MU-GeNeRF: Multi-view Uncertainty-guided Generalizable Neural Radiance Fields for Distractor-aware Scene

Supplementary Material

7. More Experimental Settings

This supplementary material consists of two parts: 1) additional experimental details of MU-GeNeRF, included to offer a more thorough understanding of the training and evaluation process; and 2) provide extended experiments on hyperparameter sensitivity analysis, uncertainty modeling, and distractor awareness to further validate the effectiveness of the proposed method.

7.1. Model Design

In this section, we provide additional details on the experimental setup of MU-GeNeRF, including the network architecture design, dataset processing strategies, and the configurations of baseline methods used for comparison.

Feature Extraction To effectively capture structural details and semantic context, we employ a VGG-based convolutional network to extract multi-level features from source views $\{I_n^s\}_{n=1}^N$. These features are fused into a compact structural representation $\{F_n^s\}_{n=1}^N$ through channel mapping and spatial alignment, thereby supporting novel view synthesis and Source-view Uncertainty modeling. For the target view I^t, we utilize the pre-trained DINOv2 ViT-S/14 model [31] to extract dense semantic features F^t. This model exhibits strong cross-view consistency and semantic discriminability, providing reliable semantic guidance for Target-view Uncertainty modeling.

Network Architecture Our feed-forward network $\mathcal{G}_\theta$ is primarily built upon the key Transformer modules from VolRecon [35] and ReTR [20], enabling robust multi-view feature aggregation and efficient neural rendering. Specifically, the View-Transformer explicitly models cross-view structural consistency, followed by the Render-Transformer, which in turn produces attention-based weights to replace conventional volume rendering. Through this pipeline, multi-view projected features f_n^s are fused with color samples c_n^s and spatial information (x_k, d_k), and are subsequently decoded by two independent MLP heads into the color c_k and Source-view Uncertainty β_k^s for each sampled point. This design effectively enhances the model's ability to capture cross-view consistency and visibility variations, demonstrating strong robustness in complex scene reconstruction and uncertainty estimation.

The decoder $\mathcal{F}_\theta$ consists of several lightweight convolutional network layers and MLPs. It progressively compresses DINO features F^t through multiple convolutional layers with nonlinear activation functions, producing a dense Target-view Uncertainty map β^t, wherein a Softplus activation ensures the outputs remainpositive, and bilinear upsampling aligns the uncertainty map β^t to the spatial resolution of the target view I^t.

7.2. Uncertainty Regularization

To ensure numerical stability in uncertainty estimation, we impose a lower bound on the model's raw predicted uncertainty $\tilde{\beta}(r)$ to prevent extremely small values during early training. Specifically, the final uncertainty $\beta(r)$ is computed as follows:

$$\beta(r) = \beta_{\min} + \mathrm{softplus}(\tilde{\beta}(r)) \tag{9}$$

Here, $\mathrm{softplus}(x) = \log(1 + e^x)$ is a smooth nonlinear activation function, and $\beta_{\min} > 0$ is a predefined constant that ensures all uncertainty values remain positive and above a critical stability threshold.

7.3. Dataset Processing

RobustNeRF dataset The RobustNeRF dataset [36] comprises five indoor scenes—Statue, Android, Crab1, Crab2, BabyYoda. Each scene is split into a training set with dynamic, non-continuous distractors and a clean validation set without distractors. We use Crab1, Crab2, and BabyYoda for generalization training, and perform per-scene fine-tuning and evaluation on Statue and Android. Due to the lack of temporal continuity and the prevalence of distractors in this dataset, conventional source-view selection strategies based on frame proximity or keypoint matching are no longer applicable. Therefore, we adopt a camera-pose-based strategy to select source views: for each target view, we first compute the angular and translational distances to all candidate views, then normalize them independently and combine them using weights of 0.7 and 0.3, and finally select the Top-N views with the lowest combined scores as source views.

On-the-go dataset The On-the-go dataset [35] comprises both indoor and outdoor scenes containing various dynamic objects (e.g., pedestrians, toys, and robots), with occlusion ratios ranging from 5% to 30%. We selected four scenes for per-scene fine-tuning and evaluation: Corner and Patio, which are categorized as medium occlusion scenes, and Spot and Patio-High, which are high occlusion scenes. The remaining eight scenes were used for generalization training. The source view selection strategy follows the same procedure as in the RobustNeRF dataset. Due to the large

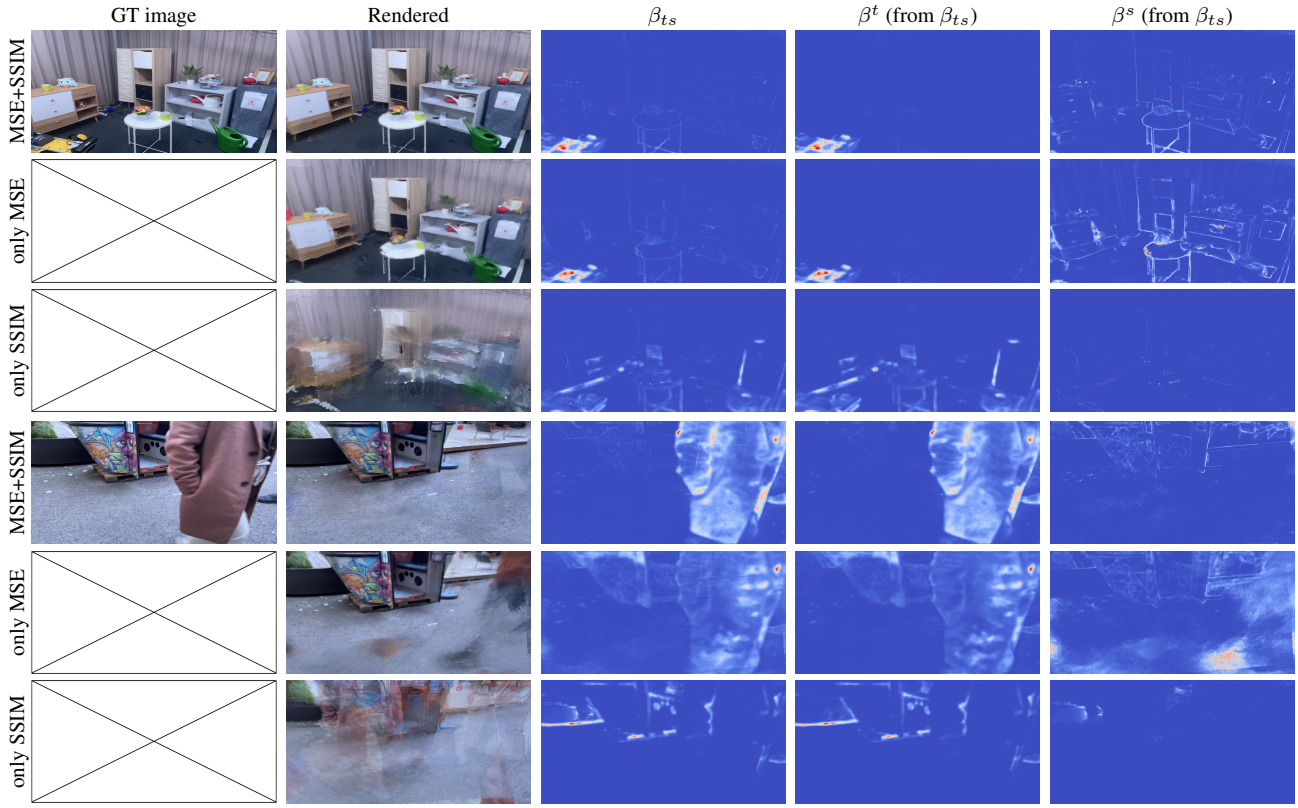

Figure 10. **Comparison of loss term ablation experiments,** showing the rendering results and uncertainty visualizations using MSE loss only, SSIM loss only, and the combination of both losses.

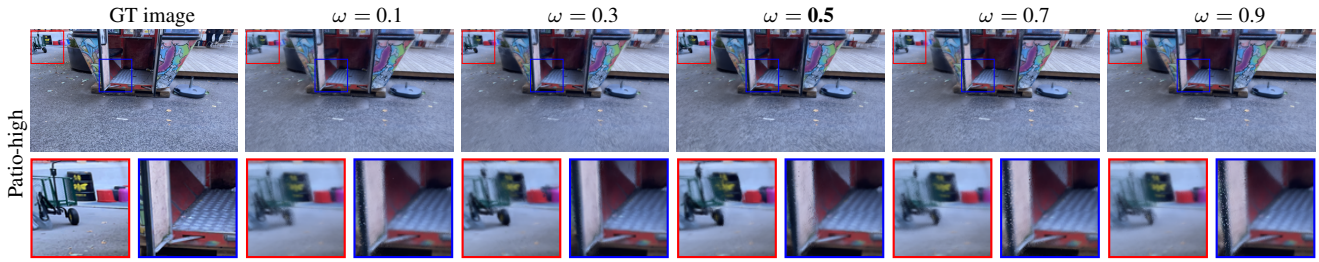

Figure 11. **Qualitative Comparison under Different Weight Settings.**

baseline variations between images in this dataset, it is difficult to select source views that fully cover the field of view for some target images. To ensure a reliable evaluation, we removed a limited number of validation images with the following indices: Spot–174, 175, 176, 177; Corner–8, 9, 11, 113, 114; Patio–81, 89, 90, 91, 92, 93; Patio-High–0, 12, 13, 14, 21, 22, 23. Since the Spot scene's validation set contains relatively few images, the number of source views during evaluation is limited to four, while it remains eight for the other scenes.

7.4. Baseline

To ensure a fair comparison, we adjusted each baseline according to the characteristics of the dataset and implementation constraints. For ReTR, we removed depth supervision, as no ground-truth depth is available. For MuRF [51], in addition to fine-tuning our self-trained generalization model, we also incorporated the officially released MuRF-mixdata pretrained model, which is specifically recommended for real-world scenarios. For UP-NeRF [15], which was originally designed for unposed reconstruction, we provided ground-truth camera poses to ensure consistent experimental settings. For NeRF on-the-go [35], due to its uncertainty estimation module being highly sensitive to image

resolution, we could only use the original resolutions image (1080×1920) as input during training, while validation was conducted at a unified resolution of 320×640.

ω	PSNR↑	SSIM↑	LPIPS↓
0.1	20.18	0.550	0.323
0.3	20.21	0.552	0.317
0.5	**20.33**	**0.564**	**0.311**
0.7	20.15	0.548	0.328
0.9	20.03	0.539	0.336

Table 4. **Quantitative Analysis of Weight Sensitivity.**

8. More Experimental Results

8.1. Loss Term Analysis

We presents an ablation study analyzing the role of each loss term in the heteroscedastic reconstruction loss $\mathcal{L}_{\text{Multi-uncer}}$. As shown in Fig. 10, training with MSE loss alone lacks spatial contextual awareness and the model is prone to interference from isolated noise, resulting in rendered results with blur and artifacts. Simultaneously, Multi-view Uncertainty modeling also fails to properly disentangle reconstruction errors—β^t still misjudges static regions, while β^s cannot accurately capture structural inconsistencies across source views. Conversely, using SSIM loss alone deprives the model of pixel-level color supervision, causing the color estimation to collapse and severe degradation in overall rendering quality; in this case, the Multi-view Uncertainty modeling completely collapses, and both β^t and β^s fail to function properly. By contrast, combining the two losses leverages their complementary strengths: the MSE loss provides essential pixel-level color supervision, while the SSIM loss introduces spatial coherence constraints. Together, they ensure the reliable modeling of Multi-view Uncertainty β_{ts}, ultimately achieving higher-quality geometric reconstruction.

8.2. Parameter Sensitivity Analysis

We conduct a sensitivity study on the fusion weight ω for Multi-view Uncertainty β_{ts}, with results summarized in Tab. 4 and Fig. 11. Within the range $\omega \in [0.3, 0.5]$, the PSNR varies by only ±0.15 dB and the LPIPS error varies by no more than 6%, indicating strong robustness of our method to small perturbations across this interval. With increasing ω, which assigns dominant influence to the Target-view Uncertainty β^t, slight noise and a loss of high-frequency details appear near geometric edges. We hypothesize that this is due to the weakened constraint of β^s on cross-view structural conflicts, which reduces the model's reliability when handling complex geometric structures. Conversely, when ω decreases to 0.1, the weight of β^s becomes excessive, resulting in slightly blurred renderings,

Ablation	Corner			Patio-High		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
MLP gating	17.52	0.565	0.523	14.35	0.435	0.537
w/o GMM	19.46	0.614	0.461	19.17	0.486	0.450
Ours	**21.77**	**0.669**	**0.338**	**20.33**	**0.564**	**0.311**

Table 5. **Quantitative comparison of ablation components.**

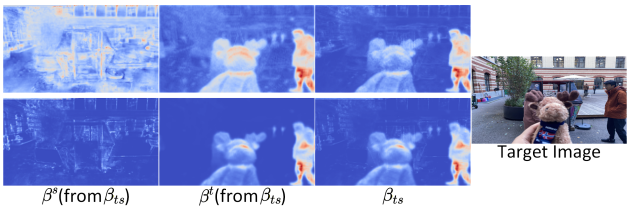

Figure 12. **Uncertainty modeling via MLP gating.** Top: *MLP gating*; Bottom: *Ours*.

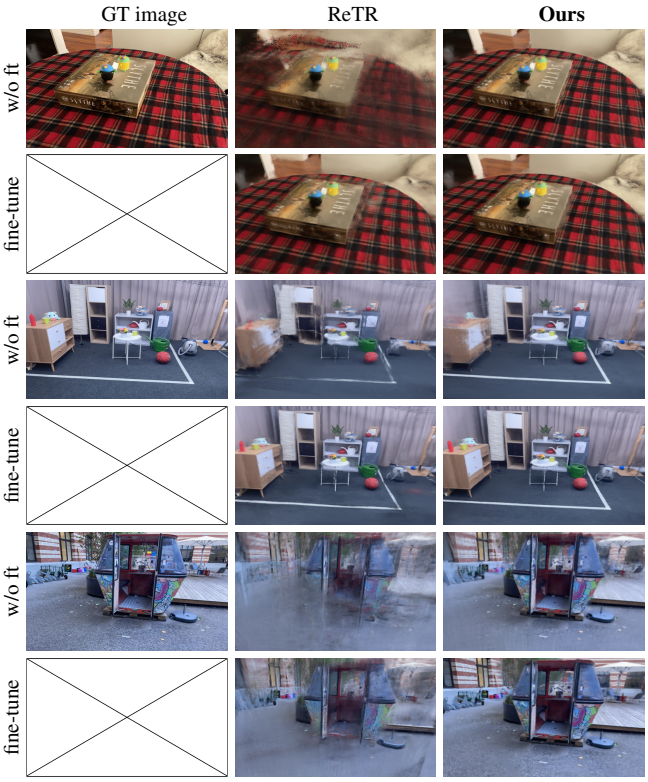

Figure 13. **Comparison of rendering results before and after fine-tuning.**

possibly because the role of β^t in locating distractors is limited, thereby affecting local reconstruction quality. We also explored MLP gating, but found that it led to a shortcut solution: the easily optimized β^t branch absorbs most of the errors, misidentifying static structures as distractors. This undermines our decoupling design based on error sources, causing β^s to become disordered as well (Fig. 12, Tab. 5).

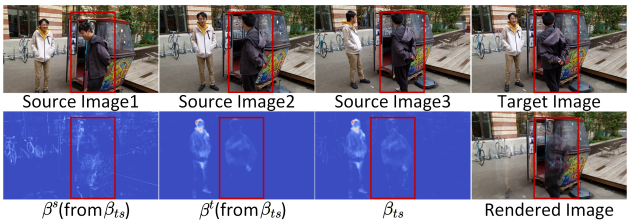

Figure 14. **Failure cases of β^t.** When a distractor appears across views with subtle motion, it exhibits high geometric consistency, which significantly impairs the accuracy of β^t.

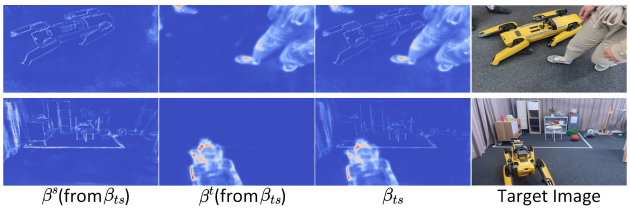

Figure 15. **Scene-dependent distractor awareness.** Top: *Spot*; Bottom: *Corner*.

8.3. Comparison of Generalization Performance

Fig. 13 presents a rendering comparison between ReTR and our method, both before and after fine-tuning. It can be observed that ReTR struggles to reliably aggregate cross-view information during generalization inference, producing renderings with noticeable artifacts and structural anomalies, along with poor geometric integrity. Although the basic geometric structure can be recovered after fine-tuning, residual noise and local distortions remain. This is because ReTR lacks a distractor-aware mechanism and cannot effectively distinguish reliable observations from transient distractors in the training data, leading the learned scene priors to be easily contaminated by distractors, which in turn limits its cross-scene generalization capability. In contrast, our method is able to reconstruct structurally complete scene geometry after generalized training. After fine-tuning, the detail quality is further improved, ultimately producing clear and coherent rendering results. This fully demonstrates the robust generalization advantage of the proposed method in scenes containing distractors.

8.4. Analysis of Uncertainty Modeling

We clarify that distractor awareness is fundamentally driven by multi-view geometric consistency rather than memorizing specific semantic categories. Thus, $\mathcal{F}_\theta$ does not learn a hard "semantic-to-distractor" mapping; high β^t is triggered only when semantic objects violate geometric consistency. This implies that whether the same object acts as a distractor across different scenes is not determined by DINO-provided semantic priors, but depends on its multi-view consistency within a specific scene, which also explains the failure cases in Fig. 14. Toy experiment (Fig. 15) confirms this: during joint training on Corner and Spot, the "dog" in Spot appears as a multi-view consistent static object, with its error absorbed by β^s; while the "dog" in Corner acts as a randomly appearing distractor, exhibiting high β^t— despite their semantic equivalence in DINO, this does not lead to systematic misjudgment by β^t. We admit that $\mathcal{F}_\theta$ does not possess generalizability and cannot a priori distinguish transient from static objects (which is also infeasible in principle), yet this design effectively identifies distractors.

Fig. 16 provides a multi-scene comparison that further validates the significant differences in distractor suppression and geometric reconstruction among various uncertainty modeling strategies. First, methods lacking a distractor-aware mechanism (e.g., ReTR and our "w/o β" variant) exhibit obvious artifacts and local distortions, indicating that the standard GeNeRF aggregation framework fails to distinguish transient distractors from true scene geometry. Second, the variant that models only Source-view Uncertainty β^s can alleviate cross-view structural conflicts, but its inability to perceive transient distractors in the target view leads to insufficient suppression and residual artifacts. Conversely, the variant that models only Target-view Uncertainty β^t can localize distractors but tends to misjudge inconsistent static structures as distractors, causing local detail loss. Finally, the complete Multi-view Uncertainty modeling β_{ts} leverages both semantic awareness from the target view and structural consistency across source views, enabling robust geometry reconstruction under diverse distractor patterns—demonstrating its necessity and superiority in complex distractor environments.

Additionally, ablation experiments (Tab. 5) confirm the necessity of our GMM formulation, showing that direct pixel-level prediction of β^s via MLP leads to significant performance degradation. Since β^s inherently reflects 3D structural inconsistencies, GMM effectively preserves this critical spatial distribution information.

8.5. Distractor-Aware Evaluation

Fig. 17 demonstrates the our method's ability to perceive and suppress transient distractors during training. The Multi-view Uncertainty maps accurately localizes the distractor regions in the target view with high uncertainty, while the view-aggregation stage effectively eliminates their influence, ultimately producing clean and clear rendering results. This indicates that even without explicit detection or annotations, our method can perceive and suppress transient distractors, thereby providing stable and reliable supervision signals for the geometry reconstruction.

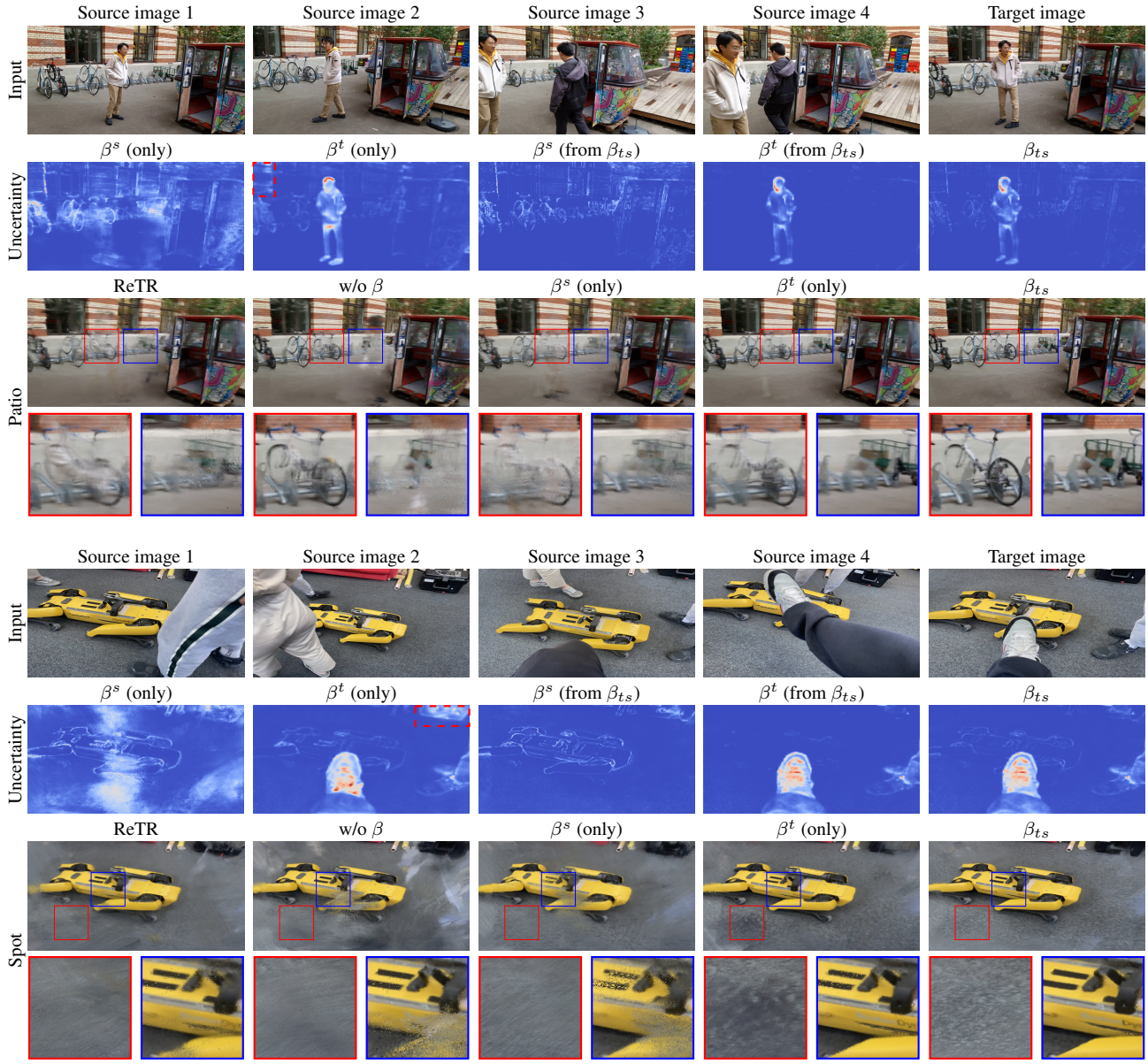

Figure 16. **Qualitative comparison of rendering results in distractor-containing scenes.**

Figure 17. **Visualization results of distractor awareness.** From top to bottom are: input target view, the Multi-view Uncertainty results, and the rendered results after distractor removal.