**A. Mạng Cảm Biến Lai Đa Phương Thức (Hybrid Sensing Network - HSN)**

Để đáp ứng nhu cầu giám sát đám đông trong các khu vực quy mô lớn, một mạng HSN gồm nhiều thiết bị WiFi sniffer và một số lượng nhỏ camera đã được xây dựng, như minh họa ở Hình 1. HSN được triển khai linh hoạt phù hợp với điều kiện thực tế, không cần tuân thủ các điều kiện tiên quyết nghiêm ngặt, nhờ đó tăng tính thực tiễn và giảm chi phí bổ sung.

Ví dụ, HSN có thể được xây dựng bằng cách thêm một vài camera vào mạng WiFi sniffer đã có sẵn, hoặc triển khai thêm nhiều thiết bị WiFi giá rẻ cho hệ thống giám sát video đã có.

Các sniffer WiFi có thể là thiết bị chuyên dụng hoặc bộ phát WiFi lập trình được. Mỗi thiết bị chạy một script để đánh dấu thời gian thu thập dữ liệu và gửi định kỳ lên máy chủ. Camera có thể là camera giám sát phổ thông, truyền video theo thời gian thực về máy chủ thông qua mạng có dây. Việc đặt camera nên ưu tiên các khu vực đông người để tăng độ chính xác khi kết hợp nhiều phương thức.

Do phạm vi quan sát của camera bị giới hạn, phần lớn khu vực sẽ được WiFi bao phủ. Các vùng mà hai phương thức cùng bao phủ sẽ cung cấp dữ liệu cho việc học sự tương quan không gian–thời gian giữa tín hiệu toàn cục (WiFi) và tín hiệu cục bộ (video). Từ đó, việc sử dụng cross-attention sẽ giúp học mối quan hệ giữa hai phương thức, cải thiện kết quả đếm người.

### **B. Phương pháp đếm đám đông đa phương thức được đề xuất**

Phần này giải thích cách tận dụng dữ liệu đa phương thức được thu thập bởi HSN để thực hiện đếm đám đông chính xác.

#### ****1) Tổng quan:****

Phương pháp đếm đám đông đa phương thức được đề xuất chủ yếu gồm hai giai đoạn, như minh họa ở Hình 2.

* **Căn chỉnh và tiền xử lý dữ liệu đa phương thức**:  
  Thứ nhất, để khắc phục tính ngẫu nhiên của cảm biến WiFi thụ động (PWS) và đáp ứng yêu cầu định dạng đầu vào cho mô-đun trích xuất đặc trưng, chúng tôi lần lượt sử dụng cơ chế cửa sổ trượt và bản đồ mật độ cảm nhận WiFi (WDM) để lọc dữ liệu WiFi trong một khoảng thời gian xác định, sau đó biểu diễn nó dưới dạng đại diện vững chắc cho phương thức WiFi [46].  
  Thứ hai, để căn chỉnh theo thời gian và giảm chi phí tính toán, chúng tôi lấy mẫu các khung hình chính, thực hiện chọn vùng quan tâm (ROI), cắt và thay đổi kích thước dữ liệu video với cùng một bước thời gian, nhằm thu được các khung hình hiệu quả.
* **Trích xuất, tối ưu hóa và kết hợp đặc trưng đa phương thức dựa trên HDANet**:  
  Thứ nhất, theo mức độ trừu tượng của các đặc trưng liên quan đến đám đông trong dữ liệu đã được tiền xử lý của hai phương thức, các mạng CNN với kiến trúc khác nhau được sử dụng để trích xuất các đặc trưng đám đông tương ứng.  
  Thứ hai, để khai thác đầy đủ mối tương quan ngữ cảnh nội phương thức, cơ chế **self-attention** được sử dụng để tối ưu hóa các đặc trưng trong từng phương thức.  
  Thứ ba, cơ chế **cross-attention** và mạng nơ-ron truyền thẳng (FFN) được sử dụng để khớp và khái quát hóa các mối quan hệ giữa các phương thức.  
  Cuối cùng, một mạng perceptron nhiều lớp (MLP) được sử dụng để chọn lọc thích nghi các đặc trưng self-attention và cross-attention, và đưa ra đầu ra đếm đám đông ở phạm vi cục bộ và/hoặc toàn cục.

Dưới đây là bản dịch **sát nghĩa** phần **III-B. The Proposed Multi-Modal Crowd Counting Method** trong bài báo:

### **B. Phương pháp đếm đám đông đa phương thức được đề xuất**

Phần này giải thích cách tận dụng dữ liệu đa phương thức được thu thập bởi HSN để thực hiện đếm đám đông chính xác.

#### ****1) Tổng quan:****

Phương pháp đếm đám đông đa phương thức được đề xuất chủ yếu gồm hai giai đoạn, như minh họa ở Hình 2.

* **Căn chỉnh và tiền xử lý dữ liệu đa phương thức**:  
  Thứ nhất, để khắc phục tính ngẫu nhiên của cảm biến WiFi thụ động (PWS) và đáp ứng yêu cầu định dạng đầu vào cho mô-đun trích xuất đặc trưng, chúng tôi lần lượt sử dụng cơ chế cửa sổ trượt và bản đồ mật độ cảm nhận WiFi (WDM) để lọc dữ liệu WiFi trong một khoảng thời gian xác định, sau đó biểu diễn nó dưới dạng đại diện vững chắc cho phương thức WiFi [46].  
  Thứ hai, để căn chỉnh theo thời gian và giảm chi phí tính toán, chúng tôi lấy mẫu các khung hình chính, thực hiện chọn vùng quan tâm (ROI), cắt và thay đổi kích thước dữ liệu video với cùng một bước thời gian, nhằm thu được các khung hình hiệu quả.
* **Trích xuất, tối ưu hóa và kết hợp đặc trưng đa phương thức dựa trên HDANet**:  
  Thứ nhất, theo mức độ trừu tượng của các đặc trưng liên quan đến đám đông trong dữ liệu đã được tiền xử lý của hai phương thức, các mạng CNN với kiến trúc khác nhau được sử dụng để trích xuất các đặc trưng đám đông tương ứng.  
  Thứ hai, để khai thác đầy đủ mối tương quan ngữ cảnh nội phương thức, cơ chế **self-attention** được sử dụng để tối ưu hóa các đặc trưng trong từng phương thức.  
  Thứ ba, cơ chế **cross-attention** và mạng nơ-ron truyền thẳng (FFN) được sử dụng để khớp và khái quát hóa các mối quan hệ giữa các phương thức.  
  Cuối cùng, một mạng perceptron nhiều lớp (MLP) được sử dụng để chọn lọc thích nghi các đặc trưng self-attention và cross-attention, và đưa ra đầu ra đếm đám đông ở phạm vi cục bộ và/hoặc toàn cục.

#### ****2) Tiền xử lý dữ liệu đa phương thức:****

Chúng tôi sẽ mô tả phương pháp tiền xử lý dữ liệu của hai phương thức riêng biệt như sau:





















