Z. Li
This thesis is dedicated to exploring locality for visual recognition. We consider traditional computer vision challenges like image classification and segmentation, attribute classification, action recognition and object tracking, which are not fully aware of locality, and ask ourselves the question: How can visual recognition profit from locality? We start with deriving an image representation for efficient classification and segmentation by decomposing objects locally. Then, we move on to exploit locality for attribute learning of objects, enabling zero-shot classification and segmentation. Beyond static images, the thesis also seeks to localize objects and actions in videos. We first propose a video representation based on neural networks which is able to capture the spatio-temporal correlations and attentions of actions in a video by relying on action class labels only. Finally, we make a connection between natural language and object tracking by approaching the problem of tracking by natural language specification. A brief summary of each chapter is provided as follows:

Chapter 2: Segmenting, classifying and searching objects locally by Codemaps. Starting from the observation that both the state-of-the-art image representations, as well as object classification machines are composed of summations of constituting terms, we derive an image representation that decomposes the classification scores of individual local elements in the image and their normalization coefficients. Therefore, we are able to leverage heavy, state-of-the-art feature encodings and reconstruct the classification score of arbitrary regions including normalization, which is vital for successful recognition. As a result, the decomposition facilitates state-of-the-art semantic segmentation and classification, while bringing significant efficiency gain.

Chapter 3: Attributes make more sense on localized objects. We revisit attribute-based representations, approaching them from the perspective of locality. Where existing work considers attributes of objects either for the global image or for the parts of the object, we propose to learn and infer attributes on segments localizing the entire object. Object-level attributes suffer less from accidental content around the object and accidental image conditions such as partial occlusions, scale changes and viewpoint changes. A joint learning framework is developed for simultaneous object classification and segment proposal ranking, solely on the basis of attributes. We experimentally show that our attribute learning on segmented objects leads to far more accurate results over full-image and part-based methods. Consequently, it also allows for zero-shot classification and segmentation, where we use attribute descriptions of unseen classes for localizing their instances in novel images and classifying them accordingly.

Chapter 4: VideoLSTM convolves, attends and flows for action recognition in videos. Rather than adapting the video to the peculiarities of established recurrent or convolutional network architectures, we aim to adapt the architecture to fit the requirements of the video medium. Therefore, we present a network architecture VideoLSTM, which integrates convolution, attention and flow, for end-to-end sequence learning of actions in videos. To exploit both the spatial and temporal correlations in a video, we first hardwire convolutions in the soft-Attention LSTM to capture its spatio-temporal dynamics. We also introduce motion-based attention which guides better the attention towards the relevant spatial-temporal locations. In the end, we demonstrate how the attention generated from our VideoLSTM can be exploited for action localization by relying on action class labels only. Experimental studies on challenging datasets outline
the merits of our VideoLSTM against other LSTM architectures for action classification, while revealing promising results for action localization.

**Chapter 5: Tracking objects by natural language specification.** We make a connection between natural language and object tracking. Instead of specifying the target object in the first frame of a video by a bounding box, we strive to track the object specified by a natural language description. Tracking by natural language specification allows for a novel type of human-machine interaction in object tracking and adapts to large appearance variations. To approach this new problem, we define three variants of tracking by language specification: one relying on lingual target specification only, one relying on visual target specification based on language, and one leveraging their joint potential. And we illustrate that enriching standard tracking by bounding box with our language specification helps against drift. Finally, we sketch new tracking scenarios in surveillance and other live video streams that become feasible with a language specification of the target, such as randomly start tracking or simultaneous multi-video tracking.