DATA SCIENCE INTERVIEW PREPARATION

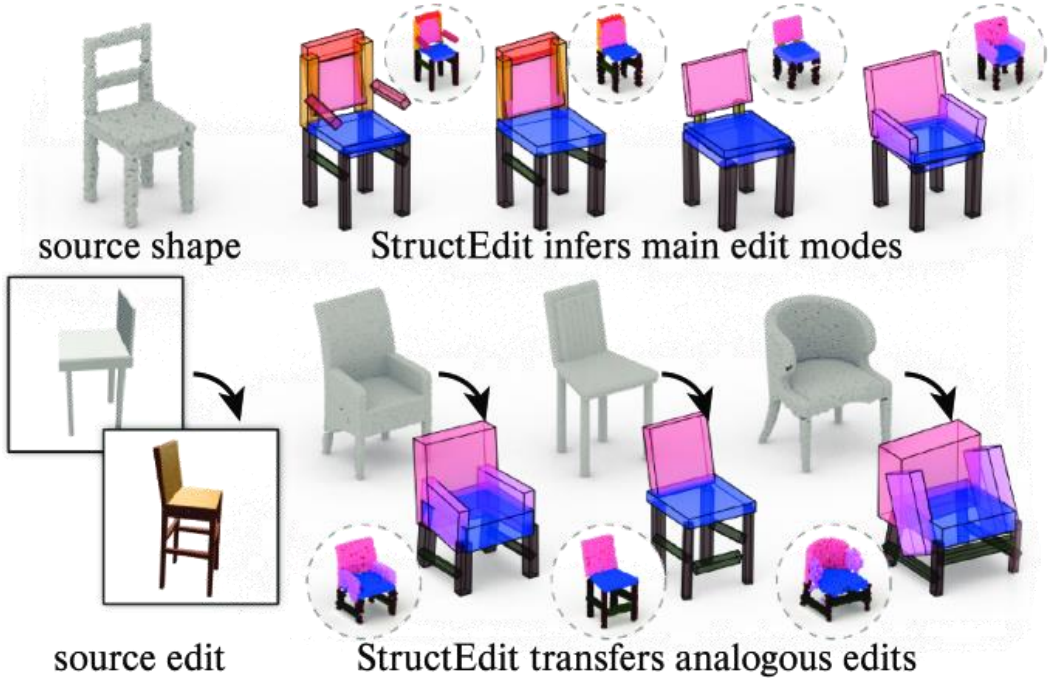
(30 Days of Interview Preparation)

Day28

Q1. Explain StructEdit(Learning Structural Shape Variations).

Answer:

The shapes of the 3D objects exhibit remarkable diversity, both in their compositional structure in terms of parts, as well as in geometry of the elements themselves. Yet human are remarkably skilled at imagining meaningful shape variation even from the isolated object instances. For example, having seen a new chair, we can easily imagine its *natural* changes with the different height back, a wider seat, with or without armrests, or with a diverse base. In this article, we investigate how to learn such shape variations directly from the 3D data. Specifically, given the shape collection, we are interested in two sub-problems: first, for any given shape, we want to discover main modes of edits, which can be inferred directly from shape collection; and second, given an example edit on one shape, we want to transfer edit to another shape in the group, as a form of analogy-based edit transfer. This ability is useful in several settings, including the design of individual 3D models, the consistent modification of the 3D model families, and the fitting of CAD models to noisy and incomplete 3D scans.



Above Fig: **Edit generation and transfer with StructEdit.**

We present the StructEdit, a method that learns the distribution of *shape differences* between structured objects that can be used to generate an ample variety of edits (in a first row); and accurately transfer edits between different purposes and across different modalities (on the second row). Edits can be both geometric and topological.

There are many challenges in capturing space of shape variations. First, individual shape can have different representations as image, surface meshes, or point clouds; second, one needs the unified setting for representing both continuous deformations as well as structural changes; third, shape edits are not directly expressed but are only implicitly contained in shape collections; and finally, learning the space of structural variations that is applicable to more than the single shape amounts to learning mappings between different shape edit distributions, since different shapes have various types and numbers of parts (like tables with or without leg bars).

In much of the existing literature on 3D machine learning(ML), 3D shapes are mapped to points in the representation space whose coordinates encode latent features of each shape. In such representation, shape edits are encoded as vectors in that same space – in other words, as differences between points representing shapes. Equivalently, we can think of forms as "anchored" vectors rooted at origin, while shape differences are "floating" vectors that can be transported around in shape space. This type of vector space arithmetic is commonly used [wu2016learning, achlioptas2017learning, wang2018global, gao2018automatic, xia2015realtime, Villegas_2018_CVPR], for example, in performing analogies, where the vector that is the difference of possible point A from point B is added to point C to produce an analogous point D. The challenge with this view in our setting is that while Euclidean spaces are perfectly homogeneous and vectors can be comfortably transported and added to points anywhere, shape spaces are far or less so. While for continuous variations, a vector space model has some plausibility, this is not so for structural variations: the "add arms" vector does not make sense for the point representing a chair that already has arms. We take the different approach. We consider embedding shapes differences or deltas *directly in their own latent space*, separate from general shape embedding space. Encoding and decoding such shape differences is always done through a VAE(variational autoencoder), in the context of the given source shape, itself encoded through the part hierarchy. This has the number of key advantages: (i) allows compact encodings of shape deltas, since in general, we aim to describe local variation; (ii) encourages network to abstract commonalities in shape variations across shape space; and (iii) adapts the edit to the provided source shape, suppressing the mode that are semantically implausible.

We have extensively evaluated the *StructEdit* on publicly available shape data sets. We introduce the new synthetic dataset with ground truth shape edits to quantitatively evaluate our method and compare it against baseline alternative. We then provide evaluation results on PartNet dataset [mo2019partnet] and provide ablation studies. Finally, we demonstrates that extension of our method allows the handling of both images and point cloud as shape sources, can predict plausible edit modes from single shape examples, and can also transfer example shape edit on one shape to other shapes in the collection.

Q2. EmpGAN: Multi-resolution Interactive Empathetic Dialogue Generation

Answer:

As a vital part of human intelligence, emotional perceptivity is playing elemental role in various social communication scenarios, such as., education and healthcare systems. Recently, sensitive conversation generation has received an increasing amount of attention to address emotion factors in an end-to-end framework. However, as li2018syntactically revealed, that conventional emotional conversation system aims to produce more emotion-rich responses according to the specific user-input emotion, which inevitably leads to the psychological inconsistency problem.

Studies on social psychology suggest that empathy is the crucial step towards a more humanized human-machine conversation, which improves emotional perceptivity in emotion-bonding social activities. To design the intelligent automatic dialogue system, it is essential to make a chatbot empathetic within dialogues. Therefore, in this paper, we focus on a task of *empathetic dialogue generation*, which automatically tracks and understands the user's emotion at each turn in multi-turn dialogue scenarios.

Despite the achieved successes, obstacles to establishing the empathetic conversational system are still far beyond current signs of progress:

- Merely considering the sentence-level emotion while neglecting more precise token-level feelings may lead to insufficient emotion perceptivity. It is challenging to capture nuances of human emotion accurately without modeling multi-granularity emotion factors in the dialogue generation.
- Merely relying on the dialogue history but overlooking the potential of user feedback for the generated responses further aggravates the deficiencies above, which causes undesirable reactions.

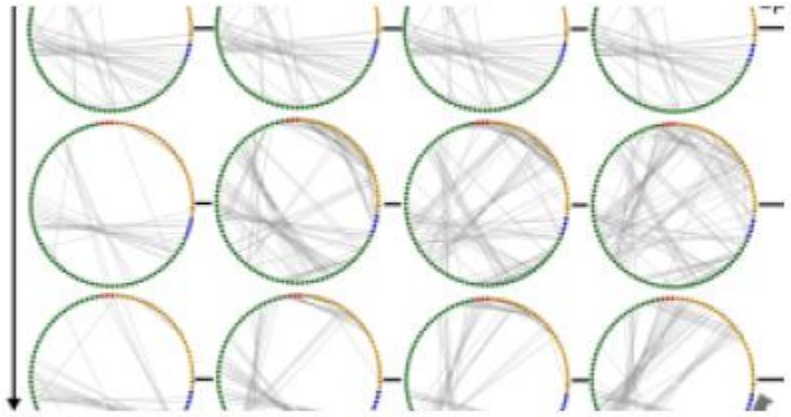
In this paper, we propose the multi-resolution adversarial empathetic dialogue generation model, named EmpGAN, to address the above challenges through generating more empathetic and appropriate responses. To capture nuances of user feelings sufficiently, EmpGAN make responses by taking both coarse-grained sentence-level and fine-grained token-level emotions into account. The response generator in EmpGAN dynamically understands sentiments along with a conversation to perceive a user's emotion states in multi-turn conversations. Furthermore, an interactive adversarial learning framework is augmented to take user feedback into account thoughtfully, where two interactive discriminators identify whether the generated responses evoke emotion perceptivity regarding both the dialogue history and user emotions.

In particular, the EmpGAN contains the empathetic generator and two interactive inverse discriminators. The empathetic generator is composed of three components: (i) A semantic understanding module based on Seq2Seq(sequence to sequence) neural networks that maintain the multi-turn semantic context. (ii) A multi-resolution emotion perception model captures the fine and coarse-grained emotion factors of each dialogue turn to build the emotional framework. (iii) An empathetic response decoder combines semantic and emotional context to produce appropriate responses in terms of both meaning and emotion. The two interactive inverse discriminator additionally incorporate the user feedback and corresponding emotional feedback as inverse supervised signal to induce the generator to produce a more empathetic response.

Q3. G-TAD: Sub-Graph Localization for Temporal Action Detection

Answer:

Video understanding has gained much attention from both academia and industry over recent years, given the rapid growth of videos published in online platforms. Temporal action detection is one of exciting but challenging tasks in this area. It involves detecting start and the end frames of action instances, as well as predicting their class label. This is onerous, especially in long untrimmed videos.



Video context is an important cue to detect actions effectively. Here, we refer to mean as frames that are outside the target action but carry valuable indicative information of it. Using video context to infer potential actions is natural for human beings. Empirical evidence shows that human can reliably predict or guess the occurrence of the specific type of work by only looking at short video snippets where the action does not happen. Therefore, incorporating context into temporal action detection has become important strategy to boost detection accuracy in the recent literature. Researchers have proposed various ways to take advantage of the video context, such as extending temporal action boundaries by the pre-

defined ratio, using dilated convolution to encode meaning into features, and aggregating definition feature implicitly by way of the Gaussian curve. All these methods only utilize temporal context, which follows or precedes an action instance in its immediate secular neighborhood. However, real-world videos vary dramatically in temporal extent, action content, and even editing preferences. The use of such temporal contexts does not fully exploit precious merits of the video context, and it may also impair detection accuracy if not adequately designed for underlying videos.

So, what properties characterize the desirable video context for accurate action detection? First, setting should be semantically or grammatically correlated to the target action other than merely temporally located in its vicinity. Imagine a case where we manually stitch an action clip into some irrelevant frames; the abrupt scene change surrounding the action would not benefit the action detection. On the other hand, snippets located at a distance from an operation but containing similar semantic content might provide indicative hints for detecting the action. Second, context should be content-adaptive rather than manually pre-defined. Considering the vast variation of videos, a framework that helps to identify different action instances could be changed in lengths and locations based on the video content. Third, context should be based on multiple semantic levels, since using only one form/level of meaning is unlikely to generalize well.

In this paper, we endow video context with all the above properties by casting action detection as a sub-graph localization problem based on a graph convolutional network (GCN). We represent each video sequence as the graph, each snippet as a node, each snippet-snippet correlation as an edge, and target actions associated with context as sub-graphs, as shown in Fig. 1. The meaning of a snippet is considered to be all snippets connected to it by an edge in a video graph. We define two types of edges — temporal corners and semantic edges, each corresponding to temporal context and grammatical context, respectively. Temporal edges exist between each pair of neighboring snippets, whereas semantic edges are dynamically learned from the video features at each GCN layer. Hence, the multi-level context of each snippet is gradually aggregated into the features of the snippet throughout the entire GCN. ResNeXt inspires the structure of each GCN block, so we name this GCN-based feature extractor GCNeXt.

The pipeline of our proposed Graph-Temporal Action Detection method, dubbed G-TAD, is analogous to faster R-CNN in object detection. There are two critical designs in G-TAD. First, GCNeXt, which generates context-enriched features, corresponds to the backbone network, analogous to a series of Convolutional Neural Network (CNN) layers in faster R-CNN. Second, to mimic RoI(region of interest) alignment in faster R-CNN, we design the sub-graph alignment (SGAlign) layer to generate a fixed-size representation for each sub-graph and embed all sub-graphs into same Euclidean space. Finally, we apply a classifier on the features of each sub-graph to obtain detection results. We summarize our contributions as follows.

(1) We present a novel GCN-based video model to exploit video context for effective temporal action detection fully. Using this video GCN representation, we can adaptively incorporate multi-level semantic meaning into the features of each snippet.

(2) We propose G-TAD, a new sub-graph detection framework, to localize actions in video graphs. G-TAD includes two main modules: GCNeXt and SGAlign. GCNeXt performs graph convolutions on video graphs, leveraging both temporal and semantic context. SGAlign re-arranges sub-graph features in the embedded space suitable for detection.

(3) G-TAD achieves state-of-the-art(SOTA) performance on two popular action detection benchmarks. On ActityNet-1.3, it achieves an average mAP of 34.09%. On THUMOS-14, it reaches 40.16%mAP@0.5, beating all contemporary one-stage methods.

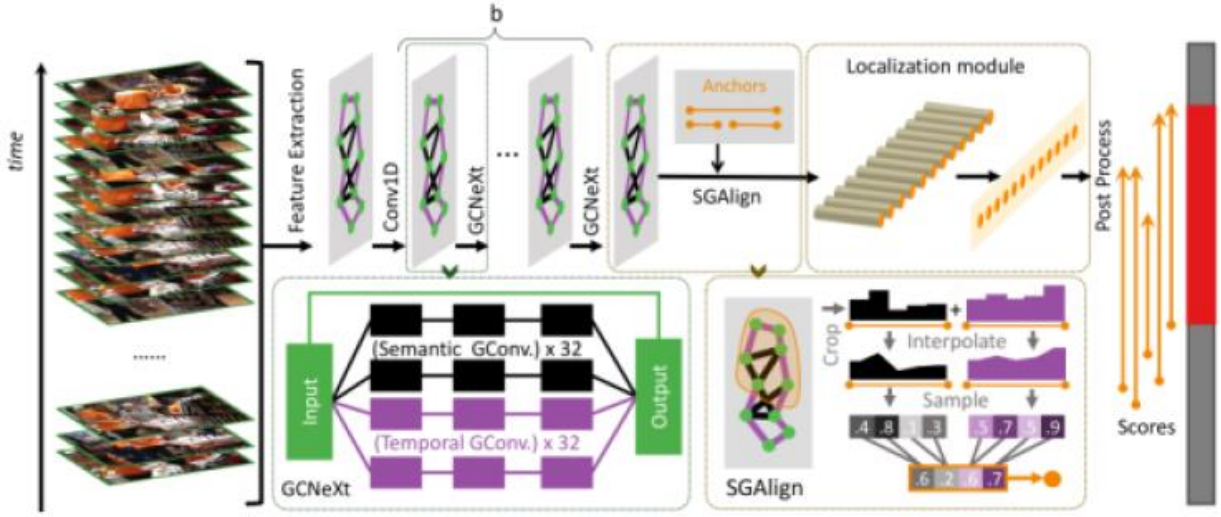
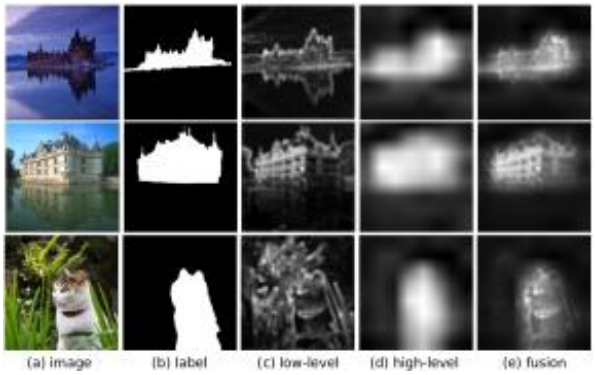


Fig: **Overview of G-TAD architecture.** The input of G-TAD is the sequence of snippet features. We first extract features using b=3 GCNeXt blocks, which gradually aggregate both temporal and multi-level semantic context. Semantic context, encoded in semantic edges, is dynamically learned from elements at each GCNeXt layer. Then we feed extracted features into the SGAlign layer, where sub-graphs defined by the set of anchors are transformed to a fixed-size representation in the Euclidean space. Finally, the localization module scores and ranks the sub-graphs for detection.

Q4. What is F3Net?

Answer:



F3Net is a combination of Fusion, Feedback, and Focus for Salient object detection (SOD) aims to estimate the significant visual regions of images or videos and often serves as the pre-processing step for many downstream vision tasks. Earlier SOD algorithms mainly rely on heuristic priors (*e.g.,* color, texture and contrast) to generate saliency maps. However, these hand-craft features can hardly capture high-level semantic relations and context information. Thus they are not robust enough to complex scenarios. Recently, convolutional neural networks (CNNs) have demonstrated its powerful feature extraction capability in visual feature representation. Many CNNs-based models have achieved remarkable progress and pushed the performance of SOD to a new level. These models adopt the encoder-decoder architecture, which is simple in structure and computationally efficient. The encoder usually is made up of a pre-trained classification model (e.g., ResNet and VGG), which can extract multiple features of different semantic levels and resolutions. In the decoder, extracted features are combined to generate saliency maps.

However, there remain two significant challenges in accurate SOD. First, features of different levels have different distribution characteristics. High-level features have rich semantics but lack precise location information. Low-level features have rich details but full of background noises. To generate better saliency maps, multi-level features are combined. However, without delicate control of the information flow in the model, some redundant features, including noises from low-level layers and coarse boundaries from high-level layers, will pass in and possibly result in performance degradation. Second, most of the existing models use binary cross-entropy that treats all pixels equally. Intuitively, different pixels deserve different weights, *e.g.*, pixels at the boundary are more discriminative and should be attached with more importance. Various boundary losses have been proposed to enhance the boundary detection accuracy, but considering only the boundary pixels is not comprehensive enough since there are lots of pixels near the boundaries prone to wrong predictions. These pixels are also essential and should be assigned with

larger weights. In consequence, it is essential to design a mechanism to reduce the impact of inconsistency between features of different levels and assign larger weights to those significant pixels.

To address the above challenges, we proposed a novel SOD framework, named F3Net, which achieves remarkable performance in producing high-quality saliency maps. First, to mitigate the discrepancy between features, we design a cross-feature module (CFM), which fuses elements of different levels by element-wise multiplication. Different from addition and concatenation, CFM takes a selective fusion strategy, where redundant information will be suppressed to avoid the contamination between features, and important features will complement each other. Compared with traditional fusion methods, CFM can remove background noises and sharpen boundaries, as shown in Fig. 1. Second, due to downsampling, high-level features may suffer from information loss and distortion, which can not be solved by CFM. Therefore, we develop the cascaded feedback decoder (CFD) to refine these features iteratively. CFD contains multiple sub-decoders, each of which includes both bottom-up and top-down processes. For the bottom-up method, multi-level features are aggregated by CFM gradually. For the top-down process, aggregated features are feedback into previous features to refine them. Third, we propose the pixel position-aware loss (PPA) to improve the commonly used binary cross-entropy loss, which treats all pixels equally. Pixels located at boundaries or elongated areas are more complicated and discriminating. Paying more attention to these hard pixels can further enhance model generalization. PPA loss assigns different weights to different pixels, which extends binary cross-entropy. The weight of each pixel is determined by its surrounding pixels. Hard pixels will get larger weights, and easy pixels will get smaller ones.

To demonstrate the performance of F3Net, we report experimental results on five popular SOD datasets and visualize some saliency maps. We conduct a series of ablation studies to evaluate the effect of each module. Quantitative indicators and visual results show that F3Net can obtain significantly better local details and improved saliency maps. Codes have been released. In short, our main contributions can be summarized as follows:

- We introduce the cross feature module to fuse features of different levels, which can extract the shared parts between features and suppress each other's background noises and complement each other's missing parts.
- We propose the cascaded feedback decoder for SOD, which can feedback features of both high resolutions and high semantics to previous ones to correct and refine them for better saliency maps generation.
- We design pixel position-aware loss to assign different weights to different positions. It can better mine the structure information contained in the features and help the network focus more on detail regions.

- Experimental results demonstrate that the proposed model F3Net achieves the state-of-the-art performance on five datasets in terms of six metrics, which proves the effectiveness and superiority of the proposed method.

Q5.Natural Language Generation using Reinforcement Learning with External Rewards

Answer:

We aim to develop models that are capable of generating language across several genres of text, conversational texts, and restaurant reviews. After all, humans are adept at both. Extant NLG(natural language generation) models work on either conversational text (like movie dialogues) or longer text (e.g., stories, reviews) but not both. Also, while the state-of-the-art(SOTA) in this field has advanced quite rapidly, current model is prone to generate language that is short, dull, off-context. More importantly, a generated language may not adequately reflect affective content of the input. Indeed, humans are already adept at this task, as well. To address these research challenges, we propose the RNN-LSTM architecture that uses an encoder-decoder network. We also use reinforcement learning(RL) that incorporates internal and external rewards. Specifically, we use emotional appropriateness as an internal reward for the NLG(Natural Language Generation) system – so that the emotional tone of generated language is consistent with the emotional tone of prior context fed as input to model. We also effectively incorporate usefulness scores as external rewards in our model. Our main contribution is the use of distantly labeled data in architecture that generates coherent, affective content and we test the architecture across two different genres of text.

What are the problem statement and their intuition?

Our aim is to take advantage of reinforcement learning(RL) and external rewards during the process of language generation. Complementary to this goal, we also aim to generate language that has same emotional tone as the other input. Emotions are recognized as functional in decision-making by influencing motivation and action selection. However, the external feedback and rewards are hard to come by for language generation; these would need to be provided through crowdsourcing judgment on generated responses *during* generation process, which makes process time-consuming and impractical. To overcome this problem, we look for distance labeling and use labels provided in training set as a proxy for human judgment on generated responses. Particularly, we incorporate usefulness scores in a restaurant review corpus as the proxy for external feedback.

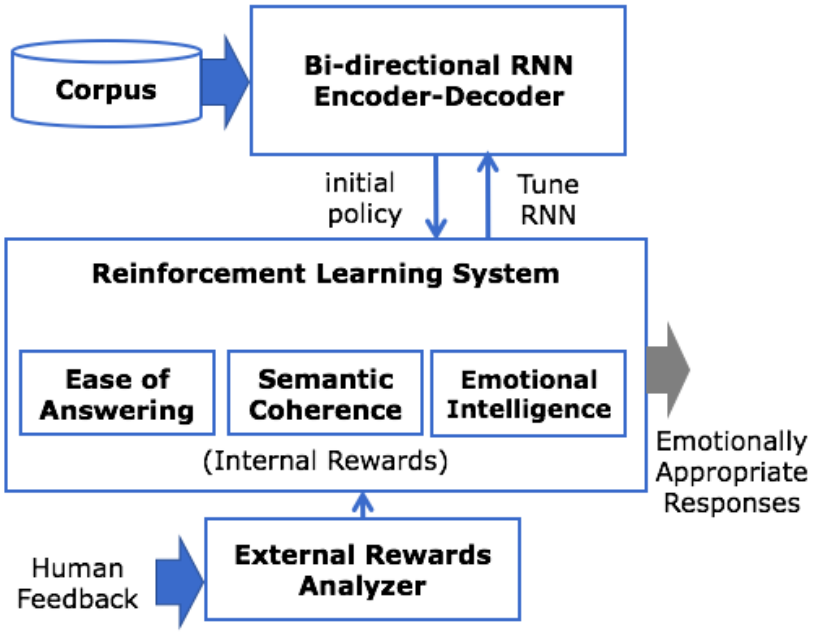


Fig. 1: Overall Architecture of the system showing internal and external rewards using reinforcement learning

Q6. LaFIn: Generative Landmark Guided Face Inpainting

Answer:

Image inpainting (*a.k.a.* image completion) refers to the process of reconstructing lost or deteriorated regions of images, which can be applied to, as a fundamental component, various tasks such as image restoration and editing. Undoubtedly, one expects the completed result to be realistic so that the reconstructed regions can be hardly perceived. Compared with natural scenes like oceans and lawns, manipulating faces, the focus of this work, is more challenging. Because the faces have much stronger topological structure and attribute consistency to preserve. Figure 1 shows three such examples. Very often, given the observed clues, human beings can easily infer what the lost parts possibly, although inexactly, look like. As a consequence, a slight violation of the topological structure and the attribute consistency in the reconstructed face highly likely leads to a significant perceptual flaw. The following defines the problem:

Definition:

Face Inpainting. Given a face image, I *with corrupted regions masked by* M. *Let* $\overline{\mathbf{M}}$ *designate the complement of* M *and* $\circ$ *the Hadamard product. The goal is to fill the target part with semantically meaningful and visually continuous information to the observed part. In other words, the completed*

result ^I:=M∘^I+¯¯¯M∘I *should preserve the topological structure among face components such as eyes, nose, and mouth, and the attribute consistency on like pose gender, ethnicity, and expression.*

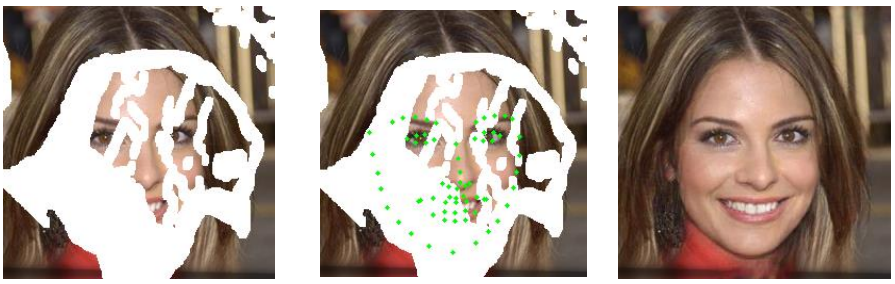


Figure 1: Three face completion results by our method. From left to right: corrupted inputs, plus landmarks predicted from the inputs, and our final results, respectively.

Q7. Image2StyleGAN++: How to Edit the Embedded Images?

Answer:

(i) (ii) (iii) (iv)

From above fig: (i) and (ii): input images; (iii): the "two-face" generated by naively copying the left half from (i) and the right half from (ii); (iv): the "two-face" created by our Image2StyleGAN++ framework.

Recent GANs demonstrated that synthetic images could be generated with very high quality. This motivates research into embedding algorithms that embed a given photograph into a GAN latent space. Such embedding algorithms can be used to analyze the limitations of GANs, do image inpainting, local image editing, global image transformations such as image morphing and expression transfer, and few-shot video generation.

In this paper, we propose to extend a very recent embedding algorithm, Image2StyleGAN. In particular, we would like to improve this previous algorithm in three aspects. First, we noticed that the embedding quality could be further improved by including Noise space optimization into embedding framework. The key insight here is that stable Noise space optimization can only be conducted if optimization is done sequentially with W+ space and not jointly. Second, we would like to improve capabilities of the embedding algorithm to increase the local control over the embedding. One way to improve local authority is to include mask in embedding algorithm with undefined content. The goal of the embedding algorithm should be to find a plausible embedding for everything outside the mask, while filling in reasonable semantic content in the masked pixels.

Similarly, we would like to provide the option of approximate embeddings, where the specified pixel colors are only a guide for the embedding. In this way, we aim to achieve high-quality embeddings that can be controlled by user scribbles. In the third technical part of the paper, we investigate the combination of embedding algorithm and direct manipulations of the activation maps (called activation tensors in our article).

Q8. oops! Predicting Unintentional Action in Video

Answer:

From just a glance at the video, we can often tell whether a person's action is intentional or not. For example, the Below figure shows a person attempting to jump off a raft, but unintentionally tripping into the sea. In a classic series of papers, developmental psychologist Amanda Woodward demonstrated that children learn this ability to recognize the intentionality of action during their first year. However, predicting the intention behind action has remained elusive for machine vision. Recent advances in action recognition have primarily focused on predicting the physical motions and atomic operations in the video, which captures the means of action but not the intent of action.

We believe a key limitation for perceiving visual intentionality has been the lack of realistic data with natural variation of intention. Although there are now extensive video datasets for action recognition, people are usually competent, which causes datasets to be biased towards successful outcomes. However, this bias for success makes discriminating and localizing visual intentionality challenging for both learning and quantitative evaluation.

Fig: **The oops! Dataset: Each pair of frames shows an example of intentional and unintentional action in our dataset. By crawling publicly available ‘‘fail’’ videos from the web, we can create a diverse and in-the-wild dataset of accidental action. For example, at the bottom-left corner shows a man failing to see gate arm, and at the top-right shows two children playing competitive games where it is inevitable; one person will fail to accomplish their goal.**

We introduce a new annotated video dataset that is abundant with unintentional action, which we have collected by crawling publicly available ‘‘fail’’ videos from the web. From the above figure shows some examples, which cover in-the-wild situations for both intentional and unintentional action. Our video dataset, which we will publicly release, is both large (over 50 hours of video) and diverse (covering hundreds of scenes and activities). We annotated videos with the temporal location at which the video transitions from intentional to unintentional action. We define three tasks on this dataset: classifying the intentionality of action, localizing the change from intentional to unintentional, and forecasting onset of unintentional action shortly into the future.

To tackle these problems, we investigate several visual clues for learning with minimal labels to recognize intentionality. First, we propose a novel self-supervised task to learn to predict the speed of the video, which is incidental supervision available in all unlabeled videos for learning the action representation. Second, we explore the predictability of temporal context as a clue to learn features, as unintentional action often deviates from expectation. Third, we study an order of events as a clue to recognize intentionality, since intentional action usually precedes unintentional action.

Experiments and visualizations suggest that unlabeled video has intrinsic perceptual clues to recognize intentionality. Our results show that, while each self-supervised task is useful, and learning to predict the speed of video helps the most. By ablating model and design choices, our analysis also suggests that models do not rely solely on low-level motion clues to solve unintentional action prediction. Moreover, although human's consistency in our dataset is high, there is still a large gap in performance between our models and human agreement, underscoring that analyzing human goals from videos remains the

fundamental challenge in computer vision(OpenCV). We hope this dataset of unintentional and unconstrained action can provide the pragmatic benchmark of progress.

Q9. FairyTED: A Fair Rating Predictor for TED Talk Data

Answer:

In recent times, artificial intelligence is being used for inconsequential decision making. Governments make use of it in the criminal justice system to predict recidivism [brennan2009evaluating, tollenaar2013method], which affects the decision about bail, sentencing, and parole. Various firms are also using machine learning algorithms to examine and filter resumes of job applicants [nguyen2016hirability, chen2017automated, naim2016automated], which is crucial for the growth of a company. Machine learning algorithms are also being used to evaluate human's social skills, such as presentation performance [Chen2017a, Tanveer2015], essay grading.

To solve such decision-making problems, machine learning algorithms are trained on massive datasets that are usually collected in the wild. Due to difficulties in the manual curation or adjustment over large datasets, the data likely capture unwanted bias towards the underrepresented group based on race, gender, or ethnicity. Such bias results in unfair decision-making systems, leading to unwanted and often catastrophic consequences to human life and society. For example, the recognition rates of pedestrians in autonomous vehicles are reported to be not equally accurate for all groups of people [wilson2019predictive]. Matthew et al. [kay2015unequal]showed that societal bias gets reflected in the machine learning algorithms through a biased dataset and causes representational harm for occupations. Face recognition is not as useful for people with different skin tones. Dark-skinned females have 43 times higher detection error than light-skinned males.

In this work, we propose a predictive framework that tackles the issue of designing a fair prediction system from biased data. As an application scenario, we choose the problem of fair rating prediction in the TED talks. TED talks cover a wide variety of topics and influence the audience by educating and inspiring them. Also, it consists of speakers from a diverse community with imbalances in age, gender, and ethnic attributes. The ratings are provided by spontaneous visitors to the TED talk website. A machine learning algorithm trained solely from the audience ratings will have a possibility of the predicted score being biased by sensitive attributes of the speakers.

It is a challenging problem because numerous factors drive human behavior and hence have huge variability. It is challenging to know the way these factors interact with each other. Also, uncovering the true interaction model may not be feasible and often expensive. Even though the sharing platforms such as YouTube, Massive Open Online Courses (MOOC), or ted.com make it possible to collect a large amount of observational data, these platforms do not correct for bias and unfair ratings.

In this work, we utilize *causal models* [pearl2009causal] to define possible dependencies between attributes of the data. We then address the problem of not knowing true interaction model by averaging outputs of predictors across several possible causes. Further, using these causal models, we generate *counterfactual samples* of sensitive attributes. These counterfactual samples are the key components in our fair prediction framework (adapted from kusner2017counterfactual russell2017worlds) and help reducing bias in ratings wrt sensitive attributes. Finally, we introduce the novel metric to quantify degree of fairness employed by our FairyTED pipeline. To best of our knowledge, FairyTED is first fair prediction pipeline for public speaking datasets and can be applied to any dataset of similar grounds. Apart from theoretical contribution, our work also has practical implications in helping both the viewers and organizers make informed and unbiased choices for the selection of talks and speakers.