Tạp chí điện tử
Khoa học và Công nghệ Trường Đại học Công nghệ Đông Á
Website Tạp chí: https: //vjai.org.vn

A Novel Architecture for Crowd-Counting Models

Pham Thi Loan
Faculty of Information Technology, East Asia University of Technology

ARTICLE INFO

Keywords:
crowd counting, convolutional neural networks, image processing, deep learning.

ABSTRACT

Crowd counting is the task of estimating the number of people in an image. Each training image contains multiple individuals annotated with point-level labels. Existing crowd counting methods typically smooth each annotated point using a Gaussian kernel or estimate the probability of each pixel corresponding to an annotated point. In this paper, we propose a novel crowd counting architecture capable of handling perspective distortion by effectively utilizing multiple features generated during the encoding process. Unlike previous approaches, our method avoids extracting additional multi-scale features, thereby significantly reducing the overall computational cost. To achieve this goal, we further improve the existing multi-scale fusion mechanism and introduce a new channel reduction block. Experimental results on the ShanghaiTech dataset demonstrate that our method outperforms state-of-the-art approaches with similar computational complexity.

* Corresponding author
Email: loan@eaut.edu.vn
DOI: https://doi.org/10.65153/g2qxfw28
Received: 18/11/2025; Received in revised form: 30/03/2026; Accepted: 03/04/2026
Available online: 03/04/2026
Published by: East Asia University of Technology

Tạp chí điện tử
Khoa học và Công nghệ Trường Đại học Công nghệ Đông Á
Website Tạp chí: https: //vjai.org.vn

Xây dựng một kiến trúc mới cho mô hình đếm đám đông

Phạm Thị Loan
Khoa Công nghệ thông tin, Trường Đại học Công nghệ Đông Á

THÔNG TIN BÀI BÁO

Từ khóa:
đếm đám đông, mạng nơ-ron tích chập đa lớp, xử ly hình ảnh, học sâu.

TÓM TẮT

Đếm đám đông là nhiệm vụ đếm số người trong một hình ảnh. Mỗi hình ảnh huấn luyện chứa nhiều người được đánh dấu bằng một dấu chấm. Các phương pháp đếm đám đông hiện có phải làm mịn mỗi điểm rơi được chú thích bằng hàm Gauss hoặc ước tính khả năng xảy ra của mỗi pixel cho điểm được chú thích. Trong bài báo này, chúng tôi đề xuất một kiến trúc đếm đám đông mới có thể xử lý biến dạng phối cảnh bằng cách sử dụng thông minh nhiều đặc trưng được tạo ra trong quá trình mã hóa. Không giống như các phương pháp trước đây, phương pháp của chúng tôi tránh trích xuất các đặc trưng đa tỷ lệ bổ sung, giúp giảm đáng kể tổng khối lượng tính toán. Để đạt được mục đích này, chúng tôi cũng đã cải tiến cơ chế hợp nhất đa tỷ lệ hiện có và đưa ra một khối giảm kênh mới. Các thí nghiệm trên cơ sở dữ liệu ShanghaiTech đã chứng minh rằng phương pháp của chúng tôi có thể vượt trội hơn các phương pháp tiên tiến có độ phức tạp tính toán tương tự.

1. MỞ ĐẦU

Đếm đám đông nhằm mục đích tự động ước tính số lượng cá thể có mặt trong một cảnh từ hình ảnh hoặc video. Chúng ta có thể áp dụng điều này vào nhiều lĩnh vực, chẳng hạn như kiểm soát giao thông [1], nghiên cứu sinh học [2] và giám sát khoảng cách xã hội [3]. Qua nhiều năm, các mô hình đếm đám đông đã phát triển từ việc sử dụng các mô hình hồi quy cổ điển như rừng ngẫu nhiên [4] và quy trình Gaussian [5], sang các mạng nơ-ron tích chập hiệu suất cao (CNN) [6–9]. Một mặt, đếm đám đông nhằm mục đích tự động ước tính số lượng cá thể có mặt trong một cảnh từ hình ảnh hoặc video. Nó có thể được áp dụng trong nhiều lĩnh vực, chẳng hạn như kiểm soát giao thông [1], nghiên cứu sinh học [2], và gần đây là giám sát khoảng cách xã hội [3]. Trong những năm qua, các mô hình đếm đám đông đã phát triển từ việc sử dụng các mô hình hồi quy cổ điển, chẳng hạn như rừng ngẫu nhiên [4] và quy trình Gaussian [5], sang các mạng nơ-ron tích chập (CNN) hiệu suất cao [6–9]. Các mạng sâu này thường áp dụng phương pháp mã hóa-giải mã: Đầu tiên, một hình ảnh được đưa vào bộ mã hóa để học các biểu diễn dữ liệu (bản đồ đặc trưng). Sau đó, bộ giải mã khai thác mô hình cấp

* Tác giả liên hệ
Email: loan@eaut.edu.vn
DOI: https://doi.org/10.65153/g2qxfw28
Ngày nhận: 18/11/2025; Ngày nhận bản sửa: 30/03/2026; Ngày chấp nhận: 03/04/2026
Ngày online: 03/04/2026
Đơn vị xuất bản: Trường Đại học Công nghệ Đông Á

cao nhất (đầu ra từ lớp cuối cùng của bộ mã hóa) để tạo ra bản đồ mật độ, là sự phân bố của đám đông. Vì các khối tích chập và khối gộp của mạng VGG [10], trong đó mỗi lớp khai thác các hạt nhân có kích thước cố định, tạo thành hầu hết các bộ mã hóa, nên kích thước của các trường tiếp nhận vẫn không đổi trên bản đồ đặc trưng được mã hóa cuối cùng. Do đó, biểu diễn này chỉ có thể xử lý hình ảnh đám đông có tỷ lệ tương tự. Tuy nhiên, con người thường được mô tả ở nhiều kích cỡ khác nhau do góc nhìn của máy ảnh. Do đó, đối tượng được mã hóa cũng nên có kích thước trường tiếp nhận khác nhau để mô hình hóa sự thay đổi tỷ lệ.

Cấu trúc đa cột [6], [11] đã được đề xuất để giải quyết vấn đề này. Tuy nhiên, gần đây đã được chứng minh rằng các đặc điểm từ mỗi cột gần như giống hệt nhau, và việc huấn luyện các mô hình sâu loại này có thể rất kém hiệu quả [7]. Do đó, để giải quyết vấn đề tỷ lệ này, các phương pháp tiên tiến [8], [12], [13] sử dụng một mô-đun đa tỷ lệ để xử lý thêm biểu diễn được mã hóa và tạo ra một bản đồ đặc điểm với các kích thước trường tiếp nhận khác nhau. Tuy nhiên, chiến lược như vậy bỏ qua việc các bản đồ đặc điểm được trích xuất bởi các lớp mã hóa nông hơn đã cung cấp thông tin về các tỷ lệ khác nhau, và việc tận dụng các thành phần bổ sung khiến mô hình tổng thể tốn kém hơn về mặt tính toán. Do đó, đóng góp của chúng tôi trong bài báo này là một cơ chế đa tỷ lệ mới, giải quyết vấn đề tỷ lệ bằng cách tận dụng phần lớn các đặc điểm được tạo ra từ bộ mã hóa để tránh các mô-đun trích xuất đặc điểm bổ sung và giữ cho chi phí tính toán ở mức thấp. Thiết kế này kết hợp một phạm vi toàn diện các kích thước trường tiếp nhận (từ 6 đến 192), bao phủ hầu hết tất cả các tỷ lệ có thể mà một người có thể mô tả trong hình ảnh đám đông. Các thí nghiệm trên hai cơ sở dữ liệu chuẩn [6] chứng minh rằng mô hình của chúng tôi có thể đạt được kết quả tiên tiến nhất hoặc tương đương với ít phép toán dấu phẩy động hơn đáng kể.

Mặt khác, trong [14], các tác giả giới thiệu và phân tích các phương pháp tiêu chuẩn trong lĩnh vực này, nhấn mạnh các phương pháp đếm dựa trên học sâu. Các phương pháp hiện có được phân loại thành bốn loại: dựa trên phát hiện, dựa trên hồi quy, dựa trên mạng nơ-ron tích chập và dựa trên video. Bên cạnh đó, trong [15], tác giả trình bày một mô hình mạng nơ-ron sâu mới cho các hệ thống hỗ trợ máy bay không người lái, trong đó hình ảnh từ camera của máy bay không người lái được xử lý cho các hoạt động đếm đám đông thông minh. Kiến trúc đề xuất của chúng tôi sử dụng đạo hàm mô hình khái niệm ResNet để ước tính đám đông trong ảnh. Trái ngược với các phương pháp tăng cường miền trước đây, trong [16], tác giả sử dụng AutoML để tìm phép biến đổi tốt nhất trên nguồn sẽ phục vụ tốt nhất cho tác vụ hạ lưu. Để giảm bớt khó khăn trong việc căn chỉnh, chúng tôi thực hiện căn chỉnh chi tiết cho tiền cảnh và hậu cảnh riêng biệt. Tác giả đã kiểm tra phương pháp này với năm chuẩn mực đếm đám đông trong thế giới thực và vượt trội hơn các phương pháp hiện có một cách đáng kể.

Đặc biệt, trong DM-count [17], những đóng góp đáng kể là việc sử dụng các phương pháp chuẩn bị và huấn luyện dữ liệu, và giới hạn lỗi của DM-generalization Count chặt chẽ hơn so với các phương pháp làm mịn Gaussian. DM-Count vượt trội hơn các phương pháp tiên tiến trước đây trên hai tập dữ liệu đếm quy mô lớn với biên độ lớn về Sai số tuyệt đối trung bình. Tuy nhiên, chúng tôi nhận thấy mô hình xây dựng còn hạn chế. Để làm được điều này, chúng tôi đã sử dụng một mô hình đơn giản hơn với mô hình VGG16 cho thuật toán, giúp phương pháp của chúng tôi đơn giản hơn và đạt được hiệu suất cao hơn.

Trong bài báo này, chúng tôi đề xuất một kiến trúc đếm đám đông mới khai thác trực tiếp các đặc trưng đa mức được sinh ra trong quá trình mã hóa nhằm xử lý hiệu quả hiện tượng biến dạng phối cảnh và thay đổi tỷ lệ. Khác với các phương pháp hiện có thường sử dụng các

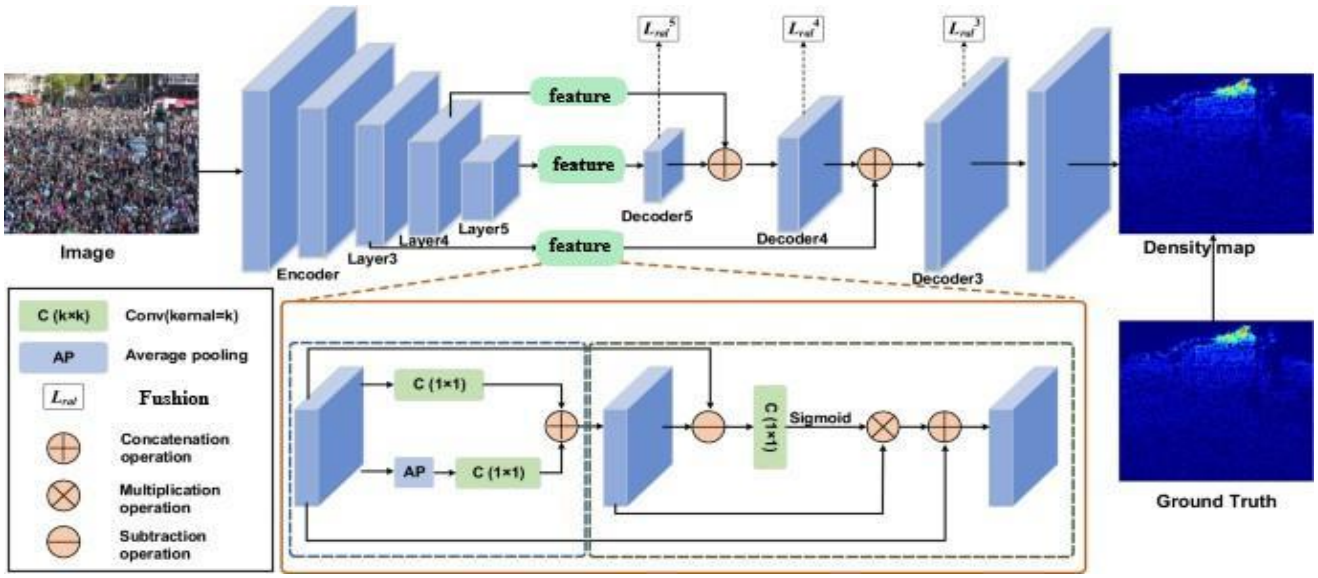
mô-đun trích xuất đa thang độ bổ sung, mô hình đề xuất có các điểm mới sau: (i) chiến lược hợp nhất đặc trưng phân nhóm theo cấu trúc phân cấp, tận dụng trực tiếp các đặc trưng từ bộ mã hóa với các kích thước trường tiếp nhận khác nhau; (ii) cơ chế gán trọng số dựa trên đặc trưng tương phản, cho phép hợp nhất thích ứng các đặc trưng đa thang độ theo tầm quan trọng không gian; (iii) mô-đun giảm kênh nhẹ kết hợp tích chập giãn nở và lớp cổ chai, giúp giảm chi phí tính toán nhưng vẫn bảo toàn thông tin ngữ cảnh quan trọng. Nhờ các thiết kế này, mô hình đề xuất đạt được hiệu năng cạnh tranh hoặc vượt trội trong khi vẫn duy trì độ phức tạp tính toán thấp. Phần sau đây giải thích phần còn lại của bài luận: Phần 2 thảo luận về các công việc liên quan. Phần 3 giải thích quy trình đề xuất. Phần 4 bao gồm đánh giá hiệu suất. Phần 5 kết thúc bằng thảo luận về những phát hiện và kế hoạch của chúng tôi.

2. CÁC VẤN ĐỀ LIÊN QUAN

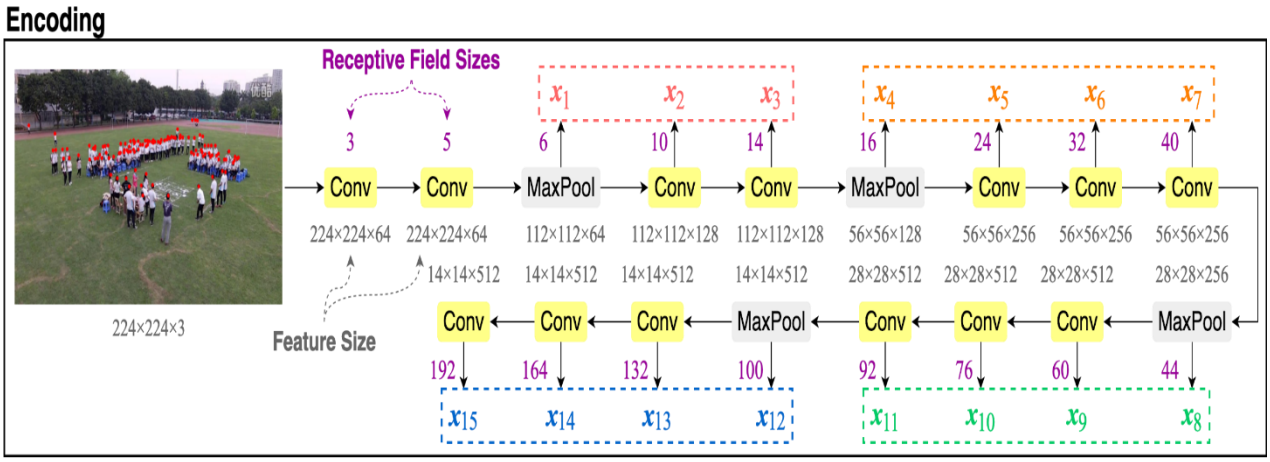
Các phương pháp đếm đám đông có thể được chia thành ba loại: phát hiện rồi đếm, hồi quy đếm trực tiếp và ước tính bản đồ mật độ. Các phương pháp đếm đám đông ban đầu [18], [19], [20], [21] dựa trên các bộ phát hiện đối tượng. Đồng thời, các công trình sau này tránh chúng vì chúng nhạy cảm với sự che khuất và những nỗ lực to lớn cần thiết để chú thích các hộp giới hạn. Một số phương pháp không dựa trên phát hiện này [4], [5], [22], [23] coi việc đếm đám đôngnhư một bài toán hồi quy: chúng học các biểu diễn đặc trưng cấp thấp mà từ đó tổng số đếm được hồi quy trực tiếp. Các tổn thất huấn luyện của các phương pháp này chỉ phụ thuộc vào số đếm thực tế (một số vô hướng) và không xem xét phân phối mật độ đám đông, do đó có khả năng khái quát hóa kém. Do đó, các mô hình này đã sớm được thay thế bởi các thuật toán [24], [25], [8], [12] thay vào đó dự đoán mật độ đám đông, và các phương pháp dựa trên mật độ này chủ yếu dựa vào CNN.

Vì các vị trí không gian 3D phải được chiếu lên không gian 2D trong khi hình ảnh đám đông RGB được chụp từ thế giới thực, nên người có thể được mô tả ở các kích thước khác nhau do sự biến dạng phối cảnh. Sự thay đổi về tỷ lệ có thể ảnh hưởng nghiêm trọng đến ước tính mật độ. Để giải quyết vấn đề này, trong [26], thông tin hình học bổ sung được khai thác để điều chỉnh mô hình của họ cho các cảnh khác nhau, nhưng thông tin này không phải lúc nào cũng được cung cấp. Do đó, các phương pháp sau này có xu hướng học tỷ lệ của đám đông một cách ngầm định. Ví dụ, Hydra-CNN [25] chia một hình ảnh thành một kim tự tháp các mảng, mỗi mảng đại diện cho một tỷ lệ khác nhau và được đưa vào một đầu mã hóa khác. Sau đó, tất cả các đặc điểm được mã hóa được nối lại mà không cần xử lý thêm và được sử dụng để tạo bản đồ mật độ. Cách tiếp cận này bỏ qua thực tế là tỷ lệ thay đổi liên tục trên toàn bộ hình ảnh. CAN [8] được đề xuất để giải quyết vấn đề này. Toàn bộ mô hình chỉ liên quan đến một bộ mã hóa, do đó, một mô-đun gộp kim tự tháp không gian [27] được tận dụng để giúp nó nhận biết tỷ lệ. Sau đó, các đặc trưng được tính trung bình theo các trọng số có thể học được để đảm bảo kích thước trường tiếp nhận của phân đoạn hợp nhất thay đổi mượt mà.

Tuy nhiên, kiến trúc này kém hiệu quả hơn vì nó không khai thác các thành phần cấp thấp được trích xuất trong quá trình mã hóa. Các biểu diễn này và bản đồ đặc trưng cấp cao cuối cùng có thể cung cấp thông tin về các tỷ lệ khác nhau vì chúng có kích thước trường tiếp nhận khác nhau. Ngoài ra, mô-đun đa tỷ lệ trong [8] chỉ sử dụng bốn kích thước bộ lọc, do đó bao phủ một phạm vi tỷ lệ hạn chế.



Hình 1. *Sơ đồ luồng đề xuất*



Hình 2. *Bước 01*

3. Phương pháp đề xuất

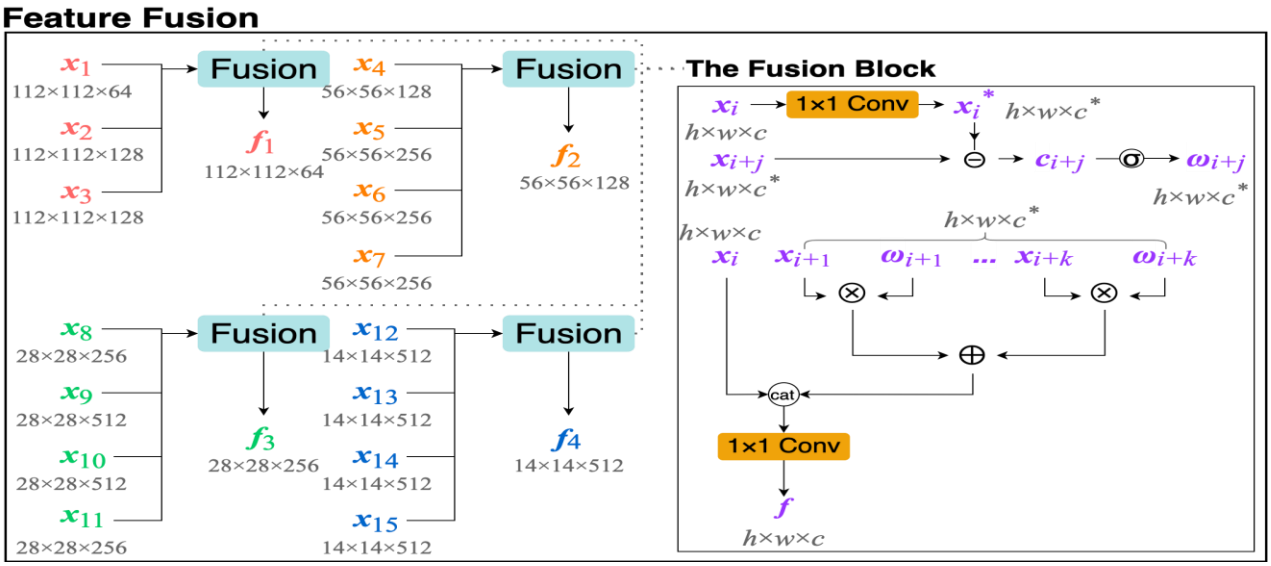
3.1. Mã hóa

Theo các thông lệ tiêu chuẩn [6], [7] trong lĩnh vực này, VGG-16 [10] được sử dụng làm bộ mã hóa cho mô hình của chúng tôi. Các lớp max-pooling và fully-connecting cuối cùng được loại bỏ vì chúng chịu trách nhiệm cho việc dự đoán lớp. Do đó, bộ mã hóa bao gồm 13 lớp tích chập và bốn lớp max-pooling. Vì hai bản đồ đặc trưng được trích xuất trước thao tác max-pooling đầu tiên không đủ thông tin, nên chỉ các đặc trưng được học sau đó mới được bảo toàn và hợp nhất. Do đó, như thể hiện trong Hình 5, bộ mã hóa xuất ra tổng cộng 15 bản đồ đặc trưng, được ký hiệu là $x_{i,j} = 1, 2, ..., 15$. Các bản đồ đặc trưng này được chia thành bốn nhóm theo chiều cao và chiều rộng của chúng: $x_1 - x_3$, $x_4 - x_7$, $x_8 - x_{11}$, and $x_{12} - x_{15}$.

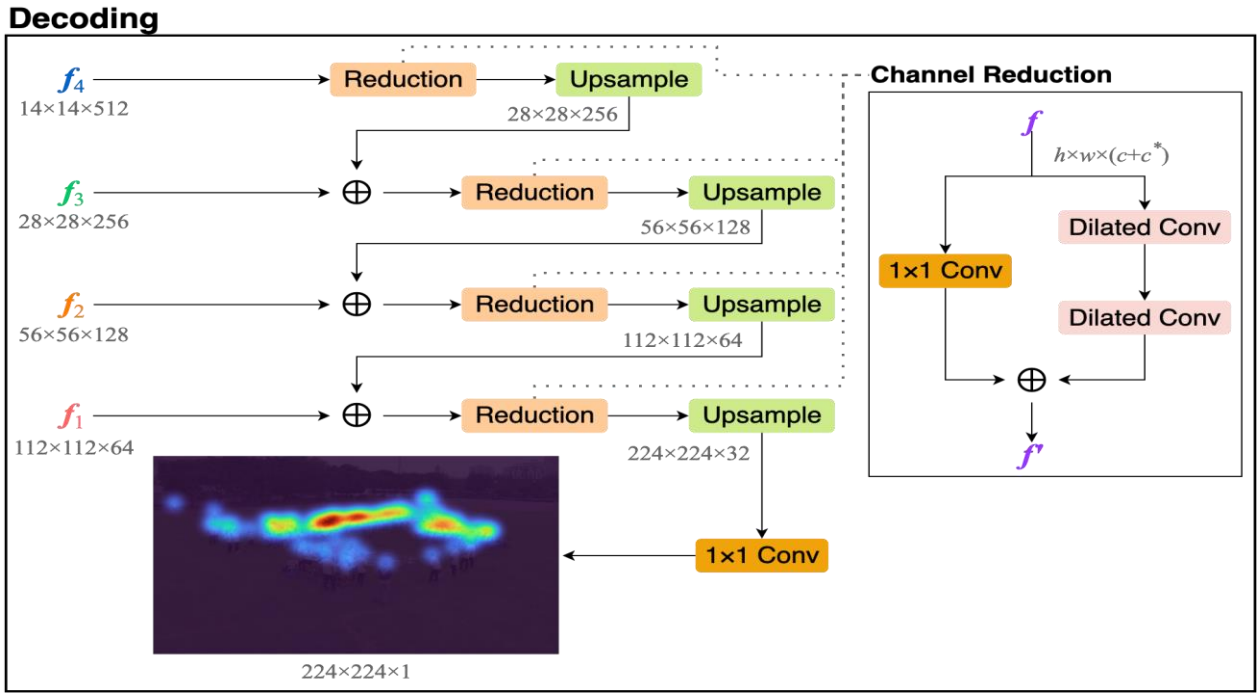
Hình 2 minh họa Bước 01 của mô hình đề xuất, trong đó các bản đồ đặc trưng được trích xuất từ các lớp mã hóa khác nhau được phân nhóm dựa trên độ phân giải không gian. Trong mỗi nhóm, các đặc trưng tương phản được tính toán giữa các đặc trưng nông và sâu nhằm nắm bắt sự thay đổi tỷ lệ cục bộ. Các đặc trưng tương phản này sau đó được sử dụng để sinh ra các bản đồ trọng số không gian, cho phép mô hình nhấn mạnh các kích thước trường tiếp nhận phù hợp tại từng vị trí ảnh.

Hình 3 trình bày Bước 02, trong đó các đặc trưng đa thang độ đã được hợp nhất trong từng nhóm sẽ tiếp tục được kết hợp theo cấu trúc phân cấp. Quá trình hợp nhất bắt đầu từ

nhóm đặc trưng có độ phân giải thấp nhất và được thực hiện dần lên các nhóm có độ phân giải cao hơn thông qua phép nội suy và cộng đặc trưng. Cách tiếp cận này giúp duy trì ngữ cảnh toàn cục đồng thời bảo toàn các chi tiết không gian quan trọng.



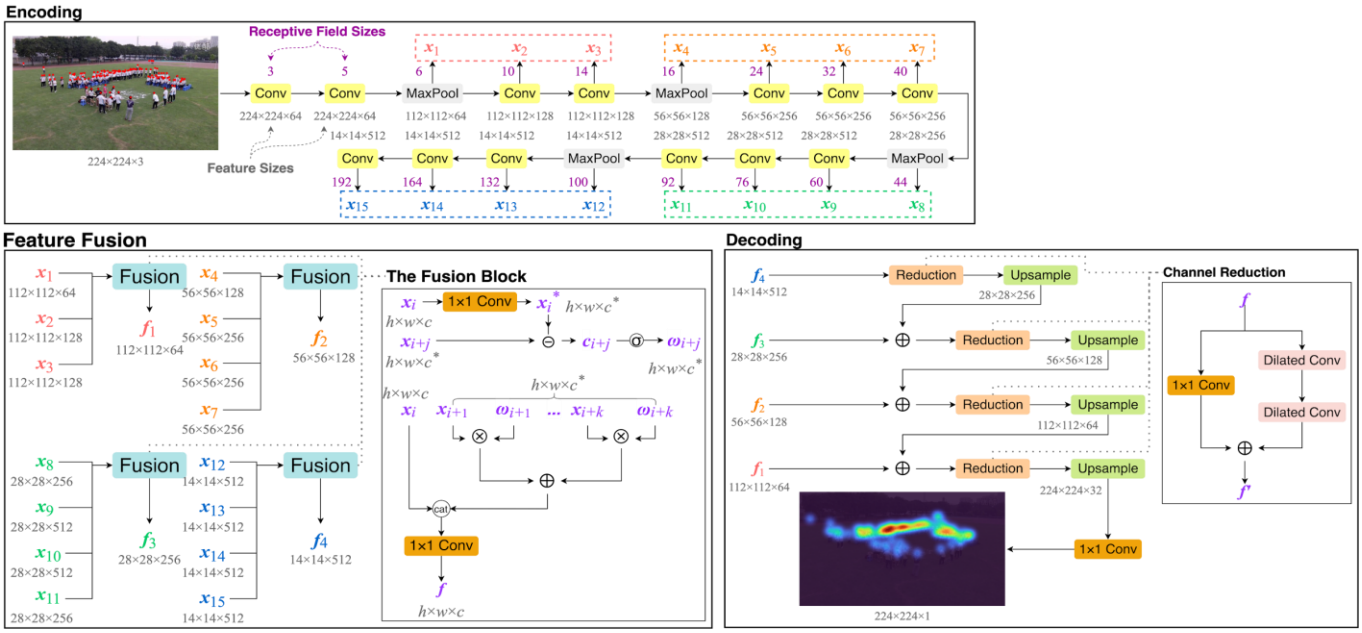
Hình 3. *Bước 02*



Hình 4. *Bước 03*

Hình 4 minh họa Bước 03 của quá trình giải mã, trong đó một mô-đun giảm kênh mới được áp dụng trước khi hợp nhất các đặc trưng. Mô-đun này kết hợp tích chập giãn nở với lớp cổ chai nhằm giảm số lượng kênh nhưng vẫn giữ được thông tin ngữ cảnh nổi bật. Các đặc trưng sau khi được tinh chỉnh sẽ được giải mã dần để tạo ra bản đồ mật độ đám đông cuối cùng.

Như minh họa trong Hình 2–4, phương pháp đề xuất bao gồm ba giai đoạn chính: (1) mã hóa đặc trưng sử dụng bộ mã hóa VGG-16 đã được điều chỉnh; (2) hợp nhất đặc trưng đa thang độ theo nhóm dựa trên cơ chế tương phản; (3) giải mã phân cấp kết hợp mô-đun giảm kênh thích ứng để tạo ra bản đồ mật độ cuối cùng.



Hình 5. *Mô hình đề xuất*

3.2. Tính năng

Giả định x_i , x_{i+1} , ..., x_{i+k} là một nhóm bản đồ đặc điểm quan tâm, trong đó x_i được tạo ra bởi một nhóm tối đa lớp và có kênh c, trong khi phần còn lại được tạo ra bởi các lớp tích chập và có kênh c∗. Để lắp ráp một biểu diễn nhận biết tỷ lệ từ chúng, chúng tôi áp dụng một chiến lược tương tự như chiến lược được đề xuất trong [7], và phương pháp của chúng tôi được minh họa trong 3. Do tỷ lệ thay đổi liên tục trên một hình ảnh, nên kích thước trường tiếp nhận của đặc trưng đa tỷ lệ đầu ra, mô hình hóa thứ hạng, cũng phải liên tục về mặt không gian. Để đạt được mục tiêu này chúng tôi hợp nhất các đặc trưng này thông qua phương pháp trung bình có trọng số, trong đó các bản đồ trọng số được học từ tầm quan trọng không gian của từng đặc trưng.

Đầu tiên, vì x_i và x_{i+j} ($j\epsilon$ [1,⋯, k]) có thể không có sự tích hợp thông tin không gian lân cận, x_i ít bị ảnh hưởng bởi nhiễu nền. Với lợi thế như vậy, tất cả các biểu diễn khác đều được so sánh với nó. Cụ thể, theo [8], chúng tôi đề xuất một khái niệm tương tự để so sánh các đặc điểm:

$$c_j = x_{i+j} - x_i^* \quad (1)$$

Với j = 1,..., k. Vì x^* và x_{i+j} (j = 1,..., k) có kích thước trường tiếp nhận khác nhau, các tính năng tương phản c_j kết hợp sự khác biệt giữa bất kỳ vị trí không gian nào và các điểm ảnh lân cận. Do đó, chúng có thể hỗ trợ việc học ranh giới của mỗi người, xác định tầm quan trọng tương đối của từng kích thước trường tiếp nhận. Sau đó, chúng tôi tính toán các trọng số ω_j như sau:

$$\omega_j = \sigma(c_j) \quad (2)$$

Trong đó σ biểu thị hàm sigmoid 1. Sử dụng các trọng số này, ω_i , ω_{i+1} , ω_{i+k} , sau đó chúng tôi kết hợp x_i , x_{i+1} , ..., x_{i+k} thích ứng và nối bản đồ tính năng trung bình với x_i. Cuối cùng, chúng tôi thêm một lớp nút thắt để giảm thiểu tính toán. Quá trình này có thể được diễn đạt như sau:

$$f = Conv\left([x_i | \sum_{j=1}^{k} w_j \odot x_{i+j}]\right) \quad (3)$$

Trong đó f, Conv, "|," và $\odot\odot$ lần lượt biểu thị đặc điểm hợp nhất, lớp cổ chai, nối từng kênh và tích từng phần tử.

3.3. Giải mã

Như thể hiện trong Hình 2, chúng tôi biểu thị các tính năng đa thang độ pha đầu tiên được tạo ra từ bốn nhóm thông qua (3) là f_1 , f_2 , f_3 ,

and f_4 . Các tính năng này được hợp nhất theo thứ tự ngược lại (xem Hình 3). Để kết hợp f_4 và f_3 thông qua phép cộng, trước tiên chúng ta biến đổi hình dạng của f_4. Cụ thể, số lượng kênh của f_4 được hạ thấp trong khi kích thước không gian của nó được mở rộng. Theo truyền thống, một hạt nhân tích chập từng điểm đơn được sử dụng để giảm số lượng kênh. Tuy nhiên, lấy cảm hứng từ tích chập giãn nở, vốn đã được chứng minh là có khả năng trích xuất thông tin nổi bật sâu hơn mà vẫn duy trì độ phân giải không gian [7], chúng tôi đề xuất một mô-đun giảm kênh mới. Mô-đun hai luồng của chúng tôi bao gồm một khối tích chập giãn nở và một lớp cổ chai, và hai cột được kết nối thông qua phép cộng. Để phù hợp với độ phân giải không gian của f_3, chúng tôi sử dụng phép nội suy song tuyến tính để tăng gấp đôi kích thước của f_4 sau khi giảm kênh. Sau đó, hai đặc điểm này có thể được hợp nhất bằng phép cộng. Đặc điểm hợp nhất mới sau đó được kết hợp với f_2 theo một mô hình tương tự — đầu tiên chúng tôi sửa đổi kích thước của nó thông qua mô-đun giảm kênh được đề xuất và nội suy song tuyến tính và sau đó thêm nó vào f_2. Quá trình này được thực hiện tương tác cho đến khi tất cả các đặc trưng được hợp nhất. Cuối cùng, chúng tôi đưa đặc trưng hợp nhất cuối cùng vào lớp đầu ra để tạo ra ước tính.

4. Đánh giá

4.1. Dữ liệu

Chúng tôi sử dụng bộ dữ liệu ShanghaiTech A & B [5] để đánh giá và so sánh mô hình. Trong ShanghaiTech A, có 482 hình ảnh đám đông được thu thập từ Internet. Ba trăm hình ảnh tạo thành bộ dữ liệu huấn luyện, và phần còn lại tạo thành bộ dữ liệu kiểm tra. Các cảnh trong bộ dữ liệu này rất đông đúc, với số lượng trung bình khoảng 501. Ngoài ra, vì hình ảnh có độ phân giải khác nhau (giá trị chiều cao và chiều rộng dao động từ 182 đến 1024), việc huấn luyện các mô hình trên bộ dữ liệu này có thể phức tạp. Quy mô của ShanghaiTech B lớn hơn (tổng cộng 716 trường hợp; 400 để huấn luyện và 316 để kiểm tra), với số lượng trung bình khoảng 123. Hình ảnh từ bộ dữ liệu này được chụp từ góc nhìn giám sát trên một phố mua sắm và do đó ít đông đúc hơn. Ngoài ra, xét đến việc những hình ảnh này có độ phân giải cố định (768 × 1024), bộ dữ liệu này phù hợp hơn cho các ứng dụng thực tế, ví dụ: giám sát an toàn công cộng.

Trong nghiên cứu này, chúng tôi tiến hành các thử nghiệm mở rộng trên các chuẩn mực công khai sau đây để đếm số lượng người tham gia. Shanghaitech: Bộ dữ liệu này chứa 1.198 hình ảnh với 330.165 cá nhân được chú thích. Bộ dữ liệu bao gồm hai phần: Phần A bao gồm 482 hình ảnh về các cảnh đám đông đông đúc, trong đó 300 hình ảnh được sử dụng để huấn luyện và 182 hình ảnh để thử nghiệm, và Phần B bao gồm 716 hình ảnh về các cảnh đám đông thưa thớt, với 400 hình ảnh để tập luyện và phần còn lại để thử nghiệm.

4.2. Experimental Settings

Trong những năm gần đây, các hàm mất mát dựa trên lý thuyết xác suất, chẳng hạn như định lý Bayes và khoảng cách Wasserstein, đã được chứng minh là giúp các mô hình đạt được khả năng khái quát hóa đáng kể hơn và do đó ngày càng phổ biến. Chúng tôi sử dụng hàm mất mát DM-Count [17] để giám sát quá trình huấn luyện phương pháp của mình. Một trình tối ưu hóa Adam [38] với tốc độ học 1e-5 và kích thước lô hai được tận dụng để tối ưu hóa. Chúng tôi sử dụng các phân chia dữ liệu mặc định để so sánh công bằng với các phương pháp khác. Do một số ảnh có kích thước quá lớn, từ mỗi ảnh đầu vào, hai bản vá có kích thước 384 × 512 được cắt và sử dụng để huấn luyện. Mô hình của chúng tôi và quá trình huấn luyện của nó được triển khai trong khung PyTorch [39] 1.10, và nền tảng huấn luyện là một máy chủ với GPU NVIDIA và HĐH Ubuntu 22.04 LTS.

Bảng 1. *So sánh mô hình hoặc phương pháp của chúng tôi với các mô hình hiện đại có quy mô tương tự.*

	SHB			SHA	
Các phương pháp	MAE	RMSE	Các phương pháp	MAE	RMSE
CSRNet [7]	10.6	16.0	DoReFa [28]	80.02	124.1
CAN [8]	7.8	12.2	QAT [29]	75.5	128.09
BL [9]	7.7	12.7	L1Filter [30]	85.18	135.82
DM-Count [17]	7.4	11.8	CP [31]	82.05	130.65
Đề xuất	6.9	11.8	AGP [32]	78.51	125.83
			FitNets [33]	87.32	140.34
			DML [34]	85.23	138.1
			NST [35]	76.26	116.57
			AT [36]	74.65	127.06
			AB [37]	75.73	123.28
			Đề xuất	71.55	114.4

4.3. Kết quả thực nghiệm

Tiếp nối các nghiên cứu trước đây [6], [8], [17], chúng tôi áp dụng Sai số tuyệt đối trung bình (MAE) và Sai số bình phương trung bình căn bậc hai (RMSE) để đánh giá hiệu suất đếm đám đông về mặt định lượng. Cụ thể, chúng được định nghĩa như sau:

$$MAE = \frac{1}{N}\sum_{\{i=1\}}^{N} ||P_i - G_i||$$

$$RMSE = \sqrt{\left\{\frac{1}{N}\sum_{\{i=1\}}^{N} ||P_i - G_i||^2\right\}}$$

trong đó N là số lượng hình ảnh thử nghiệm, P_i và G_i là số lượng dự đoán và số lượng thực tế của i^{th} hình ảnh, tương ứng.

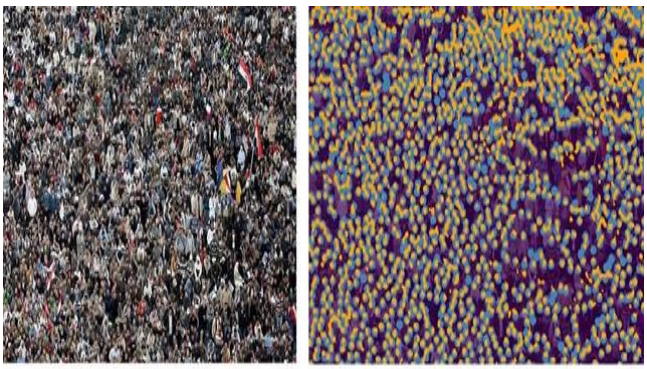
Phương pháp của chúng tôi được so sánh với các mô hình tiên tiến có độ phức tạp tính toán tương tự (được định lượng bằng số phép nhân và phép cộng liên quan đến suy luận trên ảnh RGB 1080 × 1920). Cụ thể, CSRNet [6], CAN [7], BL [8] và DM-Count [17], tất cả đều sử dụng VGG [10] làm bộ mã hóa, được đưa vào để so sánh. CSRNet sử dụng phép tích chập giãn nở trong giải mã để trích xuất độ nổi bật. Ngoài đặc điểm này, CAN còn bao gồm một khối gộp kim tự tháp không gian để tạo ra các đặc trưng đa tỷ lệ. BL và DM-Count không có cải tiến nào về kiến trúc mô hình, và đóng góp của chúng là các hàm mất mát đặc biệt mới dựa trên lý thuyết xác suất. Bảng 1 hiển thị các so sánh hiệu suất chi tiết.

Một mặt, đối với trường hợp ShanghaiTech A, một phần do khó khăn trong quá trình huấn luyện, kết quả phương pháp của chúng tôi không phải là tốt nhất nhưng vẫn tương đương. Chúng tôi tóm tắt hiệu suất của các thuật toán nén khác nhau trên Shanghaitech Part-A. Cụ thể, khi lượng tử hóa các tham số của CSRNet với 8 bit, DoReFa và QAT lần lượt đạt được MAE là 80,02/75,50. Khi chúng tôi sử dụng thiết lập chính thức của CP để cắt tỉa CSRNet, mô hình nén chấp nhận MAE là 82,05. Để duy trì cùng số lượng tham số của 1/4-CSRNet, L1Filter và AGP đã cắt tỉa 93,75% tham số, và MAE của chúng cao hơn 78.

Hơn nữa, sáu phương pháp chưng cất, bao gồm cả phương pháp của chúng tôi, được áp dụng để chưng cất CSRNet thành 1/4-CSRNet. Như có thể thấy, phương pháp của chúng tôi đạt hiệu suất tốt nhất với cả MAE và RMSE. Những ưu điểm về mặt định lượng và định tính này là do Intra-PT và Inter-RT được thiết kế riêng có thể

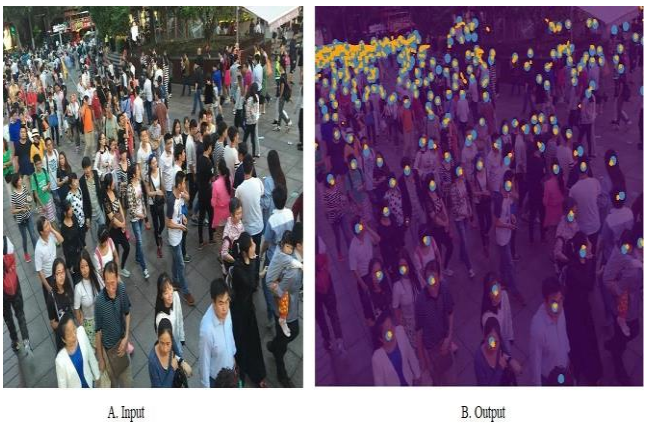
chưng cất hoàn toàn kiến thức của mạng lưới giáo viên. Phương pháp của chúng tôi phù hợp nhất với nhiệm vụ đếm đám đông trong số các thuật toán nén hiện có.

Mặt khác, tại ShanghaiTech B, phương pháp của chúng tôi vượt trội hơn các mô hình được đánh giá khác theo cả hai tiêu chí. Những kết quả này rất đáng chú ý, đặc biệt khi xét đến độ phức tạp tính toán thấp của mô hình chúng tôi.



Hình 6. *GT: 1603; Pred: 1634.79; RE: 1.98%*

Kết quả định tính của phương pháp đề xuất: Kết quả định tính của Mô hình đề xuất sử dụng phương pháp của chúng tôi để trực quan hóa ba bản đồ mật độ thực tế và ước tính của chúng. Hai hình ảnh trên cùng là từ ShanghaiTech A, và hai hình ảnh dưới cùng là từ ShanghaiTech B. Chú thích thực tế được đánh dấu bằng các chấm màu xanh lam, và các sắc thái màu cam biểu thị mật độ ước tính. Chúng tôi cũng báo cáo 'GT', tức là số lượng thực tế của 2 người trong ảnh, và 'Pred', tức là số lượng dự đoán. 'RE' biểu thị sai số tương đối tương ứng.



Hình 7. *GT: 300; Pred: 302.10; RE: 0.70%*

Bảng 2. *Tác động của phép tích chập từng điểm và số lớp tích chập giãn nở trong mô-đun giảm kênh.*

Point-wise Conv	Dilated Conv No.			MSE	RMSE
	2	1	0		
✓	✓			6.9	11.8
✓		✓		7.6	13.0
✓			✓	8.8	15.0
C	✓			9.5	15.9

Hiệu quả của các tính năng tương phản: Để xác nhận rằng các đặc điểm tương phản có thể tăng cường hiệu suất của mô hình, chúng tôi tạo ra một biến thể cho mô hình của mình, điểm khác biệt duy nhất so với phương pháp của chúng tôi là chiến lược hợp nhất đặc điểm. Trong biến thể này, trọng số được tạo trực tiếp từ các đặc điểm đã mã hóa x_{i+j} thay vì các tính năng tương phản c_{i+j} . Chúng tôi huấn luyện biến thể này trên ShanghaiTech B [6] với cùng một thiết lập. MAE và RMSE của biến thể này lần lượt là 7,6 và 12,9, cao hơn đáng kể so với phương pháp của chúng tôi (6,9 và 11,8).

Nghiên cứu cắt bỏ: Phần này chứng minh tính hiệu quả của mô hình giảm kênh được đề xuất trong Phần 3.3. Các thí nghiệm được tiến hành trên ShanghaiTech B [6]. Kết quả trong Bảng 2 chứng minh rằng lớp tích chập từng điểm và hai lớp tích chập giãn nở nối tiếp là không thể thiếu. Về mặt lý thuyết, chúng hoạt động bổ sung cho nhau trong việc giảm số lượng kênh và trích xuất độ nổi bật.

Bốn trường hợp từ bộ dữ liệu thử nghiệm của ShanghaiTech A & B được mô tả trong Hình 6, 7 và 8. Cột bên trái hiển thị các ảnh đầu vào ban đầu, và cột bên phải minh họa bản đồ mật độ thực tế và bản đồ mật độ dự đoán, được biểu thị lần lượt bằng các chấm xanh lam và sắc cam. Ví dụ ở hàng đầu tiên chứng minh rằng mô hình của chúng tôi có thể hoạt động tốt trong các

trường hợp mật độ cao. Mặc dù trong ảnh 600x900 này, có hơn 1.600 người, mô hình của chúng tôi vẫn đạt được dự đoán chính xác với sai số tương đối nhỏ (1,98%). Các hàng khác chứng minh rằng mô hình của chúng tôi có thể xử lý hiệu quả các thay đổi về tỷ lệ. Trong những ảnh này, đám đông ở phần dưới của cảnh có tỷ lệ lớn hơn, trong khi đám đông ở phần trên có tỷ lệ nhỏ hơn. Mô hình của chúng tôi có thể đưa ra dự đoán chính xác cho cả hai trường hợp với sai số nhỏ (lần lượt là 0,95%, 0,70% và 0,05%). Do đó, cả bốn trường hợp này đều khẳng định khả năng đếm mạnh mẽ của mô hình.

Mặc dù các bảng kết quả chủ yếu so sánh với các phương pháp tiêu biểu sử dụng bộ mã hóa VGG nhằm đảm bảo tính công bằng, trong những năm gần đây đã xuất hiện nhiều nghiên cứu mới tập trung vào các hàm mất mát tiên tiến, thích nghi miền và căn chỉnh đặc trưng cho bài toán đếm đám đông. So với các hướng tiếp cận này, phương pháp đề xuất trong bài báo tập trung vào thiết kế kiến trúc hiệu quả và khai thác tối đa các đặc trưng sẵn có trong bộ mã hóa, từ đó đạt được hiệu năng cạnh tranh mà không cần bổ sung các mô-đun đa thang độ phức tạp hoặc giám sát bổ sung. Điều này giúp mô hình phù hợp hơn cho các kịch bản ứng dụng thực tế với tài nguyên tính toán hạn chế.

5. Kết luận

Trong bài báo này, chúng tôi đã đề xuất một kiến trúc đếm đám đông mới, sử dụng thông minh nhiều đặc điểm được tạo ra trong quá trình mã hóa để xử lý hiện tượng méo góc nhìn. Không giống như các phương pháp hiện có, phương pháp của chúng tôi tránh được việc trích xuất thêm các đặc điểm đa tỷ lệ, do đó giảm đáng kể tổng khối lượng tính toán. Để đạt được mục đích này, chúng tôi cũng đã cải tiến cơ chế hợp nhất đa tỷ lệ hiện có và thiết kế một khối giảm kênh mới. Các thí nghiệm trên cơ sở dữ liệu ShanghaiTech đã chứng minh rằng phương pháp của chúng tôi có thể vượt trội hơn các phương pháp tiên tiến có độ phức tạp tính toán tương tự. Là một phần trong công việc tương lai, chúng tôi đang nghiên cứu việc tính toán thông tin ngữ cảnh trong các đặc điểm được hợp nhất trong quá trình giải mã. Dữ liệu như vậy có thể giúp xử lý hiệu quả hơn các thay đổi tỷ lệ, giống như quá trình hợp nhất giai đoạn đầu của các phần tử được mã hóa.

Tài liệu tham khảo

[1] T. Moranduzzo and F. Melgani, “Automatic car counting method for unmanned aerial vehicle images,” IEEE Transactions on Geoscience and Remote Sensing, vol. 52, no. 3, pp. 1635–1647, 2014.

[2] H. Lu, Z. Cao, Y. Xiao, B. Zhuang, and C. Shen, “Tasselnet: Counting maize tassels in the wild via local counts regression network,” CoRR, vol. abs/1707.02290, 2017. [Online]. Available: http://arxiv.org/abs/1707.02290

[3] I. J. C. Valencia, E. P. Dadios, A. M. Fillone, J. C. V. Puno, R. G. Baldovino, and R. K. C. Billones, “Vision-based crowd counting and social distancing monitoring using tiny-yolov4 and deepsort,” in 2021 IEEE International Smart Cities Conference (ISC2), 2021, pp. 1–7.

[4] V. Lempitsky and A. Zisserman, “Learning to count objects in images,” in Proceedings of the 23rd International Conference on Neural Information Processing Systems - Volume 1, ser. NIPS’10. Red Hook, NY, USA: Curran Associates Inc., 2010, p. 1324–1332.

[5] A. B. Chan and N. Vasconcelos, “Bayesian poisson regression for crowd counting,” in 2009 IEEE 12th International Conference on Computer Vision, 2009, pp. 545–551.

[6] Y. Zhang, D. Zhou, S. Chen, S. Gao, and Y. Ma, “Single-image crowd counting via multi-column convolutional neural network,” in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 589–597.

[7] Y. Li, X. Zhang, and D. Chen, “Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes,” CoRR, vol.

abs/1802.10062, 2018. [Online]. Available: http://arxiv.org/abs/1802.10062

[8] W. Liu, M. Salzmann, and P. Fua, "Context-aware crowd counting," CoRR, vol. abs/1811.10452, 2018. [Online]. Available: http://arxiv.org/abs/1811.10452

[9] Z. Ma, X. Wei, X. Hong, and Y. Gong, "Bayesian loss for crowd count estimation with point supervision," CoRR, vol. abs/1908.03684, 2019. [Online]. Available: http://arxiv.org/abs/1908.03684

[10] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014. [Online]. Available: https://arxiv.org/abs/1409.1556

[11] X. Cao, Z. Wang, Y. Zhao, and F. Su, "Scale aggregation network for accurate and efficient crowd counting," in Computer Vision – ECCV 2018, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Cham: Springer International Publishing, 2018, pp. 757–773.

[12] P. Thanasutives, K. Fukui, M. Numao, and B. Kijsirikul, "Encoder-decoder based convolutional neural networks with multi-scale-aware modules for crowd counting," CoRR, vol. abs/2003.05586, 2020. [Online]. Available: https://arxiv.org/abs/2003.05586

[13] M. Wang, H. Cai, J. Zhou, and M. Gong, "Stochastic multi-scale aggregation network for crowd counting," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 2008–2012.

[14] B. Li, H. Huang, A. Zhang, P. Liu, and C. Liu, "Approaches on crowd counting and density estimation: A review," *Pattern Anal. Appl.*, vol. 24, no. 3, p. 853–874, aug 2021. [Online]. Available: https://doi.org/10.1007/s10044-021-00959-z

[15] M. Wo´zniak, J. Si-lka, and M. Wieczorek, "Deep learning based crowd counting model for drone assisted systems," in *Proceedings of the 4th ACM MobiCom Workshop on Drone Assisted Wireless Communications for 5G and Beyond*, ser. DroneCom '21. New York, NY, USA: Association for Computing Machinery, 2021, p. 31–36. [Online]. Available: https://doi.org/10.1145/3477090.3481054.

[16] S. Gong, S. Zhang, J. Yang, D. Dai, and B. Schiele, "Bi-level alignment for cross-domain crowd counting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 7542–7550.

[17] B. Wang, H. Liu, D. Samaras, and M. H. Nguyen, "Distribution matching for crowd counting," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 1595–1607. [Online]. Available: https://proceedings.neurips.cc/paper/2020/file/118bd558033a1016fcc82560c65cca5f-Paper.pdf

[18] S.-F. Lin, J.-Y. Chen, and H.-X. Chao, "Estimation of number of people in crowded scenes using perspective transformation," *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, vol. 31, no. 6, pp. 645–654, 2001.

[19] M. Li, Z. Zhang, K. Huang, and T. Tan, "Estimating the number of people in crowded scenes by mid based foreground segmentation and head-shoulder detection," in *2008 19th International Conference on Pattern Recognition*, 2008, pp. 1–4.

[20] W. Ge and R. T. Collins, "Marked point processes for crowd counting," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 2913–2920.

[21] L. Wen, D. Du, P. Zhu, Q. Hu, Q. Wang, L. Bo, and S. Lyu, "Detection, tracking, and counting meets drones in crowds: A benchmark," *CoRR*, vol. abs/2105.02440, 2021. [Online]. Available: https://arxiv.org/abs/2105.02440

[22] K. Chen, C. C. Loy, S. Gong, and T. Xiang, "Feature mining for localised crowd counting," in *British Machine Vision Conference*, 2012.

[23] K. Chen, S. Gong, T. Xiang, and C. C. Loy, "Cumulative attribute space for age and crowd density estimation," in *2013 IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 2467–2474.

[24] V. A. Sindagi and V. M. Patel, "Generating high-quality crowd density maps using contextual pyramid cnns," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 1879–1888.

[25] D. On˜oro-Rubio and R. J. Lo´pez-Sastre, "Towards perspective-free object counting with deep learning," in *Computer Vision – ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham: Springer International Publishing, 2016, pp. 615–629.

[26] D. Kang, D. Dhar, and A. Chan, "Incorporating side information by adaptive convolution," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper/2017/file/e7e23670481ac78b3c4122a99ba60573-Paper.pdf

[27] S. R. J. S. Kaiming He, Xiangyu Zhang, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *CoRR*, vol. abs/1406.4729, 2014. [Online]. Available: http://arxiv.org/abs/1406.4729

[28] S. Zhou, Z. Ni, X. Zhou, H. Wen, Y. Wu, and Y. Zou, "Dorefa-net: Training low bitwidth convolutional neural networks with low bitwidth gradients," *CoRR*, vol. abs/1606.06160, 2016. [Online]. Available: http://arxiv.org/abs/1606.06160

[29] B. Jacob, S. Kligys, B. Chen, M. Zhu, M. Tang, A. G. Howard, H. Adam, and D. Kalenichenko, "Quantization and training of neural networks for efficient integer-arithmetic-only inference," *CoRR*, vol. abs/1712.05877, 2017. [Online]. Available: http://arxiv.org/abs/1712.05877

[30] H. Li, A. Kadav, I. Durdanovic, H. Samet, and H. P. Graf, "Pruning filters for efficient convnets," *CoRR*, vol. abs/1608.08710, 2016. [Online]. Available: http://arxiv.org/abs/1608.08710

[31] Y. He, X. Zhang, and J. Sun, "Channel pruning for accelerating very deep neural networks," *CoRR*, vol. abs/1707.06168, 2017. [Online]. Available: http://arxiv.org/abs/1707.06168

[32] M. Zhu and S. Gupta, "To prune, or not to prune: exploring the efficacy of pruning for model compression," *arXiv e-prints*, p. arXiv:1710.01878, Oct. 2017.

[33] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio, "Fitnets: Hints for thin deep nets," 2014. [Online]. Available: https://arxiv.org/abs/1412.6550

[34] Y. Zhang, T. Xiang, T. M. Hospedales, and H. Lu, "Deep mutual learning," *CoRR*, vol. abs/1706.00384, 2017. [Online]. Available: http://arxiv.org/abs/1706.00384

[35] Z. Huang and N. Wang, "Like what you like: Knowledge distill via neuron selectivity transfer," *CoRR*, vol. abs/1707.01219, 2017. [Online]. Available: http://arxiv.org/abs/1707.01219

[36] S. Zagoruyko and N. Komodakis, "Paying more attention to attention:Improving the performance of convolutional neural networks via attention transfer," *CoRR*, vol. abs/1612.03928, 2016. [Online]. Available: http://arxiv.org/abs/1612.03928

[37] B. Heo, M. Lee, S. Yun, and J. Y. Choi, "Knowledge transfer via distillation of activation boundaries formed by hidden neurons," *CoRR*, vol. abs/1811.03233, 2018. [Online]. Available: http://arxiv.org/abs/1811.03233

[38] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *International Conference on Learning Representations*, 12 2014.

[39] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Ko¨pf, E. Z. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," *CoRR*, vol.abs/1912.01703, 2019. [Online]. Available: http://arxiv.org/abs/1912.01703