Pushing the Frontiers of Computer Vision
CVPR Conference Report 2018

June 18 – 22, 2018
Salt Lake City, Utah, USA.

Written by: Tassilo Klein and Frederik Pahde

THE BEST RUN SAP

The 2018 conference on Computer Vision and Pattern Recognition (CVPR) took place between June 18 - 22 in Salt Lake City, Utah. As the premier and highly competitive conference in the realm of computer vision, CVPR provides a platform for a diverse group of academics, researchers, technologists, industrial giants and high-tech start-ups to showcase the field's latest innovations.

CVPR this year has shown significant growth; making it the largest CVPR conference with more than 6,000 attendees. Known for its diligent and high-quality review process, CVPR received 3309 conference paper submissions this year, out of which only 979 papers were accepted. Additionally, the conference hosted 21 tutorials, 48 workshops, the annual doctoral consortium, along with an industrial exhibition that featured around 150 companies.

The conference has incited numerous stimulating discussions and showcased a wide range of novel papers and presentations. Machine learning, particularly, was at the forefront of CVPR this year, scoring 24% of total research with a total of 233 papers submitted on the topic. Research on object recognition and scene understanding has also dominated this year's conference with 202 papers alone.

As one of CVPR's official sponsors, the SAP Leonardo Machine Learning Research team contributed to the discussion with our recent research project focusing on multimodality as an effective approach to address the shortcomings of deep learning models combining visual and natural language. Our paper "Cross-modal Hallucination for Few-shot Fine-grained Recognition" was a part of the workshop on Fine-Grained Visual Categorization. The paper proposes a multimodal approach that addresses the lack of sufficient data for model training. Our multimodal benchmark approach employs a two-phase training process with images and text descriptions to better train the model to understand and identify visual classifiers. Moreover, our research partners from the University of Pittsburgh presented their work on "Deep Ordinal Regression Network for Monocular Depth Estimation". Their proposed deep ordinal regression network (DORN) achieves state-of-the-art results and outperforms the existing methods for depth estimation by far.

In this report, we have put together a summary of the conference's main trends and highlights, along with our own selection of what we deem must-read papers.

A Glance at the Main Trends and Highlights

Multimodality: Bridging the Gap between Visual and Natural Language

Multimodality was one of the most noticeable trends at this year's CVPR, particularly in vision and language models, such as Visual Question Answering (VQA) and Visual Dialog (VisDial) systems. Visual and natural language models are still undergoing several testbeds and various shortcomings have been highlighted. One such shortcoming is the lack of an integrative multi-modal approach that would allow for the improvement of interpretability and perception ensuring that the systems learn to generalize.

Visual Question Answering (VQA): Through this task, a system is given an image and a natural question about the content of the image and asked to produce a natural language answer (to the image-question pair). Answers can be provided in the form of multiple choice, e.g. the system is given 2 − 4 choices and has to determine which option is most likely to be the correct answer or in terms of filling in blanks, where the system would need to generate an appropriate word for a given blank position.

Visual Dialog (VisDial): The system engages in a meaningful dialog about visual content with humans in conversational language. More precisely, given an image, a dialog history and a follow-up question about the image, the system has to answer questions about the content displayed.

Take a look at our recent blog post and planned ECCV workshop for more details.

Synthetic Data, Self-Supervision and the Future of AI in the Medical Field

Another topic that is gaining more attention is the usage of sophisticated synthetic data from environments that mimic the real world with high fidelity, in conjunction with domain adaptation with real data; rendering big data curation unnecessary. Similarly, the topic of self-supervision is gaining momentum. In self-supervised learning, the training labels are directly determined from the input data, thus no manual data annotation is required. One example is solving puzzles, e.g. an image is cut into pieces that are shuffled and the neural network has to learn which parts belong together. Another example is using unlimited amount of color video data, in which the data may be converted to grayscale and assigning the machine the task of recoloring the images.

Computer vision papers concerned with the medical field still constitute a small niche. However, the number of related papers are increasing as the topic continues to gain traction. The topics covered in this area of research are multi-modality of patients' images and text reports as well as segmentation.

"Good Citizen of CVPR": Skills & Ethics of the CVPR Community

A great initiative this year was the panel 'Good Citizen of CVPR', which focused on establishing a CVPR community culture and code of ethics. The panel included a variety of sessions on research, writing and presentation skills, as well as topics such as representation, inclusiveness and building up a community based on mentorship and leadership.

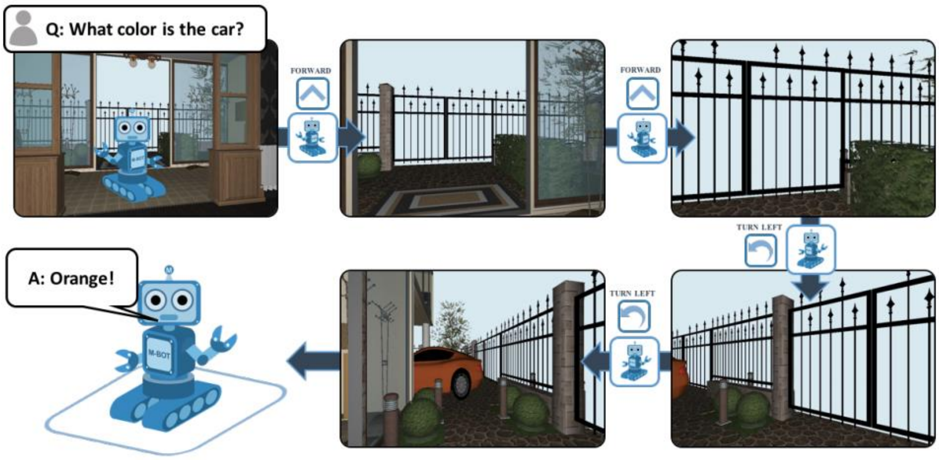
Our Selection of Must-read Papers

In this section, we have selected some of the papers presented at the conference and added our personal insights of why they are of particular interest to the computer vision community.

- **Embodied Question Answering**

Das et al., 2017

Abstract:
We present a new AI task – Embodied Question Answering (EmbodiedQA) – where an agent is spawned at a random location in a 3D environment and asked a question ('What color is the car?'). In order to answer, the agent must first intelligently navigate to explore the environment, gather necessary visual information through first-person (egocentric) vision, and then answer the question ('orange'). EmbodiedQA requires a range of AI skills – language understanding, visual recognition, active perception, goal-driven navigation, commonsense reasoning, long-term memory, and grounding language into actions. In this work, we develop a dataset of questions and answers in House3D environments, evaluation metrics, and a hierarchical model trained with imitation and reinforcement learning.



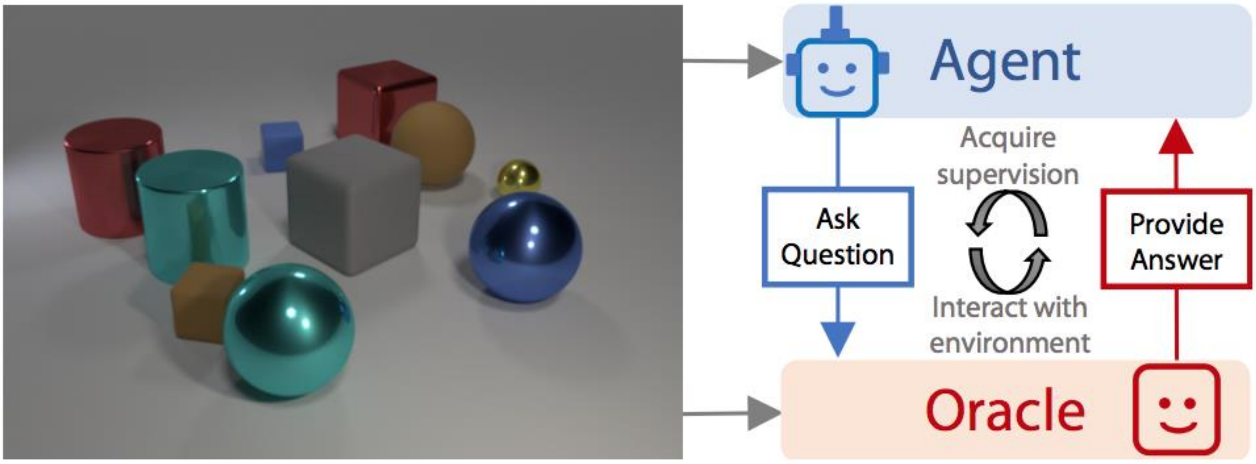
Source 1 (Das et al., 2017)

Why we think it's interesting:
Advancing standard VQA models was the focus of many papers in CVPR this year. "Embodied QA" aims at reaching towards the goal of creating fully intelligent agents that can actively perceive, naturally communicate in an environment-grounded dialogue and act and execute commands. Through a goal-driven intelligent navigation of a 3D setting, the agent is asked to answer questions based on object recognition and visual grounding and understanding. Interestingly, the agent solely uses egocentric vision to navigate its surroundings, i.e., it's not provided with a map and trained only via "raw sensory input" (pixels and words) and must rely on common sense in navigating an unfamiliar environment.

- **Learning by Asking Questions**

Misra et al., 2018

Abstract:
We introduce an interactive learning framework for the development and testing of intelligent visual systems, called learning-by-asking (LBA). We explore LBA in context of the Visual Question Answering (VQA) task. LBA differs from standard VQA training in that most questions are not observed during training time, and the learner must ask questions it wants answers to. Thus, LBA more closely mimics natural learning and has the potential to be more data-efficient than the traditional VQA setting. We present a model that performs LBA on the CLEVR dataset, and show that it automatically discovers an easy-to-hard curriculum when learning interactively from an oracle. Our LBA generated data consistently matches or outperforms the CLEVR train data and is more sample efficient. We also show that our model asks questions that generalize to state-of-the-art VQA models and to novel test time distributions.



Source 2: (Misra et al., 2017)

Why we think it’s interesting:
Standard VQA models passively rely on large static datasets; unlike the interactive nature of human learning that is more sample efficient and less redundant. Learning by asking (LBA) fills this research gap by introducing a more interactive VQA model that mimics natural learning. LBA trains the agent to learn like a human by evaluating its prior acquired knowledge and asking “good” and “relevant” questions that maximize the learning signal from each image-question pair sent to the oracle. The paper also shows how interactive questioning significantly reduces redundancy and the required number of training samples to achieve accuracy by 40%.

- Decorrelated Batch Normalization

Huang et al., 2018

Abstract:
Batch Normalization (BN) is capable of accelerating the training of deep models by centering and scaling activations within mini-batches. In this work, we propose Decorrelated Batch Normalization (DBN), which not just centers and scales activations but whitens them. We explore multiple whitening techniques, and find that PCA whitening causes a problem we call stochastic axis swapping, which is detrimental to learning. We show that ZCA whitening does not suffer from this problem, permitting successful learning. DBN retains the desirable qualities of BN and further improves BN's optimization efficiency and generalization ability. We design comprehensive experiments to show that DBN can improve the performance of BN on multilayer perceptrons and convolutional neural networks. Furthermore, we consistently improve the accuracy of residual networks on CIFAR-10, CIFAR-100, and ImageNet.

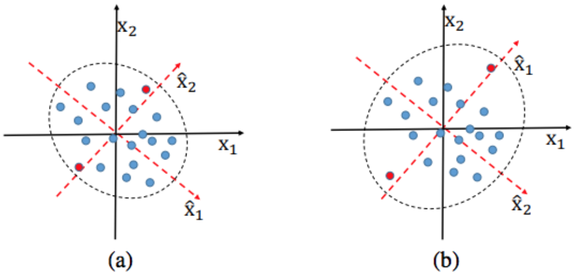


Figure 1. Illustration that PCA whitening suffers from stochastic axis swapping. (a) The axis alignment of PCA whitening in the initial iteration; (b) The axis alignment in another iteration.

Source 3 (Huang et al., 2018)

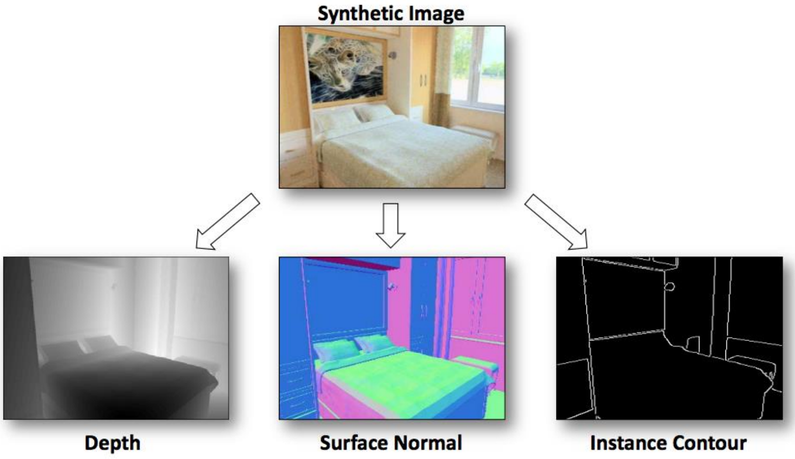
Why we think it's interesting:
Batch normalization (BN) has become a standard technique for deep learning in order to optimize the learning process. In contrast to BN, which effectively just centers and rescales the layer inputs, the proposed paper also decorrelates the input which further improves the performance. Extending previous research, the authors utilize the differentiability of the eigen-decomposition to perform backpropagation during training.

- Cross-Domain Self-supervised Multi-task Feature Learning using Synthetic Imagery

Ren and Lee, 2017

Abstract:
In human learning, it is common to use multiple sources of information jointly. However, most existing feature learning approaches learn from only a single task. In this paper, we propose a novel multi-task deep network to learn generalizable high-level visual representations. Since multi-task learning requires annotations for multiple properties of the

same training instance, we look to synthetic images to train our network. To overcome the domain difference between real and synthetic data, we employ an unsupervised feature space domain adaptation method based on adversarial learning. Given an input synthetic RGB image, our network simultaneously predicts its surface normal, depth, and instance contour, while also minimizing the feature space domain differences between real and synthetic data. Through extensive experiments, we demonstrate that our network learns more transferable representations compared to single-task baselines. Our learned representation produces state-of-the-art transfer learning results on PASCAL VOC 2007 classification and 2012 detection.



Source 4: (Ren and Lee, 2017)

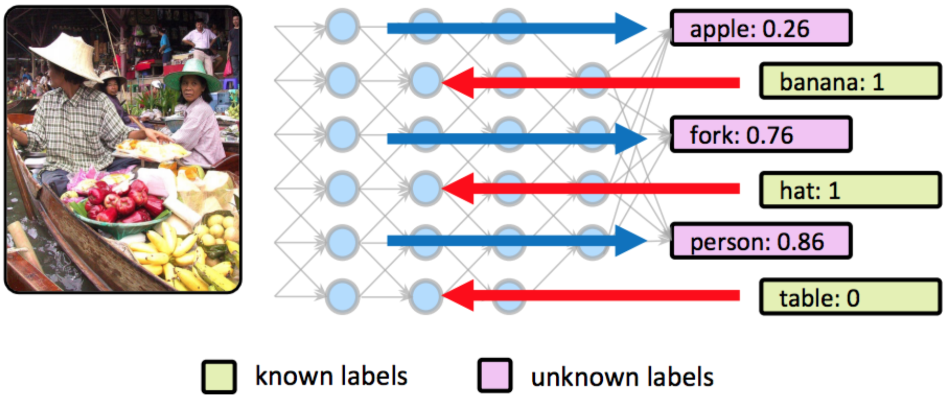
Why we think it's interesting:
The human ability to learn simultaneously from various information sources is still lacking in most existing feature learning approaches. This paper addresses this gap by proposing an original multi-task deep learning network that uses synthetic imagery to better learn visual representations in a cross-modal setting. Training the network through synthetic images dramatically reduces data annotations needed for multitask learning, which is costly and time-consuming. To bridge the cross-domain gap between real and synthetic data, adversarial learning is employed in "an unsupervised feature-level domain adaptation" method, which enhances performance upon the transfer of acquired visual features knowledge to real world tasks.

- **Feedback-prop: Convolutional Neural Network Inference under Partial Evidence**

Wang et al., 2018

Abstract:
We propose an inference procedure for deep convolutional neural networks (CNNs) when partial evidence is available. Our method consists of a general feedback-based propagation approach (feedback-prop) that boosts the prediction accuracy for an arbitrary set of unknown target labels when the values for a non-overlapping arbitrary set of target labels are known. We show that existing models trained in a multi-label or multi-task setting can readily take advantage of feedback-prop without any retraining or fine-tuning. Our feedback-

prop inference procedure is general, simple, reliable, and works on different challenging visual recognition tasks. We present two variants of feedback-prop based on layer-wise and residual iterative updates. We experiment using several multi-task models and show that feedback-prop is effective in all of them. Our results unveil a previously unreported but interesting dynamic property of deep CNNs. We also present an associated technical approach that takes advantage of this property for inference under partial evidence in general visual recognition tasks.



Source 5 (Wang et al.,2018)

Why we think it's interesting:
Typical CNN-based models for categorization with multiple targets do not allow the usage of partial evidence in scenarios where some labels are known. This paper proposes "Feedback-based propagation" to overcome this gap. Classical CNNs use back-propagation during training and forward-propagation for inference. In contrast, the proposed method uses feedback-based propagation where both forward- and back-propagation can share information using intermediate neural activations. This allows for CNN inference under partial evidence, such that known labels can be employed to improve classification results for unknown labels.

- **Empirical Study of the Topology and Geometry of Deep Networks**

Fawzi et al., 2018

Abstract:
The goal of this paper is to analyze the geometric properties of deep neural network image classifiers in the input space. We specifically study the topology of classification regions created by deep networks, as well as their associated decision boundary. Through a systematic empirical study, we show that state-of-the-art deep nets learn connected classification regions, and that the decision boundary in the vicinity of datapoints is flat along most directions. We further draw an essential connection between two seemingly unrelated properties of deep networks: their sensitivity to additive perturbations of the inputs, and the curvature of their decision boundary. The directions where the decision boundary is curved in fact characterize the directions to which the classifier is the most vulnerable. We finally

leverage a fundamental asymmetry in the curvature of the decision boundary of deep nets, and propose a method to discriminate between original images, and images perturbed with small adversarial examples. We show the effectiveness of this purely geometric approach for detecting small adversarial perturbations in images, and for recovering the labels of perturbed images.

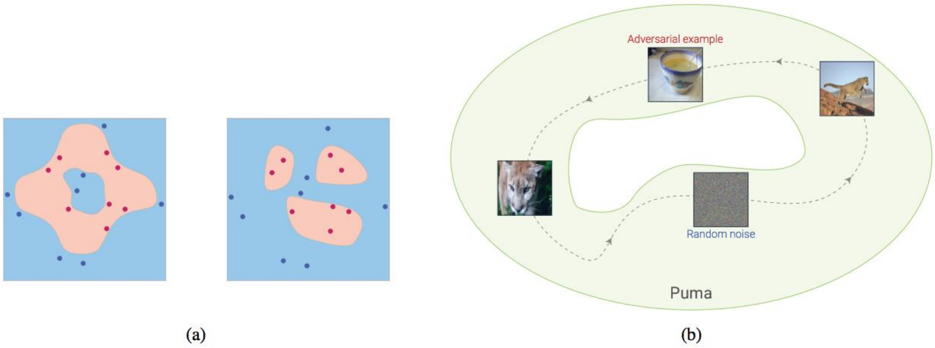


Figure 1: **(a)** Disconnected versus connected yet complex classification regions. **(b)** All four images are classified as puma. There exists a path between two images classified with the same label.

Source 6 (Fawzi et al., 2018))

Why we think it's interesting:
This paper studies the topology of classification regions created by deep networks, as well as their associated decision boundary. This is of particular interest since, compared to other central features of deep networks such as generalization, there has been little emphasis on this research area. The authors presented that state-of-the-art deep nets have the ability to learn connected classification regions. Furthermore, it is intriguing to learn that the decision boundary in the vicinity of natural data points is flat along most directions; whereas some curved directions are shared across data points. With respect to adversarial perturbations, these shared directions are where the deep networks are most vulnerable. Additionally, curvature asymmetry for real data points is employed for detecting adversarial perturbed samples from original samples. In fact, this purely geometrical approach is a unique way to enhance deep neural network image classifiers' resistance against perturbations.

****BEST PAPER AWARD****

- **Taskonomy: Disentangling Task Transfer Learning**

Zamir et al., 2018

Abstract:
Do visual tasks have a relationship, or are they unrelated? For instance, could having surface normals simplify estimating the depth of an image? Intuition answers these questions positively, implying existence of a structure among visual tasks. Knowing this structure has notable values; it is the concept underlying transfer learning and provides a principled way for identifying redundancies across tasks, e.g., to seamlessly reuse supervision among related tasks or solve many tasks in one system without piling up the complexity.

We propose a fully computational approach for modeling the structure of space of visual tasks. This is done via finding (first and higher-order) transfer learning dependencies across a dictionary of twenty six 2D, 2.5D, 3D, and semantic tasks in a latent space. The product is a computational taxonomic map for task transfer learning. We study the consequences of this structure, e.g. nontrivial emerged relationships, and exploit them to reduce the demand for labeled data. For example, we show that the total number of labeled datapoints needed for solving a set of 10 tasks can be reduced by roughly 2/3 (compared to training independently) while keeping the performance nearly the same. We provide a set of tools for computing and probing this taxonomical structure including a solver that users can employ to devise efficient supervision policies for their use cases.

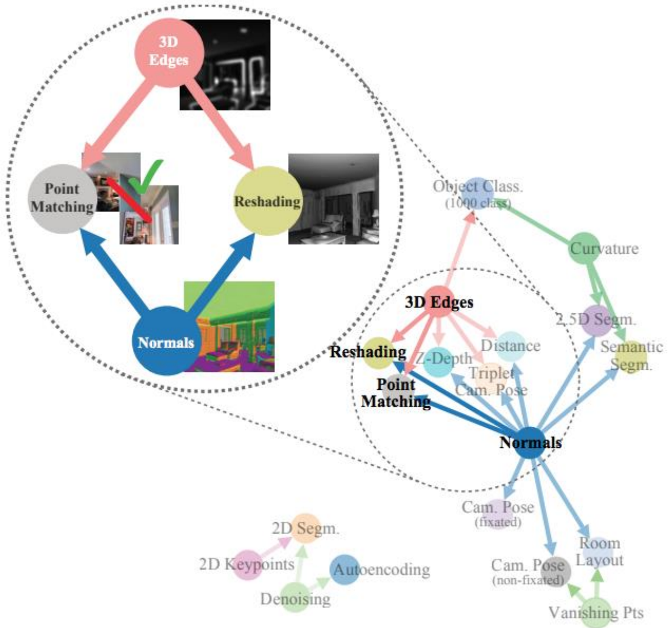


Figure 1: **A sample task structure discovered by the computational task taxonomy (*taskonomy*).** It found that, for instance, by combining the learned features of a surface normal estimator and occlusion edge detector, good networks for reshading and point matching can be rapidly trained with little labeled data.

Source 7 (Zamir et al., 2018))

Why we think it's interesting:
Using a fully sophisticated computational approach, the paper proposes "a computational taxonomic map" that interconnects and correlates relationships and transfer learning dependencies between different tasks to facilitate task transfer learning in a more orchestrated manner. Identifying redundancies across tasks can be exploited for new tasks by simply re-using existing networks in conjunction with feature transfer functions. As a result, the amount of required labeled data can be dramatically reduced as only a couple of iterations of fine-tuning might be necessary in order to obtain a high level of accuracy.

- **Learning to Segment Every Thing**

Hu et al., 2018

Abstract:

Most methods for object instance segmentation require all training examples to be labeled with segmentation masks. This requirement makes it expensive to annotate new categories and has restricted instance segmentation models to ∼100 well-annotated classes. The goal of this paper is to propose a new partially supervised training paradigm, together with a novel weight transfer function, that enables training instance segmentation models on a large set of categories all of which have box annotations, but only a small fraction of which have mask annotations. These contributions allow us to train Mask R-CNN to detect and segment 3000 visual concepts using box annotations from the Visual Genome dataset and mask annotations from the 80 classes in the COCO dataset. We evaluate our approach in a controlled study on the COCO dataset. This work is a first step towards instance segmentation models that have broad comprehension of the visual world.

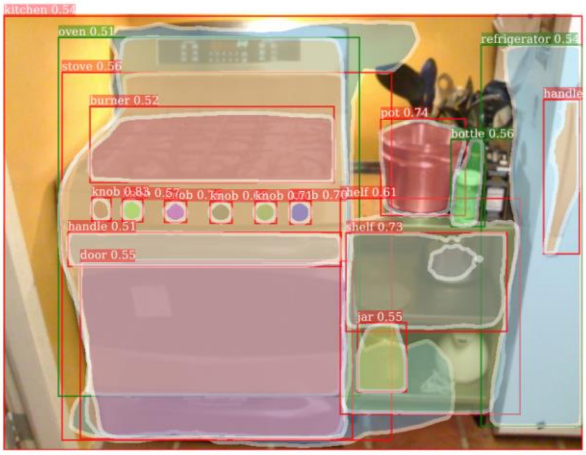


Figure 1. **We explore training instance segmentation models with partial supervision**: a subset of classes (green boxes) have instance mask annotations during training; the remaining classes (red boxes) have only bounding box annotations. This image shows output from our model trained for 3000 classes from Visual Genome, using mask annotations from only 80 classes in COCO.

Source 8 (Hu et al., 2018)

Why we think it's interesting:

High-quality instance segmentation models require heavy supervision and large sets of instance segmentation annotations which makes it time and cost inefficient. The paper addresses this issue by proposing a new partially supervised training framework that exploits bounding box annotations for a large group of categories, in conjunction with instance mask annotations across a small set of categories to maximize the scaling of state-of-the-art instance segmentation models. Interestingly, it is shown that the proposed approaches manage to learn 3000 visual concepts using box annotations from the Visual Genome dataset and mask annotations from the 80 classes in the COCO dataset.

- Who Let The Dogs Out? Modeling Dog Behavior From Visual Data

Ehsani et al., 2018

Abstract:
We study the task of directly modelling a visually intelligent agent. Computer vision typically focuses on solving various subtasks related to visual intelligence. We depart from this standard approach to computer vision; instead we directly model a visually intelligent agent. Our model takes visual information as input and directly predicts the actions of the agent. Toward this end we introduce DECADE, a dataset of ego-centric videos from a dog's perspective as well as her corresponding movements. Using this data we model how the dog acts and how the dog plans her movements. We show under a variety of metrics that given just visual input we can successfully model this intelligent agent in many situations. Moreover, the representation learned by our model encodes distinct information compared to representations trained on image classification, and our learned representation can generalize to other domains. In particular, we show strong results on the task of walkable surface estimation and scene classification by using this dog modelling task as representation learning. Code is available at https://github.com/ehsanik/dogTorch.

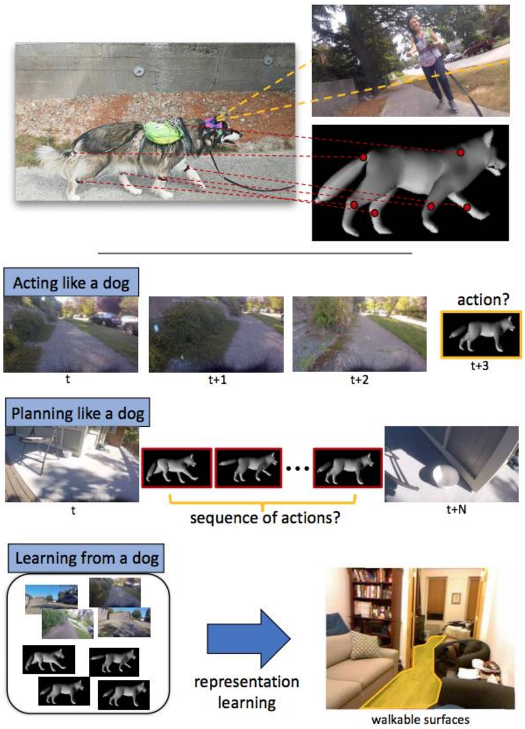


Figure 1. We address three problems: (1) *Acting like a dog*: where the goal is to predict the future movements of the dog given a sequence of previously seen images. (2) *Planning like a dog*: where the goal is to find a sequence of actions that move the dog between the locations of the given pair of images. (3) *Learning from a dog*: where we use the learned representation for a third task (e.g., walkable surface estimation).

Source 9 (Ehsani et al., 2018)

Why we think it's interesting:
The paper addresses the shortcoming of current computer vision agents and how they only solve subtasks related to visual intelligence. The authors bring computer vision closer to visual intelligence by modeling of a visually intelligent agent (i.e., a dog) that can understand visual information and act and perform tasks within its visual environment. Overall, it is very exciting to see that AI is able to replicate certain animal achievements and that as these models continually improve, we will see other types of animal visual recognition capabilities replicated as well.

- **Training Deep Networks with Synthetic Data: Bridging the Reality Gap by Domain Randomization**

Tremblay et al., 2018

Abstract:
We present a system for training deep neural networks for object detection using synthetic images. To handle the variability in real-world data, the system relies upon the technique of domain randomization, in which the parameters of the simulator—such as lighting, pose, object textures, etc.—are randomized in non-realistic ways to force the neural network to learn the essential features of the object of interest. We explore the importance of these parameters, showing that it is possible to produce a network with compelling performance using only non-artistically-generated synthetic data. With additional fine-tuning on real data, the network yields better performance than using real data alone. This result opens up the possibility of using inexpensive synthetic data for training neural networks while avoiding the need to collect large amounts of hand-annotated real-world data or to generate high-fidelity synthetic worlds—both of which remain bottlenecks for many applications. The approach is evaluated on bounding box detection of cars on the KITTI dataset.

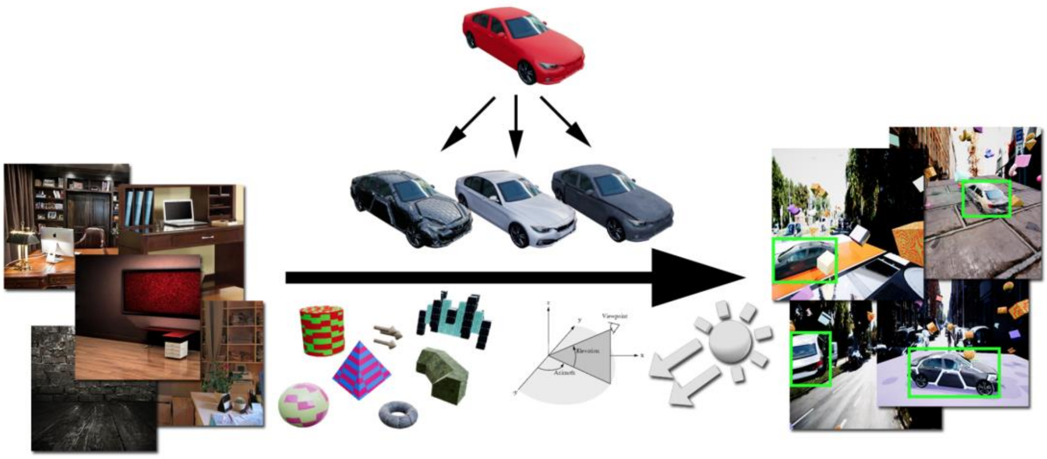


Figure 1. Domain randomization for object detection. Synthetic objects (in this case cars, top-center) are rendered on top of a random background (left) along with random flying distractors (geometric shapes next to the background images) in a scene with random lighting from random viewpoints. Before rendering, random texture is applied to the objects of interest as well as to the flying distractors. The resulting images, along with automatically-generated ground truth (right), are used for training a deep neural network.

Source 10 (Tremblay et al., 2018)

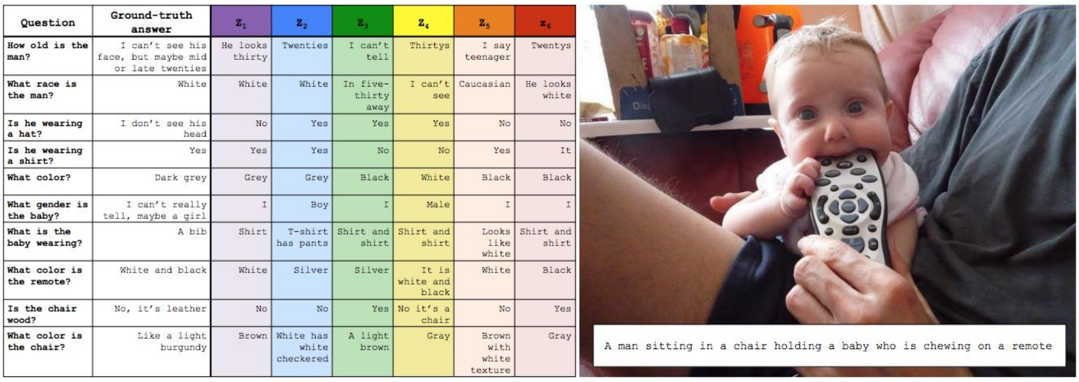
Why we think it’s interesting:
The paper proposes a refined approach for training deep neural networks data for real objects detection relying on domain randomization of synthetic. Domain randomization reduces the need for high-quality simulated datasets by intentionally and randomly disturbing the environment’s textures to force the network to focus and identify the main features of the object. To augment the process’ performance, additional training on real data in conjunction with synthetic data is performed, which bridges the reality gap, and therefore yielding better performance results. Different approaches were proposed to exploit the potentials of synthetic data, which makes it exciting to see how this area will further advance.

- **FLIPDIAL: A Generative Model for Two-Way Visual Dialogue**

Massiceti et al., 2018

Abstract:

We present FlipDial, a generative model for visual dialogue that simultaneously plays the role of both participants in a visually-grounded dialogue. Given context in the form of an image and an associated caption summarising the contents of the image, FlipDial learns both to answer questions and put forward questions, capable of generating entire sequences of dialogue (question-answer pairs) which are diverse and relevant to the image. To do this, FlipDial relies on a simple but surprisingly powerful idea: it uses convolutional neural networks (CNNs) to encode entire dialogues directly, implicitly capturing dialogue context, and conditional VAEs to learn the generative model. FlipDial outperforms the state-of-the-art model in the sequential answering task (one-way visual dialogue) on the VisDial dataset by 5 points in Mean Rank using the generated answers. We are the first to extend this paradigm to full two-way visual dialogue, where our model is capable of generating both questions and answers in sequence based on a visual input, for which we propose a set of novel evaluation measures and metrics.



Question	Ground-truth answer	z_1	z_2	z_3	z_4	z_5	z_6
How old is the man?	I can’t see his face, but maybe mid or late twenties	He looks thirty	Twenties	I can’t tell	Thirtys	I say teenager	Twentys
What race is the man?	White	White	White	In five-thirty away	I can’t see	Caucasian	He looks white
Is he wearing a hat?	I don’t see his head	No	Yes	Yes	Yes	No	No
Is he wearing a shirt?	Yes	Yes	Yes	No	No	Yes	It
What color?	Dark grey	Grey	Grey	Black	White	Black	Black
What gender is the baby?	I can’t really tell, maybe a girl	I	Boy	I	Male	I	I
What is the baby wearing?	A bib	Shirt	T-shirt has pants	Shirt and shirt	Shirt and shirt	Looks like white	Shirt and shirt
What color is the remote?	White and black	White	Silver	Silver	It is white and black	White	Black
Is the chair wood?	No, it’s leather	No	No	Yes	No it’s a chair	No	Yes
What color is the chair?	Like a light burgundy	Brown	White has white checkered	A light brown	Gray	Brown with white texture	Gray

A man sitting in a chair holding a baby who is chewing on a remote

Figure 1: Diverse answers generated by FLIPDIAL in the one-way visual dialogue (1VD) task. For a given time step (row), each column shows a *generated* answer to the current question. Answers are obtained by decoding a latent z_i sampled from the conditional prior – with conditions being the image, caption and dialogue history up until that time step.

Source 11 (Massiceti et al., 2018)

Why we think it’s interesting:
The proposed model (FlipDial) outperforms existing state-of-the art models for the one-way visual dialogue by defining a generative model with CNNs to encode entire dialogues. Furthermore, FlipDial is capable of generating both questions and answers based on visual context; setting a new benchmark for the two-way visual dialogue. The model also employs new evaluation metrics that are more similar to real life situations, through which the quality of generated dialogues (i.e. accuracy of answers, relevancy of questions, etc.) are assessed via the predicted dialogue history and not ground truth dialogue history.

- **AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks**

Xu et al., 2017

Abstract:

In this paper, we propose an Attentional Generative Adversarial Network (AttnGAN) that allows attention-driven, multi-stage refinement for fine-grained text-to-image generation. With a novel attentional generative network, the AttnGAN can synthesize fine-grained details at different subregions of the image by paying attentions to the relevant words in the natural language description. In addition, a deep attentional multimodal similarity model is proposed to compute a fine-grained image-text matching loss for training the generator. The proposed AttnGAN significantly outperforms the previous state of the art, boosting the best reported inception score by 14.14% on the CUB dataset and 170.25% on the more challenging COCO dataset. A detailed analysis is also performed by visualizing the attention layers of the AttnGAN. It for the first time shows that the layered attentional GAN is able to automatically select the condition at the word level for generating different parts of the image.

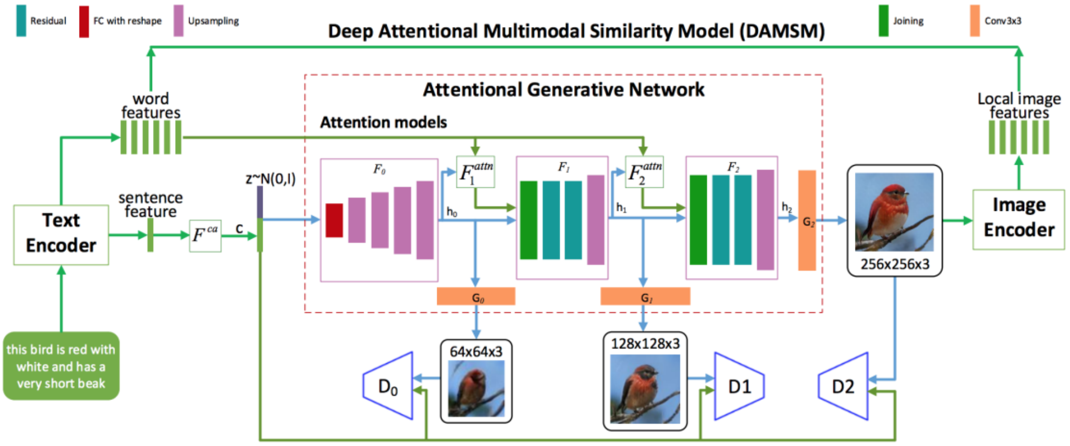


Figure 2. The architecture of the proposed AttnGAN. Each attention model automatically retrieves the conditions (*i.e.*, the most relevant word vectors) for generating different sub-regions of the image; the DAMSM provides the fine-grained image-text matching loss for the generative network.

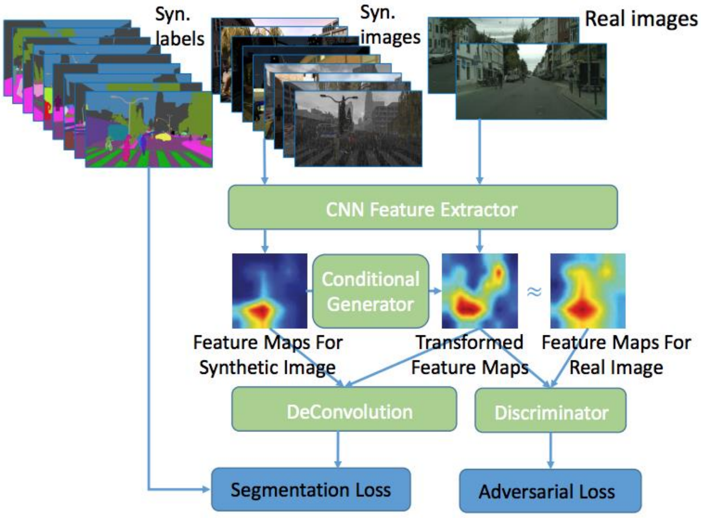
Source 12 (Xu et al., 2017)

Why we think it's interesting:
Current state-of-the-art GAN-based text-to-image generation models encode textual descriptions only on the sentence level and overlook fine-grained information on the word-level that would improve the quality of generated images. AttnGAN proposes a novel word-level attention mechanism which is far more impressive in producing complex scenes.

- **Conditional Generative Adversarial Network for Structured Domain Adaptation**

Hong et al., 2018

Abstract:
In recent years, deep neural nets have triumphed over many computer vision problems, including semantic segmentation, which is a critical task in emerging autonomous driving and medical image diagnostics applications. In general, training deep neural nets requires a humongous amount of labeled data, which is laborious and costly to collect and annotate. Recent advances in computer graphics shed light on utilizing photo-realistic synthetic data with computer generated annotations to train neural nets. Nevertheless, the domain mismatch between real images and synthetic ones is the major challenge against harnessing the generated data and labels. In this paper, we propose a principled way to conduct structured domain adaption for semantic segmentation, i.e., integrating GAN into the FCN framework to mitigate the gap between source and target domains. Specifically, we learn a conditional generator to transform features of synthetic images to real-image like features, and a discriminator to distinguish them. For each training batch, the conditional generator and the discriminator compete against each other so that the generator learns to produce real-image like features to fool the discriminator; afterwards, the FCN parameters are updated to accommodate the changes of GAN. In experiments, without using labels of real image data, our method significantly outperforms the baselines as well as state-of-the-art methods by 12% ~ 20% mean IoU on the Cityscapes dataset.



Source 13 (Hong et al., 2018)

Why we think it’s interesting:
Recent research investigates new methods to capitalize on synthetic datasets as a cost-efficient and scalable alternative in training deep neural networks, but the cross-domain gap remains a major challenge. This paper offers a new method to overcome the challenges of cross-domain differences in semantic segmentation models with a structured domain adaptation method. Unlike unsupervised domain adaptation, the method does not assume the existence of a cross-domain common feature space and rather employs a conditional generator and a discriminator. Therefore, a conditional GAN is integrated into the CNN framework that transfers features of synthetic images to real-image like features. The method results outperform previous models; highlighting the growing potential of synthetic datasets in advancing vision tasks.

- **On the Importance of Label Quality for Semantic Segmentation**

Zlateski et al., 2018

Abstract:
Convolutional networks (ConvNets) have become the dominant approach to semantic image segmentation. Producing accurate, pixel--level labels required for this task is a tedious and time consuming process; however, producing approximate, coarse labels could take only a fraction of the time and effort. We investigate the relationship between the quality of labels and the performance of ConvNets for semantic segmentation. We create a very large synthetic dataset with perfectly labeled street view scenes. From these perfect labels, we synthetically coarsen labels with different qualities and estimate human--hours required for producing them. We perform a series of experiments by training ConvNets with a varying number of training images and label quality. We found that the performance of ConvNets mostly depends on the time spent creating the training labels. That is, a larger coarsely--annotated dataset can yield the same performance as a smaller finely--annotated one. Furthermore, fine--tuning coarsely pre--trained ConvNets with few finely-annotated labels can yield comparable or superior performance to training it with a large amount of finely-annotated labels alone, at a fraction of the labeling cost. We demonstrate that our result is also valid for different network architectures, and various object classes in an urban scene.

Why we think it’s interesting:
This paper also investigates new methods and factors that influence the performance of ConvNets for semantic segmentation without relying on large sets of refined annotated labels. The authors proved that the amount of time used in producing labels is more crucial than the labels’ quality. Furthermore, training the network on a large set of coarsely annotated labels, followed by a fine-tuning phase using a small set of finely-annotated labels actually outperforms ConvNets solely trained on a large set of finely-annotated labels. This work is particularly influential as it shows that quality is irrelevant to superior performance, which would make the training process more cost and labor efficient.

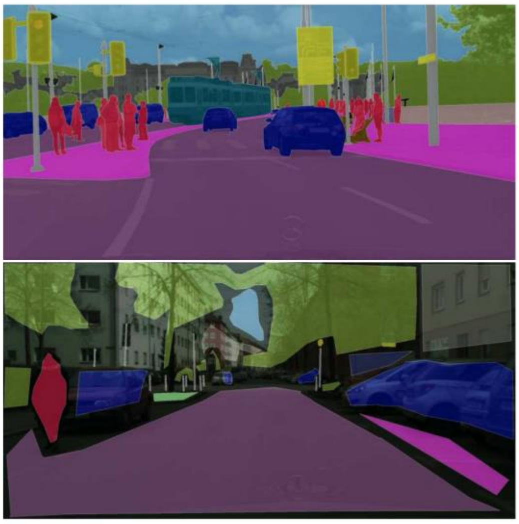


Figure 1: A finely annotated (top) and a corsely annotated (bottom) image from the CityScape's dataset.

Source 14 (Zlateski et al., 2018)

- **IQA: Visual Question Answering in Interactive Environments**

Gordon et al., 2018

Abstract:
We introduce Interactive Question Answering (IQA), the task of answering questions that require an autonomous agent to interact with a dynamic visual environment. IQA presents the agent with a scene and a question, like: "Are there any apples in the fridge?" The agent must navigate around the scene, acquire visual understanding of scene elements, interact with objects (e.g. open refrigerators) and plan for a series of actions conditioned on the question. Popular reinforcement learning approaches with a single controller perform poorly on IQA owing to the large and diverse state space. We propose the Hierarchical Interactive Memory Network (HIMN), consisting of a factorized set of controllers, allowing the system to operate at multiple levels of temporal abstraction. To evaluate HIMN, we introduce IQUAD V1, a new dataset built upon AI2- THOR, a simulated photo-realistic environment of configurable indoor scenes with interactive objects. IQUAD V1 has 75,000 questions, each paired with a unique scene configuration. Our experiments show that our proposed model outperforms popular single controller based methods on IQUAD V1. For sample questions and results, please view our video: https://youtu.be/pXd3C-1jr98.

Why we think it's interesting:
This paper addresses one of the shortcomings of standard VQA models, which are mostly passive and do not train a fully intelligent agent; capable of navigating, interacting and performing tasks within its environment. To fill this research gap, the paper proposes Interactive Question Answering (IQA) that uses a multi-level controller method with a semantic spatial memory and collects a rich dataset of simulated realistic scenes and a wide range of questions to evaluate the model. The proposed model advances standard VQA towards the ultimate goal of creating fully visually intelligent agents.

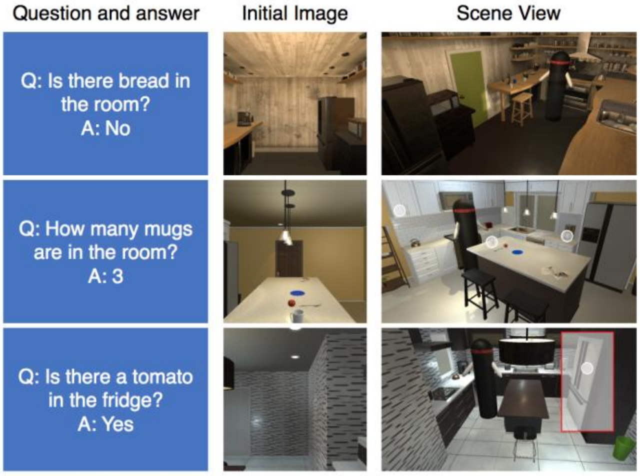

Question and answer
Initial Image
Scene View
Q: Is there bread in the room? A: No
Q: How many mugs are in the room? A: 3
Q: Is there a tomato in the fridge? A: Yes

Figure 1. Samples from IQUAD V1: Each row shows a question paired with the agent's initial view and a scene view of the environment (which is not provided to the agent). In the scene view, the agent is shown in black, and the locations of the objects of interest for each question are outlined. Note that none of the questions can be answered accurately given only the initial image.

Source 15 (Gordon et al., 2018)

- **Guide Me: Interacting with Deep Networks**

Rupprecht et al., 2018

Abstract:

Interaction and collaboration between humans and intelligent machines has become increasingly important as machine learning methods move into real-world applications that involve end users. While much prior work lies at the intersection of natural language and vision, such as image captioning or image generation from text descriptions, less focus has been placed on the use of language to guide or improve the performance of a learned visual processing algorithm. In this paper, we explore methods to flexibly guide a trained convolutional neural network through user input to improve its performance during inference. We do so by inserting a layer that acts as a spatio-semantic guide into the network. This guide is trained to modify the network's activations, either directly via an energy minimization scheme or indirectly through a recurrent model that translates human language queries to interaction weights. Learning the verbal interaction is fully automatic and does not require manual text annotations. We evaluate the method on two datasets, showing that guiding a pre-trained network can improve performance, and provide extensive insights into the interaction between the guide and the CNN.

Why we think it's interesting:

The paper proposes an original approach to enhance the performance of a pre-trained convolutional neural network (CNN) through employing an additional function to the network, called "spatio-semantic guide". This guide enables an interactive dialogue between a human user and the CNN and translates the human's feedback into actual changes in the network's activations. By facilitating simultaneous user feedback, the network can adjust its deductions on the spot without the need for additional training of the network's parameters.

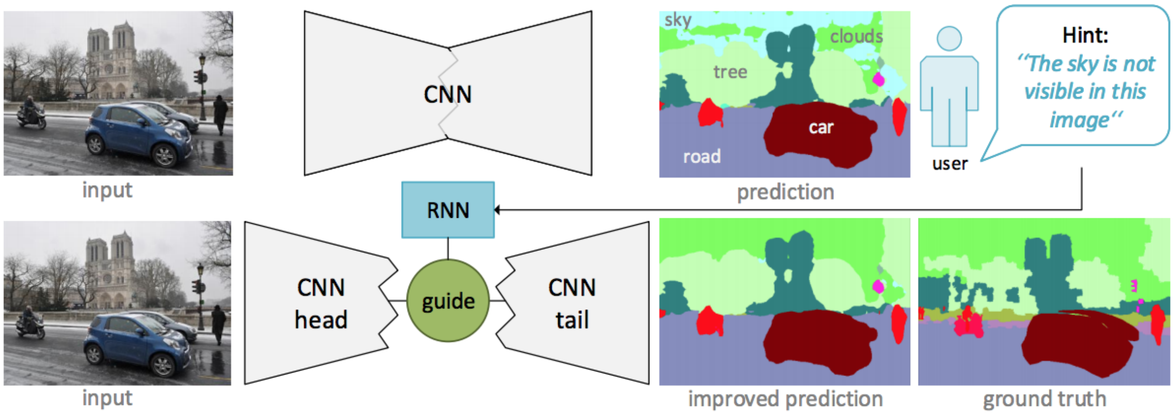


Figure 1. **Overview.** We introduce a system that is able to refine predictions of a CNN by injecting a guiding block into the network. The guiding can be performed using natural language through an RNN to process the text. In this example, the original network had difficulties to differentiate between the `sky` and the `cloud` classes. The user indicates that there is no sky and the prediction is updated, without any CNN weight updates and thus without additional training.

Source 16 (Rupprecht et al., 2018)

- Don’t Just Assume; Look and Answer: Overcoming Priors for Visual Question Answering

Agrawal et al., 2018

Abstract:
A number of studies have found that today’s Visual Question Answering (VQA) models are heavily driven by superficial correlations in the training data and lack sufficient image grounding. To encourage development of models geared towards the latter, we propose a new setting for VQA where for every question type, train and test sets have different prior distributions of answers. Specifically, we present new splits of the VQA v1 and VQA v2 datasets, which we call Visual Question Answering under Changing Priors (VQACP v1 and VQA-CP v2 respectively). First, we evaluate several existing VQA models under this new setting and show that their performance degrades significantly compared to the original VQA setting. Second, we propose a novel Grounded Visual Question Answering model (GVQA) that contains inductive biases and restrictions in the architecture specifically designed to prevent the model from ‘cheating’ by primarily relying on priors in the training data. Specifically, GVQA explicitly disentangles the recognition of visual concepts present in the image from the identification of plausible answer space for a given question, enabling the model to more robustly generalize across different distributions of answers. GVQA is built off an existing VQA model – Stacked Attention Networks (SAN). Our experiments demonstrate that GVQA significantly outperforms SAN on both VQA-CP v1 and VQA-CP v2 datasets. Interestingly, it also outperforms more powerful VQA models such as Multimodal Compact Bilinear Pooling (MCB) in several cases. GVQA offers strengths complementary to SAN when trained and evaluated on the original VQA v1 and VQA v2 datasets. Finally, GVQA is more transparent and interpretable than existing VQA models.

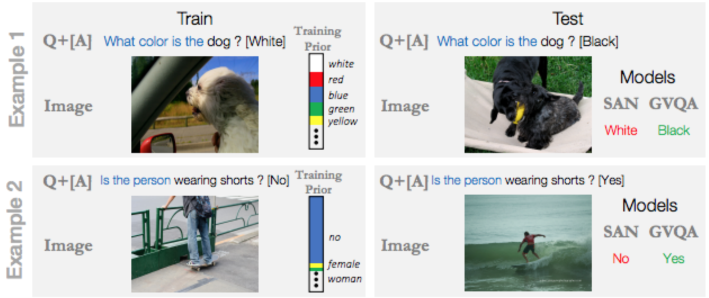


Figure 1: Existing VQA models, such as SAN [39], tend to largely rely on strong language priors in train sets, such as, the prior answer (*'white'*, *'no'*) given the question type (*'what color is the'*, *'is the person'*). Hence, they suffer significant performance degradation on test image-question pairs whose answers (*'black'*, *'yes'*) are not amongst the majority answers in train. We propose a novel model (GVQA), built off of SAN that explicitly grounds visual concepts in images, and consequently significantly outperforms SAN in a setting with mismatched priors between train and test.

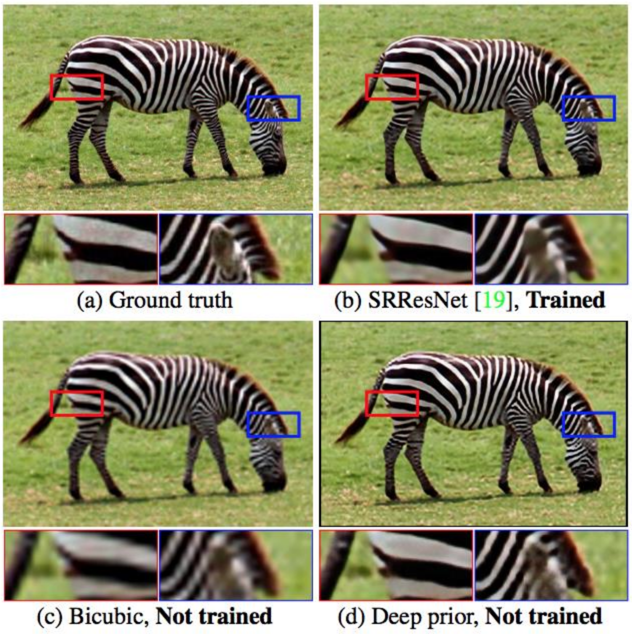
Source 17 (Agrawal et al., 2018)

Why we think it's interesting:
Effectively evaluating the performance of current state-of-the-art VQA models and prevent them from relying on biased training priors is an area that's still under development. To that end, Grounded Visual Question Answering model (GVQA) offers a new method that directly dissociates the objects recognized from plausible prior answers, forcing the model to be more visually grounded. With the excellent results the paper has reported and the current community focus on this line of research, it's promising to expect future innovative methods further advancing VQA models.

- Deep Image Prior

Ulyanov et al., 2018

Abstract: Deep convolutional networks have become a popular tool for image generation and restoration. Generally, their excellent performance is imputed to their ability to learn realistic image priors from a large number of example images. In this paper, we show that, on the contrary, the structure of a generator network is sufficient to capture a great deal of low-level image statistics prior to any learning. In order to do so, we show that a randomly-initialized neural network can be used as a handcrafted prior with excellent results in standard inverse problems such as denoising, super-resolution, and inpainting. Furthermore, the same prior can be used to invert deep neural representations to diagnose them, and to restore images based on flash-no flash input pairs.
Apart from its diverse applications, our approach highlights the inductive bias captured by standard generator network architectures. It also bridges the gap between two very popular families of image restoration methods: learning-based methods using deep convolutional networks and learning-free methods based on handcrafted image priors such as self-similarity.



(a) Ground truth

(b) SRResNet [19], **Trained**

(c) Bicubic, **Not trained**

(d) Deep prior, **Not trained**

Figure 1: **Super-resolution using the deep image prior.** Our method uses a randomly-initialized ConvNet to upsample an image, using its structure as an image prior; similar to bicubic upsampling, this method does not require learning, but produces much cleaner results with sharper edges. In fact, our results are quite close to state-of-the-art super-resolution methods that use ConvNets learned from large datasets. The deep image prior works well for all inverse problems we could test.

Source 18 (Ulyanov et al., 2018)

Why we think it's interesting:
In contrast to the common belief that the success of neural networks mainly comes from their strong ability to learn from data, this paper demonstrates the importance of the structure of the network for building good image priors. The paper proposes a decoder network as prior for imaging tasks. Interestingly enough, the authors show that a generator network is adequate to capture a large amount of low-level image statistics prior to any learning. Specifically, in this approach, the neural network is interpreted as a parametrization of the image. It is shown that fitting the weights to one visually degraded image alone is enough to obtain a rich enough network (image representation) that can serve as a generic tool for tasks such as denoise, image restoration etc. by using the learned prior as regularizer indicator function, i.e. indicator function of images that can be produced from a random noise vector by a deep convolutional net of a certain architecture. The authors also use the approach to investigate the information content retained at different levels of the network by producing so-called natural pre-images, i.e. images that map to the same latent representation. Intriguingly, using the deep image prior as a regularizer, the pre-image obtained from even very deep layers still captures a large amount of information.

- **Matching Adversarial Networks**

Máttyus and Urtasun, 2018

Abstract:

Generative Adversarial Nets (GANs) and Conditonal GANs (CGANs) show that using a trained network as loss function (discriminator) enables to synthesize highly structured outputs (e.g. natural images). However, applying a discriminator network as a universal loss function for common supervised tasks (e.g. semantic segmentation, line detection, depth estimation) is considerably less successful. We argue that the main difficulty of applying CGANs to supervised tasks is that the generator training consists of optimizing a loss function that does not depend directly on the ground truth labels. To overcome this, we propose to replace the discriminator with a matching network taking into account both the ground truth outputs as well as the generated examples. As a consequence, the generator loss function also depends on the targets of the training examples, thus facilitating learning. We demonstrate on three computer vision tasks that this approach can significantly outperform CGANs achieving comparable or superior results to task-specific solutions and results in stable training. Importantly, this is a general approach that does not require the use of task-specific loss functions.

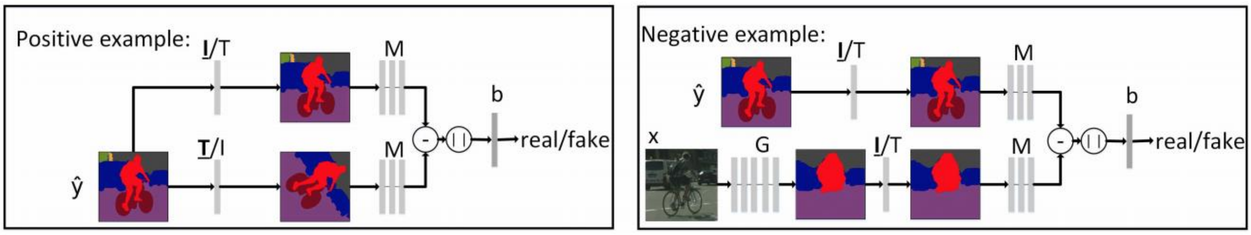


Figure 1: Our MatAN discriminator is a siamese network: (left) positive examples, (right) negative ones. The input to the siamese network is passed through a perturbation T or through the identity transformation I. The configurations of T and I result in different training behavior. The drawing shows the case when the perturbation is only applied to one branch of the positive samples.

Source 19 (Máttyus and Urtasun, 2018)

Why we think it’s interesting:

Despite the success of GANs, no considerable success has been reported on the usage of their discriminator network as a universal loss function for common supervised tasks such as semantic segmentation. The paper highlights the reason behind this, namely that the loss function does not directly depend on the ground truth labels during generator training, which leads to random production of samples from the data distribution without correlating the input-output relations in a supervised fashion. To overcome this, the paper proposes replacing the discriminator with a matching network taking into account both the ground truth outputs as well as the generated examples, which is facilitated by a Siamese network architecture.

- **iVQA: Inverse Visual Question Answering**

Liu et al, 2018

Abstract:

We propose the inverse problem of Visual question answering (iVQA), and explore its suitability as a benchmark for visuo-linguistic understanding. The iVQA task is to generate a question that corresponds to a given image and answer pair. Since the answers are less informative than the questions, and the questions have less learnable bias, an iVQA model needs to better understand the image to be successful than a VQA model. We pose question generation as a multi-modal dynamic inference process and propose an iVQA model that can gradually adjust its focus of attention guided by both a partially generated question and the answer. For evaluation, apart from existing linguistic metrics, we propose a new ranking metric. This metric compares the ground truth question's rank among a list of distractors, which allows the drawbacks of different algorithms and sources of error to be studied. Experimental results show that our model can generate diverse, grammatically correct and content correlated questions that match the given answer.

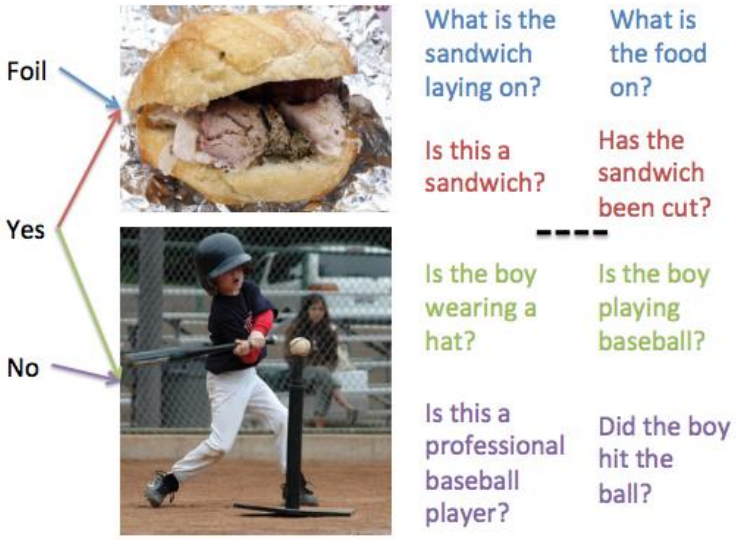

Foil
Yes
No
What is the sandwich laying on?
What is the food on?
Is this a sandwich?
Has the sandwich been cut?
Is the boy wearing a hat?
Is the boy playing baseball?
Is this a professional baseball player?
Did the boy hit the ball?

Figure 1. Illustration of iVQA task: Input answers and images along with the top questions generated by our model.

Source 20 (Liu et al, 2018)

Why we think it's interesting:

iVQA joins the other models that aim to improve the performance of standard VQA models by focusing on developing visual grounding. „This paper inverses the popular VQA task, such that the target is to generate a question given an image/answer pair"The learning biases of standard VQAs undermine the evaluation process. iVQA uses partially generated questions with less biased learning priors corresponding to an image-answer pair to achieve more visual grounding.

- **Classifier Discrepancy for Unsupervised Domain Adaptation**

Saito et al., 2018

Abstract:
In this work, we present a method for unsupervised domain adaptation. Many adversarial learning methods train domain classifier networks to distinguish the features as either a source or target and train a feature generator network to mimic the discriminator. Two problems exist with these methods. First, the domain classifier only tries to distinguish the features as a source or target and thus does not consider task-specific decision boundaries between classes. Therefore, a trained generator can generate ambiguous features near class boundaries. Second, these methods aim to completely match the feature distributions between different domains, which is difficult because of each domain's characteristics. To solve these problems, we introduce a new approach that attempts to align distributions of source and target by utilizing the task-specific decision boundaries. We propose to maximize the discrepancy between two classifiers' outputs to detect target samples that are far from the support of the source. A feature generator learns to generate target features near the support to minimize the discrepancy. Our method outperforms other methods on several datasets of image classification and semantic segmentation. The codes are available at https://github. com/mil-tokyo/MCD_DA.

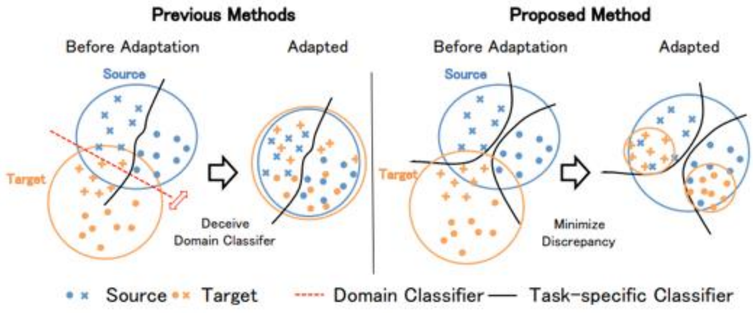


Figure 1. (Best viewed in color.) Comparison of previous and the proposed distribution matching methods.. **Left**: Previous methods try to match different distributions by mimicing the domain classifier. They do not consider the decision boundary. **Right**: Our proposed method attempts to detect target samples outside the support of the source distribution using task-specific classifiers.

Source 21 (Saito et al., 2018)

Why we think it's interesting:
Training deep learning-based models relies on large annotated datasets, which requires lots of resources. Despite achieving state-of-the-art performance in many visual recognition tasks, cross-domain differences still constitute a big challenge. To transfer knowledge across domains, the proposed approach uses a novel adversarial learning method for domain adaptation without a need for any labeling information from the target domain. The authors observe that, along with adversarial training, minimizing the discrepancy between the probability estimates from two classifiers for samples from a target domain can produce class-discriminative features for various tasks from classification to semantic segmentation.

- **Other Interesting Papers:**

- Look, Imagine and Match: Improving Textual-Visual Cross-Modal Retrieval with Generative Models
- Low-Shot Learning with Imprinted Weights
- Generative Adversarial Perturbations
- Transparency by Design: Closing the Gap Between Performance and Interpretability in Visual Reasoning
- Efficient Optimization for Rank-based Loss Functions
- Deep Layer Aggregation
- Neural Baby Talk
- Few-Shot Image Recognition by Predicting Parameters from Activations
- Iterative Visual Reasoning Beyond Convolutions
- Low-Shot Learning from Imaginary Data
- Differential Attention for Visual Question Answering
- VirtualHome: Simulating Household Activities via Programs
- Deep Unsupervised Saliency Detection: A Multiple Noisy Labeling Perspective
- Context Encoding for Semantic Segmentation
- Practical Block-wise Neural Network Architecture Generation
- Defense against Universal Adversarial Perturbations
- Maximum Classifier Discrepancy for Unsupervised Domain Adaptation
- Learning from Synthetic Data: Addressing Domain Shift for Semantic Segmentation
- Connecting Pixels to Privacy and Utility: Automatic Redaction of Private Information in Images
- Are You Talking to Me? Reasoned Visual Dialog Generation through Adversarial Learning
- Learning Answer Embeddings for Visual Question Answering
- Focal Visual-Text Attention for Visual Question Answering
- Unsupervised Textual Grounding: Linking Words to Image Concepts
- Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering
- Fooling Vision and Language Models Despite Localization and Attention Mechanism
- Tips and Tricks for Visual Question Answering: Learnings from the 2017 Challenge
- Towards Human-Machine Cooperation: Self-supervised Sample Mining for Object Detection
- Interpretable Convolutional Neural Networks
- Visual Question Generation as Dual Task of Visual Question Answering
- Data Distillation: Towards Omni-Supervised Learning
- Creating Capsule Wardrobes from Fashion Images
- Boosting Self-Supervised Learning via Knowledge Transfer
- Deep Mutual Learning

Bridging the Gap between Theory and Application

In addition to the wide range of academic and technical research presented at the conference, the industrial exhibition this year has also witnessed substantial growth. Alongside the research, several companies showcased their newest industrial innovations; from self-driving cars and robotics to a plethora of other solutions employing machine learning, 3D vision, virtual reality, video analytics, and more. With its continuous growth and success in bridging the gap between theory and application, CVPR continues to push the frontiers of computer vision.

www.sap.com/contactsap

The information contained herein may be changed without prior notice. Some software products marketed by SAP SE and its distributors contain proprietary software components of other software vendors. National product specifications may vary.

These materials are provided by SAP SE or an SAP affiliate company for informational purposes only, without representation or warranty of any kind, and SAP or its affiliated companies shall not be liable for errors or omissions with respect to the materials. The only warranties for SAP or SAP affiliate company products and services are those that are set forth in the express warranty statements accompanying such products and services, if any. Nothing herein should be construed as constituting an additional warranty.

In particular, SAP SE or its affiliated companies have no obligation to pursue any course of business outlined in this document or any related presentation, or to develop or release any functionality mentioned therein. This document, or any related presentation, and SAP SE's or its affiliated companies' strategy and possible future developments, products, and/or platform directions and functionality are all subject to change and may be changed by SAP SE or its affiliated companies at any time for any reason without notice. The information in this document is not a commitment, promise, or legal obligation to deliver any material, code, or functionality. All forward-looking statements are subject to various risks and uncertainties that could cause actual results to differ materially from expectations. Readers are cautioned not to place undue reliance on these forward-looking statements, and they should not be relied upon in making purchasing decisions.

THE BEST RUN SAP®