Supplementary Material

In this supplementary material, we further describe the details of our study not included in the main manuscript.

A. Architecture Details

In Fig. 2 of the main manuscript, we introduced SimSaC which consists of four major components: 1) a two-stream feature pyramid, 2) correlation layers estimating pixel-level similarity scores given a feature pair, 3) correspondence map decoder (CMD) estimating the scene flow w, and 4) mis-correspondence map decoder (MMD) estimating the change mask M. In this section, we further describe the details of the CMDs and MMDs in terms of the feature pyramid level l. In the following sub-sections, we use the notation of a convolutional block as a composition of the 2D convolution, batch norm and ReLU (Conv-BN-ReLU).

A.1. Correspondence Map Decoder at level l (CMD^l)

Inspired by GLU-Net [61] and PWCNet [60], our CMD^l consists of five feed-forward convolutional blocks with a spatial kernel of 3×3, and their number of feature channels are 128, 128, 96, 64, and 32, respectively. A refinement sub-network is adopted at the end of the convolutional block since this design boosts performance significantly in the optical flow estimation task [60].

More specifically, at the level $l = 1$, the correspondence map decoder CMD^l infers the flow w^l as,

$$\begin{aligned} w^l &= \text{CMD}^l(c^l), \\ c^l &= \text{Corr}_{global}(F_r^l, F_q^l), \end{aligned} \tag{8}$$

where c^l is a global correlation volume given F_r^l and F_q^l, the reference and query feature maps at level l, respectively. At the end of the convolutional blocks, we scale the output to image coordinates and convert to a displacement field, without any activation.

At the level $l > 1$, the correspondence map decoder CMD^l infers the flow w^l as follows:

$$\begin{aligned} w^l &= \text{CMD}^l(c^l, \text{up}(w^{l-1})) \\ c^l &= \text{Corr}_{local}(\tilde{F}_r^l, F_q^l). \end{aligned} \tag{9}$$

Here, c^l is a local correlation volume given $\tilde{F}_r^l$ and F_q^l, the aligned reference and query feature maps at level l, respectively, and up(w^{l-1})) refers to the up-sampled flow from the previous level, $l-1$. At the end of the convolutional blocks, we scale the output to image coordinates and convert to a displacement field, without any activation.

A.2. Mis-correspondence Map Decoder at level l (MMD^l)

MMD^l consists of five feed-forward convolutional blocks with a spatial kernel of 3×3, and their number of feature channels are 128, 128, 96, 64, and 32, respectively. More specifically, at the level $l = 1$, the mis-correspondence map decoder MMD^l infers the mis-correspondence map (change mask) M^l as follows:

$$\begin{aligned} M^l &= \text{MMD}^l(c^l, F_r^l, F_q^l) \\ c^l &= \text{Corr}_{global}(F_r^l, F_q^l), \end{aligned} \tag{10}$$

where c^l is a global correlation volume given F_r^l and F_q^l, the reference and query feature maps at level l, respectively.

At the level $l > 1$, the mis-correspondence map decoder MMD^l infers the mis-correspondence map M^l as follows:

$$\begin{aligned} M^l &= \text{MMD}^l(\bar{c}^l, \tilde{F}_r^l, F_q^l) \\ \bar{c}^l &= c^l \odot (\sigma(\text{up}(M^{l-1}))) \\ c^l &= \text{Corr}_{local}(\tilde{F}_r^l, F_q^l) \end{aligned} \tag{11}$$

Here, c^l is a local correlation volume given $\tilde{F}_r^l$ and F_q^l, the aligned reference and query feature maps at level l, respectively. $\bar{c}^l$ is a masked correlation volume, where possibly mis-correspondent pixels remain non-zero using the up-sampled mis-correspondemce map up(w^{l-1})) from the previous level, $l-1$; $\odot$ and σ represent element-wise multiplication and the sigmoid activation, respectively.

B. More Qualitative Results

Here, we additionally provide qualitative examples of the baseline networks and the proposed SimSaC architecture. Figures 5, 6, 7, 8, 9, and 10 depict change detection results on ChangeSim-normal, ChangeSim-dusty-air, ChangeSim-low-illumination, VL-CMU-CD, PCD-GSV, and PCD-TSUNAMI, respectively. We describe qualitative comparisons in the following.

Camera viewpoint. The proposed SimSaC architecture effectively handles the cases where the camera viewpoint difference is large. However, other models almost fail when the difference is large. In the failure case, false negative cases are more common than false positives.

Unseen domain. The proposed SimSaC architecture is robust to domain shifts such as dusty-air and low-illumination. Note that the two splits (dusty-air and low-illumination) are only used for evaluation. Nonetheless, baseline models do not function when domain shifts occur.

Details in a mask. The proposed SimSaC architecture properly captures the details of change masks in both wide areas and narrow areas as well. On the other hand, baseline models tend to miss change masks in narrow areas.

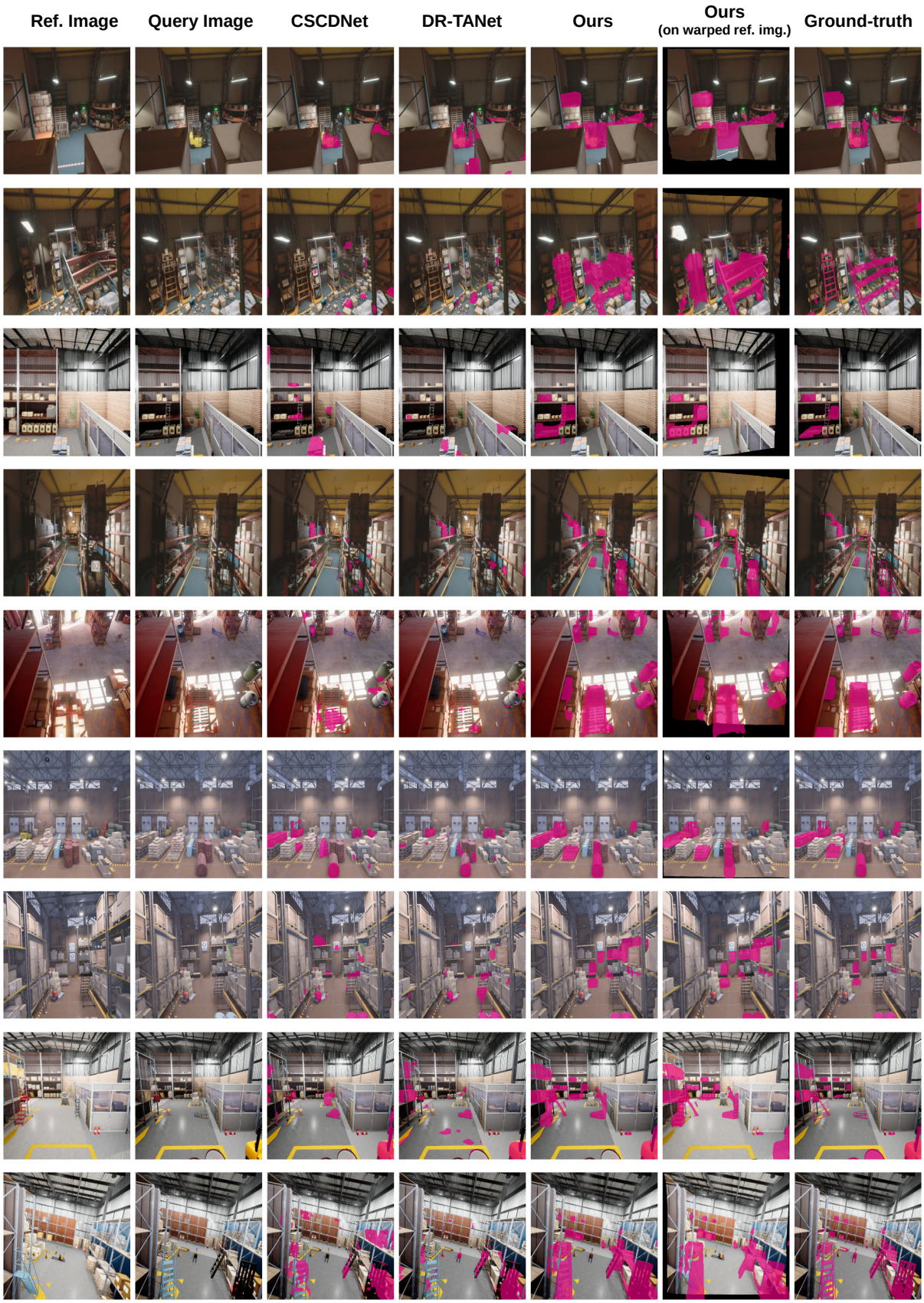

Figure 5. Qualitative results on *normal* split of the ChangeSim dataset. The change masks are marked with purple.

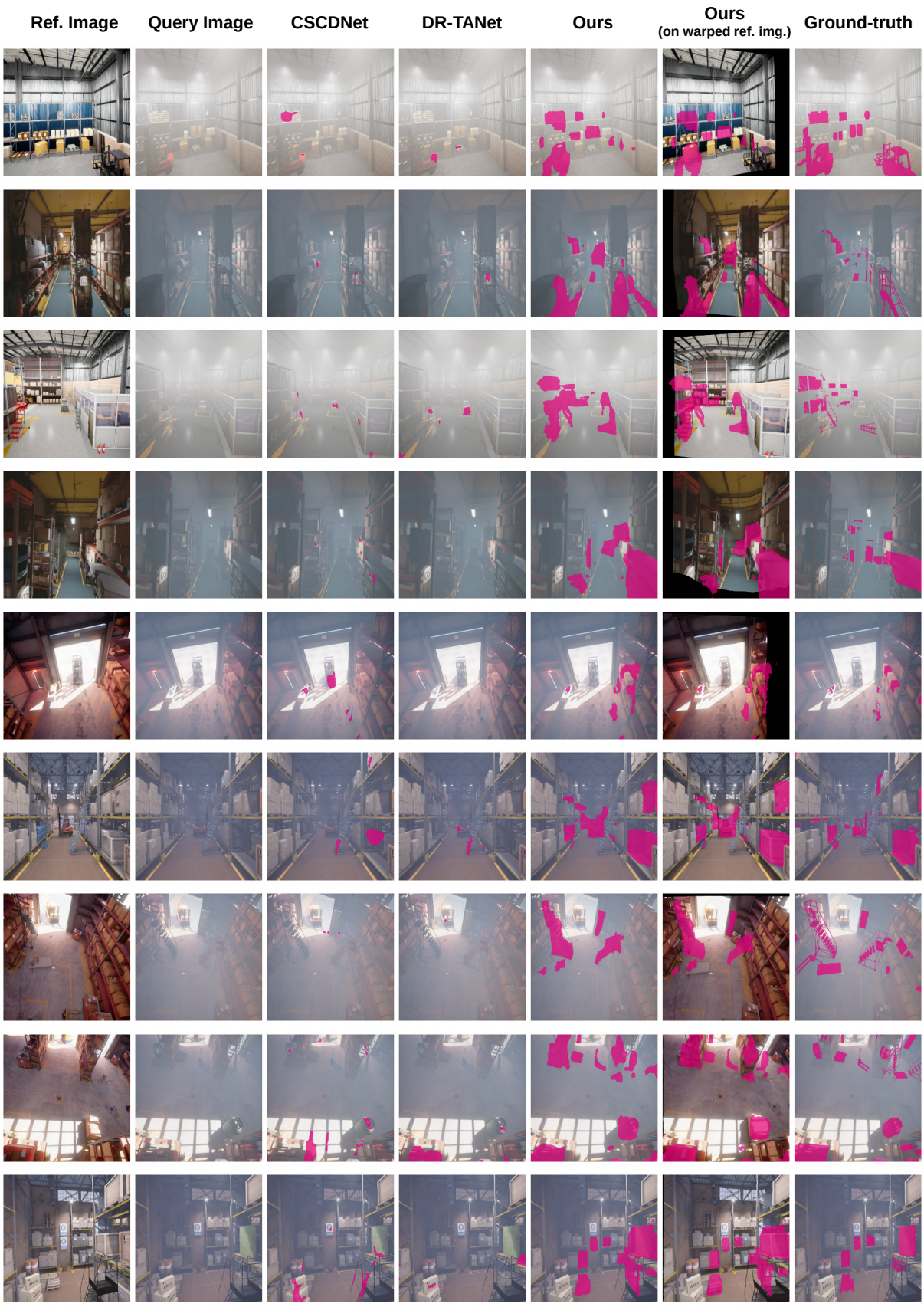

Figure 6. Qualitative results on *dusty-air* split of the ChangeSim dataset. The change masks are marked with purple.

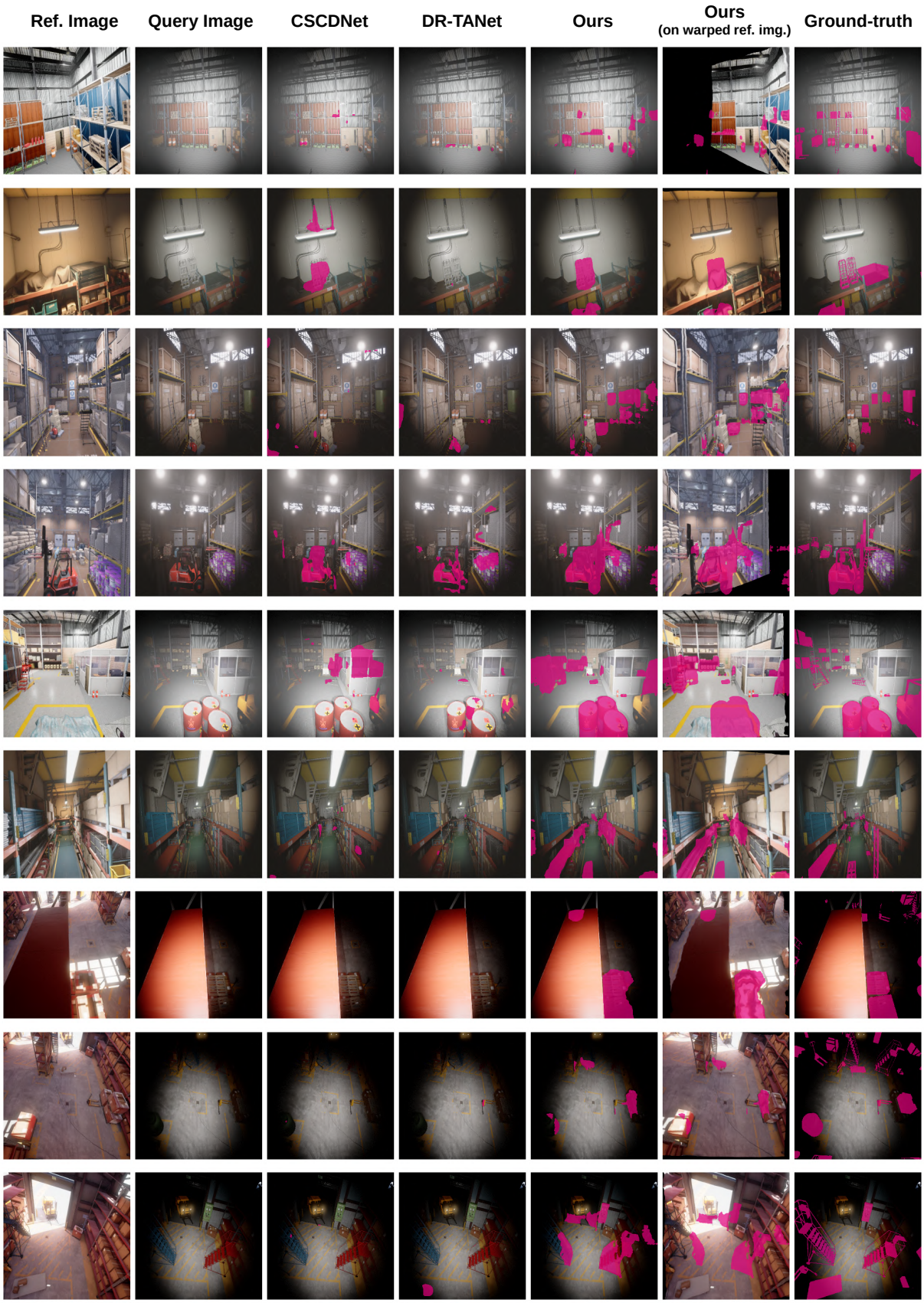

Figure 7. Qualitative results on *low-illumination* split of the ChangeSim dataset. The change masks are marked with purple.

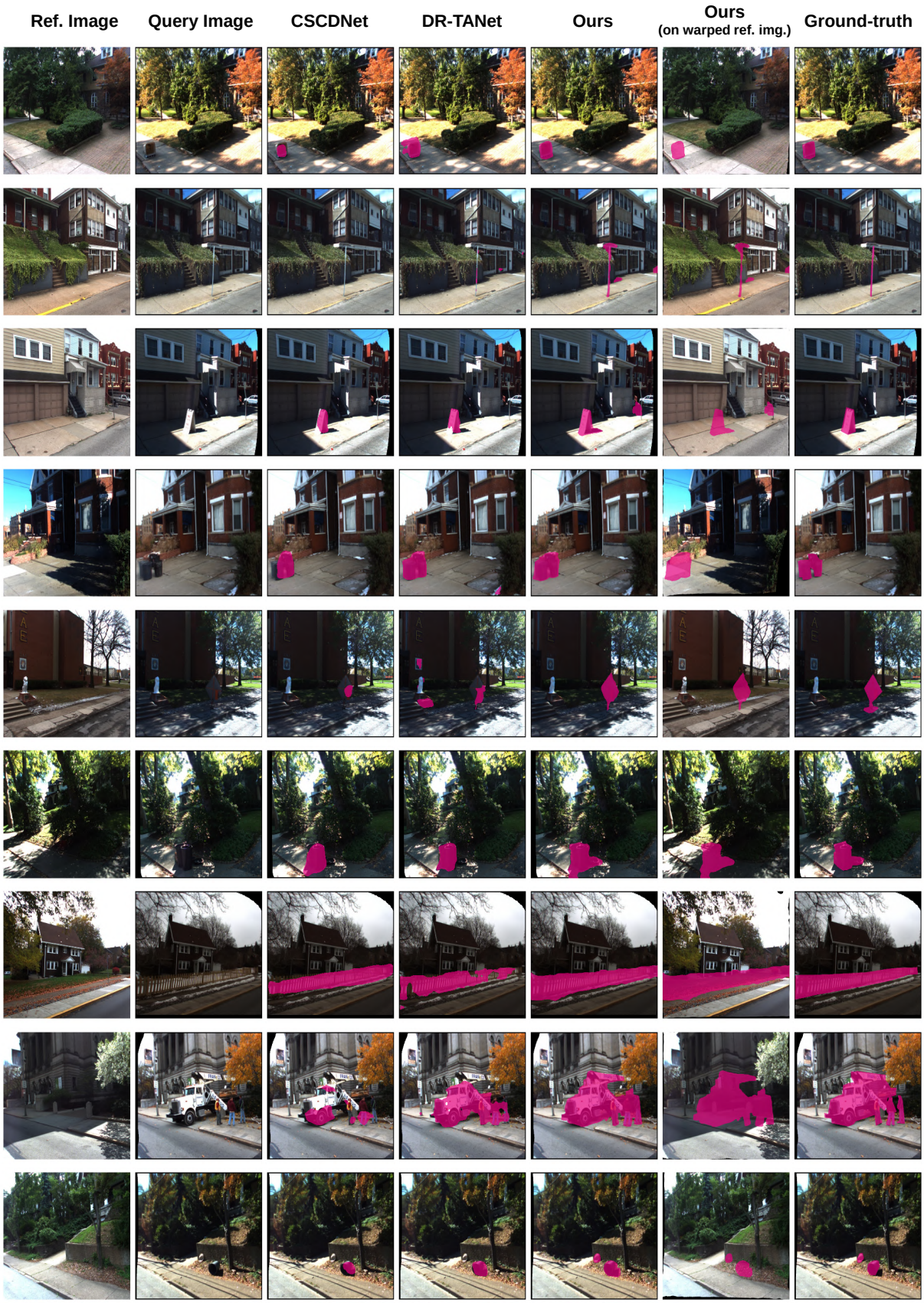

Figure 8. Qualitative results on the VL-CMU-CD dataset. The change masks are marked with purple.

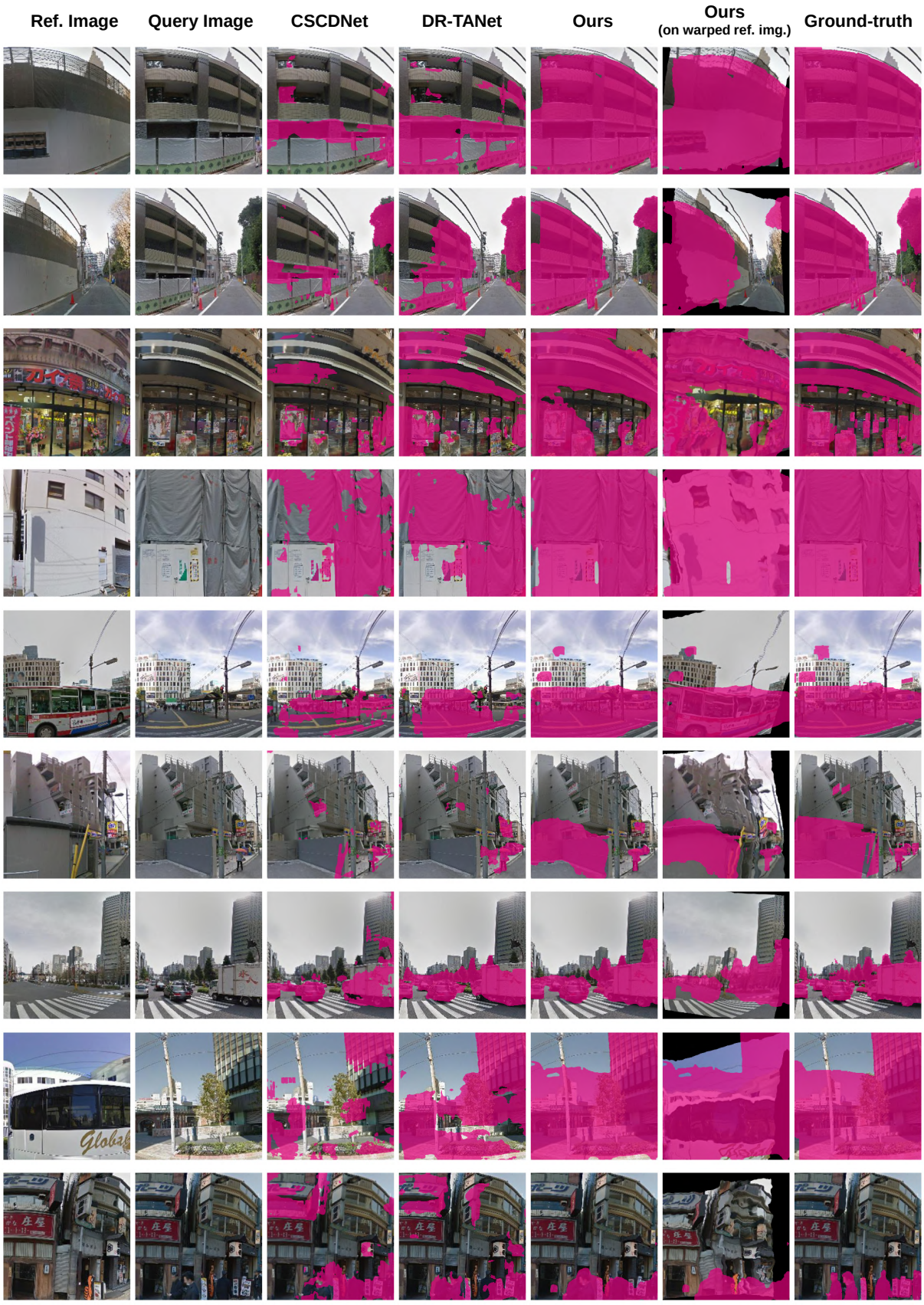

Figure 9. Qualitative results on the GSV split of the PCD dataset. The change masks are marked with purple.

Figure 10. Qualitative results on the TSUNAMI split of the PCD dataset. The change masks are marked with purple.