Supplementary Material: GitNet: Geometric Prior-based Transformation for Birds-Eye-View Segmentation

Shi Gong$^{\star 1}$, Xiaoqing Ye$^{\star 2}$, Xiao Tan^{2}, Jingdong Wang^{2}, Errui Ding^{2}, Yu Zhou$^{\star\star 1}$, and Xiang Bai^{1}

[1] Huazhong University of Science and Technology
{gongshi,yuzhou}@hust.edu.cn
[2] Baidu Inc., China

1 Ray-based Transformer

1.1 Detailed architecture

The detailed description of the Ray-based Transformer adopted in GitNet, with positional encodings passed at each attention layer, is given in Fig. 1.1. The perspective image features $\boldsymbol{F_i^j}$, *i.e.*, the $\boldsymbol{j}$-th column of the $\boldsymbol{i}$-th level of pyramid features, are passed through the transformer encoder (Column Context Augment, CCA), together with perspective positional encoding that are added to queries and keys at every multi-head self-attention layer. Then, the decoder (Ray-based Cross-Attention, RCA) receives queries, that are initialized as the pre-alined features $\boldsymbol{S_i^j}$, along with the BEV positional encoding, and the output of encoder $\widetilde{\boldsymbol{F}}_{\boldsymbol{i}}^{\boldsymbol{j}}$, along with the perspective positional encoding, and produces the refined features $\widetilde{\boldsymbol{S}}_{\boldsymbol{i}}^{\boldsymbol{j}}$ through multi-head cross-attention.

1.2 Positional Encoding

2D positional encoding As the transformer is unable to distinguish the position of elements, we add positional encoding to the *Keys* and *Queries* following [1, 2]. The 2D positional encoding map have the same shape with the input features, that denotes as $P \in \mathbb{R}^{H\times W\times d}$, where d is the channel number, H and W are the height and width of input features, at horizontal and vertical direction, respectively. We encode horizontal position in the first half of d channels, and the vertical positions in the second half channels. Suppose the $u \in [0, W)$ and $v \in [H)$ denotes the row and column index, then the horizontal positional encoding at the point of $(u, v)$is:

$$\begin{aligned} P_{\mathrm{h}}(u, v, 2i) &= \sin(u/10000^{4i/d}) \\ P_{\mathrm{h}}(u, v, 2i+1) &= \cos(u/10000^{4i/d}) \end{aligned} \tag{1}$$

[⋆] Authors contribute equally.
[⋆⋆] Corresponding author.

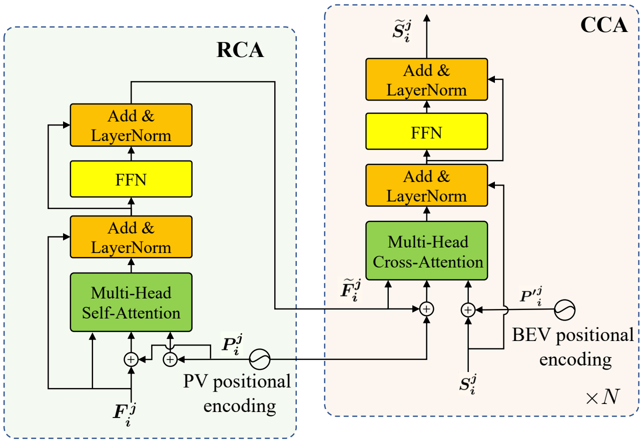

Fig. 1. Architecture of Ray-based Transformer.

The vertical positional encoding is:

$$\begin{aligned} P_{\mathrm{v}}(u, v, 2i + d/2) &= \sin(v/10000^{4i/d}) \\ P_{\mathrm{v}}(u, v, 2i + 1 + d/2) &= \cos(v/10000^{4i/d}) \end{aligned} \tag{2}$$

Where $i \in [0, d)$ is the channel index. We concatenate the positional encoding at two directions to get the whole 2D positional encoding, that is $P = \mathrm{Cat}(P_{\mathrm{h}}, P_{\mathrm{v}})$.
PV and BEV positional encoding Both our perspective-view (PV) and birds-eye-view (BEV) positional encoding follow the paradigm of 2D positional encoding detailed above. The only difference is that PV positional encoding has the same spatial size with image features, while the BEV positional encoding has spatial size of rasterized BEV map.

1.3 What does the cross-attention see?

To explore how the cross-attention module works in our framework, we visualized the attention maps of a representative sample, as shown in Fig. 1.2. As our ray-based transformer computes the cross-attention between every query point in the BEV and the corresponding column of the perspective image features, a lateral query line cross all columns/rays (white query line in Fig. 1.2) produces the attention maps of the full perspective image features. We depict the cross-attention map from three different decoder layers, which go deeper from left to right. In the first row, the queries lie on the pedestrian crossing, 10 meters away from the camera, and in the right three columns, we can observe that the corresponding cross-attention maps mainly focus on the pedestrian crossing region of the perspective space. When the queries line move farther,20 meters away, the attention maps focus on upper regions of the perspective images. Since

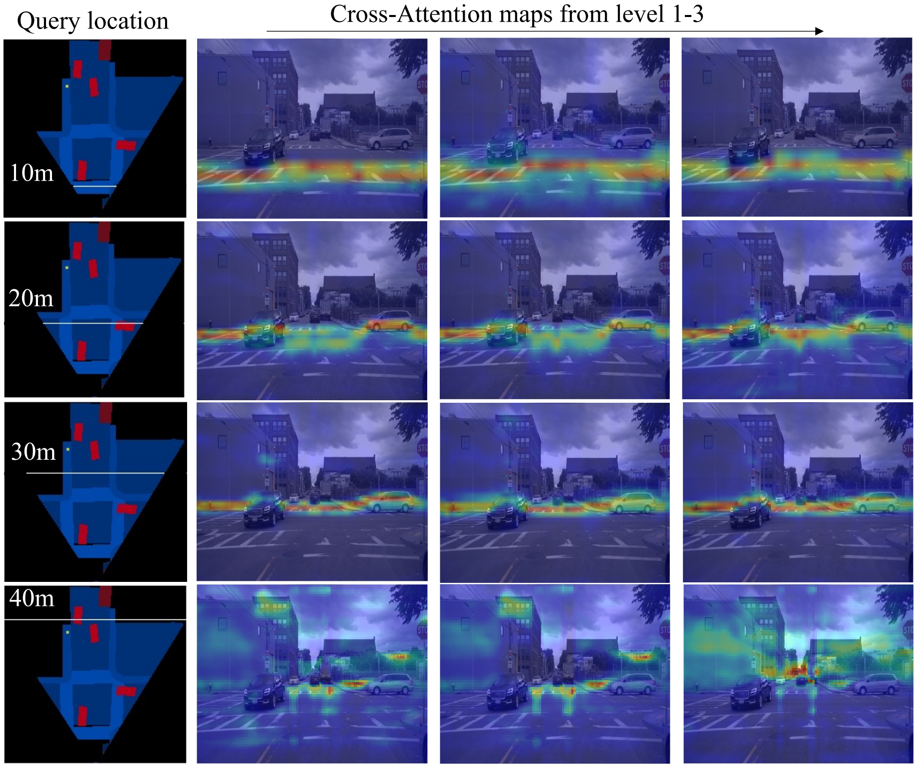

Fig. 2. Visualization of cross-attention map from three levels of decoder layers. In the first column, the white line marked in BEV semantic labels denotes the queries location at 10, 20, 30 and 40 meters away from the camera. The three columns on the right denote the cross-attention map from the level-1 to level-3 of the transformer decoder layers. For intuitive comparison, we superimpose the input RGB images onto the cross-attention maps.

our pre-alignment module provides visibility-aware pre-aligned features to the transformer, the invisible regions can be further refined by aggregating contextual information from other visible regions. This can be supported by the observation that the attention maps of invisible regions tend to disperse over an extensive region, while attentions maps of visible ground mostly focus a certain point. In the fourth row, the query line is 40 meters away, and the attention maps are scattered in the invisible regions that occluded by the building and cars.

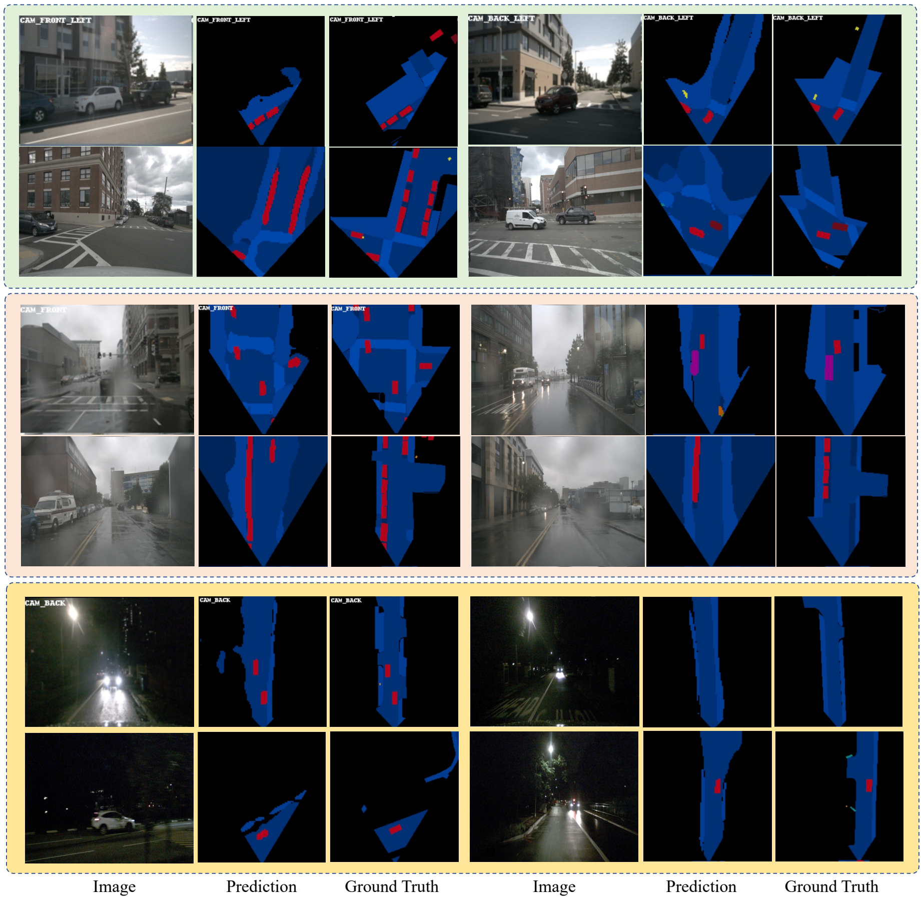

Fig. 3. Visualization of samples under three conditions including light, rainy and dark.
.

2 Robustness under Different Conditions

In Fig. 1.3, we depict additional visualization results by selecting samples from the validation set of nuScenes including three different weather conditions: light, rainy and dark. For the purpose of practical use, our model must be able to handle these various conditions. From the Fig. 1.3, our model can perform well in the light conditions (the first group), and under the more challenging rainy condition (the second group), our model can segment almost all other cars and the complete layout of the crossroads. Under the dark condition (the third group), our model also succeed to segment the forwarding cars, and right sidewalk which cannot be seen clearly even for human.

References

1. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European Conference on Computer Vision(ECCV). pp. 213–229 (2020)
2. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in Neural Information Processing Systems(NIPS) **30** (2017)