Research Projects

Chi Keung Tang

FDNeRF: Semantics-Driven Face Reconstruction, Prompt Editing and Relighting with Diffusion Models. The ability to create high-quality 3D faces from a single image has become increasingly important with wide applications in video conferencing, AR/VR, and advanced video editing in movie industries. In this paper, we propose Face Diffusion NeRF (FDNeRF), a new generative method to reconstruct high-quality Face NeRFs from single images, complete with semantic editing and relighting capabilities. FDNeRF utilizes high-resolution 3D GAN inversion and expertly trained 2D latent-diffusion model, allowing users to manipulate and construct Face NeRFs in zero-shot learning without the need for explicit 3D data. With carefully designed illumination and identity preserving loss, as well as multi-modal pre-training, FDNeRF offers users unparalleled control over the editing process enabling them to create and edit face NeRFs using just single-view images, text prompts, and explicit target lighting. The advanced features of FDNeRF have been designed to produce more impressive results than existing 2D editing approaches that rely on 2D segmentation maps for editable attributes. Experiments show that our FDNeRF achieves exceptionally realistic results and unprecedented flexibility in editing compared with state-of-the-art 3D face reconstruction and editing methods.

Distill Gold from Massive Ores: Efficient Dataset Distillation via Critical Samples Selection. Data-efficient learning has drawn significant attention, especially given the current trend of large multi-modal models, where dataset distillation can be an effective solution. However, the dataset distillation process itself is still very inefficient. In this work, we model the distillation problem with reference to information theory. Observing that severe data redundancy exists in dataset distillation, we argue to put more emphasis on the utility of the training samples. We propose a family of methods to exploit the most valuable samples, which is validated by our comprehensive analysis of the optimal data selection. The new strategy significantly reduces the training cost and extends a variety of existing distillation algorithms to larger and more diversified datasets, e.g., in some cases only **0.04%** training data is sufficient for comparable distillation performance. Moreover, our strategy consistently enhances the performance, which may open up new analyses on the dynamics of distillation and networks. Our method is able to extend the distillation algorithms to much larger-scale datasets and more heterogeneous datasets, e.g., ImageNet-1K and Kinetics-400. Our code will be made publicly available.

Segment Anything in High Quality. The recent Segment Anything Model (SAM) represents a big leap in scaling up segmentation models, allowing for powerful zero-shot capabilities and flexible prompting. Despite being trained with 1.1 billion masks, SAM's mask prediction quality falls short in many cases, particularly when dealing with objects that have intricate structures. We propose HQ-SAM, equipping SAM with the ability to accurately segment any object, while maintaining SAM's original promptable design, efficiency, and zero-shot generalizability. Our careful design reuses and preserves the pre-trained model weights of SAM, while only introducing minimal additional parameters and computation. We design a learnable High-Quality Output Token, which is injected into SAM's mask decoder and is responsible for predicting the high-quality mask. Instead of only applying it on mask-decoder features, we first fuse them with early and final ViT features for improved mask details. To train our introduced learnable parameters, we compose a dataset of 44K fine-grained masks from several sources. HQ-SAM is only trained on the introduced detaset of 44k masks, which takes only 4 hours on 8 GPUs. We show the efficacy of HQ-SAM in a suite of 9 diverse segmentation datasets across different downstream tasks, where 7 out of them are evaluated in a zero-shot transfer protocol. Our code and models will be released.

BiMatting: Efficient Video Matting via Binarization. Real-time video matting on edge devices faces significant computational resource constraints, limiting the widespread use of video matting in applications such

as online conferences and short-form video production. Binarization is a powerful compression approach that greatly reduces computation and memory consumption by using 1-bit parameters and bitwise operations. However, binarization of the video matting model is not a straightforward process, and our empirical analysis has revealed two primary bottlenecks: severe representation degradation of the encoder and massive redundant computations of the decoder. To address these issues, we propose **BiMatting**, an accurate and efficient video matting model using binarization. Specifically, we construct shrinkable and dense topologies of the binarized encoder block to enhance the extracted representation. We sparsify the binarized units to reduce the low-information decoding computation. Through extensive experiments, we demonstrate that BiMatting outperforms other binarized video matting models, including state-of-the-art (SOTA) binarization methods, by a significant margin. Our approach even performs comparably to the full-precision counterpart in visual quality. Furthermore, BiMatting achieves remarkable savings of 12.4× and 21.6× in computation and storage, respectively, showcasing its potential and advantages in real-world resource-constrained scenarios.

UniBoost: Unsupervised Unimodal Pre-training for Boosting Zero-shot Vision-Language Tasks. Large-scale joint training of multimodal models, e.g., CLIP, have demonstrated great performance in many vision-language tasks. However, image-text pairs for pre-training are restricted to the intersection of images and texts, limiting their ability to cover a large distribution of real-world data, where noise can also be introduced as misaligned pairs during pre-processing. Conversely, unimodal models trained on text or image data alone through unsupervised techniques can achieve broader coverage of diverse real-world data and are not constrained by the requirement of simultaneous presence of image and text. In this paper, we demonstrate that using large-scale unsupervised unimodal models as pre-training can enhance the zero-shot performance of image-text pair models. Our thorough studies validate that models pre-trained as such can learn rich representations of both modalities, improving their ability to understand how images and text relate to each other. Our experiments show that unimodal pre-training outperforms state-of-the-art CLIP-based models by 6.5% (52.3% $\rightarrow$ 58.8%) on PASCAL-5^i and 6.2% (27.2% $\rightarrow$ 33.4%) on COCO-20^i semantic segmentation under zero-shot setting respectively. By learning representations of both modalities, unimodal pre-training offers broader coverage, reduced misalignment errors, and the ability to capture more complex features and patterns in the real-world data resulting in better performance especially for zero-shot vision-language tasks.

Deceptive-NeRF: Enhancing NeRF Reconstruction using Pseudo-Observations from Diffusion Models. This paper introduces Deceptive-NeRF, a new method for enhancing the quality of reconstructed NeRF models using synthetically generated pseudo-observations, capable of handling sparse input and removing floater artifacts. Our proposed method involves three key steps: 1) reconstruct a coarse NeRF model from sparse inputs; 2) generate pseudo-observations based on the coarse model; 3) refine the NeRF model using pseudo-observations to produce a high-quality reconstruction. To generate photo-realistic pseudo-observations that faithfully preserve the identity of the reconstructed scene while remaining consistent with the sparse inputs, we develop a rectification latent diffusion model that generates images conditional on a coarse RGB image and depth map, which are derived from the coarse NeRF and latent text embedding from input images. Extensive experiments show that our method is effective and can generate perceptually high-quality NeRF even with very sparse inputs.