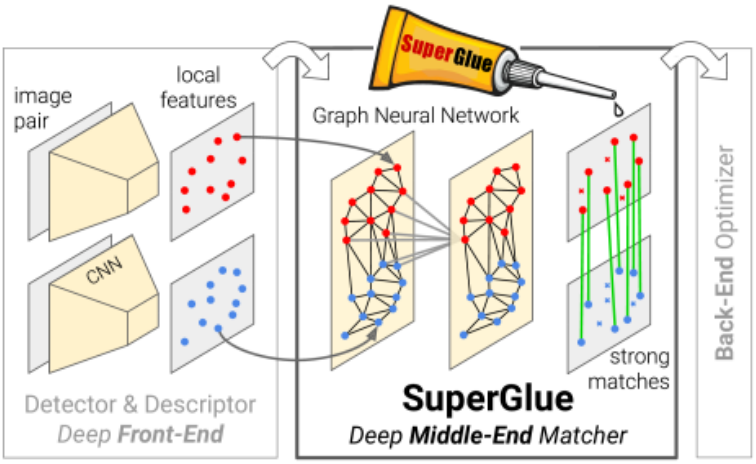
DATA SCIENCE INTERVIEW PREPARATION

(30 Days of Interview Preparation)

Day29

Q1. What is SuperGlue?

Answer:



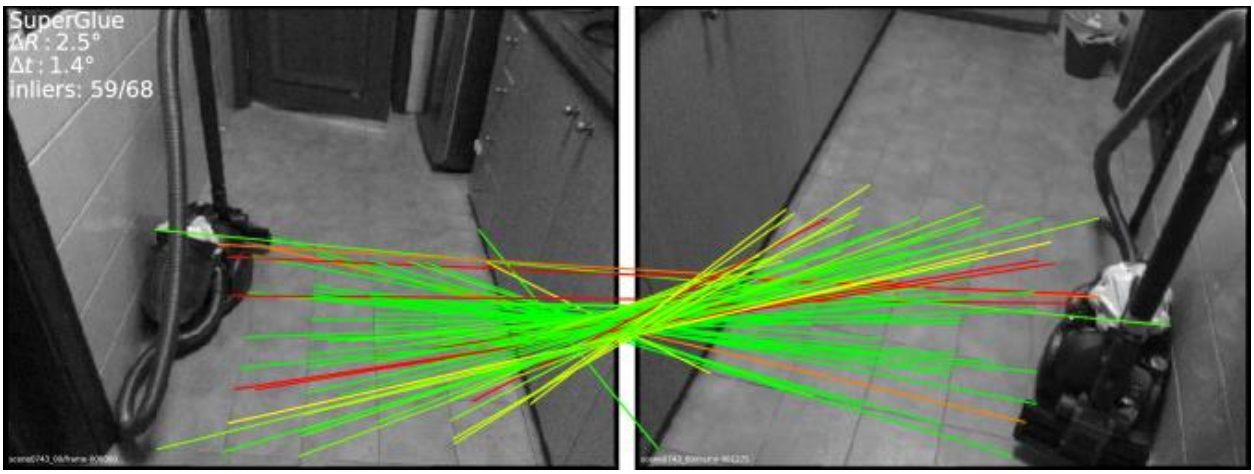
SuperGlue is a Learning Feature Matching with Graph Neural Networks. Correspondences between points in images are essential for estimating 3D structure and camera poses in geometric computer vision(OpenCV) tasks such as SLAM(Simultaneous Localization and Mapping) and SfM(Structure-from-Motion). Such correspondences are generally estimated by matching local features, the process called as data association. Broad viewpoint and lighting changes, occlusion, blur, and lack of texture are factors that make 2D-to-2D data association particularly challenging.

In this paper, we present new way of thinking about feature matching problem. Instead of learning better task-agnostic local features followed by simple matching heuristics and tricks, we propose to determine the matching process from pre-existing local features using a novel neural architecture called SuperGlue. In the context of SLAM, which typically decomposes the problem into the visual feature extraction *front-end* and the bundle adjustment or poses estimation *back-end*, our network lies directly in middle – SuperGlue is a learnable *middle-end* (see in above Figure).

In this work, *learning feature matching* is viewed as finding partial assignment between two sets of local feature. We revisit classical graph-based strategy of matching by solving the linear assignment problem, which, when relaxed to the optimal transport problem, can be solved differentiably. The cost function of this optimization is predicted by a GNN(Graph Neural Network). Inspired by success of the Transformer, it uses self- (intra-image) and cross- (inter-image) attention to leveraging both spatial relationships of

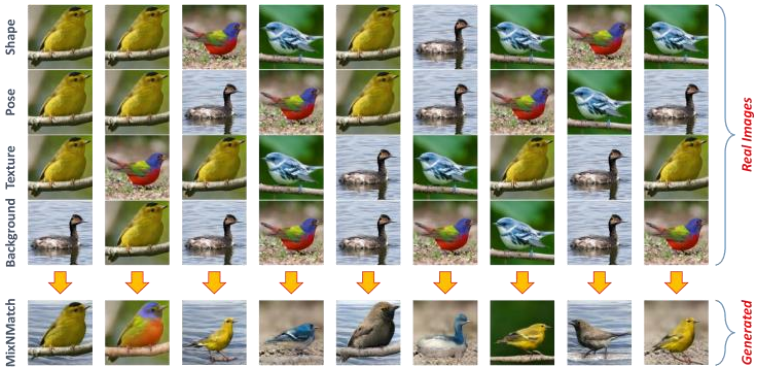
keypoints and their visual appearance. This formulation enforces assignment structure of the prediction while enabling cost to learn complex priors, handling occlusion, and non-repeatable keypoints. Our method is trained end-to-end from images pair – we learn priors for pose estimation from large annotated dataset, enabling SuperGlue to reason about 3D scene and assignment. Our work can be applied to a variety of multiple-view geometry problems that require high-quality features correspondences (see in below Figure).

We show superiority of SuperGlue compared to both handcrafted matches and learned inlier classifiers. When combined with SuperPoint, a deep front-end, SuperGlue advances the state-of-the-art on the tasks of indoor and outdoor pose estimation and paves the way towards end-to-end deep SLAM.



Q2. What is MixNMatch?

Answer:



It is a Multifactor Disentanglement and Encoding for Conditional Image Generation. Consider the real image of the yellow bird in the above Fig, First column. What would a bird look like in a different background, say that of a duck? How about in the different texture, perhaps that of the rainbow textured bird in the second column? What if we wanted to keep its texture but changes its shape to that of rainbow bird and background and pose to that of duck, as in the 3rd column? How about sampling shape, pose, texture, and experience from 4 different reference images and combining them to create entirely new image (last column)

Problem.

While research in conditional image generation has made tremendous progress, no actual work can simultaneously disentangle *background*, *object pose*, *shape*, and *texture* with minimal supervision, so that these factors can be combined from *multiple real images* for fine-grained controllable image generations. Learning disentangled representations with minimal supervision is the extremely challenging problem since the underlying factors that give rise to the data are often highly correlated and intertwined. Work that disentangles *two* such factors, by taking as input 2 reference images, e.g., one for appearance and another for pose, do exist [huang-eccv2018, joo-cvpr18, lee-eccv18, lorenz-cvpr2019, xiao-iccv2019], but they cannot disentangle other factor such as pose vs. shape or foreground vs. background appearance. Since only two factors can be controlled, these approaches cannot arbitrarily change,e.g., the object's background, shape, and texture, while keeping its pose the same. Others require intense supervision in the form of keypoint or pose or mask annotations [peng-iccv2017, Balakrishnan-cvpr2018, ma-cvpr2018, esser-cvpr2018], which limit their scalability and still fall short of disentangling all of four factors outlined above.

Our proposed conditional generative model, *MixNMatch*, aim to fill this void. MixNMatch learns to disentangle and encode background, object pose, shape, and texture latent factors from the real images, and importantly, does so with minimal human supervision. This allows, e.g., each factor to be extracted from a different actual image, and then combined for mix-and-match image generation; see in above fig. During training, MixNMatch only requires a loose bounding box around the object to the model background but requires no other supervision for modeling the object's pose, shape, and texture.

Q3. FAN: Feature Adaptation Network

Answer:

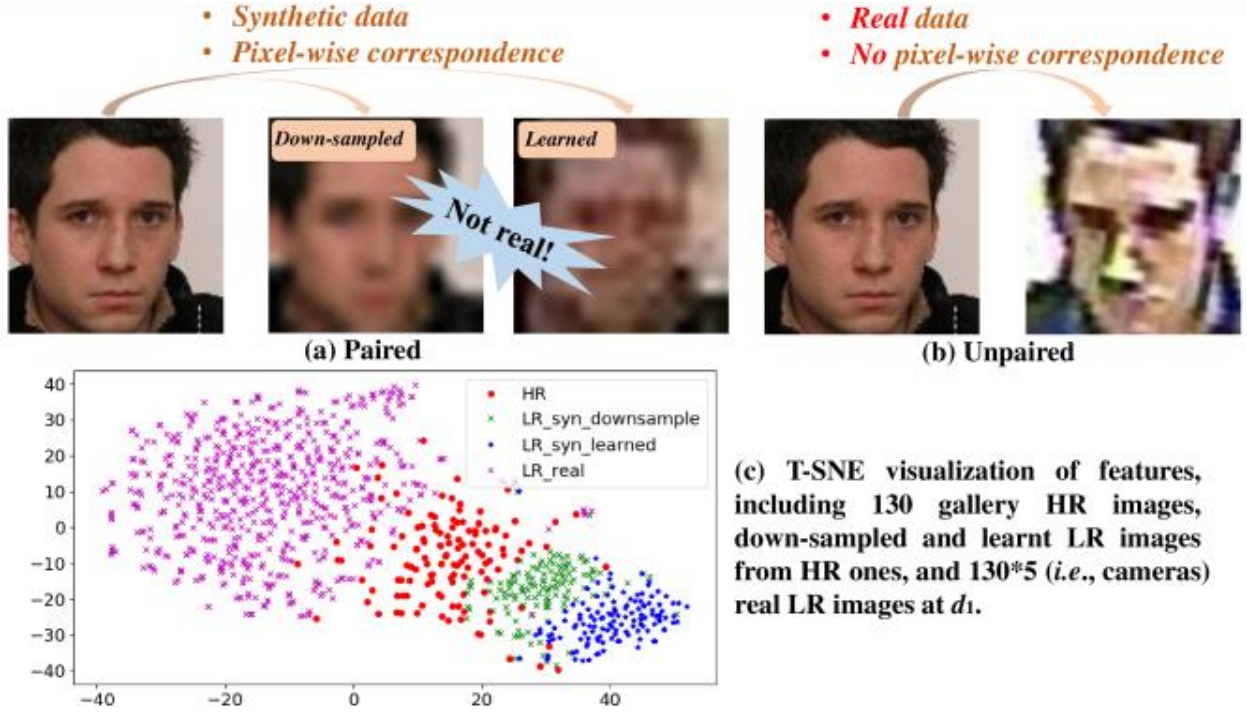
Figure: Visual results on four datasets. Vertically we show input in row first and our results in row third. For LFW and SCface datasets, we show the ground truth and gallery images in second row, respectively. For WIDER FACE and QMUL-SurFace datasets which do not have ground truth high-resolution images, we compare with two state-of-the-art(SOTA) methods: Bulat et al. [bulatyang2018learn] and FSRGAN [CT-FSRNet-2018] in row 2, respectively.

It is used for Surveillance Face Recognition and Normalization. Surveillance Face Recognition (FR) is a challenge and a significant problem yet less studied. The performance on conventional benchmarks such as LFW [LFWTech] and IJB-A have been greatly improved by state-of-the-art (SOTA) (Face Recognition(FR) methods [wang2018cosface, wen2016discriminative, deng2019arcface], which still suffer when applied to surveillance Face Recognition(FR). One intuitive approach is to perform Face Super-Resolution (FSR) on surveillance face to enhance facial details. However, existing Face Super-Resolution(FSR) methods are problematic to handle surveillance faces, because they usually ignore the *identity* information and require to *paired* training data. Preserving identity information is more crucial for surveillance of all face than recovering other information, e.g., background, Pose, Illumination, Expression (PIE).

In this work, we study surveillance face recognition(FR) and normalization. Specifically, given the surveillance face image, we aim to learn robust identity features for Face recognition(FR). Meanwhile, the feature are used to generate a normalized face with enhanced facial details and neutral PIE. Our normalization is performed mainly on the aspect of the resolution. While sharing same goal as traditional SR, it differs in removing the pixel-to-pixel correspondence between original and super-resolved images, as required by conventional SR. Therefore, we term it as face normalization. For same reason, we

compare ours to FSR instead of prior normalization methods operating on pose or expression. To the best of our knowledge, this is a *first* work to study surveillance face normalization.

We propose the novel Feature Adaptation Network(FAN) to jointly perform face recognition and normalization, which has 3 advantages over conventional FSR. i) Our joint learning scheme can benefit each other, while most FSR methods do not consider a recognition task. ii) Our framework enables training with both paired and unpaired data while conventional SR methods only support paired training. iii) Our approach simultaneously improves resolution and alleviates the background and PIE from real surveillance faces while traditional methods only act on recommendation. Examples in below Fig. One demonstrates the superiority of FAN over SOTA SR methods.



(c) T-SNE visualization of features, including 130 gallery HR images, down-sampled and learnt LR images from HR ones, and 130*5 (*i.e.*, cameras) real LR images at d_1.

Our Feature Adaptation Network (FAN) consists of 2 stages. In first stage, we adopt disentangled features learning to learns both identity and non-identity characteristics mainly from high-resolution(HR) images, which are combined as input to the decoder for pixel-wise face recovering. In second stage, we propose feature adaptation to facilitate the feature further learning from the low-resolution (LR) images by approximating feature distribution between the low-resolution and high-resolution identity encoders. There are two advantages to use Feature Adaption Network(FAN) for surveillance facial-recognition(FR) and normalization. First, Feature Adaption Network (FAN) focuses on learning disentangled identity features from Low-resolution(LR) images, which is better for facial recognition (FR) than extracting features from super-resolved faces [tran2017disentangled, zhang2018facesr, wu2016j]. 2nd, our adaptation is performed in disentangled identity feature space, which enables training with unpaired data without pixel-to-pixel

correspondences. As shown in the last fig., the synthetic paired data used in prior works [CBN_ECCV16, CT-FSRNet-
2018, bulatyang2018learn, wu2016j, zhang2018facesr, DRRN, MemNet_ICCV17, rad2019srobb]can not accurately reflect difference between real low-resolution(LR) and high-resolution(HR in-the-wild faces, which is also observed in [cai2019toward].

Furthermore, to better handle surveillance faces with the unknown and diverse resolution, we propose the Random Scale Augmentation (RSA) method that enables the network to learn all kinds of scales during training. Prior FSR [CT-FSRNet-2018, CBN_ECCV16, URDGN_ECCV16] methods either *artificially* generate the LR images from the HR ones by simple *down-sampling*, or *learn* the degradation mapping via a Convolutional Neural Network (CNN). However, their common drawback is to learn reconstruction under *fixed* scales, which may greatly limit their applications to surveillance faces. In contrast, our RSA efficiently alleviates the constraint on scale variation.

Q5. WSOD with PSNet and Box Regression

Answer:

The object detection task is to find objects belonging to specified classes and their locations in images. Benefiting from the rapid development of deep learning(DL) in recent years, the fully supervised object detection task has made significant progress. However, fully supervised task requires instance-level annotation for training, which costs lot of time and resources. Unlabeled or images labeled datasets cannot be effectively used by fully supervised method. On another hand, image-level annotated datasets are easy to generate and can even be automatically generated by web search engines. To effectively utilize these readily available datasets, we focus on weakly-supervised object detection(WSOD) tasks. The WSOD task only takes image-level annotations to train instance-level object detection network, which is different from the fully supervised object detection task.

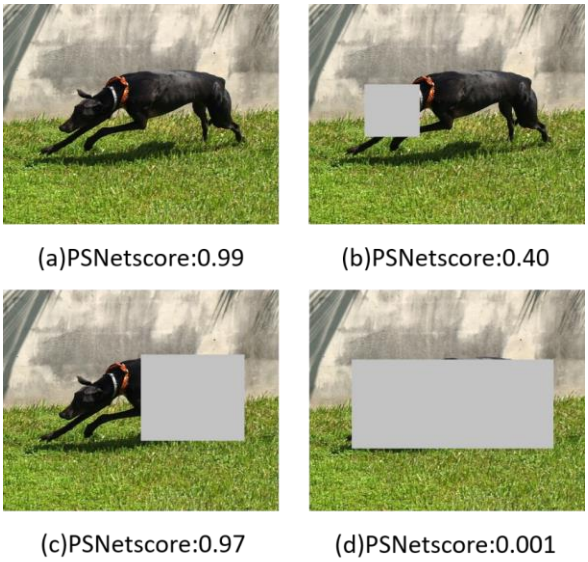
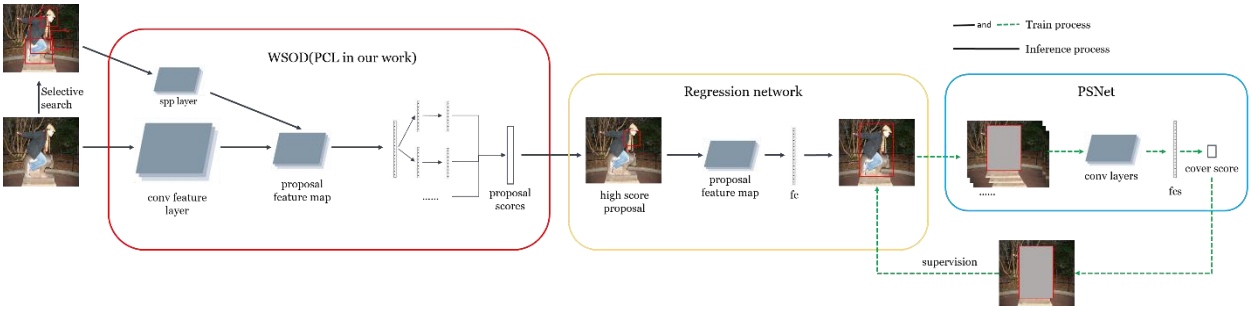


Fig.: Examples of PSNet outputs: (i) a dog without proposal occlusion, (ii) a dog whose head is occluded by the proposal box, (iii) a dog that proposal covers part of the body, and (iv) proposal completely cover the entire dog. If proposal does not completely include the whole dog, PSNet gives a high score. If proposal ultimately consists of the whole dog, PSNet gives a low score.

There are 3 main methods for weakly supervised object detection: The first is to update detector and pseudo labels from inaccurate pseudo labels iteratively; The second is to construct an end-to-end network that can take image-level annotation as supervision to train this object detection network. The third two-stage method is that taking an algorithm to optimize pseudo labels from other WSOD networks and training a fully supervised object detection network. In addition, according to different modes of proposing proposals, each of above methods can be divided into 2 classes: one is to propose proposals based on feature map that predicts probability of each pixel belonging to each class, and then get the possible instances and their locations in image; The second is detector-based method that uses the trained detector to identify multiple proposals and determines whether each proposal belongs to a specific object class or not. Comparing the effects of these methods, the end-to-end detector-based approach performs well, and our work follows this series of process.

The earliest end-to-end detector-based WSOD network is WSDDN Bilen and Vedaldi (2016), which trains a two-streams network to predict the classification accuracy of each proposal and its contributions to each class. The results of the two streams are combined to get the image classification score so that the WSDDN can take advantage of image-label annotations for training. Subsequent other work aims to improve performance of this network, like adding more classification streams, using the clustering method, adding a fully supervised module, and so on. The end-to-end detector-based approach has 2 drawbacks: one is that context information cannot be fully used to classify proposal; The second is that the most discriminative parts of the object may be detected instead of the entire object.



To make full use of the context information of the proposal and avoid finding only the most discriminative part, we design a new network structure that adds a box regression branch to the traditional WSOD network. In the previous WSOD network, there is usually no box regression part, while this branch plays an essential role in fully supervised object detection networks. The box regression network can adjust position and scale of proposal, make it closer to the ground truth. In the fully supervised object detection task, we can use the instance-level label as supervision to train box regression network; but in WSOD task, network cannot obtain the instance-level annotation and thus cannot train this branch. To obtain reliable instance annotation to train the regression network, we designed the proposal scoring network named PSNet that can detect whether proposal completely covers the object. The PSNet is specially trained multi-label classification network. Even if the object in the image is occluded or incomplete, the PSNet can detect the presence of the object. The PSNet can be used to evaluate images without proposal area. If the proposal completely covers whole object, rest of the image will not contain information about it. We use PSNet to evaluate the output of the WSOD network, and then select appropriate proposals as pseudo labels to train box regression network. Examples of the output of PSNet are shown in the above Figure.

Q6. Autonomous Driving Assistance Systems (ADAS) and Vehicle Automation.

Answer:

Vehicles are being equipped with increasingly complex autonomous driving assistance systems (ADAS) that take over parts of driving tasks previously performed by the human driver. There are several different ADAS technologies in vehicles, starting from basics that have been in vehicles for several years, such as automatic windscreen wipers and anti-lock braking systems. More advanced techniques are already on

the road today, where both the longitudinal (braking/accelerating, e.g., adaptive cruise control) and lateral (steering, e.g., assisted lane-keeping) control of the vehicle is shifting to ADAS. Further enhanced levels of automated driving functionality include autopilot (Tesla), intellisafe (Volvo), and Distronic plus steering assist (Mercedes). Overall this fast pace of market penetration of ADAS in vehicles has not allowed drivers to develop understanding of new systems over an extended period.

The most common taxonomy to capture the development of ADAS technology in cars are SAE's levels of automation sae. This approach is based on six levels of automation, ranging from no automation (level 0) to full automation (level 5). In particular, in levels 2/3, the automated system can take partial control of vehicle, where level 2 expectations of the human driver are to monitor the system and intervene appropriately, while the level 3 expectation of the human driver is to intervene appropriately upon a request from the system. Today most ADAS technology equipped cars are at level 1, in which progression to partial/semi-automation (level 2/3) with in-built ADAS technology in even lower-priced car models is becoming more common. Also, level 2/3 automation will likely be reality for some time to come, given that fuller automation (4/5) is emerging slowly without clear market deployment roadmap.

One of main challenges that arise in level 2/3 automation is transition of control from the ADAS to the human driver, often referred to as the "handover problem." This transition is, according to social factors and safety research, a phase where human attention and reliability is critical, but where humans tend to underperform in those respects son2017situation. E.g., research has indicated that automatic cruise control technology leads to a reduction in mental workload and, thus, to problems with regaining control of the vehicle in failure scenarios stanton1998vehicle. Additionally, a common misconception concerning ADAS technology is that when more automation is introduced, human error will disappear atlantic2015save, which may give rise to the problematic idea that driver training is not necessarily needed. However, social factors research advises against not training for the use of new sophisticated automation technology lee2006human; salas2006design; saetren2015effects, as humans in the technology loop will still be needed for use, maintenance or design of the technology. It may even be that increased automation increases the level of competence required for the driver, as the driver must know both how to handle system manually, for instance, if the sensors in a car stop working due to bad weather, in addition to knowing how to control and supervise the advanced automation technology.

In our previous work rismani2018qualitative, we performed a qualitative survey and found that the handover problem is challenging, and it is unclear to drivers how this could best be handled securely. Furthermore, drivers were worried about the implications of vehicle automation due to lack of knowledge and experience of level 2/3 systems and seemed concerned about the kind of training and licensing that accompanies these developments in vehicle automation. The lack of certainty around training and licensing concerning emerging ADAS technologies is a relevant ethical concern, as it exposes a gap in regulation and industry best practices that have not been the focus of much research to date.

This lack of certainty around driver training and licensing wrt level 2/3 automation systems underscores the need to understand better the following research questions: (i) What are drivers' awareness of ADAS in their vehicles, (ii) How knowledgeable are drivers about ADAS in their vehicles, and (iii) How willing are drivers to engage or use ADAS in their vehicles? Overall we expect to see people's engagement or use pattern of ADAS technologies in their vehicle correlate to their awareness and knowledge of those techniques.

Previous work has looked at driver perception of ADAS and vehicle automation, including understanding learner drivers' perspective of Blind Spot Detection(BSD) and Adaptive Cruise Control(ACC) systems. That work found that driver's awareness, use, and perceived safety of Blind Spot Detection(BSD) was higher than that of ACC tsapi_introducing_2015, and contributed to a greater understanding of driver preparation and acceptance of ADAS crump2016differing, and how drivers learn and prefer to learn about ADAS, and what their expectations are regarding ADAS and vehicle automation hoyos2018consumer.

To answer our research questions, we performed a quantitative public survey of issues specific to the public's awareness, knowledge, and use of ADAS technologies in level 2/3 automation. Also, based on previous work tsapi_introducing_2015; crump2016differing; hoyos2018consumer, we analyzed gender and age relationships as well as income and type of training with regards to our research questions above.

Q7. Robot Learning and Execution of Collaborative Manipulation Plans from YouTube Videos.

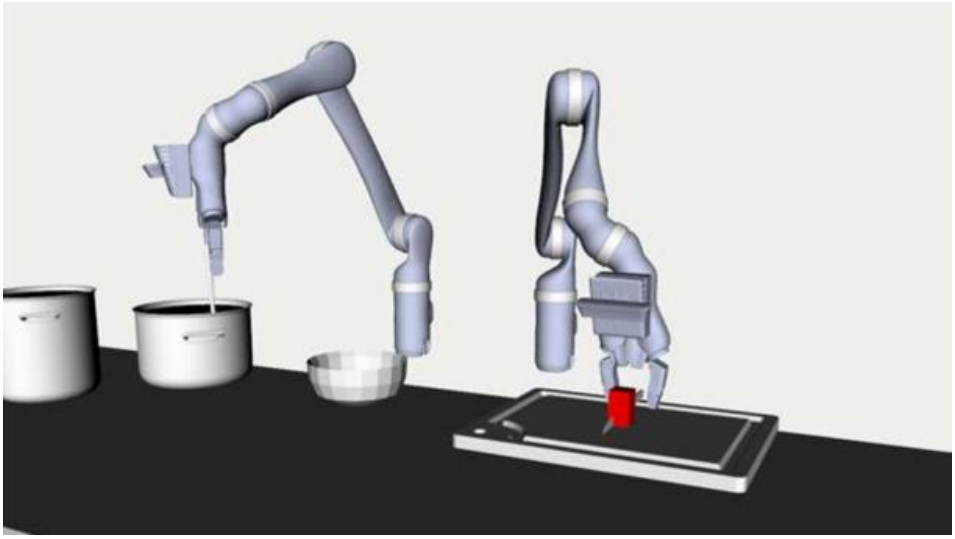
Answer:

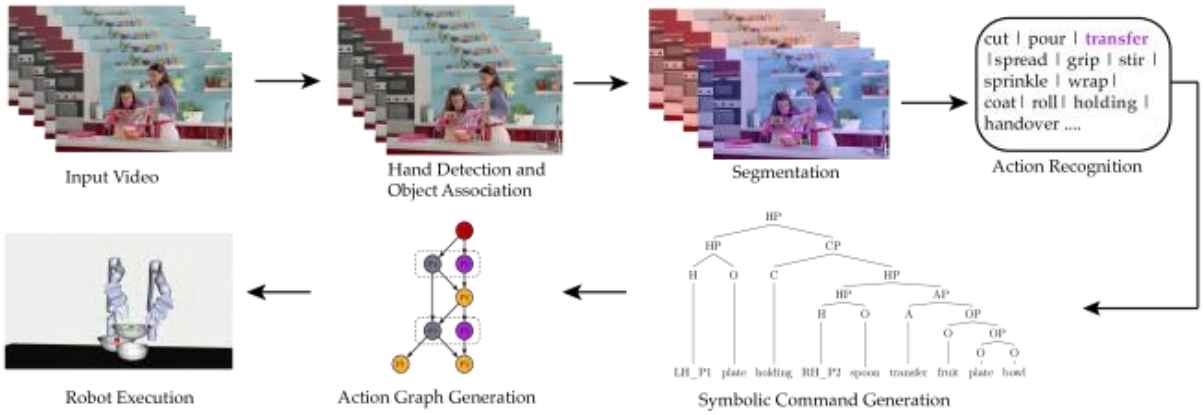
We focus on problem of learning collaborative action plans for robot. Our goal is to have robot "watch" unconstrained videos on web, extract the action sequences shown in the videos and convert them to an executable plan that it can perform either independently or as part of a human-robot or robot-robot team.

Learning from online videos is hard, particularly in collaborative settings: it requires recognizing the actions executed, together with manipulated tools and objects. In many collaborative tasks, these actions include handing objects over or holding object for the other person to manipulate. There is a very large variation in how the actions are performed and collaborative actions may overlap spatially and temporally.

In our previous work [hejia_isrr19], we proposed a system for learning activities performed by two humans collaborating at the cooking task. The system implements a collaborative action grammar built upon action grammar initially proposed by Yang et al. [yang2015robot]. Qualitative analysis in 12 clips showed that parsing these clips with grammar results in human-interpretable tree structures representing

a variety of single and collaborative actions. The clips were manually segmented and were approximate 100 frames each.





In this paper, we generalize this work with a framework for *generating single and collaborative action trees from full-length YouTube videos lasting several minutes* and *concatenating the trees in an action graph that is executable by one or more robotic arms*.

The framework takes as input YouTube video showing collaborative tasks from start to end. We assumes that objects in video are annotated with label and bounding boxes, e.g., by running the YOLOv3 algorithm. We also think a skill library that associates a detected action with skill-specific motion primitives. We focus on cooking tasks because of the variety of manipulation actions and their importance in-home service robotics.

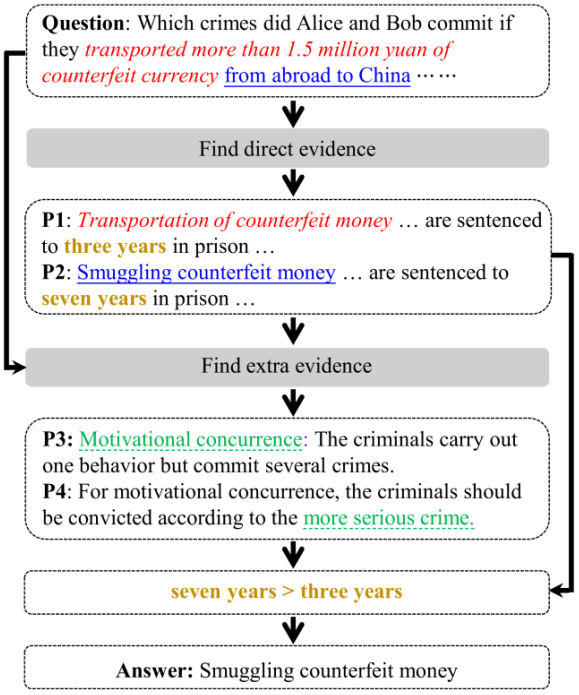
In second fig. shows the components of proposed framework. We rely on insight that hands are main driving force of manipulation actions. We detect the human hands in the video and use the hand trajectories to split the video into clips. We then associate objects and hands spatially and temporally to recognize the actions and generate human-interpretable robot commands. Finally, we propose an open-sourced platform for creating and executing an action graph. We provide a quantitative analysis of performance in two YouTube videos of 13401 frames in total and a demonstration in the simulation of robots learning and performing the actions of the third video of 2421 frames correctly.

While the extracted action sequences are executed in the open-loop manner and thus do not withstand real-world failures or disturbances, we find that this work brings us the step closer to having robots generate and execute variety of semantically meaningful plans from watching videos online.

Q8. JEC-QA: A Legal-Domain Question Answering Dataset

Legal **Q**uestion **A**nswering (**LQA**) aims to provide explanations, advice, or solutions for legal issues. A qualified LQA system can not only demonstrates a professional consulting service for unskilled humans but also help professionals to improve work efficiency and analyze real cases more accurately, which makes LQA an important NLP application in the legal domain. Recently, many researchers attempt to

build LQA systems with machine learning techniques and neural networks. Despite these efforts in employing advanced NLP models, LQA is still confronted with the following two significant challenges. The first is that there is less qualified LQA dataset, which limits the research. The second is that the cases and questions in the legal domain are very complex and rigorous. As shown in Table 1, most problems in LQA can be divided into two typical types: the knowledge-driven questions (KD-questions) and case-analysis questions (CA-questions). KD-questions focus on the understanding of specific legal concepts, while CA-questions concentrate more on the analysis of real cases. Both types of questions require sophisticated reasoning ability and text comprehension ability, which makes LQA a hard task in NLP.



To get a better understanding of these reasoning abilities, we show a question of JEC-QA in Fig. 1describing a criminal behavior that results in two crimes. The models must understand "Motivational Concurrence" to reason out further evidence rather than lexical-level semantic matching. Moreover, the models must have the ability of multi-paragraph reading and multi-hop reasoning to combine the direct evidence and the additional evidence to answer the question, while numerical analysis is also necessary for comparing which crime is more dangerous. We can see that answering one question will need multiple reasoning abilities in both retrieving and answering, makes JEC-QA a challenging task.

To investigate the challenges and characteristics of LQA, we design a unified OpenQA framework and implement seven representative neural methods of reading comprehension. By evaluating the

performance of these methods on JEC-QA, we show that even the best approach can only achieve about 25% and 29% on KD-questions and CA-questions, respectively, while skilled humans and unskilled humans can reach 81% and 64% accuracies on JEC-QA. The experimental results show that existing OpenQA methods suffer from the inability of complex reasoning on JEC-QA as they cannot well understand legal concepts and handle multi-hop logic.

Q9. SpoC: Spoofing Camera Fingerprints

Answer:

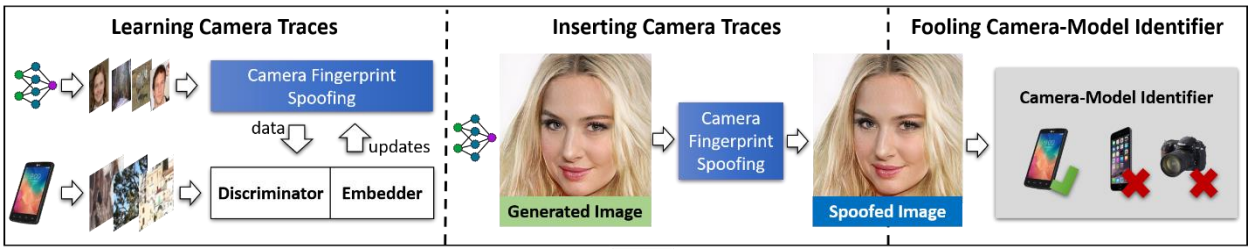


Figure 1: *SpoC* learns to spoof camera fingerprints. It can be used to insert camera traces to a generated image. Experiments show that we can fool state-of-the-art camera-model identifiers that were not seen during training.

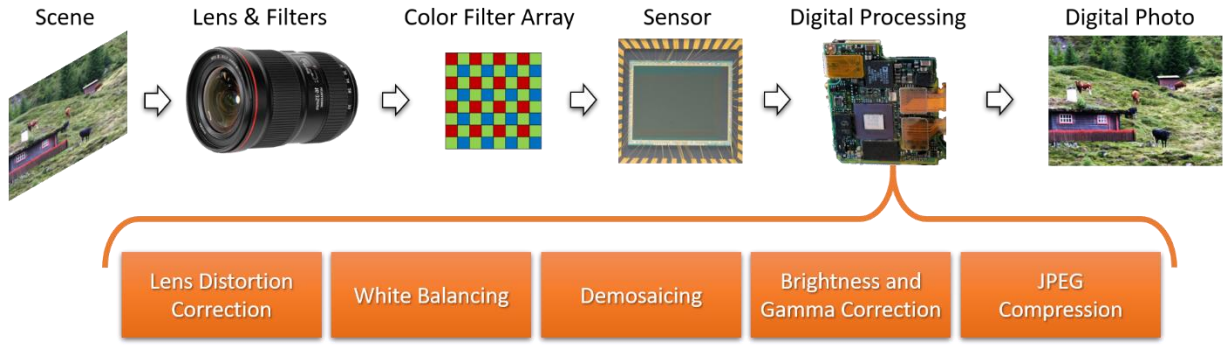


Figure 2: A digital image of a scene contains camera-related traces of the image formation process that could act as a fingerprint of a camera model. The used lenses and filters, the sensor, and the manufacturer-specific digital processing pipelines result in unique patterns. These patterns can be used to identify camera models.

There have been astonishing advances in synthetic media generation in the last few years, thanks to deep learning, and in particular to Generative Adversarial Networks (GANs). This technology-enabled a

significant improvement in the level of realism of generated data, increasing both resolution and quality. Nowadays, powerful methods exist for creating an image from scratch, and for changing its style or only some specific attributes. These methods are beneficial, especially on faces, and allow one to change the expression of a person easily or to modify its identity through face-swapping. This manipulated visual content can be used to build more effective fake news. It has been estimated that the average number of reposts for a report containing an image is 11 times larger than for those without images. This raises serious concerns about the trustworthiness of digital content, as testified by the growing attention to the profound fake phenomenon.

The research community has responded to this threat by developing several forensic detectors. Some of them exploit high-level artifacts, like asymmetries in the color of the eyes, or anomalies arising from an imprecise estimation of the underlying geometry. However, technology improves so fast that these visual artifacts will soon disappear. Other approaches rely on the fact that any acquisition device leaves distinctive traces on each captured image, because of its hardware, or its signal processing suite. They allow associating a media with its acquisition device at various levels, from the type of source (camera, scanner, etc.), to its brand/model (e.g., iPhone6 vs. iPhone7), to the individual device. A primary impulse to this field has been given by the seminal work of Lukàs et al., where it has been shown that reliable device identification is possible based on the camera photo-response non-uniformity (PRNU) pattern. This pattern is due to tiny imperfections in the silicon wafer used to manufacture the imaging sensor and can be considered as a type of device fingerprint.

Beyond extracting fingerprints that contain device-related traces, it is also possible to recover camera model fingerprints. These are related to the internal digital acquisition pipeline, including operations like demosaicing, color balancing, and compression, whose details differ according to the brand and specific model of the camera (See Fig.2). Such differences help attribute images to their source camera, but can also be used to highlight better anomalies caused by image manipulations. The absence of such traces, or their modification, is a strong clue that the image is synthetic or has been manipulated in some way. Detection algorithms, however, must confront with the capacity of an adversary to fool them. This applies to any classifier and is also very well known in forensics, where many counter-forensics methods have been proposed in the literature. Indeed, forensics and counter-forensics go hand in hand, a competition that contributes to improving the level of digital integrity over time.

In this work, we propose a method to synthesize traces of cameras using a generative approach that is agnostic to the detector (i.e., not just targeted adversarial noise). We achieve this by training a conditional generator to jointly fool an adversarial discriminator network as well as a camera embedding network. To this end, the proposed method injects the distinctive traces of a target camera model in synthetic images, while reducing the first generation traces themselves, leading all tested classifiers to attribute such images to the target camera (’targeted attack’).