arXiv:2602.06460v1 [cs.SD] 6 Feb 2026

使用更少通道的肌电转语音

Injune Hwang[⋆] *Jaejun Lee*[⋆] *Kyogu Lee*[⋆†‡]

[⋆]大韩民国首尔国立大学情报与信息系
[†]大韩民国首尔国立大学人工智能研究所
[‡]大韩民国首尔国立大学人工智能跨学科项目

摘要

表面肌电图（肌电图）是无声语音接口的一种很有前景的模态，但其有效性很大程度上取决于传感器布置和通道可用性。在本研究中，我们探讨了单个及组合肌电通道对语音重建性能的贡献。研究结果表明，尽管某些肌电通道单独提供的信息更多，但利用通道间互补关系的子集能取得最佳性能。我们还分析了通道消融条件下的音素分类准确率，并观察到反映相关肌肉解剖功能的可解释模式。为解决通道缩减导致的性能下降问题，我们使用随机通道丢弃在完整的 8 通道数据上预训练模型，再在通道数缩减后的子集上对模型进行微调。在 4 至 6 通道设置下，微调的表现始终优于从头训练；最佳丢弃策略则因通道数量而异。这些结果表明，通过预训练和通道感知设计，可以缓解传感器缩减造成的性能下降，从而推动轻量、实用的肌电图无声语音系统的发展。

索引词— 肌电转语音、无声语音、表面肌电图（sEMG）、通道缩减

1. 引言

肌电图（EMG）已成为无声语音接口的一种前景广阔的模态，它通过测量面部和颈部肌肉活动来解码语音，而无需依赖声学信号。这种方法为无法通过声带发声的人提供了交流机会。

随着深度学习的兴起，肌电转语音技术取得了显著进展。Gaddy 等人 [1] 推出了首个广泛使用的肌电数据集，并提出了一种单说话人肌电转语音合成系统。在此数据集的基础上，后续研究探索了肌电转文本识别 [2]以及扩展至多说话人合成 [3,4]等方向。

尽管这些研究取得了显著的学术进展，但在实现实际部署之前仍面临若干挑战。尤其是，传统研究依赖覆盖近半张脸的全通道肌电传感器阵列。从易用性和可及性的角度来看，开发轻便、用户友好的肌电转语音设备至关重要。因此，在减少通道数量的同时保持较高的识别性能，是一项关键目标。

本研究使用 Gaddy等人 [1]提出的广泛采用的数据集，系统地考察肌电转语音中的通道效率。我们首先通过每次迭代移除一个通道，进行贪心分析，以考察性能下降情况，并进一步评估子通道组合，以判断各通道的相对重要性。为深入了解这一问题，我们分析音素级性能变化，以估计各个通道对发音的贡献。此外，我们证明，使用减少通道的输入对基于全通道数据预训练的模型进行微调，其效果始终优于从头开始训练。

最后，我们证明，在预训练期间加入随机通道丢弃，能够在减通道场景下进一步提升性能。这一策略还能进一步增强泛化能力，尤其是在有限通道输入上进行微调时。

为便于理解，我们将贡献概述如下：

- 我们通过贪心消除和穷举子集评估，对肌电图通道的重要性进行了全面分析，识别出主导性通道贡献和互补性通道贡献。
- 我们在 7 通道设置下开展了音素级消融分析，以揭示特定通道如何影响不同音素类别（例如擦音、塞音）的识别。
- 我们提出了一种微调策略，在预训练期间加入随机通道丢弃；与从头开始训练相比，该策略在减通道场景下始终能提升性能。

2. 相关工作

2.1. 基于肌电图的语音合成

Gaddy等人 [5],提出了一个广泛使用的肌电数据集，其中包含来自单个说话人的八通道录音，涵盖有声肌电（在可听语音期间采集）和无声肌电（在不可听的发音期间采集）。由于只有有声肌电具有平行声学数据，他们提出了一种训练策略，利用动态时间规整（DTW）损失将无声肌电与相应的有声语音对齐。该方法将这种对齐方式与基于 Transformer 的肌电图编码器相结合，由后者预测梅尔频谱图，再通过预训练声码器将其转换为语音，从而在根据无声肌电合成语音方面取得了良好效果。在此框架基础上，Scheck 等人 [3] 通过采用语音转换框架，将该方法扩展到多说话人设置 [6] ； 该框架使用 HuBERT [7]表征作为编码器目标。尽管凭借其解耦能力，这一扩展方法在多说话人场景中表现出色，但在单说话人设置下，Gaddy 基于梅尔频谱图的方法仍然更胜一筹。

值得注意的是，Gaddy 的数据集已成为无声语音领域的标准基准，并被近期的多项研究采用。例如，Ren 等人 [8]提出了一种基于扩散的肌电转语音

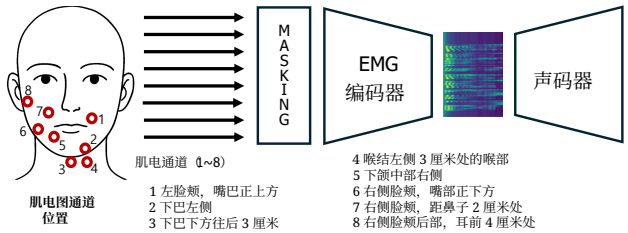


图 1。肌电转语音框架概览。掩蔽模块仅出现在我们的微调变体中。肌电通道位置根据下文所述内容进行示意。

模型，Ullah 等人 [9] 提出了一种专为无声肌电设计的优化训练框架。此外，Wu 等人 [10] 利用该数据集探索了用于个性化静默语音解码的深度提示调优，而 Benster 等人 [11] 则借此研究了肌电与声学特征之间的跨模态表征。这些工作共同凸显了 Gaddy 的数据集对于推动无声语音合成研究仍具有重要价值和实用性。

2.2. 在减少肌电通道条件下进行合成

已有多项研究探讨了通道缩减对肌电转语音性能的影响。Gaddy 等 [1] 研究了随机选取的 35 个四通道子集——70 种可能组合的一半——从而对通道重要性提供了部分见解。Lee 等 [12] 根据通道与发音特征之间的相关性选取了 2 至 4 个通道，并报告了性能提升；然而，仍需开展更广泛的实证验证。尽管这些工作凸显了肌电通道缩减的可行性，但其研究结果范围有限，验证也不够充分。相比之下，我们的工作对通道效率和音素级贡献进行了更系统、更全面的研究。

3. 方法

3.1. 减少 EMG 通道条件下的分析

我们假设，并非所有肌电通道对语音合成的贡献都相同。相反，通道之间可能存在相互关系，这表明它们的贡献具有相互依赖性和冗余性。这意味着，在需要缩减通道时，应优先保留某些通道；或者在通道数量受限时，某些通道子集可能带来更好的性能。为此，我们使用了广泛采用的八通道肌电数据集 [5] 及其配套的合成框架 [13],，该框架在单说话人设置下表现出色。整体框架如图 1 所示。为了评估各肌电通道对合成性能的相对贡献，我们设计了三项互补实验，从不同角度评估通道重要性。

3.1.1. 反向消除

我们研究了逐步减少通道数量的影响，以评估每个通道的相对贡献，并确定性能开始下降的阈值。我们没有穷举评估所有可能的通道组合，而是采用贪心反向消除策略：从完整的八通道集合开始，

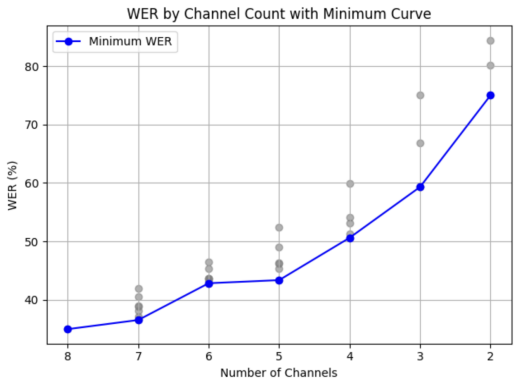


图 2。使用不同数量肌电通道的模型的词错误率。每个点代表一种模型配置，蓝线连接各通道数量下表现最佳的配置。

每次移除一个通道，形成 8 个候选子集，选出性能最佳的子集，再迭代重复这一过程。尽管该方法未考虑通道之间的组合交互或协同效应，但它能快速近似评估各通道的重要性。

3.1.2. 4 通道组合的穷举评估

我们评估了从 8 个可用通道中选取所有可能的四通道组合的性能（共 70 种组合）。这使我们能够评估首次实验中由贪心方法确定的四通道子集的相对性能，并计算每个通道在所有组合中的平均词错误率。通过分析某些通道出现在性能最佳组合中的频率，我们还深入了解了它们的相对重要性及互补关系。这项穷举实验能够与先前的研究直接比较，例如 Gaddy 等人测试的 35 个随机四通道子集 [1]，以及 Lee 等人提出的基于发音动机的四通道选择。 [12]

3.1.3. 七通道变体的音素错误分析 *nts*

我们分析了分别移除八个通道中的每一个时的音素分类性能（即所有七通道子集）。此项分析旨在评估每个通道对不同音素类别识别的影响。为单独考察肌电编码的作用，我们绕过语音合成和自动语音识别阶段，直接根据肌电编码器的输出评估音素预测结果。为深入分析，我们依据语言学和发音特征对音素进行分组：辅音与元音、浊音与清音、按发音方式和发音部位划分的辅音，以及按舌位高低和舌位前后划分的元音。这种分类方式有助于考察各个通道如何促进更精细的语音区别。

3.2. 使用通道丢弃进行微调

为应对通道减少条件下的性能下降，我们使用在全部八个通道上训练的预训练模型权重来初始化模型。目标是通过微调，使该模型适用于通道更少的场景。

在八通道预训练期间，我们采用通道丢弃（通道掩蔽），以提升模型对通道减少

表1。四通道子集中词错误率（%）排名前10。

子集	1	2	3	4	5	6	7	8	WER
1356	1	0	1	0	1	1	0	0	47.2
2357	0	1	1	0	1	0	1	0	47.3
1346	1	0	1	1	0	1	0	0	47.7
1238	1	1	1	0	0	0	0	1	48.3
1235	1	1	1	0	1	0	0	0	48.4
2347	0	1	1	1	0	0	1	0	48.6
1236	1	1	1	0	0	1	0	0	48.8
1345	1	0	1	1	1	0	0	0	49.0
1245	1	1	0	1	1	0	0	0	49.6
2367	0	1	1	0	0	1	1	0	49.6
数量	7	7	9	4	5	4	3	1	

条件的泛化能力（图 1）。通过丢弃来增强鲁棒性是各领域广泛采用的策略 [14, 15]；同样，我们采用通道丢弃，以提升模型对不同通道可用性的泛化能力。对于每条训练语句，我们以固定概率随机掩蔽各个通道，并独立地为每个通道采样。通道丢弃仅在训练期间应用，不在推理时应用。

形式上，设 $\boldsymbol{X} \in \mathbb{R}^{C\times T}$ 为具有 C 个通道和 T 个时间步的输入肌电信号。我们定义丢弃掩码 $\boldsymbol{m} \in \{0,1\}^C$ 如下：

$$m_c \sim \text{Bernoulli}(1-p) \quad \text{for } c = 1, \ldots, C$$

其中 p 为丢弃概率。该掩码在每次前向传播中独立应用于每个通道。掩码后的输入 $\tilde{\boldsymbol{X}}$ 随后通过以下方式计算：

$$\tilde{\boldsymbol{X}} = \boldsymbol{m} \otimes \boldsymbol{X}$$

其中 $\otimes$ 表示掩码 $\boldsymbol{m}$ 沿时间维度的逐通道广播。我们选择了丢弃概率$p \in \{0, 0.125, 0.25\}$，分别对应于使用全部通道，或平均保留8个通道中的7个或6个。选择这些值是为了模拟真实的传感器故障，同时避免可能导致训练不稳定的过度信息损失。

4. 实验

4.1. 数据集与预处理

我们使用 Gaddy 等人发布的公开肌电数据集开展实验 [5],，该数据集已成为静音语音研究中广泛使用的基准。该数据集包含约 20 小时的单一受试者肌电记录，其中包括 16 小时的开放词汇数据和 4 小时的封闭词汇数据。由于本研究聚焦于连续语音场景中的开放集泛化和通道鲁棒性，因此我们排除了封闭词汇部分。肌电信号由放置在面部和颈部的八个表面电极记录（见图 1）。每条语音都配有同步的音频波形，数据集还包含经过时间对齐的音素和词级转写。我们遵循 GitHub 仓库[1], 中提供的标准预处理流程，并使用梅尔频谱图作为声学目标。

4.2. 模型架构与训练目标

我们采用了 Gaddy 等人 [13],提出的模型架构，该架构结合了卷积层和 Transformer 层来进行序列

[1]https://github.com/dgaddy/silent_speech

表 2。包含各通道的所有四通道子集的通道平均词错误率（%）。

Ch.	3	2	1	5	6	4	7	8
WER	51.4	52.3	52.6	52.8	53.1	53.7	53.8	54.8

建模。在编码器输出之上，使用了两个独立的线性投影层：一个用于预测梅尔频谱图，另一个用于对音素进行分类。在训练目标方面，我们遵循 Gaddy 的设置，包括梅尔频谱图重建损失和音素分类损失。使用较少的肌电通道训练模型时，我们保持整个架构不变，仅调整第一个卷积层的输入通道大小，以匹配可用通道的数量。

使用 HiFi-GAN [16], 将预测的梅尔频谱图转换为波形，然后使用 Whisper 自动语音识别模型（中型） [17]进行转录。将生成的转录文本与真实文本进行比较，以计算词错误率（WER）。

所有实验均使用我们的开源代码库。[2]

5. 结果与分析

5.1. 反向消除

我们采用反向消除方法分析通道重要性，每次移除一个通道并测量模型性能。如图 2 所示，通道按以下顺序被消除：通道 6、7、8、5、4 和 1，最后保留通道 2 和 3，因此它们最为关键。逐一移除各通道时，对词错误率的影响（从负面影响最大到最小）依次为：通道 3、2、7、1、5、4、8 和 6。仅使用单个通道时，词错误率超过 100%，因此无法进行有意义的定量比较。有趣的是，一些通道数较少的子集表现优于较大的子集，这表明某些通道可能提供冗余甚至带有噪声的信息，从而阻碍泛化。随着通道数减少，不同通道组合之间的性能差异增大，这表明在少通道设置下，模型对通道选择更加敏感。

5.2. 四通道组合穷举评估

为补充贪心反向消除分析，我们对全部 70 种可能的 4 通道组合进行了穷举评估。如表 1 所示，其中至少有 10 种组合的表现优于贪心方法选出的子集（通道 1、2、3 和 4）。值得注意的是，不包含通道 2 时，通道 5 或 6 常被选作替代通道。同样，排除通道 1 时，通道 7 常出现在前十组合中。这表明某些通道携带互补信息，可根据组合情况相互补偿。根据大致的空间位置，通道 1 位于面部左上方，通道 7 位于右上方，通道 2 位于左下方，通道 5 和 6 位于右下方。这种空间多样性可能有助于捕捉相互重叠或互补的肌肉激活区域。有趣的是，在反向消除中最先被剔除的通道 6，出现在多个表现最佳的组合中。相反，在贪心分析中被认为最关键的通道 2，却未出现在部分排名靠前的组合中。根据所有组合中各通道出现时的平均词错误率

[2]https://github.com/SPJune/SS_通过_通道

表 3。8通道基线和表现最差的单通道消融中各音素类别的错误率。

类别	8 通道音素错误率	最差音素错误率	最关键通道
总音素错误率	16.0	17.1	8
元音	22.6	24.1	7
辅音	20.9	22.7	3
静音	4.5	5.8	8
浊音	21.3	23.2	7
清音	22.4	25.3	3
流音发音方式	12.7	16.9	1
擦音发音方式	19.4	23.2	3
鼻音发音方式	21.0	23.5	7
塞音发音方式	26.4	27.6	7
双唇发音部位	21.6	25.2	8
齿龈发音部位	21.4	23.0	3
唇齿发音部位	17.9	22.5	2
软腭发音部位	18.5	21.5	1
高元音	21.7	28.8	7
中元音	23.7	25.1	7
低元音	22.6	22.7	3
前元音	24.4	28.8	7
央元音	24.9	25.7	8
后元音	17.7	20.7	6
圆唇元音	17.8	20.7	6
非圆唇元音	23.6	24.8	7

各通道按以下顺序排名：3、2、5、1、6、4、7 和 8（表 2）。对前十子集进行频率分析发现，通道 3 出现频率最高，其次是通道 1 和 2，再接着是 5、4、6、7 和 8。最后， [1]中报告的表现最佳子集（通道 1、2、3、8）在我们的结果中排名第

4 [12] 1, , 而 2, 3, 5 5 （ ）中的基于发音动作的子集排名第。以上发现凸显了穷举子集评估的价值：它能够识别贪心或启发式方法无法发现的最优配置。

5.3. 7 通道变体的音素错误分析

为考察各个肌电图通道对不同音素类别的贡献，我们测量了所有 7 通道子集的音素错误率（PER），结果汇总于表 3。为确保统计可靠性，我们排除了样本少于 3,000 个（约占语料库的 2%）的音素类别，包括塞擦音、滑音、后齿龈音、声门音、唇软腭音和腭音。移除通道 8 后，整体音素错误率增幅最大。该通道位于咬肌后部上方，可能捕捉到下颌运动和闭唇动作。移除该通道会降低双唇音、静音片段和央元音的识别表现。通道 7 位于颧大肌附近，会影响高前元音识别，表明它在塑造脸颊和口腔轮廓方面发挥作用。通道 3 位于胸骨舌骨肌附近，会影响清擦音和低元音，这与其参与舌部和喉部定位相一致。通道 2（降口角肌）会影响唇齿音素，而通道 6（口轮匝肌）会影响圆唇元音。通道 1 对流音和软腭音音素识别的贡献尚不明确。重要的是，移除某些通道后性能没有下降，并不一定意味着这些通道无关紧要；这可能表明存在信号冗余，或其他传感器起到了补偿作用。有趣的是，在合成评估中被认为不太关键的通道 7 和 8，却在音素辨别中发挥了关键作用。

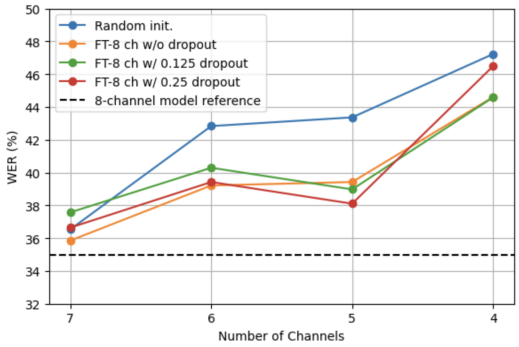


图 3。通道丢弃和微调对不同通道配置下词错误率的影响。以 8 通道的性能作为参考。

这凸显了通道在语音合成与音素识别任务中的效用差异。

5.4. 使用通道丢弃进行微调

为缓解通道数量减少时的性能下降，我们对从 8 通道预训练基线初始化的模型进行了微调，并采用了我们提出的通道丢弃策略（第 3.2 节）。微调针对表现最佳的通道子集进行：7 个通道为 (1, 2, 3, 4, 5, 7, 8)；6 个通道为 (1, 2, 3, 4, 5, 8 1, 3, 5, 6 5 1, 2, 3, 4, 5 4 4)；通道为 ( )；通道为 ( )。-通道子集依据穷举搜索结果选出，其余子集则依据反向消除选出。如图 3 所示，在 7 通道设置下，不同初始化方法的性能相当。相比之下，在 4、5 和 6 通道设置下，微调模型的表现始终优于从头训练的模型，这表明即使部署通道数较少的模型，全通道预训练仍然有用。有趣的是，表现最佳的丢弃设置因通道数而异：对于 6 和 4 个通道，不进行丢弃（$p = 0$）时结果最佳；而对于 5 个通道，较高的丢弃率能提升性能。值得注意的是，尽管从头训练时，5 通道模型的表现不如 6 通道模型，但微调后这一趋势发生了逆转。这表明，模型性能并非仅由可用通道数量决定，还受到预训练动态与特定通道配置之间相互作用的影响，凸显了在传感器限制下采用自适应微调策略的重要性。

6. 结论

我们通过贪心消除、穷举子集评估和音素级分析，研究了肌电通道缩减对无声语音合成的影响。结果表明，单独重要性较高的通道组合起来后未必最优，这凸显了互补性通道交互的重要性。使用随机通道丢弃对预训练模型进行微调，在 4–6 个通道的设置下始终能提升性能，且优于从头训练的模型。这些发现表明，通道感知训练策略即使在肌电传感器数量较少的情况下，也能实现稳健的性能，为打造更实用的无声语音系统铺平了道路。

7. 致谢

本研究部分得到韩国政府（科学技术信息通信部）资助的 IITP 信息通信技术规划与评估研究院项目支持，项目编号：
[2022 220641，50], No. RS -II %
[2021 211343, 5]
NO.RS- -II %；另获韩国政府（科学技术信息通信部）资助的韩国国家研究基金会（NRF）项目支持，项目编号： [No. RS-2024-00461617, 45%]。

8. 参考文献

[1] David Marshall Gaddy，《无声语音的数字化发声》，加利福尼亚大学伯克利分校，2022年。[2] Qinhan Hou、Stefano Van Gogh、Kevin Scheck、Zhao Ren、Tanja Schultz、Michael Wand 和 J¨urgen Schmidhuber，“emg2vec：基于肌电图的无声语音接口中的自监督预训练”，载于*2024*年*IEEE*生物医学工程学会第*46*届年度国际会议（*EMBC*）。IEEE，2024年，第1–5页。[3] Kevin Scheck 和 Tanja Schultz，“通过软语音单元预测，从肌电信号合成多说话人语音”，载于*ICASSP 2023-2023 IEEE*国际声学、语音与信号处理会议（*ICASSP*）。IEEE，2023年，第1–5页。[4] Jaejun Lee、Yoori Oh 和 Kyogu Lee，“无声发声：仅使用面部输入的多说话人无声语音发声”，载于*ICASSP 2025-2025 IEEE*国际声学、语音与信号处理会议（*ICASSP*）。IEEE，2025年，第1–5页。[5] David Gaddy 和 Dan Klein，“无声语音的数字化发声”，*arXiv*预印本 *arXiv:2010.02960*，2020年。[6] Benjamin Van Niekerk、Marc-Andr´e Carbonneau、Julian Za¨ıdi、Matthew Baas、Hugo Seut´e 和 Herman Kamper，“用于改进语音转换的离散语音单元与软语音单元比较”，载于*ICASSP 2022-2022 IEEE*国际声学、语音与信号处理会议（*ICASSP*）。IEEE，2022年，第6562–6566页。[7] Wei-Ning Hsu、Benjamin Bolte、Yao-Hung Hubert Tsai、Kushal Lakhotia、Ruslan Salakhutdinov 和 Abdelrahman Mohamed，“HuBERT：通过隐藏单元的掩码预测进行自监督语音表征学习”，*IEEE/ACM*音频、语音与语言处理汇刊，第29卷，第3451–3460页，2021年。[8] Zhao Ren、Kevin Scheck、Qinhan Hou、Stefano van Gogh、Michael Wand 和 Tanja Schultz，“Diff-ETS：学习用于肌电图到语音转换的扩散概率模型”，载于*2024*年*IEEE*生物医学工程学会第*46*届年度国际会议（*EMBC*）。IEEE，2024年，第1–4页。[9] Shan Ullah 和 Deok-Hwan Kim，“用于最小化语音转肌电转换中软语音损失的优化肌电编码器”，载于*2024*年*IEEE*大数据与智能计算国际会议（*BigComp*）。IEEE，2024年，第215–218页。[10] Peter Wu、Bohan Yu、Kevin Scheck、Alan W Black、Aditi S Krishnapriyan、Irene Y Chen、Tanja Schultz、Shinji Watan- abe 和 Gopala K Anumanchipalli，“基于多模态发音表征的深度语音合成”，*arXiv*预印本 *arXiv:2412.13387*，2024年。[11] Tyler Benster、Guy Wilson、Reshef Elisha、Francis R Wil- lett 和 Shaul Druckmann，“通过跨模态方法和大语言模型增强识别实现无声语音识别”，*arXiv*预印本 *arXiv:2403.05583*，2024年。[12]Jihwan Lee、Kevin Huang、Kleanthis Avramidis、Simon Pistrosch、Monica Gonzalez-Machorro、Yoonjeong Lee、Bj¨orn Schuller、Louis Goldstein 和 Shrikanth Narayanan，“语音产生过程中基于表面肌电图的发音特征预测”，*arXiv*预印本 *arXiv:2505.13814*，2025年。[13] David Gaddy 和 Dan Klein，“一种改进的无声语音发声模型”，*arXiv*预印本 *arXiv:2106.01933*，2021年。[14] Jintao Guo、Lei Qi、Yinghuan Shi 和 Yang Gao，“位置丢弃：一种用于域泛化的渐进式逐层、逐通道丢弃方法”，*ACM*多媒体计算、通信与应用汇刊，第20卷，第3期，第1–23页，2023年。[15] Mingjie He、Jie Zhang、Shiguang Shan、Xiao Liu、Zhongqin Wu 和 Xilin Chen，“用于遮挡人脸识别的局部性感知逐通道丢弃”，*IEEE*图像处理汇刊，第31卷，第788–798页，2021年。[16] Jungil Kong、Jaehyeon Kim 和 Jaekyoung Bae，“HiFi-GAN：用于高效、高保真语音合成的生成对抗网络”，神经信息处理系统进展，第33卷，第17022–17033页，2020年。[17] Alec Radford、Jong Wook Kim、Tao Xu、Greg Brockman、Christine McLeavey 和 Ilya Sutskever，“通过大规模弱监督实现鲁棒语音识别”，载于国际机器学习会议。PMLR，2023年，第28492–28518页。