LaTeX Guidelines for Author Response

1. Introduction

The reviewers generally acknowledge the extensibility of our framework, well evaluated experiments, and the novelty of introducing ergodic state to perform segmentation and recognition within the same framework. Due to space constraints, we address only the major criticisms below.

1.1. R2

Who else has experimented on the SYNTHIA dataset? Or are the authors the first ones? Are the authors releasing a specific publicly-available prediction challenge based on SYNTHIA? Currently, experiments on the SYNTHIA dataset [1, 2] focus on single frame, static images for semantic labeling or scene representation. To our knowledge, our paper is the first attempt to utilize continuous video streams for the prediction problems. are required for the purpose of car trajectory prediction. We collect car trajectory prediction dataset based on the SYN- THIA [29] dataset.

Where are comparisons with other activity forecasting/trajectory prediction works? I think the authors should reimplement SocialLSTM for 2D data (Stanford UAV dataset seems to be a 2D standard prediction task – just throw away images and just use the xy plane coordinates, e.g. center of detected bounding box). SocialLSTM used Avg. disp. Error, Avg. non-linear disp. error , and final disp. Error What was the justification for the metrics the authors chose? In the SocialLSTM papers Related Work section, there are about 20 sources that focus on Activity forecasting The details about the fusion of location and semantic information into the LSTM hidden state seemed missing (was a bit vague). A figure of the architecture of the SEG-LSTM would go a long way in clarifying this (or are the authors using Xu et al.'s "End-to-end Learning of Driving Models from Large-scale Video Datasets" FCN-LSTM from Trevor Darrell's group?). How is this more than just an engineering pipeline, combining tracking (distilling bounding boxes to coordinates), and then adding an LSTM? SocialLSTM already did the second part in principle. I was a bit confused about why the tracking-by-detection framework was novel – how is it different from Tracking-learning-detection (TLD), from Z. Kalal, K. Mikolajczyk, and J. Matas. In TPAMI 2012? Why are scene semantics privileged (which you also call auxiliary) information? If the pipeline relies on fusing the location and semantic stream, if privileged information is missing at test time, then wont the pipeline break? (privileged information is usually present at train time, but absent at test time). Can multiple objects be tracked at once, or just one? The video in the supplementary material only showed one object being tracked. What was the average number of objects in each SYNTHIA scene frame? The focus of the paper seems a bit unclear: apparently, the goal is to predict cars future odometry given previous egomotion visual input. But the paper is about prediction of other vehicles trajectories, not about predicting ones own car odometry (which is different). Odometry: Odometry is the use of motion sensors to determine the robot's change in position relative to some known position (not other robots' positions) https://groups.csail.mit.edu/drl/courses/cs54-2001s/odometry.html Visual Odometry – Estimating the motion of a camera in real time using sequential images (i.e., egomotion) http://www.cs.toronto.edu/~urtasun/courses/CSC2541/ (but we are not trying to estimate the motion of a camera-mounted vehicle here). I felt like the detection/tracking theme slightly bogs down the main focus of prediction in the paper. I think refocusing it on prediction would strengthen the argument. Line 124 - The authors discuss a "lack of proper metrics for evaluating the temporal prediction result." How are the current metrics poorly suited to the task? What would be better?

1.2. Old

After receiving paper reviews, authors may optionally submit a rebuttal to address the reviewers' comments, which will be limited to a **one page** PDF file. Please follow the steps and style guidelines outlined below for submitting your author response.

Note that the author rebuttal is optional and, following similar guidelines to previous CVPR conferences, it is meant to provide you with an opportunity to rebut factual errors or to supply additional information requested by the reviewers. It is NOT intended to add new contributions (theorems, algorithms, experiments) that were not included in the original submission and were not requested by the reviewers. You may optionally add a figure, graph or proof to your rebuttal to better illustrate your answer to the reviewers' comments.

The rebuttal must adhere to the same blind-submission as the original submission and must comply with this rebuttal-formatted template.

Author responses must be no longer than 1 page in length including any references and figures. Overlength responses will simply not be reviewed. This includes responses where the margins and formatting are deemed to have been significantly altered from those laid down by this style guide. Note that this LaTeX guide already sets figure captions and references in a smaller font.

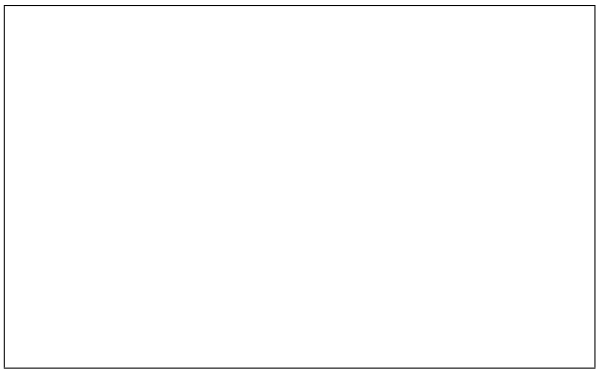

Figure 1. Example of caption. It is set in Roman so that mathematics (always set in Roman: $B \sin A = A \sin B$) may be included without an ugly clash.

2. Formatting your Response

Make sure to update the paper title and paper ID in the appropriate place in the tex file.

All text must be in a two-column format. The total allowable width of the text area is $6\frac{7}{8}$ inches (17.5 cm) wide by $8\frac{7}{8}$ inches (22.54 cm) high. Columns are to be $3\frac{1}{4}$ inches (8.25 cm) wide, with a $\frac{5}{16}$ inch (0.8 cm) space between them. The top margin should begin 1.0 inch (2.54 cm) from the top edge of the page. The bottom margin should be 1-1/8 inches (2.86 cm) from the bottom edge of the page for 8.5×11-inch paper; for A4 paper, approximately 1-5/8 inches (4.13 cm) from the bottom edge of the page.

Please number all of your sections and any displayed equations. It is important for readers to be able to refer to any particular equation.

Wherever Times is specified, Times Roman may also be used. Main text should be in 10-point Times, single-spaced. Section headings should be in 10 or 12 point Times. All paragraphs should be indented 1 pica (approx. 1/6 inch or 0.422 cm). Figure and table captions should be 9-point Roman type as in Figure 1.

2.1. References

List and number all bibliographical references in 9-point Times, single-spaced, at the end of your response. When referenced in the text, enclose the citation number in square brackets, for example [**?**]. Where appropriate, include the name(s) of editors of referenced books.

2.2. Illustrations, graphs, and photographs

All graphics should be centered. Please ensure that any point you wish to make is resolvable in a printed copy of the response. Resize fonts in figures to match the font in the body text, and choose line widths which render effectively in print. Many readers (and reviewers), even of an electronic copy, will choose to print your response in order to read it. You cannot insist that they do otherwise, and therefore must not assume that they can zoom in to see tiny details on a graphic.

When placing figures in LaTeX, it's almost always best to use `\includegraphics`, and to specify the figure width as a multiple of the line width as in the example below

```
\usepackage[dvips]{graphicx} ...
\includegraphics[width=0.8\linewidth]
                {myfile.eps}
```

References

[1] D. Hernandez-Juarez, L. Schneider, A. Espinosa, D. Vazquez, A. M. Lopez, U. Franke, M. Pollefeys, and J. C. Moure. Slanted stixels: Representing san franciscos steepest streets. In *British Machine Vision Conference*, 2017. 1

[2] G. Ros, L. Sellart, J. Materzynska, D. Vazquez, and A. M. Lopez. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 1