Paper Name	Paper subject
Learning to Match Aerial Images with Deep Attentive Architectures	defines an ultra-wide baseline image matching problem (matching differnt parts of an image with different parts of another image when both of them are captured from a same object with different viewpoints) as a classification task and uses a deep-learening based approach to tackle the problem. (fine-tunning pre-trained AlexNet model on the data, siamese architecture for feature extraction and a binary classifier, Spatial Transformer modules)
Efficient Indexing of Billion-Scale datasets of deep descriptors	introduces a new dataset of one billion descriptors based on DNNs and reveal the relative inefficiency of IMI-based indexing for such descriptors compared to SIFT data, two new indexing structures are proposed, Non-Orthogonal Inverted Multi-Index (NO-IMI) and Generalized Non-Orthogonal Inverted Multi-Index (GNO-IMI).
Analyzing Classifiers: Fisher Vectors and Deep Neural Networks	problem: complex nonlinear image classification models, (1) extension of Layer-wise Relevance Propagation (LRP) also for Fisher vector (FV) classifiers and then use it as analysis tool to (2) quantify the importance of context for classification (also using heatmapping technique that determines what pixels in the image are used by a classifier to support its decision) , (3) qualitatively compare DNNs against FV classifiers in terms of important image regions and (4) detect potential flaws and biases in data.
Towards Open Set Deep Networks	deep network are easily fooled with images humans do not consider meaningful. The closed set nature of deep networks forces them to choose from one of the known classes leading to such artifacts, while Recognition in the real world is open set, i.e. the recognition system should reject unknown/unseen classes at test time. They present a methodology to adapt deep networks for open set recognition, by introducing a new model layer, OpenMax, which estimates the probability of an input being from an unknown class. OpenMax allows rejection of "fooling" and unrelated open set images presented to the system.
Weakly Supervised Deep Detection Networks	problem: Weakly supervised learning of object detection (WSDDN). They address this problem by exploiting the power of deep convolutional neural networks pre-trained on large-scale image-level classification tasks by proposing an architecture that modifies one such network to operate at the level of image regions, performing simultaneously region selection and classification. The model, which is a simple and elegant end-to-end architecture, outperforms standard data augmentation and fine-tuning techniques for the task of image-level classification. it works on top of a SPP layer. (better than traditional fine-tuning techniques to improve the performance of a pre-trained CNN on the problem of image classification)
iLab-20M: A large-scale controlled object dataset to investigate deep learning	They introduce a large-scale synthetic dataset(iLab-20M), which is freely and publicly available, and use it to answer several fundamental questions regarding selectivity and invariance properties of convolutional neural networks. They study: 1) invariance and selectivity of different CNN layers, 2) knowledge transfer from one object category to another, 3) systematic or random sampling of images to build a train set, 4) domain adaptation from synthetic to natural scenes, and 5) order of knowledge delivery to CNNs. **Concolusions: i) the representation learned in pool5 layer is selective to parameters while fc7 layer is not, ii) the knowledge obtained from some parameters is easier to be transferred to unseen object categories, iii) random sampling strategy leads to better generalization since more instance level variations can be captured, iv) simple cross application of one dataset to another results in above chance accuracy but does not improve performance, and v) it would be advantageous to feed the network with data that has been sorted according to complexities of different dimensions. This can lead to layer-wise training of CNNs for learning different invariances in different layers.**
DCAN: Deep Contour-Aware Networks for Accurate Gland Segmentation	A unified multi-task learning framework to solve this challenging problem of accurate segmentation of glands from histology images where multi-level contextual features from the hierarchical architecture are explored with auxiliary supervision for accurate gland segmentation. When incorporated with multi-task regularization during the training, the discriminative capability of intermediate features can be further improved. Also depict clear contours simultaneously for separating clustered objects, which further boosts the gland segmentation performance.
Deep SimNets	Generalization of CNNs by two newly defined operators: 1) A similarity function and 2) A log-mean-exp function. The SimNet contains a higher abstraction level compared to a traditional ConvNet. (significant gain in accuracy over ConvNets when computational resources at run-time are limited). A higher abstraction level for the basic network building blocks carries with it the advantage of obtaining higher accuracies with small networks
Fine-grained Categorization and Dataset Bootstrapping using Deep Metric Learning with Humans in the Loop	The proposed deep metric learning scheme (an iterative framework for fine-grained visual categorization) handles 1) lack of training data, 2) large number of fine-grained categories, and 3) high intra-class vs. low inter-class variance in fine-grained visual categorization. (Although we adopt an effective and efficient online triplet sampling strategy, the training process could still be slow, which is a limitation of our method)

DeepCAMP: Deep Convolutional Action & Attribute Mid-Level Patterns	A novel CNN that mines mid-level image patches for fine-grained classification for recognition of human actions and the determination of human attributes (ACTION CLASSIFICATION) that learns discriminative patch groups. On the one hand we pay attention to contextual information in an original fashion. On the other hand, we let an iteration of feature learning and patch clustering purify the set of dedicated patches that we use. (datasets: PASCAL VOC 2012 Action and Stanford 40 Actions & the Berkeley Attributes of People dataset)
Longitudinal Face Modeling via Temporal Deep Restricted Boltzmann Machines	A deep model approach for modeling the face aging process that has non-linear variations present in different stages of face development. The Temporal Deep Restricted Boltzmann Machines based age progression model together with the prototype faces are then constructed to learn the aging transformation between faces in the sequence. (datasets: FG-NET, Cross-Age Celebrity Dataset (CACD) and MORPH, and our collected large-scale aging database named AginG Faces in the Wild (AGFW))
WELDON: Weakly Supervised Learning of Deep Convolutional Neural Networks	Subject: Image Classification, automatically selecting relevant image regions from weak annotations, e.g. global image labels, using Multiple Instance Learning paradigm, i.e. negative evidence scoring and top instance selection. The deep CNN is trained to optimize Average Precision, and fine-tuned on the target dataset with efficient computations due to convolutional feature sharing.
Occlusion Boundary Detection via Deep Exploration of Context	A novel approach based on convolutional neural networks (CNNs) and conditional random fields (CRFs) to improve occlusion boundary detection via enhanced exploration of contextual information (e.g., local structural boundary patterns, observations from surrounding regions, and temporal context).
Learning Local Image Descriptors with Deep Siamese and Triplet Convolutional Networks by Minimizing Global Loss Functions	Subject: Image Classification, (patch matching based on learning using triplet and siamese networks trained with a combination of triplet loss and global loss applied to mini-batches), first they propose the use of triplet networks for the problem of local image descriptor learning. Furthermore, they also propose the use of a global loss that minimises the overall classification error in the training set, which can improve the generalisation capability of the model. They also demonstrate that a combination of the triplet and global losses produces the best embedding in the field, using this triplet network. Finally, we also show that the use of the central-surround siamese network trained with the global loss produces the best result of the field on the UBC dataset.
Learning to Co-Generate Object Proposals with a Deep Structured Network	Subject: Object Detection & object proposal algorithms, introduces a deep structured network (consists of a fully-connected Conditional Random Field (CRF) built on top of a set of deep Convolutional Neural Networks) that jointly predicts the objectness scores and the bounding box locations of multiple object candidates. To train our deep structured network, we develop an end-to-end learning algorithm.
Deep Residual Learning for Image Recognition (from ResNet Group)	The technical explanation of the ResNet and its learning notes. (read it carefully)
Exemplar-Driven Top-Down Saliency Detection via Deep Association	They propose in this paper a locate-by-exemplar strategy and a two-stage deep model to learn the intra-class association between the exemplars and query objects. The proposed method outperforms different baselines and the category-specific models. Furthermore, the learned model is a universal model and offers great generalization to unseen objects. The network learnt with more exemplars achieves more robust association quantitatively and qualitatively.
Deep Compositional Captioning: Describing Novel Object Categories without Paired Training Data	Subject: Image Captioning, generating descriptions of novel objects which are not present in paired image-sentence datasets by leveraging large object recognition datasets and external text corpora and by transferring knowledge between semantically similar concepts. The Deep Compositional Captioner (DCC) has distinct advantages over existing image and video captioning approaches for generating descriptions of new objects in context.
Learning Transferrable Knowledge for Semantic Segmentation with Deep Convolutional Neural Network	Subject: Semantic Segmentation, a novel weakly-supervised semantic segmentation algorithm based on Deep Convolutional Neural Network that exploits auxiliary segmentation annotations available for different categories to guide segmentations on images with only image-level class labels, To make segmentation knowledge transferrable across categories, they design a decoupled encoder-decoder architecture with attention model.

Learning Deep Representation for Imbalanced Classification	imbalance classification: most data belong to a few majority classes, while the minority classes only contain a scarce amount of instances. typical solutions are class re-sampling or cost-sensitive training. More discriminative deep representation can be learned by enforcing a deep network to maintain both inter-cluster and inter-class margins proposing quintuplet sampling with triple-header loss .
FireCaffe: near-linear acceleration of deep neural network training on compute clusters	scales deep neural network training across a cluster of GPUs. The key consideration here is to reduce communication overhead wherever possible. Pillars: 1) network hardware that achieves high band-width between GPU servers, 2) a number of communication algorithms using reduction trees and 3) increase the batch size to reduce the total quantity of communication during DNN training. Finally and they identify hyperparameters that allow them to reproduce the small-batch accuracy while training with large batch sizes.
A Hierarchical Deep Temporal Model for Group Activity Recognition	A deep temporal model based on LSTM (Short Long-term memory) to capture dynamics of the individual people representing the activity. (Datasets: the Collective Activity Dataset and a new volleyball dataset).
Structural-RNN: Deep Learning on Spatio-Temporal Graphs	An approach for combining the power of high-level spatio-temporal graphs and sequence learning success of Recurrent Neural Networks (RNNs). They develop a scalable method for casting an arbitrary spatio-temporal graph as a rich RNN mixture that is feedforward, fully differentiable, and jointly trainable. (it's an interesting and foundamental work, read it)
Multi-view Deep Network for Cross-view Classification	A multi-view deep network (MvDN) for Cross-view recognition that intends to classify samples between different views consists of two sub-networks, view-specific sub-network attempting to remove view-specific variations and the following common sub-network attempting to obtain common representation shared by all views.
Accurate Image Super-Resolution Using Very Deep Convolutional Networks	Single image super-resolution (SISR) is the process of generating a high-resolution (HR) image given a low-resolution (LR) image, using a very deep convolutional network inspired by VGG-net, By cascading small filters many times in a deep network structure, contextual information over large image regions is exploited in an efficient way by learning residuals only and use extremely high learning rates enabled by adjustable gradient clipping to optimize a very deep network fast and ensure the training stability. This approach is readily applicable to other image restoration problems such as denoising and compression artifact removal.
Learning to Select Pre-trained Deep Representations with Bayesian Evidence Framework	Bayesian (LS-SVM) evidence framework is proposed to facilitate transfer learning from pre-trained deep convolutional neural networks (CNNs) by selecting the best performing CNN out of multiple candidates to transfer deep CNN models pre-trained on specific image classification tasks to another tasks. It also provides a good solution to identify the best ensemble of heterogeneous CNNs through a greedy algorithm.
Deep Hand: How to Train a CNN on 1 Million Hand Images When Your Data Is Continuous and Weakly Labelled	Subject: video recognition task, A new approach to learning a frame-based classifier on weakly labelled sequence data by embedding a CNN within an iterative EM algorithm. (no previous work has explored expectation maximization without Gaussian mixture models to exploit weak sequence labels for sign language recognition)
Saliency Unified: A Deep Architecture for simultaneous Eye Fixation Prediction and Salient Object Segmentation	Subject: saliency detection and eye fixation prediction, They propose a deep convolutional neural network (CNN) capable of predicting eye fixations and segmenting salient objects in a unified framework simultaneously. (predicting eye fixation saliency maps requires the model to estimate the saliency score for every pixel in a given test image.) (low-level and high-level semantics are necessary for salient object segmentation)
Deep Saliency with Encoded Low level Distance Map and High Level Features	Subject: saliency detection, hand-crafted features can provide complementary information to enhance performance of saliency detection that utilizes only high level features. Proposed method utilizes both high level and low level features for saliency detection under a unified deep learning framework. The high level features are extracted using the VGG-net, and the low level features are compared with other parts of an image to form a low level distance map.

Comparative Deep Learning of Hybrid Representations for Image Recommendations	Subject: image recommendations, Hybrid Representations are effective representations of not only images but also preferences and intents of users over images. This work proposes a dual-net deep network, in which the two sub-networks map input images and preferences of users into a same latent semantic space, and then the distances between images and users in the latent space are calculated to make decisions. They further propose a comparative deep learning (CDL) method to train the deep network that embraces much more training data than naive deep learning, and thus achieves superior performance than the latter, with no cost of increasing network complexity.
Deep Contrast Learning for Salient Object Detection	For Salient object detection existing CNN-based methods operate at the patch level instead of the pixel level, therefore aliency maps are typically blurry, especially near the boundary of salient objects. Furthermore, image patches are treated as independent samples even when they are overlapping, giving rise to significant redundancy in computation and storage. In this paper, we propose an end-to-end deep contrast network to overcome the aforementioned limitations that has a pixel-level fully convolutional stream and a segment-wise spatial pooling stream. A fully connected CRF model can be optionally incorporated to further improve spatial coherence and contour localization in the fused result from these two streams.
Deep Structured Scene Parsing by Learning with Image Descriptions	Subject: scene understanding (How to parse the scene image into a structured configuration that finely accords with human perception), They We propose a deep architecture consisting of two networks: i) a convolutional neural network (CNN) extracting the image representation for pixelwise object labeling and ii) a recursive neural network (RNN) discovering the hierarchical object structure and the inter-object relations. Rather than relying on elaborative user annotations (e.g., manually labeling semantic maps and relations), they train their deep model in a weakly-supervised manner by leveraging the descriptive sentences of the training images.
Efficient Piecewise Training of Deep Structured Models for Semantic Segmentation	Subject: semantic Segmentation, how to improve semantic segmentation through the use of contextual information, (semantic segmentation datasets: NYUDv2, PASCAL VOC 2012, PASCAL-Context, and SIFT-flow)
Learning Compact Binary Descriptors with Unsupervised Deep Neural Networks	A new unsupervised deep learning approach called DeepBit to learn compact binary descriptor for efficient visual object matching. This approach does not require labeled data during learning.
Visualizing and Understanding Deep Texture Representations	This work conducts a systematic evaluation of recent CNN-based texture descriptors for recognition and attempts to understand the nature of invariances captured by these representations. They propose a technique to visualize pre-images, providing a means for understanding categorical properties that are captured by these representations. The key challenge is that the approach is computationally expensive.
Deep Relative Distance Learning: Tell the Difference Between Similar Vehicles	Subject: vehicle search from a large-scale image or video database (vehicle re-identification), This work proposed a Deep Relative Distance Learning (DRDL) method which exploits a two-branch deep convolutional network to project raw vehicle images into an Euclidean space where distance can be directly used to measure the similarity of arbitrary two vehicles. They also present a carefully-organized large-scale image database "VehicleID". Another recently-released vehicle model classification dataset is "CompCars".
Deep Supervised Hashing for Fast Image Retrieval	Subject: Deep Supervised Hashing (DSH), a new hashing method to learn compact (similarity-preserving) binary codes for highly efficient image retrieval on large-scale datasets. They devise a CNN architecture that takes pairs of images (similar/dissimilar) as training inputs and encourages the output of each image to approximate discrete values (e.g. +1/-1). Also, a loss function is elaborately designed to maximize the discriminability of the output space by encoding the supervised information from the input image pairs, and simultaneously imposing regularization on the real-valued outputs to approximate the desired discrete values. (Datasets: CIFAR-10 & NUS-WIDE)
DHSNet: Deep Hierarchical Saliency Network for Salient Object Detection	Subject: saliency detection, They propose an end-to-end deep hierarchical saliency network (DHSNet) based on convolutional neural networks for detecting salient objects in which HSNet first makes a coarse global prediction by automatically learning various global structured saliency cues, including global contrast, objectness, compactness, and their optimal combination. Then a hierarchical recurrent convolutional neural network (HRCNN) is adopted to further hierarchically and progressively refine the details of saliency maps step by step via integrating local context information. The proposed architecture achieves a real-time speed of 23 FPS on modern GPUs.
Learning Relaxed Deep Supervision (RDS) for Better Edge Detection	Subject: Edge Detection, 1. Capturing relaxed labels from simple detectors 2. Merge them with the general ground truth to generate the Relaxed Deep Supervision (RDS). 3. Employ the RDS to supervise the edge network following a coarse-to-fine paradigm. (Datasets: BSDS500 & NYUD)

VLAD 3 : Encoding Dynamics of Deep Features for Action Recognition	Subject: Video Representation & Action Recognition (discrimination of complex activities that share sub-actions, and when dealing with untrimmed videos), they propose a representation that accounts for different levels of video dynamics. It captures short-term dynamics with deep convolutional neural network features, relying on linear dynamic systems (LDS) to model medium-range dynamics. (Datasets: UCF101 & THUMOS15)
Efficient Deep Learning for Stereo Matching	Subject: Stereo Matching, convolutional neural networks perform extremely well for stereo estimation, current architectures rely on siamese networks which exploit concatenation followed by further processing layers, requiring a minute of GPU computation per image pair. This paper proposes a matching network which is able to produce very accurate results in less than a second of GPU computation. They train their network by treating the problem as multi-class classification, where the classes are all possible disparities. This allows us to get calibrated scores, which result in much better matching performance.
Composition-preserving Deep Photo Aesthetics Assessment	Subject: aesthetics assessment, neural network only takes the fixed-size input, therefore, input images need to be transformed via cropping, scaling, or padding, which often damages image composition, This paper presents a composition-preserving deep ConvNet method that directly learns aesthetics features from the original input images without any image transformations. (Dataset: large-scale aesthetics assessment benchmark (AVA)) This scene-aware MNA-CNN has three enabling features. First, it uses an adaptive spatial pooling layer upon regular convolutional and pooling layers. This adaptive spatial pooling layer has a fixed-size output while having a variable receptive field size to handle images with different sizes and aspect ratios. Second, it uses multiple sub-networks to capture aesthetics features at multiple scales. Finally, it uses a scene-aware aggregation layer to combine these sub-networks into a powerful one.
Deep Exemplar 2D-3D Detection by Adapting from Real to Rendered Views	Subject: An end-to-end CNN for 2D-3D exemplar detection (2D-3D matching) that matches an object in a 2D image with a rendered 3D model to reveal some 3D information of the 2D object in the image such as hidden surfaces and object pose. 1. instance detection 2. object category detection (Datasets: IKEA & Pascal VOC)
DeepFool: a simple and accurate method to fool deep neural networks	Neural Network architectures have been shown to be unstable to small, well sought, perturbations of the images. Authors propose the DeepFool algorithm to efficiently compute perturbations that fool deep networks, and thus reliably quantify the robustness of these classifiers. It is based on an iterative linearization of the classifier to generate minimal perturbations that are sufficient to change classification labels that provides an efficient and accurate way to evaluate the robustness of classifiers and to enhance their performance by proper fine-tuning. The proposed approach can therefore be used as a reliable tool to accurately estimate the minimal perturbation vectors, and build more robust classifiers.
Blockout: Dynamic Model Selection for Hierarchical Deep Networks	Blockout is a novel generalization of stochastic regularization with parameters that can be learned during training, essentially allowing for automatic model selection within a class of hierarchical network structures. Hierarchical deep networks learn separate features for subsets of related categories. They propose Blockout, a method for regularization and model selection that simultaneously learns both the model architecture and parameters, demonstrating improved classification accuracy, better regularization performance, faster training, and the clear emergence of hierarchical network structures. Blockout allows for end-to-end learning of more complex hierarchical architectures.
Deep Decision Network for Multi-Class Image Classification	Subject: Multi-Class Image Classification & Network Building, This work presents a novel Deep Decision Network (DDN) that provides an alternative approach towards building an efficient deep learning network. During the learning phase, starting from the root network node, DDN automatically builds a network that splits the data into disjoint clusters of classes which would be handled by the subsequent expert networks. This results in a tree-like structured network driven by the data identifying the group of classes that are hard to classify and require more attention when compared to others. DDN also has the ability to make early decisions thus making it suitable for time-sensitive applications. (Datasets: CIFAR-10 & CIFAR-100)
Factors in Finetuning Deep Model for Object Detection with Long-tail Distribution	Subject: Finetuning on object detection task, Classes with more samples have higher impact on the feature learning, so it is better to make the sample number more uniform across classes. They proposed to cluster objects into visually similar class groups and learn deep representations for these groups separately using a cascaded hierarchical feature learning scheme in which the knowledge from the group with large number of classes is transferred for learning features in its sub-groups to improve the effectiveness of the learned features. (Code is available)
Shallow and Deep Convolutional Networks for Saliency Prediction	Subject: saliency prediction, A completely (end-to-end) data-driven approach is proposed to addresses the saliency detection by training a convolutional neural network (convnet). A loss function that measures the Euclidean distance of the predicted saliency map with the provided ground truth. Two designs are proposed: a shallow convnet trained from scratch, and a another deeper solution whose first three layers are adapted from another network trained for classification. (Saliency prediction datasets: SALICON & MIT300)
Hedged Deep Tracking	Subject: Visual tracking, In this paper, Authors propose a novel CNN based tracking framework, which takes full advantage of features from different CNN layers and uses an adaptive Hedge method to hedge several CNN based trackers into a single stronger one.

Hierarchically Gated Deep Networks for Semantic Segmentation	Subject: Semantic Segmentation, While image structures usually have various scales, it is difficult to use a single scale to model the spatial contexts for all individual pixels. Multi-scale Convolutional Neural Networks (CNNs) and their variants have made striking success for modeling the global scene structure for an image. However, they are limited in labeling fine-grained local structures like pixels and patches, since spatial contexts might be blindly mixed up without appropriately customizing their scales. Authors develop a novel paradigm of multi-scale deep network (Hierarchically Gated Deep Network (HGDN)) to model spatial contexts surrounding different pixels at various scales by customizing a suitable scale for each pixel.
MDL-CW: A Multimodal Deep Learning Framework with Cross Weights	Subject: multimodal representation learning, All the previous deep models contain separate modality-specific networks and find a shared representation on top of those networks. Therefore, they only consider high level interactions between modalities to find a joint representation for them. Authors proposed a multimodal deep learning framework (MDL-CW) that exploits the cross weights between representation of modalities, and try to gradually learn interactions of the modalities in a deep network manner that provides more intra-modality information and introduce a multi-stage pre-training method that is based on the properties of multi-modal data. Then they try to reconstruct the representation of each modality at a given level, with representation of other modalities in the previous layer. (Datasets: PASCAL-sentence & SUN-Attributes)
Learning Deep Representations of Fine-Grained Visual Descriptions	zero-shot visual recognition formulate learning as a joint embedding problem of images and side information. In these formulations the current best complement to visual features are attributes with these limitations: 1. finer-grained recognition requires commensurately more attributes 2. attributes do not provide a natural language interface The proposed models train end-to-end to align with the fine-grained and category-specific content of images to overcome these limitations. (Dataset: UCSD Birds 200-2011 dataset)
Deep Reflectance Maps	Subject: Reflectance Map Estimation, A convolutional neural architecture to estimate reflectance maps of specular materials in natural lighting conditions as an end-to-end learning formulation that directly predicts a reflectance map from the image itself using an indirect scheme that first predicts surface orientation and afterwards predicts the reflectance map by a learning-based sparse data interpolation. New challenge definition: Specular MAterials on SHapes with complex IllumiNation (SMASHINg) "Project has a website"
Refining Architectures of Deep Convolutional Neural Networks	Is the selected CNN optimal for the dataset in terms of accuracy and model size? This work presents a novel strategy that alters the architecture of a given CNN for a specified dataset, to potentially enhance the original accuracy while possibly reducing the model size using stretching (increases the number of hidden units (nodes) in a given CNN layer) and symmetrical splitting (symmetrical split of say K between two layers separates the input and output channels into K equal groups and connects only the corresponding input-output channel groups). (Datasets: SUN Attributes & CAMIT-NSAD) ,(Architectures: GoogleNet and VGG-11)
Object Skeleton Extraction in Natural Images by Fusing Scale-associated Deep Side Outputs	Subject: Object Detection & object skeleton Extraction. object skeleton extraction in natural images is a very challenging problem, as it requires the extractor to be able to capture both local and global image context to determine the intrinsic scale of each skeleton pixel. Existing methods rely on per-pixel based multi-scale feature computation, which results in difficult modeling and high time consumption. Authors present a fully convolutional network with multiple scale-associated side outputs to address this problem by introducing a scale-associated side output to each stage. By pointing out the relationship between the receptive field sizes of the sequential stages in the network and the skeleton scales they can capture, the responses of the multiple scale-associated side outputs are then fused in a scale-specific way to localize skeleton pixels with multiple scales effectively.
First Person Action Recognition Using Deep Learned Descriptors	Authors propose convolutional neural networks (CNNs) for end to end learning and classification of wearer's actions. The proposed network makes use of egocentric cues by capturing hand pose, head motion and saliency map. It's a three-stream architecture which uses egocentric cues in the first stream and complements it with pre-trained spatial and temporal streams from third person video analysis.
DeepHand: Robust Hand Pose Estimation by Completing a Matrix Imputed with Deep Features	They propose DeepHand to estimate the 3D pose of a hand using depth data from commercial 3D sensors. Authors discriminatively train convolutional neural networks to output a low dimensional activation feature given a depth map. Their database of activation features supplements large viewpoint coverage and their hierarchical estimation of pose parameters is robust to occlusions while achieving real time performance.
Gradual DropIn of Layers to Train Very Deep Neural Networks	**Gradual DropIn** is *dynamically growing a neural network during training*. (an untrainable deep network starts as a trainable shallow network and newly added layers are slowly, organically added during training, thereby increasing the network's depth. *adding DropIn Layers*). Consequently, deep networks, which are untrainable with conventional methods, will converge with DropIn layers interspersed in the architecture. DropIn provides regularization during training in an analogous way as dropout. (Datasets: MNIST, CIFAR and ImageNet) (Archtectures: LeNet, AlexNet and VGG-16) *if the shallow network is trainable, then the deeper network, where additional layers are added by a DropIn layer, is also trainable.* By asynchronous DropIn, layers are added starting at different iterations.
Deep Metric Learning via Lifted Structured Feature Embedding	Subject: Learning the distance metric between pairs of examples. It's a deep feature embedding and metric learning algorithm for taking full advantage of the training batches in the neural network training by lifting the vector of pairwise distances within the batch to the matrix of pairwise distances. (Datasets: CUB-200-2011, CARS196 and Stanford Online Products datasets) (Architecture: GoogLeNet) (Souce code is available at: https://github.com/rksltnl/Deep-Metric-Learning-CVPR16)

Deep Sliding Shapes for Amodal 3D Object Detection in RGB-D Images	Subject: amodal *3D object detection* in RGB-D images, which aims to produce a 3D bounding box of an object in metric form at its full extent. *Deep Sliding Shapes* is a 3D ConvNet formulation (3D ConvNet pipeline) that takes a 3D volumetric scene from a RGB-D image as input and outputs 3D object bounding boxes. They proposed the first 3D Region Proposal Network (RPN) to learn objectness from geometric shapes and the first joint Object Recognition Network (ORN) to extract geometric features in 3D and color features in 2D.
Deep Canonical Time Warping (DCTW)	Subject: time-series Analysis, in which the temporal alignment of time-series is a crucial challenge. The vast majority of algorithms oriented towards the temporal alignment of time-series are applied directly on the observation space, or utilise simple linear projections. Thus, they fail to capture complex, hierarchical non-linear representations which may prove to bebeneficial towards the task of temporal alignment, particularly when dealing with multi-modal data (e.g., aligning visual and acoustic information). So authors proposed Deep Canonical Time Warping (DCTW), a method which automatically learns complex non-linear representations of multiple time-series and discovers a hierarchical non-linear feature transformation for multiple sequences for temporal alignment of highly heterogeneous features, such as acoustic and visual features.
Deep Gaussian Conditional Random Field Network: A Model-based Deep Network for Discriminative Denoising	Subject: Image Denoising. An end-to-end trainable deep network architecture for image denoising based on a Gaussian Conditional Random Field (GCRF) model. In contrast to the existing discriminative denoising methods that train a separate model for each individual noise level, the proposed deep network explicitly models the input noise variance and hence is capable of handling a range of noise levels that consists of two sub-networks: (i) a parameter generation network that generates the pairwise potential parameters based on the noisy input image (ii) an inference network whose layers perform the computations involved in an iterative GCRF inference procedure. (Datasets: Berkeley segmentation and PASCALVOC datasets)
D 3 : Deep Dual-Domain Based Fast Restoration of JPEG-Compressed Images	The authors proposed a Deep Dual-Domain (D 3) based fast restoration model to remove artifacts of JPEG compressed images. They took into consideration both the prior knowledge of the JPEG compression scheme, and the successful practice of the sparsity-based dual-domain approach. They further design the One-Step Sparse Inference (1-SI) module, as an efficient and light-weighted feed-forward approximation of sparse coding.
Learning Deep Structure-Preserving Image-Text Embeddings	Subject: image-to-text and text-to-image retrieval, This paper proposes a method for learning joint embeddings of images and text using a two-branch neural network with multiple layers of linear projections followed by nonlinearities. The network is trained using a large-margin objective that combines cross-view ranking constraints with within-view neighborhood structure preservation constraints inspired by metric learning literature. (Datasets: Flickr30K and MSCOCO image-sentence datasets)
Studying Very Low Resolution Recognition Using Deep Networks	Subject: visual recognition on low-resolution images, Typically, the region of interest (ROI) in a Very Low Resolution Recognition (VLRR) problem can be smaller than 16 × 16 pixels, and is challenging to be recognized even by human experts. Taking advantage of techniques primarily in super resolution, domain adaptation and robust regression, we formulate a dedicated deep learning method and demonstrate how these techniques are incorporated step by step. The resulting Robust Partially Coupled Networks achieves feature enhancement and recognition simultaneously. The effectiveness of the proposed models is evaluated on three different VLRR tasks, including face identification, digit recognition and font recognition.
Constrained Deep Transfer Feature Learning and its Applications	Subject: Feature learning with deep models for data representation and classification, Deep feature learning, typically requires a large amount of training data, which may not be feasible for some application domains. Existing transfer learning methods typically perform one-shot transfer learning and often ignore the specific properties that the transferred data must satisfy. Authors introduced a constrained deep transfer feature learning method to perform simultaneous transfer learning and feature learning by performing transfer learning in a progressively improving feature space iteratively in order to better narrow the gap between the target domain and the source domain for effective transfer of the data from source domain to target domain. Furthermore, they proposed to exploit the target domain knowledge and incorporate such prior knowledge as constraint during transfer learning to ensure that the transferred data satisfies certain properties of the target domain. They apply it to thermal feature learning for eye detection by transferring from the visible domain. They also applied the proposed method for [illegible] as a second application.
Learning Deep Feature Representations with Domain Guided Dropout for Person Re-identification	In this work, they presented a pipeline for learning deep feature representations from multiple domains with Convolutional Neural Networks (CNNs) where none of them are large enough to provide abundant data variations. When training a CNN with data from all the domains, some neurons learn representations shared across several domains, while some others are effective only for a specific one. Based on this important observation, they proposed a Domain Guided Dropout algorithm to improve the feature learning procedure. (Datasets: CUHK01,CUHK03 and PRID) see the explanations to understand differences of the domains.
Deep Interactive Object Selection	Subject: Interactive Object Selection by combining user interactions with current deep learning models. Authors proposed a novel deep-learning-based algorithm which has much better understanding of objectness and can reduce user interactions to just a few clicks. Their algorithm transforms user-provided positive and negative clicks into two Euclidean distance maps which are then concatenated with the RGB channels of images to compose (image, user interactions) pairs. They generate many of such pairs by combining several random sampling strategies to model users' click patterns and use them to finetune deep Fully Convolutional Networks (FCNs). Finally the output probability maps of their FCN-8s model is integrated with graph cut optimization to refine the boundary segments. (Dataset: PASCAL segmentation dataset)
End-to-End Learning of Deformable Mixture of Parts and Deep Convolutional Neural Networks for Human Pose Estimation	Subject: Human Pose Estimation, (learning better feature representations and capturing contextual relationships) It is difficult to incorporate domain prior knowledge such as geometric relationships among body parts into DCNNs. They proposed to combines DCNNs with the expressive deformable mixture of parts. They explicitly incorporate domain prior knowledge into the framework, which greatly regularizes the learning process and enables the flexibility of our framework for loopy models or tree-structured models.

Joint **Unsupervised** Learning of Deep Representations and Image Clusters	Authors proposed a recurrent framework for joint unsupervised learning of deep representations and image clusters where successive operations in a clustering algorithm are expressed as steps in a recurrent process. In this work image clustering is conducted in the forward pass, while representation learning in the backward pass in such a way that ***good representations are beneficial to image clustering and clustering results provide supervisory signals to representation learning.*** (Code is available at: https://github.com/jwyang/joint-unsupervised-learning)
Highlight Detection with Pairwise Deep Ranking for First-Person Video Summarization	Subject: first-person video summarization or discovery of moments of user's major or special interest (i.e., highlights) in a video Authors proposed a pairwise deep ranking model that employs deep learning techniques to learn the relationship between high-light and non-highlight video segments in which video time-lapse plays the highlight (non-highlight) segments at low (high) speed rates, while the video skimming assembles the sequence of segments with the highest scores.
Efficient Training of Very Deep Neural Networks for Supervised Hashing	Subject: very deep supervised hashing (VDSH), Authors proposed a training algorithm inspired by alternating direction method of multipliers (ADMM) that decomposes the training process into independent layer-wise local updates through auxiliary variables. This training algorithm always converges and its computational complexity is linearly proportional to the number of edges in the networks. (train DNNs with 64 hidden layers and 1024 nodes per layer for supervised hashing in about 3 hours using a single GPU) (Datasets: datasets: MNIST, CIFAR-10 and NUS-WIDE)
Instance-Level Segmentation for *Autonomous Driving* with Deep Densely Connected MRFs	Subject: pixel-wise instance-level labeling of a monocular image in the context of autonomous driving, (a globally consistent instance labeling of the image) Authors proposed a densely connected Markov random field and show how to encode various intuitive potentials in a way that is amenable to efficient mean field inference. (Benchmark: KITTI)
Occlusion-free Face Alignment: Deep Regression Networks Coupled with De-corrupt AutoEncoders	Subject: Face alignment or facial landmark detection under partial occlussion, The performance of face alignment system degenerates severely when ***occlusions*** occur. In this work, authors propose a novel face alignment method, which cascades several Deep Regression networks coupled with De-corrupt Autoencoders (denoted as DRDA) to explicitly handle partial occlusion problem in which de-corrupt autoencoder network can automatically recover the genuine appearance for the occluded parts and the recovered parts can be leveraged together with those non-occluded parts for more accurate alignment. Moreover, this method can localize occluded regions rather than merely predict whether the landmarks are occluded.
Picking Deep Filter Responses for Fine-grained Image Recognition	Recognizing fine-grained sub-categories (e.g. birds and dogs) is extremely challenging due to the highly localized and subtle differences in some specific parts. Authors proposed an automatic fine-grained recognition approach which is free of any object / part annotation at both training and testing stages in the from of a unified framework based on two steps of deep filter response picking. Picking steps: 1) to find distinctive filters which respond to specific patterns significantly and consistently, and learn a set of part detectors via iteratively alternating between new positive sample mining and part model retraining. 2) to pool deep filter responses via spatially weighted combination of Fisher Vectors. Then they conditionally pick deep filter responses to encode them into the final representation.
Deep Region and Multi-label Learning for Facial Action Unit Detection	Subject: facial Action Unit (AU) detection, They proposed a Deep Region learning and multi-label learning (DRML), in which a novel region layer uses feed-forward functions to induce important facial regions, forcing the learned weights to capture structural information of the face. The complete network is end-to-end trainable, and automatically learns representations robust to variations inherent within a local region. (Benchmarks: BP4D and DISFA),
Improving the Robustness of Deep Neural Networks via Stability Training	Subject: solving output instability of deep neural networks against small perturbations, Authors presented a general stability training method to stabilize deep networks against small input distortions that result from various types of common image processing, such as compression, rescaling, and cropping. In general, it makes the output of a neural network more robust by training a model to be constant on images that are copies of the input image with small perturbations. As such, their method can enable higher performance on noisy visual data than a network without stability training. (Architecture: Inception architecture (GoogLeNet Blocks))
Learning Deep Features for Discriminative Localization	Subject: object localization by investigation of the global average pooling layer & their localization ability, Authors proposed a general technique called Class Activation Mapping (CAM) for CNNs with global average pooling that is able to localize the discriminative image regions despite just being trained for solving classification task. This enables classification-trained CNNs to learn to perform object localization, without using any bounding box annotations.
A Key Volume Mining Deep Framework for Action Recognition	Most existing deep frameworks equally treat every volume i.e. spatial-temporal video clip, and directly assign a video label to all volumes sampled from it. However, within a video, discriminative actions may occur sparsely in a few key volumes, and most other volumes are irrelevant to the labeled action category. Training with a large proportion of irrelevant volumes will hurt performance. Authors propose a key volume mining deep framework to identify key volumes and conduct classification simultaneously. Specifically, their framework is trained is optimized in an alternative way integrated to the forward and backward stages of Stochastic Gradient Descent (SGD). In the forward pass, our network mines key volumes for each action class. In the backward pass, it updates network parameters with the help of these mined key volumes. In addition, they proposed "Stochastic out" to model key volumes from multi-modalities, and an effective yet simple "unsupervised key volume proposal" method for high quality volume sampling.

Fast Training of Triplet-based Deep Binary Embedding Networks	Here, authors proposed to formulate high-order binary codes learning as a multi-label classification problem by explicitly separating learning into two interleaved stages: 1) They designed a large-scale high-order binary codes inference algorithm to reduce the high-order objective to a standard binary quadratic problem such that graph cuts can be used to efficiently infer the binary codes which serve as the labels of each training datum. 2) They proposed to map the original image to compact binary codes via carefully designed deep convolutional neural networks (CNNs) and the hashing function fitting can be solved by training binary CNN classifiers. Finally, An incremental/interleaved optimization strategy is proffered to ensure that these two steps are interactive with each other during training for better accuracy.