A Year in Computer Vision

The M Tank

Website: http://themtank.org/

Contact: info@themtank.com

Note: This document is intended for educational purposes only. Any information contained within is representative of the editors professional views. This piece contains a number of academic publications for which references are provided where appropriate.

Edited for The M Tank by

Benjamin F. Duffy
&
Daniel R. Flynn

Table of Contents

Introduction

Computer Vision typically refers to the scientific discipline of giving machines the ability of sight, or perhaps more colourfully, enabling machines to visually analyse their environments and the stimuli within them. This process typically involves the evaluation of an image, images or video. The British Machine Vision Association (BMVA) defines Computer Vision as "*the automatic extraction, analysis and understanding of useful information from a single image or a sequence of images.*"[1]

The term *understanding* provides an interesting counterpoint to an otherwise mechanical definition of vision, one which serves to demonstrate both the significance and complexity of the Computer Vision field. True understanding of our environment is not achieved through visual representations alone. Rather, visual cues travel through the optic nerve to the primary visual cortex and are interpreted by the brain, in a highly stylised sense. The interpretations drawn from this sensory information encompass the near-totality of our natural programming and subjective experiences, i.e. how evolution has wired us to survive and what we learn about the world throughout our lives.

In this respect, *vision* only relates to the transmission of images for interpretation; while *computing* said images is more analogous to thought or cognition, drawing on a multitude of the brain's faculties. Hence, many believe that Computer Vision, a true understanding of visual environments and their contexts, paves the way for future iterations of Strong Artificial Intelligence, due to its cross-domain mastery.

However, put down the pitchforks as we're still very much in the embryonic stages of this fascinating field. This piece simply aims to shed some light on 2016's biggest Computer Vision advancements. And hopefully ground some of these advancements in a healthy mix of expected near-term societal-interactions and, where applicable, tongue-in-cheek prognostications of the end of life as we know it.

While our work is always written to be as accessible as possible, sections within this particular piece may be oblique at times due to the subject matter. We do provide rudimentary definitions throughout, however, these only convey a facile understanding of key concepts. In keeping our focus on work produced in 2016, often omissions are made in the interest of brevity.

One such glaring omission relates to the functionality of Convolutional Neural Networks (hereafter CNNs or ConvNets), which are ubiquitous within the field of Computer Vision.

[1] British Machine Vision Association (BMVA). 2016. What is computer vision? *[Online]* Available at: http://www.bmva.org/visionoverview [Accessed 21/12/2016]

The success of AlexNet[2] in 2012, a CNN architecture which blindsided ImageNet competitors, proved instigator of a de facto revolution within the field, with numerous researchers adopting neural network-based approaches as part of Computer Vision's new period of 'normal science'.[3]

Over four years later and CNN variants still make up the bulk of new neural network architectures for vision tasks, with researchers reconstructing them like legos; a working testament to the power of both open source information and Deep Learning. However, an explanation of CNNs could easily span several postings and is best left to those with a deeper expertise on the subject and an affinity for making the complex understandable.

For casual readers who wish to gain a quick grounding before proceeding we recommend the first two resources below. For those who wish to go further still, we have ordered the resources below to facilitate that:

- **What a Deep Neural Network thinks about your #selfie** from Andrej Karpathy is one of our favourites for helping people understand the applications and functionalities behind CNNs.[4]
- **Quora: "what is a convolutional neural network?" -** Has no shortage of great links and explanations. Particularly suited to those with no prior understanding.[5]
- **CS231n: Convolutional Neural Networks for Visual Recognition** from Stanford University is an excellent resource for more depth.[6]
- **Deep Learning** (Goodfellow, Bengio & Courville, 2016) provides detailed explanations of CNN features and functionality in Chapter 9. The textbook has been kindly made available for free in HTML format by the authors.[7]

For those wishing to understand more about Neural Networks and Deep Learning in general we suggest:

---

[2] Krizhevsky, A., Sutskever, I. and Hinton, G. E. 2012. ImageNet Classification with Deep Convolutional Neural Networks, *NIPS 2012: Neural Information Processing Systems*, Lake Tahoe, Nevada. Available: http://www.cs.toronto.edu/~kriz/imagenet_classification_with_deep_convolutional.pdf

[3] Kuhn, T. S. 1962. *The Structure of Scientific Revolutions*. 4th ed. United States: The University of Chicago Press.

[4] Karpathy, A. 2015. What a Deep Neural Network thinks about your #selfie. *[Blog] Andrej Karpathy Blog*. Available: http://karpathy.github.io/2015/10/25/selfie/ [Accessed: 21/12/2016]

[5] Quora. 2016. What is a convolutional neural network? *[Online]* Available: https://www.quora.com/What-is-a-convolutional-neural-network [Accessed: 21/12/2016]

[6] Stanford University. 2016. Convolutional Neural Networks for Visual Recognition. *[Online] CS231n*. Available: http://cs231n.stanford.edu/ [Accessed 21/12/2016]

[7] Goodfellow et al. 2016. Deep Learning. *MIT Press*. *[Online]* http://www.deeplearningbook.org/ [Accessed: 21/12/2016] Note: Chapter 9, Convolutional Networks [Available: http://www.deeplearningbook.org/contents/convnets.html]

- **Neural Networks and Deep Learning** (Nielsen, 2017) is a free online textbook which provides the reader with a really intuitive understanding of the complexities of Neural Networks and Deep Learning. Even just completing chapter one should greatly illuminate the subject matter of this piece for first-timers.[8]

As a whole this piece is disjointed and spasmodic, a reflection of the authors' excitement and the spirit in which it was intended to be utilised, section by section. Information is partitioned using our own heuristics and judgements, a necessary compromise due to the cross-domain influence of much of the work presented.

We hope that readers benefit from our aggregation of the information here to further their own knowledge, regardless of previous experience.

From all our contributors,

The M Tank

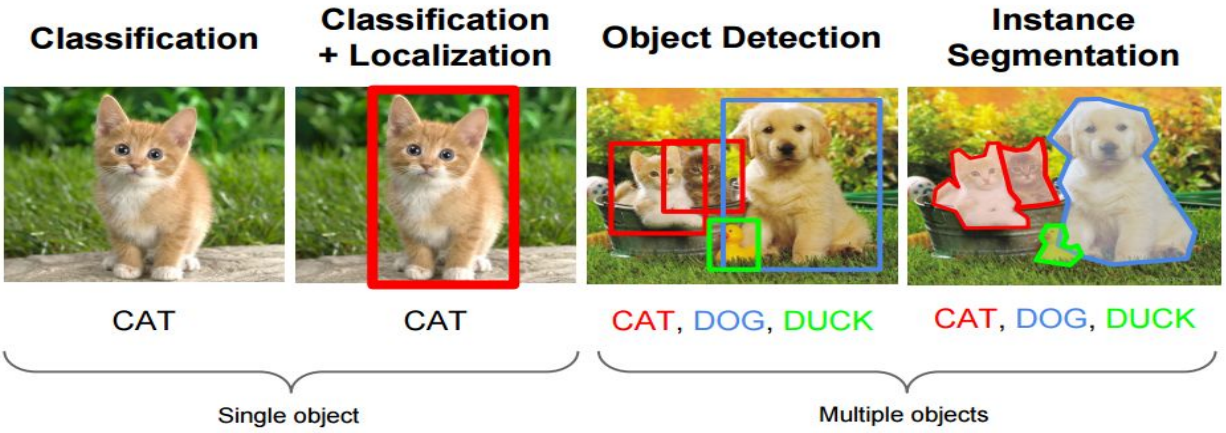
[8] Nielsen, M. 2017. Neural Networks and Deep Learning. *[Online] EBook*. Available: http://neuralnetworksanddeeplearning.com/index.html [Accessed: 06/03/2017].

Part One: Classification/Localisation, Object Detection, Object Tracking

Classification/Localisation

The task of classification, when it relates to images, generally refers to assigning a label to the whole image, e.g. 'cat'. Assuming this, Localisation may then refer to finding where the object is in said image, usually denoted by the output of some form of bounding box around the object. Current classification/localisation techniques on ImageNet[9] have likely surpassed an ensemble of trained humans.[10] For this reason, we place greater emphasis on subsequent sections of the blog.

Figure 1: Computer Vision Tasks



Source: Fei-Fei Li, Andrej Karpathy & Justin Johnson (2016) cs231n, Lecture 8 - Slide 8, *Spatial Localization and Detection* (01/02/2016). Available: http://cs231n.stanford.edu/slides/2016/winter1516_lecture8.pdf

However, the introduction of larger datasets with an increased number of classes[11] will likely provide new metrics for progress in the near future. On that point, François Chollet, the creator of Keras,[12] has applied new techniques, including the popular architecture Xception, to an internal google dataset with over 350 million multi-label images containing 17,000 classes.[13,14]

---

[9] ImageNet refers to a popular image dataset for Computer Vision. Each year entrants compete in a series of different tasks called the ImageNet Large Scale Visual Recognition Challenge (ILSVRC). Available: http://image-net.org/challenges/LSVRC/2016/index

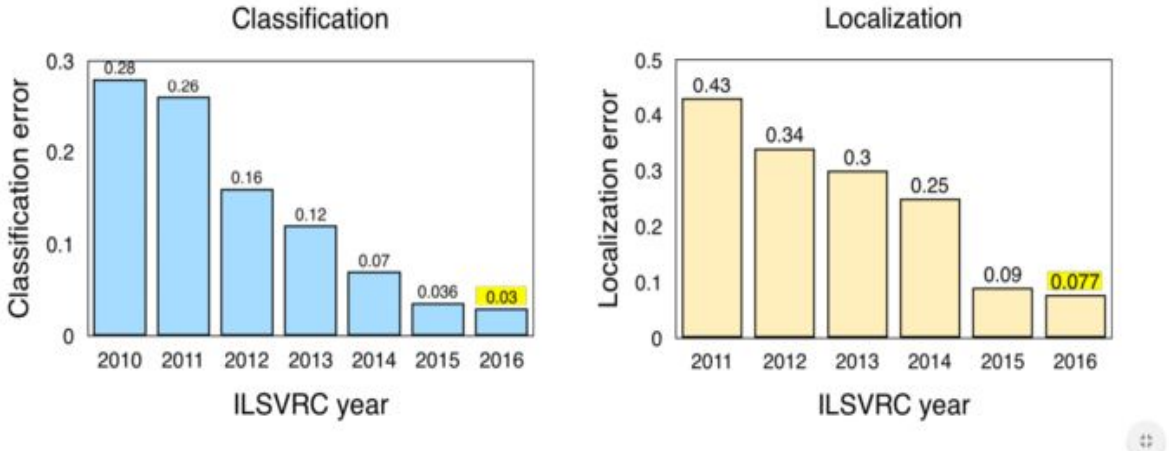
[10] See "*What I learned from competing against a ConvNet on ImageNet*" by Andrej Karpathy. The blog post details the author's journey to provide a human benchmark against the ILSVRC 2014 dataset. The error rate was approximately 5.1% versus a then state-of-the-art GoogLeNet classification error of 6.8%. Available: http://karpathy.github.io/2014/09/02/what-i-learned-from-competing-against-a-convnet-on-imagenet/

[11] See new datasets later in this piece.

[12] Keras is a popular neural network-based deep learning library: https://keras.io/

[13] Chollet, F. 2016. Information-theoretical label embeddings for large-scale image classification. *[Online] arXiv: 1607.05691*. Available: **arXiv:1607.05691v1**

Figure 2: Classification/Localisation results from ILSVRC (2010-2016)



Note: ImageNet Large Scale Visual Recognition Challenge (ILSVRC). The change in results from 2011-2012 resulting from the AlexNet submission. For a review of the challenge requirements relating to Classification and Localization see: http://www.image-net.org/challenges/LSVRC/2016/index#comp
Source: Jia Deng (2016). *ILSVRC2016 object localisation: introduction, results*. Slide 2. Available: http://image-net.org/challenges/talks/2016/ILSVRC2016_10_09_clsloc.pdf

Interesting takeaways from the ImageNet LSVRC (2016):

- **Scene Classification** refers to the task of labelling an image with a certain scene class like 'greenhouse', 'stadium', 'cathedral', etc. ImageNet held a Scene Classification challenge last year with a subset of the Places2[15] dataset: 8 million images for training with 365 scene categories.
  Hikvision[16] won with a 9% top-5 error with an ensemble of deep Inception-style networks, and not-so-deep residuals networks.

- **Trimps-Soushen** won the ImageNet Classification task with 2.99% top-5 classification error and 7.71% localisation error. The team employed an ensemble for classification (averaging the results of Inception, Inception-Resnet, ResNet and Wide Residual Networks models[17]) and Faster R-CNN for localisation based on the labels.[18] The dataset was distributed across 1000 image classes with 1.2 million images provided as training data. The partitioned test data compiled a further 100 thousand unseen images.

---

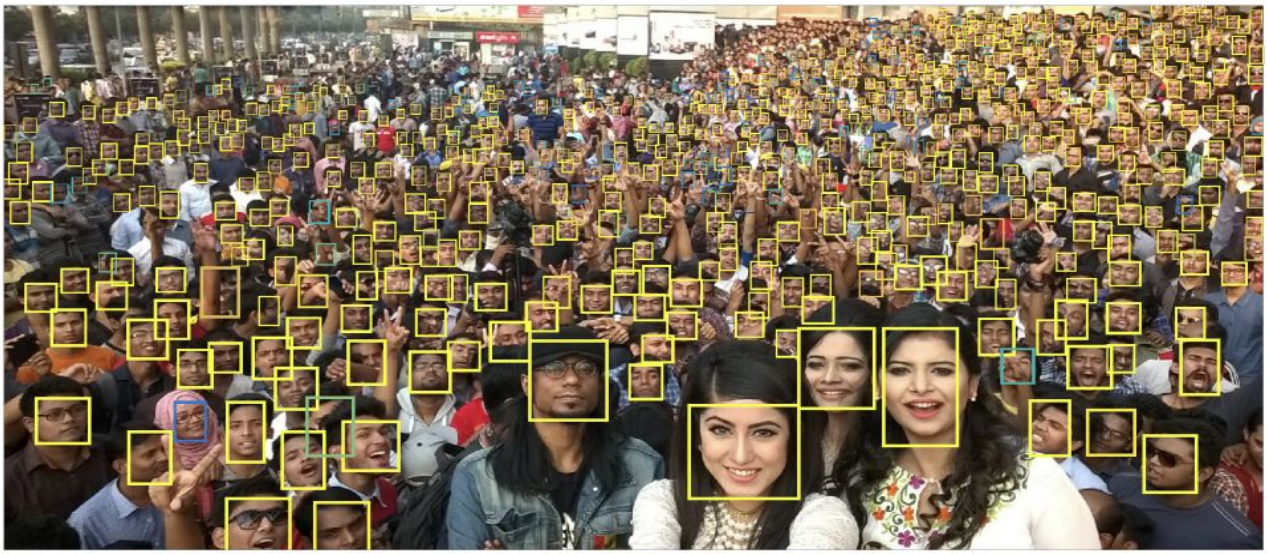
[14] Chollet, F. 2016. Xception: Deep Learning with Depthwise Separable Convolutions. *[Online] arXiv:1610.02357*. Available: **arXiv:1610.02357v2**
[15] Places2 dataset, details available: http://places2.csail.mit.edu/. See also new datasets section.
[16] Hikvision. 2016. Hikvision ranked No.1 in Scene Classification at ImageNet 2016 challenge. *[Online] Security News Desk*. Available: http://www.securitynewsdesk.com/hikvision-ranked-no-1-scene-classification-imagenet-2016-challenge/ [Accessed: 20/03/2017].
[17] See Residual Networks in Part Four of this publication for more details.
[18] Details available under team information Trimps-Soushen from: http://image-net.org/challenges/LSVRC/2016/results

- **ResNeXt** by Facebook came a close second in top-5 classification error with 3.03% by using a new architecture that extends the original ResNet architecture. [19]

Object Detection

As one can imagine the process of **Object Detection** does exactly that, detects objects within images. The definition provided for object detection by the ILSVRC 2016[20] includes outputting bounding boxes and labels for individual objects. This differs from the classification/localisation task by applying classification and localisation to many objects instead of just a single dominant object.

Figure 3: Object Detection With Face as the Only Class



Note: Picture is an example of face detection, Object Detection of a single class. The authors cite one of the persistent issues in Object Detection to be the detection of small objects. Using small faces as a test class they explore the role of scale invariance, image resolution, and contextual reasoning.

Source: Hu and Ramanan (2016, p. 1)[21]

One of 2016's major trends in **Object Detection** was the shift towards a quicker, more efficient detection system. This was visible in approaches like YOLO, SSD and R-FCN as a move towards sharing computation on a whole image. Hence, differentiating themselves from the costly subnetworks associated with Fast/Faster R-CNN

[19] Xie, S., Girshick, R., Dollar, P., Tu, Z. & He, K. 2016. Aggregated Residual Transformations for Deep Neural Networks. *[Online] arXiv: 1611.05431*. Available: **arXiv:1611.05431v1**

[20] ImageNet Large Scale Visual Recognition Challenge (2016), Part II, Available: http://image-net.org/challenges/LSVRC/2016/#det [Accessed: 22/11/2016]

[21] Hu and Ramanan. 2016. Finding Tiny Faces. *[Online] arXiv: 1612.04402.* Available: **arXiv:1612.04402v1**

techniques. This is typically referred to as 'end-to-end training/learning' and features throughout this piece.

The rationale generally is to avoid having separate algorithms focus on their respective subproblems in isolation as this typically increases training time and can lower network accuracy. That being said this end-to-end adaptation of networks typically takes place after initial sub-network solutions and, as such, is a retrospective optimisation. However, Fast/Faster R-CNN techniques remain highly effective and are still used extensively for object detection.

- **SSD: Single Shot MultiBox Detector**[22] utilises a single Neural Network which encapsulates all the necessary computation and eliminates the costly proposal generation of other methods. It achieves "*75.1% mAP, outperforming a comparable state of the art Faster R-CNN model*" (Liu et al. 2016).

- One of the most impressive systems we saw in 2016 was from the aptly named "**YOLO9000: Better, Faster, Stronger**"[23], which introduces the YOLOv2 and YOLO9000 detection systems.[24] YOLOv2 vastly improves the initial YOLO model from mid-2015,[25] and is able to achieve better results at very high FPS (up to 90 FPS on low resolution images using the original GTX Titan X). In addition to completion speed, the system outperforms Faster RCNN with ResNet and SSD on certain object detection datasets.

  YOLO9000 implements a joint training method for detection and classification extending its prediction capabilities beyond the labelled detection data available i.e. it is able to detect objects that it has never seen labelled detection data for. The YOLO9000 model provides real-time object detection across 9000+ categories, closing the dataset size gap between classification and detection. Additional details, pre-trained models and a video showing it in action is available here.[26]

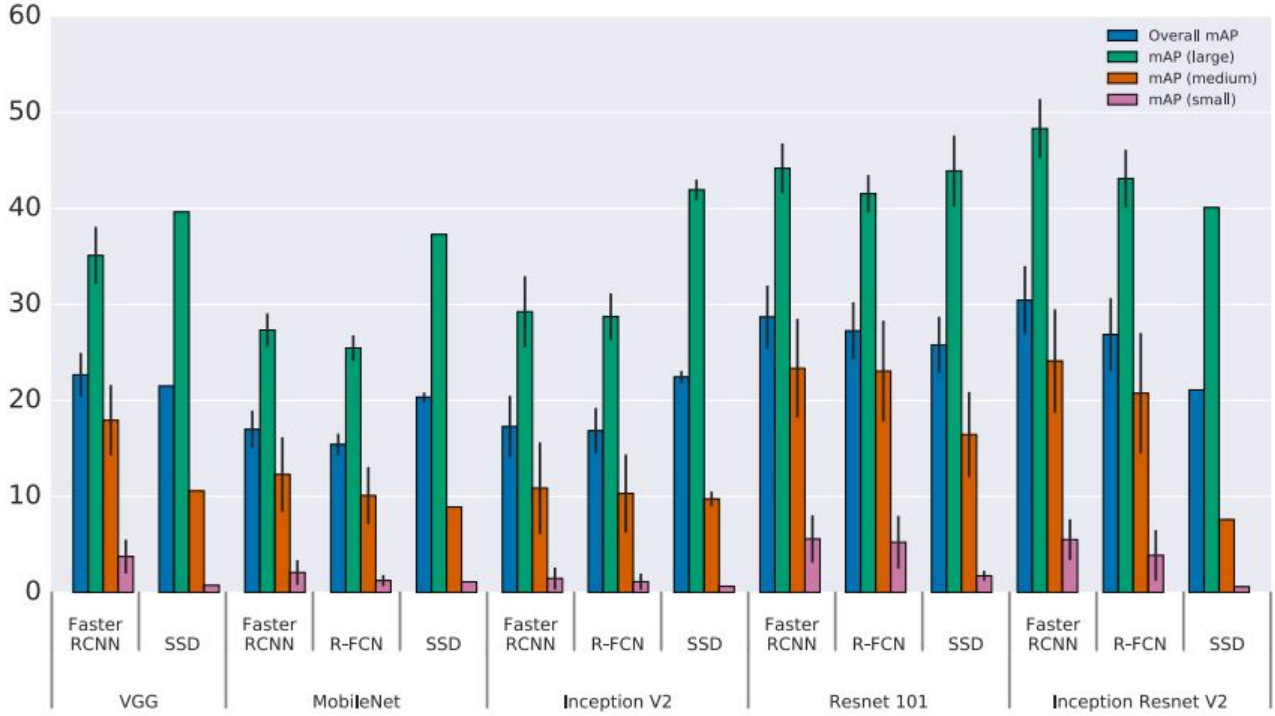
- **Feature Pyramid Networks for Object Detection**[27] comes from FAIR[28] and capitalises on the "*inherent multi-scale, pyramidal hierarchy of deep*

[22] Liu et al. 2016. SSD: Single Shot MultiBox Detector. *[Online] arXiv: 1512.02325v5*. Available: **arXiv:1512.02325v5**
[23] Redmon, J. Farhadi, A. 2016. YOLO9000: Better, Faster, Stronger. *[Online] arXiv: 1612.08242v1*. Available: **arXiv:1612.08242v1**
[24] YOLO stands for "You Only Look Once".
[25] Redmon et al. 2016. You Only Look Once: Unified, Real-Time Object Detection. *[Online] arXiv: 1506.02640*. Available: **arXiv:1506.02640v5**
[26] Redmon. 2017. YOLO: Real-Time Object Detection. *[Website] pjreddie.com*. Available: *https://pjreddie.com/darknet/yolo/* [Accessed: 01/03/2017].
[27] Lin et al. 2016. Feature Pyramid Networks for Object Detection. *[Online] arXiv: 1612.03144*. Available: **arXiv:1612.03144v1**
[28] Facebook's Artificial Intelligence Research

convolutional networks to construct feature pyramids with marginal extra cost", meaning that representations remain powerful without compromising speed or memory. Lin et al. (2016) achieve state-of-the-art (hereafter SOTA) single-model results on COCO[29]. Beating the results achieved by winners in 2016 when combined with a basic Faster R-CNN system.

- **R-FCN: Object Detection via Region-based Fully Convolutional Networks**:[30] This is another method that avoids applying a costly per-region subnetwork hundreds of times over an image by making the region-based detector fully convolutional and sharing computation on the whole image. "*Our result is achieved at a test-time speed of 170ms per image, 2.5-20x faster than the Faster R-CNN counterpart*" (Dai et al., 2016).

Figure 4: Accuracy tradeoffs in Object Detection



Note: Y-axis displays mAP (mean Average Precision) and the X-axis displays meta-architecture variability across each feature extractor (VGG, MobileNet...Inception ResNet V2). Additionally, mAP small, medium and large describe the average precision for small, medium and large objects, respectively. As such accuracy is "*stratified by object size, meta-architecture and feature extractor*" and "*image resolution is fixed to 300*". While Faster R-CNN performs comparatively well in the above sample, it is worth noting that the meta-architecture is considerably slower than more recent approaches, such as R-FCN.
Source: Huang et al. (2016, p. 9)[31]

[29] Common Objects in Context (COCO) image dataset
[30] Dai et al. 2016. R-FCN: Object Detection via Region-based Fully Convolutional Networks. *[Online] arXiv: 1605.06409*. Available: **arXiv:1605.06409v2**
[31] Huang et al. 2016. Speed/accuracy trade-offs for modern convolutional object detectors. *[Online] arXiv: 1611.10012*. Available: **arXiv:1611.10012v1**

Huang et al. (2016)[32] present a paper which provides an in depth performance comparison between R-FCN, SSD and Faster R-CNN. Due to the issues around accurate comparison of Machine Learning (ML) techniques we'd like to point to the merits of producing a standardised approach here. They view these architectures as 'meta-architectures' since they can be combined with different kinds of feature extractors such as ResNet or Inception.

The authors study the trade-off between accuracy and speed by varying meta-architecture, feature extractor and image resolution. The choice of feature extractor for example produces large variations between meta-architectures.

The trend of making object detection cheap and efficient while still retaining the accuracy required for real-time commercial applications, notably in autonomous driving applications, is also demonstrated by SqueezeDet[33] and PVANet[34] papers. While a Chinese company, DeepGlint, provides a good example of object detection in operation as a CCTV integration, albeit in a vaguely Orwellian manner: Video.[35]

Results from ILSVRC and COCO Detection Challenge

COCO[36] (Common Objects in Context) is another popular image dataset. However, it is comparatively smaller and more curated than alternatives like ImageNet, with a focus on object recognition within the broader context of scene understanding. The organizers host a yearly challenge for Object Detection, segmentation and keypoints. Detection results from both the ILSVRC[37] and the COCO[38] Detection Challenge are;

- **ImageNet LSVRC Object Detection from Images (DET):** CUImage 66% meanAP. Won 109 out of 200 object categories.
- **ImageNet LSVRC Object Detection from video (VID):** NUIST 80.8% mean AP
- **ImageNet LSVRC Object Detection from video with tracking:** CUvideo 55.8% mean AP

---

[32] ibid

[33] Wu et al. 2016. SqueezeDet: Unified, Small, Low Power Fully Convolutional Neural Networks for Real-Time Object Detection for Autonomous Driving. *[Online] arXiv: 1612.01051*. Available: **arXiv:1612.01051v2**

[34] Hong et al. 2016. PVANet: Lightweight Deep Neural Networks for Real-time Object Detection. *[Online] arXiv: 1611.08588v2*. Available: **arXiv:1611.08588v2**

[35] DeepGlint Official. 2016. DeepGlint CVPR2016. *[Online] Youtube.com*. Available: https://www.youtube.com/watch?v=xhp47v5OBXQ [Accessed: 01/03/2017].

[36] COCO - Common Objects in Common. 2016. *[Website]* Available: http://mscoco.org/ [Accessed: 04/01/2017].

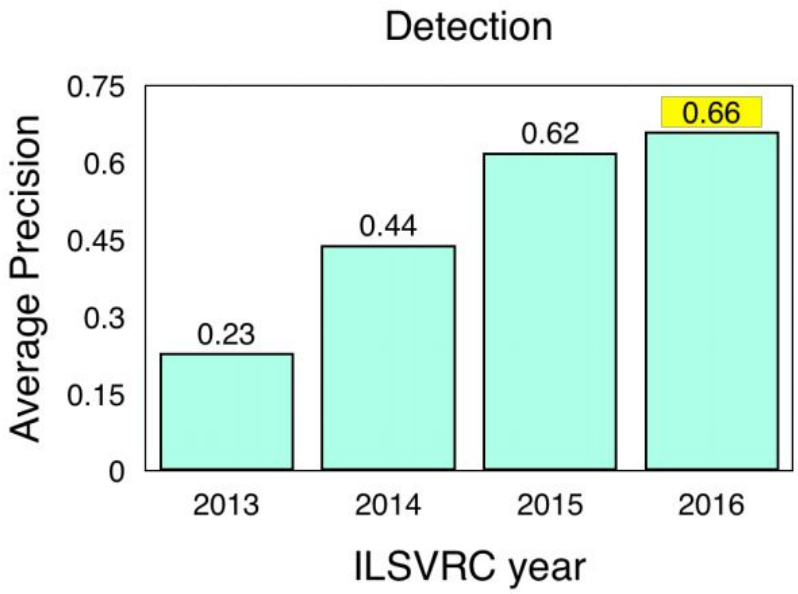
[37] ILSRVC results taken from: ImageNet. 2016. Large Scale Visual Recognition Challenge 2016. *[Website] Object Detection*. Available: http://image-net.org/challenges/LSVRC/2016/results [Accessed: 04/01/2017].

[38] COCO Detection Challenge results taken from: COCO - Common Objects in Common. 2016. Detections Leaderboard *[Website] mscoco.org*. Available: http://mscoco.org/dataset/#detections-leaderboard [Accessed: 05/01/2017].

- **COCO 2016 Detection Challenge (bounding boxes):** G-RMI (Google) 41.5% AP (4.2% absolute percentage increase from 2015 winner MSRAVC)

In review of the detection results for 2016, ImageNet stated that the 'MSRAVC 2015 set a very high bar for performance [introduction of ResNets to competition]. Performance on all classes has improved across entries. Localization improved greatly in both challenges. High relative improvement on small object instances' (ImageNet, 2016).[39]

Figure 5: ILSVRC detection results from images (2013-2016)



Note: ILSVRC Object Detection results from images (DET) (2013-2016).
Source: ImageNet. 2016. [Online] *Workshop Presentation, Slide 2*. Available: http://image-net.org/challenges/talks/2016/ECCV2016_ilsvrc_coco_detection_segmentation.pdf

Object Tracking

Refers to the process of following a specific object of interest, or multiple objects, in a given scene. It traditionally has applications in video and real-world interactions where observations are made following an initial **object detection**; the process is crucial to autonomous driving systems for example.

[39] ImageNet. 2016. [Online] *Workshop Presentation*, Slide 31. Available: http://image-net.org/challenges/talks/2016/ECCV2016_ilsvrc_coco_detection_segmentation.pdf [Accessed: 06/01/2017].

- **Fully-Convolutional Siamese Networks for Object Tracking**[40] combines a basic tracking algorithm with a Siamese network, trained end-to-end, which achieves SOTA and operates at frame-rates in excess of real-time. This paper attempts to tackle the lack of richness available to tracking models from traditional online learning methods.

- **Learning to Track at 100 FPS with Deep Regression Networks**[41] is another paper which attempts to ameliorate the existing issues with online training methods. The authors produce a tracker which leverages a feed-forward network to learn the generic relationships surrounding object motion, appearance and orientation which effectively track novel objects without online training. Provides SOTA on a standard tracking benchmark while also managing "*to track generic objects at 100 fps*" (Held et al., 2016).

  Video of GOTURN (Generic Object Tracking Using Regression Networks) available: Video[42]

- **Deep Motion Features for Visual Tracking**[43] merge hand-crafted features, deep RGB/appearance features (from CNNs), and deep motion features (trained on optical flow images) to achieve SOTA. While deep motion features are commonplace in Action Recognition and Video Classification, the authors claim this is the first time they are used for visual tracking. The paper was also awarded Best Paper in ICPR 2016, for "Computer Vision and Robot Vision" track.

  "*This paper presents an investigation of the impact of deep motion features in a tracking-by-detection framework. We further show that hand-crafted, deep RGB, and deep motion features contain complementary information. To the best of our knowledge, we are the first to propose fusing appearance information with deep motion features for visual tracking. Comprehensive experiments clearly suggest that our fusion approach with deep motion features outperforms standard methods relying on appearance information alone.*"

- **Virtual Worlds as Proxy for Multi-Object Tracking Analysis**[44] approaches the lack of true-to-life variability present in existing video-tracking benchmarks and

---

[40] Bertinetto et al. 2016. Fully-Convolutional Siamese Networks for Object Tracking. *[Online] arXiv: 1606.09549*. Available: https://arxiv.org/abs/1606.09549v2
[41] Held et al. 2016. Learning to Track at 100 FPS with Deep Regression Networks. *[Online] arXiv: 1604.01802*. Available: https://arxiv.org/abs/1604.01802v2
[42] David Held. 2016. GOTURN - a neural network tracker. *[Online] YouTube.com*. Available: https://www.youtube.com/watch?v=kMhwXnLgT_I [Accessed: 03/03/2017].
[43] Gladh et al. 2016. Deep Motion Features for Visual Tracking. *[Online] arXiv: 1612.06615*. Available: **arXiv:1612.06615v1**
[44] Gaidon et al. 2016. Virtual Worlds as Proxy for Multi-Object Tracking Analysis. *[Online] arXiv: 1605.06457*. Available: **arXiv:1605.06457v1**

datasets. The paper proposes a new method for real-world cloning which generates rich, virtual, synthetic, photo-realistic environments from scratch with full-labels that overcome some of the sterility present in existing datasets. The generated images are automatically labelled with accurate ground truth allowing a range of applications aside from object detection/tracking, such as depth and optical flow.

- **Globally Optimal Object Tracking with Fully Convolutional Networks**[45] deals with object variance and occlusion, citing these as two of the root limitations within object tracking. "*Our proposed method solves the object appearance variation problem with the use of a Fully Convolutional Network and deals with occlusion by Dynamic Programming*" (Lee et al., 2016).

Part Two: Segmentation, Super-res/Colourisation/Style Transfer, Action Recognition

Segmentation

Central to Computer Vision is the process of Segmentation, which divides whole images into pixel groupings which can then be labelled and classified. Moreover, Semantic Segmentation goes further by trying to semantically understand the role of each pixel in the image e.g. is it a cat, car or some other type of class? Instance Segmentation takes this even further by segmenting different instances of classes e.g. labelling three different dogs with three different colours. It is one of a barrage of Computer Vision applications currently employed in autonomous driving technology suites.

Perhaps some of the best improvements in the area of segmentation come courtesy of FAIR, who continue to build upon their DeepMask work from 2015.[46] DeepMask generates rough 'masks' over objects as an initial form of segmentation. In 2016, Fair introduced SharpMask[47] which refines the 'masks' provided by DeepMask, correcting the loss of detail and improving semantic segmentation. In addition to this, MultiPathNet [48] identifies the objects delineated by each mask.

---

[45] Lee et al. 2016. Globally Optimal Object Tracking with Fully Convolutional Networks. *[Online] arXiv: 1612.08274*. Available: **arXiv:1612.08274v1**

[46] Pinheiro, Collobert and Dollar. 2015. Learning to Segment Object Candidates. *[Online] arXiv: 1506.06204*. Available: **arXiv:1506.06204v2**

[47] Pinheiro et al. 2016. Learning to Refine Object Segments. *[Online] arXiv: 1603.08695*. Available: **arXiv:1603.08695v2**

[48] Zagoruyko, S. 2016. A MultiPath Network for Object Detection. *[Online] arXiv: 1604.02135v2.* Available: **arXiv:1604.02135v2**

"*To capture general object shape, you have to have a high-level understanding of what you are looking at (DeepMask), but to accurately place the boundaries you need to look back at lower-level features all the way down to the pixels (SharpMask).*" - Piotr Dollar, 2016.[49]

Figure 6: Demonstration of FAIR techniques in action

Note: The above pictures demonstrate the segmentation techniques employed by FAIR. These include the application of DeepMask, SharpMask and MultiPathNet techniques which are applied in that order. This process allows accurate segmentation and classification in a variety of scenes.
Source: Dollar (2016).[50]

Video Propagation Networks[51] attempt to create a simple model to propagate accurate object masks, assigned at first frame, through the entire video sequence along with some additional information.

In 2016, researchers worked on finding alternative network configurations to tackle the aforementioned issues of scale and localisation. DeepLab[52] is one such example of this which achieves encouraging results for semantic image segmentation tasks. Khoreva et

[49] Dollar, P. 2016. Learning to Segment. *[Blog] FAIR*. Available: https://research.fb.com/learning-to-segment/
[50] Dollar, P. 2016. Segmenting and refining images with SharpMask. *[Online] Facebook Code*. Available: https://code.facebook.com/posts/561187904071636/segmenting-and-refining-images-with-sharpmask/
[51] Jampani et al. 2016. Video Propagation Networks. *[Online] arXiv: 1612.05478*. Available: **arXiv:1612.05478v2**
[52] Chen et al., 2016. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *[Online] arXiv: 1606.00915*. Available: **arXiv:1606.00915v1**

al. (2016)[53] build on Deeplab's earlier work (circa 2015) and propose a weakly supervised training method which achieves comparable results to fully supervised networks.

Computer Vision further refined the network sharing of useful information approach through the use of end-to-end networks, which reduce the computational requirements of multiple omni-directional subtasks for classification. Two key papers using this approach are:

- **100 Layers Tiramisu[54]** is a fully-convolutional DenseNet which connects every layer, to every other layer, in a feed-forward fashion. It also achieves SOTA on multiple benchmark datasets with fewer parameters and training/processing.
- **Fully Convolutional Instance-aware Semantic Segmentation[55]** performs instance mask prediction and classification jointly (two subtasks).
  COCO Segmentation challenge winner MSRA. 37.3% AP.
  9.1% absolute jump from MSRAVC in 2015 in COCO challenge.

While **ENet**,[56] a DNN architecture for real-time semantic segmentation, is not of this category, it does demonstrate the commercial merits of reducing computation costs and giving greater access to mobile devices.

Our work wishes to relate as much of these advancements back to tangible public applications as possible. With this in mind, the following contains some of the most interesting **healthcare application** of segmentation in 2016;

- A Benchmark for Endoluminal Scene Segmentation of Colonoscopy Images[57]
- 3D fully convolutional networks for subcortical segmentation in MRI: A large-scale study[58]
- Semi-supervised Learning using Denoising Autoencoders for Brain Lesion Detection and Segmentation[59]

---

[53] Khoreva et al. 2016. Simple Does It: Weakly Supervised Instance and Semantic Segmentation. *[Online] arXiv: 1603.07485v2.* Available: **arXiv:1603.07485v2**
[54] Jégou et al. 2016. The One Hundred Layers Tiramisu: Fully Convolutional DenseNets for Semantic Segmentation. *[Online] arXiv: 1611.09326v2.* Available: **arXiv:1611.09326v2**
[55] Li et al. 2016. Fully Convolutional Instance-aware Semantic Segmentation. *[Online] arXiv: 1611.07709v1*. Available: **arXiv:1611.07709v1**
[56] Paszke et al. 2016. ENet: A Deep Neural Network Architecture for Real-Time Semantic Segmentation. *[Online] arXiv: 1606.02147v1*. Available: **arXiv:1606.02147v1**
[57] Vázquez et al. 2016. A Benchmark for Endoluminal Scene Segmentation of Colonoscopy Images. *[Online] arXiv: 1612.00799*. Available: **arXiv:1612.00799v1**
[58] Dolz et al. 2016. 3D fully convolutional networks for subcortical segmentation in MRI: A large-scale study. *[Online] arXiv: 1612.03925*. Available: **arXiv:1612.03925v1**
[59] Alex et al. 2017. Semi-supervised Learning using Denoising Autoencoders for Brain Lesion Detection and Segmentation. *[Online] arXiv: 1611.08664*. Available: **arXiv:1611.08664v4**

- 3D Ultrasound image segmentation: A Survey[60]
- A Fully Convolutional Neural Network based Structured Prediction Approach Towards the Retinal Vessel Segmentation[61]
- 3-D Convolutional Neural Networks for Glioblastoma Segmentation[62]

One of our favourite quasi-medical segmentation applications is **FusionNet**[63]- a deep fully residual convolutional neural network for image segmentation in connectomics[64] benchmarked against SOTA electron microscopy (EM) segmentation methods.

Super-resolution, Style Transfer & Colourisation

Not all research in Computer Vision serves to extend the pseudo-cognitive abilities of machines, and often the fabled malleability of neural networks, as well as other ML techniques, lend themselves to a variety of other novel applications that spill into the public space. Last year's advancements in Super-resolution, Style Transfer & Colourisation occupied that space for us.

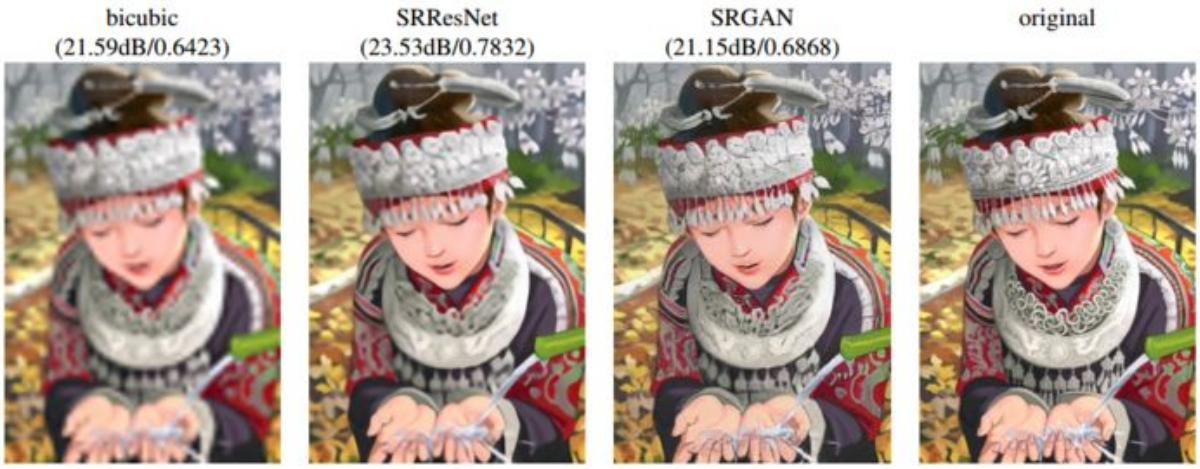
Super-resolution refers to the process of estimating a high resolution image from a low resolution counterpart, and also the prediction of image features at different magnifications, something which the human brain can do almost effortlessly. Originally super-resolution was performed by simple techniques like bicubic-interpolation and nearest neighbours. In terms of commercial applications, the desire to overcome low-resolution constraints stemming from source quality and realisation of 'CSI Miami' style image enhancement has driven research in the field. Here are some of the year's advances and their potential impact:

- **Neural Enhance**[65] is the brainchild of Alex J. Champandard and combines approaches from four different research papers to achieve its Super-resolution method.

[60] Mozaffari and Lee. 2016. 3D Ultrasound image segmentation: A Survey. *[Online] arXiv: 1611.09811*. Available: **arXiv:1611.09811v1**
[61] Dasgupta and Singh. 2016. A Fully Convolutional Neural Network based Structured Prediction Approach Towards the Retinal Vessel Segmentation. *[Online] arXiv: 1611.02064*. Available: **arXiv:1611.02064v2**
[62] Yi et al. 2016. 3-D Convolutional Neural Networks for Glioblastoma Segmentation. *[Online] arXiv: 1611.04534*. Available: **arXiv:1611.04534v1**
[63] Quan et al. 2016. FusionNet: A deep fully residual convolutional neural network for image segmentation in connectomics. *[Online] arXiv: 1612.05360*. Available: **arXiv:1612.05360v2**
[64] Connectomics refers to the mapping of all connections within an organism's nervous system, i.e. neurons and their connections.
[65] Champandard, A.J. 2017. Neural Enhance (latest commit 30/11/2016). *[Online] Github*. Available: https://github.com/alexjc/neural-enhance [Accessed: 11/02/2017]

- **Real-Time Video Super Resolution** was also attempted in 2016 in two notable instances.[66,67]
- **RAISR:** Rapid and Accurate Image Super-Resolution[68] from Google avoids the costly memory and speed requirements of neural network approaches by training filters with low-resolution and high-resolution image pairs. RAISR, as a learning-based framework, is two orders of magnitude faster than competing algorithms and has minimal memory requirements when compared with neural network-based approaches. Hence super-resolution is extendable to personal devices. There is a research blog available here.[69]

Figure 7: Super-resolution SRGAN example



Note: From left to right: bicubic interpolation (the objective worst performer for focus), Deep residual network optimised for MSE, deep residual generative adversarial network optimized for a loss more sensitive to human perception, original High Resolution (HR) image. Corresponding peak signal to noise ratio (PSNR) and structural similarity (SSIM) are shown in two brackets. [4 x upscaling] The reader may wish to zoom in on the middle two images (SRResNet and SRGAN) to see the difference between image smoothness vs more realistic fine details.
Source: Ledig et al. (2017)[70]

The use of Generative Adversarial Networks (GANs) represent current SOTA for Super-resolution:

[66] Caballero et al. 2016. Real-Time Video Super-Resolution with Spatio-Temporal Networks and Motion Compensation. *[Online] arXiv: 1611.05250*. Available: **arXiv:1611.05250v1**
[67] Shi et al. 2016. Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network. *[Online] arXiv: 1609.05158*. Available: **arXiv:1609.05158v2**
[68] Romano et al. 2016. RAISR: Rapid and Accurate Image Super Resolution. *[Online] arXiv: 1606.01299*. Available: **arXiv:1606.01299v3**
[69] Milanfar, P. 2016. Enhance! RAISR Sharp Images with Machine Learning. *[Blog] Google Research Blog*. Available: https://research.googleblog.com/2016/11/enhance-raisr-sharp-images-with-machine.html [Accessed: 20/03/2017].
[70] ibid

- **SRGAN**[71] provides photo-realistic textures from heavily downsampled images on public benchmarks, using a discriminator network trained to differentiate between super-resolved and original photo-realistic images.

  Qualitatively SRGAN performs the best, although SRResNet performs best with peak-signal-to-noise-ratio (PSNR) metric but SRGAN gets the finer texture details and achieves the best Mean Opinion Score (MOS). *"To our knowledge, it is the first framework capable of inferring photo-realistic natural images for 4× upscaling factors."*[72] All previous approaches fail to recover the finer texture details at large upscaling factors.

- **Amortised MAP Inference for Image Super-resolution**[73] proposes a method for calculation of Maximum a Posteriori (MAP) inference using a Convolutional Neural Network. However, their research presents three approaches for optimisation, all of which GANs perform markedly better on real image data at present.

Figure 8: Style Transfer from Nikulin & Novak

Note: Transferring different styles to a photo of a cat (original top left).
Source: Nikulin & Novak (2016)

[71] Ledig et al. 2017. Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network. *[Online] arXiv: 1609.04802*. Available: **arXiv:1609.04802v3**
[72] ibid
[73] Sønderby et al. 2016. Amortised MAP Inference for Image Super-resolution. *[Online] arXiv: 1610.04490*. Available: **arXiv:1610.04490v1**

Undoubtedly, **Style Transfer** epitomises a novel use of neural networks that has ebbed into the public domain, specifically through last year's facebook integrations and companies like Prisma[74] and Artomatix[75]. Style transfer is an older technique but converted to a neural networks in 2015 with the publication of a Neural Algorithm of Artistic Style.[76] Since then, the concept of style transfer was expanded upon by Nikulin and Novak[77] and also applied to video,[78] as is the common progression within Computer Vision.

Figure 9: Further examples of Style Transfer

Note: The top row (left to right) represent the artistic style which is transposed onto the original images which are displayed in the first column (Woman, Golden Gate Bridge and Meadow Environment). Using conditional instance normalisation a single style transfer network can capture 32 style simultaneously, five of which are displayed here. The full suite of images in available in the source paper's appendix. This work will feature in the International Conference on Learning Representations (ICLR) 2017.
Source: Dumoulin et al. (2017, p. 2)[79]

Style transfer as a topic is fairly intuitive once visualised; take an image and imagine it with the stylistic features of a different image. For example, in the style of a famous

[74] Prisma. 2017. *[Website] Prisma*. Available: https://prisma-ai.com/ [Accessed: 01/04/2017].
[75] Artomatix. 2017. *[Website] Artomatix*. Available: https://services.artomatix.com/ [Accessed: 01/04/2017].
[76] Gatys et al. 2015. A Neural Algorithm of Artistic Style. *[Online] arXiv: 1508.06576*. Available: **arXiv:1508.06576v2**
[77] Nikulin & Novak. 2016. Exploring the Neural Algorithm of Artistic Style. *[Online] arXiv: 1602.07188*. Available: **arXiv:1602.07188v2**
[78] Ruder et al. 2016. Artistic style transfer for videos. *[Online] arXiv: 1604.08610*. Available: **arXiv:1604.08610v2**
[79] ibid

painting or artist. This year Facebook released Caffe2Go,[80] their deep learning system which integrates into mobile devices. Google also released some interesting work which sought to blend multiple styles to generate entirely unique image styles: Research blog [81] and full paper.[82]

Besides mobile integrations, style transfer has applications in the creation of game assets. Members of our team recently saw a presentation by the Founder and CTO of Artomatix, Eric Risser, who discussed the technique's novel application for content generation in games (texture mutation, etc.) and, therefore, dramatically minimises the work of a conventional texture artist.

Colourisation is the process of changing monochrome images to new full-colour versions. Originally this was done manually by people who painstakingly selected colours to represent specific pixels in each image. In 2016, it became possible to automate this process while maintaining the appearance of realism indicative of the human-centric colourisation process. While humans may not accurately represent the true colours of a given scene, their real world knowledge allows the application of colours in a way which is consistent with the image and another person viewing said image.

The process of colourisation is interesting in that the network assigns the most likely colouring for images based on its understanding of object location, textures and environment, e.g. it learns that skin is pinkish and the sky is blueish.

Three of the most influential works of the year are as follows:

- Zhang et al.[83] produced a method that was able to successfully fool humans on 32% of their trials. Their methodology is comparable to a "colourisation Turing test."

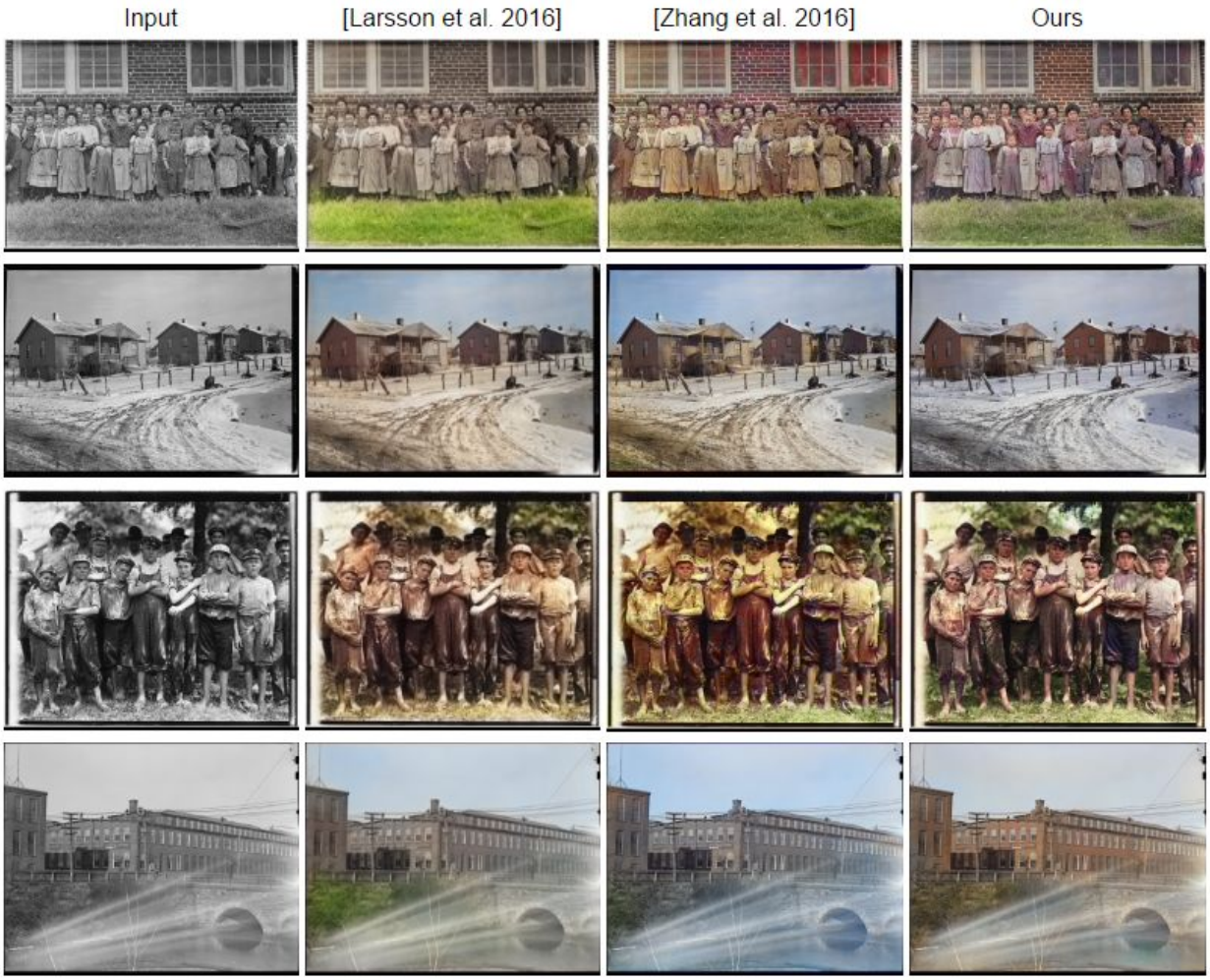
- Larsson et al.[84] fully automate their image colourisation system using Deep Learning for Histogram estimation.

---

[80] Jia and Vajda. 2016. Delivering real-time AI in the palm of your hand. *[Online] Facebook Code*. Available: https://code.facebook.com/posts/196146247499076/delivering-real-time-ai-in-the-palm-of-your-hand/ [Accessed: 20/01/2017].
[81] Dumoulin et al. 2016. Supercharging Style Transfer. *[Online] Google Research Blog*. Available: https://research.googleblog.com/2016/10/supercharging-style-transfer.html [Accessed: 20/01/2017].
[82] Dumoulin et al. 2017. A Learned Representation For Artistic Style. *[Online] arXiv: 1610.07629*. Available: **arXiv:1610.07629v5**
[83] Zhang et al. 2016. Colorful Image Colorization. *[Online] arXiv: 1603.08511*. Available: **arXiv:1603.08511v5**
[84] Larsson et al. 2016. Learning Representations for Automatic Colorization. *[Online] arXiv: 1603.06668*. Available: **arXiv:1603.06668v2**

- Finally, Lizuka, Simo-Serra and Ishikawa[85] demonstrate a colourisation model also based upon CNNs. The work outperformed the existing SOTA, we [the team] feel as though this work is qualitatively best also, appearing to be the most realistic. Figure 10 provides comparisons, however the image is taken from Lizuka et al.

Figure 10: Comparison of Colourisation Research



Note: From top to bottom - column one contains the original monochrome image input which is subsequently colourised through various techniques. The remaining columns display the results generated by other prominent colourisation research in 2016. When viewed from left to right, these are Larsson et al.[78] 2016 (column two), Zhang et al.[77] 2016 (Column three), and Lizuka, Simo-Serra and Ishikawa.[79] 2016, also referred to as "ours" by the authors (Column four). The quality difference in colourisation is most evident in row three (from the top) which depicts a group of young boys. We believe Lizuka et al.'s work to be qualitatively superior (Column four).
Source: Lizuka et al. 2016[86]

[85] Lizuka, Simo-Serra and Ishikawa. 2016. Let there be Color!: Joint End-to-end Learning of Global and Local Image Priors for Automatic Image Colorization with Simultaneous Classification. *[Online] ACM Transaction on Graphics (Proc. of SIGGRAPH), 35(4):110*. Available: http://hi.cs.waseda.ac.jp/~iizuka/projects/colorization/en/
[86] ibid

"*Furthermore, our architecture can process images of any resolution, unlike most existing approaches based on CNN.*"

In a test to see how natural their colourisation was, users were given a random image from their models and were asked, "does this image look natural to you?"

Their approach achieved 92.6%, the baseline achieved roughly 70% and the ground truth (the actual colour photos) were considered 97.7% of the time to be natural.

Action Recognition

The task of action recognition refers to the both the classification of an action within a given video frame, and more recently, algorithms which can predict the likely outcomes of interactions given only a few frames before the action takes place. In this respect we see recent research attempt to imbed context into algorithmic decisions, similar to other areas of Computer Vision. Some key papers in this space are:

- **Long-term Temporal Convolutions for Action Recognition**[87] leverages the spatio-temporal structure of human actions, i.e. the particular movement and duration, to correctly recognise actions using a CNN variant. To overcome the sub-optimal temporal modelling of longer term actions by CNNs, the authors propose a neural network with long-term temporal convolutions (LTC-CNN) to improve the accuracy of action recognition. Put simply, the LTCs can look at larger parts of the video to recognise actions. Their approach uses and extends 3D CNNs 'to enable action representation at a fuller temporal scale'.

  "*We report state-of-the-art results on two challenging benchmarks for human action recognition UCF101 (92.7%) and HMDB51 (67.2%).*"

- **Spatiotemporal Residual Networks for Video Action Recognition**[88] apply a variation of two stream CNN to the task of action recognition, which combines techniques from both traditional CNN approaches and recently popularised Residual Networks (ResNets). The two stream approach takes its inspiration from a neuroscientific hypothesis on the functioning of the visual cortex, i.e. separate pathways recognise object shape/colour and movement. The authors

[87] Varol et al. 2016. Long-term Temporal Convolutions for Action Recognition. *[Online] arXiv: 1604.04494*. Available: **arXiv:1604.04494v1**
[88] Feichtenhofer et al. 2016. Spatiotemporal Residual Networks for Video Action Recognition. *[Online] arXiv: 1611.02155*. Available: **arXiv:1611.02155v1**

combine the classification benefits of ResNets by injecting residual connections between the two CNN streams.

"*Each stream initially performs video recognition on its own and for final classification, softmax scores are combined by late fusion. To date, this approach is the most effective approach of applying deep learning to action recognition, especially with limited training data. In our work we directly convert image ConvNets into 3D architectures and show greatly improved performance over the two-stream baseline.*" - 94% on UCF101 and 70.6% on HMDB51. Feichtenhofer et al. made improvements over traditional improved dense trajectory (iDT) methods and generated better results through use of both techniques.

- **Anticipating Visual Representations from Unlabeled Video**[89] is an interesting paper, although not strictly action classification. The program predicts the action which is likely to take place given a sequence of video frames up to one second before an action. The approach uses visual representations rather than pixel-by-pixel classification, which means that the program can operate without labeled data, by taking advantage of the feature learning properties of deep neural networks.[90]

  "*The key idea behind our approach is that we can train deep networks to predict the visual representation of images in the future. Visual representations are a promising prediction target because they encode images at a higher semantic level than pixels yet are automatic to compute. We then apply recognition algorithms on our predicted representation to anticipate objects and actions*".

- The organisers of the **Thumos Action Recognition Challenge**[91] released a paper describing the general approaches for Action Recognition from the last number of years. The paper also provides a rundown of the Challenges from 2013-2015, future directions for the challenge and ideas on how to give computers a more holistic understanding of video through Action Recognition. We hope that the Thumos Action Recognition Challenge returns in 2017 after its (seemingly) unexpected hiatus.

[89] Vondrick et al. 2016. Anticipating Visual Representations from Unlabeled Video. *[Online] arXiv: 1504.08023*. Available: **arXiv:1504.08023v2**
[90] Conner-Simons, A., Gordon, R. 2016. Teaching machines to predict the future. *[Online] MIT NEWS*. Available: https://news.mit.edu/2016/teaching-machines-to-predict-the-future-0621 [Accessed: 03/02/2017].
[91] Idrees et al. 2016. The THUMOS Challenge on Action Recognition for Videos "in the Wild". *[Online] arXiv: 1604.06182*. Available: **arXiv:1604.06182v1**

Part Three: Toward a 3D understanding of the world

"*A key goal of Computer Vision is to recover the underlying 3D structure from 2D observations of the world.*" - Rezende et al. (2016, p. 1)[92]

In Computer Vision, the classification of scenes, objects and activities, along with the output of bounding boxes and image segmentation is, as we have seen, the focus of much new research. In essence, these approaches apply computation to gain an 'understanding' of the 2D space of an image. However, detractors note that a 3D understanding is imperative for systems to successfully interpret, and navigate, the real world.

For instance, a network may locate a cat in an image, colour all of its pixels and classify it as a cat. But does the network fully understand where the cat in the image is, in the context of the cat's environment?

One could argue that the computer learns very little about the 3D world from the above tasks. Contrary to this, humans understand the world in 3D even when examining 2D pictures, i.e. perspective, occlusion, depth, how objects in a scene are related, etc. Imparting these 3D representations and their associated knowledge to artificial systems represents one of the next great frontiers of Computer Vision. A major reason for thinking this is that, generally;

"*the 2D projection of a scene is a complex function of the attributes and positions of the camera, lights and objects that make up the scene. If endowed with 3D understanding, agents can abstract away from this complexity to form stable, disentangled representations, e.g., recognizing that a chair is a chair whether seen from above or from the side, under different lighting conditions, or under partial occlusion.*"[93]

However, 3D understanding has traditionally faced several impediments. The first concerns the problem of both 'self and normal occlusion' along with the numerous 3D shapes which fit a given 2D representation. Understanding problems are further compounded by the inability to map different images of the same structures to the same 3D space, and in the handling of the multi-modality of these representations.[94] Finally, ground-truth 3D datasets were traditionally quite expensive and difficult to obtain which, when coupled with divergent approaches for representing 3D structures, may have led to training limitations.

---

[92] Rezende et al. 2016. Unsupervised Learning of 3D Structure from Images. *[Online] arXiv: 1607.00662*. Available: **arXiv:1607.00662v1**
[93] ibid
[94] ibid

We feel that the work being conducted in this space is important to be mindful of. From the embryonic, albeit titillating early theoretical applications for future AGI systems and robotics, to the immersive, captivating applications in augmented, virtual and mixed reality which will affect our societies in the near future. We cautiously predict exponential growth in this area of Computer Vision, as a result of lucrative commercial applications, which means that soon computers may start reasoning about the world rather than just about pixels.

3D Objects

This first section is a tad scattered, acting as a catch-all for computation applied to objects represented with 3D data, inference of 3D object shape from 2D images and Pose Estimation; determining the transformation of an object's 3D pose from 2D images.[95] The process of reconstruction also creeps in ahead of the following section which deals with it explicitly. However, with these points in mind, we present the work which excited our team the most in this general area:

- **OctNet: Learning Deep 3D Representations at High Resolutions**[96] continues the recent development of convolutional networks which operate on 3D data, or Voxels (which are like 3D pixels), using 3D convolutions. OctNet is 'a novel 3D representation which makes deep learning with high-resolution inputs tractable'. The authors test OctNet representations by 'analyzing the impact of resolution on several 3D tasks including 3D object classification, orientation estimation and point cloud labeling.' The paper's central contribution is its exploitation of sparsity in 3D input data which then enables much more efficient use of memory and computation.

- **ObjectNet3D: A Large Scale Database for 3D Object Recognition**[97] - contributes a database for 3D object recognition, presenting 2D images and 3D shapes for 100 object categories. '*Objects in the images in our database [taken from ImageNet] are aligned with the 3D shapes [taken from the ShapeNet repository], and the alignment provides both accurate 3D pose annotation and the closest 3D shape annotation for each 2D object.*' Baseline experiments are provided on: Region proposal generation, 2D object detection, joint 2D detection

[95] Pose Estimation can refer to either just an object's orientation, or both orientation and position in 3D space.

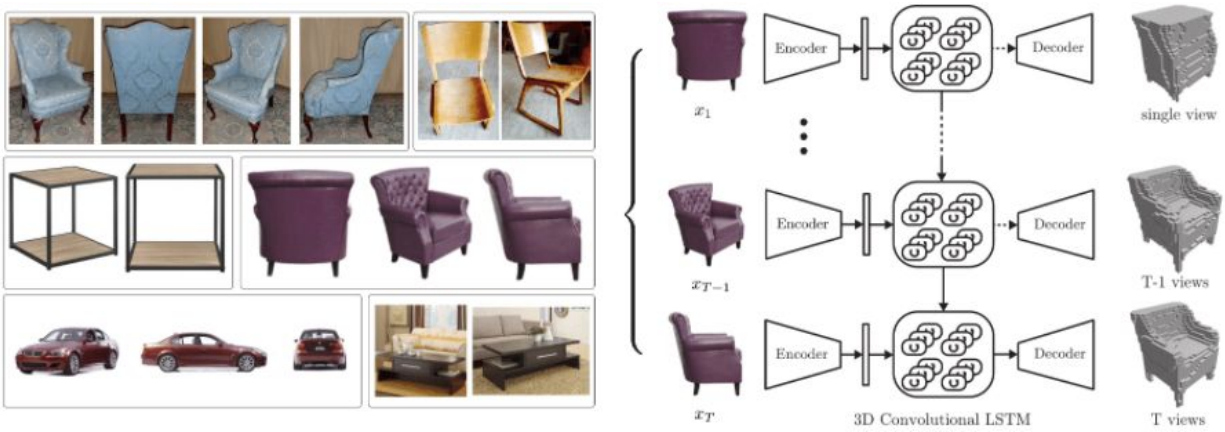
[96] Riegler et al. 2016. OctNet: Learning Deep 3D Representations at High Resolutions. *[Online] arXiv: 1611.05009*. Available: **arXiv:1611.05009v3**

[97] Xiang et al. 2016. ObjectNet3D: A Large Scale Database for 3D Object Recognition. *[Online] Computer Vision and Geometry Lab, Stanford University (cvgl.stanford.edu)*. Available from: http://cvgl.stanford.edu/projects/objectnet3d/

and 3D object pose estimation, and image-based 3D shape retrieval.

- **3D-R2N2: A Unified Approach for Single and Multi-view 3D Object Reconstruction**[98] - creates a reconstruction of an object 'in the form of a 3D occupancy grid using single or multiple images of object instance from arbitrary viewpoints.' Mappings from images of objects to 3D shapes are learned using primarily synthetic data, and the network can train and test without requiring 'any image annotations or object class labels'. The network comprises a 2D-CNN, a 3D Convolutional LSTM (an architecture newly created for purpose) and a 3D Deconvolutional Neural Network. How these different components interact and are trained together end-to-end is a perfect illustration of the layering capable with Neural Networks.

Figure 11: Example of 3D-R2N2 functionality



Note: Images taken from Ebay (left) and an overview of the functionality of 3D-R2N2 (right).
Note from source: Some sample images of the objects we [the authors] wish to reconstruct - notice that views are separated by a large baseline and objects' appearance shows little texture and/or are non-lambertian. (b) An overview of our proposed 3D-R2N2: The network takes a sequence of images (or just one image) from arbitrary (uncalibrated) viewpoints as input (in this example, 3 views of the armchair) and generates voxelized 3D reconstruction as an output. The reconstruction is incrementally refined as the network sees more views of the object.
Source: Choy et al. (2016, p. 3)[99]

3D-R2N2 generates 'rendered images and voxelized models' using ShapeNet models and facilitates 3D object reconstruction where structure from motion (SfM) and simultaneous localisation and mapping (SLAM) approaches typically fail:

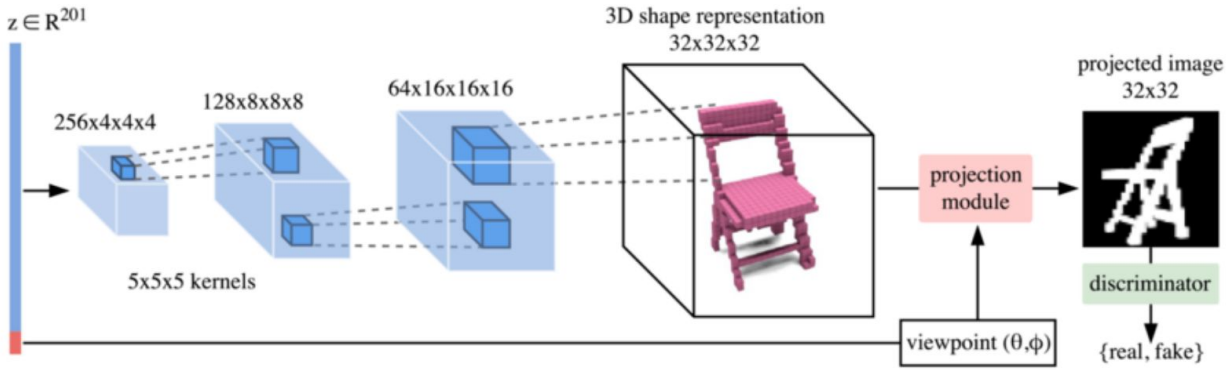
"*Our extensive experimental analysis shows that our reconstruction framework i) outperforms the state-of-the-art methods for single view reconstruction, and ii)*

[98] Choy et al. 2016. 3D-R2N2: A Unified Approach for Single and Multi-view 3D Object Reconstruction. *[Online] arXiv: 1604.00449*. Available: **arXiv:1604.00449v1**
[99] ibid

enables the 3D reconstruction of objects in situations when traditional SFM/SLAM methods fail."

- **3D Shape Induction from 2D Views of Multiple Objects**[100] uses "*Projective Generative Adversarial Networks*" (PrGANs), which train a deep generative model allowing accurate representation of 3D shapes, with the discriminator only being shown 2D images. The projection module captures the 3D representations and converts them to 2D images before passing to the discriminator. Through iterative training cycles the generator improves projections by improving the 3D voxel shapes it generates.

Figure 12: PrGAN architecture segment



Note from source: The PrGAN architecture for generating 2D images of shapes. A 3D voxel representation (32^3) and viewpoint are independently generated from the input z (201-d vector). The projection module renders the voxel shape from a given viewpoint (θ, φ) to create an image. The discriminator consists of 2D convolutional and pooling layers and aims to classify if the input image is generated or real.
Source: Gadhelha et al. (2016, p. 3)[101]

In this way the inference ability is learned through an unsupervised environment:

"*The addition of a projection module allows us to infer the underlying 3D shape distribution without using any 3D, viewpoint information, or annotation during the learning phase.* "

Additionally, the internal representation of the shapes can be interpolated, meaning discrete commonalities in voxel shapes allow transformations from object to object, e.g. from car to aeroplane.

[100] Gadelha et al. 2016. 3D Shape Induction from 2D Views of Multiple Objects. *[Online] arXiv: 1612.058272*. Available: **arXiv:1612.05872v1**
[101] ibid

- **Unsupervised Learning of 3D Structure from Images**[102] presents a completely unsupervised, generative model which demonstrates 'the feasibility of learning to infer 3D representations of the world' for the first time. In a nutshell the DeepMind team present a model which "*learns strong deep generative models of 3D structures, and recovers these structures from 3D and 2D images via probabilistic inference*", meaning that inputs can be both 3D and 2D.

  DeepMind's strong generative model runs on both volumetric and mesh-based representations. The use of Mesh-based representations with OpenGL allows more knowledge to be built in, e.g. how light affects the scene and the materials used. "*Using a 3D mesh-based representation and training with a fully-fledged black-box renderer in the loop enables learning of the interactions between an object's colours, materials and textures, positions of lights, and of other objects*." [103]

  The models are of high quality, capture uncertainty and are amenable to probabilistic inference, allowing for applications in 3D generation and simulation. The team achieve the first quantitative benchmark for 3D density modelling on 3D MNIST and ShapeNet. This approach demonstrates that models may be trained end-to-end unsupervised on 2D images, requiring no ground-truth 3D labels.

Human Pose Estimation and Keypoint Detection

Human Pose Estimation attempts to find the orientation and configuration of human body parts. 2D Human Pose Estimation, or Keypoint Detection, generally refers to localising body parts of humans e.g finding the 2D location of the knees, eyes, feet, etc. However, 3D Pose Estimation takes this even further by finding the orientation of the body parts in 3D space and then an optional step of shape estimation/modelling can be performed. There has been a tremendous amount of improvement across these sub-domains in the last few years.

In terms of competitive evaluation "*the COCO 2016 Keypoint Challenge involves simultaneously detecting people and localizing their keypoints*".[104] The European

[102] Rezende et al. 2016. Unsupervised Learning of 3D Structure from Images. *[Online] arXiv: 1607.00662*. Available: **arXiv:1607.00662v1**
[103] Colyer, A. 2017. Unsupervised learning of 3D structure from images. *[Blog] the morning paper*. Available: https://blog.acolyer.org/2017/01/05/unsupervised-learning-of-3d-structure-from-images/ [Accessed: 04/03/2017].
[104] COCO. 2016. Welcome to the COCO 2016 Keypoint Challenge! *[Online] Common Objects in Common (mscoco.org)*. Available: http://mscoco.org/dataset/#keypoints-challenge2016 [Accessed: 27/01/2017].

Convention on Computer Vision (ECCV)[105] provides more extensive literature on these subjects, however we would like to highlight:

- **Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields**.[106] This method set SOTA performance on the inaugural MSCOCO 2016 keypoints challenge with 60% average precision (AP) and won the best demo award at ECCV, video: Video[107]

- **Keep it SMPL: Automatic Estimation of 3D Human Pose and Shape from a Single Image**.[108] This method first predicts 2D body joint locations and then uses another model called SMPL to create the 3D body shape mesh, which allows it to understand 3D aspects working from 2D pose estimation. The 3D mesh is capable of capturing both pose and shape, versus previous methods which could only find 2D human pose. The authors provide an excellent video analysis of their work here: Video[109]

"*We describe the first method to automatically estimate the 3D pose of the human body as well as its 3D shape from a single unconstrained image. We estimate a full 3D mesh and show that 2D joints alone carry a surprising amount of information about body shape. The problem is challenging because of the complexity of the human body, articulation, occlusion, clothing, lighting, and the inherent ambiguity in inferring 3D from 2D*".[110]

Reconstruction

As mentioned, a previous section presented some examples of reconstruction but with a general focus on objects, specifically their shape and pose. While some of this is technically reconstruction, the field itself comprises many different types of reconstruction, e.g. scene reconstruction, multi-view and single view reconstruction, structure from motion (SfM), SLAM, etc. Furthermore, some reconstruction approaches leverage additional (and multiple) sensors and equipment, such as Event or RGB-D cameras, and can often layer multiple techniques to drive progress.

---

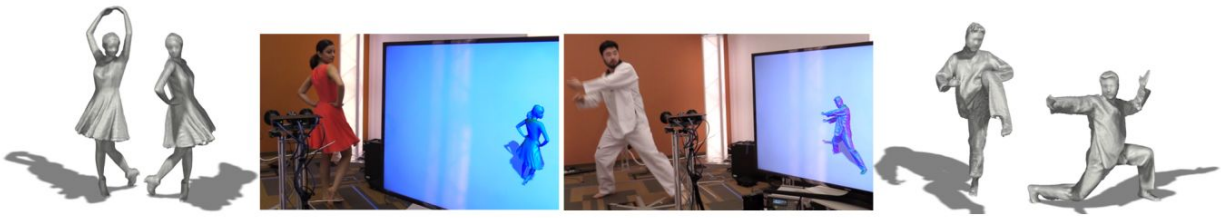
[105] ECCV. 2016. Webpage. *[Online] European Convention on Computer Vision (www.eccv2016.org)*. Available: http://www.eccv2016.org/main-conference/ [Accessed: 26/01/2017].
[106] Cao et al. 2016. Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. *[Online] arXiv: 161108050*. Available: **arXiv:1611.08050v1**
[107] Zhe Cao. 2016. Realtime Multi-Person 2D Human Pose Estimation using Part Affinity Fields, CVPR 2017 Oral. *[Online] YouTube.com*. Available: https://www.youtube.com/watch?v=pW6nZXeWlGM [Accessed: 04/03/2017].
[108] Bogo et al. 2016. Keep it SMPL: Automatic Estimation of 3D Human Pose and Shape from a Single Image. *[Online] arXiv: 1607.08128*. Available: **arXiv:1607.08128v1**
[109] Michael Black. 2016. SMPLify: 3D Human Pose and Shape from a Single Image (ECCV 2016). *[Online] YouTube.com*. Available: https://www.youtube.com/watch?v=eUnZ2rjxGaE [Accessed: 04/03/2017].
[110] ibid

The result? Whole scenes can be reconstructed non-rigidly and change spatio-temporally, e.g. a high-fidelity reconstruction of yourself, and your movements, updated in real-time.

As identified previously, issues persist around the mapping of 2D images to 3D space. The following papers present a plethora of approaches to create high-fidelity, real-time reconstructions:

- **Fusion4D: Real-time Performance Capture of Challenging Scenes**[111] veers towards the domain of Computer Graphics, however the interplay between Computer Vision and Graphics cannot be overstated. The authors' approach uses RGB-D and Segmentation as inputs to form a real-time, multi-view reconstruction which is outputted using Voxels.

Figure 13: Fusion4D examples from real-time feed



Note from source: "*We present a new method for real-time high quality 4D (i.e. spatio-temporally coherent) performance capture, allowing for incremental non-rigid reconstruction from noisy input from multiple RGBD cameras. Our system demonstrates unprecedented reconstructions of challenging non-rigid sequences, at real-time rates, including robust handling of large frame-to-frame motions and topology changes.*"
Source: Dou et al. (2016, p. 1)[112]

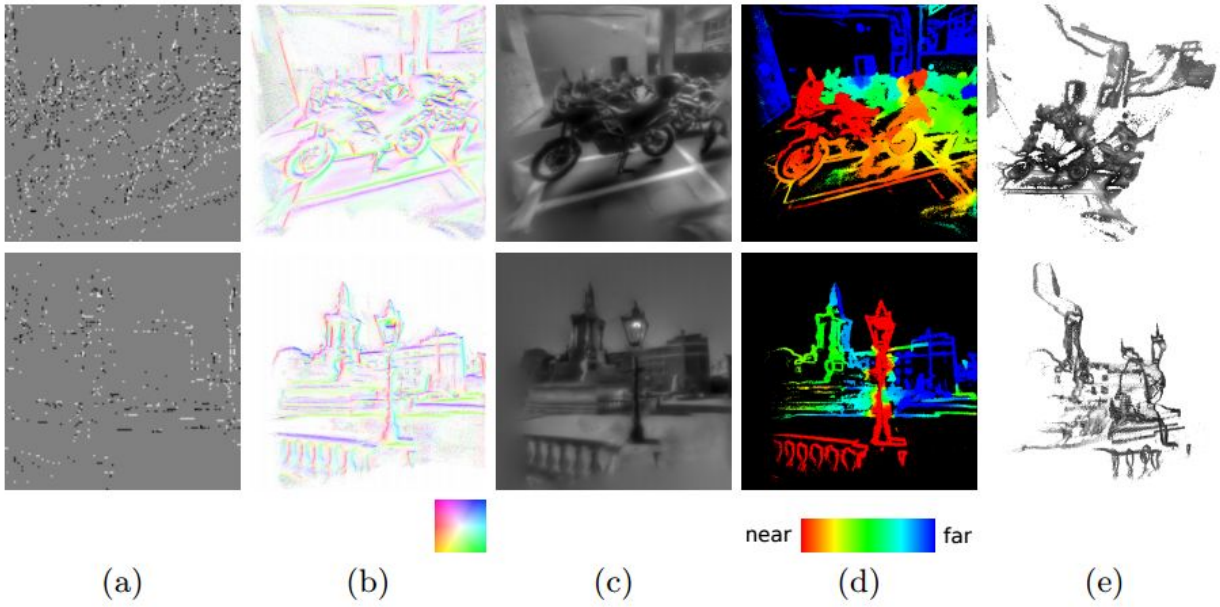
Fusion4D creates real-time, high fidelity voxel representations which have impressive applications in virtual reality, augmented reality and telepresence. This work from Microsoft will likely revolutionise motion capture, possibly for live sports. An example of the technology in real-time use is available here: Video[113]

[111] Dou et al. 2016. Fusion4D: Real-time Performance Capture of Challenging Scenes. *[Online] SamehKhamis.com*. Available: http://www.samehkhamis.com/dou-siggraph2016.pdf
[112] ibid
[113] Microsoft Research. 2016. Fusion4D: Real-time Performance Capture of Challenging Scenes. *[Online] YouTube.com*. Available: https://www.youtube.com/watch?v=2dkcJ1YhYw4&feature=youtu.be [Accessed: 04/03/2017].

For an astounding example of telepresence/holoportation by Microsoft, see here: Video[114]

- **Real-Time 3D Reconstruction and 6-DoF Tracking with an Event Camera**[115] won best paper at the European Convention on Computer Vision (ECCV) in 2016. The authors propose a novel algorithm capable of tracking 6D motion and various reconstructions in real-time using a single Event Camera.

Figure 14: Examples of the Real-Time 3D Reconstruction



Note from source: Demonstrations in various settings of the different aspects of our joint estimation algorithm. (a) visualisation of the input event stream; (b) estimated gradient keyframes; (c) reconstructed intensity keyframes with super resolution and high dynamic range properties; (d) estimated depth maps; (e) semi-dense 3D point clouds.
Source: Kim et al. (2016, p. 12)[116]

The Event camera is gaining favour with researchers in Computer Vision due to its reduced latency, lower power consumption and higher dynamic range when compared to traditional cameras. Instead of a sequence of frames outputted by a regular camera, the event camera outputs "*a stream of asynchronous spikes, each with pixel location, sign and precise timing, indicating when individual pixels record a threshold log intensity change.*"[117]

---

[114] I3D Past Projects. 2016. holoportation: virtual 3D teleportation in real-time (Microsoft Research). *[Online] YouTube.com*. Available: https://www.youtube.com/watch?v=7d59O6cfaM0&feature=youtu.be [Accessed: 03/03/2017].
[115] Kim et al. 2016. Real-Time 3D Reconstruction and 6-DoF Tracking with an Event Camera. *[Online] Department of Computer, Imperial College London (www.doc.ic.ac.uk)*. Available: https://www.doc.ic.ac.uk/~ajd/Publications/kim_etal_eccv2016.pdf
[116] ibid
[117] Kim et al. 2014. Simultaneous Mosaicing and Tracking with an Event Camera. *[Online] Department of Computer, Imperial College London (www.doc.ic.ac.uk)*. Available: https://www.doc.ic.ac.uk/~ajd/Publications/kim_etal_bmvc2014.pdf

For an explanation of event camera functionality, real-time 3D reconstruction and 6-DoF tracking, see the paper's accompanying video here: Video[118]

This approach is incredibly impressive when one considers the real-time image rendering and depth estimation involved using a single view-point:
"*We propose a method which can perform real-time 3D reconstruction from a single hand-held event camera with no additional sensing, and works in unstructured scenes of which it has no prior knowledge.*"

- **Unsupervised CNN for Single View Depth Estimation: Geometry to the Rescue[119]** proposes an unsupervised method for training a deep CNN for single view depth prediction with results comparable to SOTA using supervised methods. Traditional deep CNN approaches for single view depth prediction require large amounts of manually labelled data, however unsupervised methods again demonstrate their value by removing this necessity. The authors achieve this "*by training the network in a manner analogous to an autoencoder*", using a stereo-rig.

Other uncategorised 3D

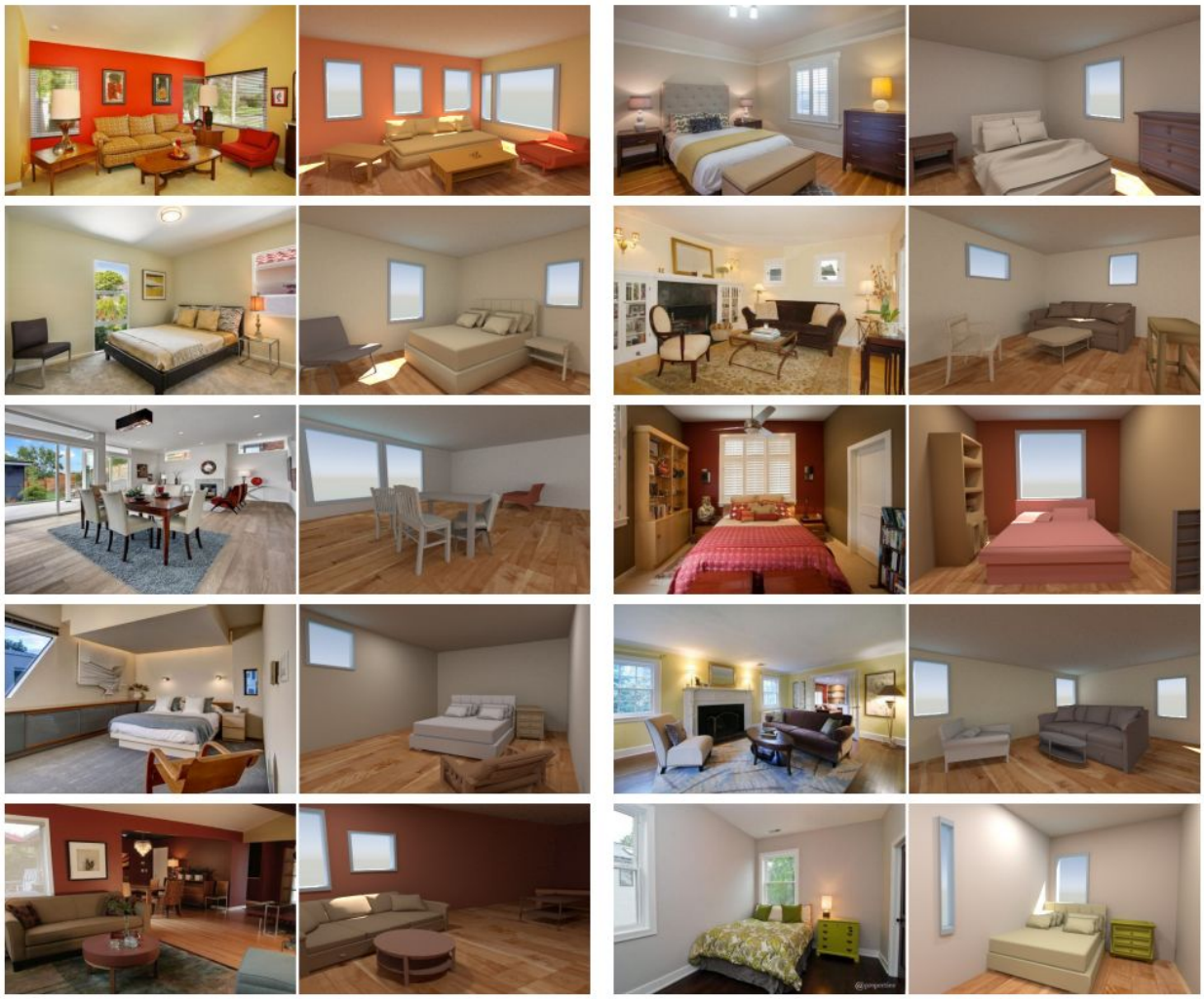
- **IM2CAD[120]** describes the process of transferring an 'image to CAD model', CAD meaning computer-assisted design, which is a prominent method used to create 3D scenes for architectural depictions, engineering, product design and many other fields.

  "*Given a single photo of a room and a large database of furniture CAD models, our goal is to reconstruct a scene that is as similar as possible to the scene depicted in the photograph, and composed of objects drawn from the database.*"

  The authors present an automatic system which 'iteratively optimizes object placements and scales' to best match input from real images. The rendered scenes validate against the original images using metrics trained using deep CNNs.

[118] Hanme Kim. 2017. Real-Time 3D Reconstruction and 6-DoF Tracking with an Event. *[Online] YouTube.com*. Available: https://www.youtube.com/watch?v=yHLyhdMSw7w [Accessed: 03/03/2017].
[119] Garg et al. 2016. Unsupervised CNN for Single View Depth Estimation: Geometry to the Rescue. *[Online] arXiv: 1603.04992*. Available: **arXiv:1603.04992v2**
[120] Izadinia et al. 2016. IM2CAD. *[Online] arXiv: 1608.05137*. Available: **arXiv:1608.05137v1**

Figure 15: Example of IM2CAD rendering bedroom scene



Note: Left: input image. Right: Automatically created CAD model from input.
Note from source: The reconstruction results. In each example the left image is the real input image and the right image is the rendered 3D CAD model produced by IM2CAD.
Source: Izadinia et al. (2016, p. 10)[121]

Why care about IM2CAD?

The issue tackled by the authors is one of the first meaningful advancements on the techniques demonstrated by Lawrence Roberts in 1963, which allowed inference of a 3D scene from a photo using a known-object database, albeit in the very simple case of line drawings.

"*While Robert's method was visionary, more than a half century of subsequent research in Computer Vision has still not yet led to practical extensions of his approach that work reliably on realistic images and scenes.*"

The authors introduce a variant of the problem, aiming to reconstruct a high

[121] ibid

fidelity scene from a photo using '*objects taken from a database of 3D object models*' for reconstruction.

The process behind IM2CAD is quite involved and includes:

- A Fully Convolutional Network that is trained end-to-end to find Geometric Features for Room Geometry Estimation.
- Faster R-CNN for Object Detection.
- After finding the objects within the image, CAD Model Alignment is completed to find the closest models within the ShapeNet repository for the detected objects. For example, the type of chair, given shape and approximate 3D pose. Each 3D model is rendered to 32 viewpoints which are then compared with the bounding box generated in object detection using deep features[122].
- Object Placement in the Scene
- Finally Scene Optimization further refines the placement of the objects by optimizing the visual similarity between the camera views of the rendered scene and input image.

Again in this domain, ShapeNet proves invaluable:
"*First, we leverage ShapeNet, which contains millions of 3D models of objects, including thousands of different chairs, tables, and other household items. This dataset is a game changer for 3D scene understanding research, and was key to enabling our work.*"

- **Learning Motion Patterns in Videos**[123] proposes to solve the issue of determining object motion independent of camera movement using synthetic video sequences to teach the networks. "*The core of our approach is a fully convolutional network, which is learnt entirely from synthetic video sequences, and their ground-truth optical flow and motion segmentation.*" The authors test their approach on the new moving object segmentation dataset called DAVIS,[124] as well as the Berkeley motion segmentation dataset and achieve SOTA on both.

- **Deep Image Homography Estimation**[125] comes from the Magic Leap team, a secretive US startup working in Computer Vision and Mixed Reality. The authors reclassify the task of homography estimation as '*a learning problem*' and present two deep CNNs architectures which form "*HomographyNet: a regression network which directly estimates the real-valued homography parameters, and a*

[122] Yet more neural network spillover
[123] Tokmakov et al. 2016. Learning Motion Patterns in Videos. *[Online] arXiv: 1612.07217*. Available: **arXiv:1612.07217v1**
[124] DAVIS. 2017. DAVIS: Densely Annotated Video Segmentation. *[Website] DAVIS Challenge*. Available: http://davischallenge.org/ [Accessed: 27/03/2017].
[125] DeTone et al. 2016. Deep Image Homography Estimation. *[Online] arXiv: 1606.03798*. Available: **arXiv:1606.03798v1**

classification network which produces a distribution over quantized homographies."

The term homography comes from projective geometry and refers to a type of transformation that maps one plane to another. '*Estimating a 2D homography from a pair of images is a fundamental task in computer vision, and an essential part of monocular SLAM systems*'.

The authors also provide a method for producing a "*seemingly infinite dataset*", from existing datasets of real images such as MS-COCO, which offsets some of data requirements of deeper networks. They manage to create "*a nearly unlimited number of labeled training examples by applying random projective transformations to a large image dataset*".

- **gvnn: Neural Network Library for Geometric Computer Vision**[126] introduces a new neural network library for Torch, a popular computing framework for machine learning. Gvnn aims to 'bridge the gap between classic geometric computer vision and deep learning'. The gvnn library allows developers to add geometric capabilities to their existing networks and training methods.

  "*In this work, we build upon the 2D transformation layers originally proposed in the spatial transformer networks and provide various novel extensions that perform geometric transformations which are often used in geometric computer vision.*"

  "*This opens up applications in learning invariance to 3D geometric transformation for place recognition, end-to-end visual odometry, depth estimation and unsupervised learning through warping with a parametric transformation for image reconstruction error.*"

In summation

Throughout this section we cut a swath across the field of 3D understanding, focusing primarily on the areas of Pose Estimation, Reconstruction, Depth Estimation and Homography. But there is considerably more superb work which will go unmentioned by us, constrained as we are by volume. And so, we hope to have provided the reader with a valuable starting point, which is to say by no means an absolute.

A large portion of the highlighted work may be classified under Geometric Vision, which generally deals with measuring real-world quantities like distances, shapes, areas and

[126] Handa et al. 2016. gvnn: Neural Network Library for Geometric Computer Vision. *[Online] arXiv: 1607.07405.* Available: **arXiv:1607.07405v3**

volumes directly from images. Our heuristic is that recognition-based tasks focus more on higher level semantic information than typically concerns applications in Geometric Vision. However, often we find that much of these different areas of 3D understanding are inextricably linked.

One of the largest Geometric problems is that of simultaneous localisation and mapping (SLAM), with researchers considering whether SLAM will be in the next problems tackled by Deep Learning. Skeptics of the so-called 'universality' of deep learning, of which there are many, point to the importance and functionality of SLAM as an algorithm:

"*Visual SLAM algorithms are able to simultaneously build 3D maps of the world while tracking the location and orientation of the camera.*"[127] The geometric estimation portion of the SLAM approach is not currently suited to deep learning approaches and end-to-end learning remains unlikely. SLAM represents one of the most important algorithms in robotics and was designed with large input from the Computer Vision field. The technique has found its home in applications like Google Maps, autonomous vehicles, AR devices like Google Tango[128] and even the Mars Rover.

That being said, Tomasz Malisiewicz delivers the anecdotal aggregate opinion of some prominent researchers on the issue, who agree "*that semantics are necessary to build bigger and better SLAM systems.*"[129] This potentially shows promise for future applications of Deep Learning in the SLAM domain.

We reached out to Mark Cummins, co-founder of Plink and Pointy, who provided us with his thoughts on the issue. Mark completed his PhD on SLAM techniques:

"*The core geometric estimation part of SLAM is pretty well solved by the current approaches, but the high-level semantics and the lower-level system components can all benefit from deep learning. In particular:*

- *Deep learning can greatly improve the quality of map semantics - i.e. going beyond poses or point clouds to a full understanding of the different kind of objects or regions in the map. This is much more powerful for many applications, and can also help with general robustness (for example through better handling*

[127] Malisiewicz. 2016. The Future of Real-Time SLAM and Deep Learning vs SLAM. *[Blog] Tombone's Computer Vision Blog*. Available: http://www.computervisionblog.com/2016/01/why-slam-matters-future-of-real-time.html [Accessed: 01/03/2017].
[128] Google. 2017. Tango. *[Website] get.google.com*. Available: https://get.google.com/tango/ [Accessed: 23/03/2017].
[129] ibid

of dynamic objects and environmental changes).

- *At a lower level, many components can likely be improved via deep learning. Obvious candidates are place recognition / loop closure detection / relocalization, better point descriptors for sparse SLAM methods, etc*

Overall the structure of SLAM solvers probably remains the same, but the components improve. It is possible to imagine doing something radically new with deep learning, like throwing away the geometry entirely and have a more recognition-based navigation system. But for systems where the goal is a precise geometric map, deep learning in SLAM is likely more about improving components than doing something completely new."

In summation, we believe that SLAM is not likely to be completely replaced by Deep Learning. However, it is entirely likely that the two approaches may become complements to each other going forward. If you wish to learn more about SLAM, and its current SOTA, we wholeheartedly recommend Tomasz Malisiewicz's blog for that task: The Future of Real-Time SLAM and Deep Learning vs SLAM[130]

Part Four: ConvNet Architectures, Datasets, Ungroupable Extras

ConvNet Architectures

ConvNet architectures have recently found many novel applications outside of Computer Vision, some of which will feature in our forthcoming publications. However, they continue to feature prominently in Computer Vision, with architectural advancements providing improvements in speed, accuracy and training for many of the aforementioned applications and tasks in this paper.

For this reason, ConvNet architectures are of fundamental importance to Computer Vision as a whole. The following features some noteworthy ConvNet architectures from 2016, many of which take inspiration from the recent success of ResNets.

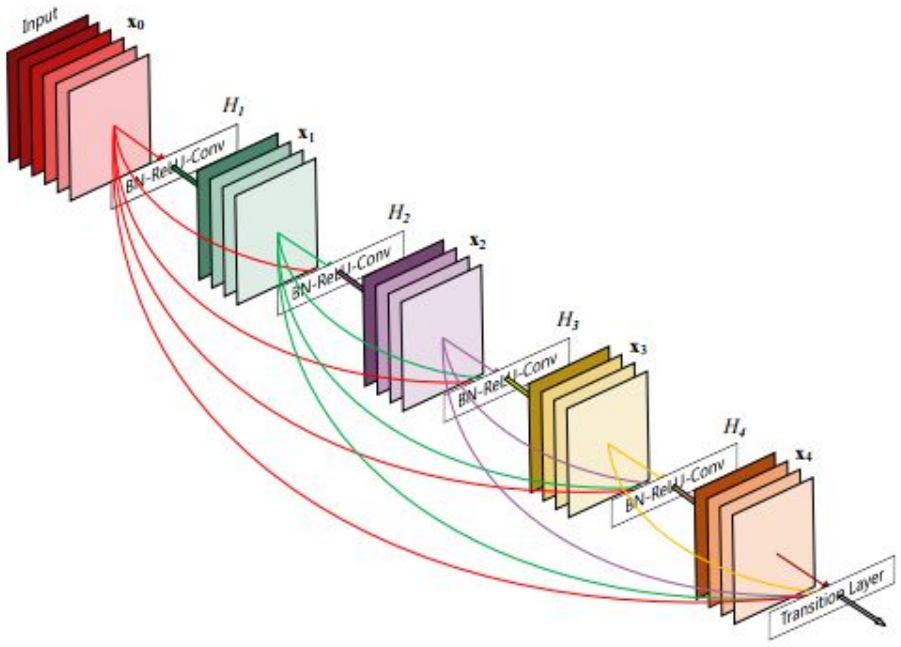
[130] Malisiewicz. 2016. The Future of Real-Time SLAM and Deep Learning vs SLAM. *[Blog] Tombone's Computer Vision Blog*. Available: http://www.computervisionblog.com/2016/01/why-slam-matters-future-of-real-time.html [Accessed: 01/03/2017].

- **Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning**[131] - present Inception v4, a new Inception architecture which builds on the Inception v2 and v3 from the end of 2015.[132] The paper also provides an analysis of using residual connections for training Inception Networks along with some Residual-Inception hybrid networks.

- **Densely Connected Convolutional Networks**[133] or "DenseNets" take direct inspiration from the identity/skip connections of ResNets. The approach extends this concept to ConvNets by having each layer connect to every other layer in a feed forward fashion, sharing feature maps from previous layers as inputs, thus creating DenseNets.

  "*DenseNets have several compelling advantages: they alleviate the vanishing-gradient problem, strengthen feature propagation, encourage feature reuse, and substantially reduce the number of parameters*".[134]

Figure 16: Example of DenseNet Architecture



Note: A 5-layer dense block with a growth rate of $k = 4$. Each layer takes all preceding feature-maps as input.
Source: Huang et al. (2016)[135]

[131] Szegedy et al. 2016. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. *[Online] arXiv: 1602.07261*. Available: **arXiv:1602.07261v2**
[132] Szegedy et al. 2015. Rethinking the Inception Architecture for Computer Vision. *[Online] arXiv: 1512.00567*. Available: **arXiv:1512.00567v3**
[133] Huang et al. 2016. Densely Connected Convolutional Networks. *[Online] arXiv: 1608.06993*. Available: **arXiv:1608.06993v3**
[134] ibid
[135] ibid

The model was evaluated on CIFAR-10, CIFAR-100, SVHN and ImageNet; it achieved SOTA on a number of them. Impressively, DenseNets achieve these results while using less memory and with reduced computational requirements. There are multiple implementations (Keras, Tensorflow, etc) here.[136]

- **FractalNet Ultra-Deep Neural Networks without Residuals**[137] - utilises interacting subpaths of different lengths, without pass-through or residual connections, instead altering internal signals using filter and nonlinearities for transformations.

  "*FractalNets repeatedly combine several parallel layer sequences with different numbers of convolutional blocks to obtain a large nominal depth, while maintaining many short paths in the network*".[138]

  The network achieved SOTA performance on CIFAR and ImageNet, while demonstrating some additional properties. For instance, they call into question the role of residuals in the success of extremely deep ConvNets, while also providing insight into the nature of answers attained by various subnetwork depths.

- **Lets keep it simple: using simple architectures to outperform deeper architectures**[139] focuses on creating a simplified mother architecture. The architecture achieved SOTA results, or parity with existing approaches, on 'datasets such as CIFAR10/100, MNIST and SVHN with simple or no data-augmentation'. We feel their exact words provide the best description of the motivation here:

  "*In this work, we present a very simple fully convolutional network architecture of 13 layers, with minimum reliance on new features which outperforms almost all deeper architectures with 2 to 25 times fewer parameters. Our architecture can be a very good candidate for many scenarios, especially for use in embedded devices."*

  "It can be furthermore compressed using methods such as DeepCompression and thus its memory consumption can be decreased drastically. We intentionally tried to create a mother architecture with minimum reliance on new features

[136] Liuzhuang13. 2017. Code for Densely Connected Convolutional Networks (DenseNets). *[Online] github.com*. Available: https://github.com/liuzhuang13/DenseNet [Accessed: 03/04/2017].
[137] Larsson et al. 2016. FractalNet: Ultra-Deep Neural Networks without Residuals. *[Online] arXiv: 1605.07648*. Available: **arXiv:1605.07648v2**
[138] Huang et al. 2016. Densely Connected Convolutional Networks. *[Online] arXiv: 1608.06993*. Available: **arXiv:1608.06993v3**, pg. 1.
[139] Hossein HasanPour et al. 2016. Lets keep it simple: using simple architectures to outperform deeper architectures. *[Online] arXiv: 1608.06037*. Available: **arXiv:1608.06037v3**

proposed recently, to show the effectiveness of a well-crafted yet simple convolutional architecture which can then later be enhanced with existing or new methods presented in the literature."[140]

Here are some additional techniques which complement ConvNet Architectures:

- **Swapout: Learning an ensemble of deep architectures**[141] generalises dropout and stochastic depth methods to prevent co-adaptation of units, both in a specific layer and across network layers. The ensemble training method samples from multiple architectures including "*dropout, stochastic depth and residual architectures*". Swapout outperforms ResNets of identical network structure on the CIFAR-10 and CIFAR-100 and can be classified as a regularisation technique.

- **SqueezeNet**[142] posits that smaller DNNs offer various benefits, from less computationally taxing training to easier information transmission to, and operation on, devices with limited storage or processing power. SqueezeNet is a small DNN architecture which achieves 'AlexNet-level accuracy with significantly reduced parameters and memory requirements using model compression techniques which make it 510x smaller than AlexNet.'

A Rectified Linear Unit (ReLU) is traditionally the dominant activation function for all Neural Networks. However, here are some recent alternatives:

- **Concatenated Rectified Linear Units (CRelu)**[143]
- **Exponential Linear Units (ELUs)**[144] from the close of 2015
- **Parametric Exponential Linear Unit (PELU)**[145]

Moving towards equivariance in ConvNets

ConvNets are translation invariant - meaning they can identify the same features in multiple parts of an image. However, the typical CNN isn't rotation invariant - meaning that if a feature or the whole image is rotated then the network's performance suffers. Usually ConvNets learn to (sort of) deal with rotation invariance through data augmentation (e.g. purposefully rotating the images by small random amounts during

[140] ibid
[141] Singh et al. 2016. Swapout: Learning an ensemble of deep architectures. *[Online] arXiv: 1605.06465*. Available: **arXiv:1605.06465v1**
[142] Iandola et al. 2016. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size. *[Online] arXiv: 1602.07360*. Available: **arXiv:1602.07360v4**
[143] Shang et al. 2016. Understanding and Improving Convolutional Neural Networks via Concatenated Rectified Linear Units. *[Online] arXiv: 1603.05201*. Available: **arXiv:1603.05201v2**
[144] Clevert et al. 2016. Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs). *[Online] arXiv: 1511.07289*. Available: **arXiv:1511.07289v5**
[145] Trottier et al. 2016. Parametric Exponential Linear Unit for Deep Convolutional Neural Networks. *[Online] arXiv: 1605.09332*. Available: **arXiv:1605.09332v3**

training). This means the network gains slight rotation invariant properties without specifically designing rotation invariance into the network. This means that rotation invariance is fundamentally limited in networks using current techniques. This is an interesting parallel with humans who also typically fare worse at recognising characters upside down, although there is no reason for machines to suffer this limitation.

The following papers tackle **rotation-invariant ConvNets**. While each approach has novelties, they all improve rotation invariance through more efficient parameter usage leading to eventual global rotation equivariance:

- **Harmonic CNNs**[146] replace regular CNN filters with 'circular harmonics'.
- **Group Equivariant Convolutional Networks (G-CNNs)**[147] uses G-Convolutions, which are a new type of layer that "*enjoys a substantially higher degree of weight sharing than regular convolution layers and increases the expressive capacity of the network without increasing the number of parameters.*"
- **Exploiting Cyclic Symmetry in Convolutional Neural Networks**[148] presents four operations as layers which augment neural network layers to partially increase rotational equivariance.
- **Steerable CNNs**[149] - Cohen and Welling build on the work they did with **G-CNNs**, demonstrating that "*steerable architectures"* outperform residual and dense networks on the CIFARs. They also provide a succinct overview of the invariance problem:

  "*To improve the statistical efficiency of machine learning methods, many have sought to learn invariant representations. In deep learning, however, intermediate layers should not be fully invariant, because the relative pose of local features must be preserved for further layers. Thus, one is led to the idea of* ***equivariance****: a network is equivariant if the representations it produces transform in a predictable linear manner under transformations of the input. In other words, equivariant networks produce representations that are steerable. Steerability makes it possible to apply filters not just in every position (as in a standard convolution layer), but in every pose, thus allowing for increased parameter sharing.*"[107]

[146] Worrall et al. 2016. Harmonic Networks: Deep Translation and Rotation Equivariance. *[Online] arXiv: 1612.04642*. Available: **arXiv:1612.04642v1**
[147] Cohen & Welling. 2016. Group Equivariant Convolutional Networks. *[Online] arXiv: 1602.07576*. Available: **arXiv:1602.07576v3**
[148] Dieleman et al. 2016. Exploiting Cyclic Symmetry in Convolutional Neural Networks. *[Online] arXiv: 1602.02660*. Available: **arXiv:1602.02660v2**
[149] Cohen & Welling. 2016. Steerable CNNs. *[Online] arXiv: 1612.08498*. Available: **arXiv:1612.08498v1**

Residual Networks

Figure 17: Test-Error Rates on CIFAR Datasets

method				CIFAR-10(%)	CIFAR-100(%)
NIN[19]				8.81	35.68
DSN[18]				8.22	34.57
FitNet[22]				8.39	35.04
Highway[30]				7.72	32.39
All-CNN[27]				7.25	33.71
ELU[2]				6.55	24.28
method	depth	k,(w)	#parameters		
resnet[7]	110	1	1.7M	6.43 (6.61±0.16)	25.16
	1202	1	19.4M	7.93	27.82
pre-resnet[9]	110	1	1.7M	6.37	-
	164	1	1.7M	5.46	24.33
	1001	1	10.2M	4.62 (4.69±0.20)†	22.71 (22.68±0.22)
stoch-depth[12]	110	1	1.7M	5.25	24.58
	1001	1	10.2M	4.91	-
swapout[26]	20	1,(2)	1.1M	6.58	25.86
	32	1,(4)	7.43M	4.76	22.72
wide-resnet[34]	40	1,(4)	8.7M	4.97	22.89
	16	1,(8)	11.0M	4.81	22.07
	28	1,(10)	36.5M	4.17	20.50
DenseNet[11]†	100	1	7.0M	4.10	20.20
	100	1	27.2M	3.74	**19.25**
multi-resnet [ours]†	200	5	10.2M	**4.35** (4.36±0.04)	**20.42** (20.44±0.15)
	398	5	20.4M	3.92	20.59
	26	2,(10)	72M	3.96	**19.45**
	26	4,(10)	145M	**3.73**	19.60

Note: Yellow highlight indicates that these papers feature within this piece. Pre-resnet refers to "*Identity Mappings in Deep Residual Networks*" (see following section). Furthermore, while not included in the table we believe that *"**Learning Identity Mappings with Residual Gates***" produced some of the lowest error rates of 2016 with 3.65% and 18.27% on CIFAR-10 and CIFAR-100, respectively.
Source: Abdi and Nahavandi (2016, p. 6)[150]

Residual Networks and their variants became incredibly popular in 2016, following the success of Microsoft's ResNet,[151] with many open source versions and pre-trained models now available. In 2015, ResNet won 1st place in ImageNet's Detection, Localisation and Classification tasks as well as in COCO's Detection and Segmentation challenges. Although questions still abound about depth, ResNets tackling of the vanishing gradient problem provided more impetus for the "*increased depth produces superior abstraction*" philosophy which underpins much of Deep Learning at present.

ResNets are often conceptualised as an ensemble of shallower networks, which somewhat counteract the hierarchical nature of Deep Neural Networks (DNNs) by

[150] Abdi, M., Nahavandi, S. 2016. Multi-Residual Networks: Improving the Speed and Accuracy of Residual Networks. *[Online] arXiv: 1609.05672*. Available: **arXiv:1609.05672v3**
[151] He et al. 2015. Deep Residual Learning for Image Recognition. *[Online] arXiv: 1512.03385*. Available: **arXiv:1512.03385v1**

running shortcut connections parallel to their convolutional layers. These shortcuts or ***skip connections*** mitigate vanishing/exploding gradient problems associated with DNNs, by allowing easier back-propagation of gradients throughout the network layers. For more information there is a Quora thread available here.[152]

Residual Learning, Theory and Improvements

- **Wide Residual Networks**[153] is now an extremely common ResNet approach. The authors conduct an experimental study on the architecture of ResNet blocks, and improve residual network performance by increasing the width and reducing the depth of the networks, which mitigates the diminishing feature reuse problem. This approach produces new SOTA on multiple benchmarks including 3.89% and 18.3% on CIFAR-10 and CIFAR-100 respectively. The authors show that a '16-layer-deep wide ResNet performs as well or better in accuracy and efficiency than many other ResNets (including 1000 layer networks)'.

- **Deep Networks with Stochastic Depth**[154] essentially applies dropout to whole layers of neurons instead of to bunches of individual neurons. "*We start with very deep networks but during training, for each mini-batch, randomly drop a subset of layers and bypass them with the identity function.*" Stochastic depth allows quicker training and better accuracy even when training networks greater than 1200 layers.

- **Learning Identity Mappings with Residual Gates**[155] - "*by using a scalar parameter to control each gate, we provide a way to learn identity mappings by optimizing only one parameter.*" The authors use these Gated ResNets to improve the optimisation of deep models, while providing 'high tolerance to full layer removal' such that 90% of performance remains following significant removal at random. Using Wide Gated ResNets the model achieves 3.65% and 18.27% error on CIFAR- 10 and CIFAR-100, respectively.

- **Residual Networks Behave Like Ensembles of Relatively Shallow Networks** [156] - ResNets can be viewed as collections of many paths, which don't strongly

[152] Quora. 2017. What is an intuitive explanation of Deep Residual Networks? *[Website]* www.quora.com. Available: https://www.quora.com/What-is-an-intuitive-explanation-of-Deep-Residual-Networks [Accessed: 03/04/2017].
[153] Zagoruyko, S. and Komodakis, N. 2017. Wide Residual Networks. *[Online] arXiv: 1605.07146*. Available: **arXiv:1605.07146v3**
[154] Huang et al. 2016. Deep Networks with Stochastic Depth. *[Online] arXiv: 1603.09382*. Available: **arXiv:1603.09382v3**
[155] Savarese et al. 2016. Learning Identity Mappings with Residual Gates. *[Online] arXiv: 1611.01260*. Available: **arXiv:1611.01260v2**
[156] Veit, Wilber and Belongie. 2016. Residual Networks Behave Like Ensembles of Relatively Shallow Networks. *[Online] arXiv: 1605.06431*. Available: **arXiv:1605.06431v2**

depend upon one another and hence reinforce the notion of ensemble behaviour. Furthermore, residual pathways vary in length with the short paths contributing to gradient during training while the deeper paths don't factor in this stage.

- **Identity Mappings in Deep Residual Networks**[157] comes as an improvement from the original Resnet authors: Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun. Identity mappings are shown to allow 'forward and backward signals to be propagated between any ResNet block when used as the skip connections and after-addition activation'. The approach improves generalisation, training and results "*using a 1001-layer ResNet on CIFAR-10 (4.62% error) and CIFAR-100, and a 200-layer ResNet on ImageNet.*"

- **Multi-Residual Networks: Improving the Speed and Accuracy of Residual Networks**[158] again advocates for the ensemble behaviour of ResNets and favours a wider-over-deeper approach to ResNet architecture. "*The proposed multi-residual network increases the number of residual functions in the residual blocks.*" Improved accuracy produces 3.73% and 19.45% error on CIFAR-10 and CIFAR-100, respectively. The table presented in Fig. 17 was taken from this paper, and more up-to-date versions are available which consider the work produced in 2017 thus far.

Other residual theory and improvements

Although a relatively recent idea, there is quite a considerable body of work being created around ResNets presently. The following represents some additional theories and improvements which we wished to highlight for interested readers:

- Highway and Residual Networks learn Unrolled Iterative Estimation[159]
- Residual Networks of Residual Networks: Multilevel Residual Networks[160]
- Resnet in Resnet: Generalizing Residual Architectures[161]
- Wider or Deeper: Revisiting the ResNet Model for Visual Recognition[162]

[157] He at al. 2016. Identity Mappings in Deep Residual Networks. *[Online] arXiv: 1603.05027*. Available: **arXiv:1603.05027v3**
[158] Abdi, M., Nahavandi, S. 2016. Multi-Residual Networks: Improving the Speed and Accuracy of Residual Networks. *[Online] arXiv: 1609.05672*. Available: **arXiv:1609.05672v3**
[159] Greff et al. 2017. Highway and Residual Networks learn Unrolled Iterative Estimation. *[Online] arXiv: 1612. 07771*. Available: **arXiv:1612.07771v3**
[160] Abdi and Nahavandi. 2017. Multi-Residual Networks: Improving the Speed and Accuracy of Residual Networks. *[Online] 1609.05672*. Available: **arXiv:1609.05672v4**
[161] Targ et al. 2016. Resnet in Resnet: Generalizing Residual Architectures. *[Online] arXiv: 1603.08029*. Available: **arXiv:1603.08029v1**
[162] Wu et al. 2016. Wider or Deeper: Revisiting the ResNet Model for Visual Recognition. *[Online] arXiv: 1611.10080*. Available: **arXiv:1611.10080v1**

- Bridging the Gaps Between Residual Learning, Recurrent Neural Networks and Visual Cortex[163]
- Convolutional Residual Memory Networks[164]
- Identity Matters in Deep Learning[165]
- Deep Residual Networks with Exponential Linear Unit[166]
- Weighted Residuals for Very Deep Networks[167]

Datasets

The significance of rich datasets for all facets of machine learning cannot be overstated. Hence, we feel it is prudent to include some of the largest advancements in this domain. To paraphrase Ben Hamner, the CTO and co-founder of Kaggle, '*a new dataset can make a thousand papers flourish*',[168] that is to say the availability of data can promote new approaches, as well as breath new life into previously ineffectual techniques.

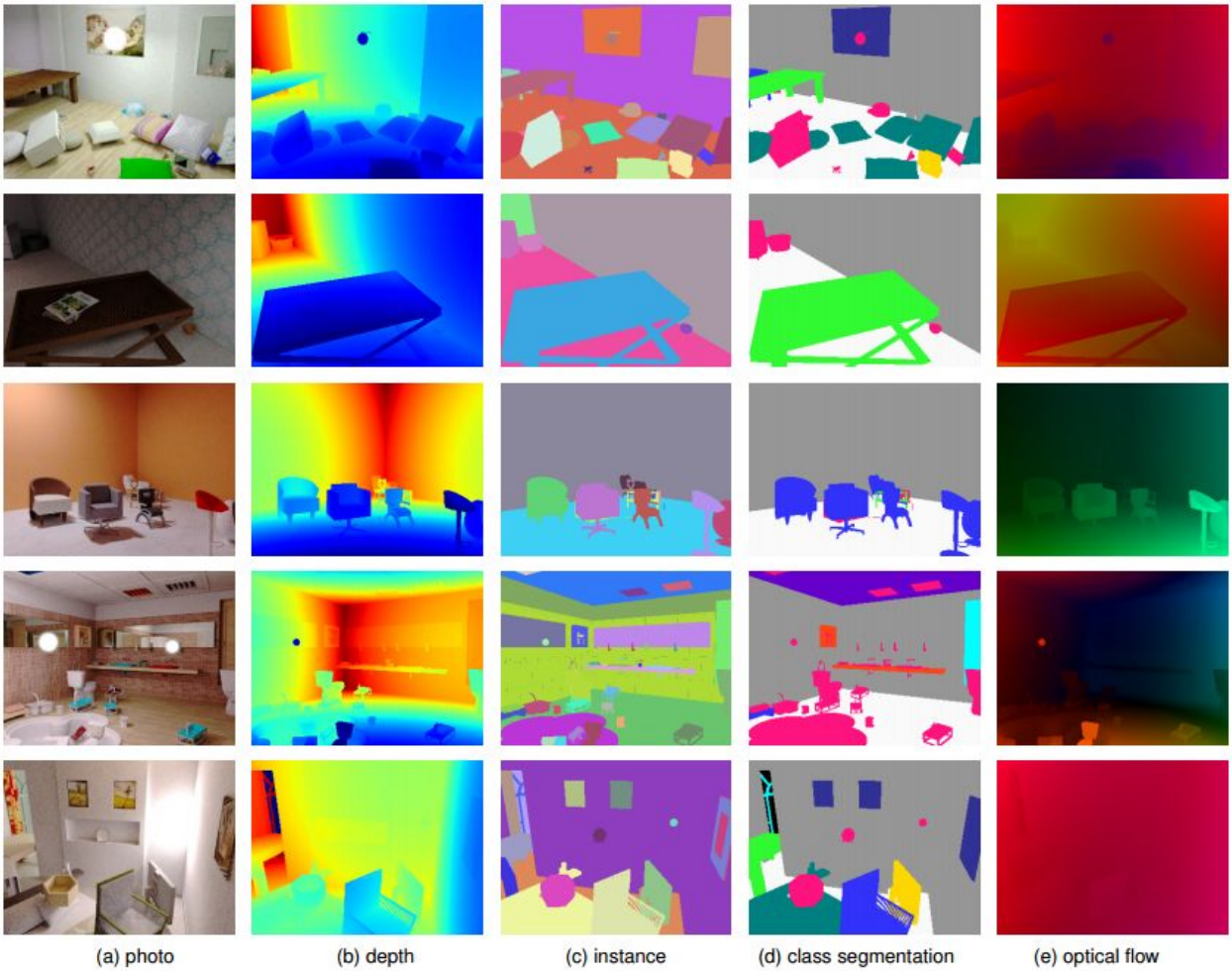
In 2016, traditional datasets such as ImageNet[169], Common Objects in Context (COCO)[170], the CIFARs[171] and MNIST[172] were joined by a host of new entries. We also noted the rise of synthetic datasets spurred on by progress in graphics. Synthetic datasets are an interesting work-around of the large data requirements for Artificial Neural Networks (ANNs). In the interest of brevity, we have selected our (subjective) most important *new* datasets for 2016:

- **Places2[173]** is a scene classification dataset, i.e. the task is to label an image with a scene class like 'Stadium', 'Park', etc. While prediction models and image understanding will undoubtedly be improved by the Places2 dataset, an interesting finding from networks that are trained on this dataset is that in the process of learning to classify scenes, the network learns to detect objects in

[163] Liao and Poggio. 2016. Bridging the Gaps Between Residual Learning, Recurrent Neural Networks and Visual Cortex. *[Online] arXiv: 1604.03640*. Available: **arXiv:1604.03640v1**
[164] Moniz and Pal. 2016. Convolutional Residual Memory Networks. *[Online] arXiv: 1606.05262*. Available: **arXiv:1606.05262v3**
[165] Hardt and Ma. 2016. Identity Matters in Deep Learning. *[Online] arXiv: 1611.04231*. Available: **arXiv:1611.04231v2**
[166] Shah et al. 2016. Deep Residual Networks with Exponential Linear Unit. *[Online] arXiv: 1604.04112*. Available: **arXiv:1604.04112v4**
[167] Shen and Zeng. 2016. Weighted Residuals for Very Deep Networks. *[Online] arXiv: 1605.08831*. Available: **arXiv:1605.08831v1**
[168] Ben Hamner. 2016. Twitter Status. *[Online] Twitter.* Available: https://twitter.com/benhamner/status/789909204832227329
[169] ImageNet. 2017. Homepage. *[Online]* Available: http://image-net.org/index [Accessed: 04/01/2017]
[170] COCO. 2017. Common Objects in Common Homepage. *[Online]* Available: http://mscoco.org/ [Accessed: 04/01/2017]
[171] CIFARs. 2017. The CIFAR-10 dataset. *[Online]* Available: https://www.cs.toronto.edu/~kriz/cifar.html [Accessed: 04/01/2017]
[172] MNIST. 2017. THE MNIST DATABASE of handwritten digits. *[Online]* Available: http://yann.lecun.com/exdb/mnist/ [Accessed: 04/01/2017]
[173] Zhou et al. 2016. Places2. *[Online]* Available: http://places2.csail.mit.edu/ [Accessed: 06/01/2017]

them without ever being explicitly taught this. For example, that bedrooms contain beds and that sinks can be in both kitchens and bathrooms. This means that the objects themselves are lower level features in the abstraction hierarchy for the classification of scenes.

Figure 18: Examples from SceneNet RGB-D



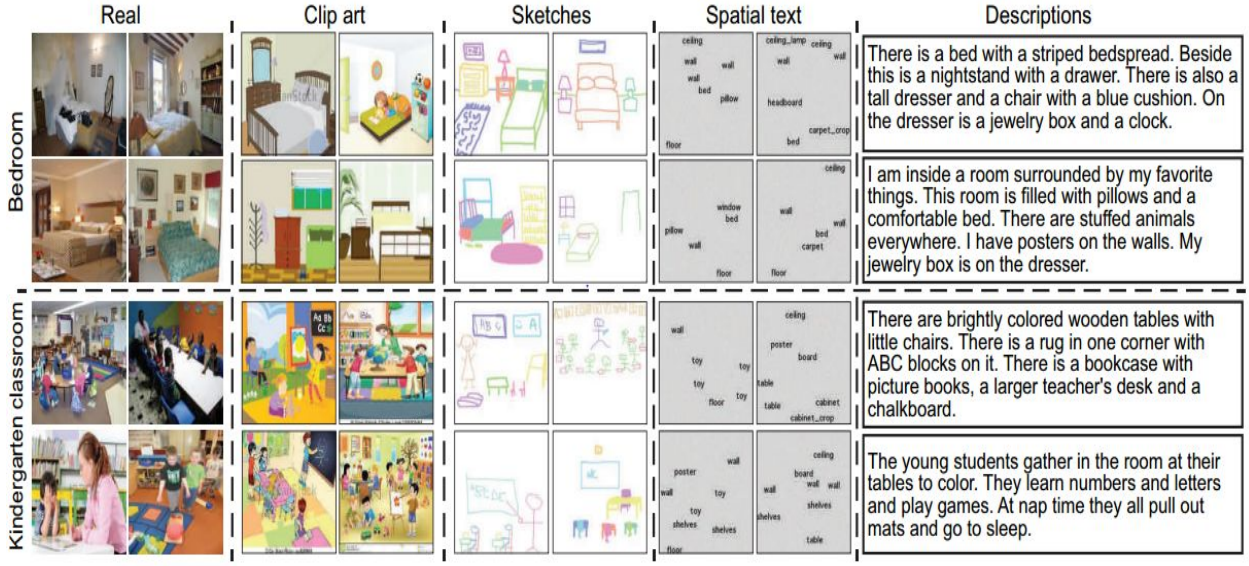
Note: Examples taken from SceneNet RGB-D, a dataset with 5M Photorealistic Images of Synthetic Indoor Trajectories with Ground Truth. The photo (a) is rendered through computer graphics with available ground truth for specific tasks from (b) to (e). Creation of synthetic datasets should aid the process of domain adaptation. Synthetic datasets are somewhat pointless if the knowledge learned from them cannot be applied to the real world. This is where domain adaptation comes in, which refers to this transfer learning process of moving knowledge from one domain to another, e.g. from synthetic to real-world environments. Domain adaptation has recently been improving very rapidly again highlighting the recent efforts in transfer learning. Columns (c) vs (d) show the difference between instance and semantic/class segmentation.
Source: McCormac et al. (2017)[174]

[174] ibid

- **SceneNet RGB-D**[175] - This synthetic dataset expands on the original SceneNet dataset and provides pixel-perfect ground truth for scene understanding problems such as semantic segmentation, instance segmentation, and object detection, and also for geometric computer vision problems such as optical flow, depth estimation, camera pose estimation, and 3D reconstruction. The dataset granularizes the chosen environment by providing pixel-perfect representations.

- **CMPlaces**[176] is a cross-modal scene dataset from MIT. The task is to recognize scenes across many different modalities beyond natural images and in the process hopefully transfer that knowledge across modalities too. Some of the modalities are: Real, Clip Art, Sketches, Spatial Text (words written which correspond to spatial locations of objects) and natural language descriptions. The paper also discusses methods for how to deal with this type of problem with cross-modal convolutional neural networks.

Figure 19: CMPlaces cross-modal scene representations



Note: Taken from the CMPlaces paper showing two examples, bedrooms and kindergarten classrooms, across different modalities. Conventional Neural Network approaches learn representations that don't transfer well across modalities and this paper attempts to generate a shared representation "agnostic of modality".
Source: Aytar et al. (2016)[177]

In CMPlaces we see explicit mention of transfer learning, domain invariant representations, domain adaptation and multi-modal learning, all of which serve to demonstrate further the current undertow of Computer Vision research. The

[175] McCormac et al. 2017. SceneNet RGB-D: 5M Photorealistic Images of Synthetic Indoor Trajectories with Ground Truth. *[Online] arXiv: 1612.05079*v3. Available: **arXiv:1612.05079v3**
[176] Aytar et al. 2016. Cross-Modal Scene Networks. *[Online] arXiv: 1610.09003*. Available: **arXiv:1610.09003v1**
[177] ibid

authors focus on trying to find "*domain/modality-independent representations*", which could correspond to the higher level abstractions where humans draw their unified representations from. For instance take 'cat' across its various modalities, humans see the word 'cat' in writing, a picture drawn in a sketchbook, a real world-image or mentioned in speech but we still have the same unified representation abstracted at a higher level above these modalities.

"Humans are able to leverage knowledge and experiences independently of the modality they perceive it in, and a similar capability in machines would enable several important applications in retrieval and recognition".

- **MS-Celeb-1M**[178] contains images of one million celebrities with ten million training images in a training set for Facial Recognition.

- **Open Images**[179] comes courtesy of Google Inc. and comprises ~9 million URLs to images complete with multiple labels, a vast improvement over typical single label images. Open images spans 6000 categories, a large improvement over the 1000 classes offered previously by ImageNet (with less focus on canines) and should prove indispensable to the Machine Learning community.

- **YouTube-8M**[180] also comes courtesy of Google with 8 million video URLs, 500,000 hours of video, 4800 classes, Avg. 1.8 Labels per video. Some examples of the labels are: 'Arts & Entertainment', 'Shopping' and 'Pets & Animals'. Video datasets are much more difficult to label and collect hence the massive value this dataset provides.

  That being said, advancements in image understanding, such as segmentation, object classification and detection have brought video understanding to the fore of research. However, prior to this dataset release there was a real lack in the variety and scale of real-world video datasets available. Furthermore, this dataset was just recently updated,[181] and this year in association with Kaggle, Google is organising a video understanding competition as part of CVPR 2017.[182]

---

[178] Guo et al. 2016. MS-Celeb-1M: A Dataset and Benchmark for Large-Scale Face Recognition. *[Online] arXiv: 1607.08221*. Available: **arXiv:1607.08221v1**
[179] Open Images. 2017. Open Images Dataset. *[Online] Github*. Available: https://github.com/openimages/dataset [Accessed: 08/01/2017]
[180] Abu-El-Haija et al. 2016. YouTube-8M: A Large-Scale Video Classification Benchmark. *[Online] arXiv: 1609.08675*. Available: **arXiv:1609.08675v1**
[181] Natsev, P. 2017. An updated YouTube-8M, a video understanding challenge, and a CVPR workshop. Oh my!. *[Online] Google Research Blog*. Available: https://research.googleblog.com/2017/02/an-updated-youtube-8m-video.html [Accessed: 26/02/2017].
[182] YouTube-8M. 2017. CVPR'17 Workshop on YouTube-8M Large-Scale Video Understanding. *[Online] Google Research*. Available: https://research.google.com/youtube8m/workshop.html [Accessed: 26/02/2017].

General information about YouTube-8M: here[183]

Ungroupable extras and interesting trends

As this piece draws to a close, we lament the limitations under which we had to construct it. Indeed, the field of Computer Vision is too expansive to cover in any real, meaningful depth, and as such many omissions were made. One such omission is, unfortunately, almost everything that didn't use Neural Networks. We know there is great work outside of NNs, and we acknowledge our own biases, but we feel that the impetus lies with these approaches currently, and our subjective selection of material for inclusion was predominantly based on the reception received from the research community at large (and the results speak for themselves).

We would also like to stress that there are hundreds of other papers in the above topics, and this amalgam of topics is not curated as a definitive, but rather hopes to encourage interested parties to read further along the entrances we provide. As such, this final section acts as a catch all for some of the other applications we loved, trends we wished to highlight and justifications we wanted to make to the reader.

Applications/use cases

- Applications for the blind from Facebook[184] and hardware from Baidu.[185]

- Emotion detection combines facial detection and semantic analysis, and is growing rapidly. There are 20+ APIs currently available.[186]

[183] Google. 2017. YouTube-8M Dataset. *[Online] research.google.com*. Available: https://research.google.com/youtube8m/ [Accessed: 04/03/2017].
[184] Wu, Pique & Wieland. 2016. Using Artificial Intelligence to Help Blind People 'See' Facebook. *[Online] Facebook Newsroom*. Available: http://newsroom.fb.com/news/2016/04/using-artificial-intelligence-to-help-blind-people-see-facebook/ [Accessed: 02/03/2017].
[185] Metz. 2016. Artificial Intelligence Finally Entered Our Everyday World. *[Online] Wired*. Available: https://www.wired.com/2016/01/2015-was-the-year-ai-finally-entered-the-everyday-world/ [Accessed: 02/03/2017].
[186] Doerrfeld. 2015. 20+ Emotion Recognition APIs That Will Leave You Impressed, and Concerned. *[Online] Nordic Apis*. Available: http://nordicapis.com/20-emotion-recognition-apis-that-will-leave-you-impressed-and-concerned/ [Accessed: 02/03/2017].

- Extracting roads from aerial imagery,[187] land use classification from aerial maps and population density maps.[188]

- Amazon Go further raised the profile of Computer Vision by demonstrating a queue-less shopping experience,[189] although there remain some functional issues at present.[190]

- There is a huge volume of work being done for Autonomous Vehicles that we largely didn't touch. However, for those wishing to delve into general market trends, there's an excellent piece by Moritz Mueller-Freitag of Twenty Billion Neurons about the German auto industry and the impact of autonomous vehicles.[191]

- Other interesting areas: Image Retrieval/Search,[192] Gesture Recognition, Inpainting and Facial Reconstruction.

- There is considerable work around Digital Imaging and Communications in Medicine (DICOM) and other medical applications, especially related to imaging. For instance, there have been (and still are) numerous Kaggle detection competitions (lung cancer, cervical cancer), some with large monetary incentives, in which algorithms attempt to outperform specialists at the classification/detection tasks in question.

  However, while work continues on improving the error rates of these algorithms their value as a tool for medical practitioners appears increasingly evident. This is particularly striking when we consider the performance improvements in breast cancer detection achieved by combining AI systems[193] with medical specialists.[194]

---

[187] Johnson, A. 2016. Trailbehind/DeepOSM - Train a deep learning net with OpenStreetMap features and satellite imagery. *[Online] Github.com*. Available: https://github.com/trailbehind/DeepOSM [Accessed: 29/03/2017].
[188] Gros and Tiecke. 2016. Connecting the world with better maps. *[Online] Facebook Code*. Available: https://code.facebook.com/posts/1676452492623525/connecting-the-world-with-better-maps/ [Accessed: 02/03/2017].
[189] Amazon. 2017. Frequently Asked Questions - Amazon Go. *[Website] Amazon.com*. Available: https://www.amazon.com/b?node=16008589011 [Accessed: 29/03/2017].
[190] Reisinger, D. 2017. Amazon's Cashier-Free Store Might Be Easy to Break. *[Online] Fortune Tech*. Available: http://fortune.com/2017/03/28/amazon-go-cashier-free-store/ [Accessed: 29/03/2017].
[191] Mueller-Freitag, M. 2017. Germany asleep at the wheel? *[Blog] Twenty Billion Neurons - Medium.com*. Available: https://medium.com/twentybn/germany-asleep-at-the-wheel-d800445d6da2
[192] Gordo et al. 2016. Deep Image Retrieval: Learning global representations for image search. *[Online] arXiv: 1604.01325*. Available: **arXiv:1604.01325v2**
[193] Wang et al. 2016. Deep Learning for Identifying Metastatic Breast Cancer. *[Online] arXiv: 1606.05718*. Available: **arXiv:1606.05718v1**
[194] Rosenfeld, J. 2016. AI Achieves Near-Human Detection of Breast Cancer. *[Online] Mentalfloss.com*. Available: http://mentalfloss.com/article/82415/ai-achieves-near-human-detection-breast-cancer [Accessed: 27/03/2017].

In this instance, robot-human symbiosis produces accuracy far greater than the sum of its parts at 99.5%.

This is just one example of the torrent of medical applications currently being pursued by the deep learning/machine learning communities. Some cynical members of our team jokingly make light of these attempts as a means to ingratiate society to the idea of AI research as a ubiquitous, benevolent force. But as long as the technology helps the healthcare industry, and it is introduced in a safe and considered manner, we wholeheartedly welcome such advances.

Hardware/markets

- Growing markets for Robotic Vision/Machine Vision (separate fields) and potential target markets for IoT. A personal favourite of ours is the use of Deep Learning, a Raspberry Pi and TensorFlow by a farmer's son to sort cucumbers in Japan based on unique producer heuristics for quality, e.g. shape, size and colour.[195] This produced massive decreases in human-time spent by his mother sorting cucumbers.

- The trend of shrinking compute requirements and migrating to mobile is evident, but it's also complemented by steep hardware acceleration. Soon we'll see pocket sized CNNs and Vision Processing Units (VPUs) everywhere. For instance, the Movidius Myriad2 is used in Google's Project Tango and drones.[196] The Movidius Fathom stick,[197] which also uses the Myriad2's technology, allows users to add SOTA Computer Vision performance to consumer devices. The Fathom stick, which has the physical properties of a USB stick, brings the power of a Neural Network to almost any device: Brains on a stick.

- Sensors and systems that use something other than visible light. Examples include radar, thermographic cameras, hyperspectral imaging, sonar, magnetic resonance imaging, etc.

- Reduction in cost of LIDAR, which use light and radar to measure distances, and offer many advantages over normal RGB cameras. There are many LIDAR devices for currently less than $500.

---

[195] Sato, K. 2016. How a Japanese cucumber farmer is using deep learning and TensorFlow. *[Blog] Google Cloud Platform*. Available: https://cloud.google.com/blog/big-data/2016/08/how-a-japanese-cucumber-farmer-is-using-deep-learning-and-tensorflow

[196] Banerjee, P. 2016. The Rise of VPUs: Giving eyes to machines. *[Online]* *www.digit.in*. Available: http://www.digit.in/general/the-rise-of-vpus-giving-eyes-to-machines-29561.html [Accessed: 22/03/2017.

[197] Movidius. 2017. Embedded Neural Network Compute Framework: Fathom. *[Online] Movidius.com*. Available: https://www.movidius.com/solutions/machine-vision-algorithms/machine-learning [Accessed: 03/03/2017].

- Hololens and the near-countless other Augmented Reality headsets[198] entering the market.

- **Project Tango by Google[199]** represents the next big commercialisation of SLAM. Tango is an augmented reality computing platform, comprising both novel software and hardware. Tango allows the detection of mobile device position, relative to the world, without the use of GPS or other external information while simultaneously mapping the area around the device in 3D.

  Corporate partners Lenovo brought affordable Tango enabled phones to market in 2016, allowing hundreds of developers to begin creating applications for the platform. Tango employs the following software technologies: Motion Tracking, Area Learning, and Depth Perception.

Omissions based on forthcoming publications

There is also considerable, and increasing overlap between Computer Vision techniques and other domains in Machine Learning and Artificial Intelligence. These other domains and hybrid use cases are the subject of The M Tank's forthcoming publications and, as with the whole of this piece, we partitioned content based on our own heuristics.

For instance, we decided to place the two integral Computer Vision tasks, Image Captioning and Visual Question Answering, in our forthcoming NLP piece along with Visual Speech Recognition because of the combination of CV and NLP involved. Whereas the application of Generative Models to images we place in our work on Generative Models. Examples included in these future works are:

- **Lip Reading**: In 2016 we saw huge lip reading advancements in programs such as LipNet[200], which combine Computer Vision and NLP into Visual Speech Recognition.

- **Generative models** applied to images will feature as part of our depiction of the violent* battle between the Autoregressive Models (PixelRNN, PixelCNN, ByteNet, VPN, WaveNet), Generative Adversarial Networks (GANs), Variational

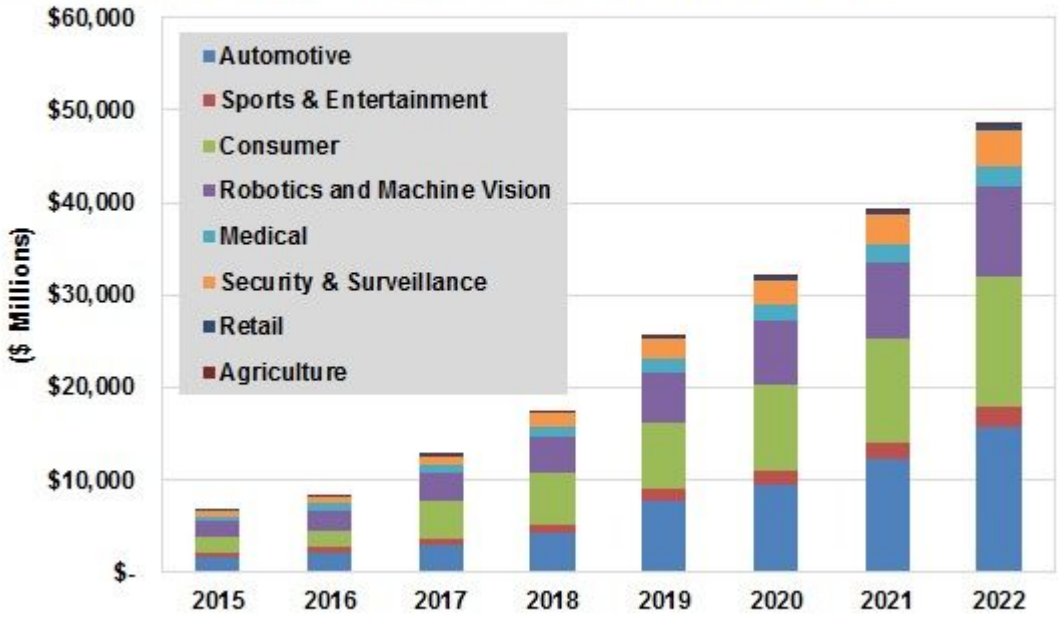
[198] Dzyre, N. 2016. 10 Forthcoming Augmented Reality & Smart Glasses You Can Buy. *[Blog] Hongkiat*. Available: http://www.hongkiat.com/blog/augmented-reality-smart-glasses/ [Accessed: 03/03/2017].
[199] Google. 2017. Tango. *[Website] get.google.com*. Available: https://get.google.com/tango/ [Accessed: 23/03/2017].
[200] Assael et al. 2016. LipNet: End-to-End Sentence-level Lipreading. *[Online] arXiv: 1611.01599*. Available: arXiv:1611.01599v2

Autoencoders and, as you should expect by this stage, all of their variants, combinations and hybrids.

*Disclaimer: The team wishes to mention that they do not condone Network on Network (NoN) violence in any form and are sympathisers to the movement towards Generative Unadversarial Networks (GUNs).[201]

In the final section, we'll offer some concluding remarks and a recapitulation of some of the trends we identified. We would hope that we were comprehensive enough to show a bird's-eye view of where the Computer Vision field is loosely situated and where it is headed in the near-term. We also would like to draw particular attention to the fact that our work does not cover January-April 2017. The blistering pace of research output means that much of this work could be outdated already; we encourage readers to go and find out whether it is for themselves. But this rapid pace of growth also brings with it lucrative opportunities as the Computer Vision hardware and software markets are expected to reach $48.6 Billion by 2022.

Figure 19: Computer Vision Revenue by Application Market[202]



Note: Estimation of Computer Vision revenue by application market spanning the period from 2015-2022. The largest growth is forecasted to come from applications within the automotive, consumer, robotics and machine vision sectors. **Source**: Tractica (2016)[203]

[201] Albanie et al. 2017. Stopping GAN Violence: Generative Unadversarial Networks. *[Online] arXiv: 1703.02528*. Available: **arXiv:1703.02528v1**

[202] Tractica. 2016. Computer Vision Hardware and Software Market to Reach $48.6 Billion by 2022. *[Website]* *www.tractica.com*. Available: https://www.tractica.com/newsroom/press-releases/computer-vision-hardware-and-software-market-to-reach-48-6-billion-by-2022/ [Accessed: 12/03/2017].

[203] ibid

Conclusion

In conclusion we'd like to highlight some of the trends and recurring themes that cropped up repeatedly throughout our research review process. First and foremost, we'd like to draw attention to the Machine Learning research community's voracious pursuit of optimisation. This is most notable in the year on year changes in accuracy rates, but especially in the intra-year changes in accuracy. We'd like to underscore this point and return to it in a moment.

Error rates are not the only fanatically optimised parameter, with researchers working on improving speed, efficiency and even the algorithm's ability to generalise to other tasks and problems in completely new ways. We are acutely aware of the research coming to the fore with approaches like one-shot learning, generative modelling, transfer learning and, as of recently, evolutionary learning, and we feel that these research principles are gradually exerting greater influence on the approaches of the best performing work.

While this last point is unequivocally meant in commendation for, rather than denigration of, this trend, one can't help but to cast their mind toward the (very) distant spectre of Artificial General Intelligence, whether merited a thought or not. Far from being alarmist, we just wish to highlight to both experts and laypersons that this concern arises from here, from the startling progress that's already evident in Computer Vision and other AI subfields. Properly articulated concerns from the public can only come through education about these advancements and their impacts in general. This may then in turn quell the power of media sentiment and misinformation in AI.

We chose to focus on a one year timeline for two reasons. The first relates to the sheer volume of work being produced. Even for people who follow the field very closely, it is becoming increasingly difficult to remain abreast of research as the number of publications grow exponentially. The second brings us back to our point on intra-year changes.

In taking a single year snapshot of progress, the reader can begin to comprehend the pace of research at present. We see improvement after improvement in such short time spans, but why? Researchers have cultivated a global community where building on previous approaches (architectures, meta-architectures, techniques, ideas, tips, wacky hacks, results, etc.), and infrastructures (libraries like Keras, TensorFlow and PyTorch, GPUs, etc.), is not only encouraged but also celebrated. A predominantly open source community with few parallels, which is continuously attracting new researchers and having its techniques reappropriated by fields like economics, physics and countless others.

It’s important to understand for those who have yet to notice, that among the already frantic chorus of divergent voices proclaiming divine insight into the true nature of this technology, there is at least agreement; agreement that this technology will alter the world in new and exciting ways. However, much disagreement still comes over the timeline on which these alterations will unravel.

Until such a time as we can accurately model the progress of these developments we will continue to provide information to the best of our abilities. With this resource we hoped to cater to the spectrum of AI experience, from researchers playing catch-up to anyone who simply wishes to obtain a grounding in Computer Vision and Artificial Intelligence. With this our project hopes to have added some value to the open source revolution that quietly hums beneath the technology of a lifetime.

With thanks,

The M Tank