The overarching theme for the research conducted under our BDD-sponsored project this year was unsupervised/self-supervised visual prediction. Prediction is one of the central goals of machine learning, and is particularly vital in the automotive context. However, much of current research deals with prediction of semantic information, which has been previously annotated by human labelers. Our research, on the other hand, focuses on directly predicting raw visual signal, bypassing the issues associated with labelling. The central benefit is that we can build and train high-capacity prediction models without the need for costly human annotation. Additionally, deep feature hierarchies that has been trained for a particular visual prediction task turn out, somewhat surprisingly, to also be useful for a number of other, often unrelated, semantic and non-semantic tasks (see below). So, as a secondary benefit of our research, we have successfully applied visual prediction for self-supervised feature learning of deep models. Below, we present several visual data prediction tasks that we have been developing this year:

**Color Prediction**
Given a grayscale photograph as input, this paper attacks the problem of hallucinating a plausible color version of the photograph. This problem is clearly underconstrained, so previous approaches have either relied on significant user interaction or resulted in desaturated colorizations. We propose a fully automatic approach that produces vibrant and realistic colorizations. We embrace the underlying uncertainty of the problem by posing it as a classification task and use class-rebalancing at training time to increase the diversity of colors in the result. The system is implemented as a feed-forward pass in a CNN at test time and is trained on over a million color images. We evaluate our algorithm using a "colorization Turing test," asking human participants to choose between a generated and ground truth color image. Our method successfully fools humans on 32% of the trials, significantly higher than previous methods. Moreover, we show that colorization can be a powerful pretext task for self-supervised feature learning, acting as a cross-channel encoder. This approach results in state-of-the-art performance on several feature learning benchmarks. 
*Publication: “Colorful Image Colorization”, Richard Zhang, Phillip Isola, Alexei A. Efros, ECCV’16*
*Code and models will be in the BDD repository*

**Spatial Context Prediction**
We present an unsupervised visual feature learning algorithm driven by context-based pixel prediction. By analogy with auto-encoders, we propose Context Encoders -- a convolutional neural network trained to generate the contents of an arbitrary image region conditioned on its surroundings. In order to succeed at this task, context encoders need to both understand the content of the entire image, as well as produce a plausible hypothesis for the missing part(s). When training context encoders, we have experimented with both a standard pixel-wise reconstruction loss, as well as a reconstruction plus an
adversarial loss. The latter produces much sharper results because it can better handle multiple modes in the output. We found that a context encoder learns a representation that captures not just appearance but also the semantics of visual structures. We quantitatively demonstrate the effectiveness of our learned features for CNN pre-training on classification, detection, and segmentation tasks. Furthermore, context encoders can be used for semantic inpainting tasks, either stand-alone or as initialization for non-parametric methods.

**Publication:** “Context Encoders: Feature Learning by Inpainting”, Deepak Pathak, Philipp Krähenbühl, Jeff Donahue, Trevor Darrell, Alexei A. Efros, CVPR'16

*Code and models are in the BDD repository*

---

**Novel View Prediction**

We address the problem of novel view synthesis: given an input image, synthesizing new images of the same object or scene observed from arbitrary viewpoints. We approach this as a learning task but, critically, instead of learning to synthesize pixels from scratch, we learn to copy them from the input image. Our approach exploits the observation that the visual appearance of different views of the same instance is highly correlated, and such correlation could be explicitly learned by training a convolutional neural network (CNN) to predict appearance flows – 2-D coordinate vectors specifying which pixels in the input view could be used to reconstruct the target view. Furthermore, the proposed framework easily generalizes to multiple input views by learning how to optimally combine single-view predictions.

**Publication:** “View Synthesis by Appearance Flow”, Tinghui Zhou, Shubham Tulsiani, Weilun Sun, Jitendra Malik, Alexei A. Efros, ECCV'16

*Code and models will be in the BDD repository*

---

**Generalized Image-to-Image prediction**

*(ongoing work)*

We investigate conditional adversarial networks as a general-purpose solution to image-to-image translation problems. These networks not only learn the mapping from input image to output image, but also learn a loss function to train this mapping. This makes it possible to apply the same generic approach to problems that traditionally would require very different loss formulations. We demonstrate that this approach is effective at synthesizing photos from label maps, reconstructing objects from edge
maps, and colorizing images, among other tasks. As a community, we no longer hand-engineer our mapping functions, and this work suggests we can achieve reasonable results without hand-engineering our loss functions either.


Code and models will be in the BDD repository

Split-Brain Autoencoders: generalized prediction-based feature learning (ongoing work)

We propose split-brain autoencoders, a straightforward modification of the traditional autoencoder architecture, for unsupervised representation learning. The method adds a split to the network, resulting in two disjoint sub-networks. Each sub-network is trained to perform a difficult task -- predicting one subset of the data channels from another. Together, the sub-networks extract features from the entire input signal. By forcing the network to solve cross-channel prediction tasks, we induce a representation within the network which transfers well to other, unseen tasks. This method achieves state-of-the-art performance on several large-scale transfer learning benchmarks.


Code and models will be in the BDD repository

Continued Funding Request for 2017

We request continued funding (at the same funding levels as in 2016) to continue our work on this multi-year effort. For 2017, we plan to focus on the following three directions:

1. Continue the on-going work on development of a **universal image-to-image translation architecture based on generative adversarial networks**. The impressive preliminary results (see above) demonstrate promise of the proposed approach and demand further investigation. E.g. while we have shown that the method can, given a single aerial/satellite image, automatically predict the map of the area (labelling roads, buildings, geographic landmarks, etc), or synthesize what a daylight scene might look like at night, we are yet to attempt using our approach for predicting appearance in the temporal direction. This will be one focus for 2017 – to try and generalize to video-to-video translation and apply it to the driving data being collected by BDD. At the same time, we will also investigate using our method for smart data augmentation, e.g. synthesizing rainy scenes from sunny scenes or snowy scenes.

2. Continue the on-going work on the development of prediction-based **Split-Brain autoencoders** for general multi-channel / multi-modal feature learning. We have shown the promise of the technique for cross-channel prediction between $L$ and $ab$ channels of a color image representation, as well as $RGB$ and $D$ channels of an RGBD representation. But we believe that
the method should be able to generalize to other modalities, such as multi-spectral, audio, and other sensor data that we might get access to (some via the BDD-sponsored data collection efforts). Indeed, the next step is to combine all the different input streams available to an automotive agent (visual, audio, positional, accelerometer, GPS, etc etc) and train a big cross-prediction model that is able to translate between all these modalities at the same time, building an extremely robust and universal representation. The main difficulty that needs to be overcome is that, as the number of domains/modalities grows, the feature capacity of each individual modality shrinks (given constant computing resources). One way we plan to tackle this problem is by applying a technique similar to distillation to the various inputs of the split-brain autoencoder to force it to learn a joint representation within the available network capacity.

3. Finally, next year we plan to make a big push towards self-supervised online learning. Most contemporary machine learning is practiced in batch mode, where a well-defined “training set” is used to train a model which is then being evaluated on a “test set”. While batch training makes a lot of sense in an academic setting, where different algorithms need to be evaluated on common benchmarks, for real-world problems, online training offers a number of important advantages. In online training, also known as life-long learning, every new piece of data (e.g. video frame) is first considered “test data” and then immediately used as “training data” to update the current model (using a self-supervised training paradigm). This offers several important advantages, particularly in the automotive domain: 1) there is no need to store the acquired training data – every data point, once used in training can be safely discarded afterwards, 2) subsequently, there is little worry of overfitting, as the model only exposed to novel data, without ever having to see the same bit of data more than once, 3) graduate domain shift is also handled gracefully by online learning, as the system slowly adjusts from one domain to the next as the incoming data distribution starts to change. Despite these advantages, online training has not seen my use in practice, mainly due to the difficulties of coming up with effective self-supervised tasks that could be used for training. Here, we believe that our work on self-supervised prediction, such as Split-Brain Autoencoders, could be a perfect candidate to try and revive the fortunes of online training.

Interaction with BDD sponsors

During 2016, the PI as well as members of his research group have given a number of talks at various BDD functions. Additionally, we have had one-on-one meetings with several members of the consortium. We will be happy to continue doing this in the coming year. All the work (software, models, data) produced under this grant has been (or in the process of being) placed in the BDD repository. We plan to continue this practice into 2017.