DESIRE: Distant Future Prediction in Dynamic Scenes with Interacting Agents

Namhoon Lee[1], Wongun Choi[2], Paul Vernaza[2], Christopher B. Choy[3], Philip H. S. Torr[1], Manmohan Chandraker[2,4]

[1]University of Oxford, [2]NEC Labs America, [3]Stanford University, [4]UCSD

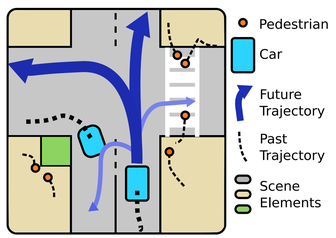
IEEE 2017 Conference on Computer Vision and Pattern Recognition

CVPR 2017

DEep Stochastic IOC RNN Encoder-decoder

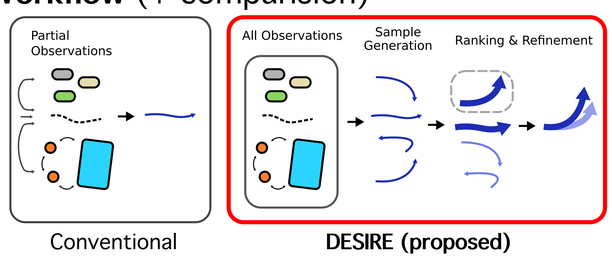
- **DESIRE** is a framework for future prediction, which predicts probable outcomes for agents in the scene in terms of trajectories given a series of past events.

Motivation (scenario + why it's hard)



- Reason from past motion history, scene context and interactions among agents.
- Account for the multi-modality nature of the future prediction.
- Foresee potential future outcomes for a strategic prediction.
- Need time-profile prediction to account for agents' influence.

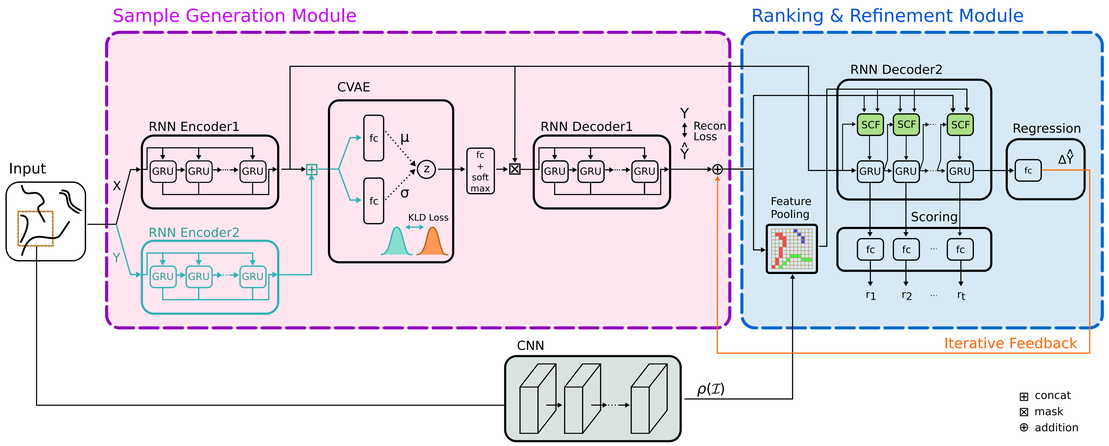
Workflow (+ comparision)



DESIRE Characteristics

- **Scalability**: The use of deep learning allows for end-to-end training and easy incorporation of multiple cues.
- **Diversity**: CVAE is combined with RNN encodings to generate stochastic prediction hypotheses to hallucinate multi-modalities.
- **Accuracy**: The IOC-based framework accumulates long-term future rewards and the refinement module learns to estimate a deformation of the trajectory, enabling more accurate predictions.

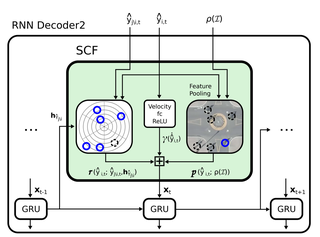
Architecture



Diverse Sample Generation with CVAE

- CVAE introduces stochastic latent variable z that are learned to encode a diverse set of predictions Y given input X, making it suitable for modeling one-to-many mapping.
- During training, Q is learned such that it gives higher probability to z that is likely to produce a reconstruction Y close to actual prediction given the full context X and Y.
  (loss terms: $\ell_{Recon} = \frac{1}{K}\sum_k \|Y_i - \hat{Y}_i^{(k)}\|$, $\ell_{KLD} = D_{KL}(Q_\phi(z_i|Y_i, X_i)\|P_\nu(z_i))$)
- At test time, z is sampled randomly from the prior distribution and decoded through the decoder network to produce a prediction hypothesis.

Scene Context Fusion



$\mathbf{x}_t = \left[\gamma(\hat{v}_{i,t}), p(\hat{y}_{i,t}; \rho(\mathcal{I})), r(\hat{y}_{i,t}; \hat{y}_{j\setminus i,t}, \mathbf{h}_{\hat{Y}_{j\setminus i}})\right]$

IOC-based Ranking and Refinement

- RNN model assigns rewards to each prediction hypothesis and measures their goodness based on the accumulated long-term rewards. (cross-entropy loss $\ell_{CE} = H(p,q)$, where $p = softmax_k(\sum_{t=1}^{T} r_{i,t}^{(k)})$, $q = softmax_k(-max_t\|\hat{Y}_{i,t}^{(k)} - Y_{i,t}\|)$)
- At the same time, prediction hypotheses get refined by learning displacements ΔY to the actual prediction Y. (regression loss $\ell_{Reg} = \frac{1}{K}\sum_k \|Y_i - \hat{Y}_i^{(k)} - \triangle\hat{Y}_i^{(k)}\|$)
- Module receives iterative feedbacks from regressed predictions and keeps adjusting so that it produces precise predictions at the end.

Experiments

Iterative refinement

Prediction results

(10% acc. for CVAE and DESIRE)

	KITTI (error in meters / miss-rate with 1m threshold)				SDD (pixel error at 1/5 resolution)			
Method	1s	2s	3s	4s	1s	2s	3s	4s
Linear	0.89 / 0.31	2.07 / 0.49	3.67 / 0.59	5.62 / 0.64	2.58	5.37	8.74	12.54
RNN ED-SI	0.56 / 0.16	1.40 / 0.44	2.65 / 0.58	4.29 / 0.65	1.51	3.56	6.04	8.80
CVAE	0.35 / 0.06	0.93 / 0.30	1.81 / 0.49	3.07 / 0.59	1.84	3.93	6.47	9.65
DESIRE-S-IT0	0.32 / 0.05	0.84 / 0.26	1.67 / 0.43	2.82 / 0.54	1.59	3.31	5.27	7.75
DESIRE-SI-IT4	**0.28 / 0.04**	**0.67 / 0.17**	**1.22 / 0.29**	**2.06 / 0.41**	**1.29**	**2.35**	**3.47**	**5.33**