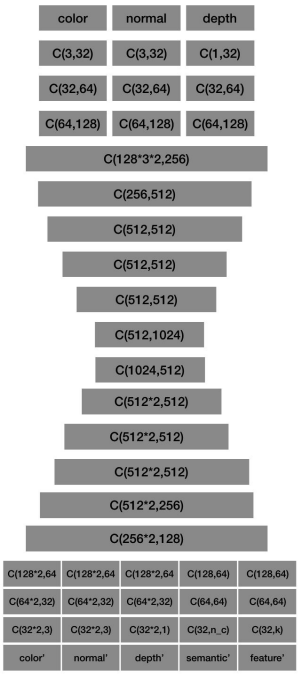

Figure 7: Completion network architecture. C(m,n) stands for m input channel and m output channel. Skip connections are added at mirroring location of the encoder and decoder network. Two sets of input(corresponding to source and transformed target scans respectively) first go through the first three layers separately, then being concatenated and pass through the rest layers.

A. More Technical Details about Our Approach

A.1. Completion Network Architecture

The completion network takes two sets of RGB-D-N (RGB, depth, and normal) as input. Three separate layers of convolution (followed by ReLU and Batchnorm) are applied to extract domain specific signal before merging. Those three preprocessing-branches are applied to both sets of RGB-D-N input. We also use skip layer to facilitate training. The overall architecture is listed as follows, where C(m,n) specify the convolution layer input/output channel.

A.2. Iteratively Reweighted Least Squares for Solving the Robust Regression Problem

In this section, we provide technical details on solving the following robust regression problem:

$$R^\star, \boldsymbol{t}^\star = \underset{R,\boldsymbol{t}}{\operatorname{argmin}} \sum_{c=(q_1,q_2)} a_c \big(\|R\boldsymbol{p}(q_1) + \boldsymbol{t} - \boldsymbol{p}(q_2)\|^2 + \|R\boldsymbol{n}(q_1) - \boldsymbol{n}(q_2)\|^2 \big)^\alpha \quad (10)$$

,where we use $\alpha = 1$ in all of our experiments. We solve (10) using reweighted non-linear least squares. Introduce an initial weight $w_c^{(0)} = a_c, c \in \mathcal{C}$. At each iteration $k \geq 0$, we first solve the following non-linear least squares:

$$\min_{R,\boldsymbol{t}} \sum_{c=(q_1,q_2)\in\mathcal{C}} w_c^{(k)} \big(\|R\boldsymbol{p}(q_1) + \boldsymbol{t} - \boldsymbol{p}(q_2)\|^2 + \|R\boldsymbol{n}(q_1) - \boldsymbol{n}(q_2)\|^2 \big). \quad (11)$$

According to [14], (10) admits a closed-form solution. Specifically, define

$$\boldsymbol{c}^{(k)}(Q_1) := \frac{\sum_{c=(q_1,q_2)\in\mathcal{C}} w_c^{(k)} \boldsymbol{p}(q_1)}{\sum_{c=(q_1,q_2)\in\mathcal{C}} w_c^{(k)}},$$

$$\boldsymbol{c}^{(k)}(Q_2) := \frac{\sum_{c=(q_1,q_2)\in\mathcal{C}} w_c^{(k)} \boldsymbol{p}(q_2)}{\sum_{c=(q_1,q_2)\in\mathcal{C}} w_c^{(k)}}.$$

The optimal translation and rotation to (11) are given by

$$\boldsymbol{t}^\star = \boldsymbol{c}^{(k)}(Q_2) - R^\star \cdot \boldsymbol{c}^{(k)}(Q_1),\ R^\star = U \mathrm{diag}(1, 1, \mathrm{sign}(M)) V^T,$$

where U and V are given by the singular value decomposition of

$$M = U\Sigma V^T = \sum_{(q_1,q_2)\in\mathcal{C}} w_c^{(k)} \big(\overline{\boldsymbol{p}}(q_1)\overline{\boldsymbol{p}}(q_1)^T + \overline{\boldsymbol{n}}(q_1)\overline{\boldsymbol{n}}(q_1)^T \big),$$

and where

$$\overline{\boldsymbol{p}}(q_1) = \boldsymbol{p}(q_1) - \boldsymbol{c}^{(k)}(Q_1),\ \overline{\boldsymbol{p}}(q_2) = \boldsymbol{p}(q_2) - \boldsymbol{c}^{(k)}(Q_2).$$

After obtaining the new optimal transformation $R^\star, \boldsymbol{t}^\star$, we update the weight $w_c^{(k+1)}$ associated with correspondence c at iteration $k+1$ as $w_c^{(k+1)} :=$

$$\frac{1}{(\epsilon^2 + \|R\boldsymbol{p}(q_1) + \boldsymbol{t} - \boldsymbol{p}(q_2)\|^2 + \|R\boldsymbol{n}(q_1) - \boldsymbol{n}(q_2)\|^2)^{2-\alpha}}$$

where ϵ is a small constant to address the issue of division by zero.

In our experiments, we used 5 reweighting operations for solving (10).

A.3. Implementation Details

Implementation details of the completion network. We used a combination of 5 source of information(color,normal,depth,semantic label,feature) to supervise the completion network. Specifically, we use

$$loss_{recon} = \lambda_c loss_c + \lambda_n loss_n + \lambda_d loss_d + \lambda_s loss_s + \lambda_f loss_f$$

, where we use l_1 loss for color, normal, depth, l_2 loss for feature, and cross-entropy loss for semantic label. We use $\lambda_c, \lambda_n, \lambda_d, \lambda_f = 1$, $\lambda_s = 0.1$. We trained for 100k iterations using a single GTX 1080Ti. We use Adam optimizer with initial learning rate 0.0002.

	SUNCG				Matterport				ScanNet			
	Rotation		Trans.		Rotation		Trans.		Rotation		Trans.	
	Median	Mean	Median	Mean	Median	Mean	Median	Mean	Median	Mean	Median	Mean
nr	4.51	26.25	0.21	0.62	4.85	22.33	0.22	0.60	12.90	33.89	0.36	0.61
r	1.54	23.36	0.10	0.54	2.51	18.69	0.10	0.49	7.11	**30.40**	0.23	**0.57**
sm	2.65	25.6	0.18	0.64	3.15	20.23	0.20	0.60	7.10	35.32	0.17	**0.57**
r+sm	**1.32**	**19.36**	**0.06**	**0.48**	**1.45**	**13.9**	**0.04**	**0.34**	**5.47**	32.38	**0.12**	**0.57**

Table 2: Ablation study for pairwise matching. nr: Directly apply the closed-form solution [14] without reweighted procedure. r: reweighted least square, sm: spectral method, r+sm: alternate between reweighted least square and spectral method.

B. Additional Experimental Results

Figure 8, 9, 10 show more qualitative results on SUNCG, Matterport, and ScanNet, respectively. Table 2 gives a detailed ablation study of our proposed pairwise matching algorithm. We compare against three variants, namely, direct regression(nr) using [14], reweighted least squares(r) (using the robust norm), and merely using spectral matching (sm). We can see that the combination of reweighted least squares and spectral matching gives the best result.

We also applied the idea of learning weights for correspondence from data [36]. Since [36] addresses a different problem of estimating the functional matrix between a pair of RGB images, we tried applying the idea on top of reweighted least squares (r) of our approach, namely, by replacing the reweighting scheme described in Section A.2 by a small network for predicting the correspondence weight. However, we find this approach generalized poorly on testing data. In contrast, we found that the spectral matching approach, which leverages geometric constraints that are specifically designed for matching 3D data, leads to additional boost in performance.

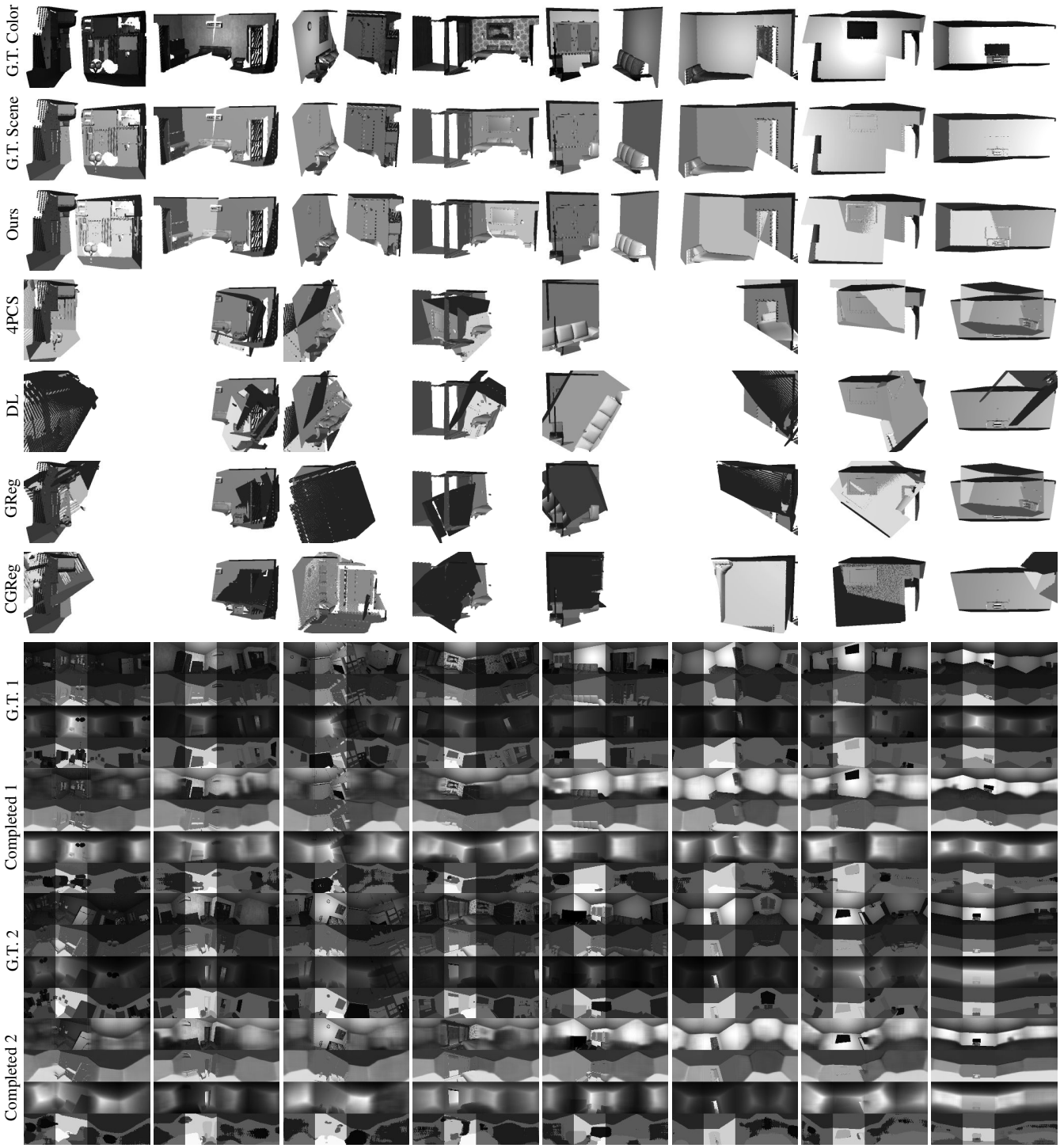

Figure 8: SUNCG qualitative results. From top to bottom: ground-truth color and scene geometry, our pose estimation results (two input scans in red and green), baseline results (4PCS, DL, GReg and CGReg), ground-truth scene RGBDNS and completed scene RGBDNS for two input scans. The unobserved regions are dimmed.

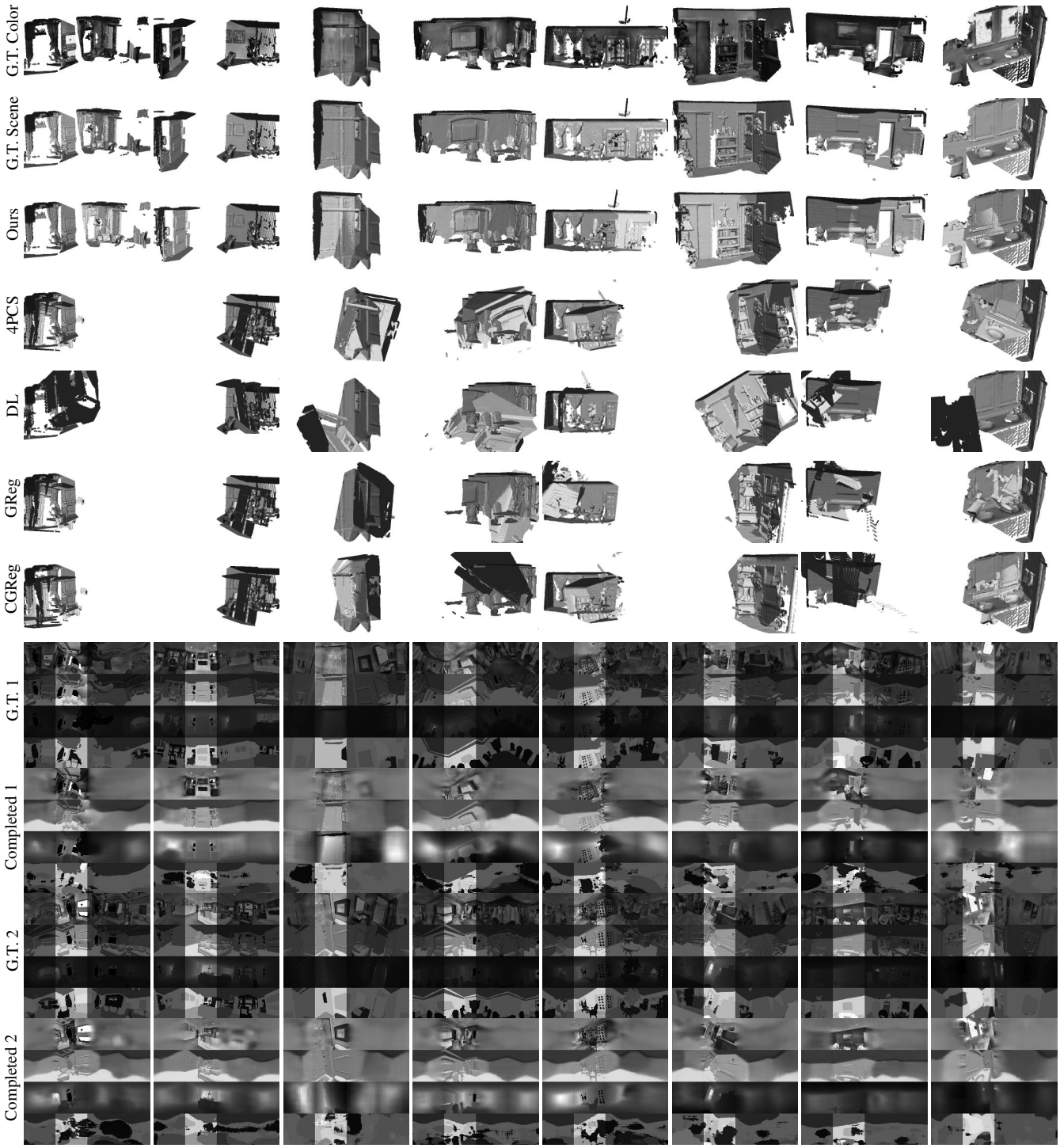

Figure 9: Matterport qualitative results. From top to bottom: ground-truth color and scene geometry, our pose estimation results (two input scans in red and green), baseline results (4PCS, DL, GReg and CGReg), ground-truth scene RGBDNS and completed scene RGBDNS for two input scans. The unobserved regions are dimmed.

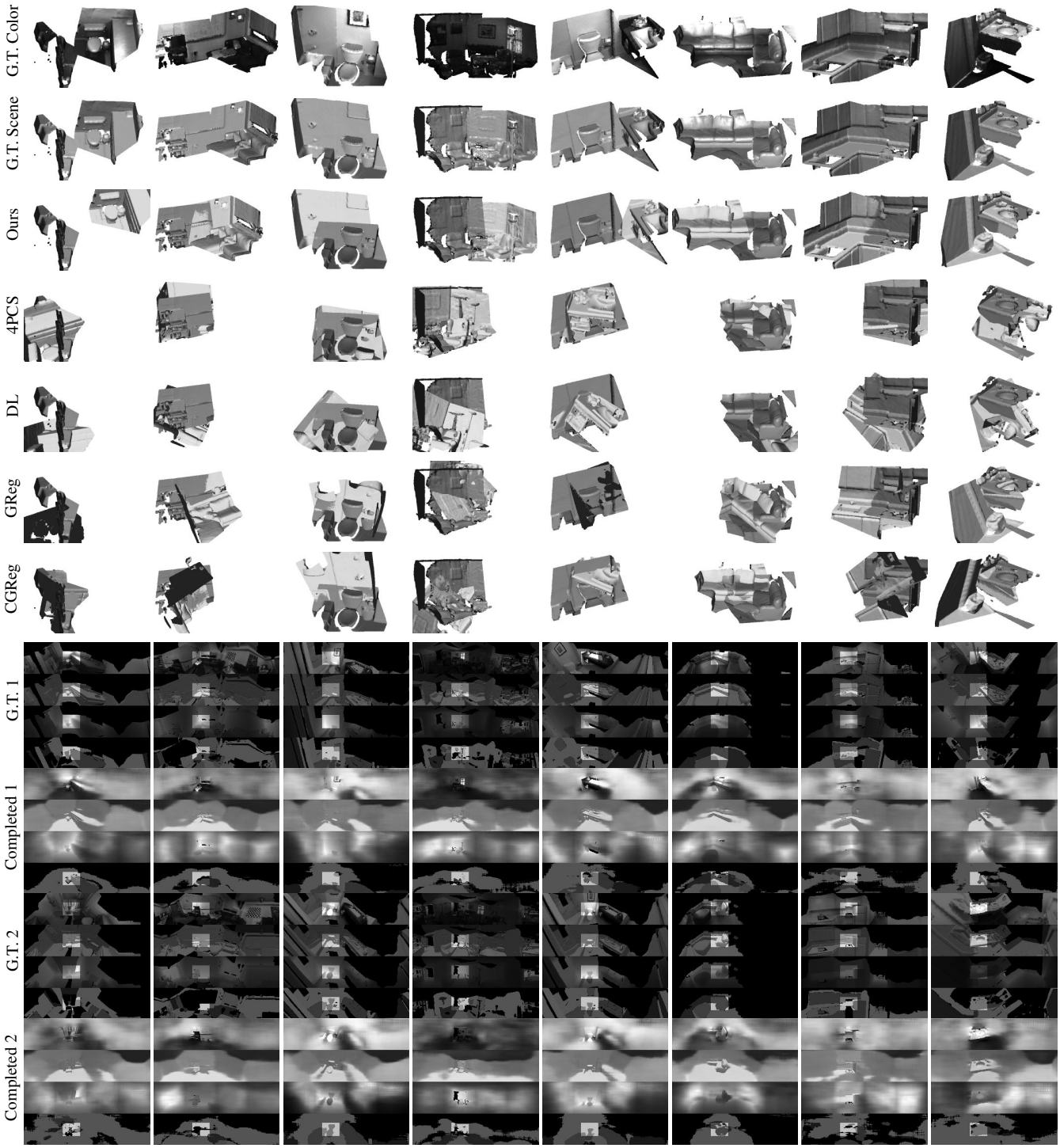

Figure 10: ScanNet qualitative results. From top to bottom: ground-truth color and scene geometry, our pose estimation results (two input scans in red and green), baseline results (4PCS, DL, GReg and CGReg), ground-truth scene RGBDNS and completed scene RGBDNS for two input scans. The unobserved regions are dimmed.