Object Centric Learning

- Putting the Object Back into Video Object Segmentation

- EAGLE: Eigen Aggregation Learning for Object-Centric Unsupervised Semantic Segmentation

- Guided Slot Attention for Unsupervised Video Object Segmentation

- SPOT: Self-Training with Patch-Order Permutation for Object-Centric Learning with Autoregressive Transformers

- Adaptive Slot Attention: Object Discovery with Dynamic Slot Number

-

OOD Generalization

- Rethinking Prior Information Generation with CLIP for Few-Shot Segmentation

- Unlocking the Potential of Pre-trained Vision Transformers for Few-Shot Semantic Segmentation through Relationship Descriptors

- LocLLM: Exploiting Generalizable Human Keypoint Localization via Large Language Model

- Cross-Domain Few-Shot Segmentation via Iterative Support-Query Correspondence Mining

- Exploring Regional Clues in CLIP for Zero-Shot Semantic Segmentation

- A Dual-Augmentor Framework for Domain Generalization in 3D Human Pose Estimation

- Segment Every Out-of-Distribution Object

- Emergent Open-Vocabulary Semantic Segmentation from Off-the-shelf Vision-Language Models

- Grounding Everything: Emerging Localization Properties in Vision-Language Transformers

- One-Shot Open Affordance Learning with Foundation Models

- Fast Adaptation for Human Pose Estimation via Meta-Optimization

- Open-World Semantic Segmentation Including Class Similarity

- Training-Free Open-Vocabulary Segmentation with Offline Diffusion-Augmented Prototype Generation

- Style Blind Domain Generalized Semantic Segmentation via Covariance Alignment and Semantic Consistence Contrastive Learning

- Active Domain Adaptation with False Negative Prediction for Object Detection

- Disentangled Prompt Representation for Domain Generalization

- Towards Generalizing to Unseen Domains with Few Labels

- Prompt-Driven Dynamic Object-Centric Learning for Single Domain Generalization

- Domain-Agnostic Mutual Prompting for Unsupervised Domain Adaptation

- Unbiased Faster R-CNN for Single-source Domain Generalized Object Detection

- Source-Free Domain Adaptation with Frozen Multimodal Foundation Model

- Exploring Region-Word Alignment in Built-in Detector for Open-Vocabulary Object Detection

- ID-like Prompt Learning for Few-Shot Out-of-Distribution Detection

- A Bayesian Approach to OOD Robustness in Image Classification

- Part-aware Unified Representation of Language and Skeleton for Zero-shot Action Recognition

- Improving Generalized Zero-Shot Learning by Exploring the Diverse Semantics from External Class Names

- Hyperbolic Learning with Synthetic Captions for Open-World Detection

- Leveraging Vision-Language Models for Improving Domain Generalization in Image Classification

- Split to Merge: Unifying Separated Modalities for Unsupervised Domain Adaptation

Self-Supervised Learning

- Diffuse Attend and Segment: Unsupervised Zero-Shot Segmentation using Stable Diffusion

- LAFS: Landmark-based Facial Self-supervised Learning for Face Recognition

- PSDPM: Prototype-based Secondary Discriminative Pixels Mining for Weakly Supervised Semantic Segmentation

- Self-Supervised Multi-Object Tracking with Path Consistency

Human Pose Estimation

- DiffusionRegPose: Enhancing Multi-Person Pose Estimation using a Diffusion-Based End-to-End Regression Approach

- Video-Based Human Pose Regression via Decoupled Space-Time Aggregation

- RTMO: Towards High-Performance One-Stage Real-Time Multi-Person Pose Estimation

Zero(Few)-Shot Learning

-

Disentanglement & Compositional Learning

- Compositional Chain-of-Thought Prompting for Large Multimodal Models

- Iterated Learning Improves Compositionality in Large Vision-Language Models

- Representing Part-Whole Hierarchies in Foundation Models by Learning Localizability Composability and Decomposability from Anatomy via Self Supervision

- Choose What You Need: Disentangled Representation Learning for Scene Text Recognition Removal and Editing

- Beyond Seen Primitive Concepts and Attribute-Object Compositional Learning

- Troika: Multi-Path Cross-Modal Traction for Compositional Zero-Shot Learning

- MCPNet: An Interpretable Classifier via Multi-Level Concept Prototypes

- CosalPure: Learning Concept from Group Images for Robust Co-Saliency Detection

- Understanding Video Transformers via Universal Concept Discovery

- Image-Text Co-Decomposition for Text-Supervised Semantic Segmentation

- EGTR: Extracting Graph from Transformer for Scene Graph Generation

Human-Object Interaction

- Bilateral Adaptation for Human-Object Interaction Detection with Occlusion-Robustness

- Open-World Human-Object Interaction Detection via Multi-modal Prompts

- Exploring Pose-Aware Human-Object Interaction via Hybrid Learning

-

Tracking

- Towards Generalizable Multi-Object Tracking

- DeconfuseTrack: Dealing with Confusion for Multi-Object Tracking

ETC

- Open-Vocabulary Attention Maps with Token Optimization for Semantic Segmentation in Diffusion Models

- Sparse Semi-DETR: Sparse Learnable Queries for Semi-Supervised Object Detection

- Zero-shot Referring Expression Comprehension via Structural Similarity Between Images and Captions