How to Get Your Paper Rejected

Wen-Hui Chen
2023/11/06

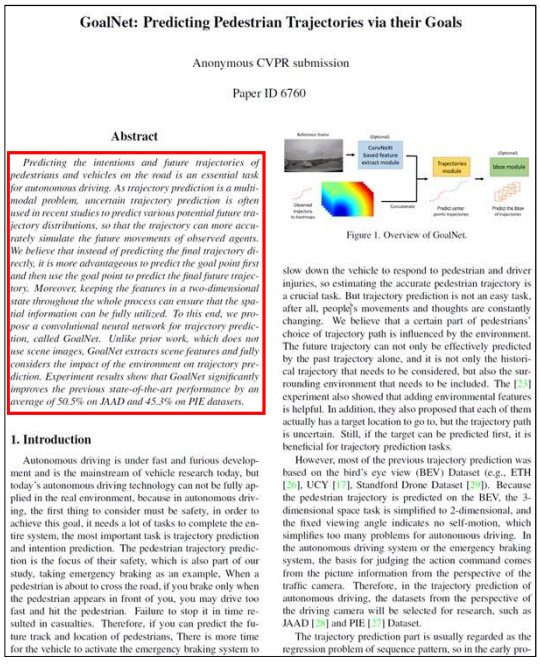
Top Publications

	文章	H5 指數	H5 中位數
1.	Nature	467	707
2.	The New England Journal of Medicine	439	876
3.	Science	424	665
4.	IEEE/CVF Conference on Computer Vision and Pattern Recognition	422	681
5.	The Lancet	368	688
6.	Nature Communications	349	456
7.	Advanced Materials	326	415
8.	Cell	316	503
9.	Neural Information Processing Systems	309	503
10.	International Conference on Learning Representations	303	563
11.	JAMA	286	476
12.	Science of The Total Environment	273	375
13.	Nature Medicine	268	459

https://scholar.google.com/citations?view_op=top_venues

References

- How to Get Your CVPR Paper Rejected? by Ming-Hsuan Yang (UC Merced), 2017.
- How to write a good CVPR submission, by Bill Freeman (MIT), 2014.



GoalNet: Predicting Pedestrian Trajectories via their Goals

Anonymous CVPR submission

Paper ID 6760

Abstract

Predicting the intentions and future trajectories of pedestrians and vehicles on the road is an essential task for autonomous driving. As trajectory prediction is a multi-modal problem, uncertain trajectory prediction is often used in recent studies to predict various potential future trajectory distributions, so that the trajectory can more accurately simulate the future movements of observed agents. We believe that instead of predicting the final trajectory directly, it is more advantageous to predict the goal point first and then use the goal point to predict the final future trajectory. Moreover, keeping the features in a two-dimensional state throughout the whole process can ensure that the spatial information can be fully utilized. To this end, we propose a convolutional neural network for trajectory prediction, called GoalNet. Unlike prior work, which does not use scene images, GoalNet extracts scene features and fully considers the impact of the environment on trajectory prediction. Experiment results show that GoalNet significantly improves the previous state-of-the-art performance by an average of 50.5% on JAAD and 45.3% on PIE datasets.

1. Introduction

Autonomous driving is under fast and furious development and is the mainstream of vehicle research today, but today's autonomous driving technology can not be fully applied in the real environment, because in autonomous driving, the first thing to consider must be safety, in order to achieve this goal, it needs a lot of tasks to complete the entire system, the most important task is trajectory prediction and intention prediction. The pedestrian trajectory prediction is the focus of their safety, which is also part of our study, taking emergency braking as an example, When a pedestrian is about to cross the road, if you brake only when the pedestrian appears in front of you, you may drive too fast and hit the pedestrian. Failure to stop it in time resulted in casualties. Therefore, if you can predict the future track and location of pedestrians, There is more time for the vehicle to activate the emergency braking system to slow down the vehicle to respond to pedestrian and driver injuries, so estimating the accurate pedestrian trajectory is a crucial task. But trajectory prediction is not an easy task, after all, people's movements and thoughts are constantly changing. We believe that a certain part of pedestrians' choice of trajectory path is influenced by the environment. The future trajectory can not only be effectively predicted by the past trajectory alone, and it is not only the historical trajectory that needs to be considered, but also the surrounding environment that needs to be included. The [23] experiment also showed that adding environmental features is helpful. In addition, they also proposed that each of them actually has a target location to go to, but the trajectory path is uncertain. Still, if the target can be predicted first, it is beneficial for trajectory prediction tasks.

Figure 1. Overview of GoalNet.

However, most of the previous trajectory prediction was based on the bird's eye view (BEV) Dataset (e.g., ETH [26], UCY [17], Standford Drone Dataset [29]). Because the pedestrian trajectory is predicted on the BEV, the 3-dimensional space task is simplified to 2-dimensional, and the fixed viewing angle indicates no self-motion, which simplifies too many problems for autonomous driving. In the autonomous driving system or the emergency braking system, the basis for judging the action command comes from the picture information from the perspective of the traffic camera. Therefore, in the trajectory prediction of autonomous driving, the datasets from the perspective of the driving camera will be selected for research, such as JAAD [28] and PIE [27] Dataset.

The trajectory prediction part is usually regarded as the regression problem of sequence pattern, so in the early pro-

- Motivation
- Problem
- Method
- Result
- Conclusion

Research Gaps?

Predicting the intentions and future trajectories of pedestrians and vehicles on the road is an essential task for autonomous driving. As trajectory prediction is a multimodal problem, uncertain trajectory prediction is often used in recent studies to predict various potential future trajectory distributions, so that the trajectory can more accurately simulate the future movements of observed agents. We believe that instead of predicting the final trajectory directly, it is more advantageous to predict the goal point first and then use the goal point to predict the final future trajectory. Moreover, keeping the features in a two-dimensional state throughout the whole process can ensure that the spatial information can be fully utilized. To this end, we propose a convolutional neural network for trajectory prediction, called GoalNet. Unlike prior work, which does not use scene images, GoalNet extracts scene features and fully considers the impact of the environment on trajectory prediction. Experiment results show that GoalNet significantly improves the previous state-of-the-art performance by an average of 50.5% on JAAD and 45.3% on PIE datasets.

- What are the research gaps relevant to the problem?
- How do your model play a role in improving pedestrian trajectory prediction?
- How did your research address the issue of uncertainty in pedestrian behavior and its impact on trajectory prediction?

Abstract

Predicting the intentions and future trajectories of pedestrians and vehicles on the road is an essential task for autonomous driving. As trajectory prediction is a multi-modal problem, uncertain trajectory prediction is often used in recent studies to predict various potential future trajectory distributions, so that the trajectory can more accurately simulate the future movements of observed agents. We believe that instead of predicting the final trajectory directly, it is more advantageous to predict the goal point first and then use the goal point to predict the final future trajectory. Moreover, keeping the features in a two-dimensional state throughout the whole process can ensure that the spatial information can be fully utilized. To this end, we propose a convolutional neural network for trajectory prediction, called GoalNet. Unlike prior work, which does not use scene images, GoalNet extracts scene features and fully considers the impact of the environment on trajectory prediction. Experiment results show that GoalNet significantly improves the previous state-of-the-art performance by an average of 50.5% on JAAD and 45.3% on PIE datasets.

1. Introduction

Autonomous driving is under fast and furious develop-

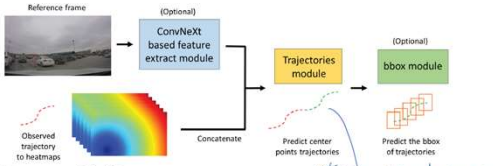


Figure 1. Overview of GoalNet.

hicle to respond to pedestrian and driver injuries, so estimating the accurate pedestrian trajectory is a crucial task. But trajectory prediction is not an easy task, after all, people's movements and thoughts are constantly changing. We believe that a certain part of pedestrians' choice of trajectory path is influenced by the environment. The future trajectory can not only be effectively predicted by the past trajectory alone, and it is not only the historical trajectory that needs to be considered, but also the surrounding environment that needs to be included. The [19] experiment also showed that adding environmental features is helpful. In addition, they also proposed that each of them actually has a target location to go to, but the trajectory path is uncertain, but if the target can be predicted first, it is very helpful for trajectory prediction tasks.

However, Most of the previous trajectory prediction was

keep the font style consistent

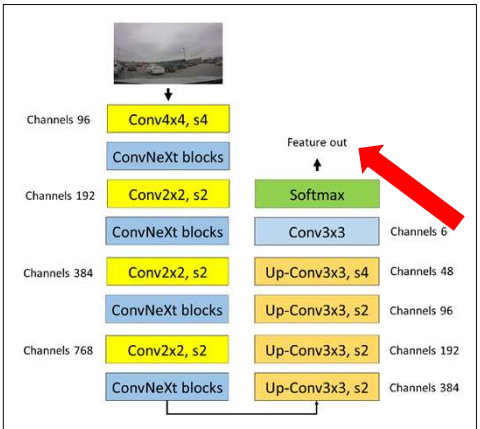


Figure 2. **GoalNet feature extract module architecture.** It consists of a ConvNeXt-T based encoder and a lightweight decoder based on Upsample convolution.

GoalNet consists of three modules, which are feature extraction, trajectory prediction, and bounding box prediction module, and we estimate the number of parameters and flops of our model as shown in Table 1.

modules	# param. parameters	FLOPs flops
GoalNet - extract	31.347M	22.770G
GoalNet - trajectory	2.131M	20.821G
GoalNet - bbox	0.174M	0.174M
Total	33.652M	43.591G

Table 1. GoalNet parameters & flops.

However, most of the previous trajectory prediction was based on the bird's eye view (BEV) Dataset (e.g., ETH [26], UCY [17], Standford Drone Dataset [29]). Because the pedestrian trajectory is predicted on the BEV, the 3-dimensional space task is simplified to 2-dimensional, and the fixed viewing angle indicates no self-motion, which simplifies too many problems for autonomous driving. In the autonomous driving system or the emergency braking system, the basis for judging the action command comes from the picture information from the perspective of the traffic camera. Therefore, in the trajectory prediction of autonomous driving, the datasets from the perspective of the driving camera will be selected for research, such as JAAD [28] and PIE [27] Dataset.

datasets

3. Proposed Method

The problem of trajectory prediction can be defined as follows. Suppose that the current time is t, and P represents the bounding box coordinates of the object with length and width (in pixels), then Pt represents the position of the object at time t. The given observation time is 15 frames, X = {Pt-14, Pt-13, ..., Pt}, our goal is to predict the position of X in the next 45 frames, Y = {Pt+1, Pt+2, ..., Pt+45}. For the prediction of stochastic trajectories, because the future is random, given an observation trajectory X, there may be multiple results, so various trajectories corresponding to these uncertain futures need to be generated to cover this uncertainty, the uncertain trajectories are expressed as Ys, and the corresponding n trajectories are generated. Ys = {Y_1, Y_2, ..., Y_n}.

We implemented the feature extraction module based on ConvNeXt [21], the structure of which is shown in Figure 2. Considering the computational load of this module and the trajectory prediction module, we chose the smallest Convnext-Tiny as the encoder. After extracting the feature map through Convnext-Tiny, the decoder uses the same but greatly simplified steps as downsample to upsample back the feature map to the size of the original RGB image and normalizes the features to probability distribution through softmax.

3.2. Trajectory module

In the trajectory prediction module, the same two-dimensional processing method as the environmental image can avoid the loss of spatial information contained in the environmental image. Some previous studies [1, 25, 34] mostly adopted the RNN-based trajectory prediction model.

[1] Alexandre Alahi, Kratarth Goel, Vignesh Ramanathan, Alexandre Robicquet, Li Fei-Fei, and Silvio Savarese. Social lstm: Human trajectory prediction in crowded spaces. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 961–971, 2016. 3, 4

[25] Seong Hyeon Park, ByeongDo Kim, Chang Mook Kang, Chung Choo Chung, and Jun Won Choi. Sequence-to-sequence prediction of vehicle trajectory via lstm encoder-decoder architecture. In *2018 IEEE intelligent vehicles symposium (IV)*, pages 1672–1678. IEEE, 2018. 4

[34] Hao Wu, Ziyang Chen, Weiwei Sun, Baihua Zheng, and Wei Wang. Modeling trajectories with recurrent neural networks. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, IJCAI'17, page 3083–3090. AAAI Press, 2017. 4

This module encoder the observed trajectory and the environment image, and the decoder predicts the goal map represented by a probability distribution, and calculates the main goal through SpatialSoftArgmax [18]. At the same

404
405
406
407

This module encodes the observed trajectory and the environmental image, while the decoder predicts the goal map represented by a probability distribution. The main goal is then calculated using SpatialSoftArgmax [18].

venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020. 2

[5] Yuning Chai, Benjamin Sapp, Mayank Bansal, and Dragomir Anguelov. Multipath: Multiple probabilistic anchor trajectory hypotheses for behavior prediction. In Leslie Pack Kaelbling, Danica Kragic, and Komei Sugiura, editors, *Proceedings of the Conference on Robot Learning*, volume 100 of *Proceedings of Machine Learning Research*, pages 86–99. PMLR, 30 Oct–01 Nov 2020. 2

[6] Ming-Fang Chang, John Lambert, Patsorn Sangkloy, Jagjeet Singh, Slawomir Bak, Andrew Hartnett, De Wang, Peter Carr, Simon Lucey, Deva Ramanan, et al. Argoverse: 3d tracking and forecasting with rich maps. In *Proceedings of national Conference on Computer Vision*, pages 2575- 2019. 2

[16] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[17] Alon Lerner, Yiorgos Chrysanthou, and Dani Lischinski. Crowds by example. *Computer Graphics Forum*, 26(3):655–664, 2007. 1

[18] Sergey Levine, Chelsea Finn, Trevor Darrell, and Pieter Abbeel. End-to-end training of deep visuomotor policies. *The Journal of Machine Learning Research*, 17(1):1334–1373, 2016. 4

[19] Junwei Liang, Lu Jiang, and Alexander Hauptmann. Simaug: Learning robust representations from 3d simulation for pedestrian trajectory prediction in unseen cameras. *arXiv preprint arXiv:2004.02022*, 2, 2020. 3

End-to-End Training of Deep Visuomotor Policies

Sergey Levine† SVLEVINE@EECS.BERKELEY.EDU
Chelsea Finn† CBFINN@EECS.BERKELEY.EDU
Trevor Darrell TREVOR@EECS.BERKELEY.EDU
Pieter Abbeel PABBEEL@EECS.BERKELEY.EDU
Division of Computer Science
University of California
Berkeley, CA 94720-1776, USA
†These authors contributed equally.

Editor: Jan Peters

Abstract

Policy search methods can allow robots to learn control policies for a wide range of tasks, but practical applications of policy search often require hand-engineered components for perception, state estimation, and low-level control. In this paper, we aim to answer the following question: does training the perception and control systems jointly end-to-end provide better performance than training each component separately? To this end, we develop a method that can be used to learn policies that map raw image observations directly to torques at the robot's motors. The policies are represented by deep convolutional neural networks (CNNs) with 92,000 parameters, and are trained using a guided policy search method, which transforms policy search into supervised learning, with supervision provided by a simple trajectory-centric reinforcement learning method. We evaluate our method on a range of real-world manipulation tasks that require close coordination between vision and control, such as screwing a cap onto a bottle, and present simulated comparisons to a range of prior policy search methods.

Keywords: Reinforcement Learning, Optimal Control, Vision, Neural Networks

1. Introduction

Robots can perform impressive tasks under human control, including surgery (Lanfranco et al., 2004) and household chores (Wyrobek et al., 2008). However, designing the perception and control software for autonomous operation remains a major challenge, even for basic

uses a simple quadratic augmented Lagrangian term, it further requires penalty terms on the gradient of the policy to account for local feedback. Our approach enforces this feedback behavior due to the higher moments included in the KL-divergence term, but does not require computing the second derivative of the policy.

5. End-to-End Visuomotor Policies

Guided policy search allows us to optimize complex, high-dimensional policies with raw observations, such as when the input to the policy consists of images from a robot's onboard camera. However, leveraging this capability to directly learn policies for visuomotor control requires designing a policy representation that is both data-efficient and capable of learning complex control strategies directly from raw visual inputs. In this section, we describe a deep convolutional neural network (CNN) model that is uniquely suited to this task. Our approach combines a novel spatial soft-argmax layer with a pretraining procedure that provides for flexibility and data-efficiency.

spatial × 10 of 32

Deep Spatial Autoencoders for Visuomotor Learning

Chelsea Finn, Xin Yu Tan, Yan Duan, Trevor Darrell, Sergey Levine, Pieter Abbeel

***Abstract*— Reinforcement learning provides a powerful and flexible framework for automated acquisition of robotic motion skills. However, applying reinforcement learning requires a sufficiently detailed representation of the state, including the configuration of task-relevant objects. We present an approach that automates state-space construction by learning a state representation directly from camera images. Our method uses a deep spatial autoencoder to acquire a set of feature points that describe the environment for the current task, such as the positions of objects, and then learns a motion skill with these feature points using an efficient reinforcement learning method based on local linear models. The resulting controller reacts continuously to the learned feature points, allowing the robot to dynamically manipulate objects in the world with closed-loop control. We demonstrate our method with a PR2 robot on tasks that include pushing a free-standing toy block, picking up a bag of rice using a spatula, and hanging a loop of rope on a hook at various positions. In each task, our method automatically learns to track task-relevant objects and manipulate their configuration with the robot's arm.**

I. Introduction

One of the fundamental challenges in applying reinforcement learning to robotic manipulation tasks is the need to define a suitable state space representation. Typically, this is handled manually, by enumerating the objects in the scene, designing perception algorithms to detect them, and feeding high-level information about each object to the algorithm. However, this highly manual process makes it difficult to apply the same reinforcement learning algorithm to a wide range of manipulation tasks in complex, unstructured environments. What if a robot could automatically identify the visual features that might be relevant for a particular task, and then learn a controller that accounts for those features? This would amount to automatically acquiring a vision system that is suitable for the current task, and would allow a range of object interaction skills to be learned with minimal human supervision. The robot would only need to see a glimpse of what task completion looks like, and could then figure out

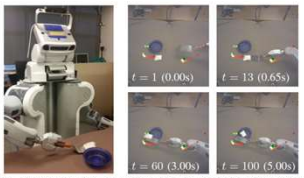


Fig. 1: PR2 learning to scoop a bag of rice into a bowl with a spatula (left) using a learned visual state representation (right).

using a spatial autoencoder architecture to learn a state representation that consists of feature points. Intuitively, these feature points encode the configurations of objects in the scene, and the spatial auto
for data-efficient learning
convolutional parameters
learned feature points cor
architecture also addresse
valued quantities such as
control than the types of
more commonly learned

In fact, our experimen
point representation can b
with an efficient trajecto
algorithm. This method
models to the states vis
estimating how the robo
joint torques) affect the st
feature points, this method can learn how the robot's actions affect the objects in the world, and the trained controllers can perform closed-loop control on the configuration of these objects, allowing the robot to move and manipulate them.

identified and localized using vision.

An overview of this procedure is provided in Figure 2. In the next section, we describe our representation learning method.

V. Unsupervised State Representation Learning from Visual Perception

The goal of state representation learning is to find a mapping $h_{\text{enc}}(I_t)$ from a high-dimensional observation I_t to a robot state representation for which it is easy to learn a control policy. We will use $\tilde{\mathbf{x}}_t$ to denote the configuration of the robot, $\mathbf{f}_t$ to denote the learned representation, and $\mathbf{s}_t = [\tilde{\mathbf{x}}_t; \mathbf{f}_t]$ to denote the final state space that combines

layer convolutional neural network with rectified linear units of the form $a_{cij} = max(0, z_{cij})$ for each channel c and pixel (i, j). We compute the spatial features from the last convolutional layer by performing a "spatial soft arg-max" operation that determines the image-space point of maximal activation in each channel of conv3. This set of maximal activation points forms our spatial feature representation and forces the autoencoder to focus on object positions. The spatial soft arg-max consists of two operations. The response maps of the third convolutional layer (conv3) are first passed through a spatial softmax $s_{cij} = e^{a_{cij}/\alpha} / \sum_{i'j'} e^{a_{ci'j'}/\alpha}$, where the temperature α is a learned parameter. Then, the expected 2D position of each softmax probability distribution $\mathbf{s}_c$ is

Entropy-driven Unsupervised Keypoint Representation Learning in Videos

Ali Younes [1] Simone Schaub-Meyer [1 2] Georgia Chalvatzaki [1 2 3]

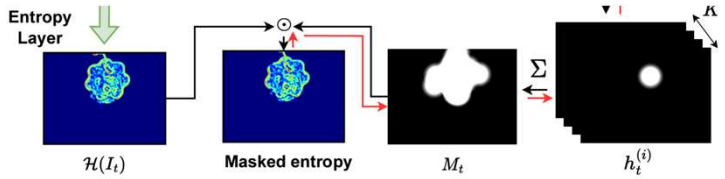


Figure 1. The architecture of our keypoint model $\Phi(I_t)$ (Section 2.2) and the masked entropy (Section 2.2.1). For an input image I_t our model $\Phi(I_t)$ outputs K feature maps $\mathrm{f}_t^{(i)}$ for each keypoint $\mathrm{k}_t^{(i)}$, $i \in \{1, \dots, K\}$. A heatmap $h_t^{(i)}$ is generated for each keypoint, while the active keypoints are aggregated to form the mask M_t. The entropy layer computes the entropy of the image $\mathcal{H}(I_t)$. Our ME loss maximizes the percentage of the entropy in the masked entropy image. Red arrows show the backward gradient flow. Only the part encircled by the dashed line is used during inference.

timating the entropy of the neighborhood region $R(x, y)$ centered at (x, y), using a normalized histogram-based discrete probability function $p(b, R(x, y))$ for each color value b in the region $R(x, y)$ summed and normalized over the color channels (details in Appendix B). The final per-pixel

It outputs K feature maps $\mathrm{f}_t^{(i)}$, each corresponding to one keypoint. The coordinates $(x_i, y_i)_t$ of the respective keypoint $k_t^{(i)}$ are obtained with a spatial soft-argmax (Levine et al., 2016). Besides predicting the coordinates, the model also assigns an activation status $s_t^{(i)} = \{0, 1\}$ per keypoint.

As a classic convolutional encoder-decoder network structure, U-Net [30] was originally designed for segmentation. Still, the trajectory is highly correlated with space, so it may bring benefits to designing a structure similar to U-Net. Y-Net [23] has proven this.

412
413
414
415
416

Batch Normalization (BN) [14] was used in the U-Net, but BN of the scene heatmap will destroy this information. So BN should be replaced with a batch independent normalization method, like Group normalization (GN) [35] or Layer Normalization (LN) [3], The GN restriction that channels must be divisible by the group severely limits model design, And we find that LN and Atrous Spatial Pyramid Pooling (ASPP) [7] match better. So we chose to use LN to substitute the BN.

We have found that LN and Atrous Spatial Pyramid Pooling (ASPP) are a better match. ◦

LN and Atrous Spatial Pyramid Pooling (ASPP) are a good match.

complement each other well.

provide better matches.

the combination of LN and Atrous Spatial Pyramid Pooling (ASPP) matches better.

Datasets. Joint Attention in Autonomous Driving (JAAD) [28] and Pedestrian Intention Estimation (PIE) [27] Dataset are used in our experiment. Trajectories for both datasets were recorded using an on-board camera, recorded and annotated at 30 frames per second (fps). JAAD contained 2,800 pedestrian trajectories out of 75,000 annotated frames, while PIE contained 1,800 pedestrian trajectories out of 293,000 annotated frames, with longer annotated trajectories and more comprehensive annotations than JAAD. We followed the JAAD and PIE standard training/testing split [27], using the same observation and prediction lengths as prior work [32, 36]. A trajectory with a length of 0.5 seconds (15 frames) is entered to generate a trajectory with a length of 0.5, 1.0, 1.5 seconds (15, 30, 45 frames).

input

Active vs. Passive voice

We input a trajectory with a duration of 0.5 seconds (15 frames) to generate trajectories with durations of 0.5, 1.0, and 1.5 seconds (15, 30, and 45 frames, respectively).

ries. All models were trained with batch size 8, Adam [16] optimizer with initial learning rate 1e-4, and a ReduceLROnPlateau LR scheduler on a single RTX2080Ti GPU.

All models were trained with a batch size of 8, the Adam optimizer [16] with an initial learning rate of 1e-4, and the ReduceLROnPlateau scheduler on a single RTX2080Ti GPU.

Negative Log Likelihood (NLL)

PIE JAAD SGNet BiTrap

Evaluation Metrics. Follow [27, 28, 32, 36], our GoalNet model uses mean square error (MSE), center mean square error (CMSE) and Center Final mean square error (CFMSE) to evaluate our performance on JAAD and PIE in pixels. ~~Where~~ MSE is calculated according to the upper left and lower right coordinates of the bounding box trajectory. CMSE and CFMSE are calculated using the center point of the bounding box trajectory.

Following / In line with

BiTrap

Methods	JAAD				PIE			
	ADE (0.5/1.0/1.5s)	C_{ADE} (1.5s)	C_{FDE} (1.5s)	NLL	ADE (0.5/1.0/1.5s)	C_{ADE} (1.5s)	C_{FDE} (1.5s)	NLL
Linear [23]	233/857/2303	1565	6111	-	123/477/1365	950	3983	-
LSTM [23]	289/569/1558	1473	5766	-	172/330/911	837	3352	-
B-LSTM [39]	159/539/1535	1447	5615	-	101/296/855	811	3259	-
FOL-X [22]	147/484/1374	1290	4924	-	47/183/584	546	2303	-
PIE_{traj} [23]	110/399/1280	1183	4780	-	58/200/636	596	2477	-
PIE_{full} [23]	-	-	-	-	-/-/556	520	2162	-
BiTraP-D	**93/378/1206**	**1105**	**4565**	-	**41/161/511**	**481**	**1949**	-
BiTraP-NP (20)	**38/94/222**	**177**	**565**	18.9	**23/48/102**	**81**	**261**	16.5
BiTraP-GMM (20)	153/250/585	501	998	**16.0**	38/90/209	171	368	**13.8**

[31] Tim Salzmann, Boris Ivanovic, Punarjay Chakravarty, and Marco Pavone. Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16*, pages 683–700. Springer, 2020. 2

[32] Chuhua Wang, Yuchen Wang, Mingze Xu, and David J Crandall. Stepwise goal-driven networks for trajectory prediction. *IEEE Robotics and Automation Letters*, 7(2):2716–2723, 2022. 2, 3, 5, 6

[33] Conghao Wong, Beihao Xia, Ziming Hong, Qinmu Peng, Wei Yuan, Qiong Cao, Yibo Yang, and Xinge You. View vertically: A hierarchical network for trajectory prediction via fourier spectrums. In *European Conference on Computer Vision*, pages 682–700. Springer, 2022. 2

[34] Hao Wu, Ziyang Chen, Weiwei Sun, Baihua Zheng, and Wei Wang. Modeling trajectories with recurrent neural networks. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, IJCAI'17, page 3083–3090. AAAI Press, 2017. 4

[35] Yuxin Wu and Kaiming He. Group normalization. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018. 4, 7

How to Get Your CVPR Paper Rejected?

Do Not

– Pay attention to the review process

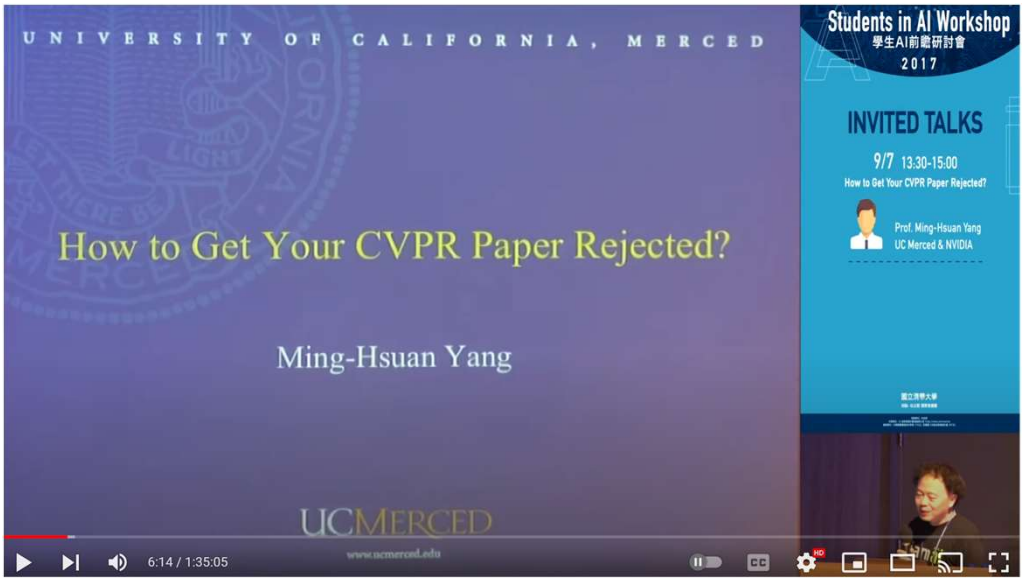
– Put yourself as a reviewer to examine your work from that perspective

– Put the work in the right context

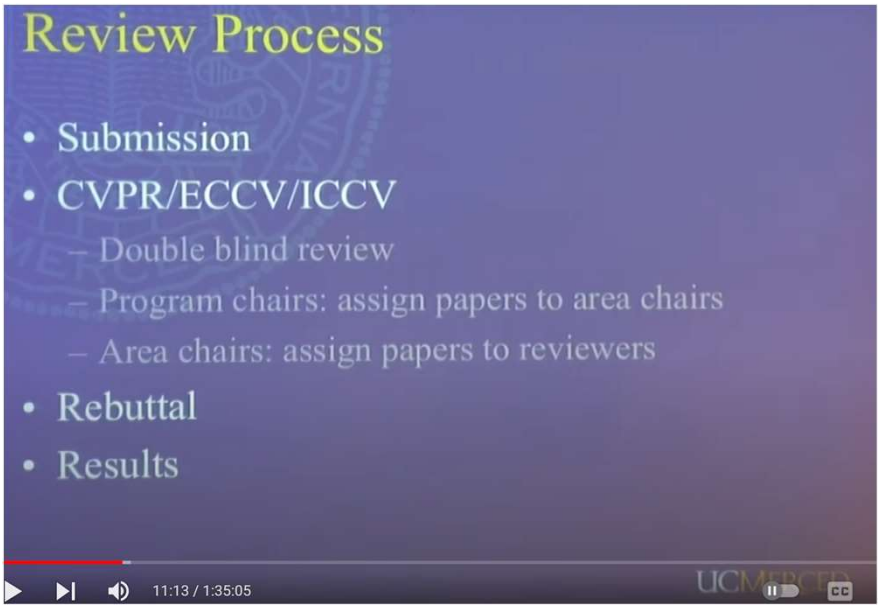
– Carry out a sufficient amount of experiments

– Compare with state-of-the-art algorithms

– Pay attention to writing



https://www.youtube.com/watch?v=jp_TGMU4ASI



Review Form

- Summary
- Overall Rating
  - Definite accept, weakly accept, borderline, weakly reject, definite reject
- Novelty
  - Very original, original, minor originality, has been done before
- Importance/relevance
  - Of broad interest, interesting to a subarea, interesting only to a small number of attendees, out of CVPR scope

UCMERCED

Review Form

- Clarity of presentation
  - Reads very well, is clear enough, difficult to read, unreadable
- Technical correctness
  - Definite correct, probably correct but did not check completely, contains rectifiable errors, has major problems
- Experimental validation
  - Excellent validation or N/A (a theoretical paper), limited but convincing, lacking in some aspects, insufficient validation
- Additional comments
- Reviewer's name

UCMERCED

Novelty

- What is new in this work?
  - Higher accuracy, significant speed-up, scale-up, ease to implement, generalization, wide application domain, connection among seemingly unrelated topics, ...
- What are the contributions (over prior art)?
- Make a compelling case with strong supporting evidence

UCMERCED

Compare with State of the Art

Need to show why one's method outperforms others, and in what way?

Writing

- Clear presentation
- Terse
- Careful about wording
- Make claims with strong evidence

GoalNet: Predicting Pedestrian Trajectories via their Goals

Anonymous CVPR submission

Paper ID 6760

Abstract

Predicting the intentions and future trajectories of pedestrians and vehicles on the road is an essential task for autonomous driving. As trajectory prediction is a multi-modal problem, uncertain trajectory prediction is often used in recent studies to predict various potential future trajectory distributions, so that the trajectory can more accurately simulate the future movements of observed agents. We believe that instead of predicting the final trajectory directly, it is more advantageous to predict the goal point first and then use the goal point to predict the final future trajectory. Moreover, keeping the features in a two-dimensional state throughout the whole process can ensure that the spatial information can be fully utilized. To this end, we propose a convolutional neural network for trajectory prediction, called GoalNet. Unlike prior work, which does not use scene images, GoalNet extracts scene features and fully considers the impact of the environment on trajectory pre-

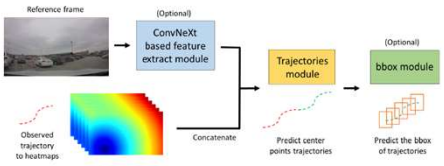


Figure 1. Overview of GoalNet.

slow down the vehicle to respond to pedestrian and driver injuries, so estimating the accurate pedestrian trajectory is a crucial task. But trajectory prediction is not an easy task, after all, people's movements and thoughts are constantly changing. We believe that a certain part of pedestrians' choice of trajectory path is influenced by the environment. The future trajectory can not only be effectively predicted by the past trajectory alone, and it is not only the historical trajectory that needs to be considered, but also the sur-

Get Results First than Writing?

- Conventional mode
  - Idea-> Do research -> Write paper
- "How to write a great research paper" by Simon Peyton Jones
  - Idea -> Write paper -> Do research
    - Forces us to be clear, focused
    - Crystallizes what we don't understand
    - Opens the way to dialogue with others: reality check, critique, and collaboration
- My take
  - Idea -> Write paper -> Do research -> Revise paper -> Do research -> Revise paper -> …

UCMERCED

Salami Publishing

貶低學者將其研究成果分散在數篇論文中以誇大發表數之行為。

- Other names: salami slicing, segmented publication
- The technical terms are: publishing the least publishable unit (LPU), smallest publishable unit (SPU), minimum publishable unit (MPU), or *publon.*
- when you divide the findings of one study into a series of shorter papers or articles

Most Important Factors

- Novelty
- Significant contributions

good idea, great results, well-written

My first drafts are so-so, but I think I re-write pretty well. Good writing is re-writing. This means you need to start writing the paper early!

The Cockroach and the Puppy with 6 toes

You try, but you can't find a way to kill this paper. While there's nothing too exciting about it, it's pretty well written, the reviews are ok, the results show an incremental improvement. Yet another kind of boring CVPR paper.

A delightful paper, but with some easy-to-point-to flaw. This flaw may not be important, but it makes it easy to kill the paper, and sometimes you have to reject that paper, even though it's so fresh and wonderful.

Quick and Easy Reasons to Reject a Paper

With the task of rejecting at least 75% of the submissions, area chairs are groping for reasons to reject a paper. Here's a summary of reasons that are commonly used:

- Do the authors promise more than they deliver?
- Are there some important references that they don't mention (and therefore they're not up on the state-of-the-art for this problem)?
- Has their main idea been done before by someone else?
- Are the results incremental (too similar to previous work)?
- Are the results believable (too different than previous work)?
- Is the paper poorly written?
- Do they make incorrect statements?

Efros's comments

A number of papers to be published this year, all developed independently, are closely related to our work. The idea of texture transfer based on variations of [6] has been proposed by several authors [9, 1, 11] (in particular, see the elegant paper by Hertzmann et.al. [11] in these proceedings). Liang et.al. [13] propose a real-time patch-based texture synthesis method very similar to ours. The reader is urged to review these works for a more complete picture of the field.

名人金句

Steve Jobs

I'm convinced about half of what separates successful entrepreneurs from the non-successful ones is pure perseverance.

我深信，成功與不成功的企業家之所以不同，有半數原因在於能否堅持下去

perseverance

noun [U] • approving

UK /ˌpɜː.sɪˈvɪə.rᵊns/ US /ˌpɝː.səˈvɪr.ᵊns/

C2

continued effort and determination

不屈不撓，堅持不懈

毅力

Conference Reviewing

If the reviewers misunderstand your paper, or if some flaw in your paper is found, you're dead."

Make it Easy to See the Main Point

- Your paper will get rejected unless you make it very clear, up front, what you think your paper has contributed.
- If you don't explicitly state the problem you're solving, the context of your problem and solution, and how your paper differs (and improves upon) previous work, you're trusting that the reviewers will figure it out.

Make it Easy to See the Main Point

You must make your paper easy to read. You've got to make it easy for anyone to tell what your paper is about, what problem it solves, why the problem is interesting, what is really new in your paper (and what isn't), why it's so neat.

Paper Organization

(1) Start by stating which problem you are addressing, keeping the audience in mind. They must care about it, which means that sometimes you must tell them why they should care about the problem.

(2) Then state briefly what the other solutions are to the problem, and why they aren't satisfactory. If they were satisfactory, you wouldn't need to do the work.

(3) Then explain your own solution, compare it with other solutions, and say why it's better.

(4) At the end, talk about related work where similar techniques and experiments have been used, but applied to a different problem.

What Reviewers Look For

- Again, stating the problem and its context is important.
- But what you want to do here is to state the "implications" of your solution. Sure it's obvious....to you.
- But you run the risk of misunderstanding and rejection if you don't spell it out explicitly in your Introduction.

- The most dangerous mistake you can make when writing your paper is assuming that the reviewer will understand the point of your paper.
- The complaint is often heard that the reviewer did not understand what an author was trying to say.

Write a dynamite introduction