Supplementary Material for: On Exposing the Challenging Long Tail in Future Prediction of Traffic Actors

Osama Makansi Özgün Çiçek Yassine Marrakchi Thomas Brox
University of Freiburg
makansio,cicek,marrakch,brox@cs.uni-freiburg.de

1. Visualization Plots

Figure 1 and 2 show the comparison between our method and different baselines where each circle indicates the performance of one method. These figures illustrate better the improvements gained by our method (dashed arrows).

2. Feature Space Visualization

Figure 3 shows the projection of the feature space using tSNE [7] on three different datasets with different input modalities and views. For each dataset, we show the feature space embedding without our joint optimization (i.e, only the supervised loss) and with our joint optimization (i.e, additionally utilizing the contrastive loss). Note how our approach reshapes the feature space by pushing the challenging scenarios to be closer so that they can benefit each other as also shown in our quantitative results.

3. Effect of the Strength of the Contrastive Loss

In Table 1 we show a study for the importance of the contrastive loss (λ) used in our approach (Eq. (4)). Using a small factor leads to small improvements on the challenging scenarios as the force of reshaping the feature space is rather weak. On the other hand, using a very large factor yields worse results as the network focuses more on reshaping the feature space and ignores the important cues for the actual task which are learned from the supervised loss. Note that this study is used only to show the effect of the weight of the contrastive loss. In our main results, we use the validation set to select the best value for λ.

4. More Qualitative Results

We provide more qualitative results from our approach in Figure 4, Figure 5 and Figure 6 for the ETH-UYC, nuScenes (bird's-eye view) and nuScenes/Waymo (egocentric view) datasets, respectively.

	ETH-UCY (AVG)			
	All	Top 3%	Top 2%	Top 1%
Traj++ EWTA (ours)	**0.16/0.32**	0.47/1.07	0.51/1.13	0.42/0.87
+ contrastive ($\lambda = 20$)	0.17/0.33	0.47/1.04	0.50/1.07	0.43/0.84
+ contrastive ($\lambda = 50$)	**0.16/0.32**	**0.46/1.03**	**0.48/1.03**	**0.38/0.71**
+ contrastive ($\lambda = 100$)	0.17/0.32	0.48/1.04	0.52/1.10	0.50/0.97

Table 1. Study of the hyper-parameter λ on the ETH-UCY dataset. While small λ yields small improvement on the challenging scenarios, large λ yields larger errors on the challenging scenarios.

5. Detailed Quantitative Results

Table 2 show a detailed comparison between our method and the resampling/reweighting baselines across all datasets on all metrics and difficulties. This support our findings that these baselines tend to bias the challenging cases (overfitting) while our approach maintain the average performance and improves largely on the challenging cases.

6. Baselines Implementation Details

In order to use state-of-the-art methods for long-tail classification, we map the regression task to a classification task by assigning classes to training samples based on the error of the Kalman filter. In particular, we group the errors into bins and assign the same class to all samples in each bin. To alleviate the issue of having classes with only one sample, we group all samples with a score greater than a specific threshold into the same bin. This yields 13, 36, 331 classes for ETH-UCY, nuScenes bird's eye view and nuScenes egocentric view, respectively. For all baselines (including our method), we use the same joint training scheme where two heads (classification and regression) are trained on top of the feature embedding. For the LDAM baseline [1], we experiment with different scaling factors and use the best setting $s = 1$. Following BAGS [4], we split the classes into 4 homogeneous groups to ensure that all classes from the same group have roughly the same number of items and use a sampling ration of 8 to ensure that all groups contribute to the mini-batch during training.

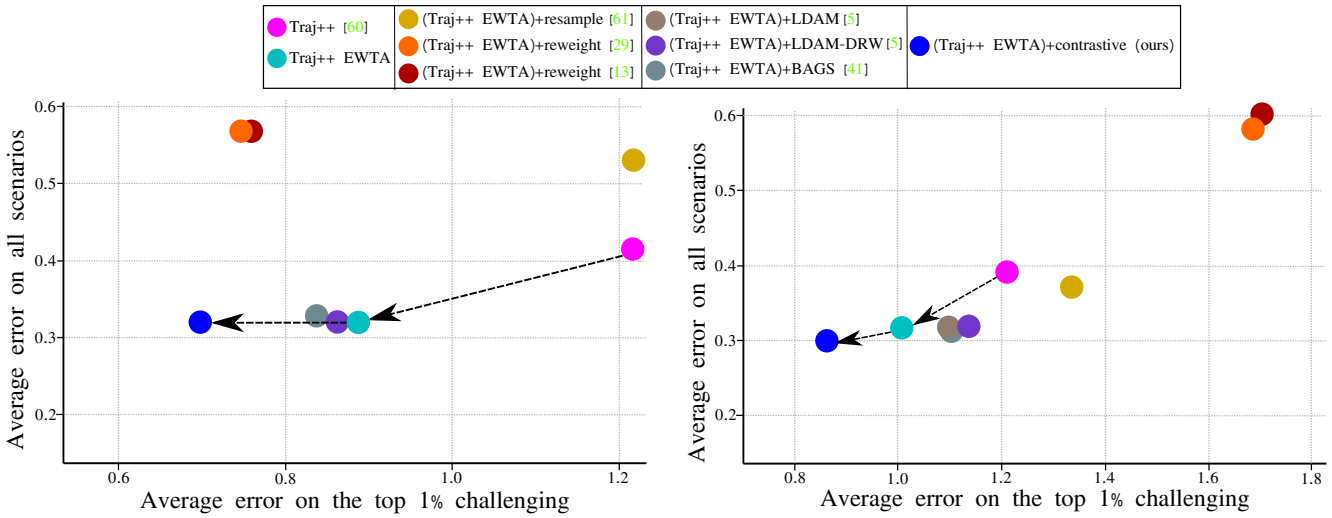


Figure 1. Average vs. Top 1% error comparison on the **ETH-UCY dataset** (left) and the **nuScenes bird's eye view** (right). Our base method of integrating EWTA with the backbone of Trajectron++ (cyan) outperforms the previous state-of-the-art (magenta). Joint learning with the contrastive loss (blue) yields large improvements on the challenging scenarios while not reducing the overall average accuracy. The improvements are indicated by dashed arrows. While the resampling/reweighting baselines also improve on the hard cases, they increase the average error a lot (overfitting). The model-based baselines for long-tailed (LDAM and BAGS) yield only small improvements on ETH-UCY or worse performance on nuScenes bird's eye view.

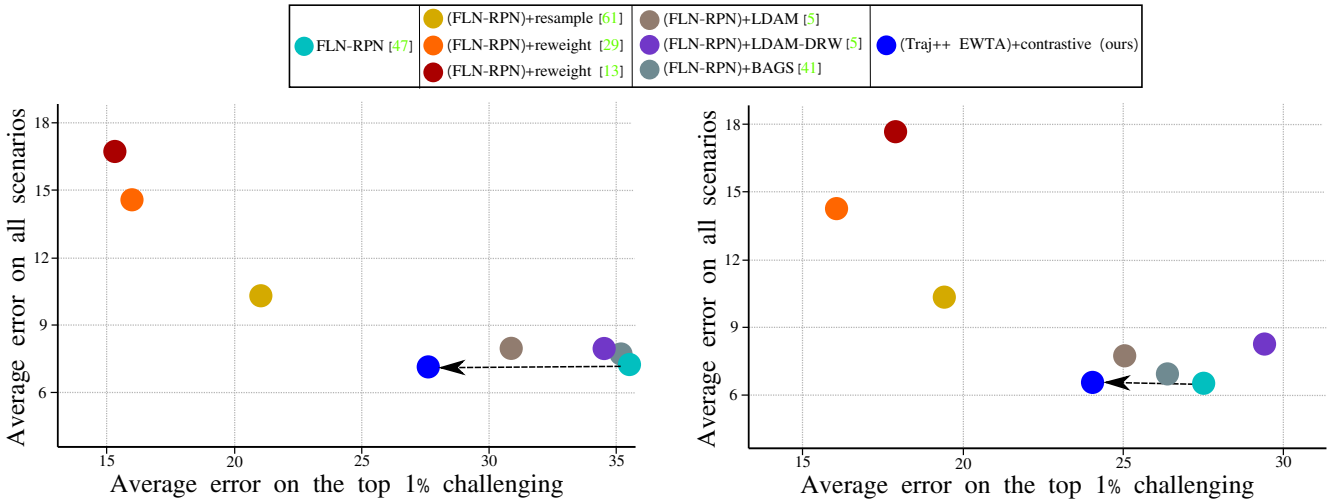


Figure 2. Average vs. Top 1% error comparison on the **nuScenes egocentric view dataset** (left) and the **Waymo open dataset** (right). Our approach utilizing the contrastive loss (blue) yields a significant improvement on the challenging scenarios while not reducing the overall average accuracy. The improvements are indicated by dashed arrows. While the resampling/reweighting baselines also improve on the hard cases, they increase the average error a lot (overfitting). The model-based baselines for long-tailed (LDAM and BAGS) yield smaller improvements than our method.

References

[1] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Aréchiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. In *NeurIPS*, 2019. 1

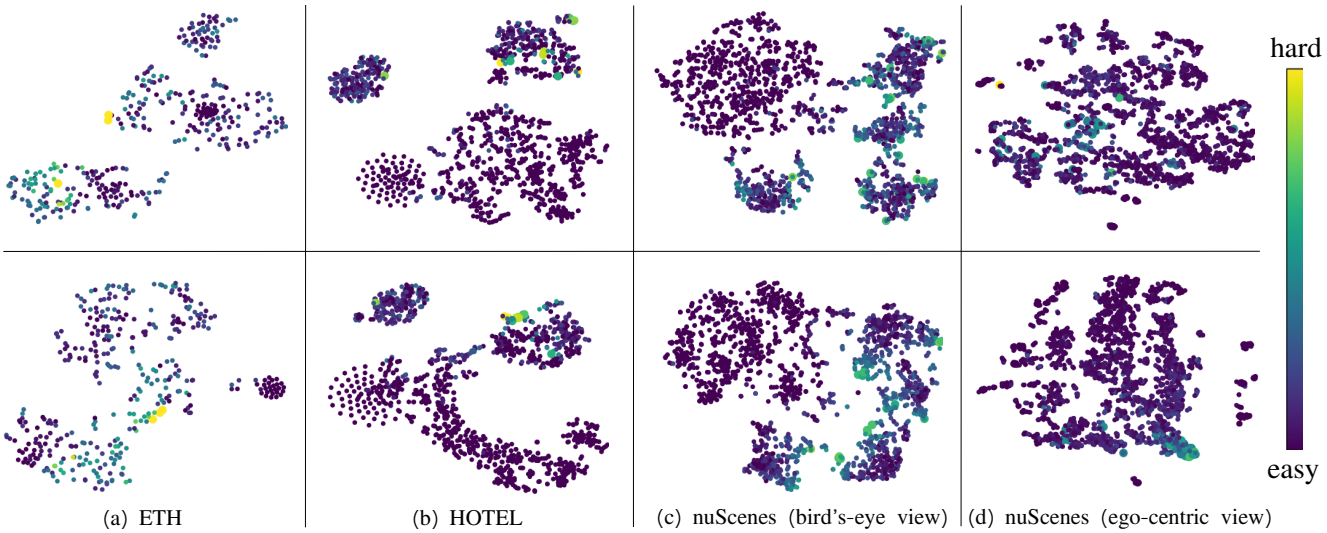
[2] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge J. Belongie. Class-balanced loss based on effective number of samples. In *CVPR*, 2019. 4

[3] Chen Huang, Yining Li, Chen Change Loy, and Xiaoou Tang. Learning deep representation for imbalanced classification. In *CVPR*, 2016. 4

[4] Yu Li, Tao Wang, Bingyi Kang, Sheng Tang, Chunfeng Wang, Jintao Li, and Jiashi Feng. Overcoming classifier imbalance for long-tail object detection with balanced group softmax. In *CVPR*, 2020. 1

[5] O. Makansi, Ö. Çiçek, K. Buchicchio, and T. Brox. Multi-modal future localization and emergence prediction for objects in egocentric view with a reachability prior. In *CVPR*, 2020. 4

[6] Li Shen, Zhouchen Lin, and Qingming Huang. Relay back-



(a) ETH (b) HOTEL (c) nuScenes (bird's-eye view) (d) nuScenes (ego-centric view)

Figure 3. Plot of the feature space using tSNE [7] on three different datasets (a and b are different scenes from the ETH-UCY dataset). **Top.** Training only with the supervised regression loss. **Bottom.** The resulting feature space when trained jointly with the contrastive loss. Large brighte circles indicate the top 1% challenging scenarios. The darker the color of the sample, the easier it is.

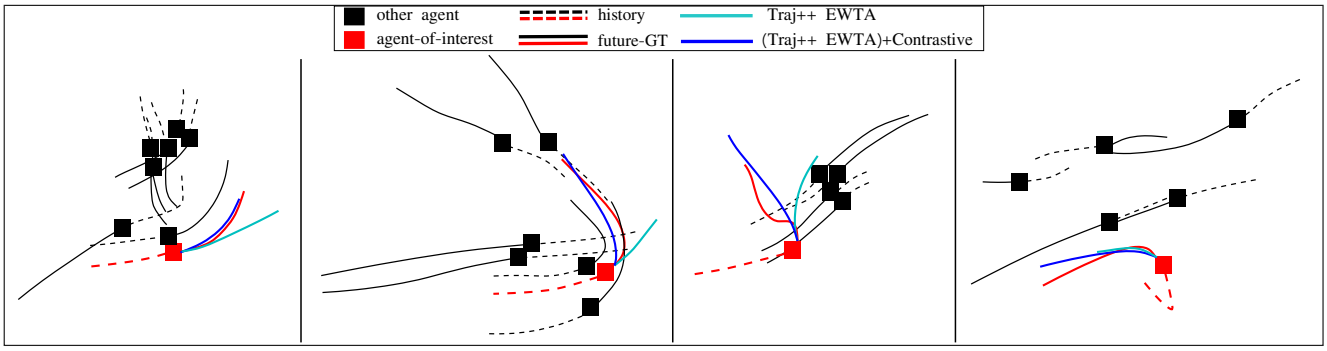


Figure 4. More results from our approach on the ETH-UCY dataset. For all these challenging scenarios, our approach reasons successfully about the social relations to other pedestrians and yields better prediction than the baseline.

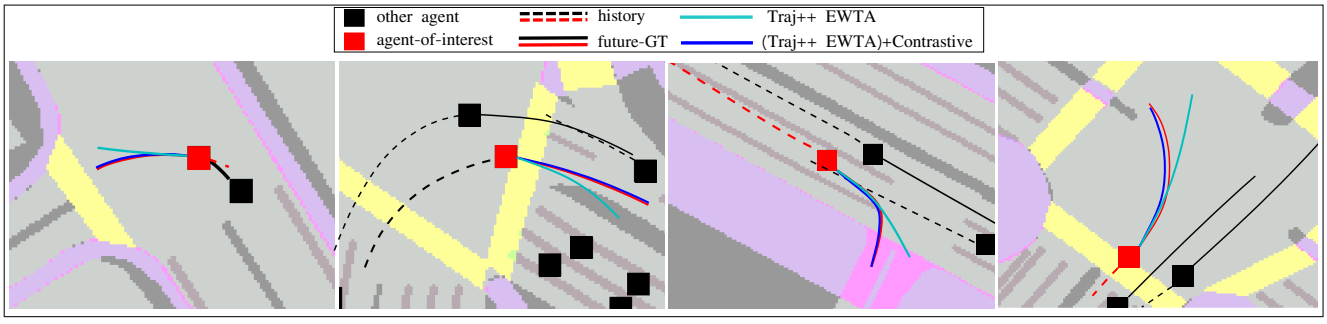
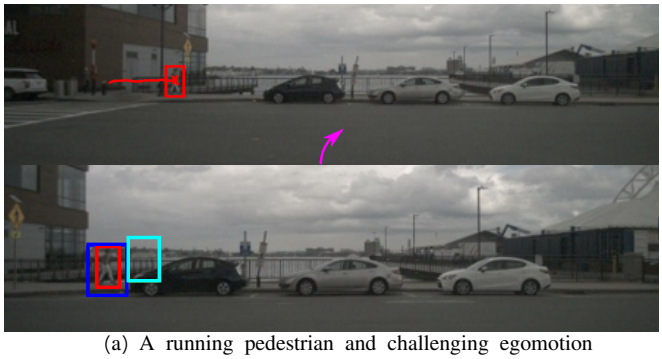


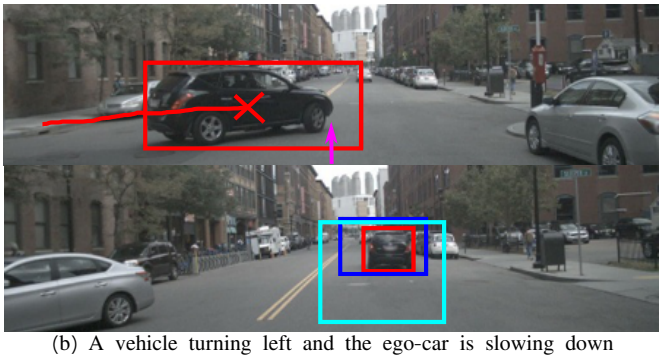
Figure 5. More results from our approach on the nuScenes dataset (bird's-eye view). For all these challenging scenarios, our approach reasons successfully about the semantic cues and predicts the correct trajectory.

propagation for effective learning of deep convolutional neural networks. In *ECCV*, 2016. 4

[7] L.J.P. van der Maaten and G.E. Hinton. Visualizing high-dimensional data using t-sne. *Journal of Machine Learning Research*, 2008. 1, 3

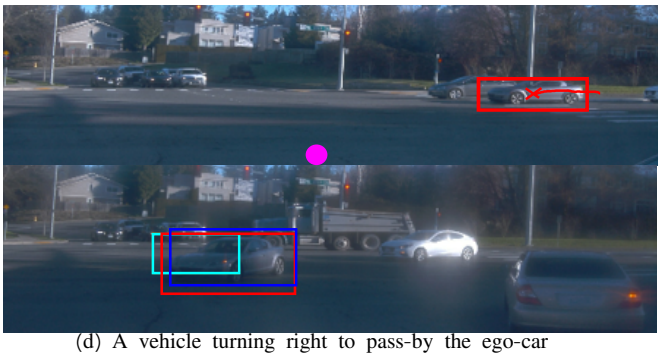


(a) A running pedestrian and challenging egomotion



(b) A vehicle turning left and the ego-car is slowing down

(c) A walking pedestrian and challenging egomotion



(d) A vehicle turning right to pass-by the ego-car

Figure 6. More results from our approach on both egocentric view datasets: nuScenes (a-b) and Waymo (c-d). For each example, we show both the last observed image (top) and the future image (bottom) along with the predictions (FLN-RPN [5] and Ours) and the ground truth. We visualize the best hypothesis for each method. The future egomotion is also shown as arrow indicating the motion of the ego-car.

	ETH-UCY				nuScenes-Bird's Eye View				nuScenes Egocentric View				Waymo Open Dataset			
	All	Top 3%	Top 2%	Top 1%	All	Top 3%	Top 2%	Top 1%	All	Top 3%	Top 2%	Top 1%	All	Top 3%	Top 2%	Top 1%
Baseline	0.16/0.32	0.47/1.07	0.51/1.13	0.42/0.87	0.19/0.32	0.48/0.88	0.50/0.88	0.59/1.02	7.10	29.98	31.13	36.16	6.39	24.87	25.49	27.32
+ resample [6]	0.25/0.53	0.56/1.16	0.61/1.24	0.61/1.22	0.21/0.37	0.55/0.98	0.61/1.07	0.78/1.33	10.20	18.90	19.37	21.62	10.48	19.46	18.91	19.69
+ reweight [3]	0.28/0.56	**0.41/0.78**	**0.44/0.81**	0.43/0.76	0.33/0.58	0.74/1.28	0.80/1.38	0.99/1.67	14.47	15.33	15.42	16.20	14.00	**17.01**	**16.80**	**16.44**
+ reweight [2]	0.28/0.56	0.43/0.83	0.45/0.86	0.44/0.78	0.34/0.60	0.75/1.33	0.80/1.42	0.99/1.71	16.54	**15.29**	**15.34**	**15.46**	17.43	20.34	19.40	18.79
+ contrastive	**0.16/0.32**	0.46/1.03	0.48/1.03	**0.38/0.71**	**0.18/0.30**	**0.44/0.73**	**0.46/0.72**	**0.54/0.85**	**7.04**	25.05	25.26	27.49	6.49	22.36	22.72	24.09

Table 2. Comparison to the common resampling/reweighting techniques on the four datasets. For each method, we show the min-FDE/min-ADE over all samples and over top 1-3% challenging samples. Our method yields large improvements on the challenging ones while maintaining the average. This is in contrast to the reweighting/resampling baselines, which lead to much worse performance on average (see the error increase on the 'All' columns). Baseline indicates Traj++ EWTA for bird's eye view and FLN-RPN [5] for egocentric view.