# Thu 2024.03.21

## GraphBEV: Towards Robust BEV Feature Alignment for Multi-Modal 3D Object Detection

*Ziying Song, Lei Yang, Shaoqing Xu, Lin Liu, Dongyang Xu, Caiyan Jia, Feiyang Jia, Li Wang*

Integrating LiDAR and camera information into Bird's-Eye-View (BEV) representation has emerged as a crucial aspect of 3D object detection in autonomous driving. However, existing methods are susceptible to the inaccurate calibration relationship between LiDAR and the camera sensor. Such inaccuracies result in errors in depth estimation for the camera branch, ultimately causing misalignment between LiDAR and camera BEV features. In this work, we propose a robust fusion framework called Graph BEV. Addressing errors caused by inaccurate point cloud projection, we introduce a Local Align module that employs neighbor-aware depth features via Graph matching. Additionally, we propose a Global Align module to rectify the misalignment between LiDAR and camera BEV features. Our Graph BEV framework achieves state-of-the-art performance, with an mAP of 70.1\%, surpassing BEV Fusion by 1.6\% on the nuscenes validation set. Importantly, our Graph BEV outperforms BEV Fusion by 8.3\% under conditions with misalignment noise.

link: http://arxiv.org/abs/2403.11848v1

## Reinforcement Learning with Latent State Inference for Autonomous On-ramp Merging under Observation Delay

*Amin Tabrizian, Zhitong Huang, Peng Wei*

This paper presents a novel approach to address the challenging problem of autonomous on-ramp merging, where a self-driving vehicle needs to seamlessly integrate into a flow of vehicles on a multi-lane highway. We introduce the Lane-keeping, Lane-changing with Latent-state Inference and Safety Controller (L3IS) agent, designed to perform the on-ramp merging task safely without comprehensive knowledge about surrounding vehicles' intents or driving styles. We also present an augmentation of this agent called AL3IS that accounts for observation delays, allowing the agent to make more robust decisions in real-world environments with vehicle-to-vehicle (V2V) communication delays. By modeling the unobservable aspects of the environment through latent states, such as other drivers' intents, our approach enhances the agent's ability to adapt to dynamic traffic conditions, optimize merging maneuvers, and ensure safe interactions with other vehicles. We demonstrate the effectiveness of our method through extensive simulations generated from real traffic data and compare its performance with existing approaches. L3IS shows a 99.90% success rate in a challenging on-ramp merging case generated from the real US Highway 101 data. We further perform a sensitivity analysis on AL3IS to evaluate its robustness against varying observation delays, which demonstrates an acceptable performance of 93.84% success rate in 1-second V2V communication delay.

link: http://arxiv.org/abs/2403.11852v2

## denoiSplit: a method for joint image splitting and unsupervised denoising

*Ashesh Ashesh, Florian Jug*

In this work we present denoiSplit, a method to tackle a new analysis task, i.e. the challenge of joint semantic image splitting and unsupervised denoising. This dual approach has important applications in fluorescence microscopy, where semantic image splitting has important applications but noise does generally hinder the downstream analysis of image content. Image splitting involves dissecting an image into its distinguishable semantic structures. We show that the current state-of-the-art method for this task struggles in the presence of image noise, inadvertently also distributing the noise across the predicted outputs. The method we present here can deal with image noise by integrating an unsupervised denoising sub-task. This integration results in improved semantic image unmixing, even in the presence of notable and realistic levels of imaging noise. A key innovation in denoiSplit is the use of specifically formulated noise models and the suitable adjustment of KL-divergence loss for the high-dimensional hierarchical latent space we are training.

We showcase the performance of denoiSplit across 4 tasks on real-world microscopy images. Additionally, we perform qualitative and quantitative evaluations and compare results to existing benchmarks, demonstrating the effectiveness of using denoiSplit: a single Variational Splitting Encoder-Decoder (VSE) Network using two suitable noise models to jointly perform semantic splitting and denoising.

link: http://arxiv.org/abs/2403.11854v1

### GPT-4 as Evaluator: Evaluating Large Language Models on Pest Management in Agriculture

*Shanglong Yang, Zhipeng Yuan, Shunbao Li, Ruoling Peng, Kang Liu, Po Yang*

In the rapidly evolving field of artificial intelligence (AI), the application of large language models (LLMs) in agriculture, particularly in pest management, remains nascent. We aimed to prove the feasibility by evaluating the content of the pest management advice generated by LLMs, including the Generative Pre-trained Transformer (GPT) series from OpenAI and the FLAN series from Google. Considering the context-specific properties of agricultural advice, automatically measuring or quantifying the quality of text generated by LLMs becomes a significant challenge. We proposed an innovative approach, using GPT-4 as an evaluator, to score the generated content on Coherence, Logical Consistency, Fluency, Relevance, Comprehensibility, and Exhaustiveness. Additionally, we integrated an expert system based on crop threshold data as a baseline to obtain scores for Factual Accuracy on whether pests found in crop fields should take management action. Each model's score was weighted by percentage to obtain a final score. The results showed that GPT-3.4 and GPT-4 outperform the FLAN models in most evaluation categories. Furthermore, the use of instruction-based prompting containing domain-specific knowledge proved the feasibility of LLMs as an effective tool in agriculture, with an accuracy rate of 72%, demonstrating LLMs' effectiveness in providing pest management suggestions.

link: http://arxiv.org/abs/2403.11858v1

### Exploring Multi-modal Neural Scene Representations With Applications on Thermal Imaging

*Mert Özer, Maximilian Weiherer, Martin Hundhausen, Bernhard Egger*

Neural Radiance Fields (NeRFs) quickly evolved as the new de-facto standard for the task of novel view synthesis when trained on a set of RGB images. In this paper, we conduct a comprehensive evaluation of neural scene representations, such as NeRFs, in the context of multi-modal learning. Specifically, we present four different strategies of how to incorporate a second modality, other than RGB, into NeRFs: (1) training from scratch independently on both modalities; (2) pre-training on RGB and fine-tuning on the second modality; (3) adding a second branch; and (4) adding a separate component to predict (color) values of the additional modality. We chose thermal imaging as second modality since it strongly differs from RGB in terms of radiosity, making it challenging to integrate into neural scene representations. For the evaluation of the proposed strategies, we captured a new publicly available multi-view dataset, ThermalMix, consisting of six common objects and about 360 RGB and thermal images in total. We employ cross-modality calibration prior to data capturing, leading to high-quality alignments between RGB and thermal images. Our findings reveal that adding a second branch to NeRF performs best for novel view synthesis on thermal images while also yielding compelling results on RGB. Finally, we also show that our analysis generalizes to other modalities, including near-infrared images and depth maps. Project page: https://mert-o.github.io/ThermalNeRF/.

link: http://arxiv.org/abs/2403.11865v1

### View-Consistent 3D Editing with Gaussian Splatting

*Yuxuan Wang, Xuanyu Yi, Zike Wu, Na Zhao, Long Chen, Hanwang Zhang*

The advent of 3D Gaussian Splatting (3DGS) has revolutionized 3D editing, offering efficient, high-fidelity rendering and enabling precise local manipulations. Currently, diffusion-based 2D editing models are harnessed to modify multi-view rendered images, which then guide the editing of

3DGS models. However, this approach faces a critical issue of multi-view inconsistency, where the guidance images exhibit significant discrepancies across views, leading to mode collapse and visual artifacts of 3DGS. To this end, we introduce View-consistent Editing (VcEdit), a novel framework that seamlessly incorporates 3DGS into image editing processes, ensuring multi-view consistency in edited guidance images and effectively mitigating mode collapse issues. VcEdit employs two innovative consistency modules: the Cross-attention Consistency Module and the Editing Consistency Module, both designed to reduce inconsistencies in edited images. By incorporating these consistency modules into an iterative pattern, VcEdit proficiently resolves the issue of multi-view inconsistency, facilitating high-quality 3DGS editing across a diverse range of scenes.

link: http://arxiv.org/abs/2403.11868v2

### Rapidly Deployable Intelligent 5G Aerial Neutral Host Networks: an O-RAN-Based Approach

*Yi Chu, David Grace, Josh Shackleton, Andy White, David Hunter, Hamed Ahmadi*

Arxiv is acting weird and throwing error: "Bad character(s) in field Abstract." for no reason. Please refer to the manuscript.

link: http://arxiv.org/abs/2403.11869v1

### IDF-CR: Iterative Diffusion Process for Divide-and-Conquer Cloud Removal in Remote-sensing Images

*Meilin Wang, Yexing Song, Pengxu Wei, Xiaoyu Xian, Yukai Shi, Liang Lin*

Deep learning technologies have demonstrated their effectiveness in removing cloud cover from optical remote-sensing images. Convolutional Neural Networks (CNNs) exert dominance in the cloud removal tasks. However, constrained by the inherent limitations of convolutional operations, CNNs can address only a modest fraction of cloud occlusion. In recent years, diffusion models have achieved state-of-the-art (SOTA) proficiency in image generation and reconstruction due to their formidable generative capabilities. Inspired by the rapid development of diffusion models, we first present an iterative diffusion process for cloud removal (IDF-CR), which exhibits a strong generative capabilities to achieve component divide-and-conquer cloud removal. IDF-CR consists of a pixel space cloud removal module (Pixel-CR) and a latent space iterative noise diffusion network (IND). Specifically, IDF-CR is divided into two-stage models that address pixel space and latent space. The two-stage model facilitates a strategic transition from preliminary cloud reduction to meticulous detail refinement. In the pixel space stage, Pixel-CR initiates the processing of cloudy images, yielding a suboptimal cloud removal prior to providing the diffusion model with prior cloud removal knowledge. In the latent space stage, the diffusion model transforms low-quality cloud removal into high-quality clean output. We refine the Stable Diffusion by implementing ControlNet. In addition, an unsupervised iterative noise refinement (INR) module is introduced for diffusion model to optimize the distribution of the predicted noise, thereby enhancing advanced detail recovery. Our model performs best with other SOTA methods, including image reconstruction and optical remote-sensing cloud removal on the optical remote-sensing datasets.

link: http://dx.doi.org/10.1109/TGRS.2024.3378720

### CO3: Low-resource Contrastive Co-training for Generative Conversational Query Rewrite

*Yifei Yuan, Chen Shi, Runze Wang, Liyi Chen, Renjun Hu, Zengming Zhang, Feijun Jiang, Wai Lam*

Generative query rewrite generates reconstructed query rewrites using the conversation history while rely heavily on gold rewrite pairs that are expensive to obtain. Recently, few-shot learning is gaining increasing popularity for this task, whereas these methods are sensitive to the inherent noise due to limited data size. Besides, both attempts face performance degradation when there exists language style shift between training and testing cases. To this end, we study low-resource generative conversational query rewrite that is robust to both noise and language style shift. The core idea is to utilize massive unlabeled data to make further improvements via a contrastive

co-training paradigm. Specifically, we co-train two dual models (namely Rewriter and Simplifier) such that each of them provides extra guidance through pseudo-labeling for enhancing the other in an iterative manner. We also leverage contrastive learning with data augmentation, which enables our model pay more attention on the truly valuable information than the noise. Extensive experiments demonstrate the superiority of our model under both few-shot and zero-shot scenarios. We also verify the better generalization ability of our model when encountering language style shift.

link: http://arxiv.org/abs/2403.11873v1

### Benchmarking Analytical Query Processing in Intel SGXv2

*Adrian Lutsch, Muhammad El-Hindi, Matthias Heinrich, Daniel Ritter, Zsolt István, Carsten Binnig*

The recently introduced second generation of Intel SGX (SGXv2) lifts memory size limitations of the first generation. Theoretically, this promises to enable secure and highly efficient analytical DBMSs in the cloud. To validate this promise, in this paper, we conduct the first in-depth evaluation study of running analytical query processing algorithms inside SGXv2. Our study reveals that state-of-the-art query operators like radix joins and SIMD-based scans can indeed achieve high performance inside SGXv2 enclaves. These operations are orders of magnitude faster than joins optimized for the discontinued SGXv1 hardware. However, substantial performance overheads are still caused by subtle hardware and software differences influencing code execution inside an SGX enclave. We investigate these differences and propose new optimizations to bring the performance inside the enclave on par with native code execution outside an enclave.

link: http://arxiv.org/abs/2403.11874v1

### Towards Real-Time Fast Unmanned Aerial Vehicle Detection Using Dynamic Vision Sensors

*Jakub Mandula, Jonas Kühne, Luca Pascarella, Michele Magno*

Unmanned Aerial Vehicles (UAVs) are gaining popularity in civil and military applications. However, uncontrolled access to restricted areas threatens privacy and security. Thus, prevention and detection of UAVs are pivotal to guarantee confidentiality and safety. Although active scanning, mainly based on radars, is one of the most accurate technologies, it can be expensive and less versatile than passive inspections, e.g., object recognition. Dynamic vision sensors (DVS) are bio-inspired event-based vision models that leverage timestamped pixel-level brightness changes in fast-moving scenes that adapt well to low-latency object detection. This paper presents F-UAV-D (Fast Unmanned Aerial Vehicle Detector), an embedded system that enables fast-moving drone detection. In particular, we propose a setup to exploit DVS as an alternative to RGB cameras in a real-time and low-power configuration. Our approach leverages the high-dynamic range (HDR) and background suppression of DVS and, when trained with various fast-moving drones, outperforms RGB input in suboptimal ambient conditions such as low illumination and fast-moving scenes. Our results show that F-UAV-D can (i) detect drones by using less than <15 W on average and (ii) perform real-time inference (i.e., <50 ms) by leveraging the CPU and GPU nodes of our edge computer.

link: http://arxiv.org/abs/2403.11875v1

### Deep Bayesian Future Fusion for Self-Supervised, High-Resolution, Off-Road Mapping

*Shubhra Aich, Wenshan Wang, Parv Maheshwari, Matthew Sivaprakasam, Samuel Triest, Cherie Ho, Jason M. Gregory, John G. Rogers III, Sebastian Scherer*

The limited sensing resolution of resource-constrained off-road vehicles poses significant challenges towards reliable off-road autonomy. To overcome this limitation, we propose a general framework based on fusing the future information (i.e. future fusion) for self-supervision. Recent approaches exploit this future information alongside the hand-crafted heuristics to directly supervise the targeted downstream tasks (e.g. traversability estimation). However, in this paper, we opt for a more general line of development - time-efficient completion of the highest resolution (i.e. 2cm per pixel) BEV map in a self-supervised manner via future fusion, which can be used for any

downstream tasks for better longer range prediction. To this end, first, we create a high-resolution future-fusion dataset containing pairs of (RGB / height) raw sparse and noisy inputs and map-based dense labels. Next, to accommodate the noise and sparsity of the sensory information, especially in the distal regions, we design an efficient realization of the Bayes filter onto the vanilla convolutional network via the recurrent mechanism. Equipped with the ideas from SOTA generative models, our Bayesian structure effectively predicts high-quality BEV maps in the distal regions. Extensive evaluation on both the quality of completion and downstream task on our future-fusion dataset demonstrates the potential of our approach.

link: http://arxiv.org/abs/2403.11876v1

### Efficient Training of Learning-Based Thermal Power Flow for 4th Generation District Heating Grids

*Andreas Bott, Mario Beykirch, Florian Steinke*

Thermal power flow (TPF) is an important task for various control purposes in 4 Th generation district heating grids with multiple decentral heat sources and meshed grid structures. Computing the TPF, i.e., determining the grid state consisting of temperatures, pressures, and mass flows for given supply and demand values, is classically done by solving the nonlinear heat grid equations, but can be sped up by orders of magnitude using learned models such as neural networks. We propose a novel, efficient scheme to generate a sufficiently large training data set covering relevant supply and demand values. Instead of sampling supply and demand values, our approach generates training examples from a proxy distribution over generator and consumer mass flows, omitting the iterations needed for solving the heat grid equations. The exact, but slightly different, training examples can be weighted to represent the original training distribution. We show with simulations for typical grid structures that the new approach can reduce training set generation times by two orders of magnitude compared to sampling supply and demand values directly, without loss of relevance for the training samples. Moreover, learning TPF with a training data set is shown to outperform sample-free, physics-aware training approaches significantly.

link: http://arxiv.org/abs/2403.11877v1

### InTeX: Interactive Text-to-texture Synthesis via Unified Depth-aware Inpainting

*Jiaxiang Tang, Ruijie Lu, Xiaokang Chen, Xiang Wen, Gang Zeng, Ziwei Liu*

Text-to-texture synthesis has become a new frontier in 3D content creation thanks to the recent advances in text-to-image models. Existing methods primarily adopt a combination of pretrained depth-aware diffusion and inpainting models, yet they exhibit shortcomings such as 3D inconsistency and limited controllability. To address these challenges, we introduce InteX, a novel framework for interactive text-to-texture synthesis. 1) InteX includes a user-friendly interface that facilitates interaction and control throughout the synthesis process, enabling region-specific repainting and precise texture editing. 2) Additionally, we develop a unified depth-aware inpainting model that integrates depth information with inpainting cues, effectively mitigating 3D inconsistencies and improving generation speed. Through extensive experiments, our framework has proven to be both practical and effective in text-to-texture synthesis, paving the way for high-quality 3D content creation.

link: http://arxiv.org/abs/2403.11878v1

### Unimodal Multi-Task Fusion for Emotional Mimicry Prediciton

*Tobias Hallmen, Fabian Deuser, Norbert Oswald, Elisabeth André*

In this study, we propose a methodology for the Emotional Mimicry Intensity (EMI) Estimation task within the context of the 6th Workshop and Competition on Affective Behavior Analysis in-the-wild. Our approach leverages the Wav2Vec 2.0 framework, pre-trained on a comprehensive podcast dataset, to extract a broad range of audio features encompassing both linguistic and paralinguistic elements. We enhance feature representation through a fusion technique that integrates individual features with a global mean vector, introducing global contextual insights into our analysis. Additionally, we incorporate a pre-trained valence-arousal-dominance (VAD) module from the

Wav2Vec 2.0 model. Our fusion employs a Long Short-Term Memory (LSTM) architecture for efficient temporal analysis of audio data. Utilizing only the provided audio data, our approach demonstrates significant improvements over the established baseline.

link: http://arxiv.org/abs/2403.11879v1