Re-Depth Anything: Test-Time Depth Refinement via Self-Supervised Re-lighting

Supplementary Material

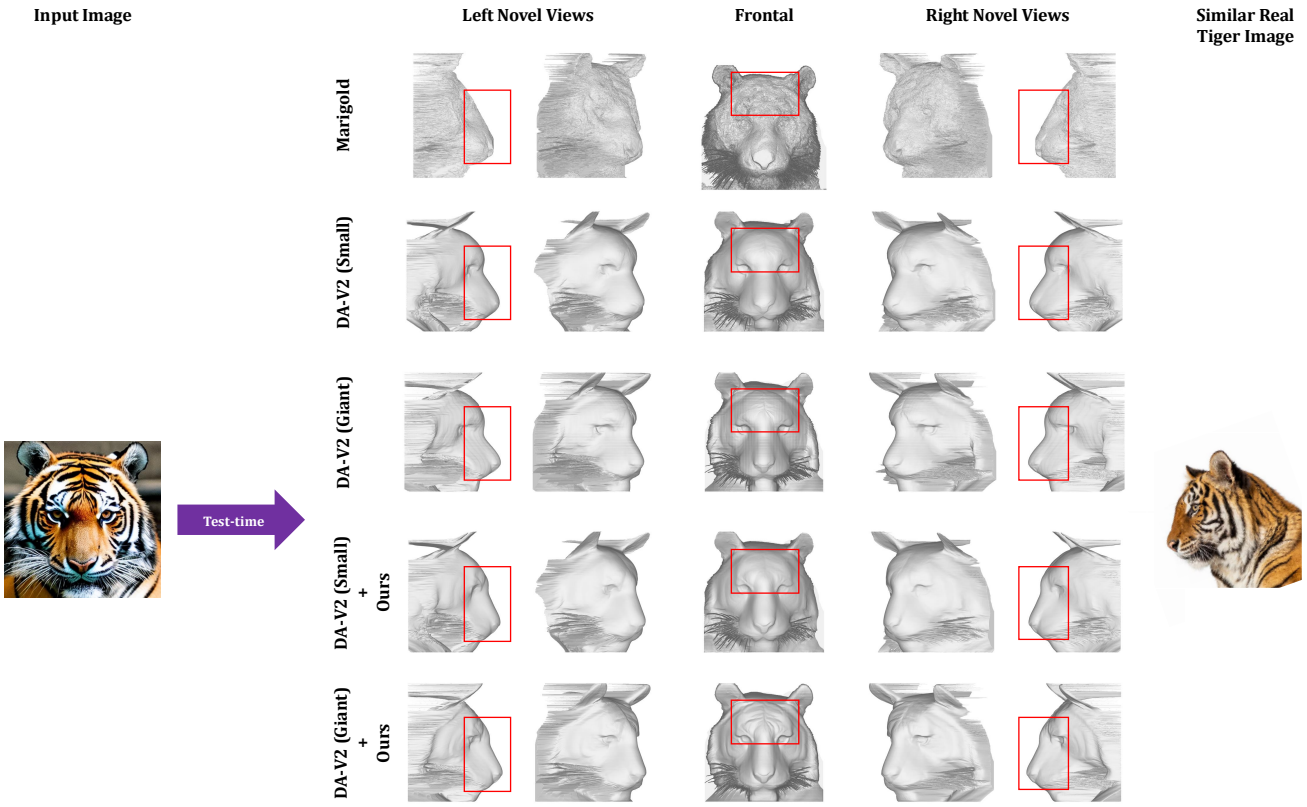

Figure 7. **Biased predictions by baselines and our correction via re-lighting. Top:** Marigold and the larger DA-V2 variants struggle with the tiger image in ways similar to the small variant shown in the main paper teaser. **Bottom:** Our method applies to both the large and small variants of DA-V2, correcting the overall shape in each case and adding more detail when using the giant variant.

This supplemental document provides additional qualitative and quantitative comparisons, justifies the choice of DA-V2 as the baseline, and explains the challenge of comparing state-of-the-art depth estimation models due to mismatched depth representations and evaluation protocols. Some references refer to the main paper.

7. Additional Comparisons

Fig. 7 exemplifies how different depth estimation methods struggle with the tiger image from the main paper. The reconstructed dog-like shapes indicate a bias in the training data rather than an issue with the model architecture or depth representation. Our re-lighting is agnostic to training and corrects the bias from a dog-like to a tiger-like shape, regardless of whether the giant or small variant of DA-V2 is used. Quantitative improvements over other methods are listed in Table 4 and discussed further below.

We present additional qualitative results in Figs. 9, 10, 11, 12, and 13. These figures also illustrate several remaining limitations of our method, highlighted by red boxes and discussed in the main document. Fig. 8 further shows the per-sample change in SqRel relative to DA-V2, highlighting that our method delivers more frequent and larger-magnitude improvements across all three datasets.

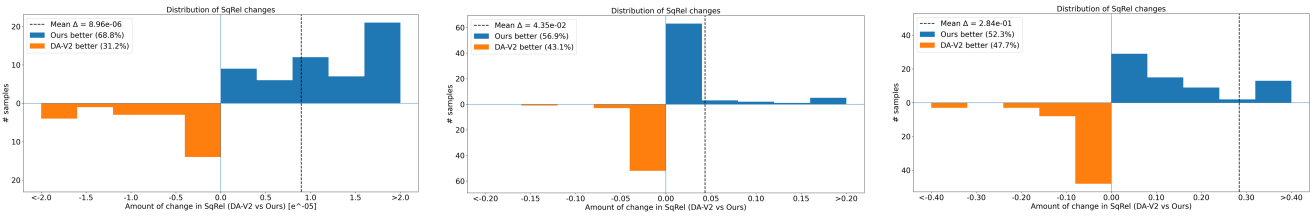

Figure 8. Histograms of per-sample change in SqRel with respect to DA-V2 on samples from CO3D, ETH3D, and KITTI. The x-axis shows $\Delta\text{SqRel} = \text{SqRel}_{\text{DA-V2}} - \text{SqRel}_{\text{Ours}}$ (for CO3D, scaled by 10^{-5} as indicated on the axis): blue bars on the right correspond to samples where our method achieves lower SqRel than DA-V2, and orange bars on the left correspond to samples where DA-V2 is better. The dashed vertical line marks the mean ΔSqRel for each dataset. On CO3D, the distribution is clearly skewed toward positive changes, with relatively few and smaller-magnitude failures. ETH3D and KITTI both exhibit a heavier positive tail. Overall, our method improves SqRel on more samples than it degrades, and the gains are larger in magnitude than the occasional losses, leading to a positive mean ΔSqRel on all three datasets.

8. Results with the DA-V2 Giant Backbone

We used the ViT-S backbone for the main experiments because it performs almost as well as the larger ones while being significantly smaller (in terms of embedding size) and therefore more efficient to optimize. We also found that the giant model with the ViT-G encoder suffers from similar biases, such as the dog-like prediction for the tiger image in Fig. 7. Therefore, we used the smaller model for development and for the major comparisons.

Table 4. **Quantitative comparison with state-of-the-art monocular depth estimation methods.** DA-V2 and DA3 outperform the other baselines on CO3D and ETH3D, respectively, and our method further improves their performance. Therefore, we view our method as a self-supervised refinement strategy applicable to various base models. We compared different alignment metrics (see Sec. 10) and highlighted the comparable ones in gray and black, respectively. We advocate using the protocol marked in solid black.

Dataset	Method	Prediction			*Higher is better* ↑			*Lower is better* ↓					
		Rel. Disp.	Rel. Depth	Abs. Depth	δ_1	δ_2	δ_3	AbsRel	RMSE	log10	RMSE log	SI log	SqRel
CO3D	Marigold$_{\text{ls-depth}}$		✓		1.0	1.0	1.0	0.00276	0.0685	0.001200	0.00366	0.366	0.000320
	Marigold$_{\text{ls-depth-disp}}$		✓		1.0	1.0	1.0	0.00275	0.0685	0.001196	0.00366	0.366	0.000320
	DepthPro$_{\text{ls-depth}}$			✓	1.0	1.0	1.0	0.00242	0.0625	0.001053	0.00334	0.334	0.000286
	DepthPro$_{\text{ls-depth-disp}}$			✓	1.0	1.0	1.0	0.00241	0.0625	0.001046	0.00334	0.334	0.000286
	DA-V2$_{\text{ls-disp-depth}}$	✓			**1.0**	**1.0**	**1.0**	0.00227	0.0602	0.000985	0.00321	0.321	0.000244
	DA-V2 + Ours$_{\text{ls-disp-depth}}$	✓			**1.0**	**1.0**	**1.0**	**0.00223**	**0.0588**	**0.000968**	**0.00314**	**0.314**	**0.000235**
	DA3$_{\text{ls-depth}}$		✓		**1.0**	**1.0**	**1.0**	0.00251	0.0667	0.00108	0.00357	0.357	0.000317
	DA3 + Ours$_{\text{ls-depth}}$		✓		**1.0**	**1.0**	**1.0**	**0.00238**	**0.0637**	**0.00103**	**0.00341**	**0.341**	**0.000294**
	DA-V2$_{\text{ls-disp}}$	✓			**1.0**	**1.0**	**1.0**	0.00226	0.0602	0.000981	0.00321	0.321	0.000244
	DA-V2 + Ours$_{\text{ls-disp}}$	✓			**1.0**	**1.0**	**1.0**	**0.00222**	**0.0588**	**0.000964**	**0.00314**	**0.314**	**0.000235**
KITTI	Marigold$_{\text{ls-depth}}$		✓		0.889	0.978	0.992	0.109	3.86	0.047	0.162	16.1	0.53
	Marigold$_{\text{ls-depth-disp}}$		✓		0.876	0.953	0.980	0.110	5.63	0.051	0.181	17.9	0.84
	DepthPro$_{\text{ls-depth}}$			✓	0.937	0.987	0.995	0.086	2.74	0.037	0.132	13.0	0.30
	DepthPro$_{\text{ls-depth-disp}}$			✓	0.896	0.963	0.983	0.096	6.65	0.045	0.186	18.3	3.21
	DA-V2$_{\text{ls-disp-depth}}$	✓			0.568	0.796	0.902	0.305	7.01	0.118	0.348	33.6	2.49
	DA-V2 + Ours$_{\text{ls-disp-depth}}$	✓			**0.593**	**0.818**	**0.917**	**0.283**	**6.71**	**0.110**	**0.319**	**30.7**	**2.20**
	DA-V2$_{\text{ls-disp}}$	✓			0.818	0.937	0.974	0.323	130	0.062	0.256	25.4	1756
	DA-V2 + Ours$_{\text{ls-disp}}$	✓			**0.823**	**0.940**	**0.976**	**0.276**	**105**	**0.060**	**0.243**	**24.1**	**1335**
ETH3D	Marigold$_{\text{ls-depth}}$		✓		0.963	0.994	0.998	0.058	0.476	0.0255	0.101	10.0	0.083
	Marigold$_{\text{ls-depth-disp}}$		✓		0.952	0.988	0.995	0.072	4.666	0.0318	0.178	17.7	30.25
	DepthPro$_{\text{ls-depth}}$			✓	0.966	0.993	0.997	0.058	0.498	0.0237	0.077	7.69	0.158
	DepthPro$_{\text{ls-depth-disp}}$			✓	0.962	0.990	0.994	0.065	5.489	0.0299	0.206	20.5	19.9
	DA-V2$_{\text{ls-disp-depth}}$	✓			0.884	0.956	0.978	0.113	0.955	0.0448	0.153	15.1	0.391
	DA-V2 + Ours$_{\text{ls-disp-depth}}$	✓			**0.898**	**0.965**	**0.982**	**0.104**	**0.875**	**0.0413**	**0.143**	**14.1**	**0.347**
	DA3$_{\text{ls-depth}}$		✓		**0.983**	**0.998**	**0.999**	0.0461	0.373	0.0194	0.0631	6.27	0.100
	DA3 + Ours$_{\text{ls-depth}}$		✓		**0.983**	0.997	**0.999**	**0.0458**	**0.370**	**0.0193**	**0.0628**	**6.24**	**0.099**
	DA-V2$_{\text{ls-disp}}$	✓			**0.968**	0.991	0.995	0.198	35.69	**0.0253**	0.0998	9.93	1339
	DA-V2 + Ours$_{\text{ls-disp}}$	✓			**0.968**	**0.992**	**0.996**	**0.148**	**23.27**	**0.0253**	**0.0941**	**9.36**	**850.9**

At the time the paper was first written, the DA-V2 Giant weights were still available on Hugging Face via the following link, even though they had been removed from the official GitHub repository.

9. Results with the DA3 Mono Large

We evaluated the effectiveness of our refinement strategy on the state-of-the-art Depth Anything 3 (DA3) [21] (DA3MONO-LARGE), which predicts depth instead of disparity and employs a ViT-L encoder. Our method improves upon DA3 across both the CO3D and ETH3D datasets (see Table 4), setting a new state-of-the-art. Qualitative comparisons in Fig. 11 further show that our approach significantly enhances details. Note that while the model predicts depth, we visualize the disparity maps for better contrast and to maintain visual consistency with other figures in this paper.

For DA3 experiments, we use a scaled orthographic camera, with its scale factor optimized from an initial value of 7.0. Additionally, we set a learning rate of 2×10^{-4} for the embeddings and 1×10^{-6} for the DPT weights. The regularization weight λ_1 is set to 1.0 for ETH3D's indoor/outdoor scenes and 10.0 for CO3D's close-up objects, respectively. Unless explicitly noted, all other hyperparameters are identical to those used in our DA-V2 experiments. Normal MSE in the main document is calculated from spatial depth gradients over pixels with valid neighboring depth values.

10. Evaluation Protocol Details

Since monocular depth estimation is inherently scale-ambiguous, various depth representations and corresponding evaluation protocols have been proposed to support different invariances. Unfortunately, even among relative depth estimation methods, there are discrepancies in the evaluation protocols used.

The classical approach is to align the relative depth prediction with the ground-truth depth in a least-squares sense before applying a range of metrics, which may include log transformations to reduce the impact of uncertainties. The benefit of aligning to the ground truth is that it provides interpretable results in metric space that are comparable to methods using absolute depth prediction.

DA-V2 deviated from this path by predicting disparity instead of relative depth. Hence, it was evaluated directly in disparity space using the same alignment procedure and metrics. However, this makes it incomparable to models that evaluate on depth. An established alternative is to perform alignment in disparity space and then apply the metrics after converting disparity back to depth. However, least-squares fitting in disparity handles outliers very differently than when computed in depth. For instance, in disparity space, far estimates are less pronounced, leading to an unfair comparison with methods that perform alignment directly in depth space.

To provide an as-fair-as-possible comparison consistent with widely used evaluation protocols, we followed the procedure described in the main document: we first align to obtain absolute disparity and then perform a second alignment to minimize least-squares errors in the same space used by methods that output relative depth. We refer to this as least-squares disparity-and-depth (ls-disp-depth) alignment. To shed light on the effect of the different alignment methods, we compare them in Table 10. The least-squares alignment performed directly in disparity space followed by depth conversion (ls-disp), as used in [12], performs better on CO3D than ls-disp-depth, but worse on the other two datasets. To further analyze this effect for methods predicting depth, we also mapped depth predictions into disparity space for a second alignment (ls-depth-disp) and observed the same trend. This experiment highlights the importance of the alignment procedure, and we conclude that the fairest comparison is to perform alignment in the same (depth) space, i.e., to use ls-disp-depth for methods operating on disparity and ls-depth for methods outputting depth directly. Note that the initial alignment in disparity space (ls-disp-depth) is inevitable for methods that output disparity, as it is required to obtain absolute disparity before converting disparity to depth.

11. Baseline Selection

To determine the suitability of DA-V2 as a baseline, we evaluated several existing methods (including Marigold, DepthPro, and DA-V2) on the CO3D dataset. We selected CO3D as our benchmark because its objects exhibit high detail and are relatively close to the camera (e.g., a toy truck instead of a real truck in KITTI). As a result, the depth estimates are more reliable, which aligns well with our goal of detail refinement, as motivated in the main document.

On CO3D, DA-V2 consistently outperformed all other methods. Marigold (a diffusion-based foundational model for depth estimation) captures coarse geometric structure; however, its reconstructed mesh, as shown in Fig. 7, exhibits significant high-frequency noise and artifacts. DepthPro provides an estimate of absolute scale, but it is often misled by toy versions of real objects (e.g., a toy truck), and even after normalizing for scale (and shift), it does not outperform DA-V2. This justifies our selection of DA-V2 as the baseline. At the time of the paper's first draft, DA3 had not been publicly released.

Notably, Table 4 shows that DA-V2 performs better on CO3D but slightly worse on the other two datasets. This is consistent with the evaluations in DepthPro and Marigold, which report improvements over DA-V2 on these datasets. These results support the visual observation that DA-V2 excels at predicting fine details while lagging slightly in capturing overall scene composition and the relative scale of

Table 5. **Ablation** of the camera model and b parameter on the CO3D dataset.

Method	AbsRel ↓	RMSE ↓	log10 ↓	RMSE log ↓	SI log ↓	SqRel ↓
Ours ($\mathbf{K}_{\text{orth}}$, $b = 0.1$)	**0.00223**	**0.0588**	**0.000968**	**0.00314**	**0.314**	**0.000235**
Ours-Perspective-Fixed ($\mathbf{K}_{\text{persp}}$, $b = 0.1$)	0.00224	0.0591	0.000973	0.00315	0.315	0.000236
Ours-Perspective-Opt ($\mathbf{K}_{\text{persp}}$, $b_{\text{init}} = 0.01$)	0.00225	0.0592	0.000976	0.00316	0.316	0.000237

objects.

Another closely related work is BetterDepth [61], which, similar to our approach, focuses on refining the output of a pre-trained depth foundation model. However, its implementation is not publicly available.

12. Ablation - Perspective Camera

We conduct an ablation study on the camera model (Eq. 2 in the main paper) and the choice of $b = ms$ (Eq. 5 in the main paper). We compare a scaled orthographic camera, with its scale factor optimized from an initial value of 7.0, against a perspective camera, with its focal length jointly optimized from an initial value of 2.0. As shown in Table 5, the scaled orthographic model with $b = 0.1$ yields the lowest errors. We therefore adopt this configuration for all main experiments.

13. SfS Implementation Details

We implement a simple shape from shading algorithm similar to [11]. We assume a Lambertian surface, with constant albedo and a single light at infinity such that the light direction $\mathbf{l}$ is unknown yet constant across all pixels. We consider only brightness variations and therefore convert the input image to grayscale and set the incoming light intensity to $L_{\text{in}} = \max(\mathbf{I})$, where $\mathbf{I}$ is the input image.

Under these assumptions, the image formation model reduces to a Lambertian dot product between the surface normal and the light direction:

$$\hat{\mathbf{I}}(u, v) = \max\big(0,\ \mathbf{N}(u, v) \cdot \mathbf{l}\big),$$

where $\hat{\mathbf{I}}$ is the rendered image.

Motivated by the shape from intensity gradient technique [59], we initialize the normals using image gradients,

$$\mathbf{N}(u, v) = \big(-\mathbf{I}_u(u, v),\ -\mathbf{I}_v(u, v),\ 1\big),$$

where $\mathbf{I}_u$ and $\mathbf{I}_v$ denote the partial derivatives of $\mathbf{I}$ with respect to the spatial coordinates u and v. We compute these derivatives using Scharr filters [38].

To optimize the light direction and surface normals, we minimize a combination of a photometric loss and a smoothness loss,

$$\mathcal{L} = \mathcal{L}_{\text{smooth}} + \lambda\, \mathcal{L}_{\text{photo}},$$

where λ is a regularization parameter. The photometric loss minimizes the difference between the input and the rendered images,

$$\mathcal{L}_{\text{photo}} = \frac{1}{|\Omega|} \sum_{(u,v)\in\Omega} \big(\mathbf{I}(u, v) - \hat{\mathbf{I}}(u, v)\big)^2,$$

where Ω denotes the valid region defined by the object mask. The smoothness loss penalizes spatial variation in the normals,

$$\mathcal{L}_{\text{smooth}} = \frac{1}{|\Omega|} \sum_{(u,v)\in\Omega} \left(\|\nabla \mathbf{N}_u(u, v)\|_2^2 + \|\nabla \mathbf{N}_v(u, v)\|_2^2\right),$$

where $\mathbf{N}_u$ and $\mathbf{N}_v$ denote the first two components of the normal field.

This simple shading model closely matches the relighting procedure used in our main method, highlighting the benefit of our shading-based augmentation. Unlike full photometric reconstruction, which produces artifacts at texture boundaries or under complex real-world illumination, our re-lighting refinement succeeds with a simple and robust illumination model.

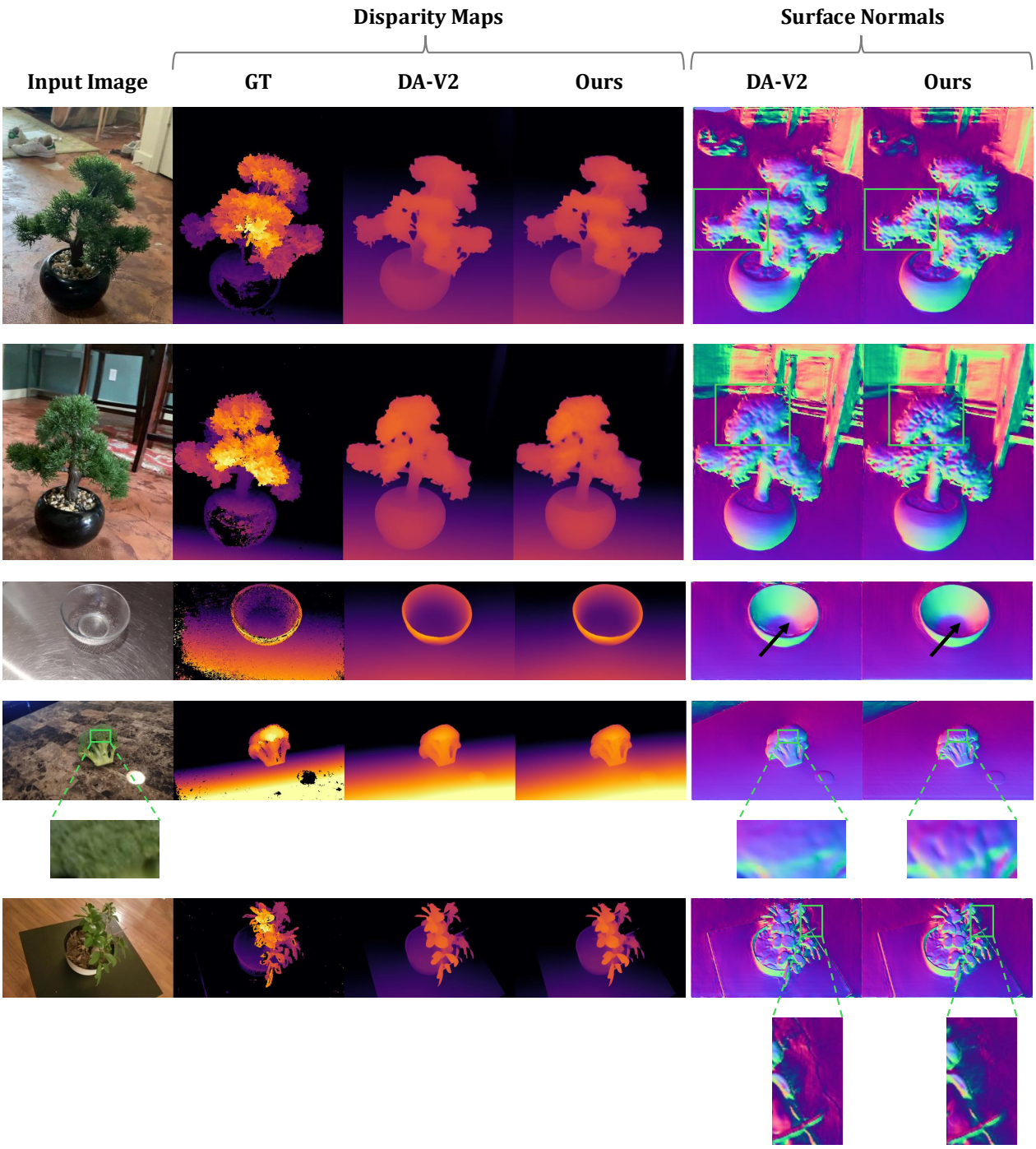

Figure 9. **Additional qualitative comparison on the CO3D dataset.**

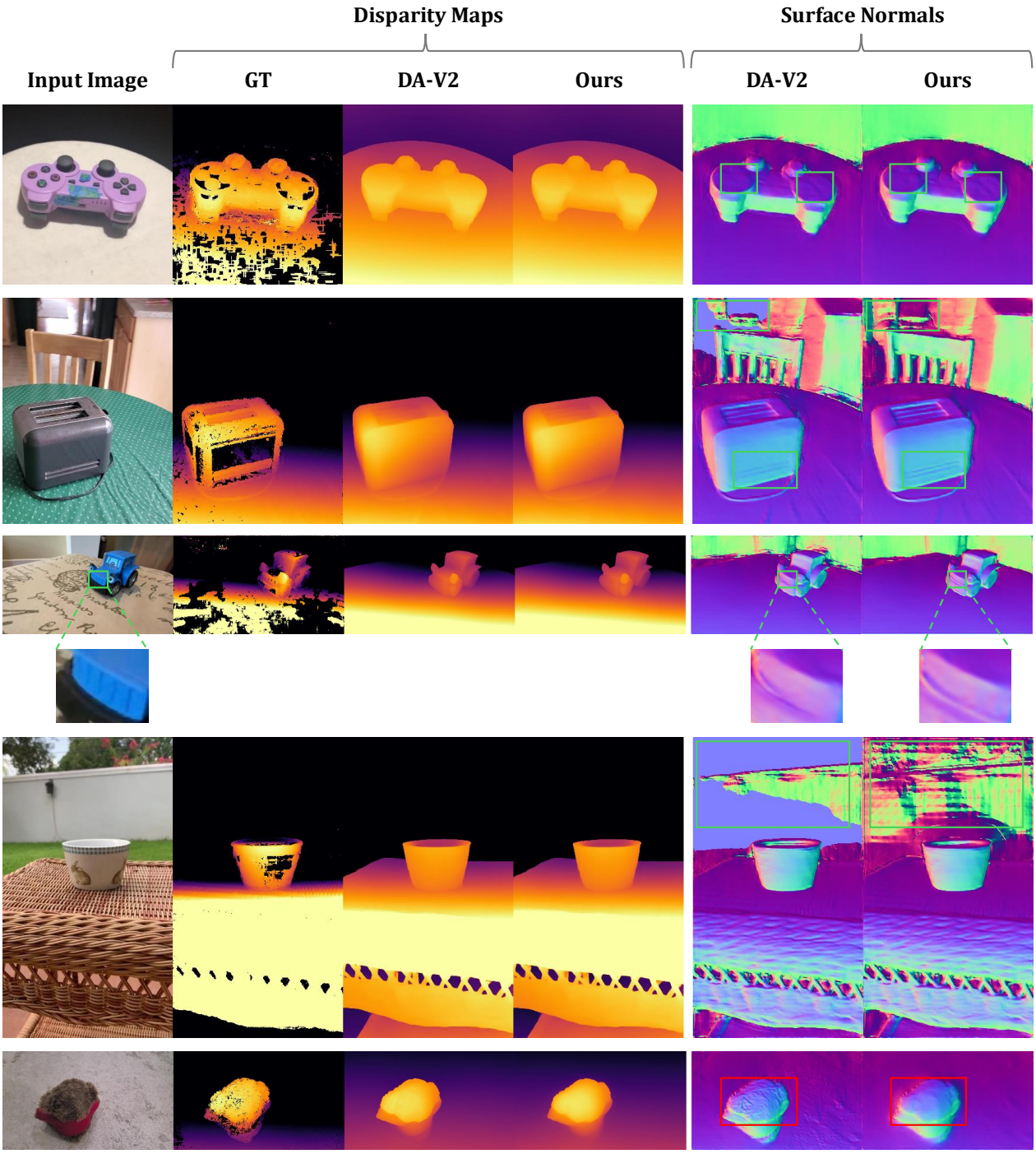

Figure 10. **Additional qualitative comparison on the CO3D dataset.** Red squares highlight occasional oversmoothing, a limitation of our method.

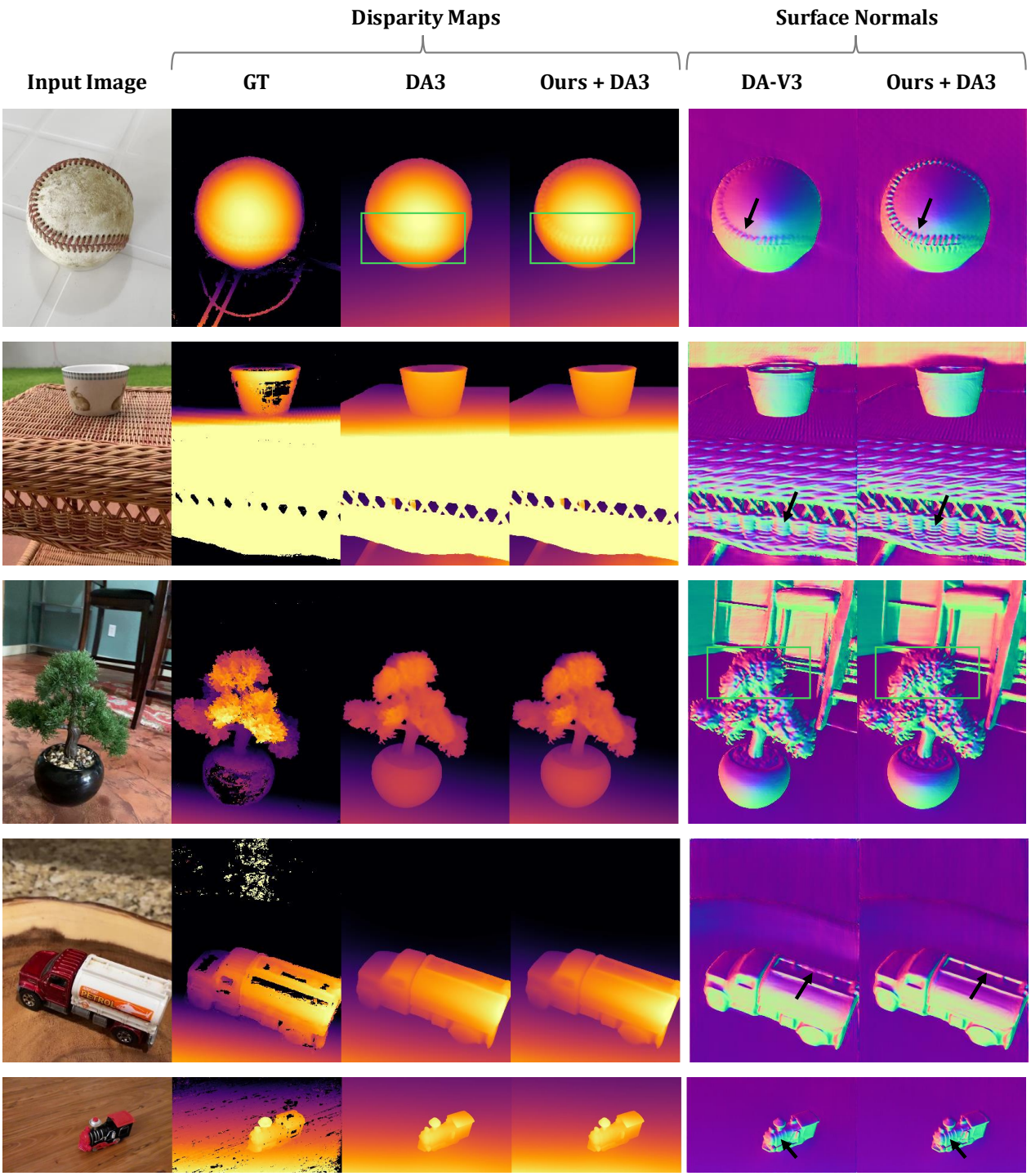

Figure 11. **Qualitative comparison against DA3 on the CO3D dataset.**

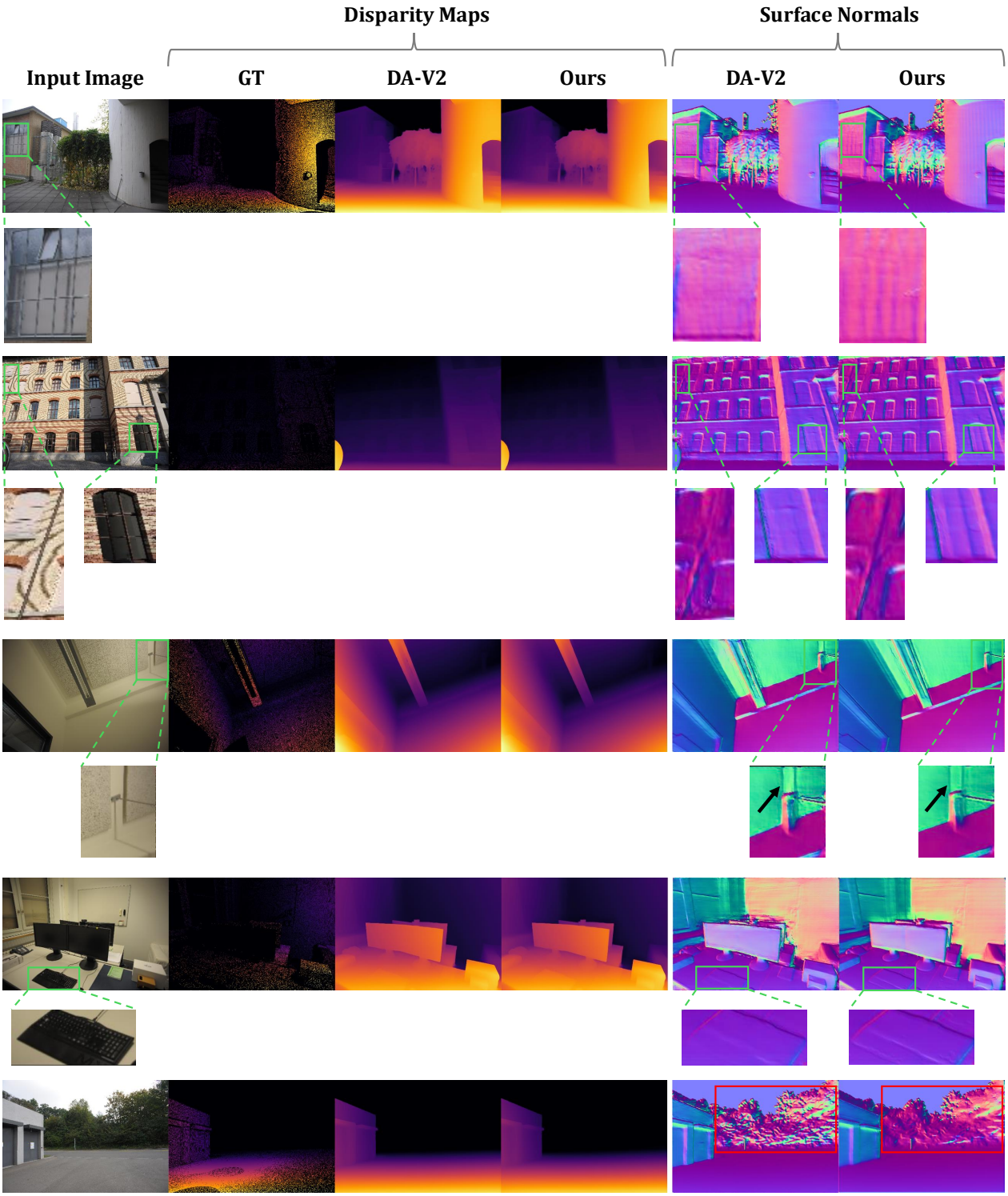

Figure 12. **Additional qualitative comparison on the ETH3D dataset.** Red squares highlight occasional oversmoothing, a limitation of our method.

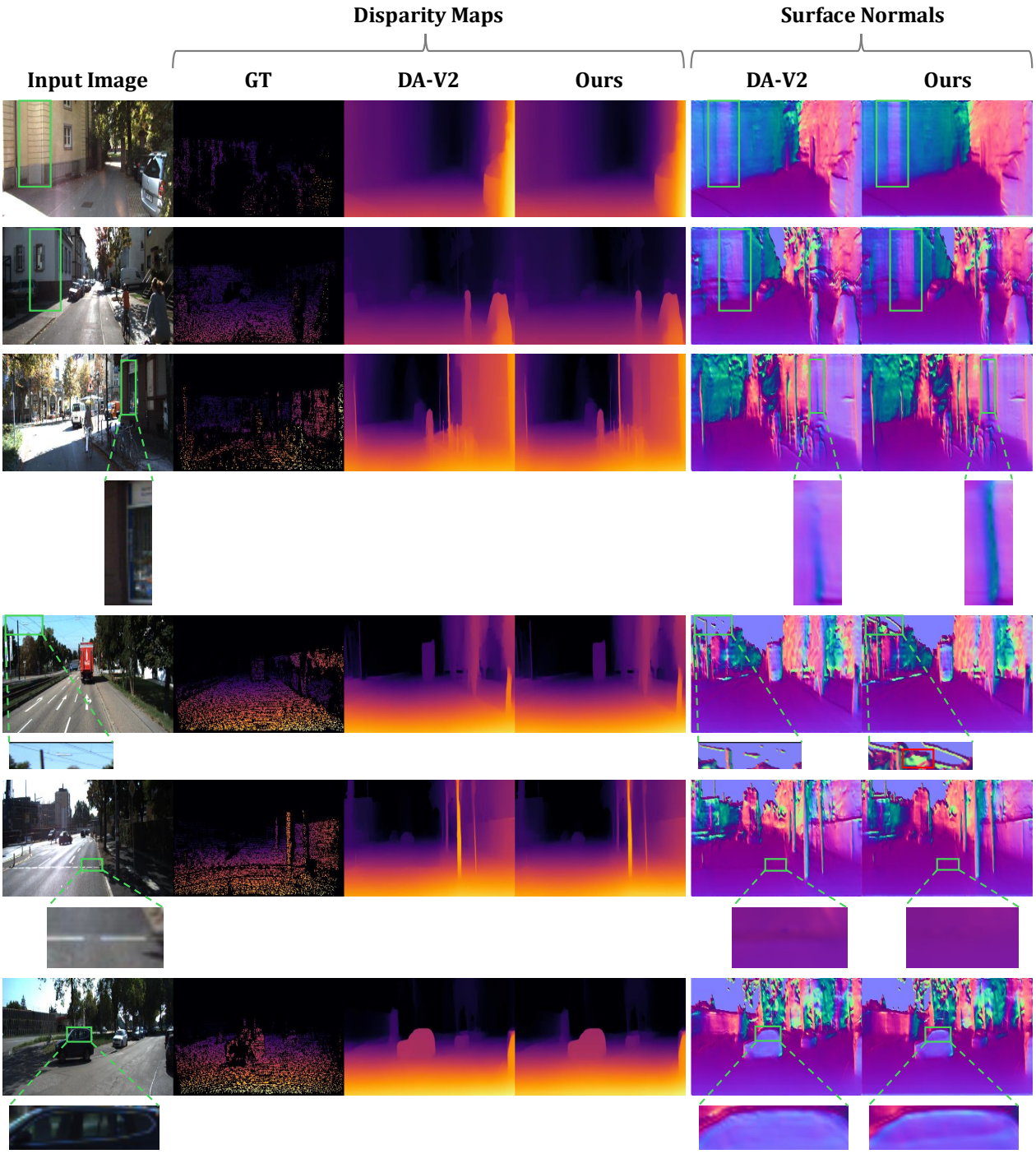

Figure 13. **Additional qualitative comparison on the KITTI dataset.** Red squares highlight occasional oversmoothing, a limitation of our method.