Highlighted Papers and YouTube Results

Paper: Towards RAW Object Detection in Diverse Conditions

Paper link: https://arxiv.org/abs/2411.15678

YouTube Result:

1. **Motivation**: The motivation behind the study is to improve the robustness of object detection algorithms when faced with adverse weather conditions, as existing models tend to fail or perform poorly in such scenarios due to being primarily trained on clear weather datasets.

2. **Novelty**: The novel aspects of the study include the use of synthetic adverse weather images generated from diverse text prompts created with the help of ChatGPT, allowing for a wide range of driving scenarios to be depicted.

3. **Main Findings**: The main findings indicate that traditional object detection models, like Faster R-CNN, struggle significantly in adverse weather conditions, as demonstrated by their inability to detect vehicles accurately in rainy settings. The study aims to address these shortcomings through fine-tuning models with the newly synthesized images.

4. **Video Title**: Time To Shine: Fine-Tuning Object Detection Models With Synthetic Adverse Weather Images

5. **Video Link**: [Watch here](https://www.youtube.com/watch?v=6ko_LsO7b24)

Paper: NTClick: Achieving Precise Interactive Segmentation With Noise-tolerant Clicks

Paper link: Not found on arXiv

YouTube Result:

I don't know.

Paper: Decoupled Distillation to Erase: A General Unlearning Method for Any Class-centric Tasks

Paper link: https://arxiv.org/abs/2503.23751

YouTube Result:

I don't know.

Paper: Towards Zero-Shot Anomaly Detection and Reasoning with Multimodal Large Language Models

Paper link: https://arxiv.org/abs/2504.13399

YouTube Result:

I don't know.

Paper: Cross-View Completion Models are Zero-shot Correspondence Estimators

Paper link: https://arxiv.org/abs/2412.09072

YouTube Result:

I don't know.

Paper: PlanarSplatting: Accurate Planar Surface Reconstruction in 3 Minutes

Paper link: https://arxiv.org/abs/2412.03451

YouTube Result:

I don't know.

Paper: Prior-free 3D Object Tracking

Paper link: https://arxiv.org/abs/2502.10606

YouTube Result:

I don't know.

Paper: Gradient-Guided Annealing for Domain Generalization

Paper link: https://arxiv.org/abs/2502.20162

YouTube Result:

I don't know.

Paper: Assessing and Learning Alignment of Unimodal Vision and Language Models

Paper link: https://arxiv.org/abs/2412.04616

YouTube Result:

I don't know.

Paper: BEVDiffuser: Plug-and-Play Diffusion Model for BEV Denoising with Ground-Truth Guidance

Paper link: https://arxiv.org/abs/2502.19694

YouTube Result:

I don't know.

Paper: HaWoR: World-Space Hand Motion Reconstruction from Egocentric Videos

Paper link: https://arxiv.org/abs/2501.02973

YouTube Result:

I don't know.

Paper: ALIEN: Implicit Neural Representations for Human Motion Prediction under Arbitrary Latency

Paper link: Not found on arXiv

YouTube Result:

I don't know.

Paper: SkillMimic: Learning Basketball Interaction Skills from Demonstrations

Paper link: https://arxiv.org/abs/2408.15270

YouTube Result:

I don't know.

Paper: Multitwine: Multi-Object Compositing with Text and Layout Control

Paper link: https://arxiv.org/abs/2502.05165

YouTube Result:

I don't know.

Paper: You See it, You Got it: Learning 3D Creation on Pose-Free Videos at Scale

Paper link: https://arxiv.org/abs/2412.06699

YouTube Result:

I don't know.

Paper: SP3D: Boosting Sparsely-Supervised 3D Object Detection via Accurate Cross-Modal Semantic Prompts

Paper link: https://arxiv.org/abs/2503.06467

YouTube Result:

I don't know.

Paper: Structured 3D Latents for Scalable and Versatile 3D Generation

Paper link: https://arxiv.org/abs/2412.01506

YouTube Result:

I don't know.

Paper: Augmented Deep Contexts for Spatially Embedded Video Coding

Paper link: Not found on arXiv

YouTube Result:

I don't know.

Paper: Learning Phase Distortion with Selective State Space Models for Video Turbulence Mitigation

Paper link: https://arxiv.org/abs/2504.02697

YouTube Result:

I don't know.

Paper: SeedVR: Seeding Infinity in Diffusion Transformer Towards Generic Video Restoration

Paper link: https://arxiv.org/abs/2501.01320

YouTube Result:

I don't know.

Paper: COUNTS: Benchmarking Object Detectors and Multimodal Large Language Models under Distribution Shifts

Paper link: https://arxiv.org/abs/2504.10158

YouTube Result:

I don't know.

Paper: Memories of Forgotten Concepts

Paper link: https://arxiv.org/abs/2412.01207

YouTube Result:

I don't know.

Paper: Revisiting MAE Pre-training for 3D Medical Image Segmentation

Paper link: https://arxiv.org/abs/2410.23132

YouTube Result:

I don't know.

Paper: CH3Depth: Efficient and Flexible Depth Foundation Model with Flow Matching

Paper link: Not found on arXiv

YouTube Result:

I don't know.

Paper: Lessons and Insights from a Unifying Study of Parameter-Efficient Fine-Tuning (PEFT) in Visual Recognition

Paper link: https://arxiv.org/abs/2409.16434

YouTube Result:

I don't know.

Paper: Hyperbolic Safety-Aware Vision-Language Models

Paper link: https://arxiv.org/abs/2503.12127

YouTube Result:

I don't know.

Paper: v-CLR: View-Consistent Learning for Open-World Instance Segmentation

Paper link: https://arxiv.org/abs/2504.01383

YouTube Result:

I don't know.

Paper: Satellite Observations Guided Diffusion Model for Accurate Meteorological States at Arbitrary Resolution

Paper link: https://arxiv.org/abs/2502.07814

YouTube Result:

I don't know.

Paper: ESC: Erasing Space Concept for Knowledge Deletion

Paper link: https://arxiv.org/abs/2504.02199

YouTube Result:

I don't know.

Paper: Goku: Flow Based Video Generative Foundation Models

Paper link: https://arxiv.org/abs/2502.04896

YouTube Result:

1. **Motivation**: The motivation behind the study is to advance video generation technology by developing a new family of models that can achieve state-of-the-art performance in both image and video generation, leveraging open-source projects to demonstrate the potential of collaborative efforts outside large tech companies.

2. **Novelty**: The novel aspects of the study include the use of rectified flow Transformers, which provide a smoother and more direct approach to learning how to transform noise into realistic images and videos. This method allows for faster training and sharper results, as well as a better understanding of motion in videos through a shared latent space.

3. **Main Findings**: The main findings of the study show that Goku achieves exceptional performance metrics, scoring 76 and 83.65 on text-to-image generation benchmarks, and an impressive 84.85% on text-to-video generation. These outcomes are attributed to the large datasets used for training (36 million video-text pairs and 160 million image-text pairs) and the innovative model architecture that enhances generative capabilities.

4. **Video Title**: Goku: Flow Based Video Generative Foundation Models (Feb 2025)

5. **Video Link**: [Watch here](https://www.youtube.com/watch?v=coh1ya9Rx4A)

Paper: X-Dyna: Expressive Dynamic Human Image Animation

Paper link: https://arxiv.org/abs/2501.10021

YouTube Result:

I don't know.

Paper: MangaNinja: Line Art Colorization with Precise Reference Following

Paper link: https://arxiv.org/abs/2501.08332

YouTube Result:

1. **Motivation**: The motivation behind the study is to revolutionize the art of line art colorization, enabling artists, particularly in the anime industry, to transform intricate black and white line art into beautifully colored masterpieces more efficiently.

2. **Novelty**: The novel aspects of the study include the use of a dual Branch structure that incorporates diffusion models, patch shuffling, and an interactive point guard guidance system. This approach allows for precise color matching to specific reference images, which is a significant advancement over previous automated methods.

3. **Main Findings**: The study found that MangaNinja performs exceptionally well in colorizing line art with great accuracy, even in complex scenarios involving mismatched references and multi-character compositions. It allows users to fine-tune specific areas of colorization, providing a high degree of control and customization.

4. **Video Title**: MangaNinja: Line Art Colorization with Precise Reference Following (Paper Walkthrough)

5. **Video Link**: [Watch the video](https://www.youtube.com/watch?v=63ZK40J5nD0)

Paper: PartGen: Part-level 3D Generation and Reconstruction with Multi-view Diffusion Models

Paper link: https://arxiv.org/abs/2412.18608

YouTube Result:

1. **Motivation**: The motivation behind the study is to create a tool that allows users to generate 3D models in a modular way, enabling the creation of objects composed of meaningful, editable parts from various input sources like text, images, or existing 3D models.

2. **Novelty**: The novel aspect of the study is the use of a diffusion-based network to segment and generate 3D objects into consistent parts, which can then be assembled into a complete model. This approach allows for a high level of customization and flexibility in 3D model creation.

3. **Main Findings**: The main findings indicate that PartGen facilitates the process of creating 3D models by enabling functions such as text-to-3D and image-to-3D conversions, as well as real-world object decomposition and part editing.

4. **Video Title**: PartGen: Part-level 3D Generation and Reconstruction with Multi-View Diffusion Models

5. **Video Link**: [Watch here](https://www.youtube.com/watch?v=e9ABYNKA7tc)

Paper: DepthCrafter: Generating Consistent Long Depth Sequences for Open-world Videos

Paper link: https://arxiv.org/abs/2409.02095

YouTube Result:

1. **Motivation**: The motivation behind the study is to improve depth estimation in videos, which

involves understanding the spatial layout of scenes by estimating the distance of objects from the camera in order to create accurate 3D representations.

2. **Novelty**: The novelty of the DepthCrafter project lies in its ability to generate depth sequences specifically for open-world videos. It utilizes a specialized training strategy with paired video depth datasets to enhance the accuracy of depth estimation.

3. **Main Findings**: The main findings suggest that DepthCrafter effectively estimates depth in videos by addressing the challenges posed by varying content, motion, and length in video data. It produces depth maps that assign distance values to pixels, enabling better understanding of foreground and background objects.

4. **Video Title**: DepthCrafter - Install Locally - Generate Depth Sequences for Open-world Videos

5. **Video Link**: [Watch here](https://www.youtube.com/watch?v=anZGKW4nFe4)

Paper: Reference-Based 3D-Aware Image Editing with Triplanes

Paper link: https://arxiv.org/abs/2404.03632

YouTube Result:

I don't know.

Paper: WonderWorld: Interactive 3D Scene Generation from a Single Image

Paper link: https://arxiv.org/abs/2406.09394

YouTube Result:

I don't know.

Paper: MAtCha Gaussians: Atlas of Charts for High-Quality Geometry and Photorealism From Sparse Views

Paper link: https://arxiv.org/abs/2412.06767

YouTube Result:

I don't know.

Paper: Taming Video Diffusion Prior with Scene-Grounding Guidance for 3D Gaussian Splatting from Sparse Inputs

Paper link: https://arxiv.org/abs/2503.05082

YouTube Result:

I don't know.

Paper: MoSca: Dynamic Gaussian Fusion from Casual Videos via 4D Motion Scaffolds

Paper link: https://arxiv.org/abs/2405.17421

YouTube Result:

1. **Motivation**: The study aims to address the complex challenge of reconstructing a 4D scene from unposed, in-the-wild monocular RGB videos, which is an ill-posed inverse problem.

2. **Novelty**: The novel aspect of the study is the introduction of the MoSca representation, which is a sparse graph of SE(3) motion trajectories. This representation effectively encodes the underlying motion and allows for interpolation into a dense SE(3) deformation field.

3. **Main Findings**: The study demonstrates the ability to reconstruct and render dynamic 4D scenes that can be viewed from novel viewpoints. It showcases visual results from various casual videos and provides comparisons on benchmarks from DICE and Nvidia.

4. **Video Title**: MoSca-Version2: Dynamic Gaussian Fusion from Casual Videos via 4D Motion Scaffolds

5. **Video Link**: [Watch the video here](https://www.youtube.com/watch?v=7WrG5-xH1_k)

Paper: TFCustom: Customized Image Generation with Time-Aware Frequency Feature Guidance

Paper link: Not found on arXiv

YouTube Result:

I don't know.

Paper: DualPM: Dual Posed-Canonical Point Maps for 3D Shape and Pose Reconstruction

Paper link: https://arxiv.org/abs/2412.04464

YouTube Result:

I don't know.

Paper: EBS-EKF: Accurate and High Frequency Event-based Star Tracking

Paper link: https://arxiv.org/abs/2503.20101

YouTube Result:

I don't know.

Paper: RoboPEPP: Vision-Based Robot Pose and Joint Angle Estimation through Embedding Predictive Pre-Training

Paper link: https://arxiv.org/abs/2411.17662

YouTube Result:

1. **Motivation**: The study aims to improve the accuracy of robot pose and joint angle estimation from a single image, which is important for applications like human-robot interaction and multi-robot collaboration. Existing methods struggle to utilize the rich information from the robot's physical structure, particularly in challenging scenarios such as occlusions and truncations.

2. **Novelty**: The novel aspect of the study is the introduction of the RoboPEPP framework, which incorporates embedding predictive pre-training into a network that fuses the robot's physical model with image data. This approach enhances the model's ability to estimate joint angles and poses by better understanding the robot's structure.

3. **Main Findings**: The study found that by using masking-based embedding predictive pre-training, the network could effectively infer joint information from surrounding context, leading to improved robustness against occlusions and partial visibility issues. The method outperforms existing techniques by better utilizing the physical model of the robot.

4. **Video Title**: [CVPR 2025] RoboPEPP

5. **Video Link**: https://www.youtube.com/watch?v=pbM60-kHSdE

Paper: MonSter: Marry Monodepth to Stereo Unleashes Power

Paper link: https://arxiv.org/abs/2501.08643

YouTube Result:

1. **Motivation**: The study aims to improve depth estimation from images, specifically addressing

the challenges faced by existing stereo matching methods, which struggle in complex scenarios such as occlusions and textureless surfaces.

2. **Novelty**: The novel aspect of this study is the introduction of a dual branch architecture that combines monocular depth estimation with stereo matching, allowing the two approaches to iteratively enhance each other's performance.

3. **Main Findings**: The results of the Monster model show strong performance, ranking first on several major benchmarks. It exhibits zero-shot generalization, meaning it effectively produces accurate depth maps even for unseen data, as evidenced by its performance on the KITTI dataset compared to baseline methods.

4. **Video Title**: MonSter: Better Depth with Mono & Stereo Vision

5. **Video Link**: [Watch here](https://www.youtube.com/watch?v=65ZSR9AY8EM)

Paper: InteractAnything: Zero-shot Human Object Interaction Synthesis via LLM Feedback and Object Affordance Parsing

Paper link: Not found on arXiv

YouTube Result:

I don't know.

Paper: Image Quality Assessment: From Human to Machine Preference

Paper link: https://arxiv.org/abs/2503.10078

YouTube Result:

1. **Motivation**: The motivation behind the study is to improve image quality assessment methods

for various applications, particularly in fields like image and video processing, where traditional metrics such as mean squared error can be misleading.

2. **Novelty**: The novelty of the study lies in the exploration of advanced quality assessment techniques that do not rely solely on traditional reference images, instead allowing for the automated evaluation of image quality without a gold standard.

3. **Main Findings**: The findings indicate that while mean squared error can provide a quantifiable measure of image quality, it does not necessarily correlate with perceived quality. The study emphasizes the need for more robust metrics that align better with human perception of image quality.

4. **Video Title**: Objective image quality assessment, what's beyond - Zhou Wang

5. **Video Link**: [Watch here](https://www.youtube.com/watch?v=ibiCs_NJgCQ)

Paper: Multirate Neural Image Compression with Adaptive Lattice Vector Quantization

Paper link: Not found on arXiv

YouTube Result:

I don't know.

Paper: OpenHumanVid: A Large-Scale High-Quality Dataset for Enhancing Human-Centric Video Generation

Paper link: https://arxiv.org/abs/2412.00115

YouTube Result:

1. **Motivation**: The motivation behind the study is to improve AI-generated human movement in video generation, making it appear more natural and realistic by providing a dataset that reflects actual human behavior and movement.

2. **Novelty**: The novel aspect of the study lies in the creation of a large-scale dataset, OpenHumanVid, which specifically focuses on high-quality clips of people performing everyday activities. This approach helps to train AI models to produce more lifelike human motions, addressing the common issue of unrealistic animations.

3. **Main Findings**: The main findings indicate that the dataset helps AI learn realistic pacing and posture by filtering out low-quality videos and retaining only the best examples of human movements. However, the researchers noted that the dataset is still predominantly composed of indoor clips and that privacy regulations limited the range of content included.

4. **Video Title**: OpenHumanVid: A Large-Scale High-Quality Dataset for Enhancing Human-Centric Video Generation

5. **Video Link**: [Watch the video](https://www.youtube.com/watch?v=ZzQBvpKYZ88)

Paper: Generative Photography: Scene-Consistent Camera Control for Realistic Text-to-Image Synthesis

Paper link: https://arxiv.org/abs/2412.02168

YouTube Result:

I don't know.

Paper: SoundVista: Novel-View Ambient Sound Synthesis via Visual-Acoustic

Binding

Paper link: https://arxiv.org/abs/2504.05576

YouTube Result:

I don't know.

Paper: MambaVLT: Time-Evolving Multimodal State Space Model for Vision-Language Tracking

Paper link: https://arxiv.org/abs/2411.15459

YouTube Result:

I don't know.

Paper: Theoretical Insights in Model Inversion Robustness and Conditional Entropy Maximization for Collaborative Inference Systems

Paper link: https://arxiv.org/abs/2503.00383

YouTube Result:

I don't know.

Paper: Modeling Thousands of Human Annotators for Generalizable Text-to-Image Person Re-identification

Paper link: https://arxiv.org/abs/2503.09962

YouTube Result:

I don't know.

Paper: Graph Neural Network Combining Event Stream and Periodic Aggregation for Low-Latency Event-based Vision

Paper link: Not found on arXiv

YouTube Result:

I don't know.

Paper: Enhanced Visual-Semantic Interaction with Tailored Prompts for Pedestrian Attribute Recognition

Paper link: Not found on arXiv

YouTube Result:

I don't know.

Paper: Timestep Embedding Tells: It's Time to Cache for Video Diffusion Model

Paper link: https://arxiv.org/abs/2411.19108

YouTube Result:

I don't know.

Paper: Scaling Vision Pre-Training to 4K Resolution

Paper link: https://arxiv.org/abs/2503.19903

YouTube Result:

1. **Motivation**: The motivation behind the study is to enhance visual perception capabilities by leveraging advancements in technology to capture and process higher resolution images, specifically 4K resolution, which is increasingly vital for everyday tasks like navigation and interaction with complex software interfaces.

2. **Novelty**: The novel aspect of the study is the introduction of a method called scale selective processing (PS3), which enables pre-training of vision models at 4K resolution while maintaining

computational efficiency, a feat that has been challenging for researchers previously.

3. **Main Findings**: The study found that using PS3 allows for effective pre-training of vision models at a stunning 4K resolution, significantly improving visual detail without a proportional increase in computational resources.

4. **Video Title**: Scaling Vision Pre-Training to 4K Resolution

5. **Video Link**: [Scaling Vision Pre-Training to 4K Resolution](https://www.youtube.com/watch?v=hGyIc_wG2yw)

Paper: Multimodal Autoregressive Pre-training of Large Vision Encoders

Paper link: https://arxiv.org/abs/2411.14402

YouTube Result:

1. **Motivation**: The study aims to improve AI's ability to understand both images and text simultaneously, addressing the limitations of traditional models that excel in either domain but not both.

2. **Novelty**: The novelty lies in the introduction of multimodal autoregressive pre-training, which allows the model to predict what comes next in both images and text, resembling human-like learning patterns.

3. **Main Findings**: The researchers found that this approach outperforms existing models in tasks such as object identification and understanding relationships between objects in images, demonstrating a significant advancement in visual AI capabilities.

4. **Video Title**: Multimodal Autoregressive Pre-training of Large Vision Encoders

5. **Video Link**: [Watch Here](https://www.youtube.com/watch?v=qzpz_DKIWjM)

Paper: Learning Class Prototypes for Unified Sparse-Supervised 3D Object Detection

Paper link: https://arxiv.org/abs/2503.21099

YouTube Result:

I don't know.

Paper: Open-Canopy: Towards Very High Resolution Forest Monitoring

Paper link: https://arxiv.org/abs/2407.09392

YouTube Result:

I don't know.

Paper: HOT3D: Hand and Object Tracking in 3D from Egocentric Multi-View Videos

Paper link: https://arxiv.org/abs/2411.19167

YouTube Result:

I don't know.

Paper: Doppelgängers and Adversarial Vulnerability

Paper link: https://arxiv.org/abs/2410.13193

YouTube Result:

I don't know.

Paper: Annotation Ambiguity Aware Semi-Supervised Medical Image Segmentation

Paper link: Not found on arXiv

YouTube Result:

I don't know.

Paper: EffiDec3D: An Optimized Decoder for High-Performance and Efficient 3D Medical Image Segmentation

Paper link: Not found on arXiv

YouTube Result:

I don't know.

Paper: HotSpot: Signed Distance Function Optimization with an Asymptotically Sufficient Condition

Paper link: https://arxiv.org/abs/2411.14628

YouTube Result:

I don't know.

Paper: Hardware-Rasterized Ray-Based Gaussian Splatting

Paper link: https://arxiv.org/abs/2503.18682

YouTube Result:

I don't know.

Paper: OpticalNet: An Optical Imaging Dataset and Benchmark Beyond the Diffraction Limit

Paper link: Not found on arXiv

YouTube Result:

I don't know.

Paper: Scaling Inference Time Compute for Diffusion Models

Paper link: https://arxiv.org/abs/2505.01823

YouTube Result:

I don't know.

Paper: MAR-3D: Progressive Masked Auto-regressor for High-Resolution 3D Generation

Paper link: https://arxiv.org/abs/2503.20519

YouTube Result:

I don't know.

Paper: DashGaussian: Optimizing 3D Gaussian Splatting in 200 Seconds

Paper link: https://arxiv.org/abs/2503.18402

YouTube Result:

I don't know.

Paper: Matrix3D: Large Photogrammetry Model All-in-One

Paper link: https://arxiv.org/abs/2502.07685

YouTube Result:

I don't know.

Paper: DroneSplat: 3D Gaussian Splatting for Robust 3D Reconstruction from In-the-Wild Drone Imagery

Paper link: https://arxiv.org/abs/2503.16964

YouTube Result:

I don't know.

Paper: NeRFPrior: Learning Neural Radiance Field as a Prior for Indoor Scene Reconstruction

Paper link: https://arxiv.org/abs/2503.18361

YouTube Result:

I don't know.

Paper: QuCOOP: A Versatile Framework for Solving Composite and Binary-Parametrised Problems on Quantum Annealers

Paper link: https://arxiv.org/abs/2503.19718

YouTube Result:

I don't know.

Paper: Image Reconstruction from Readout-Multiplexed Single-Photon Detector Arrays

Paper link: https://arxiv.org/abs/2312.02971

YouTube Result:

I don't know.

Paper: Learning to Filter Outlier Edges in Global SfM

Paper link: Not found on arXiv

YouTube Result:

I don't know.

Paper: Structure-Aware Correspondence Learning for Relative Pose Estimation

Paper link: https://arxiv.org/abs/2503.18671

YouTube Result:

I don't know.

Paper: CRISP: Object Pose and Shape Estimation with Test-Time Adaptation

Paper link: https://arxiv.org/abs/2412.01052

YouTube Result:

I don't know.

Paper: CAP-Net: A Unified Network for 6D Pose and Size Estimation of Categorical Articulated Parts from a Single RGB-D Image

Paper link: https://arxiv.org/abs/2504.11230

YouTube Result:

I don't know.

Paper: Tuning the Frequencies: Robust Training for Sinusoidal Neural Networks

Paper link: https://arxiv.org/abs/2407.21121

YouTube Result:

I don't know.

Paper: Detection-Friendly Nonuniformity Correction: A Union Framework for Infrared UAV Target Detection

Paper link: https://arxiv.org/abs/2504.04012

YouTube Result:

I don't know.

Paper: SimLingo: Vision-Only Closed-Loop Autonomous Driving with Language-Action Alignment

Paper link: https://arxiv.org/abs/2503.09594

YouTube Result:

I don't know.

Paper: HSI-GPT: A General-Purpose Large Scene-Motion-Language Model for Human Scene Interaction

Paper link: Not found on arXiv

YouTube Result:

I don't know.

Paper: DiffusionDrive: Truncated Diffusion Model for End-to-End Autonomous Driving

Paper link: https://arxiv.org/abs/2411.15139

YouTube Result:

I don't know.

Paper: MPDrive: Improving Spatial Understanding with Marker-Based Prompt

Learning for Autonomous Driving

Paper link: https://arxiv.org/abs/2504.00379

YouTube Result:

I don't know.

Paper: Reasoning in Visual Navigation of End-to-end Trained Agents: A Dynamical Systems Approach

Paper link: https://arxiv.org/abs/2503.08306

YouTube Result:

I don't know.

Paper: Seurat: From Moving Points to Depth

Paper link: https://arxiv.org/abs/2504.14687

YouTube Result:

I don't know.

Paper: ManiVideo: Generating Hand-Object Manipulation Video with Dexterous and Generalizable Grasping

Paper link: https://arxiv.org/abs/2412.16212

YouTube Result:

I don't know.

Paper: InterMimic: Towards Universal Whole-Body Control for Physics-Based Human-Object Interactions

Paper link: https://arxiv.org/abs/2502.20390

YouTube Result:

1. **Motivation**: The study aims to develop a universal whole-body control system that enables physics-based human-object interactions, addressing the challenges of dynamic interactions and contact-rich scenarios.

2. **Novelty**: The novel aspect of the study is the integration of retargeting directly into the imitation process, allowing a single policy to manage diverse whole-body interaction skills without relying on external retargeting mechanisms.

3. **Main Findings**: The findings indicate that the InterMimic framework is capable of synthesizing smooth and natural interactions, effectively learning from motion capture data while correcting contact penetration artifacts. It demonstrates an ability to handle long-term interactions with various objects while maintaining reliable contact.

4. **Video Title**: InterMimic: Towards Universal Whole-Body Control for Physics-Based Human-Object Interactions

5. **Video Link**: [Watch here](https://www.youtube.com/watch?v=ZJT387dvI9w)

Paper: ClimbingCap: Multi-Modal Dataset and Method for Rock Climbing in World Coordinate

Paper link: https://arxiv.org/abs/2503.21268

YouTube Result:

I don't know.

Paper: FineVQ: Fine-Grained User Generated Content Video Quality

Assessment

Paper link: https://arxiv.org/abs/2412.19238

YouTube Result:

I don't know.

Paper: Volumetrically Consistent 3D Gaussian Rasterization

Paper link: https://arxiv.org/abs/2502.08297

YouTube Result:

I don't know.

Paper: UniReal: Universal Image Generation and Editing via Learning Real-world Dynamics

Paper link: https://arxiv.org/abs/2412.07774

YouTube Result:

1. **Motivation**: The motivation behind the study is to advance the field of image generation and editing by leveraging real-world dynamics, aiming to create a universal system that can generate and manipulate images effectively and realistically.

2. **Novelty**: The novel aspect of the study is its approach to integrating real-world dynamics into the image generation process, which enhances the realism and applicability of the generated images in various contexts.

3. **Main Findings**: The main findings of the study indicate that the proposed method successfully generates high-quality images and enables effective editing capabilities by understanding and applying the dynamics of real-world scenarios.

4. **Video Title**: UniReal: Universal Image Generation and Editing via Learning Real-world Dynamics

5. **Video Link**: [Watch here](https://www.youtube.com/watch?v=zwOedmmEGv4)

Paper: Extrapolating and Decoupling Image-to-Video Generation Models: Motion Modeling is Easier Than You Think

Paper link: https://arxiv.org/abs/2503.00948

YouTube Result:

I don't know.

Paper: SKDream: Controllable Multi-view and 3D Generation with Arbitrary Skeletons

Paper link: Not found on arXiv

YouTube Result:

I don't know.

Paper: High-Fidelity Lightweight Mesh Reconstruction from Point Clouds

Paper link: Not found on arXiv

YouTube Result:

I don't know.

Paper: Parallelized Autoregressive Visual Generation

Paper link: https://arxiv.org/abs/2504.08388

YouTube Result:

1. **Motivation**: The study aims to significantly speed up the process of AI generating images and videos, addressing the long wait times users typically experience when rendering visual content.

2. **Novelty**: The research introduces a technique called parallelized autoregressive visual generation, which allows for the prediction of visual elements in a more efficient manner, akin to predictive text but applied to images.

3. **Main Findings**: The key finding is that not all visual elements depend on each other equally, which enables the parallelization of the generation process, drastically reducing the time it takes to create images without sacrificing quality.

4. **Video Title**: Parallelized Autoregressive Visual Generation

5. **Video Link**: [Parallelized Autoregressive Visual Generation](https://www.youtube.com/watch?v=iGmhCoApmHE)

Paper: Driving by the Rules: A Benchmark for Integrating Traffic Sign Regulations into Vectorized HD Map

Paper link: https://arxiv.org/abs/2410.23780

YouTube Result:

I don't know.

Paper: TKG-DM: Training-free Chroma Key Content Generation Diffusion Model

Paper link: https://arxiv.org/abs/2411.15580

YouTube Result:

I don't know.

Paper: SCSA: A Plug-and-Play Semantic Continuous-Sparse Attention for Arbitrary Semantic Style Transfer

Paper link: https://arxiv.org/abs/2503.04119

YouTube Result:

I don't know.

Paper: Towards Explainable and Unprecedented Accuracy in Matching Challenging Finger Crease Patterns

Paper link: Not found on arXiv

YouTube Result:

I don't know.

Paper: From Words to Structured Visuals: A Benchmark and Framework for Text-to-Diagram Generation and Editing

Paper link: https://arxiv.org/abs/2411.11916

YouTube Result:

I don't know.

Paper: Digital Twin Catalog: A Large-Scale Photorealistic 3D Object Digital Twin Dataset

Paper link: https://arxiv.org/abs/2504.08541

YouTube Result:

I don't know.

Paper: FirePlace: Geometric Refinements of LLM Common Sense Reasoning for

3D Object Placement

Paper link: https://arxiv.org/abs/2503.04919

YouTube Result:

I don't know.

Paper: Seeing More with Less: Human-like Representations in Vision Models

Paper link: https://arxiv.org/abs/1812.02378

YouTube Result:

I don't know.

Paper: DistinctAD: Distinctive Audio Description Generation in Contexts

Paper link: https://arxiv.org/abs/2411.18180

YouTube Result:

I don't know.

Paper: Deep Fair Multi-View Clustering with Attention KAN

Paper link: Not found on arXiv

YouTube Result:

I don't know.

Paper: Unsupervised Continual Domain Shift Learning with Multi-Prototype Modeling

Paper link: Not found on arXiv

YouTube Result:

I don't know.

Paper: VEU-Bench: Towards Comprehensive Understanding of Video Editing

Paper link: https://arxiv.org/abs/2504.17828

YouTube Result:

I don't know.

Paper: Question-Aware Gaussian Experts for Audio-Visual Question Answering

Paper link: https://arxiv.org/abs/2503.04459

YouTube Result:

I don't know.

Paper: Instruction-based Image Manipulation by Watching How Things Move

Paper link: https://arxiv.org/abs/2412.12087

YouTube Result:

I don't know.

Paper: SnapGen: Taming High-Resolution Text-to-Image Models for Mobile Devices with Efficient Architectures and Training

Paper link: https://arxiv.org/abs/2412.09619

YouTube Result:

I don't know.

Paper: MASH-VLM: Mitigating Action-Scene Hallucination in Video-LLMs through Disentangled Spatial-Temporal Representations

Paper link: https://arxiv.org/abs/2503.15871

YouTube Result:

I don't know.

Paper: UMotion: Uncertainty-driven Human Motion Estimation from Inertial and Ultra-wideband Units

Paper link: Not found on arXiv

YouTube Result:

I don't know.

Paper: MammAlps: A Multi-view Video Behavior Monitoring Dataset of Wild Mammals in the Swiss Alps

Paper link: https://arxiv.org/abs/2503.18223

YouTube Result:

I don't know.

Paper: GroundingFace: Fine-grained Face Understanding via Pixel Grounding Multimodal Large Language Model

Paper link: Not found on arXiv

YouTube Result:

I don't know.

Paper: Exact: Exploring Space-Time Perceptive Clues for Weakly Supervised Satellite Image Time Series Semantic Segmentation

Paper link: https://arxiv.org/abs/2412.03968

YouTube Result:

I don't know.

Paper: Dr. Splat: Directly Referring 3D Gaussian Splatting via Direct Language Embedding Registration

Paper link: https://arxiv.org/abs/2502.16652

YouTube Result:

I don't know.

Paper: Style Evolving along Chain-of-Thought for Unknown-Domain Object Detection

Paper link: https://arxiv.org/abs/2503.09968

YouTube Result:

I don't know.

Paper: Olympus: A Universal Task Router for Computer Vision Tasks

Paper link: https://arxiv.org/abs/2412.09612

YouTube Result:

1. **Motivation**: The study aims to enhance computer vision capabilities by creating a universal task router that can efficiently manage various vision-related tasks using multimodal large language models (MLMs).

2. **Novelty**: The novel aspect of the study is the introduction of Olympus, which acts as a universal remote for vision tasks, capable of routing tasks to specialized tools based on natural language instructions.

3. **Main Findings**: Olympus can handle over 20 different tasks related to images, videos, and 3D objects. It is designed to perform tasks such as image generation, video editing, and 3D model conversion, demonstrating significant versatility in executing complex chains of actions.

4. **Video Title**: Olympus: Smarter Computer Vision with Task Routing

5. **Video Link**: [Olympus: Smarter Computer Vision with Task Routing](https://www.youtube.com/watch?v=Lc4iDiG-O3M)

Paper: Filter Images First, Generate Instructions Later: Pre-Instruction Data Selection for Visual Instruction Tuning

Paper link: https://arxiv.org/abs/2503.07591

YouTube Result:

1. **Motivation**: The study aims to address the resource-intensive nature of visual instruction tuning in large vision language models (LVLMs), which often require massive datasets of image-instruction pairs that can be expensive and time-consuming to create.

2. **Novelty**: The novel aspect of the study is the approach of selecting the best unlabeled images first, before generating instructions for them, thus flipping the traditional data acquisition process on its head.

3. **Main Findings**: The findings suggest that by being more selective in the image data used for instruction generation, the process can become more efficient and accessible, reducing barriers for researchers and developers working with these technologies.

4. **Video Title**: Filter Images First: Efficient Visual Instruction Tuning!

5. **Video Link**: [Filter Images First: Efficient Visual Instruction Tuning!](https://www.youtube.com/watch?v=8EePj6CbayY)

Paper: Holmes-VAU: Towards Long-term Video Anomaly Understanding at Any Granularity

Paper link: https://arxiv.org/abs/2412.06171

YouTube Result:

I don't know.

Paper: Can Machines Understand Composition? Dataset and Benchmark for Photographic Image Composition Embedding and Understanding

Paper link: Not found on arXiv

YouTube Result:

I don't know.

Paper: MLLM-as-a-Judge for Image Safety without Human Labeling

Paper link: https://arxiv.org/abs/2501.00192

YouTube Result:

I don't know.

Paper: LLMDet: Learning Strong Open-Vocabulary Object Detectors under the Supervision of Large Language Models

Paper link: https://arxiv.org/abs/2501.18954

YouTube Result:

I don't know.

Paper: EventPSR: Surface Normal and Reflectance Estimation from Photometric Stereo Using an Event Camera

Paper link: Not found on arXiv

YouTube Result:

I don't know.

Paper: OPTICAL: Leveraging Optimal Transport for Contribution Allocation in Dataset Distillation

Paper link: Not found on arXiv

YouTube Result:

I don't know.

Paper: Less is More: Efficient Model Merging with Binary Task Switch

Paper link: https://arxiv.org/abs/2412.00054

YouTube Result:

I don't know.

Paper: KAC: Kolmogorov-Arnold Classifier for Continual Learning

Paper link: https://arxiv.org/abs/2503.21076

YouTube Result:

I don't know.

Paper: Label Shift Meets Online Learning: Ensuring Consistent Adaptation with Universal Dynamic Regret

Paper link: Not found on arXiv

YouTube Result:

I don't know.

Paper: Overcoming Shortcut Problem in VLM for Robust Out-of-Distribution Detection

Paper link: Not found on arXiv

YouTube Result:

I don't know.

Paper: CCIN: Compositional Conflict Identification and Neutralization for Composed Image Retrieval

Paper link: Not found on arXiv

YouTube Result:

I don't know.

Paper: Towards Improved Text-Aligned Codebook Learning: Multi-Hierarchical Codebook-Text Alignment with Long Text

Paper link: https://arxiv.org/abs/2503.01261

YouTube Result:

I don't know.

Paper: Learning Conditional Space-Time Prompt Distributions for Video Class-Incremental Learning

Paper link: Not found on arXiv

YouTube Result:

I don't know.

Paper: Task-driven Image Fusion with Learnable Fusion Loss

Paper link: https://arxiv.org/abs/2412.03240

YouTube Result:

I don't know.

Paper: EmoDubber: Towards High Quality and Emotion Controllable Movie Dubbing

Paper link: https://arxiv.org/abs/2412.08988

YouTube Result:

I don't know.

Paper: From Faces to Voices: Learning Hierarchical Representations for High-quality Video-to-Speech

Paper link: https://arxiv.org/abs/2503.16956

YouTube Result:

I don't know.

Paper: Diffusion-based Realistic Listening Head Generation via Hybrid Motion Modeling

Paper link: Not found on arXiv

YouTube Result:

1. **Motivation**: The study aims to explore innovative methods for generating realistic listening heads using diffusion-based techniques, which can enhance the representation of human figures in digital environments.

2. **Novelty**: The novel aspect of the study lies in its hybrid motion modeling approach that

combines diffusion processes with traditional methods to create more lifelike representations of listening heads.

3. **Main Findings**: The findings suggest that the proposed method significantly improves the realism and responsiveness of listening head animations compared to existing techniques.

4. **Video Title**: Diffusion-based Realistic Listening Head Generation via Hybrid Motion Modeling Supplementary Video

5. **Video Link**: [Watch here](https://www.youtube.com/watch?v=v0pq9YqmAjw)

Paper: FreeCloth: Free-form Generation Enhances Challenging Clothed Human Modeling

Paper link: https://arxiv.org/abs/2411.19942

YouTube Result:

1. **Motivation**: The study aims to improve the modeling of animatable human avatars, specifically focusing on creating realistic representations of clothed humans based on 3D pose parameters and specific garment types.

2. **Novelty**: The novel aspect of the study is the hybrid framework that segments the human body into three distinct categories (unclothed, deformed, and generated) and employs a specialized strategy to model different clothing regions effectively, utilizing a free form part-aware generation module for loose clothing.

3. **Main Findings**: The method achieves superior visual quality and realism compared to previous works, capturing more high-fidelity details in the modeling of loose clothing. It effectively

differentiates between regions that require deformation and those that do not, enhancing the overall realism of the clothed human avatars.

4. **Video Title**: [CVPR 2025 Highlight] FreeCloth: Free-form Generation Enhances Challenging Clothed Human Modeling

5. **Video Link**: [Watch here](https://www.youtube.com/watch?v=4dyM1hjTBdQ)

Paper: Volume Tells: Dual Cycle-Consistent Diffusion for 3D Fluorescence Microscopy De-noising and Super-Resolution

Paper link: https://arxiv.org/abs/2503.02261

YouTube Result:

I don't know.

Paper: UltraFusion: Ultra High Dynamic Imaging using Exposure Fusion

Paper link: https://arxiv.org/abs/2501.11515

YouTube Result:

I don't know.

Paper: SpecTRe-GS: Modeling Highly Specular Surfaces with Reflected Nearby Objects by Tracing Rays in 3D Gaussian Splatting

Paper link: Not found on arXiv

YouTube Result:

I don't know.

Paper: Inst3D-LMM: Instance-Aware 3D Scene Understanding with Multi-modal

Instruction Tuning

Paper link: https://arxiv.org/abs/2503.00513

YouTube Result:

I don't know.

Paper: Flowing from Words to Pixels: A Noise-Free Framework for Cross-Modality Evolution

Paper link: https://arxiv.org/abs/2412.15213

YouTube Result:

I don't know.

Paper: No Pains, More Gains: Recycling Sub-Salient Patches for Efficient High-Resolution Image Recognition

Paper link: Not found on arXiv

YouTube Result:

I don't know.

Paper: Light Transport-aware Diffusion Posterior Sampling for Single-View Reconstruction of 3D Volumes

Paper link: https://arxiv.org/abs/2501.05226

YouTube Result:

I don't know.

Paper: Rethinking Personalized Aesthetics Assessment: Employing Physique Aesthetics Assessment as An Exemplification

Paper link: Not found on arXiv

YouTube Result:

I don't know.

Paper: Breaking the Memory Barrier of Contrastive Loss via Tile-Based Strategy

Paper link: Not found on arXiv

YouTube Result:

I don't know.

Paper: ArcPro: Architectural Programs for Structured 3D Abstraction of Sparse Points

Paper link: https://arxiv.org/abs/2503.02745

YouTube Result:

I don't know.

Paper: Your Large Vision-Language Model Only Needs A Few Attention Heads For Visual Grounding

Paper link: https://arxiv.org/abs/2503.06287

YouTube Result:

I don't know.

Paper: Structure from Collision

Paper link: https://arxiv.org/abs/2505.04134

YouTube Result:

1. **Motivation**: The study aims to prepare individuals for collision repair by providing knowledge

about unitized structural body repair techniques, which are essential in modern automotive repairs.

2. **Novelty**: The video discusses specific repair methods and techniques related to unitized structural bodies, which are prevalent in late-model vehicles, emphasizing modern approaches to body repair.

3. **Main Findings**: Key points include the importance of adhesive mounted glass as a structural component, the definition and implications of asymmetrical dimensions, and the acceptable techniques for sectioning structural members when repairs are needed.

4. **Video Title**: Unitized Structural Body Repair Study Guide Questions & Answers | Collision Repair State Test

5. **Video Link**: [Watch here](https://www.youtube.com/watch?v=meIa93zfPk8)

Paper: OmniSplat: Taming Feed-Forward 3D Gaussian Splatting for Omnidirectional Images with Editable Capabilities

Paper link: https://arxiv.org/abs/2412.16604

YouTube Result:

I don't know.

Paper: VideoScene: Distilling Video Diffusion Model to Generate 3D Scenes in One Step

Paper link: https://arxiv.org/abs/2504.01956

YouTube Result:

I don't know.

Paper: USP-Gaussian: Unifying Spike-based Image Reconstruction, Pose Correction and Gaussian Splatting

Paper link: https://arxiv.org/abs/2411.10504

YouTube Result:

I don't know.

Paper: SLAM3R: Real-Time Dense Scene Reconstruction from Monocular RGB Videos

Paper link: https://arxiv.org/abs/2412.09401

YouTube Result:

I don't know.

Paper: MASt3R-SLAM: Real-Time Dense SLAM with 3D Reconstruction Priors

Paper link: https://arxiv.org/abs/2412.12392

YouTube Result:

1. **Motivation**: The study aims to develop a real-time system for dense SLAM (Simultaneous Localization and Mapping) using just a single camera, leveraging a powerful 3D reconstruction prior to achieve robust results on real-world videos.

2. **Novelty**: The novel aspect of this study is the introduction of the MAS 3R SLAM system, which utilizes a 3D reconstruction prior for dense SLAM. This approach allows for real-time processing and accurate mapping with minimal hardware requirements.

3. **Main Findings**: The main findings indicate that the MAS 3R SLAM system successfully generates a dense 3D point map from real-time video input, demonstrating effective camera pose tracking and point map fusion, as illustrated by the system's output during tests on various

sequences.

4. **Video Title**: MASt3R-SLAM: Real-Time Dense 3D Mapping with Priors

5. **Video Link**: [MASt3R-SLAM: Real-Time Dense 3D Mapping with Priors](https://www.youtube.com/watch?v=fSu0X9xsOqY)

Paper: Self-Supervised Cross-View Correspondence with Predictive Cycle Consistency

Paper link: Not found on arXiv

YouTube Result:

I don't know.

Paper: BADGR: Bundle Adjustment Diffusion Conditioned by Gradients for Wide-Baseline Floor Plan Reconstruction

Paper link: https://arxiv.org/abs/2503.19340

YouTube Result:

I don't know.

Paper: Light3R-SfM: Towards Feed-forward Structure-from-Motion

Paper link: https://arxiv.org/abs/2501.14914

YouTube Result:

I don't know.

Paper: Full-DoF Egomotion Estimation for Event Cameras Using Geometric Solvers

Paper link: https://arxiv.org/abs/2503.03307

YouTube Result:

I don't know.

Paper: Active Hyperspectral Imaging Using an Event Camera

Paper link: Not found on arXiv

YouTube Result:

I don't know.

Paper: Shape Abstraction via Marching Differentiable Support Functions

Paper link: Not found on arXiv

YouTube Result:

I don't know.

Paper: Implicit Correspondence Learning for Image-to-Point Cloud Registration

Paper link: https://arxiv.org/abs/2307.07693

YouTube Result:

I don't know.

Paper: Glossy Object Reconstruction with Cost-effective Polarized Acquisition

Paper link: https://arxiv.org/abs/2504.07025

YouTube Result:

I don't know.

Paper: Universal Scene Graph Generation

Paper link: https://arxiv.org/abs/2504.01924

YouTube Result:

I don't know.

Paper: ICP: Immediate Compensation Pruning for Mid-to-high Sparsity

Paper link: Not found on arXiv

YouTube Result:

I don't know.

Paper: Can Generative Video Models Help Pose Estimation?

Paper link: https://arxiv.org/abs/2412.16155

YouTube Result:

1. **Motivation**: The study investigates whether pre-trained video generators, which have shown unexpected capabilities in 3D tasks, can be utilized for camera pose estimation, particularly in the context of generative video models that are trained on extensive datasets.

2. **Novelty**: The novel aspect of the study lies in the exploration of combining video generation models with camera pose estimation, suggesting that features from video generation can enhance performance in understanding scene geometry and viewpoint changes.

3. **Main Findings**: The findings indicate that generative models trained on 2D data, such as video diffusion models, exhibit capabilities that can be applied to 3D tasks, including camera pose estimation. The architecture introduced, which integrates video generation with camera pose estimation, shows potential for improving accuracy in pose estimation tasks.

4. **Video Title**: On Unifying Video Generation and Camera Pose Estimation

5. **Video Link**: [Watch here](https://www.youtube.com/watch?v=0oP-t5hLXUw)

Paper: Enduring, Efficient and Robust Trajectory Prediction Attack in Autonomous Driving via Optimization-Driven Multi-Frame Perturbation Framework

Paper link: Not found on arXiv

YouTube Result:

I don't know.

Paper: GEN3C: 3D-Informed World-Consistent Video Generation with Precise Camera Control

Paper link: https://arxiv.org/abs/2503.03751

YouTube Result:

1. **Motivation**: The motivation behind the study is to improve the consistency and control in AI-generated video content, addressing common issues such as objects disappearing or appearing unexpectedly and erratic camera movements that detract from the viewing experience.

2. **Novelty**: The novel aspect of the study is the introduction of a "3D cache," which acts as a memory bank of point clouds derived from depth predictions of previous frames, enabling more precise camera control and maintaining 3D consistency during video generation.

3. **Main Findings**: The main findings indicate that GEN3C effectively resolves issues related to 3D consistency and camera control by utilizing the 3D cache to inform the generation of new frames based on user-defined camera movements, allowing for smooth transitions and stable video output.

4. **Video Title**: GEN3C: 3D-Informed World-Consistent Video Generation with Precise Camera Control (Paper Walkthrough)

5. **Video Link**: [Watch here](https://www.youtube.com/watch?v=Q80Mgm-0JCM)

Paper: SpatialLLM: A Compound 3D-Informed Design towards Spatially-Intelligent Large Multimodal Models

Paper link: https://arxiv.org/abs/2505.00788

YouTube Result:

I don't know.

Paper: DriveGPT4-V2: Harnessing Large Language Model Capabilities for Enhanced Closed-Loop Autonomous Driving

Paper link: Not found on arXiv

YouTube Result:

I don't know.

Paper: OmniManip: Towards General Robotic Manipulation via Object-Centric Interaction Primitives as Spatial Constraints

Paper link: https://arxiv.org/abs/2501.03841

YouTube Result:

I don't know.

Paper: Generating 6DoF Object Manipulation Trajectories from Action Description in Egocentric Vision

Paper link: Not found on arXiv

YouTube Result:

I don't know.

Paper: Erase Diffusion: Empowering Object Removal Through Calibrating Diffusion Pathways

Paper link: https://arxiv.org/abs/2503.07026

YouTube Result:

I don't know.

Paper: GigaHands: A Massive Annotated Dataset of Bimanual Hand Activities

Paper link: https://arxiv.org/abs/2412.04244

YouTube Result:

1. **Motivation**: The study aims to address the significant challenge of capturing the complexity of everyday activities involving both hands in AI and robotics, particularly focusing on human dexterity and intricate manipulation tasks. It highlights the bottleneck in progress due to the lack of suitable large-scale, diverse, and accurately annotated 3D datasets.

2. **Novelty**: The novel aspect of the study is the introduction of Giga Hands, a massive annotated dataset specifically designed to capture natural bimanual activities with high 3D accuracy and realism. This dataset addresses the shortcomings of existing datasets, which often fall short in 3D accuracy, realism, and comprehensive annotations.

3. **Main Findings**: The study finds that existing datasets often suffer from issues like unnatural interactions, missing 3D information, inaccurate annotations, and a lack of semantic depth. Giga Hands aims to provide a solution by offering a more comprehensive and realistic dataset for training models that require detailed descriptions of bimanual activities.

4. **Video Title**: [Seminar] Advancing Large-Scale Dataset for Bimanual Hand-Object Interaction

5. **Video Link**: https://www.youtube.com/watch?v=qEi9wJ1NKqw

Paper: FRAME: Floor-aligned Representation for Avatar Motion from Egocentric Video

Paper link: https://arxiv.org/abs/2503.23094

YouTube Result:

I don't know.

Paper: Lifting Motion to the 3D World via 2D Diffusion

Paper link: https://arxiv.org/abs/2411.18808

YouTube Result:

1. **Motivation**: The study aims to generate 3D motion from 2D pose sequences extracted from monocular videos without the need for training on any 3D motion data. It addresses the challenge of limited availability of consistent multi-view data for training.

2. **Novelty**: The novel aspect of the study is the reformulation of 3D motion estimation as generating consistent multi-view 2D pose sequences. The framework leverages 2D motion diffusion to progressively establish multi-view consistency, which is a unique approach compared to traditional methods that require extensive multi-view data.

3. **Main Findings**: The approach successfully estimates complete 3D motion, including joint

rotations and root translations, using only single-view 2D sequences. It builds multi-view consistency through four stages, ultimately training a multi-view 2D motion diffusion model capable of generating consistent sequences in a single forward pass.

4. **Video Title**: Lifting Motion to the 3D World via 2D Diffusion

5. **Video Link**: [Lifting Motion to the 3D World via 2D Diffusion](https://www.youtube.com/watch?v=nffTJHUR8yw)

Paper: RGBAvatar: Reduced Gaussian Blendshapes for Online Modeling of Head Avatars

Paper link: https://arxiv.org/abs/2503.12886

YouTube Result:

1. **Motivation**: The study aims to develop an efficient method for reconstructing head avatar models from monocular video input, enabling real-time animation and rendering of avatars.

2. **Novelty**: The method presented is novel in its ability to achieve fast reconstruction (about 80 seconds for a 2-3 minute video) and real-time rendering (approximately 400 fps) while capturing finer details than existing Gaussian blend shape methods, including teeth and eyeglass reflections.

3. **Main Findings**: The study finds that their method significantly improves results for expressive facial animations and requires only 20 reduced blend shapes to effectively capture essential details like wrinkles. It also supports online reconstruction, allowing for immediate visual feedback during data capture.

4. **Video Title**: RGBAvatar Reduced Gaussian Blendshapes

5. **Video Link**: [Watch here](https://www.youtube.com/watch?v=r0Rl6t-Btlc)

Paper: EnergyMoGen: Compositional Human Motion Generation with Energy-Based Diffusion Model in Latent Space

Paper link: https://arxiv.org/abs/2412.14706

YouTube Result:

I don't know.

Paper: FIction: 4D Future Interaction Prediction from Video

Paper link: https://arxiv.org/abs/2412.00932

YouTube Result:

I don't know.

Paper: FoundHand: Large-Scale Domain-Specific Learning for Controllable Hand Image Generation

Paper link: https://arxiv.org/abs/2412.02690

YouTube Result:

I don't know.

Paper: 4Real-Video: Learning Generalizable Photo-Realistic 4D Video Diffusion

Paper link: https://arxiv.org/abs/2412.04462

YouTube Result:

I don't know.

Paper: EvEnhancer: Empowering Effectiveness, Efficiency and Generalizability for Continuous Space-Time Video Super-Resolution with Events

Paper link: Not found on arXiv

YouTube Result:

I don't know.

Paper: Mamba as a Bridge: Where Vision Foundation Models Meet Vision Language Models for Domain-Generalized Semantic Segmentation

Paper link: https://arxiv.org/abs/2504.03193

YouTube Result:

I don't know.

Paper: Balanced Rate-Distortion Optimization in Learned Image Compression

Paper link: https://arxiv.org/abs/2502.20161

YouTube Result:

I don't know.

Paper: DyFo: A Training-Free Dynamic Focus Visual Search for Enhancing LMMs in Fine-Grained Visual Understanding

Paper link: https://arxiv.org/abs/2504.14920

YouTube Result:

I don't know.

Paper: LP-Diff: Towards Improved Restoration of Real-World Degraded License Plate

Paper link: Not found on arXiv

YouTube Result:

I don't know.

Paper: CoSER: Towards Consistent Dense Multiview Text-to-Image Generator for 3D Creation

Paper link: Not found on arXiv

YouTube Result:

I don't know.

Paper: CoMM: A Coherent Interleaved Image-Text Dataset for Multimodal Understanding and Generation

Paper link: https://arxiv.org/abs/2406.10462

YouTube Result:

I don't know.

Paper: UniRestore: Unified Perceptual and Task-Oriented Image Restoration Model Using Diffusion Prior

Paper link: https://arxiv.org/abs/2501.13134

YouTube Result:

I don't know.

Paper: Optimizing for the Shortest Path in Denoising Diffusion Model

Paper link: https://arxiv.org/abs/2503.03265

YouTube Result:

1. **Motivation**: The study aims to identify the optimal procedure for corrupting images in diffusion models, as existing methods have not clearly defined this optimal path, which is crucial for improving image generation quality.

2. **Novelty**: The novel aspect of the study is the introduction of the Shortest Path Diffusion (SPD) approach, which computes the optimal corruption procedure as the shortest path in the space of distributions using the Fisher metric. This approach contrasts with traditional methods that often rely on blurring.

3. **Main Findings**: The findings indicate that the SPD outperforms previous methods based on image blurring and shows that any deviation from the shortest path negatively impacts performance. The study demonstrates that the optimal corruption corresponds to a combination of image sharpening and noise deblurring, leading to improved results in image generation tasks.

4. **Video Title**: Image generation with shortest path diffusion - ArXiv:2306.00501

5. **Video Link**: [Watch here](https://www.youtube.com/watch?v=0O4I8zY4Nzw)

Paper: TinyFusion: Diffusion Transformers Learned Shallow

Paper link: https://arxiv.org/abs/2412.01199

YouTube Result:

I don't know.

Paper: Style-Editor: Text-driven Object-centric Style Editing

Paper link: https://arxiv.org/abs/2408.08461

YouTube Result:

I don't know.

Paper: MUSt3R: Multi-view Network for Stereo 3D Reconstruction

Paper link: https://arxiv.org/abs/2503.01661

YouTube Result:

I don't know.

Paper: DPU: Dynamic Prototype Updating for Multimodal Out-of-Distribution Detection

Paper link: https://arxiv.org/abs/2411.08227

YouTube Result:

I don't know.

Paper: Focus-N-Fix: Region-Aware Fine-Tuning for Text-to-Image Generation

Paper link: https://arxiv.org/abs/2501.06481

YouTube Result:

I don't know.

Paper: DefectFill: Realistic Defect Generation with Inpainting Diffusion Model for Visual Inspection

Paper link: https://arxiv.org/abs/2503.13985

YouTube Result:

I don't know.

Paper: Circumventing Shortcuts in Audio-visual Deepfake Detection Datasets

with Unsupervised Learning

Paper link: https://arxiv.org/abs/2412.00175

YouTube Result:

I don't know.

Paper: UIBDiffusion: Universal Imperceptible Backdoor Attack for Diffusion Models

Paper link: https://arxiv.org/abs/2412.11441

YouTube Result:

I don't know.

Paper: Open-Vocabulary Functional 3D Scene Graphs for Real-World Indoor Spaces

Paper link: https://arxiv.org/abs/2503.19199

YouTube Result:

1. **Motivation**: The study addresses the limitations of current mapping techniques, such as V maps and concept fusion, which struggle to scale to larger scenes due to high storage overhead. The aim is to develop a more efficient method for mapping large-scale environments.

2. **Novelty**: The introduction of hierarchical open vocabulary 3D scene graphs (HOV-SG) for language-grounded robot navigation, which enhances 3D scene graphs by integrating open vocabulary features without explicitly modeling hierarchical semantics, is a key novel aspect of the study.

3. **Main Findings**: The study presents a two-stage pipeline for creating a segment-level open vocabulary map and constructing a scene graph that incorporates RGBD sequences and odometry.

The method improves the mapping process and enables better handling of large-scale environments.

4. **Video Title**: HOV-SG: Hierarchical Open-Vocabulary 3D Scene Graphs for Language-Grounded Robot Navigation (RSS'24)

5. **Video Link**: [Watch here](https://www.youtube.com/watch?v=GC-Q0ekO9qg)

Paper: AnySat: One Earth Observation Model for Many Resolutions, Scales, and Modalities

Paper link: https://arxiv.org/abs/2412.14123

YouTube Result:

I don't know.

Paper: All Languages Matter: Evaluating LMMs on Culturally Diverse 100 Languages

Paper link: https://arxiv.org/abs/2411.16508

YouTube Result:

I don't know.

Paper: CL-MoE: Enhancing Multimodal Large Language Model with Dual Momentum Mixture-of-Experts for Continual Visual Question Answering

Paper link: https://arxiv.org/abs/2503.00413

YouTube Result:

I don't know.

Paper: Improving Personalized Search with Regularized Low-Rank Parameter Updates

Paper link: Not found on arXiv

YouTube Result:

I don't know.

Paper: PhD: A ChatGPT-Prompted Visual Hallucination Evaluation Dataset

Paper link: https://arxiv.org/abs/2403.11116

YouTube Result:

I don't know.

Paper: O-TPT: Orthogonality Constraints for Calibrating Test-time Prompt Tuning in Vision-Language Models

Paper link: https://arxiv.org/abs/2503.12096

YouTube Result:

I don't know.

Paper: RLAIF-V: Open-Source AI Feedback Leads to Super GPT-4V Trustworthiness

Paper link: https://arxiv.org/abs/2405.17220

YouTube Result:

I don't know.

Paper: F^3OCUS - Federated Finetuning of Vision-Language Foundation Models with Optimal Client Layer Updating Strategy via Multi-objective

Meta-Heuristics

Paper link: Not found on arXiv

YouTube Result:

I don't know.

Paper: CARE Transformer: Mobile-Friendly Linear Visual Transformer via Decoupled Dual Interaction

Paper link: https://arxiv.org/abs/2411.16170

YouTube Result:

I don't know.

Paper: Coeff-Tuning: A Graph Filter Subspace View for Tuning Attention-Based Large Models

Paper link: https://arxiv.org/abs/2503.18337

YouTube Result:

I don't know.

Paper: Generative Modeling of Class Probability for Multi-Modal Representation Learning

Paper link: https://arxiv.org/abs/2503.17417

YouTube Result:

I don't know.

Paper: STING-BEE: Towards Vision-Language Model for Real-World X-ray Baggage Security Inspection

Paper link: https://arxiv.org/abs/2504.02823

YouTube Result:

I don't know.

Paper: Perceptually Accurate 3D Talking Head Generation: New Definitions, Speech-Mesh Representation, and Evaluation Metrics

Paper link: https://arxiv.org/abs/2503.20308

YouTube Result:

I don't know.

Paper: PGC: Physics-Based Gaussian Cloth from a Single Pose

Paper link: https://arxiv.org/abs/2503.20779

YouTube Result:

1. **Motivation**: The study aims to reconstruct simulation-ready garments with intricate appearances from a single pose, addressing the challenges of realistic garment representation in computer graphics.

2. **Novelty**: The study introduces a hybrid approach that combines mesh embedded 3D Gaussian splats with physics-based simulation and surface rendering techniques, allowing for better pose generalization and detailed garment representation.

3. **Main Findings**: The findings indicate that their method successfully reconstructs garment details while achieving realistic appearances through a combination of high pass and low pass shading techniques, outperforming traditional methods that either lack realism in deformation or detailed shading.

4. **Video Title**: PGC: Physics-based Gaussian Cloth from a Single Pose

5. **Video Link**: https://www.youtube.com/watch?v=Pi4Kw2wUBSU

Paper: Is this Generated Person Existed in Real-world? Fine-grained Detecting and Calibrating Abnormal Human-body

Paper link: https://arxiv.org/abs/2411.14205

YouTube Result:

I don't know.

Paper: 3D Convex Splatting: Radiance Field Rendering with 3D Smooth Convexes

Paper link: https://arxiv.org/abs/2411.14974

YouTube Result:

1. **Motivation**: The study aims to address the limitations of existing techniques in 3D radiance field rendering, specifically the challenges posed by continuous models that struggle to accurately represent flat surfaces and sharp edges in realistic scenes.

2. **Novelty**: The novel aspect of the study is the introduction of 3D convex splatting, which utilizes 3D smooth convexes. This approach allows for the modeling of both smooth transitions and sharp boundaries, striking a balance between various representation methods.

3. **Main Findings**: The main findings indicate that 3D convex splatting can effectively perform novel view synthesis and 3D reconstruction while accurately capturing sharp edges and complex

geometries, improving upon previous methods that could not handle these features well.

4. **Video Title**: 3D Convex Splatting: Radiance Field Rendering with 3D Smooth Convexes

5. **Video Link**: [Watch here](https://www.youtube.com/watch?v=5N3OFHH7lbU)

Paper: MetricGrids: Arbitrary Nonlinear Approximation with Elementary Metric Grids based Implicit Neural Representation

Paper link: https://arxiv.org/abs/2503.10000

YouTube Result:

I don't know.

Paper: ARM: Appearance Reconstruction Model for Relightable 3D Generation

Paper link: https://arxiv.org/abs/2411.10825

YouTube Result:

I don't know.

Paper: CADDreamer: CAD Object Generation from Single-view Images

Paper link: https://arxiv.org/abs/2502.20732

YouTube Result:

I don't know.

Paper: High-fidelity 3D Object Generation from Single Image with RGBN-Volume Gaussian Reconstruction Model

Paper link: https://arxiv.org/abs/2504.01512

YouTube Result:

I don't know.

Paper: Panorama Generation From NFoV Image Done Right

Paper link: https://arxiv.org/abs/2503.18420

YouTube Result:

I don't know.

Paper: World-consistent Video Diffusion with Explicit 3D Modeling

Paper link: https://arxiv.org/abs/2503.07135

YouTube Result:

I don't know.

Paper: Improving Gaussian Splatting with Localized Points Management

Paper link: https://arxiv.org/abs/2411.08373

YouTube Result:

I don't know.

Paper: Event Ellipsometer: Event-based Mueller-Matrix Video Imaging

Paper link: https://arxiv.org/abs/2411.17313

YouTube Result:

I don't know.

Paper: All-directional Disparity Estimation for Real-world QPD Images

Paper link: Not found on arXiv

YouTube Result:

I don't know.

Paper: Reconstructing People, Places, and Cameras

Paper link: https://arxiv.org/abs/2412.17806

YouTube Result:

I don't know.

Paper: SeCap: Self-Calibrating and Adaptive Prompts for Cross-view Person Re-Identification in Aerial-Ground Networks

Paper link: https://arxiv.org/abs/2503.06965

YouTube Result:

I don't know.

Paper: Sonata: Self-Supervised Learning of Reliable Point Representations

Paper link: https://arxiv.org/abs/2503.16429

YouTube Result:

I don't know.

Paper: BWFormer: Building Wireframe Reconstruction from Airborne LiDAR Point Cloud with Transformer

Paper link: Not found on arXiv

YouTube Result:

I don't know.

Paper: DexGrasp Anything: Towards Universal Robotic Dexterous Grasping with Physics Awareness

Paper link: https://arxiv.org/abs/2503.08257

YouTube Result:

1. **Motivation**: The motivation behind the study is to develop a dexterous robotic hand capable of grasping any object, which is essential for creating general-purpose embodied robots. The challenge lies in the complexity of hand degrees of freedom and the diversity of objects, making it difficult to generate robust and high-quality grasp poses.

2. **Novelty**: The novel aspect of the study is the introduction of a method called DexGrasp Anything, which integrates physical constraints into both the training and sampling phases of a diffusion-based generative model. This approach allows for more effective grasping by incorporating physics awareness into the generative process.

3. **Main Findings**: The main findings indicate that by using a diffusion model that combines object point cloud data, shadow hand pose parameters, and physical constraints, the system can progressively transform hand parameters into noise and derive cleaner and grasp-suitable configurations. This method enables effective grasping of diverse objects by guiding the noise distribution toward physically feasible configurations.

4. **Video Title**: DexGrasp Anything: Towards Universal Robotic Dexterous Grasping with Physics Awareness (CVPR 2025)

5. **Video Link**: [Watch here](https://www.youtube.com/watch?v=KNEY_LKG5Y4)

Paper: H-MoRe: Learning Human-centric Motion Representation for Action

Analysis

Paper link: https://arxiv.org/abs/2504.10676

YouTube Result:

I don't know.

Paper: Tracktention: Leveraging Point Tracking to Attend Videos Faster and Better

Paper link: https://arxiv.org/abs/2503.19904

YouTube Result:

I don't know.

Paper: Align3R: Aligned Monocular Depth Estimation for Dynamic Videos

Paper link: https://arxiv.org/abs/2412.03079

YouTube Result:

I don't know.

Paper: Meta-Learning Hyperparameters for Parameter Efficient Fine-Tuning

Paper link: https://arxiv.org/abs/2008.05984

YouTube Result:

1. **Motivation**: The study addresses the challenge of fine-tuning natural language models efficiently and accurately, particularly in the context of few-shot learning, where there is limited labeled data available for new tasks.

2. **Novelty**: The approach discussed is focused on making fine-tuning of NLP models both parameter-efficient and capable of generalizing from few examples, which is a critical requirement

for users with limited resources and data.

3. **Main Findings**: The findings suggest that by leveraging few-shot learning capabilities, models can be adapted effectively to new tasks with minimal additional data, thus making them practical for end-users of cloud-based machine learning services.

4. **Video Title**: [AutoMLConf'22]: Meta-Adapters: Parameter Efficient Few-shot Fine-tuning through Meta-Learning

5. **Video Link**: [Watch here](https://www.youtube.com/watch?v=PdB80toLiAw)

Paper: Understanding Multi-layered Transmission Matrices

Paper link: https://arxiv.org/abs/2410.23864

YouTube Result:

1. **Motivation**: The motivation behind the study of multi-layered transmission matrices is to simplify the analysis of optical systems involving multiple layers, where calculating reflection and transmission for a complex system can become mathematically challenging.

2. **Novelty**: The novel aspect of the study is the application of the transfer matrix method, which allows for a systematic and manageable approach to analyze layered optical structures by breaking down the problem into simpler, individual layer interactions rather than attempting to solve the entire system at once.

3. **Main Findings**: The main findings indicate that using the transfer matrix method enables more efficient calculations of reflection and transmission in multi-layered systems by handling one interface at a time, thus avoiding the complex geometric series that would otherwise be required for

multiple layers.

4. **Video Title**: Transfer Matrix Method Explained

5. **Video Link**: [Transfer Matrix Method Explained](https://www.youtube.com/watch?v=XuSxmb9-viY)

Paper: Good, Cheap, and Fast: Overfitted Image Compression with Wasserstein Distortion

Paper link: https://arxiv.org/abs/2412.00505

YouTube Result:

I don't know.

Paper: Visual Representation Learning through Causal Intervention for Controllable Image Editing

Paper link: Not found on arXiv

YouTube Result:

I don't know.

Paper: Boost Your Human Image Generation Model via Direct Preference Optimization

Paper link: https://arxiv.org/abs/2405.20216

YouTube Result:

I don't know.

Paper: ReNeg: Learning Negative Embedding with Reward Guidance

Paper link: https://arxiv.org/abs/2412.19637

YouTube Result:

I don't know.

Paper: STEREO: A Two-Stage Framework for Adversarially Robust Concept Erasing from Text-to-Image Diffusion Models

Paper link: https://arxiv.org/abs/2408.16807

YouTube Result:

I don't know.

Paper: Supervising Sound Localization by In-the-wild Egomotion

Paper link: Not found on arXiv

YouTube Result:

I don't know.

Paper: Cross-modal Causal Relation Alignment for Video Question Grounding

Paper link: https://arxiv.org/abs/2503.07635

YouTube Result:

I don't know.

Paper: Video-MME: The First-Ever Comprehensive Evaluation Benchmark of Multi-modal LLMs in Video Analysis

Paper link: https://arxiv.org/abs/2405.21075

YouTube Result:

I don't know.

Paper: Just Dance with pi! A Poly-modal Inductor for Weakly-supervised Video Anomaly Detection

Paper link: Not found on arXiv

YouTube Result:

I don't know.

Paper: Learning 4D Panoptic Scene Graph Generation from Rich 2D Visual Scene

Paper link: https://arxiv.org/abs/2503.15019

YouTube Result:

I don't know.

Paper: The Scene Language: Representing Scenes with Programs, Words, and Embeddings

Paper link: https://arxiv.org/abs/2410.16770

YouTube Result:

I don't know.

Paper: VL-RewardBench: A Challenging Benchmark for Vision-Language Generative Reward Models

Paper link: https://arxiv.org/abs/2503.06260

YouTube Result:

I don't know.

Paper: Spatial457: A Diagnostic Benchmark for 6D Spatial Reasoning of Large

Mutimodal Models

Paper link: Not found on arXiv

YouTube Result:

I don't know.

Paper: Not Only Text: Exploring Compositionality of Visual Representations in Vision-Language Models

Paper link: https://arxiv.org/abs/2503.17142

YouTube Result:

I don't know.

Paper: Realistic Test-Time Adaptation of Vision-Language Models

Paper link: https://arxiv.org/abs/2501.03729

YouTube Result:

1. **Motivation**: The study addresses the challenge of adapting machine learning models at test time, particularly when the distribution of input data diverges from that seen during training. This is crucial for maintaining model performance in real-world scenarios where conditions change.

2. **Novelty**: The presentation introduces the concept of supervised test-time adaptation, distinguishing between scenarios where test distributions are static versus those that change over time, which is a less explored area in the context of deep learning.

3. **Main Findings**: The findings suggest that models trained under traditional conditions may struggle when faced with data that diverges from the training set, indicating a need for methodologies that can dynamically adapt to these changes during deployment.

4. **Video Title**: ICCV 2023 Tutorial: Test-time Adaptation: Formulations, Methods and Benchmarks

5. **Video Link**: [ICCV 2023 Tutorial](https://www.youtube.com/watch?v=l584yXZfYx4)

Paper: Comprehensive Information Bottleneck for Unveiling Universal Attribution to Interpret Vision Transformers

Paper link: Not found on arXiv

YouTube Result:

I don't know.

Paper: T2ICount: Enhancing Cross-modal Understanding for Zero-Shot Counting

Paper link: https://arxiv.org/abs/2502.20625

YouTube Result:

I don't know.

Paper: WISH: Weakly Supervised Instance Segmentation using Heterogeneous Labels

Paper link: Not found on arXiv

YouTube Result:

I don't know.

Paper: Project-Probe-Aggregate: Efficient Fine-Tuning for Group Robustness

Paper link: https://arxiv.org/abs/2503.09487

YouTube Result:

I don't know.

Paper: Dataset Distillation with Neural Characteristic Function: A Minmax Perspective

Paper link: https://arxiv.org/abs/2502.20653

YouTube Result:

I don't know.

Paper: SoMA: Singular Value Decomposed Minor Components Adaptation for Domain Generalizable Representation Learning

Paper link: https://arxiv.org/abs/2412.04077

YouTube Result:

I don't know.

Paper: Free-viewpoint Human Animation with Pose-correlated Reference Selection

Paper link: https://arxiv.org/abs/2412.17290

YouTube Result:

I don't know.

Paper: Real-time High-fidelity Gaussian Human Avatars with Position-based Interpolation of Spatially Distributed MLPs

Paper link: https://arxiv.org/abs/2504.12909

YouTube Result:

1. **Motivation**: The motivation behind the study is to create high-fidelity human avatars from multi-view videos that can be animated under novel views and poses, addressing the limitations of existing methods in capturing fine details such as wrinkles.

2. **Novelty**: The novel aspects of the study include the use of a new method that achieves a significantly faster rendering speed compared to state-of-the-art methods and captures finer details than previous approaches, particularly in rendering realistic avatars under various poses.

3. **Main Findings**: The study found that their method can model high-fidelity appearance under novel poses and effectively captures details like wrinkles, which other methods, such as 3DGS avatars and mesh avatars, fail to do. The research also discusses the importance of design choices like spatially distributed MLPs and control points in improving detail capture and rendering quality.

4. **Video Title**: Real-time High-fidelity Gaussian Human Avatars with Position-based Interpolation

5. **Video Link**: https://www.youtube.com/watch?v=TeTO4tYRdjw

Paper: Material Anything: Generating Materials for Any 3D Object via Diffusion

Paper link: https://arxiv.org/abs/2411.15138

YouTube Result:

1. **Motivation**: The study aims to address the significant challenges in creating realistic materials for 3D objects, which is a major bottleneck in computer graphics. Traditional methods are time-consuming, require artistic skill, and often fail to adapt materials to varying lighting conditions.

2. **Novelty**: The novel aspect of this study is the development of a method that automates the generation of realistic materials for any 3D object, thus eliminating the need for manual material creation. This innovation has the potential to revolutionize the way 3D worlds are created and experienced.

3. **Main Findings**: The main findings highlight that the proposed approach enables the creation of hyper-realistic materials, enhancing the realism in video games, virtual reality experiences, and product prototypes. It overcomes the limitations of traditional methods by allowing materials to look natural under different lighting conditions.

4. **Video Title**: Material Anything: Generating Materials for Any 3D Object via Diffusion

5. **Video Link**: [Material Anything: Generating Materials for Any 3D Object via Diffusion](https://www.youtube.com/watch?v=lqn7v08qR0I)

Paper: Generative Densification: Learning to Densify Gaussians for High-Fidelity Generalizable 3D Reconstruction

Paper link: https://arxiv.org/abs/2412.06234

YouTube Result:

I don't know.

Paper: NexusGS: Sparse View Synthesis with Epipolar Depth Priors in 3D Gaussian Splatting

Paper link: https://arxiv.org/abs/2503.18794

YouTube Result:

1. **Motivation**: The study aims to address the limitations of prior sparse view synthesis methods

that rely on monocular depth estimation networks, which often lead to inaccuracies in depth predictions that affect scene reconstruction quality.

2. **Novelty**: The novel aspect of this study is the introduction of an epipolar depth prior, which provides a more reliable foundation for constructing accurate dense point clouds, as opposed to the traditional methods that generate depth maps with potential inaccuracies.

3. **Main Findings**: The main findings indicate that the proposed method significantly reduces artifacts such as floaters and aliasing while enhancing the detail and coherence of the reconstructed 3D surfaces. Qualitative results show superior performance on various datasets compared to existing methods, especially in large-scale scenes.

4. **Video Title**: NexusGS: Sparse View Synthesis with Epipolar Depth Priors in 3D Gaussian Splatting

5. **Video Link**: [Watch here](https://www.youtube.com/watch?v=K2foTIXzpMQ)

Paper: Event Fields: Capturing Light Fields at High Speed, Resolution, and Dynamic Range

Paper link: https://arxiv.org/abs/2412.06191

YouTube Result:

1. **Motivation**: The study is motivated by the limitations of traditional RGB cameras, which struggle to capture high-speed, high spatial resolution videos due to massive bandwidth requirements. Event cameras, in contrast, can capture sparse brightness changes asynchronously, making them suitable for high-speed imaging.

2. **Novelty**: The novel aspect of this study is the development of "event fields," which combine event cameras with light field technology. This innovative approach uses optical designs to capture light fields at high speed, spatial angular resolution, and dynamic range, addressing the need for capturing angular information in addition to spatial and temporal data.

3. **Main Findings**: The main findings indicate that event fields can successfully multiplex angular information into either the spatial or temporal dimension, allowing for enhanced imaging capabilities. The study demonstrates the use of a rectangular Kaleidoscope design for spatial multiplexing and the incorporation of a galvanometer for temporal multiplexing, enabling the capture of angular derivatives of scenes at full spatial resolution.

4. **Video Title**: Event fields: Capturing light fields at high speed, resolution, and dynamic range

5. **Video Link**: [Watch here](https://www.youtube.com/watch?v=R00mDG9ZBo8)

Paper: IncEventGS: Pose-Free Gaussian Splatting from a Single Event Camera

Paper link: https://arxiv.org/abs/2410.08107

YouTube Result:

I don't know.

Paper: HELVIPAD: A Real-World Dataset for Omnidirectional Stereo Depth Estimation

Paper link: https://arxiv.org/abs/2503.23502

YouTube Result:

I don't know.

Paper: Order-One Rolling Shutter Cameras

Paper link: https://arxiv.org/abs/2403.11295

YouTube Result:

1. **Motivation**: The study is motivated by the need to understand how rolling shutter cameras, which are commonly used in smartphones and other devices, function and how they can be modeled effectively.

2. **Novelty**: The novel aspect of the study lies in its approach to combining engineering insights with computer vision and algebraic geometry to model the behavior of rolling shutter cameras, which have become ubiquitous due to their cost-effectiveness.

3. **Main Findings**: The study discusses the implications of rolling shutter technology in cameras, noting that while it is prevalent, it introduces specific challenges in capturing images. The findings suggest that understanding these challenges can help in developing better models for camera operation.

4. **Video Title**: Kathlén Kohn (KTH)

5. **Video Link**: [Kathlén Kohn (KTH)](https://www.youtube.com/watch?v=yPLKDXRmZdA)

Paper: Towards In-the-wild 3D Plane Reconstruction from a Single Image

Paper link: Not found on arXiv

YouTube Result:

I don't know.

Paper: MATCHA: Towards Matching Anything

Paper link: Not found on arXiv

YouTube Result:

I don't know.

Paper: Simulator HC: Regression-based Online Simulation of Starting Problem-Solution Pairs for Homotopy Continuation in Geometric Vision

Paper link: https://arxiv.org/abs/2411.03745

YouTube Result:

I don't know.

Paper: GaussianUDF: Inferring Unsigned Distance Functions through 3D Gaussian Splatting

Paper link: https://arxiv.org/abs/2503.19458

YouTube Result:

I don't know.

Paper: Doppelgangers++: Improved Visual Disambiguation with Geometric 3D Features

Paper link: https://arxiv.org/abs/2412.05826

YouTube Result:

I don't know.

Paper: MITracker: Multi-View Integration for Visual Object Tracking

Paper link: https://arxiv.org/abs/2502.20111

YouTube Result:

I don't know.

Paper: Ev-3DOD: Pushing the Temporal Boundaries of 3D Object Detection with Event Cameras

Paper link: https://arxiv.org/abs/2502.19630

YouTube Result:

I don't know.

Paper: A Unified Approach to Interpreting Self-supervised Pre-training Methods for 3D Point Clouds via Interactions

Paper link: Not found on arXiv

YouTube Result:

I don't know.

Paper: Deep Change Monitoring: A Hyperbolic Representative Learning Framework and a Dataset for Long-term Fine-grained Tree Change Detection

Paper link: https://arxiv.org/abs/2503.00643

YouTube Result:

I don't know.

Paper: SplatFlow: Self-Supervised Dynamic Gaussian Splatting in Neural Motion Flow Field for Autonomous Driving

Paper link: https://arxiv.org/abs/2411.15482

YouTube Result:

I don't know.

Paper: Towards Autonomous Micromobility through Scalable Urban Simulation

Paper link: https://arxiv.org/abs/2505.00690

YouTube Result:

I don't know.

Paper: RoboTwin: Dual-Arm Robot Benchmark with Generative Digital Twins

Paper link: https://arxiv.org/abs/2504.13059

YouTube Result:

1. **Motivation**: The study addresses the challenge of training dual-arm robots for complex tasks, such as tool use and collaboration with humans. There is a recognized need for diverse and high-quality training data, which is often hard to obtain in real-world scenarios.

2. **Novelty**: The research introduces Robo Twin, a unique dataset that combines real-world robot data with synthetic data generated from virtual replicas of robots, known as digital twins. This innovative approach utilizes advanced AI technologies to create realistic training environments without the need for extensive physical setups.

3. **Main Findings**: The study demonstrates that by leveraging 3D generative models and language models, researchers can effectively generate realistic training data for dual-arm robots. This method allows for more efficient training processes, saving time and resources by reducing the reliance on real-world data collection.

4. **Video Title**: RoboTwin: Dual-Arm Robot Benchmark with Generative Digital Twins

5. **Video Link**: [RoboTwin Video](https://www.youtube.com/watch?v=7srcPIESTtw)

Paper: End-to-End HOI Reconstruction Transformer with Graph-based Encoding

Paper link: https://arxiv.org/abs/2503.06012

YouTube Result:

I don't know.

Paper: Dyn-HaMR: Recovering 4D Interacting Hand Motion from a Dynamic Camera

Paper link: https://arxiv.org/abs/2412.12861

YouTube Result:

I don't know.

Paper: MotionPRO: Exploring the Role of Pressure in Human MoCap and Beyond

Paper link: https://arxiv.org/abs/2504.05046

YouTube Result:

I don't know.

Paper: UniPose: A Unified Multimodal Framework for Human Pose Comprehension, Generation and Editing

Paper link: https://arxiv.org/abs/2411.16781

YouTube Result:

I don't know.

Paper: Unified Reconstruction of Static and Dynamic Scenes from Events

Paper link: Not found on arXiv

YouTube Result:

I don't know.

Paper: FreePCA: Integrating Consistency Information across Long-short Frames in Training-free Long Video Generation via Principal Component Analysis

Paper link: https://arxiv.org/abs/2505.01172

YouTube Result:

I don't know.

Paper: A Polarization-Aided Transformer for Image Deblurring via Motion Vector Decomposition

Paper link: Not found on arXiv

YouTube Result:

I don't know.

Paper: All-Optical Nonlinear Diffractive Deep Network for Ultrafast Image Denoising

Paper link: Not found on arXiv

YouTube Result:

I don't know.

Paper: FlexiDiT: Your Diffusion Transformer Can Easily Generate High-Quality Samples with Less Compute

Paper link: https://arxiv.org/abs/2502.20126

YouTube Result:

I don't know.

Paper: Gaze-LLE: Gaze Target Estimation via Large-Scale Learned Encoders

Paper link: https://arxiv.org/abs/2412.09586

YouTube Result:

1. **Motivation**: The study is motivated by the importance of gaze in human behavior, as it reflects individuals' engagement with their surroundings and aids in social interactions. Accurate gaze target estimation is crucial for systems that aim to understand and interpret human behavior, especially in joint attention scenarios.

2. **Novelty**: The Gaze-LLE model distinguishes itself from previous multi-branch architectures by utilizing a frozen large-scale foundational visual encoder (D2) that simplifies the overall architecture and improves efficiency. It eliminates the need for complex fusion modules by integrating head position prompting directly into a unified decoder, significantly reducing the number of learnable parameters while maintaining high performance.

3. **Main Findings**: The Gaze-LLE model demonstrates that it can effectively capture gaze cues using a streamlined architecture, which not only simplifies training but also enhances generalization across various datasets. It achieves state-of-the-art performance in gaze target estimation without relying on specialized encoders, which often complicate the training process.

4. **Video Title**: Gaze-LLE: Gaze Target Estimation via Large-Scale Learned Encoders

5. **Video Link**:

https://www.youtube.com/watch?v=KFfyjZ_4kNo

Paper: Which Viewpoint Shows it Best? Language for Weakly Supervising View Selection in Multi-view Instructional Videos

Paper link: https://arxiv.org/abs/2411.08753

YouTube Result:

I don't know.

Paper: CASAGPT: Cuboid Arrangement and Scene Assembly for Interior Design

Paper link: https://arxiv.org/abs/2504.19478

YouTube Result:

I don't know.

Paper: Revealing Key Details to See Differences: A Novel Prototypical Perspective for Skeleton-based Action Recognition

Paper link: https://arxiv.org/abs/2411.18941

YouTube Result:

I don't know.

Paper: SmartCLIP: Modular Vision-language Alignment with Identification Guarantees

Paper link: Not found on arXiv

YouTube Result:

I don't know.

Paper: Octopus: Alleviating Hallucination via Dynamic Contrastive Decoding

Paper link: https://arxiv.org/abs/2503.00361

YouTube Result:

I don't know.

Paper: ImagineFSL: Self-Supervised Pretraining Matters on Imagined Base Set for VLM-based Few-shot Learning

Paper link: Not found on arXiv

YouTube Result:

I don't know.

Paper: TIDE: Training Locally Interpretable Domain Generalization Models Enables Test-time Correction

Paper link: https://arxiv.org/abs/2411.16788

YouTube Result:

I don't know.

Paper: Multi-Label Prototype Visual Spatial Search for Weakly Supervised Semantic Segmentation

Paper link: Not found on arXiv

YouTube Result:

I don't know.

Paper: UCOD-DPL: Unsupervised Camouflaged Object Detection via Dynamic Pseudo-label Learning

Paper link: Not found on arXiv

YouTube Result:

I don't know.

Paper: SURGEON: Memory-Adaptive Fully Test-Time Adaptation via Dynamic Activation Sparsity

Paper link: https://arxiv.org/abs/2503.20354

YouTube Result:

I don't know.

Paper: ROLL: Robust Noisy Pseudo-label Learning for Multi-View Clustering with Noisy Correspondence

Paper link: Not found on arXiv

YouTube Result:

I don't know.

Paper: DAMM-Diffusion: Learning Divergence-Aware Multi-Modal Diffusion Model for Nanoparticles Distribution Prediction

Paper link: https://arxiv.org/abs/2503.09491

YouTube Result:

I don't know.

Paper: Do Computer Vision Foundation Models Learn the Low-level Characteristics of the Human Visual System?

Paper link: https://arxiv.org/abs/2502.20256

YouTube Result:

I don't know.

Paper: Context-Aware Multimodal Pretraining

Paper link: https://arxiv.org/abs/2505.03315

YouTube Result:

I don't know.

Paper: LaTexBlend: Scaling Multi-concept Customized Generation with Latent Textual Blending

Paper link: https://arxiv.org/abs/2503.06956

YouTube Result:

I don't know.

Paper: How Do I Do That? Synthesizing 3D Hand Motion and Contacts for Everyday Interactions

Paper link: https://arxiv.org/abs/2504.12284

YouTube Result:

I don't know.

Paper: Point-to-Region Loss for Semi-Supervised Point-Based Crowd Counting

Paper link: Not found on arXiv

YouTube Result:

I don't know.

Paper: BlenderGym: Benchmarking Foundational Model Systems for Graphics Editing

Paper link: https://arxiv.org/abs/2504.01786

YouTube Result:

I don't know.

Paper: Text-guided Sparse Voxel Pruning for Efficient 3D Visual Grounding

Paper link: https://arxiv.org/abs/2502.10392

YouTube Result:

I don't know.

Paper: NLPrompt: Noise-Label Prompt Learning for Vision-Language Models

Paper link: https://arxiv.org/abs/2412.01256

YouTube Result:

I don't know.

Paper: Creating Your Editable 3D Photorealistic Avatar with Tetrahedron-constrained Gaussian Splatting

Paper link: https://arxiv.org/abs/2504.20403

YouTube Result:

I don't know.

Paper: Understanding Multi-Task Activities from Single-Task Videos

Paper link: https://arxiv.org/abs/2503.18223

YouTube Result:

I don't know.

Paper: Few-shot Implicit Function Generation via Equivariance

Paper link: https://arxiv.org/abs/2501.01601

YouTube Result:

I don't know.

Paper: Efficient Motion-Aware Video MLLM

Paper link: https://arxiv.org/abs/2504.13074

YouTube Result:

I don't know.

Paper: Instant Gaussian Stream: Fast and Generalizable Streaming of Dynamic Scene Reconstruction via Gaussian Splatting

Paper link: https://arxiv.org/abs/2503.16979

YouTube Result:

I don't know.

Paper: InPO: Inversion Preference Optimization with Reparametrized DDIM for Efficient Diffusion Model Alignment

Paper link: https://arxiv.org/abs/2503.18454

YouTube Result:

I don't know.

Paper: Structure-from-Motion with a Non-Parametric Camera Model

Paper link: https://arxiv.org/abs/2309.17054

YouTube Result:

I don't know.

Paper: Galaxy Walker: Geometry-aware VLMs For Galaxy-scale Understanding

Paper link: https://arxiv.org/abs/2503.18578

YouTube Result:

I don't know.

Paper: Polarized Color Screen Matting

Paper link: Not found on arXiv

YouTube Result:

I don't know.

Paper: StyleSSP: Sampling StartPoint Enhancement for Training-free Diffusion-based Method for Style Transfer

Paper link: https://arxiv.org/abs/2501.11319

YouTube Result:

I don't know.

Paper: Detecting Backdoor Attacks in Federated Learning via Direction Alignment Inspection

Paper link: https://arxiv.org/abs/2503.07978

YouTube Result:

I don't know.

Paper: LeviTor: 3D Trajectory Oriented Image-to-Video Synthesis

Paper link: https://arxiv.org/abs/2412.15214

YouTube Result:

I don't know.

Paper: Samba: A Unified Mamba-based Framework for General Salient Object Detection

Paper link: Not found on arXiv

YouTube Result:

I don't know.

Paper: Video Depth Anything: Consistent Depth Estimation for Super-Long Videos

Paper link: https://arxiv.org/abs/2501.12375

YouTube Result:

1. **Motivation**: The motivation behind the study is to enhance the capability of depth estimation for longer videos, addressing the limitations of previous models that could only handle short video clips with lower fidelity and resolution.

2. **Novelty**: The novel aspect of the study is the ability to upload videos longer than 5 minutes and generate a comprehensive depth map for the entire video, which was not feasible with earlier models.

3. **Main Findings**: The main finding is that this new method significantly improves the process of creating depth maps for longer videos, which is crucial for various post-production tasks such as compositing and color grading.

4. **Video Title**: Big Update for utilizing AI to estimate depth on longer videos!

5. **Video Link**: [Watch here](https://www.youtube.com/watch?v=ThRGB6zSiFQ)

Paper: Distraction is All You Need for Multimodal Large Language Model Jailbreaking

Paper link: https://arxiv.org/abs/2502.10794

YouTube Result:

I don't know.

Paper: VILA-M3: Enhancing Vision-Language Models with Medical Expert Knowledge

Paper link: https://arxiv.org/abs/2411.12915

YouTube Result:

I don't know.

Paper: Empowering Vector Graphics with Consistently Arbitrary Viewing and View-dependent Visibility

Paper link: Not found on arXiv

YouTube Result:

I don't know.

Paper: SSHNet: Unsupervised Cross-modal Homography Estimation via Problem Reformulation and Split Optimization

Paper link: https://arxiv.org/abs/2409.17993

YouTube Result:

I don't know.

Paper: GenVDM: Generating Vector Displacement Maps From a Single Image

Paper link: https://arxiv.org/abs/2503.00605

YouTube Result:

1. **Motivation**: The motivation behind the study is to improve the integration of generative neural models into artistic workflows by addressing their limitations in synthesizing fine geometric details and providing the spatial and compositional controls that artists need.

2. **Novelty**: The novel aspect of the study is the introduction of GenVDM, which is the first method that generates vector displacement maps (VDMs) directly from a single image, facilitating the creation of complex geometric details in a more accessible manner.

3. **Main Findings**: The main findings indicate that GenVDM can generate VDMs from simple inputs, overcoming the challenges faced by artists in authoring VDMs manually. This advancement allows for the efficient addition of intricate details to 3D models, making it easier for artists to enhance their workflows with generative models.

4. **Video Title**: GenVDM: Generating Vector Displacement Maps From a Single Image, CVPR 2025

5. **Video Link**: [Watch the video](https://www.youtube.com/watch?v=QnLVobyZUuM)

Paper: DIV-FF: Dynamic Image-Video Feature Fields For Environment Understanding in Egocentric Videos

Paper link: https://arxiv.org/abs/2503.08344

YouTube Result:

I don't know.

Paper: DiffCAM: Data-Driven Saliency Maps by Capturing Feature Differences

Paper link: Not found on arXiv

YouTube Result:

I don't know.

Paper: Advancing Multiple Instance Learning with Continual Learning for Whole Slide Imaging

Paper link: https://arxiv.org/abs/2408.15032

YouTube Result:

I don't know.

Paper: Blurred LiDAR for Sharper 3D: Robust Handheld 3D Scanning with Diffuse LiDAR and RGB

Paper link: https://arxiv.org/abs/2411.19474

YouTube Result:

I don't know.

Paper: Cubify Anything: Scaling Indoor 3D Object Detection

Paper link: https://arxiv.org/abs/2412.04458

YouTube Result:

I don't know.

Paper: NSD-Imagery: A Benchmark Dataset for Extending fMRI Vision Decoding Methods to Mental Imagery

Paper link: Not found on arXiv

YouTube Result:

I don't know.