Dear Reviewers,

The reviewers of our paper have obviously spent considerable effort in reading our work and pointing out many directions in which the paper can be improved. We are really grateful for this and want to thank all reviewers upfront for their diligence. Following the review comments, we have strived to revise the manuscript based on the comments accordingly. Our revisions are marked in blue in the manuscript to ease reading. In the following, we describe the changes that we did to the paper in response to these comments and suggestions. To make reading easier, we mark the reviewers’ comments in *blue italic*. Our answers follow in plain text. Also, for brevity, we omitted in the text below the positive appreciations in the reviews and also comments which did not point to issues to be improved or changes to be effectuated.

Thanks again for reviewing our paper. We hope everyone stays healthy in this time!

Sincerely,

The Authors

**Summary of main changes**

* We have revised the visual designs, including
  + adopting consistent value ranges for the colormaps across multiple scales in the map view and scatterplot;
  + changing diverging colormap used in the multi-scale attribution view to the sequential colormap of Viridis;
  + changing diverging colormap used in the scatterplot to single-hue sequential colormap; and
  + updating Figs. 1 , 5 , 7, 8, 9, 10, 11, and 12 accordingly.
* We have added more discussions, including
  + limitations of the ST-ResNet model utilizing 2D convolutions, and potential future work on adopting graph convolutions;
  + more insights on opposite conclusions of RMSE and scale-independent metrics; and
  + clarification on how the exerts were intrigued.
* We have proofread the paper and made several corrections, including
  + explaining the term ‘local tract’ earlier;
  + replacing misused term ‘partition’ to ‘region’;
  + correcting the typo ‘structured’ to ‘semi-structured’ interview;
  + adding colormap legends in Fig. 6; and
  + fixing typos pointed out by the reviewers.
* We have updated the video to the latest design.

**Summary Review:**

*issues with the visual design as well as some design choices (R2, R3, R4)*

We have revised the visual design extensively according to the reviewer’s comments. The revisions include:

* adopting the same min and max values in VSUP to make the comparison across scales easier (R2, R3, R4); see Figs. 1, 5, 10, and 11.
* changing diverging colormap used in the multi-scale attribution view to the sequential colormap of Viridis (R2), which is also robust to colorblindness (R4);
* changing diverging colormap used in the scatterplot to single-hue sequential colormap (R4); and
* explaining the rationale for not ordering dots by time in multi-scale attribution view (R3).

*necessity of expansion and clarification on the user study (design study paper) (R2, R4)*

We change the typos of ‘structured interviews’ to ‘semi-structured interviews’, ‘the experts’ to ‘CR’ (R2), and clarify that the bivariate colormap has now been endorsed by all experts (R3).

*required clarification of some key terms/concepts (R2, R3)*

We explain ‘local tract’ in earlier sections of Introduction and Analytical Tasks (R2, R4), fix the inaccurate usage of the term ‘partition’ throughout the paper (R3), and correct many typos (R2).

*vagueness in evaluating expert user feedback (R2, R4)*

We add discussions on how the MAUP affects traffic prediction and expert feedbacks in Sec. 6.4 Expert Review: Applicability (R2, R3), and clarify that all experts approved the adoption of bivariate colormap (R3).

*missing explanations for key steps in data processing (R3) as well as metrics used (R2, R3, R4)*

We add discussions on the necessity of rasterization to fit in ST-ResNet models (R3). We also add more discussions on the insights regarding opposition conclusions between RMSE and scale-independent metrics, and trends of PRMSE, CORR, and uncertainty coefficient (R2, R3, R4).

*too limited discussion limitations of the approach (R3).*

We add discussions of the possible negative effects of rasterization on traffic predictions in the end of Sec. 4.2 Data Preprocessing, and the potential future work of adopting graph convolutions to address the limitation of 2D convolution in Sec. 7 Discussion (R3).

-------

Below we list responses to questions raised by individual reviewer. Detailed answers for the summary reviews can be found accordingly.

**Reviewer 3:**

*Forced coercion of base input areas (TAZ) into regular grids to conform to a uniform input format for the ST-ResNets means TAZ structure is all but obliterated, this fact and its implications are not discussed adequately. Does this make sense from a domain perspective if key information inherent in TAZ layout is removed from model input? Please expand the discussion in either 3.1 or 4.2 (Rasterization).*

Good point. Transportation domain adopts deep learning models to make traffic predictions, and what they care most is to make accurate predictions. Rasterization of TAZ partitions is a necessary step to fit in the inputs of ST-ResNet. One can easily compute the predicted flow volumes for each partition from predicted flow volumes of grid partitions by ST-ResNet. Nevertheless, as the reviewer points out, the rasterization operation obliterates the neighborhood relationships of TAZ partitions, which may cause negative effects on traffic predictions. This is indeed a limitation of deep traffic prediction models with 2D convolutions.

Following the suggestion, we have added a discussion on this point in the end of Sec. 4.2. Moreover, we have discussed potential usage of graph convolutional networks (GCN) to overcome the limitation of 2D convolution operations; see discussion of future work in Sec. 7 Discussion.

*As the paper aims to be an application/design study (and I concur), both use cases and user feedback should receive at least some more room. Several rather basic formalisms/formulae could be removed to facilitate this.*

Thanks for the suggestions. We have 1) added more discussions on the applicability and future work derived from the studies in Sec. 6.4 Expert Review and Sec. 7 Discussion, and 2) removed the formulae of Equation 1 in the previous submission.

*As a general remark, please review the use of specific terms, e.g., "single partition" vs. "entire region" in section 3.2 / T.3. Doesn't each individual partition always cover the entire region of Shenzhen? The same goes for "each partition" vs. "local tract" (vs. single grid cell?) in section 5.3.*

We actually meant ‘each region’ derived after partition, instead of ‘single partition’ and ‘each partition’. ‘Local tract’ indicates surrounding regions of a region under investigation. We have corrected inaccurate usage of the terms throughout the paper, and explained ‘local tract’ in the Introduction and Sec. 3.2 Analytical Tasks sections.

*Section 5.2 argues for the use of bivariate colormaps over juxtaposed map views. Given that bivariate color scales are not trivial to interpretet, was the concluding statement of 5.2 endorsed by the two domain expert users, and did they test both design variants?*

The design was only endorsed by our collaborating expert in the development stage. Recently, we also showed the comparison of the bivariate colormaps and side-by-side map views to the independent experts, and they also provided similar feedbacks. We made the clarification in Sec. 6.4 Expert Review.

*The Multi-scale attribution view arranges partitions according to decreasing prediction error, which is also encoded in the dot color. Using this layout, the view does not retain temporal ordering while dual-encoding the selected error metric. I would imagine that ordering by time would not only be more intuitive, it would also better facilitate detecting outlier time periods for which models struggle. In the current design, this requires selection in the map view. Please explain the rationale for this design decision.*

First, the multiscale attribution view arranges the dots according to *absolute* prediction errors, while color codes the dots according to scale-independent metrics, which are related to, but not necessarily the same with, absolute prediction errors. In fact, the experiment reveals that absolute prediction error and scale independent metrics are very different. For instance, in Fig. 1, the dots on the right side indicate their absolute prediction errors are high, but their colors are green indicating the uncertainty coefficients are low. Our visual analytics can help domain experts identify outliers, e.g., region 2 in Fig. 1, and help investigate potential reasons behind it. We now emphasized that sorting is based on absolute prediction errors.

Second, we do not order the dots by time simply because there are too many time slots (7 days x 48 slots/day = 336 slots) for each partition. The view will be cluttered if we present all time slots. Instead, we adopt ‘overview + details’ design rationale to facilitate exploration of temporal variations. Users can select a partition to view its temporal variations in all three views of the map, the scatterplot, and the multiscale attribution view.

**Reviewer 4:**

*First, the ranges of both observed flow and prediction error dimensions vary upon multiple scales and shapes of partition on the bivariate map. The color scale of flow volume has to fit its map so that each map has its own range. But inconsistency in the scale of prediction error weakens the function of value-suppressing uncertainty palette.*

We have fixed value ranges for both flow volumes and prediction errors in maps of all three scales. The color scales are consistent now. Please check revised map views in Figs. 1, 5, 6, 10 and 11.

*Besides, the standardized prediction errors in Moran’s I scatter plot are encoded into symmetric diverging color map but the distribution is much denser above zero. The diverging color scale is also utilized to present scale-independent metrics. It is not suitable for all metrics since except the correlation coefficient the other two are non-zero. In quantitative analysis, users would be easily misled under such color scales.*

Good suggestion. We have made the following changes accordingly.

* We have changed colormap in the scatter plot to a single-hue sequential colormap. Since colors in the scatter plot indicate prediction errors, we adopt the same hue of orange as that for prediction errors in the map view.
* We have changed colormap in the multi-scale attribution view to sequential colormap of Viridis, which is a perceptually-uniform color scheme widely used in many js, python, and R libraries. Note that colors in the multi-scale attribution view indicate scale-independent metrics, which are different with prediction errors used in the map view. We would like to emphasize the difference. Hence we choose viridis colormap that has minimal overlapping with colors in VSUP.

Please refer to Figs. 1, 7, 8, 9, 10, 11, and 12 for the revised visual effects.

*Second, the paper employs different metrics to measure traffic flow prediction but it does not provide enough insights to explain the opposite conclusions on different measurements. In case studies, the author does not describe the connection among the trends of* *PRMSE, uncertainty coefficient, and CORR, either. The author needs to add more evaluation on the metrics instead of simply listing them in multiple views.*

A possible reason for the opposite conclusions derived from RMSE and scale-independent metrics is because prediction errors are linearly correlated with flow volumes, and scale-independent metrics can cope with the correlations but not RMSE. The correlation can be observed in the Moran’s I scatterplot; see Fig. 7 for an example. Based on the finding, a promising future work is to improve the prediction performance using attention mechanism that emphasizes regions with high flow volumes.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 25×50 | | | 50×100 | | | 100×200 | | |
| PRMSE | CORR | U | PRMSE | CORR | U | PRMSE | CORR | U |
| Grid | 13.39 | 0.99 | 0.08 | 16.46 | 0.97 | 0.12 | 28.25 | 0.89 | 0.30 |
| TAZ | 7.64 | 0.82 | 0.32 | 5.71 | 0.91 | 0.22 | 4.84 | 0.95 | 0.16 |

Table 1: Average values of PRMSE, correlation coefficient (CORR), and uncertainty coefficient (U) under different partition shapes and scales.

We also compute the average values for PRMSE, CORR, and U under different partition shapes and scales; see the table above. From the table, PRMSE and uncertainty increase while correlation drops, indicating prediction performance drops, when scaling up for Grid partition. In contrast, PRMSE and uncertainty decrease while correlation increases, indicating prediction performance improves, when scaling up for TAZ partition. This is probably because spatial heterogeneity decreases when scaling up for TAZ partition, while that increases when scaling up for Grid partition, and deep learning models can better predict traffic of regions with low spatial heterogeneity. The insight here is to generate network inputs with low spatial heterogeneity, which may be derived by using spatial clustering based on human activities, or gradual partition.

We have added more discussions on the metrics in Sec. 7 Discussion. The table above shows average metric values, which however is in contrary to our statement of using unit visualization. Hence we decide to not put it in the paper.

*In the spatial association analysis, there needs an introduction about the traffic flow volume in local tract.*

Done. We have described ‘local tract’ earlier in the Introduction and Sec. 3.2 Analytical Tasks sections.

**Reviewer 1:**

It’s great to know the reviewer likes our work. We would like to express our great appreciation to the reviewer’s nice comments.

**Reviewer 2:**

*I was surprised that they were intrigued to find out that the MAUP affected deep traffic prediction. Intrigued in what way? Were they not aware of this problem?*

Thanks for pointing this out. All experts (including our collaborating expert) expressed their concerns on how do researchers in deep learning determine network input size, e.g., 32x32 for the whole Beijing city used in ST-ResNet. The experts would like to know if the MAUP affects predictions, and if so, how the MAUP affects predictions. Our study clearly show that the MAUP does affect prediction results, and more importantly, the study reveals how the MAUP affects deep traffic prediction. For instance, the Moran’s I scatter plot shows that prediction accuracy is more dependent on flow volumes in each partition, but not local tract. The comparison of railway station and airport under TAZ and Grid partition also proves the finding.

We have clarified the misunderstanding in Sec. 6.4 Expert Review: Applicability. Please check the revision.

*… but at the beginning of section 3.2, they said they “conducted several rounds of structured interviews with the experts.” Were there other experts besides the independent expert reviewers that were not mentioned or was there a typo in the word experts (it should be expert)? Why have you decided to use structured interviews, when usually, in this case, visualization researchers use semistructured interviews? Aren’t structured interviews probably too rigid in this situation?*

We apologize for the carelessness. These are typos. We conducted semi-structured interviews only with the collaborating researcher. We have fixed the typos.

*Only in section 5.3, when the Moran’s I is defined, is that the reader can infer that the local tract is the partitions surrounding the given partition (3x3 matrix). Is that correct? I think it needs to be clarified when task T.2 is presented, at least briefly.*

Yes. Local tract refers to the partitions surrounding the given partition (3x3 matrix). We now introduce the term in both Introduction and Sec. 3.2 Analytical Tasks.

*I noticed in the bivariate map that the min and max values of the colorscale vary across the partitions’ scales. Why haven’t you used the same min and max values for all the variations? Couldn’t this cause misinterpretation when making comparisons if the user is not paying attention to the color legend? I would assume that the color intensities represent the same values in different partitions. Not using a common scale made the comparisons more difficult in the case studies.*

Thank you for your rigorous suggestion. We have modified the color scale to the same min and max values for all scales; please see Figs. 1, 5, 10, and 11.

*Provide the color legend for the bivariate map in Figure 6.*

Fixed.

*In the multiscale attribution view, can you provide a justification for why using a green-red colormap instead of a color-blind safe palette?*

We have changed the color scheme to Viridis colormap, which is a perceptually-uniform sequential colormap adopted by many js, python, and R libraries. According to the description (<https://cran.r-project.org/web/packages/viridis/vignettes/intro-to-viridis.html>), Viridis colormap is robust to colorblindness.

*Review for typos (see attached marked-up pdf).*

Thank you for the careful reviews. We have fixed the typos and proofread the manuscript again.

End of response letter