**Vtab-1k Dataset: Visual Task Adaptation Benchmark**

* **VTAB-1K bao gồm 19 nhiệm vụ, bao quát một loạt các lĩnh vực và ngữ nghĩa. Chúng được nhóm thành ba bộ: NATURAL, SPECIALIZED và STRUCTURED. Đồng thời, mỗi dataset chỉ bao gồm 800 ảnh cho quá trình training và 200 ảnh cho validation:**
  + **Nhóm NATURAL** đại diện cho các vấn đề thị giác cổ điển. Những nhiệm vụ này chứa ảnh tự nhiên được chụp bằng các máy ảnh tiêu chuẩn. Các lớp có thể đại diện cho đối tượng chung, chi tiết hóa hoặc trừu tượng. Nhóm này bao gồm:
    - Caltech101: Bao gồm 101 lớp, phân loại đa dạng về các đối tượng như: máy bay trực thằng, voi, ghế, cua,…
    - CIFAR-100: Bao gồm 20 SuperClass, tổng cộng có 100 lớp phân loại đa dạng về: các loại động vật (ăn cỏ, dưới nước, lưỡng cư, …), con người, xe cộ, …
    - DTD: Bao gồm 47 lớp. Là bộ dữ liệu được sử dụng chủ yếu cho nhiệm vụ phân loại texture. Sự đa dạng của loại texture trong DTD bao gồm cả vật liệu tự nhiên (ví dụ: cây cỏ, da động vật, đá) và vật liệu nhân tạo (ví dụ: giấy, kim loại, nhựa).
    - Flowers102: Bao gồm 102 lớp đại diện cho 102 loại hoa như: Rose, Sunflower, Tulip, Daisy, Daffodil, Iris, …
    - Pets (Oxford-IIIT Pets): Bao gồm 37 lớp. Được sử dụng để thực hiện nhiệm vụ phân loại ảnh động vật, chủ yếu là hình ảnh của các giống thú cưng như chó (25 giống) và mèo (12 giống).
    - Sun397: Bộ dữ liệu là 1 cơ sở dữ liệu bao gồm 899 lớp và 130.519 hình ảnh. Tuy nhiên, nhóm tác giả chỉ sử dụng 397 lớp có mẫu tốt nhất để phục vụ cho việc phân loại.
    - SVHN: Bao gồm 10 lớp, mỗi lớp đại diện cho 1 số từ 0 đến 9. Được sử dụng để nhận diện số từ các ảnh chụp đường phố. Tức là được dùng để phân loại các chữ số từ 0 đến 9 được cắt ra từ hình ảnh của biển số nhà.
  + **Nhóm SPECIALIZED** cũng chứa ảnh của thế giới, nhưng được chụp thông qua thiết bị chuyên gia. Những bức ảnh này có những sự không biến đổi khác so với những nhiệm vụ NATURAL. Tuy nhiên, con người vẫn nhận biết cấu trúc bên trong, do đó, các biểu diễn thị giác chung cũng nên bắt kịp các khái niệm thị giác. Chúng ta có hai nhóm con: cảm biến từ xa và y tế.
    - Cảm biến từ xa bao gồm Resisc45 và EuroSAT: ảnh không gian của trái đất được chụp bằng vệ tinh hoặc chụp ảnh từ không trung.
      * Resisc45: Bao gồm 45 lớp: Bộ dữ liệu được thiết kế đặc biệt để phục vụ cho việc nhận diện các loại đất đai và vùng đô thị từ ảnh chụp từ vệ tinh.
      * EuroSAT: Bao gồm 10 lớp. Là một bộ dữ liệu được dùng để phân loại đất đai hay phân loại độ che phủ đất. Bộ dữ liệu dựa trên hình ảnh chụp từ vệ tinh Sentinel-2 ở các khu vực đa dạng ở Châu Âu, phân loại các loại đất đai, chẳng hạn như rừng, cánh đồng, thành phố, hồ, đường băng, sân golf, rừng cây lau, …
    - Y tế bao gồm Patch Camelyon, phát hiện métastases từ hình ảnh kính hiển vi, và Diabetic Retinopathy, phân loại retinopathy từ hình ảnh đáy mắt.
      * Patch Camelyon: Phân loại 2 lớp: "normal" (bình thường) và "tumor" (ung thư). Được sử dụng để đào tạo và đánh giá các mô hình học máy trong việc phân loại ảnh histopathology (ảnh uốn nắn) được trích xuất từ ​​các lần quét mô bệnh học của các phần hạch bạch huyết.
      * Diabetic Retinopathy: Phân loại 5 lớp: "Không có bệnh" (No DR), "DR nhẹ" (Mild DR), "DR trung bình" (Moderate DR), "DR nặng" (Severe DR), và "DR rất nặng" (Proliferative DR). Bao gồm các hình ảnh chụp võng mạc liên quan bệnh tiểu đường.
  + **Nhóm STRUCTURED** đánh giá sự hiểu biết về cấu trúc của một cảnh, ví dụ như đếm đối tượng hoặc dự đoán chiều sâu 3D. Hầu hết những nhiệm vụ này được tạo ra từ môi trường mô phỏng, cấu trúc của chúng dễ dàng để con người xác định, nhưng lĩnh vực của chúng khác biệt lớn so với các tập dữ liệu như ImageNet. Những nhiệm vụ này được thiết kế như là một bước tiến về các biểu diễn hữu ích cho kiểm soát giác quan. Chúng ta bao gồm:
    - Clevr: Hình dạng đơn giản được vẽ trong một cảnh 3D, với các đặc trưng thuộc một tập hợp bốn thuộc tính: 2 kích thước: lớn, nhỏ; 3 hình dạng: hình vuông, hình trụ, hình cầu; 2 loại vật liệu: cao su, kim loại; 8 loại màu sắc: xám, xanh, nâu, vàng, đỏ, xanh lá cây, tím, xanh lơ, tạo ra 96 sự kết hợp duy nhất. Dataset có hai nhiệm vụ:
      * Clevr-Count: đếm
      * Clevr-Dist: dự đoán chiều sâu
    - dSprites: Hình dạng đơn giản đen/trắng được vẽ trong không gian 2D được tạo theo thủ tục từ 6 yếu tố tiềm ẩn độc lập với sự thật cơ bản. Các yếu tố này là màu sắc , hình dạng , tỷ lệ , xoay , vị trí x và y của một sprite. Với hai nhiệm vụ:
      * dSpr-Loc: dự đoán vị trí
      * dSpr-Ori: định hướng.
    - SmallNORB: Bộ dữ liệu này là một tập dữ liệu được sử dụng để nhận diện đối tượng 3D dựa trên hình dạng. Nó bao gồm hình ảnh của 50 đồ chơi thuộc 5 danh mục chung: động vật bốn chân, hình người, máy bay, xe tải và ô tô. Các đối tượng được chụp ảnh bởi hai máy ảnh dưới 6 điều kiện ánh sáng khác nhau, 9 độ cao (từ 30 đến 70 độ mỗi 5 độ), và 18 hướng (từ 0 đến 340 mỗi 20 độ). Với hai nhiệm vụ:
      * sNORB-Azim: dự đoán hướng đối tượng
      * sNORB-Ele: hướng camera.
    - DMLab: Khung hình từ mê cung 3D được vẽ. Nhiệm vụ liên quan đến việc dự đoán thời gian mà một đại lý RL đã được đào tạo cần để điều hướng đến một đối tượng.
    - KITTI: Khung hình được chụp từ góc nhìn của người lái xe ô tô. Chúng ta dự đoán sâu của phương tiện gần nhất.

**Nhóm data Natural**

* **Cifar100:**
  + Bộ dữ liệu này bao gồm 20 SuperClass phân loại đa dạng về:
    - Động vật có vú sống dưới nước: hải ly, cá heo, rái cá, hải cẩu, cá voi
    - Cá: cá cảnh, cá dẹt, cá đuối, cá mập, cá hồi
    - Hoa: hoa lan, anh túc, hoa hồng, hoa hướng dương, hoa tulip
    - Hộp đựng thực phẩm: chai, bát, lon, cốc, đĩa
    - trái cây và rau quả: táo, nấm, cam, lê, ớt ngọt
    - thiết bị điện gia dụng: đồng hồ, bàn phím máy tính, đèn, điện thoại, tivi
    - Nội thất gia đình: giường, ghế, đi văng, bàn, tủ quần áo
    - côn trùng: ong, bọ cánh cứng, bướm, sâu bướm, gián
    - động vật ăn thịt lớn: gấu, báo, sư tử, hổ, sói
    - Những thứ ngoài trời nhân tạo lớn: cầu, lâu đài, nhà, đường, tòa nhà chọc trời
    - cảnh thiên nhiên ngoài trời rộng lớn: mây, rừng, núi, đồng bằng, biển
    - Động vật ăn tạp lớn và động vật ăn cỏ: lạc đà, gia súc, tinh tinh, voi, chuột túi
    - động vật có vú cỡ trung bình: cáo, nhím, possum, gấu trúc, chồn hôi
    - động vật không xương sống không côn trùng: cua, tôm hùm, ốc, nhện, giun
    - dân tộc: em bé, trai, gái, đàn ông, phụ nữ
    - Reptiles: cá sấu, khủng long, thằn lằn, rắn, rùa
    - động vật có vú nhỏ: chuột đồng, chuột, thỏ, chuột chù, sóc
    - Cây: phong, sồi, cọ, thông, liễu
    - Xe cộ 1: xe đạp, xe buýt, xe máy, xe bán tải, tàu hỏa
    - Xe cộ 2: máy cắt cỏ, tên lửa, xe điện, xe tăng, máy kéo
  + Phân loại 100 lớp, mỗi lớp có 600 ảnh (500 training, 100 testing)
* **Caltech101:**
  + Là bộ dữ liệu phân loại đa dạng về các đối tượng như: máy bay trực thằng, voi, ghế, cua,…
  + Bao gồm 101 lớp, đối với mỗi loại đối tượng, có khoảng 40 đến 800 hình ảnh, trong khi hầu hết các lớp có khoảng 50 hình ảnh. Độ phân giải của hình ảnh khoảng 300×200 pixel.
* **DTD:**
  + Là bộ dữ liệu được sử dụng chủ yếu cho nhiệm vụ phân loại texture. Sự đa dạng của loại texture trong DTD bao gồm cả vật liệu tự nhiên (ví dụ: cây cỏ, da động vật, đá) và vật liệu nhân tạo (ví dụ: giấy, kim loại, nhựa).
  + Phân loại 47 lớp, mỗi lớp gồm 120 ảnh (chia đều cho 3 tập train/val/test)
* **Flower102:** 
  + Là bộ dữ liệu phân loại đa dạng về các loại hoa như: Rose, Sunflower, Tulip, Daisy, Daffodil, Iris
  + Phân loại 102 lớp đại diện cho 102 loại hoa, mỗi lớp bao gồm từ 40 đến 258 hình ảnh.
* **Oxford-IIIT Pets:**
  + Bộ dữ liệu Oxford-IIIT - Pet được sử dụng để thực hiện nhiệm vụ phân loại ảnh động vật, chủ yếu là hình ảnh của các giống thú cưng như chó và mèo.
  + Phân loại 37 lớp, bao gồm 12 giống mèo và 25 giống chó. Có khoảng 200 hình ảnh cho mỗi lớp.
* **SVHN:**
  + Bộ dữ liệu được sử dụng để nhận diện số từ các ảnh chụp đường phố. Tức là được dùng để phân loại các chữ số từ 0 đến 9 được cắt ra từ hình ảnh của biển số nhà.
  + Phân loại 10 lớp, 1 lớp đại diện cho 1 số từ 0 đến 9. Bao gồm 73257 cho training và 26032 cho testing.
* **Sun397:**
  + Bộ dữ liệu là 1 cơ sở dữ liệu bao gồm 899 lớp và 130.519 hình ảnh. Tuy nhiên, nhóm tác giả chỉ sử dụng 397 lớp có mẫu tốt nhất để phục vụ cho việc phân loại.
  + Phân loại 397 lớp, mỗi lớp có ít nhất 100 ảnh và bao gồm tổng cộng 108.754 hình ảnh; 76,128 ảnh cho training, 10,875 cho validation and 21,750 cho testing.

**Nhóm Specialized:**

* **Patch Camelyon:**
  + Là bộ dữ liệu thuộc lĩnh vực y học, được sử dụng để đào tạo và đánh giá các mô hình học máy trong việc phân loại ảnh histopathology (ảnh uốn nắn) được trích xuất từ ​​các lần quét mô bệnh học của các phần hạch bạch huyết.
  + Phân loại 2 lớp: "normal" (bình thường) và "tumor" (ung thư). Bao gồm 327.680 hình ảnh màu; 262,144 ảnh cho training, 32,768 cho validation and 32,768 cho testing.
* **EuroSAT:**
  + Là một bộ dữ liệu được dùng để phân loại đất đai hay phân loại độ che phủ đất. Bộ dữ liệu dựa trên hình ảnh chụp từ vệ tinh Sentinel-2 ở các khu vực đa dạng ở Châu Âu, phân loại các loại đất đai, chẳng hạn như rừng, cánh đồng, thành phố, hồ, đường băng, sân golf, rừng cây lau, …
  + Được dùng để phân loại 10 lớp với tổng cộng 27000 ảnh màu.
* **Resisc45:**
  + Bộ dữ liệu được thiết kế đặc biệt để phục vụ cho việc nhận diện các loại đất đai và vùng đô thị từ ảnh chụp từ vệ tinh.
  + Bộ dữ liệu này chứa 31.500 hình ảnh, bao gồm 45 lớp cảnh với 700 hình ảnh trong mỗi lớp.

**Nhóm Structured**

* **Clevr-Count:** **Compositional Language and Elementary Visual Reasoning**
  + Là một bộ dữ liệu Trả lời câu hỏi trực quan tổng hợp. Nó chứa hình ảnh của các đối tượng kết xuất 3D; Mỗi hình ảnh đi kèm với một số câu hỏi có tính sáng tác cao thuộc các loại khác nhau. Các loại đó rơi vào 5 lớp nhiệm vụ: Exist, Count, So sánh số nguyên, Query Attribute và Compare Attribute.
  + Bộ dữ liệu CLEVR bao gồm một tập train với 70,000 hình ảnh và 700,000 câu hỏi, một tập val với 15,000 hình ảnh và 150,000 câu hỏi, cùng một tập test với 15,000 hình ảnh và 150,000 câu hỏi về đối tượng, câu trả lời, đồ thị cảnh và chương trình chức năng cho tất cả hình ảnh và câu hỏi trong tập huấn luyện và xác nhận.
* **Clevr-Dist:** ????
* **DMLab:**
  + Bộ dữ liệu Dmlab chứa các khung được quan sát bởi tác nhân hoạt động trong môi trường DeepMind Lab, được chú thích bằng khoảng cách giữa tác nhân và các đối tượng khác nhau có trong môi trường. Mục tiêu là để đánh giá khả năng của một mô hình trực quan suy luận về khoảng cách từ đầu vào trực quan trong môi trường 3D.
  + Bộ dữ liệu Dmlab bao gồm các hình ảnh màu 360x480 trong 6 lớp. Các lớp lần lượt là {gần, xa, rất xa} x {phần thưởng tích cực, phần thưởng tiêu cực}. ('test’: 22,735, 'train: 65.550; 'validation': 22,628)
* **KITTI-Dist: Karlsruhe Institute of Technology and Toyota Technological Institute**
  + là một trong những bộ dữ liệu phổ biến nhất để sử dụng trong robot di động và lái xe tự động. Nó bao gồm hàng giờ các kịch bản giao thông được ghi lại với nhiều phương thức cảm biến, bao gồm RGB độ phân giải cao, camera âm thanh nổi thang độ xám và máy quét laser 3D.
  + Phân loại 11 lớp: tòa nhà, cây, bầu trời, ô tô, biển báo, đường, người đi bộ, hàng rào, cột, vỉa hè và người đạp xe.
* **dSpr-Loc:**
* **dSpr-Ori:**
* **sNORB-Azim**
* **sNORB-Ele**