tag{4}$$

其中，ϵ为防止除零的极小常数。CVGF为语义分支和像素分支各维护一组C维可学习基础权重$W_{base}^{(s)}$，$W_{base}^{(p)} \in \mathbb{R}^C$，代表各通道在无方差调节时的默认贡献。结合重要性分数S，当前批次的瞬时权重如式(5)、式(6)所示。

$$\tilde{W}_t^{(s)} = W_{base}^{(s)} \odot (1 + S) \tag{5}$$

$$\tilde{W}_t^{(p)} = W_{base}^{(p)} \odot (1 + S) \tag{6}$$

其中，$\odot$表示逐元素乘法。乘以(1+S)意味着：所有通道至少保留基础权重，高方差通道在此基础上获得额外增益。

由于单个批次的方差估计不够稳定，直接使用$\tilde{W}_t$会导致权重在训练过程中震荡，为此引入指数移动平均进行平滑，如式(7)所示。

$$W_t = \alpha \cdot W_{t-1} + (1 - \alpha) \cdot \tilde{W}_t \tag{7}$$

其中，t为训练步，$\alpha = 0.9$。式(7)对语义分支权重$W_t^{(s)}$与像素分支权重$W_t^{(p)}$分别执行。经过多步累积后，W_t能够较好地反映训练数据上的整体通道方差分布，而非某一批次的偶然波动。

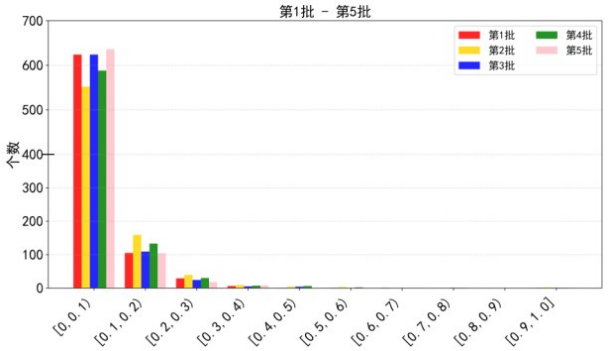
2.4.3 特征融合与复杂度分析

利用累积权重对两个分支的特征进行逐通道加权求和，如式(8)所示。

$$F_{fusion} = F_s \odot W_t^{(s)} + F_p \odot W_t^{(p)} \tag{8}$$

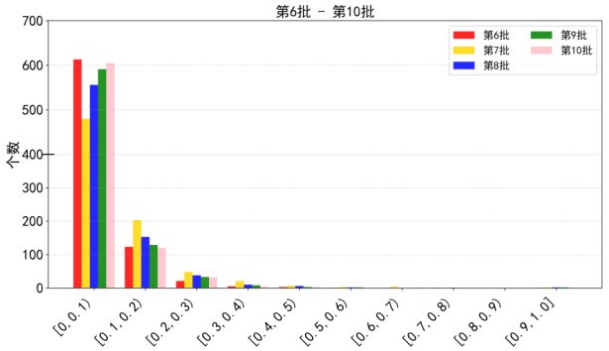
得到的$F_{fusion} \in \mathbb{R}^C$即为分类头的输入。

在参数量上，CVGF的可学习参数只有$W_{base}^{(s)}$与$W_{base}^{(p)}$两



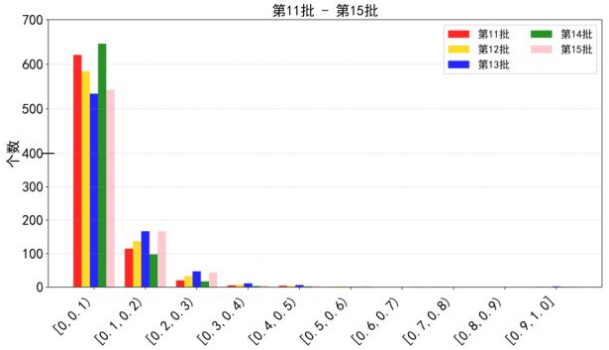
(a) Distribution of batches 1 to 5

(a) 第1–5批次分布情况



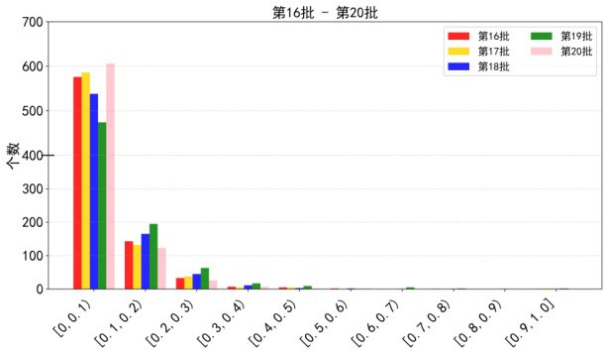
(b) Distribution of batches 6 to 10

(b) 第6–10批次分布情况



(c) Distribution of batches 11 to 15

(c) 第11–15批次分布情况



(d) Distribution of batches 16 to 20

(d) 第16–20批次分布情况

Fig. 3 Channel variance distribution

图3 通道方差分布

组向量，共$2C = 1\,536$个标量。方差权重完全由批次统计在线计算，不引入额外参数。计算复杂度为$\mathcal{O}(B \cdot C)$，远低于传统注意力融合的$\mathcal{O}(C^2)$。

3　实验设置

3.1　数据集

训练集与测试集的构建遵循PatchCraft的实验协议。本文从原始训练集中提取了8 000张包含人物（Person）类别的图像构建训练集（真实图像与虚假图像各4 000张）。测试集涵盖5类生成模型：ProGAN与训练集同源，用于验证源域检测性能；CycleGAN、GauGAN、VQDM和DALL·E 2共4类均来自不同的生成架构，用于考察跨域泛化能力。各测试集规模如表1所示。

Table 1　Test set size

表 1　测试集规模　　（张）

测试集	真实图像	生成图像	总计
ProGAN	4 000	4 000	8 000
GauGAN	5 000	5 000	10 000
CycleGAN	1 321	1 321	2 642
DALL·E 2	1 000	1 000	2 000
VQDM	6 000	6 000	12 000
总计	17 321	17 321	34 642

3.2　评价指标

本文采用3类准确率指标：总体准确率Acc、真实图像准确率Acc_r与虚假图像准确率Acc_f，定义分别如式（9）—式（11）所示。

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{9}$$

$$Acc_r = \frac{TN}{TN + FP} \tag{10}$$

$$Acc_f = \frac{TP}{TP + FN} \tag{11}$$

其中，TP与TN分别为被正确识别的生成图像与真实图像数量，FP与FN分别表示被误判为生成图像的真实图像与被误判为真实图像的生成图像数量。Acc_r偏低意味着真实图像被大量误报，Acc_f偏低则意味着生成图像被大量漏检，二者的平衡性直接关系到检测器在实际部署中的可用性。

3.3　基线方法

本文选取4种方法作为对比基线。①PatchCraft[6]：通过图像块重组与局部方差采样分析纹理统计差异，捕获底层生成指纹；②UnivFD[7]：以CLIP ViT-L/14图像编码器为特征提取器，利用跨模态语义表达进行判别；③CNNSpot[9]：基于ResNet的RGB域通用检测方法，在CNN特征上训练二分类器；④Gram[25]：利用Gram矩阵刻画特征图的通道相关性，捕获生成图像的纹理风格偏差。4者分别代表像素特征、语义特征、通用CNN特征和风格特征4种技术路线。

3.4　实现细节

实验在配备NVIDIA GeForce RTX 4080 SUPER的工作站上完成，软件环境为PyTorch。训练采用AdamW优化器，学习率1×10^{-4}，权重衰减0.025，批次大小64，最多训练20轮（Gram[25]为80轮）。像素分支与语义分支的骨干网络均使用预训练权重并冻结，只有特征对齐层、CVGF的基础权重和分类头参与更新。

4　实验与结果分析

4.1　主实验

表2列出了CVGF与4种基线方法在5类测试集上的完整对比结果（依次为Acc、Acc_r、Acc_f）。图4以热力图展示了各方法的总体准确率Acc、真实图像准确率Acc_r和虚假图像准确率Acc_f。

Table 2　Cross-domain detection performance comparison

表 2　跨域检测性能比较　　（%）

方法	ProGAN	CycleGAN	GauGAN	DALL·E 2	VQDM
PatchCraft	95.03 / 99.40 / 90.65	51.86 / 73.58 / 30.13	84.95 / 74.90 / 95.00	65.25 / 78.00 / 52.50	62.65 / 81.88 / 43.42
UnivFD	82.99 / 99.80 / 66.17	84.75 / **98.86** / 70.63	**86.95** / **98.50** / 75.40	50.80 / **99.90** / 1.70	61.10 / **99.30** / 22.90
CNNSpot	83.05 / 97.50 / 68.60	67.41 / 86.37 / 48.45	58.71 / 98.16 / 19.26	49.50 / 97.60 / 1.40	52.92 / 95.88 / 9.95
Gram	59.56 / 81.33 / 37.80	52.99 / 81.61 / 24.38	50.46 / 80.48 / 20.44	48.35 / 78.50 / 18.20	51.30 / 78.35 / 24.25
CVGF	**99.81** / **99.95** / **99.68**	**88.57** / 79.71 / **97.43**	65.80 / 34.18 / **97.42**	**74.50** / 68.70 / **80.30**	**79.17** / 81.30 / **77.03**

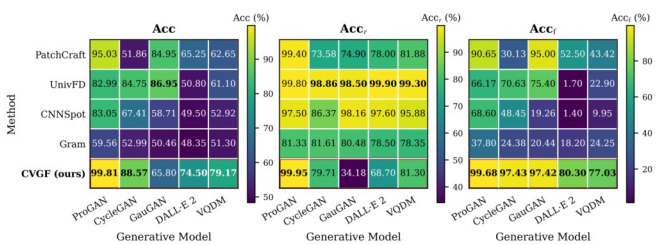


Fig. 4　Comparison of baseline experiment results

图 4　基线实验结果比较

在训练域同源的ProGAN上，CVGF的Acc达到99.81%，Acc_r与Acc_f也超过99%，说明双分支架构在源域上已具备充分的判别能力。更关键的对比体现在跨域场景：CycleGAN上CVGF的Acc为88.57%，较PatchCraft的51.86%提升了36.71个百分点；VQDM上为79.17%，较62.65%提升了16.52个百分点。这两类测试集恰好是像素特征方法的薄弱域，语义分支的引入明显填补了底层特征的盲区。在扩散模型方面，UnivFD在DALL·E 2上的Acc_f仅为1.70%，在VQDM上也只有22.90%，几乎丧失了对伪

造图像的识别能力；CVGF 则在 DALL·E 2 上将 Acc_f 提升至 80.30%，在 VQDM 上达到 77.03%，表明双分支的互补作用在扩散模型场景下同样成立。

然而，CVGF 在 GauGAN 上的 Acc_r 仅为 34.18%，拖累了该测试集的总体准确率。这一问题与语义分支的过敏感有关，将在本文 4.6 节专门讨论。

4.2 消融实验

为了隔离 CVGF 融合机制本身的贡献，本文设计了两种简化变体：直接相加（$F_{fusion} = F_s + F_p$，不引入权重）和可学习加权（$F_{fusion} = w_s F_s + w_p F_p$，$w_s$，$w_p$ 为全局可学习标量）。两种变体的网络结构和训练配置与 CVGF 完全一致，只替换融合模块。实验结果如表 3 所示。

Table 3 Comparison of fusion strategies

表 3 融合策略比较

测试集	直接相加			可学习加权			CVGF		
	Acc	Acc_r	Acc_f	Acc	Acc_r	Acc_f	Acc	Acc_r	Acc_f
ProGAN	99.80	**100.00**	99.60	**99.81**	99.95	**99.68**	**99.81**	99.95	**99.68**
GauGAN	65.56	33.44	**97.68**	65.60	33.90	97.30	**65.80**	**34.18**	97.42
CycleGAN	87.89	79.26	96.52	87.66	78.12	97.20	**88.57**	**79.71**	**97.43**
DALL·E 2	73.65	67.80	79.50	74.40	67.60	81.20	**74.50**	**68.70**	**80.30**
VQDM	78.93	79.60	78.25	79.09	79.82	**78.37**	**79.17**	**81.30**	77.03

从 Acc 来看，CVGF 在 5 个测试集上均取得最优或并列最优，其中在 CycleGAN 上较可学习加权提升 0.91 个百分点，GauGAN、DALL·E 2、VQDM 上的提升幅度为 0.08~0.20 个百分点。提升虽然不大，但方向一致，说明通道方差先验提供的引导是稳定的。在细分指标上，CVGF 的 Acc_r 在多数测试集上领先，而可学习加权的 Acc_f 在个别测试集上略高（如 DALL·E 2 上高 0.90 个百分点），反映出两种策略在判定倾向上的细微差异，即 CVGF 对真实图像更稳健，可学习加权对生成图像稍敏感。总体而言，CVGF 用 $2C$ = 1 536 个参数换取了通道级的自适应调节，比两个标量的全局加权提供了更细粒度的控制。此外，3 种策略在 ProGAN 上的表现几乎没有差异（Acc 均在 99.80% 以上）。其原因在于，训练域与测试域同源时，两个分支各自的判别力已经足够，融合策略的差异只在跨域场景下才会体现出来。

4.3 效率分析

表 4 比较了 CVGF 与各基线方法在训练效率、推理速度和可学习参数量上的差异。所有实验均在同一硬件环境（RTX 4080 SUPER）上完成。

在参数量上，CVGF 的可学习参数仅为 30.72 K，远低于需要训练完整分类器骨干的 CNNSpot（23.51 M）和 Gram（11.73 M），也显著低于 PatchCraft（119.17 K）。这得益于 CVGF 冻结了两个分支的骨干网络，实际参与训练的只有线性投影层、归一化层、CVGF 融合模块的基础权重（1 536 个参数）和分类头。

Table 4 Efficiency analysis experiment results

表 4 效率分析实验结果

方法	训练中平均每轮用时/s	平均推理每张图像用时/ms	可学习参数量/K
UnivFD	74.063 0	11.480 7	0.769
PatchCraft	193.577 8	21.819 7	119.17
CNNSpot	42.348 5	3.421 2	23 510
Gram	37.441 6	3.423 6	11 730
CVGF	227.020 5	34.982 9	30.72

在推理速度上，CVGF 的单张推理耗时为 34.98 ms，高于其他方法。这是因为 CVGF 需要同时完成像素分支和语义分支的前向传播，两个骨干网络的计算量叠加构成了推理的主要瓶颈。需要指出的是，CVGF 融合模块本身引入的计算开销极小（仅涉及方差统计和逐元素乘法，复杂度为 $\mathcal{O}(B \cdot C)$），推理时间的增长完全来自双分支架构的骨干网络，而非融合机制。在训练效率方面，如图 5 所示，CVGF 的训练损失约在第 5 轮趋于稳定，而 PatchCraft 和 CNNSpot 分别需要约 15 轮和 10 轮。考虑到 CVGF 冻结了两个骨干网络、实际可训练参数仅 30.72 K，其总训练时间（约 227×5≈1 135 s）仍显著低于 PatchCraft（约 194×15≈2 910 s）。

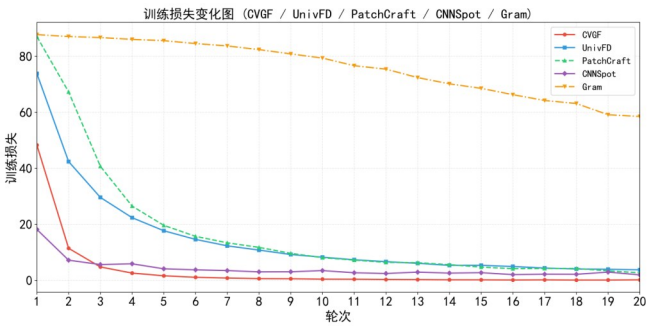


Fig. 5 Comparison of training loss variation

图 5 训练损失变化比较

4.4 鲁棒性实验

为评估 CVGF 在实际部署场景中的稳定性，本文进一步测试了高斯模糊（核大小为 7 × 7，标准差 σ = 1.5）与 JPEG 压缩（质量因子 Q = 75）两种常见扰动对 CVGF 的影响，结果如表 5（高斯模糊）和表 6（JPEG 压缩）所示。

Table 5 Detection results under Gaussian blur perturbation

表 5 高斯模糊扰动下的检测结果 （%）

测试集	Acc	Acc_r	Acc_f
ProGAN	99.10	99.40	98.80
GauGAN	67.18	41.88	92.48
CycleGAN	81.07	76.38	85.77
DALL·E 2	75.55	65.80	85.30
VQDM	70.29	74.22	66.37

Table 6 Detection results under JPEG compression perturbation

表 6 JPEG 压缩扰动下的检测结果 （%）

测试集	Acc	Acc_r	Acc_f
ProGAN	96.71	97.33	96.10
GauGAN	70.91	88.28	53.54
CycleGAN	71.04	91.67	50.42
DALL·E 2	64.20	98.70	29.70
VQDM	64.30	90.55	38.05

高斯模糊对CVGF的影响总体可控。ProGAN和DALL·E 2上的Acc变化不到1.1个百分点，CycleGAN和VQDM下降较为明显（分别为7.50和8.88个百分点），但绝对准确率仍在70%以上。一个值得关注的现象是GauGAN上的Acc在模糊后反而从65.80%上升到67.18%，可能是因为模糊操作在一定程度上平滑了CLIP语义分支对真实场景的过度响应，使Acc_r有所回升。

JPEG压缩的影响则严重得多。Acc_f在CycleGAN、VQDM、DALL·E 2上分别降至50.42%、38.05%、29.70%，生成图像的漏检率大幅上升。其原因在于，JPEG的DCT量化破坏了高频纹理痕迹，而像素分支的判别依据恰恰是高频伪影，即高频信号被压平后，像素分支基本失去了提取生成指纹的能力。与此形成对比的是，Acc_r在压缩后反而普遍提升（多数超过88%），这或许是因为压缩的同时抹平了真实图像中多样化的传感器噪声，使其特征分布更集中，更接近模型训练时见过的分布。

两组实验说明，CVGF在轻度扰动下尚能保持稳定，但面对破坏高频信号的压缩类扰动，像素分支的失效会直接拖累整体检测能力。针对这一问题，一种直接的改进思路是在训练阶段对像素分支的输入施加随机质量因子（$Q\in[50,95]$）的JPEG压缩增强，迫使像素分支在高频信号被部分破坏的条件下学习，从而发掘压缩后仍然存留的中低频判别线索。该策略不改变模型结构，实施成本较低，将作为后续研究的优先方向。

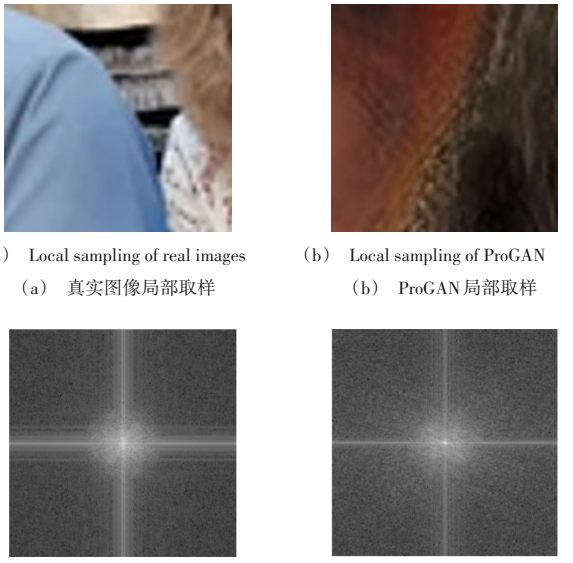
4.5　定性分析

本文选取若干典型案例，分析CVGF在不同生成模型上的检测行为。

在ProGAN测试集上，CVGF的3项指标均接近满分。观察具体样本可以发现，ProGAN生成的图像在细节区域存在明显的周期性高频纹理伪影，像素分支的总变分描述子恰好擅长捕获这类特征。同时，这些图像在宏观结构上与真实图像差异不大，语义分支给出的置信度适中，不会压制像素分支的判定。两类证据在此场景下形成了稳定的互补关系。从图6的频谱对比可以直观看到ProGAN样本在频域上的周期性异常。

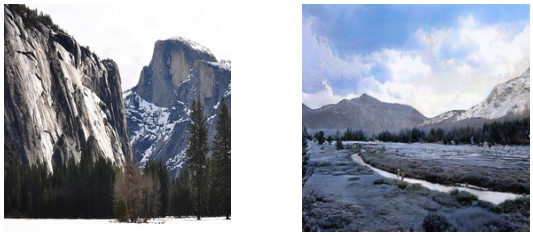
CycleGAN的情况则不同，由于它采用无配对翻译策略，像素层面的纹理迁移非常逼真，PatchCraft单独使用时Acc仅为51.86%。而风格迁移难以做到语义上完全无破绽，因为跨域迁移可能在光照方向、对象结构或场景布局上留下细微不协调，这些对于CLIP的语义编码器而言是可感知的。CVGF正是借助语义分支捕获了这类高层异常，将整体Acc拉升至88.57%。CycleGAN生成图像与真实图像比较如图7所示。

最后看一个失败案例：DALL·E 2图像经JPEG压缩后，CVGF的Acc_f从80.30%骤降至29.70%。扩散模型生成的图像在语义上与真实图像高度一致，像素分支是CVGF在该测试集上的主要判别手段。一旦JPEG量化破坏了高频信息，像素分支便无法提供有效证据。而语义分支对扩散模型本就存在漏检倾向（UnivFD在该测试集上的Acc_f仅为1.70%），此时也无法互补，两条线索同时失效，导致检测大面积失败。针对这一问题，一种可能的改进方向是在训练阶段对像素分支的输入施加随机JPEG压缩，迫使像素分支在高频信号被部分破坏的条件下学习，从而发掘压缩后仍然存留的中低频判别线索。



（a） Local sampling of real images　（b） Local sampling of ProGAN
（a） 真实图像局部取样　（b） ProGAN局部取样
（c） Spectrogram of real sample image　（d） Spectrogram of ProGAN image
（c） 真实样图频谱图　（d） ProGAN样图频谱图

Fig. 6　Local comparison of ProGAN and real images
图6　ProGAN与真实图像局部比较



（a） Real image　（b） CycleGAN image
（a） 真实图像　（b） CycleGAN图像

Fig. 7　Comparison of CycleGAN and real images
图7　CycleGAN与真实图像比较

4.6　GauGAN失效机理分析与阈值调整

如4.1节所示，CVGF在GauGAN上的Acc_r仅为34.18%，在5个测试集中最低。本文尝试分析原因并探索改进思路。

从鲁棒性实验可知，图像经JPEG压缩破坏高频细节后，Acc_r反而有所提升。这一现象与以下假设一致：GauGAN使用SPADE机制，以语义分割图引导像素生成，输出图像在空间布局、光照和纹理上都达到了很高的自然度。面对这类图像，依赖高频伪影进行判别的像素分支因找不到熟悉的纹理证据，反而被真实图像中的高频噪声所

误导，造成真实图像误判；而语义分支关注的低频全局结构特征才是区分 GauGAN 真伪图像的有效线索。当 JPEG 压缩抹除高频噪声后，像素分支的干扰被削弱，语义分支在融合中的相对贡献增大，从而提高了模型对真实图像的判别准确率。图 8 展示了几张被 CVGF 误判的 GauGAN 风格的真实图像。

Fig. 8 Realistic images in GauGAN style

图 8 GauGAN 风格的真实图像

为了验证这一判断，本文对比了语义分支和像素分支单独使用的结果（UnivFD 和 PatchCraft 的结果）。UnivFD 单独使用时在 GauGAN 上的 Acc_r 达到 98.50%，说明语义分支本身具备良好的真实图像识别能力；PatchCraft 单独使用时 Acc_r 为 74.90%，虽不及语义分支但也尚可。然而，融合后 Acc_r 骤降至 34.18%，远低于任何一个单分支的表现，说明问题出在融合环节，即 CVGF 的方差权重机制在 GauGAN 域上未能正确分配两类证据的贡献，反而放大了二者之间的分歧，导致决策偏向保守。

这种“偏保守”意味着模型对真实图像输出的伪造概率普遍偏高，使得不少真实样本越过了 0.5 的默认判定边界。一个直接的缓解办法是提高阈值：要求伪造概率超过 0.7 才判定为生成图像。两种阈值下的结果如表 7 所示。

阈值提升后，Acc_r 在多个测试集上明显回升：在 GauGAN 上提升 6.14 个百分点，在 CycleGAN 上提升 5.83 个百分点，在 DALL·E 2 上提升 9.20 百分点，符合“决策偏保守”的预期。其代价是 Acc_f 有所下降，尤其是 VQDM 和 DALL·E 2 分别下降了 6.06 和 6.00 个百分点，原因在于阈值提高放宽了对生成图像的判定门槛，漏检随之增加。这本质上是 Acc_r 与 Acc_f 之间的再平衡，具体如何取舍要看应用场景：偏重低误报的内容审核可以用更高阈值，偏重高召回的伪造检测则应降低阈值。

Table 7 Comparison of detection results under different decision thresholds

表 7 不同决策阈值下的检测结果比较 （%）

测试集		阈值=0.5			阈值=0.7	
	Acc	Acc_r	Acc_f	*Acc*	Acc_r	Acc_f
ProGAN	99.81	99.95	99.68	99.60	99.98	99.23
GauGAN	65.80	34.18	97.42	68.12	40.32	95.92
CycleGAN	88.57	79.71	97.43	90.01	85.54	94.47
DALL·E 2	74.50	68.70	80.30	76.10	77.90	74.30
VQDM	79.17	81.30	77.03	78.60	86.23	70.97

当然，阈值调整只是事后补救，并没有消除 CVGF 在 GauGAN 域上的融合权重偏差。更根本的改进需要从模型层面入手，比如：引入门控机制，根据输入图像的域特征动态调节各分支的融合权重；或者在训练阶段引入少量 GauGAN 域的标注数据对融合权重进行微调，使方差权重机制适应该域的特征分布。

5 结语

本文针对 AI 生成图像检测中单一特征空间跨域泛化不足的问题，提出了通道方差引导的语义—像素融合检测方法（CVGF）。该方法在语义—像素双分支架构的基础上，以批次通道方差作为融合权重的先验信号，仅需 $2C$ 个可学习参数即可完成异构特征的自适应整合，在多个测试集上取得了优于可学习加权融合的检测效果。

在 ProGAN、CycleGAN、GauGAN、VQDM 和 DALL·E 2 这 5 类生成模型上的实验表明，像素证据与语义证据确实具有互补性，通道方差引导机制在总体准确率上稳定优于直接相加和可学习加权两种基线策略。鲁棒性实验也暴露了 CVGF 的薄弱环节，即 JPEG 压缩会大幅破坏像素分支依赖的高频判别信号。此外，GauGAN 场景下融合权重机制未能正确分配两类证据的贡献，导致 Acc_r 偏低，尽管阈值调整可以部分缓解，但并未从根本上解决问题。

本文研究还存在一些局限，并据此提出后续改进的具体思路。针对 GauGAN 域上的融合权重偏差，可引入门控机制，根据输入图像的域特征动态调节各分支的融合权重，或通过少量域内标注数据对方差权重进行微调。针对 JPEG 压缩下像素分支鲁棒性不足的问题，可在训练阶段对像素分支的输入施加随机质量因子（$Q\in[50,95]$）的 JPEG 压缩增强，迫使其学习压缩后仍存留的中低频判别线索。此外，当前训练集仅覆盖 Person 类别且规模有限，后续有必要扩展类别覆盖并将框架拓展至局部篡改检测任务。这些改进方向的共同目标，是使 CVGF 从当前实验验证阶段走向更可靠的实际部署。

参考文献：

[1] RAMESH A, DHARIWAL P, NICHOL A, et al. Hierarchical text-conditional image generation with clip latents [DB/OL]. https://arxiv. org/abs/2204. 06125.

[2] GU S, CHEN D, BAO J, et al. Vector quantized diffusion model for text-to-

image synthesis[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition，2022：10696-10706.

[3] KARRAS T，AILA T，LAINE S，et al. Progressive growing of GANs for improved quality，stability，and variation [DB/OL]. https://arxiv. org/abs/1710. 10196.

[4] ZHU J Y，PARK T，ISOLA P，et al. Unpaired image-to-image translation using cycle-consistent adversarial networks[C]//Proceedings of the IEEE International Conference on Computer Vision，2017：2223-2232.

[5] PARK T，LIU M Y，WANG T C，et al. Semantic image synthesis with spatially-adaptive normalization [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition，2019：2337-2346.

[6] ZHONG N，XU Y，LI S，et al. PatchCraft：exploring texture patch for efficient AI-generated image detection [DB/OL]. https://arxiv. org/abs/2311. 12397.

[7] OJHA U，LI Y，LEE Y J. Towards universal fake image detectors that generalize across generative models[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition，2023：24480-24489.

[8] FRANK J，EISENHOFER T，SCHÖNHERR L，et al. Leveraging frequency analysis for deep fake image recognition[C]//International Conference on Machine Learning，2020：3247-3258.

[9] WANG S Y，WANG O，ZHANG R，et al. CNN-generated images are surprisingly easy to spot... for now[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition，2020：8695-8704.

[10] ZHANG X，KARAMAN S，CHANG S F. Detecting and simulating artifacts in GAN fake images[C]//2019 IEEE International Workshop on Information Forensics and Security(WIFS)，2019：1-6.

[11] DURALL R，KEUPER M，KEUPER J. Watch your up-convolution：Cnn based generative deep neural networks are failing to reproduce spectral distributions[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition，2020：7890-7899.

[12] CHEN S，YAO T，CHEN Y，et al. Local relation learning for face forgery detection[C]//Proceedings of the AAAI Conference on Artificial Intelligence，2021：1081-1088.

[13] ZHOU C，WANG J，LI Y，et al. Beyond semantic features：pixel-level mapping for generalized AI-generated image detection[C]//Proceedings of the AAAI Conference on Artificial Intelligence，2026：36101-36109.

[14] RICKER J，DAMM S，HOLZ T，et al. Towards the detection of diffusion model deepfakes[DB/OL]. https://arxiv. org/abs/2210. 14571.

[15] YAN J，WANG F，JIANG W，et al. NS-Net：decoupling CLIP semantic information through NULL-space for generalizable AI-generated image detection[DB/OL]. https://arxiv. org/abs/2508. 01248.

[16] XU H，CHEN Z，ZHANG H，et al. GCS-Net：a universal AI-generated visual content detection method based on CLIP[J]. Knowledge-Based Systems，2025，323：113806.

[17] COZZOLINO D，POGGI G，CORVI R，et al. Raising the bar of ai-generated image detection with clip[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition，2024：4356-4366.

[18] CORVI R，COZZOLINO D，ZINGARINI G，et al. On the detection of synthetic images generated by diffusion models[C]//Rhodes：2023 IEEE International Conference on Acoustics，Speech and Signal Processing (ICASSP 2023)，2023.

[19] YAO L，NIU S，LIAO K，et al. MSAFNet：multi-scale self-adaptive feature fusion network for AI-generated image detection[J]. Measurement Science and Technology，2025，36(9)：095402.

[20] HU J，SHEN L，SUN G. Squeeze-and-excitation networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition，2018：7132-7141.

[21] WOO S，PARK J，LEE J Y，et al. CBAM：convolutional block attention module[C]//Proceedings of the European Conference on Computer Vision (ECCV)，2018：3-19.

[22] LIN H，LUO W，WEI K，et al. Improved Xception with dual attention mechanism and feature fusion for face forgery detection[C]//2022 4th International Conference on Data Intelligence and Security(ICDIS)，2022：208-212.

[23] ZHAI T，LU K，LI J，et al. Learning spatial-frequency interaction for generalizable deepfake detection[J]. IET Image Processing，2024，18(14)：4666-4679.

[24] WANG A，CHEN H，LIN Z，et al. RepViT：revisiting mobile CNN from vit perspective[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition，2024：15909-15920.

[25] LIU Z，QI X，TORR P H S. Global texture enhancement for fake face detection in the wild[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition，2020：8060-8069.

（责任编辑：孙　娟）