Project 1: Object Detection

Brief Description

Apply object detection algorithms to rare and novel classes that do not exist in publicly available datasets like COCO. Given input images, the algorithms should return the bounding boxes and class labels of the target object.

Datasets

You should create a customized dataset, which is suitable for the target objects. Ensure the training and testing sets are well split.

Difficulty Level (1-5): 3

Project 2: Spam Email Detection/Classification

Brief Description

Design a spam email classification algorithm. You can use any publicly available datasets or create your own datasets. Given a new email from your intl.zju.edu.cn address, your algorithm can distinguish whether it is a spam.

Difficulty Level (1-5): 3

Project 3: Vision Encoding meets Language Instruction

Background

Current unified multi-task models often fail to focus on task-relevant features. Use frozen visual encoder is not enough for all the task. Even for one task, language instruction can help with visual encoding.

Description

Current attention algorithms (e.g., self-attention) are stimulus-driven and highlight all the salient objects in an image. However, intelligent agents like humans often guide their attention based on the high-level task at hand, focusing only on task-related objects. This ability of task guided top-down attention provides task-adaptive representation and helps the model generalize to various tasks.

ViT is easy to extend due to the token design. AbSViT, which is a top-down modulated ViT model that variationally approximates Analysis-by-Synthesis (AbS), and achieves controllable top-down attention. For real-world applications, AbSViT consistently improves over baselines on Vision-Language tasks such as VQA and zero-shot retrieval where language guides the top-down attention.

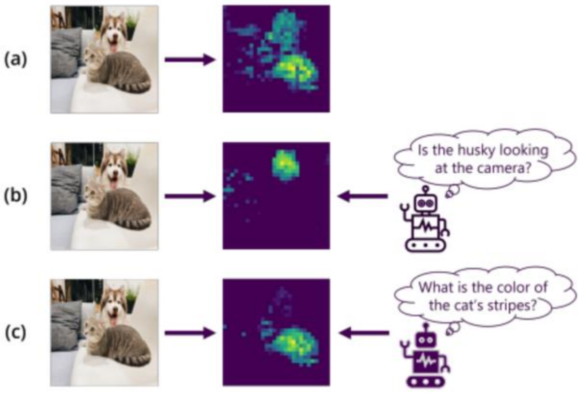

(a)
(b)
(c)
Is the husky looking at the camera?
What is the color of the cat's stripes?

Figure: Top-Down Visual Attention from Analysis by Synthesis

- Collect a task-specific attention dataset: Given an image and a task, mark the attention map provided by AbsViT. (For the same picture, when performing different tasks, the attention map is often different)

Expected outcome

Design a task-specific attention dataset: given an image and a task, mark the attention map required by the model. You should use some methods that can automatically or with very limited human labeling cost to mark the attention maps.

Dependencies
https://github.com/bfshi/AbSViT

Difficulty Level (1-5): 3

Project 4: Sports Action Quality Assessment

Background
Sports Action Quality Assessment is a computer vision task that involves evaluating the quality or performance of human actions or activities captured in video recordings. It aims to automatically assess how well a particular action, such as a sports move or a dance routine, has been executed based on visual cues within the video. This task is essential for applications like sports coaching, physical therapy, and motion analysis, as it provides valuable insights into the proficiency and correctness of human actions in a quantitative manner.

Description
Given a video recording of a human performing a specific sports action or activity which can be captured from various sources, such as sports events, training sessions, or physical therapy sessions.

You need to train a model to output is a numerical score or rating that quantitatively assesses the quality or performance of the action depicted in the video. This score indicates how well the action has been executed based on visual cues.

Dataset

AQA-7 consists of 1106 action samples from seven actions with quality scores as measured by expert human judges. 803 videos are used for training and 303 videos used for testing.

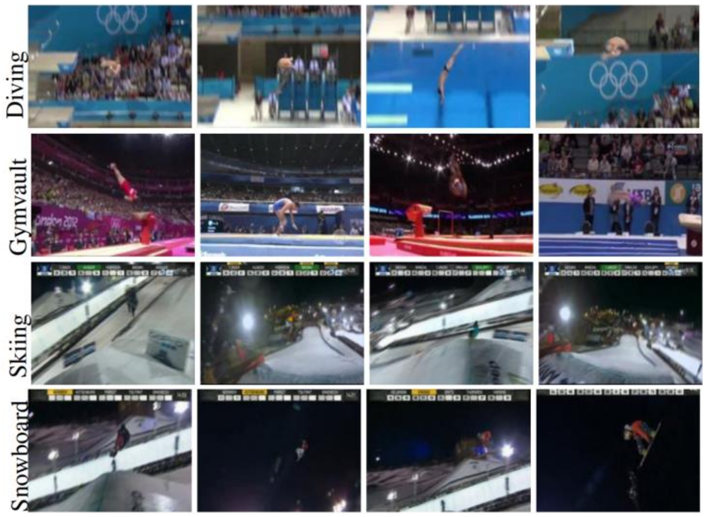


Figure AQA-7 dataset

URL: http://rtis.oit.unlv.edu/datasets.html

Expected Outcome

Design a sports action quality assessment framework to solve this problem, and evaluate this task using the regular accuracy metric.

Difficulty Level (1-5): 3

Project 5: Low-quality Video Classification on Jersy Number Recognition

Background

Low-quality video classification refers to the process of identifying and categorizing videos based on their quality. In the context of machine learning and computer vision, this typically involves training models to detect and classify videos that might be blurry, pixelated, corrupted, shaky, or have other quality-degrading characteristics. This task can be useful for consumer applications as it helps to build systems that can accurately identify and recognize players in real-time during a soccer game. This information can be used to create interactive experiences for fans, such as player statistics, player tracking, and in-game analysis. Additionally, this technology can also be used in broadcast production to add player labels and enhance the viewing experience.

Description

Given short video tracklets of soccer players that can be a few hundreds frames long, the participants are asked to identify the jersey number of each player (Players with non visible jersey numbers are annotated with the "-1" value). The task is challenging because of the low quality of the thumbnails (low resolution, high motion blur) and because the jersey numbers might be visible on a very small subset of the whole tracklet. All jersey numbers from 1 to 99, and one extra class when there is no jersey number visible in the tracklet.

Dataset

The SoccerNet Jersey Number dataset is composed of 2853 tracklets of players extracted from the SoccerNet tracking videos. The challenge set is composed of 1211 separate players tracklets with hidden annotations.

https://github.com/SoccerNet/sn-jersey

Expected Outcome

Design a video classification framework to solve this problem, and evaluate this task using the regular accuracy metric.

Difficulty Level (1-5): 4

Project 6: Scene Graph Generation

Background

Scene graph generation is a computer vision task that aims to represent the visual scene of an image as a graph, where nodes correspond to objects and edges represent relationships or interactions between those objects. The scene graph provides a structured and semantically rich representation of the visual content of an image, allowing for higher-level reasoning and understanding of the scene. This structured representation can be used for various downstream tasks, such as image captioning, visual question answering, and scene understanding.

Description

Given an image, which is a grid of pixels, you are expected to design a model to output a structured representation of the visual scene in the form of a graph. This graph consists of nodes and edges:

Nodes: Each node in the scene graph represents an object detected in the image. Nodes contain information about the object, such as its class label (e.g., "dog," "car") and its location (bounding box or pixel coordinates).

Edges: Edges represent relationships between objects. They connect pairs of object nodes and indicate how objects interact or relate to each other. Each edge also has a label that describes the nature of the relationship (e.g., "on," "next to").

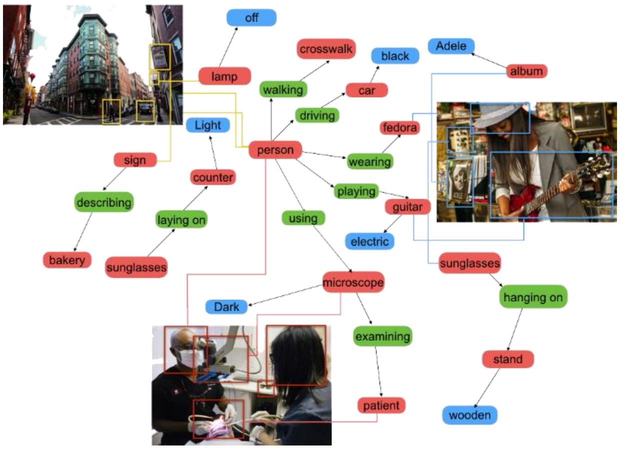


Figure Visual Genome dataset

Dataset

Visual Genome contains Visual Question Answering data in a multi-choice setting. It consists of 101,174 images from MSCOCO with 1.7 million QA pairs, 17 questions per image on average. Compared to the Visual Question Answering dataset, Visual Genome represents a more balanced distribution over 6 question types: What, Where, When, Who, Why and How. The Visual Genome dataset also presents 108K images with densely annotated objects, attributes and relationships.

[**Note**] You can only use 1/3 data of this dataset optional.

URL: https://homes.cs.washington.edu/~ranjay/visualgenome/api.html

Expected Outcome

Design a scene graph generation framework to solve this problem, and evaluate this task using the regular accuracy metric.

Difficulty Level (1-5): 4

Project 7: Sparse View 3D Reconstruction

Background

Realistic 3D object modelling, especially from limited observations, poses a significant challenge with broad applications in various vision and robotics tasks. It relies on assumptions of sparsity, where the 3D model itself is assumed to be representable by a small number of geometric primitives. To advance research in this field, researchers present the OmniObject3D dataset, a comprehensive collection of high-quality, real-scanned 3D objects with an extensive vocabulary.

Dataset

OmniObject3D covers the data of about 6,000 3D objects in more than 200 categories. At the same time, it contains a wealth of annotations, including high-precision surface mesh, point cloud, multi-view rendering images, and video captured in reality; In addition, professional scanning equipment ensures the fine shape and true texture of the object data.
Train set:

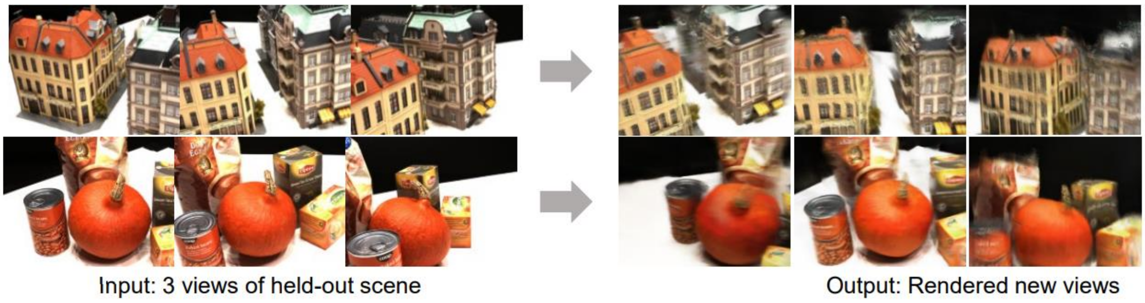
https://opendatalab.com/OpenXDLab/OmniObject3D-New/tree/main

Test set:

https://drive.google.com/file/d/1GKEa-r1__tnVKAZSF5I5uWcLh1cAllqh/view

Description

This evaluates algorithms for novel view synthesis and surface reconstruction given a few posed images of each object. The number of input images will be 1, 2, and 3, as provided in the test set. Submit the predicted novel view images and extracted point clouds in a .zip file.



Expected outcome

Design a sparse view 3D reconstruction framework to solve this problem successfully.

Difficulty Level (1-5): 4

Project 8: Referring Multi-Object Tracking with Domain Adaptation

Background

The core idea of Referring multi-object tracking (RMOT) is to employ a language expression as a semantic cue to guide the prediction of multi-object tracking. It is the work to achieve an arbitrary number of referent object predictions in videos.

Description

Referring multi-object tracking with domain adaptation needs to be based on the existing multi-object tracking model and use the language model to guide the model to achieve multi-object tracking in scenarios without ground truth.

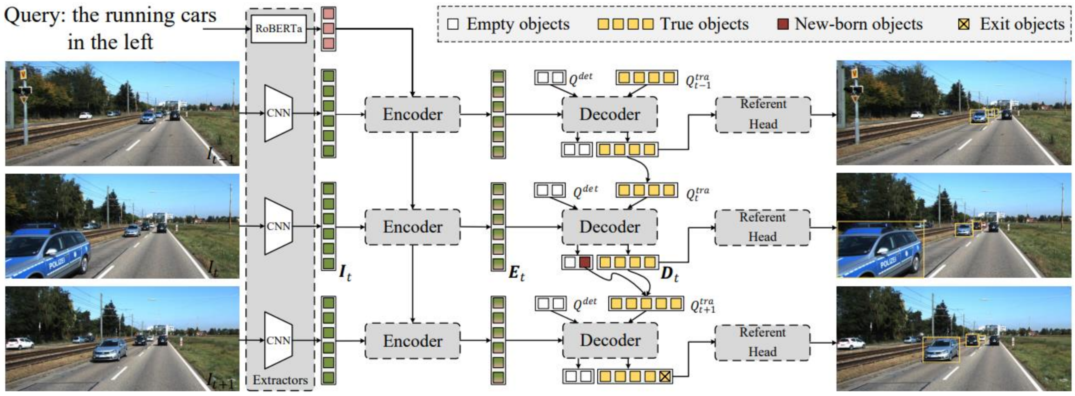


Fig. The overall architecture of TransRMOT. It is an online cross-modal tracker and includes four essential parts: feature extractors, cross-modal encoder, decoder, and referent head. The feature extractors embed the input video and the corresponding language query into feature maps/vectors. The cross-modal encoder models comprehensive visual-linguistic representation via efficient fusion. The decoder takes the visual-linguistic features, detect queries and the track queries as inputs and updates the representation of queries. The updated queries are further used to predict the referred objects by the referent head.

Expected outcome

In this project, the following tasks need to be completed step by step:

- Use the 15 videos in the refer-KITTI data set with serial numbers 01, 02, 03, 04, 06, 07, 08, 09, 10, 12, and 14 as the training set. The remaining data sets are merged with the test set (05, 11, 13) as a new test set.
- Retrain RMOT on the new training set and test on the new test set.
- Use a language model (RoBERTa, or other multi-modal models such as CLIP) to train on the **test set** and complete two multi-object tracking tasks:

    a) Give only the ground truth of the box, use the language model to design the loss function and test it on the new test set;

b) Without giving the ground truth of the box and ID, use the pre-training weights of DETR (following RMOT) to generate box proposals as pseudo labels, use the language model to design the loss function and test it on the test set.

Dependencies

- Referring MOT: https://arxiv.org/abs/2303.03366
  https://referringmot.github.io/
- DETR: https://arxiv.org/abs/2005.12872
- MOTR: https://arxiv.org/abs/2105.03247
- MOTRv2: https://arxiv.org/abs/2211.09791

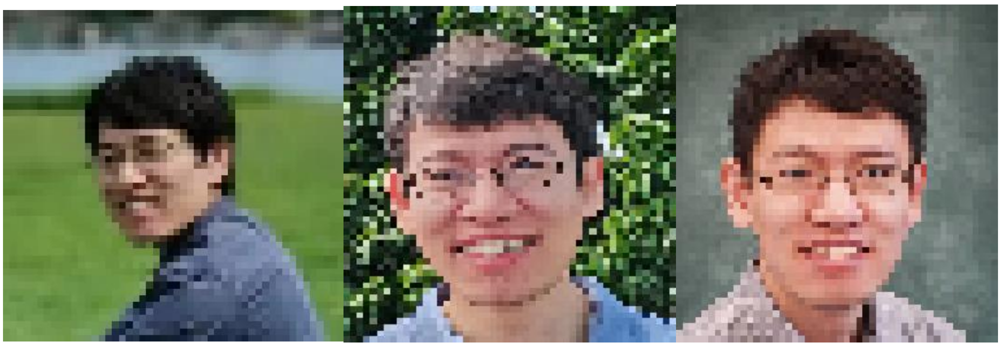
Difficulty Level (1-5): 4

Project 9: Smart Album for Multi-view Face Denoising and Super-resolution

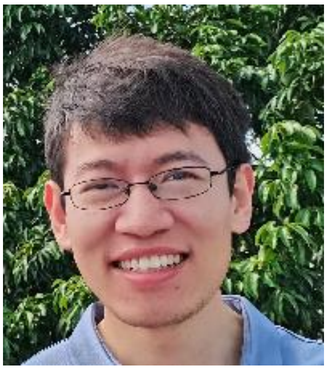
Description

Given a collection of noisy or low-resolution face photos of the same person, you are required to recover the high resolution face images. An example is shown as follows.

Input (can be many noisy and low-resolution images of the same person. May contain one high-resolution image for reference)



Output (should recover the clean one for each input image)



Data

You may use face images of yourself, your friends or download faces of celebrities from the internet. You can begin with a few thousands of images.

Expected Features and Results

1) Each output should correspond to one input image with exact the same view angle.
2) The team should give demo in the final presentation.

Challenges

1) No ground truth clean image available in the image recovery.
2) How to complete the missing information from pictures taken from other views.

Keywords for Online Search of References

Multi-view (image set-based) super-resolution

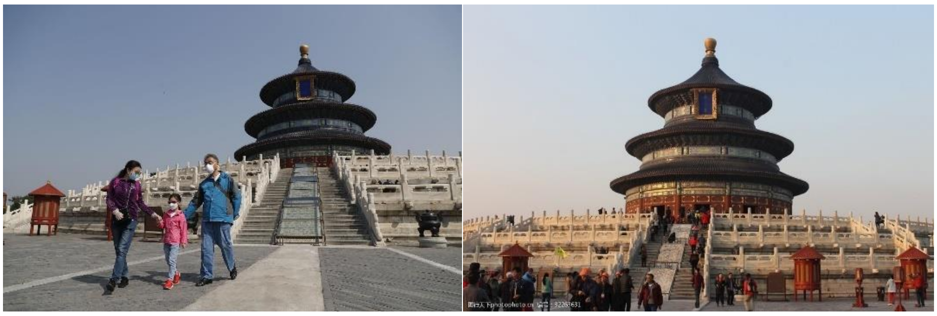
Difficulty Level (1-5): 5

Project 10: Multi-view Inpainting

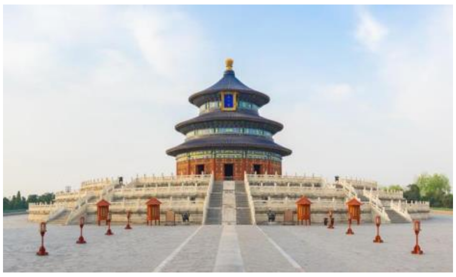
Description

Given a collection of photos captured from the same scene, you are required to recover the corresponding clean ones with all visible people removed. An example is shown as follows.

Input images (can be many images, each image should contain some foreground people)



Output (should recover the clean one for each input image)



Data

You may download similar images from the internet yourself. You can begin with a few thousands of images.

Expected Features and Results

1) The algorithm can distinguish the foreground person and the background sights automatically without manual manipulation.
2) Each output should correspond to one input image with exact the same view angle.
3) The team should give demo in the final presentation.

Challenges

1) No ground truth clean image available in the recovery.
2) How to complete the missing information from pictures taken from other views.

Keywords for Online Search of References

Video object removal, multi-view inpainting

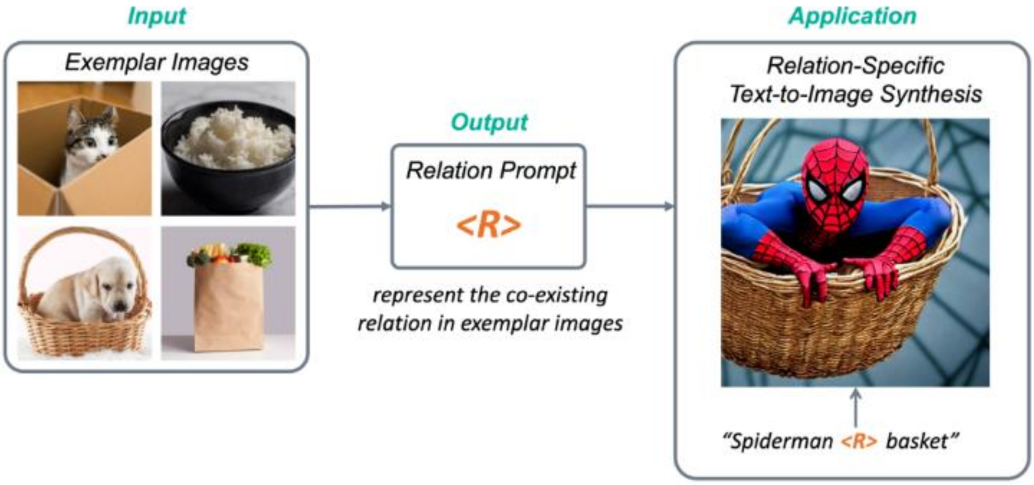
Difficulty Level (1-5): 5

Project 11: Few-shot Relation Personalization

Background

In this project, we focus on the concept of few-shot relation personalization. We are provided with several reference pictures, and these reference pictures all share a common underlying relation. The key objective is to generate new images where the objects within them interact in accordance with this shared relation.

Description


Input
Exemplar Images
Output
Relation Prompt
<R>
represent the co-existing relation in exemplar images
Application
Relation-Specific Text-to-Image Synthesis
"Spiderman <R> basket"

- We are given a set of reference images, each depicting various objects or entities in specific arrangements.
- These reference pictures are intentionally chosen to contain a common relation or interaction among the depicted objects.
- The primary challenge is to create entirely new images where objects are arranged and interact based on the shared relation observed in the reference pictures.

- The new images should effectively reflect the relation seen in the reference pictures, even though the specific objects or scenes may be different.

Expected outcome
Design a framework to solve this problem successfully.

Hint：

Few-Shot Learning, text-to-image diffusion synthesis, diffusion personalization:
This project utilizes few-shot learning techniques, meaning that we aim to train models or algorithms to understand and apply the shared relation with minimal exposure to reference pictures.
The ability to generalize from a small number of reference images to generate new compositions is a key aspect of this project.

Overall, the goal of this project is to explore and develop methods for few-shot relation personalization, allowing us to generate novel images that capture the desired relation based on a limited set of reference pictures.

Difficulty Level (1-5): 5