Paper List: Foundation Models for Computer Vision

Dominik Schnaus and Tarun Yenamandra

1 Self-Supervised Representation Learning in Computer Vision

1. Contrastive Methods

   Chen et al. A simple framework for contrastive learning of visual representations

   Zbontar et al. Barlow twins: Self-supervised learning via redundancy reduction

   He et al. Momentum contrast for unsupervised visual representation learning

2. Masked Image Methods

   He et al. Masked autoencoders are scalable vision learners

   Feichtenhofer et al. Masked autoencoders as spatiotemporal learners

   Assran et al. Self-supervised learning from images with a joint-embedding predictive architecture

3. Self-Distillation Methods

   Grill et al. Bootstrap your own latent-a new approach to self-supervised learning

   Caron et al. Emerging properties in self-supervised vision transformers

   Oquab et al. DINOv2: Learning Robust Visual Features without Supervision

4. Video Methods

   Bardes et al. V-JEPA: Latent video prediction for visual representation learning

   Venkataramanan et al. Is ImageNet worth 1 video? Learning strong image encoders from 1 long unlabelled video

2 Multi-Modal 2D Foundation Models

1. Self-Distillation Methods

   Radford et al. Learning transferable visual models from natural language supervision

   Yuan et al. Florence: A new foundation model for computer vision

2. Segmentation Models

   Kirillov et al. Segment anything

   Zou et al. Segment everything everywhere all at once

3. Image Generation Models

   Ramesh et al. Zero-shot text-to-image generation

   Nichol et al. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models

   Ramesh et al. Hierarchical text-conditional image generation with clip latents

   Rombach et al. High-resolution image synthesis with latent diffusion models

4. Video Generation Methods

   Blattmann et al. Stable Video Diffusion: Scaling Latent Video Diffusion Models to Large Datasets

   Bar-Tal et al. Lumiere: A Space-Time Diffusion Model for Video Generation

3 3D Foundation Models

1. 3D Generation from 2D Models

   Poole et al. DreamFusion: Text-to-3D using 2D Diffusion

   Chu et al. DreamScene4D: Dynamic Multi-Object Scene Generation from Monocular Videos

   Singer et al. Text-To-4D Dynamic Scene Generation

   Wang et al. ProlificDreamer: High-Fidelity and Diverse Text-to-3D Generation with Variational Score Distillation

2. Multi-view Consistent Image Generation

   Liu et al. Zero-1-to-3: Zero-shot One Image to 3D Object

   Liu et al. One-2-3-45: Any Single Image to 3D Mesh in 45 Seconds without Per-Shape Optimization

   Li et al. Instant3D: Instant Text-to-3D Generation

3. Foundation Models for Depth

   Ke et al. Repurposing Diffusion-Based Image Generators for Monocular Depth Estimation

   Yang et al. Depth Anything V2

   Bhat et al. ZoeDepth: Zero-shot Transfer by Combining Relative and Metric Depth