ChiTransformer: Towards Reliable Stereo from Cues

Supplementary Material

Qing Su
Georgia State University
qsu3@gsu.edu

Shihao Ji
Georgia State University
sji@gsu.edu

1. Depth Constraint of Fisheye Re-projection

The spherical projection model of fisheye lens is non-linear. To facilitate the re-projection process without incurring extra non-linearity, we let the model predict planar distance instead of the depth in z-axis direction. In the prediction map, vertical features of the same planar distance in one roll angle direction should reside in the same longitudinal curve determined by η in the camera coordinate system.

The fisheye sequence we used for training is synthesized with equisolid projection model shown in Figure 1. A fisheye camera is modeled as a spherical lens with radius f as its focal length, θ is the azimuth angle of the projection line $\overline{\boldsymbol{qC}}$ between object point $\boldsymbol{q}$ and camera center $\boldsymbol{C}$. The chord $\overline{\hat{\boldsymbol{q}}\boldsymbol{Z}_0}$ of θ is non-linearly projected on to the image plane with its length ρ preserved. Given the spherical coordinates (f, θ, ϕ) of point $\hat{\boldsymbol{q}}$, the coordinates of image point $\boldsymbol{q}_i$ of $\boldsymbol{q}$ are calculated as:

$$\rho = 2f \sin\frac{\theta}{2}, \tag{1}$$

$$\boldsymbol{q}_i = \begin{pmatrix} x \\ y \end{pmatrix}_i = \rho \begin{pmatrix} \cos\phi \\ \sin\phi \end{pmatrix}. \tag{2}$$

In the re-projection process, we project the pixel coordinates p_t of target image I_t to the coordinates p'_t in source image I'_t for sampling. For simplicity, we use the ideal intrinsic parameters (unit focal length, perfect image centers) and unit length horizontal translation for the synthetic sequence. The coordinates re-projection constrained by planar distance d_p is calculated through

$$(x, y)_i \rightarrow (\eta, \gamma) \tag{3}$$

$$\hat{\boldsymbol{q}} = \begin{pmatrix} \hat{x} \\ \hat{y} \\ \hat{z} \end{pmatrix} = f \begin{pmatrix} \sin\gamma\cos\eta \\ \cos\gamma \\ \sin\gamma\sin\eta \end{pmatrix} \tag{4}$$

$$\boldsymbol{q} = \frac{d_p}{\sin\gamma}\hat{\boldsymbol{q}} = d_p f \begin{pmatrix} \cos\eta \\ \cot\gamma \\ \sin\eta \end{pmatrix} \tag{5}$$

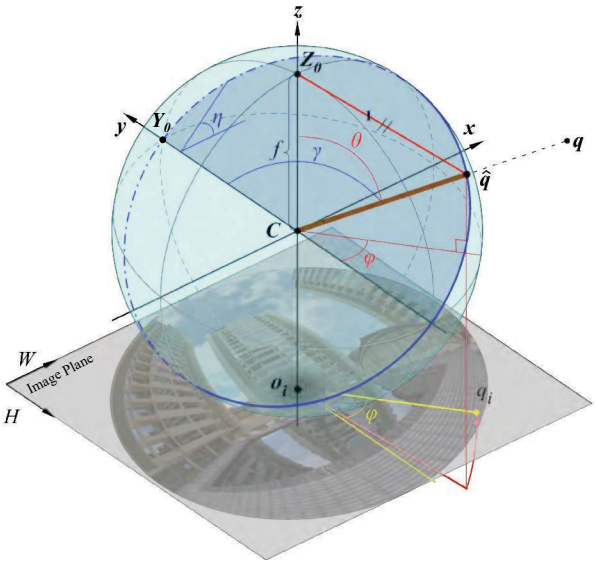

Figure 1. Equisolid Projection Model. Fisheye lens is modeled as a sphere with radius f. $\boldsymbol{Z}_0$ is the zenith of the sphere. Object point $\boldsymbol{q}$ is projected onto the sphere surface at $\hat{\boldsymbol{q}}$, whose spherical coordinates are (f, θ, ϕ). The chord $\overline{\boldsymbol{Z}_0\hat{\boldsymbol{q}}}$ of the azimuth angle θ is projected onto the image plane with its length preserved as shown in the image. The endpoint of the projected chord is then the image point $\boldsymbol{q}_i$ for $\boldsymbol{q}$. To facilitate the distance estimation, we also define the spherical coordinates (f, γ, η) with $\boldsymbol{Y}_0$ as the zenith, where γ is the azimuth angle and η is the roll angle from x-axis.

$$\boldsymbol{q}' = T\boldsymbol{q} \tag{6}$$

$$\hat{\boldsymbol{q}}' = f\frac{\boldsymbol{q}'}{\|\boldsymbol{q}'\|} \tag{7}$$

$$\hat{\boldsymbol{q}}' \sim (f, \theta, \phi)' \xrightarrow{(1),(2)} \begin{pmatrix} x \\ y \end{pmatrix}'_i \tag{8}$$

$$f = 1, \tag{9}$$

where we first construct a coordinate conversion map from image coordinates to (η, γ) coordinates. η is the roll an-

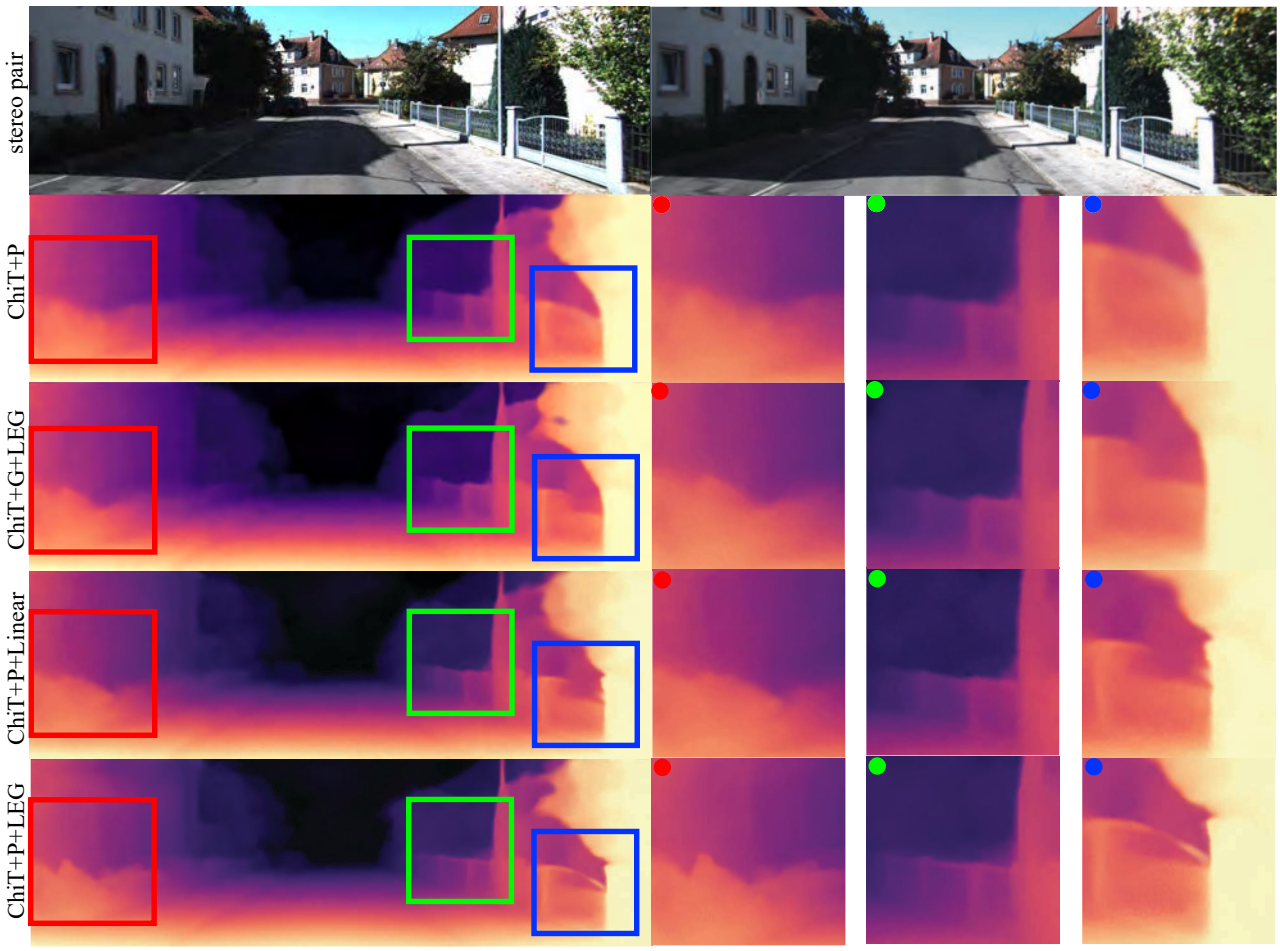

Figure 2. Qualitative results for ablation study.

gle about the y-axis and γ is the azimuth angle from the direction of y. The coordinates of point $\boldsymbol{q}$ in the target camera coordinate system is then recovered through (4) and (5), and subsequently converted to the coordinates of $\boldsymbol{q}'$ in the source camera coordinate system through relative pose transformation T in (6). The source image coordinates are then calculated through (7) and (8). The singular points at poles are replaced with the values of their vicinal pixels.

2. Ablation Study

We trained ChiTransformer-8 with different configurations to show the effectiveness of the designed features in the model. Qualitative results are shown in Figure 2, where "P" denotes the polarized attention, "G" stands for the direct learning of self-adjoint matrix G, "LEG" represents the feature of learnable epipolar geometry, and "Linear" is the single line attention zone.

3. Choice of Intermediary Variable

For rectilinear stereo, since the relationship between disparity D and depth d (i.e. $d = g(D)$) is linear, either D or d can be used as intermediary variable. However, for non-rectilinear stereo, such as unwarped fisheye stereo, $g(\cdot)$ is highly nonlinear and disparity does not necessarily take the form of 2D displacement (e.g., planar disparity). This necessitates extra nonlinear disparity-location conversion which does not benefit much the self-supervision in terms of complexity. Meanwhile, we desire direct depth output for our model. Thus, we choose depth as intermediary variable for both rectilinear and non-rectilinear stereos.

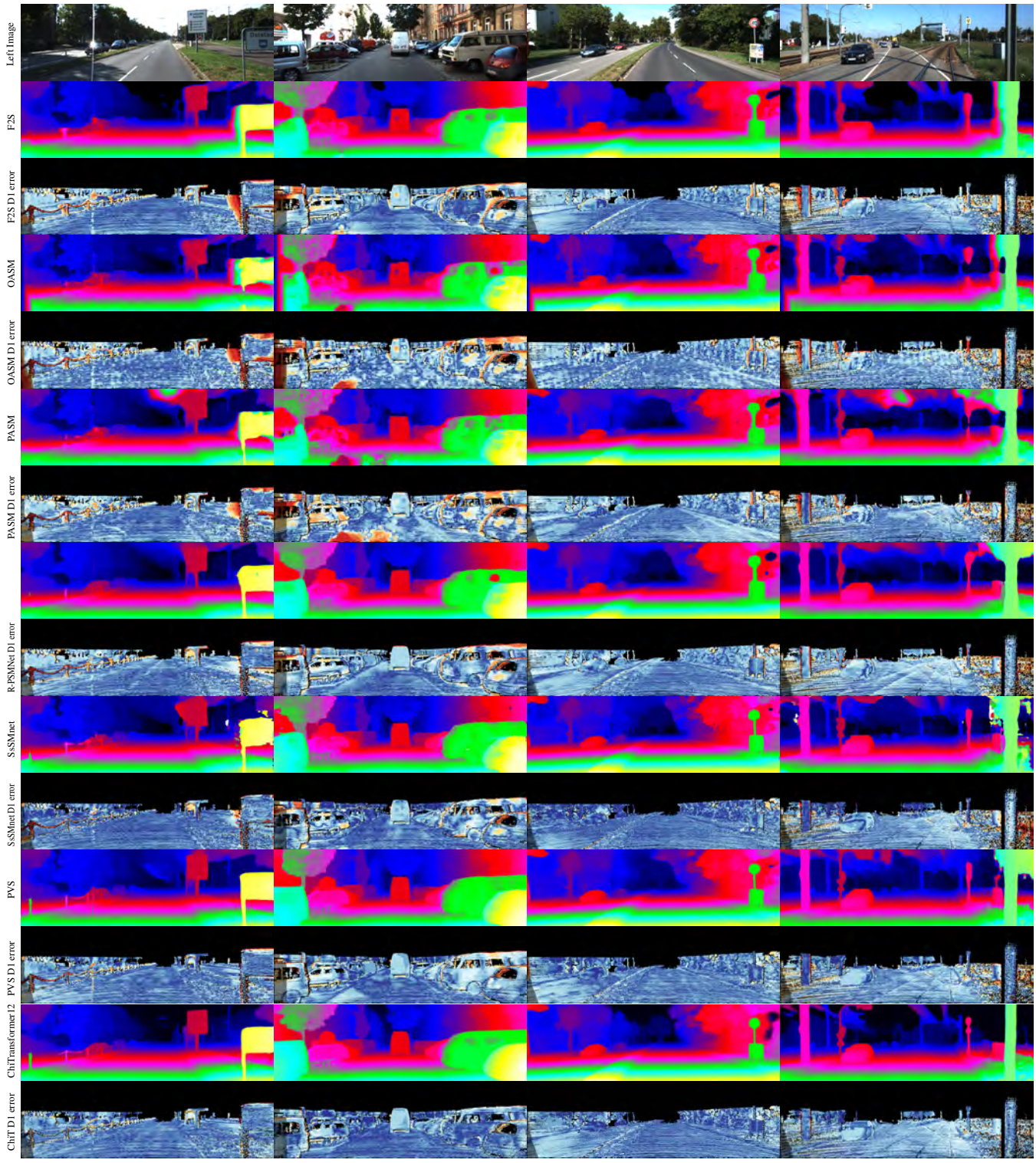

Figure 3. Qualitative results on the KITTI stereo 2015 compared with existing self-supervised stereo matching methods. The sharp depth maps generated by our model (ChiTransformer-12 in the second last row) provides more reliable estimation especially in the close range as reflected in the error map and better global coherence consistent.