

# BỨC TRANH ÂM THANH: KHÁM PHÁ SỨC MẠNH CỦA GIỌNG NÓI TIẾNG VIỆT TRONG TẠO HÌNH ẢNH

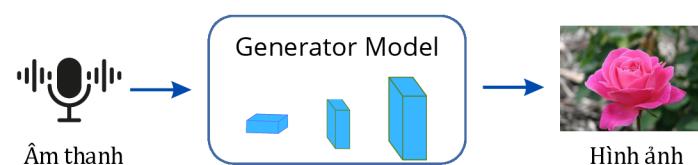
Hoàng Tiến Anh<sup>1</sup>, Dương Văn Tài<sup>1</sup>, Trần Xuân Diện<sup>1</sup>, Nguyễn Văn Anh Tuấn<sup>1</sup>

<sup>1</sup> Khoa Công nghệ Thông tin, Trường Đại học Công nghiệp Hồ Chí Minh

## TÓM TẮT

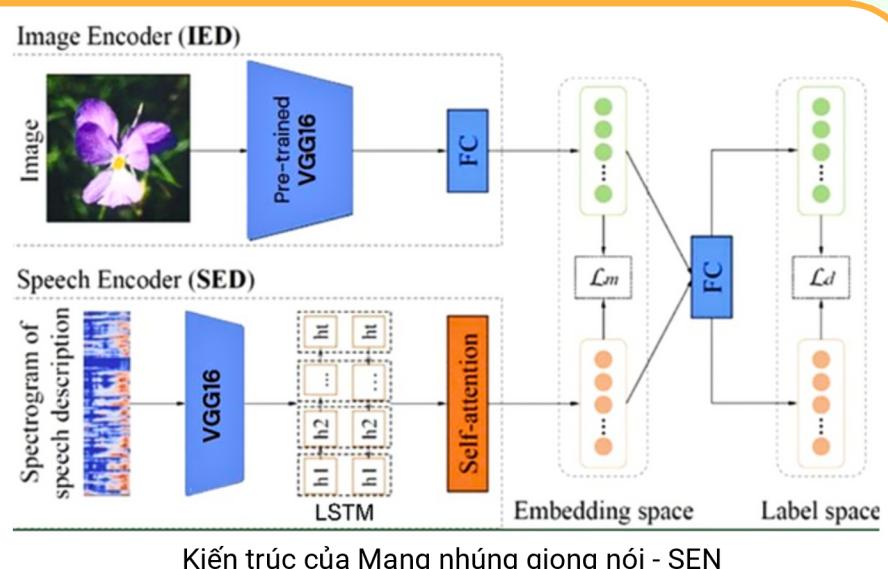
Bài báo nghiên cứu bài toán chuyển đổi giọng nói tiếng Việt thành hình ảnh (S2IG) - một bài toán khó với ngôn ngữ không có chữ viết hay lượng dữ liệu hạn chế. Chúng tôi ứng dụng phương pháp S2IGAN gồm Mạng nhúng giọng nói - SEN và Mô hình sinh xếp dày đặc được giám sát bởi các mối quan hệ - RDG.

## GIỚI THIỆU

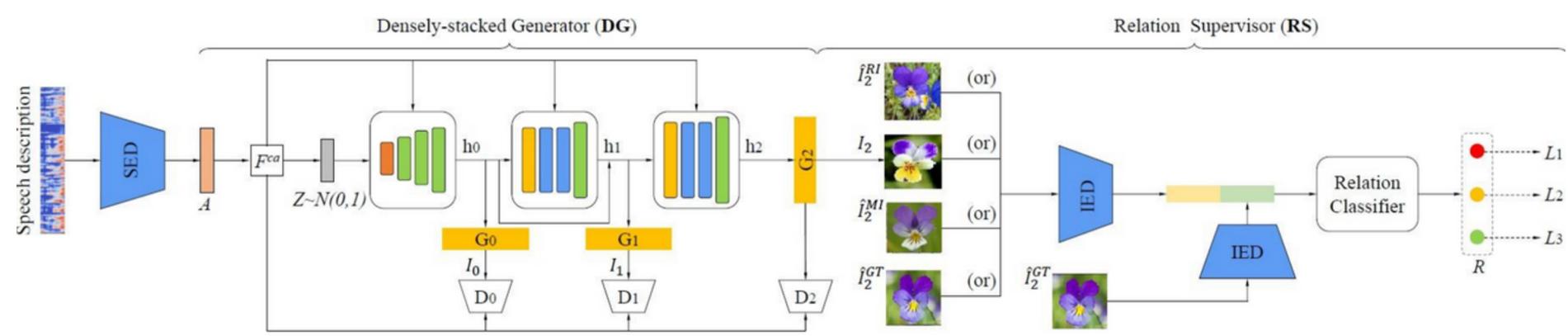


## PHƯƠNG PHÁP

Mô-đun Speech Embedding Network (SEN) được sử dụng để học cách liên kết giữa giọng nói và hình ảnh để tạo ra các embedding thống nhất trong không gian chung, trong khi mô-đun Relation-supervised Densely-Stacked Generative Model (RDG) sử dụng các embedding của SEN đó trước đó để tạo ra các hình ảnh tương ứng.



Kiến trúc của Mạng nhúng giọng nói - SEN



Kiến trúc của Mô hình sinh xếp dày đặc được giám sát bởi các mối quan hệ - RDG

## THỰC NGHIỆM

- Phần cứng: NVIDIA TESLA P100 GPU được hỗ trợ bởi Kaggle
- Dữ liệu: Oxford 102 Flower Dataset được dịch sang tiếng việt và chuyển thành âm thanh.

Datasets	Hình ảnh	Âm thanh
Train	7370	73700
Test	819	8190

Metrics	Kết quả
FID score	146.41
IS score	1.09±0.13

Kết quả đánh giá sinh ảnh

Thông số	D_loss	G_loss	KL_loss	RS_loss	Cond_loss_64	Uncond_loss
Train	0.11397	13.98112	4.16717	0.48064	4.44182	4.3817

Loss của mô hình

## KẾT LUẬN

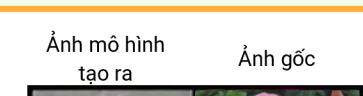
- Nghiên cứu cho thấy tiềm năng trong việc tạo hình ảnh từ giọng nói tiếng Việt, và có thể ứng dụng trong nhiều lĩnh vực khác nhau.
- Đã đạt được kết quả khả quan, nhưng cần tiếp tục cải thiện, cải tiến phương pháp để đạt kết quả thực sự chất lượng.
- Huấn luyện thêm trên nhiều tập dữ liệu khác, sử dụng giọng nói tự nhiên thay vì tổng hợp.

## KẾT QUẢ

- Các hình ảnh do mô hình tạo ra đã nắm bắt được các chi tiết chính của hình ảnh gốc, phản ánh chính xác nội dung giọng nói đầu vào.
- Chất lượng hình ảnh hơi mờ do hạn chế về tài nguyên tính toán nên nghiên cứu tập trung vào việc sinh hình ảnh 64x64.



Nội dung  
giọng nói



Ảnh mô hình  
tạo ra

Ảnh gốc

Bông hoa này có  
cánh màu hồng với  
đầu nhụy hoa



Hoa này có cánh  
màu vàng xám với  
kết cấu hơi nhăn  
và một số đốm

