Supplementary Material

Freeze-Frame with StaticNeRF: Uncertainty-Guided NeRF Map Reconstruction in Dynamic Scenes

1 Analysis of NeRF-W and Reformulation of the Problem

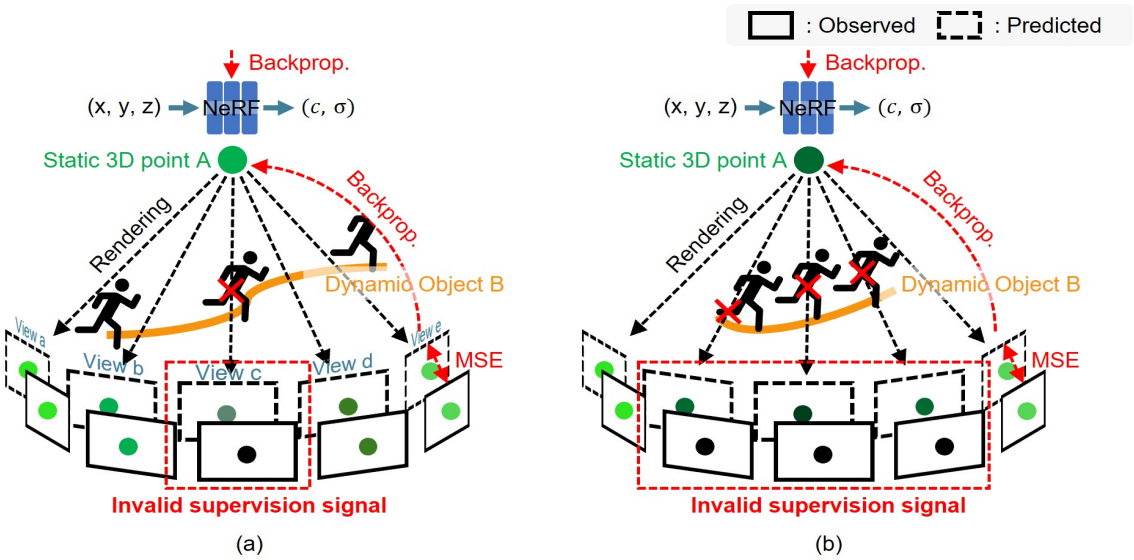

Figure A: **NeRF training in the presence of dynamic objects.** (a) When the dynamic object rarely occludes the static 3D point A, supervision is mainly provided by static pixels, enabling accurate learning of the static scene. (b) When the dynamic object frequently occludes the static point, supervision becomes dominated by dynamic pixels, leading to incorrect learning and failure to suppress the dynamic content.

Previous methods are typically based on the assumption that dynamic objects move over time, whereas static backgrounds occupy fixed positions in the world coordinate system. Building on this assumption, NeRF-W models dynamic content using aleatoric uncertainty to suppress its influence. Before analyzing how NeRF-W addresses dynamic regions, we first examine how vanilla NeRF behaves in the presence of dynamic objects.

1.1 Training Characteristics of NeRF under Scene Dynamics

Consider a static scene in which a dynamic object B moves across frames, as illustrated in Fig. A(a). NeRF is trained to reconstruct the scene, in which a static 3D point A is visible from viewpoints $\{a, b, d, e\}$ but occluded by the dynamic object in viewpoint c. Supervision for A is provided indirectly through the MSE loss between predicted and observed images, with gradients used to update the MLP parameters.

When the dynamic object occludes point A in viewpoint c, the observed color reflects the appearance of the dynamic content rather than the static background. Since A is consistently supervised from the other views, NeRF assigns it a static color. In turn, rendering viewpoint c yields a noticeable discrepancy at that pixel, leading to higher reconstruction errors in dynamic regions.

Consequently, NeRF successfully reconstructs static regions given dense and consistent observations, while dynamic regions, receiving sparse and inconsistent supervision, tend to exhibit higher reconstruction errors.

1.2 Conditions for Successful Dynamic Object Removal

NeRF-W exploits the fact that reconstruction loss tends to be higher in dynamic regions than in static regions, and leverages this property to model aleatoric uncertainty:

$$\mathcal{L}_{\text{NeRF-W}}(\mathbf{r}) = \underbrace{\frac{\overbrace{\left\| C(\mathbf{r}) - \hat{C}(\mathbf{r}) \right\|_2^2}^{\text{Reconstruction Loss}}}{2\beta(\mathbf{r})^2} + \log \beta(\mathbf{r})}_{\text{Aleatoric Uncertainty Loss}} + \underbrace{\frac{\lambda_u}{N} \sum_{i=1}^{N} \sigma_i^t}_{\text{Regularization Loss}}, \tag{1}$$

where the reconstruction error in the numerator encourages the learned uncertainty tensor $\beta(\mathbf{r})$ to increase in regions with high error. At the same time, $\beta(\mathbf{r})$ functions as a weighting term, reducing the impact of uncertain regions on the loss. This allows the model to de-emphasize unreliable areas during optimization. Through this mechanism, NeRF-W models dynamic objects as aleatoric uncertainty.

To clearly separate static backgrounds from dynamic objects, the reconstruction loss in dynamic regions must remain high throughout training. This condition is satisfied when the corresponding 3D points appear static from most viewpoints, with only few affected by dynamic occlusions, as shown in Fig. A(a). Therefore, sufficient observation of the static background is essential for successful dynamic object removal. Ultimately, the failure to suppress dynamic content can be attributed to insufficient static supervision, often caused by frequent occlusions, as depicted in Fig. A(b).

1.3 Challenges in the Learning Process

Observed pixels corresponding to static 3D points are occasionally subject to appearance variations across views, especially when non-Lambertian effects are involved. Nonetheless, such changes tend to follow predictable patterns determined by the viewing direction or rendering pipeline. In contrast, pixels affected by dynamic objects exhibit irregular variations that cannot be modeled consistently. Under ideal conditions, such irregularities should enable the model to suppress dynamic content as long as sparse clean viewpoints are available. When this separation fails despite such minimal supervision, the failure can ultimately be attributed to shortcomings in the learning process.

Before addressing this issue, we briefly revisit NeRF's optimization behavior. NeRF is trained to progressively achieve high-fidelity rendering, and NeRF-W extends this objective by additionally aiming to separate static and transient fields. This learning objective naturally leads to a gradual reduction in the extent of the transient field over time. If the transient field does not diminish, it implies that even static regions are being incorrectly explained by transient components. In a training regime where the transient field is continuously suppressed, transient content may be mistakenly absorbed into the static field. Once this occurs, it becomes difficult to reverse due to both NeRF's inherent convergence behavior and the effect of the regularization term in Eq. (1).

This problem is particularly pronounced in the early stages of training, where premature convergence can arise from overly rapid optimization. In our case, this effect is further exacerbated by both data-driven sampling and the use of multiresolution hash encoding, the latter of which is

integrated into our system and significantly accelerates convergence. Moreover, such misalignment can still occur mid-training, even with proper initialization, highlighting the need for training strategies that maintain sufficient expressiveness in the transient field throughout optimization.

2 Effect of Joint Optimization on Uncertainty Estimation

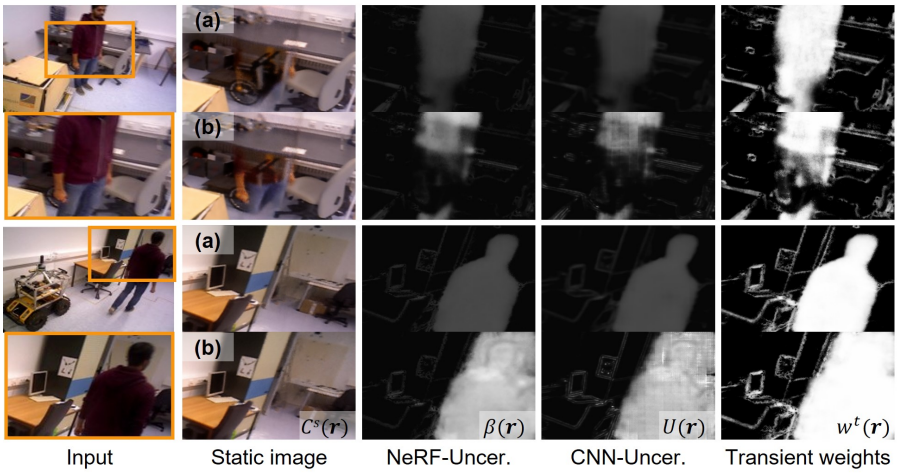

Figure B: **Effect of joint optimization on CNN-based uncertainty estimation.** (a) With joint optimization, the CNN-based uncertainty network effectively captures dynamic regions, enabling accurate static image reconstruction. (b) When the gradient flow through the tensor $U(\mathbf{r})$ in Eq. (3) is detached, the network struggles to identify dynamic regions, which degrades both CNN-based and NeRF-based uncertainty estimation.

In this study, we propose a joint optimization framework in which the semantics-aware uncertainty estimated by a CNN-based network complements the geometry-aware uncertainty predicted by NeRF. An alternative configuration treats the CNN-based network as a teacher and NeRF as a student in a unidirectional distillation scheme. For clarity, we restate the formulations of online distillation and joint optimization from the main text:

$$\mathcal{L}_{\text{distill}}(\mathbf{r}) = \frac{\|C(\mathbf{r}) * B_{11\times 11} - C^s(\mathbf{r})\|_2^2}{2U(\mathbf{r})^2} + \log U(\mathbf{r}), \tag{2}$$

$$\mathcal{L}_{\text{joint}}(\mathbf{r}) = \|\beta(\mathbf{r}) - U(\mathbf{r})\|_2^2 . \tag{3}$$

To validate that joint optimization also benefits CNN-based uncertainty estimation, we conducted the following experiments: (i) No detachment is applied to either tensor in Eq. (3), allowing both the NeRF and the CNN-based network to be jointly optimized (Fig. B(a)); (ii) Detachment is applied to the CNN-based uncertainty $U(\mathbf{r})$ in Eq. (3), such that the CNN-based network is no longer influenced by the loss (Fig. B(b)). In the second setting, the CNN is trained solely via the aleatoric uncertainty loss defined in Eq. (2) during online distillation (NeRF-to-CNN). As shown in Fig. B(b), $U(\mathbf{r})$ often exhibits spatially non-localized and dispersed activations. This degraded component, when used to supervise NeRF in the reverse direction of distillation (CNN-to-NeRF), negatively affects the estimation of NeRF-based uncertainty $\beta(\mathbf{r})$ and transient weight sum $w^t(\mathbf{r})$.

This issue stems from the limitations of the aleatoric uncertainty loss. Since the network relies on reconstruction errors between the predicted static image $C^s(\mathbf{r})$ and the input image $C(\mathbf{r})$, it struggles to estimate meaningful uncertainty when the signal is weak or ambiguous. As a result,

$U(\mathbf{r})$ may fail to highlight dynamic regions or become overly diffuse. This ill-posed behavior of aleatoric uncertainty has also been theoretically discussed in prior work [1].

By jointly leveraging both semantics-aware and geometry-aware uncertainty, our framework facilitates mutual refinement between NeRF and the CNN-based uncertainty network. This synergy not only improves the quality of NeRF's uncertainty prediction but also contributes to the stable training of the CNN-based network.

3 Effect of Patch Size on TV Loss

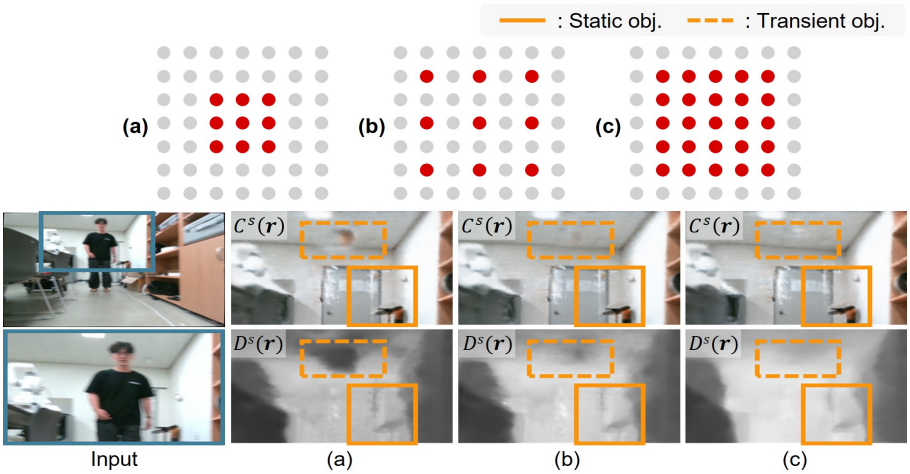

Figure C: **Effect of TV loss patch configuration on rendering.** (a) Less effective noise suppression with **a plain 11 × 11 patch**. (b) Effective noise suppression and preservation of geometric details with **a dilated 11 × 11 patch**. (c) High computational cost and loss of geometric fidelity with **a plain 44 × 44 patch**.

Even after joint optimization, subtle noise may remain in the static field. To mitigate this issue, we apply TV loss to the predicted static depth $D^s(\mathbf{r})$. For effective suppression of residual artifacts, each patch should ideally include both the static background and noise, with the background occupying the majority of the region. However, using an excessively large receptive field increases the computational cost per iteration and may degrade training efficiency. To strike a balance between receptive field size and computational efficiency, we adopt an 11×11 patch size with a dilation factor of 4.

We conducted a series of experiments with different patch configurations to analyze the effect of patch size on TV loss. As shown in Fig. C, using a plain 11×11 patch without dilation (Fig. C(a)) is less effective at reducing noise. In contrast, a large 44×44 patch (Fig. C(c)) reduces efficiency and also harms geometric details. Our proposed configuration (Fig. C(b)) effectively suppresses noise while preserving geometric fidelity.

4 Ablation Study

We conducted an ablation study on two datasets, Replica and Bonn, using both quantitative metrics and qualitative evaluations as demonstrated in Fig. D and Table 1. In Replica, GT is available even for regions occluded by dynamic objects, allowing direct evaluation of dynamic content removal. In contrast, Bonn lacks such GT, so the evaluation focuses on how well fine details are preserved in visible static areas. The results for each component are summarized as follows:

- **Joint optimization:** As shown in Replica, joint optimization effectively suppresses perturbation artifacts introduced by noisy observations in input images from other viewpoints.

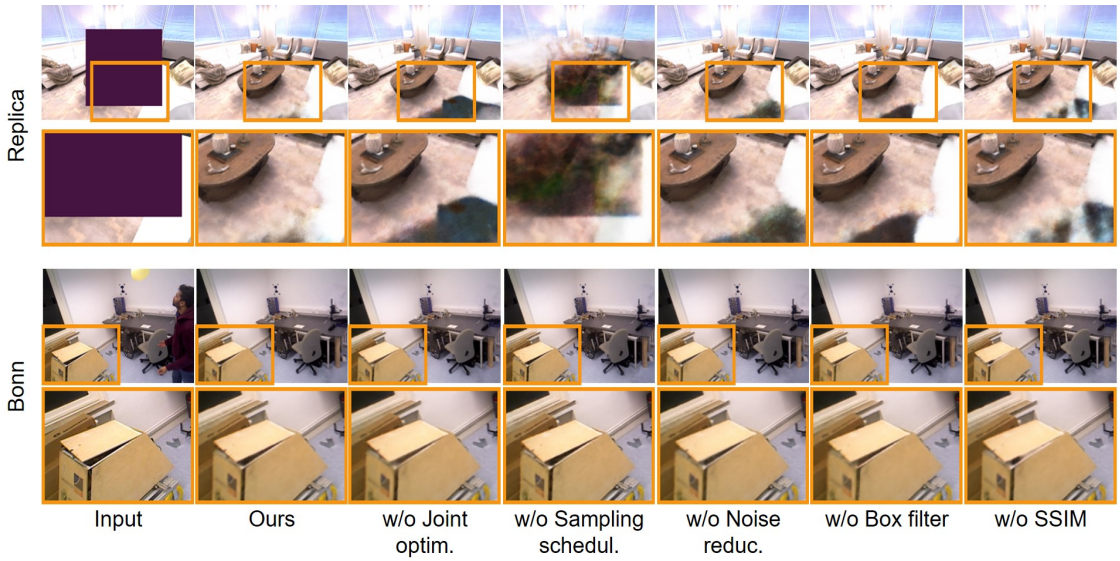

Figure D: **Ablation study of key components in our method.** We analyze the effect of excluding individual components, including joint optimization, sampling scheduling, noise reduction, box filtering for stable distillation, and SSIM loss, on the performance of dynamic object removal and static scene reconstruction.

Table 1: Quantitative Evaluation of the Ablation Study

	Ours	Joint optim.	Sampling schedul.	Noise reduc.	Box filter	SSIM
LPIPS ↓	0.250 / 0.153	0.262 / 0.155	0.399 / 0.143	0.247 / 0.154	0.240 / 0.154	0.288 / 0.161
SSIM ↑	0.892 / 0.910	0.885 / 0.910	0.807 / 0.913	0.894 / 0.910	0.896 / 0.909	0.876 / 0.908
PSNR ↑	25.94 / 29.27	24.51 / 29.13	17.27 / 29.48	25.52 / 29.23	25.49 / 28.93	25.14 / 28.81

Results are reported as Replica / Bonn.

Likewise, the results on Bonn show that it preserves fine details without degrading rendering quality.

- **Sampling scheduling:** As observed in Replica, applying data-driven sampling from the early stages of training leads to premature convergence, often resulting in a failure to remove dynamic objects. Meanwhile, Bonn shows that applying data-driven sampling in the later stages of training yields rendering quality comparable to early application, suggesting that high-fidelity results can still be achieved without early sampling.
- **Noise reduction:** Applying TV loss effectively suppresses residual noise, as observed in Replica, while results on Bonn confirm that it does not noticeably degrade fine geometry or overall rendering quality.
- **Box filter for stabilizing distillation:** Incorporating a box filter into the input image $C(\mathbf{r})$ used in the online distillation loss in Eq. (2) mitigates early-stage supervision artifacts caused by coarse geometry. Without filtering, sharp boundaries in the input may be misinterpreted as uncertainty, causing nearby static regions to be incorrectly assigned to the transient field. As shown in Replica, this leads to incomplete reconstruction of static backgrounds. A similar issue is observed in Bonn, particularly near the edges of static structures such as carts, where the use of a box filter results in more stable and accurate reconstructions.
- **Uncertainty-aware SSIM loss:** Excluding SSIM loss from training leads to notable degra-

dation in reconstruction quality, with static background restoration deteriorating in Replica and fine detail recovery reduced in Bonn.

5 Robust Dynamic Object Removal under Appearance Changes

Figure E: **Dynamic object removal under varying illumination conditions.** Compared to prior methods, our method effectively removes dynamic objects across diverse lighting conditions.

To simulate diverse illumination conditions, we applied pixel scaling and trained an image-specific appearance embedding vector that captures the lighting characteristics of each frame. The performance of dynamic object removal was evaluated under these varying illumination settings, as shown in Fig. E. Baseline methods (e.g., NeRF-W, GS-W, WG) fail to consistently remove dynamic objects under such variations. In particular, WG produces abnormal artifacts and noticeable degradation in rendering quality.

This limitation stems from the entanglement of illumination and dynamic content within the appearance embedding space. While these methods learn a per-image appearance embedding solely to model lighting conditions, dynamic objects may inadvertently be encoded into the embedding because they also exhibit image-dependent properties. Such entanglement hinders the effective separation of transient content from the static background.

In contrast, our method demonstrates robust dynamic object removal under various lighting conditions. This robustness arises from our curriculum learning strategy, which facilitates stable training and enables the network to learn both appearance representation and dynamic object removal effectively.

6 Evaluating the Generalization of the Uncertainty Network

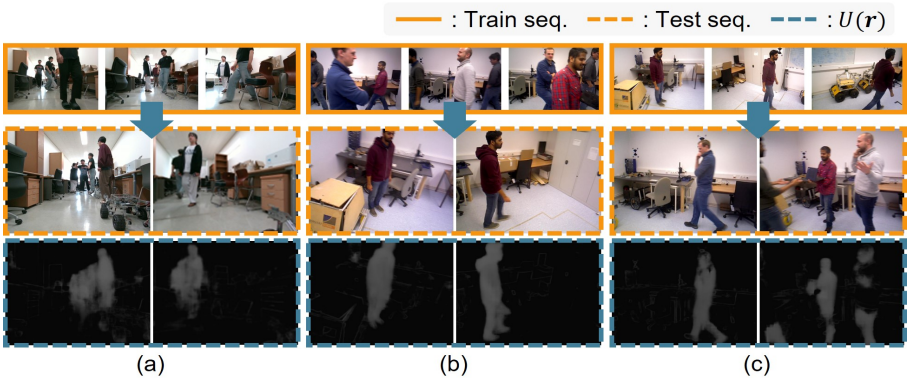

Figure F: **Generalization performance of the uncertainty network.** (a) Accurate estimation when dynamic objects at test time match those seen during training. (b) Successful generalization to unseen dynamic objects enabled by diversity in the training set. (c) Conservative predictions resulting from training with only a single dynamic object.

We evaluated the robustness of our CNN-based uncertainty network when applied to a novel query sequence captured from different viewpoints of the same 3D scene. As illustrated in Fig. F, three types of evaluation were conducted as follows:

- **(a):** While the test sequence differs from the training sequence, it contains the same dynamic objects. In this case, the uncertainty network successfully identifies the dynamic regions, demonstrating strong consistency under viewpoint changes.
- **(b):** The test sequence includes dynamic objects that are not present in the training set. Thanks to the diversity of dynamic objects in the training data, the network generalizes well and accurately detects previously unseen instances.
- **(c):** The network is trained on a single type of dynamic object within the scene. In the test sequence, dynamic objects with appearances not seen during training, especially those with color distributions similar to the background, are not effectively captured. This highlights a limitation in conservative prediction when training diversity is insufficient.

Failure cases, as shown in Fig. F(c), typically occur when dynamic objects with novel appearances and low color contrast with the background are present in the test set. In such cases, the difference from the GT–static image is subtle, leading the network to underestimate uncertainty. This is due to the conservative nature of the network, which tends to assign high uncertainty only to clearly inconsistent regions and produces low uncertainty in ambiguous cases.

The goal of our CNN-based uncertainty network is not to detect specific dynamic objects such as humans, but to identify regions that deviate from the static background learned during NeRF training. Given this objective, addressing the limitation of conservative prediction requires more than simply relying on pretrained features to improve generalization toward dynamic objects. Instead, future work should enable the network to respond not only to clearly inconsistent regions but also to moderately uncertain areas that may still reflect meaningful differences from the static background.

References

[1] Y. Zhang, J. Lin, F. Li, Y. Adler, K. Rasul, A. Schneider, and Y. Nevmyvaka, "Risk bounds on aleatoric uncertainty recovery," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2023, pp. 6015–6036.