D^{2}NeRF: Self-Supervised Decoupling of Dynamic and Static Objects from a Monocular Video

Supplementary Material

A Code and Data

All coda and data, as well as additional video results can be found on our project page: d2nerf.github.io. We also include a static copy of the code and website as zip files in the supplementary.

B Hyperparameters

As we do not have large TPUs available for training, we incorporate a light-weighted HyperNeRF [40] as the dynamic component to reduce training time. Compared to the original hyperparameters described in the paper, we reduce the number of samples per ray (64 vs. 128 in [40]), batch size (1024 vs. 6144 in [40]), and number of iterations (100k vs. 250k in [40]). We also do not apply the background regularization, as it requires a set of known background 3D points, which rely on accurate dynamic masks during COLMAP registration.

We experimentally established a set of hyperparameters applicable for a variety of scenes. In total, we used five sets of hyperparameters for the evaluation on the real-world dataset, and four on the synthetic dataset. To ensure the various scenes are fully separated into different components, we increase λ_s during training, where $\rightarrow$ indicates the value is linearly increased and $\Rightarrow$ indicates it is exponentially increased. - entry for λ_ρ indicates that the shadow field is not applied.

Table 3: **Hyperparameters –** Row 1-4 specify hyperparameters for real-world scenes containing a mixture of dynamic objects and shadows, whereas row 5 is suitable for real-world scenes with dynamic shadows only. Row 6-9 contain hyperparameters for synthetic scenes.

	k	λ_s	λ_r	λ_{σ^S}	λ_ρ	Dataset
1	1.75	$1\mathrm{e}{-4} \rightarrow 1\mathrm{e}{-2}$	$1\mathrm{e}{-3}$	0	$1\mathrm{e}{-1}$	Broom, Chicken, Curls, Pick2, Duck, Balloon, Cookie, Hand, Shark, Toy
2	3	$1\mathrm{e}{-4} \Rightarrow 1$	$1\mathrm{e}{-3}$	0	$1\mathrm{e}{-1}$	Banana (novel view)
3	2.75	$1\mathrm{e}{-5} \Rightarrow 1$	$1\mathrm{e}{-3}$	0	-	Water, Banana (decoupling)
4	2.875	$5\mathrm{e}{-4} \Rightarrow 1$	0	0	-	Pick
5	1.5	$1\mathrm{e}{-3} \Rightarrow 1$	$1\mathrm{e}{-1}$	$1\mathrm{e}{-2}$	$1\mathrm{e}{-2}$	Camera Shadow, Shadow Car
6	2	$1\mathrm{e}{-5} \Rightarrow 1$	$1\mathrm{e}{-5}$	$1\mathrm{e}{-4}$	-	Cars, Soft
7	1.75	$1\mathrm{e}{-5} \Rightarrow 0.1$	$1\mathrm{e}{-4}$	0	-	Car
8	2.5	$1\mathrm{e}{-5} \Rightarrow 1$	$1\mathrm{e}{-5}$	$1\mathrm{e}{-3}$	-	Chairs
9	2.75	$1\mathrm{e}{-4} \Rightarrow 1$	$2\mathrm{e}{-4}$	$1\mathrm{e}{-4}$	-	Bag

C Scene Decoupling – Figure 11, Figure 13, Figure 14

We demonstrate additional qualitative results on scene decoupling task on both real-world and synthetic scenes; see Figure 11, 12, 13, and 14.

D Video Segmentation – Table 4, Figure 15

As our method learns a density distribution of the dynamic objects in the scene, we can further produce an alpha mask of the objects. We therefore also evaluate the correctness of object segmentation at the image level. Existing benchmarks on video segmentation [42, 56, 24] either contain too few video frames for reliable SfM reconstruction, or do not have correct ground truth masks for all of the dynamic objects and effects in the scene. Similarly, the dataset from NeuralDiff [54] focuses on egocentric videos and the over-exaggerated difference between frames in the videos is not suitable for HyperNeRF which we use as the dynamic component. Hence, to highlight the ability of our method to decouple dynamic objects and shadows from video sequences with large viewpoint shifts,

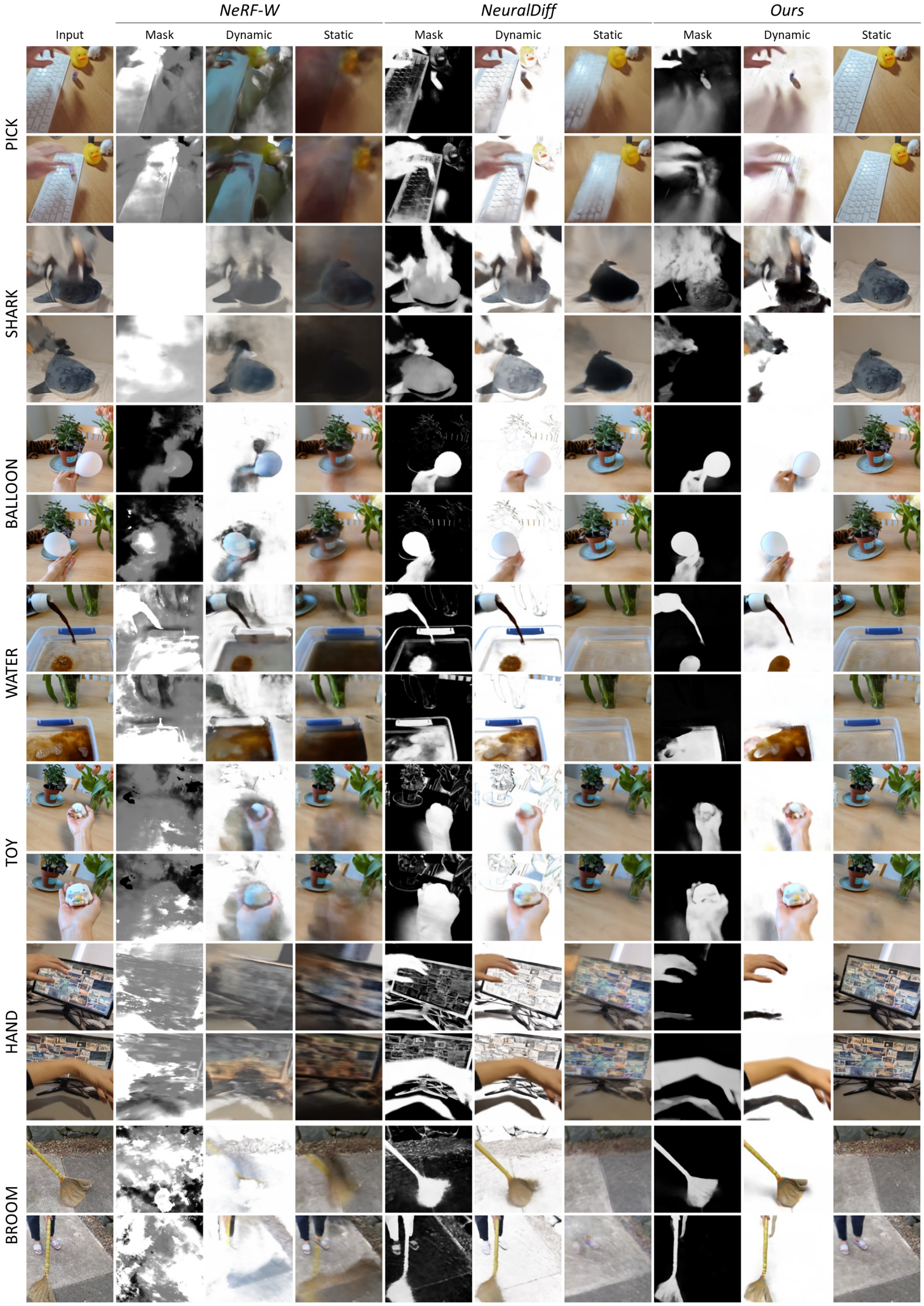

Figure 11: **Additional results on real-world scene decoupling and segmentation –** Similar to Figure 7, we show the dynamic alpha mask, dynamic part and static background respectively. Last two rows at bottom show the "broom" scene from HyperNeRF [40].

we evaluate on our synthetic dataset. In addition to the NeRF-like baselines, we also compare with Motion Grouping (MG) [64], a motion-based 2D image segmentation method. We fine-tuned the

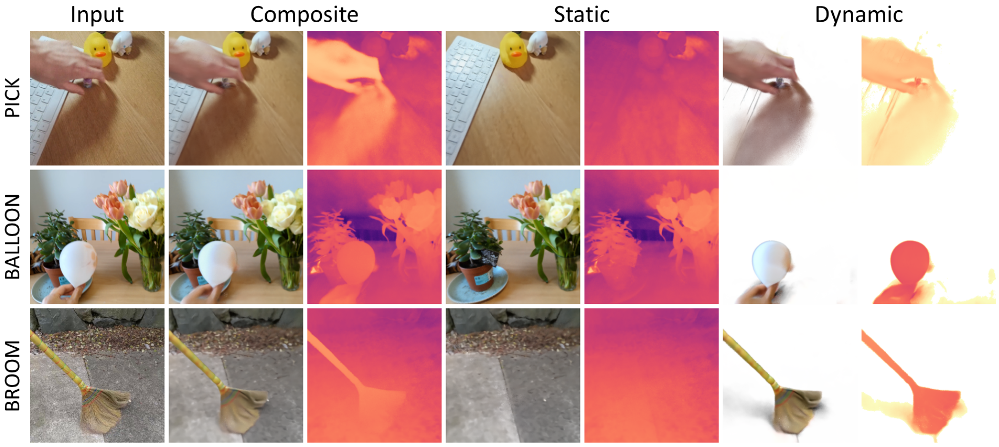

Figure 12: **Decoupled depth –** We show the disentangled geometry as depth maps for dynamic and static components respectively.

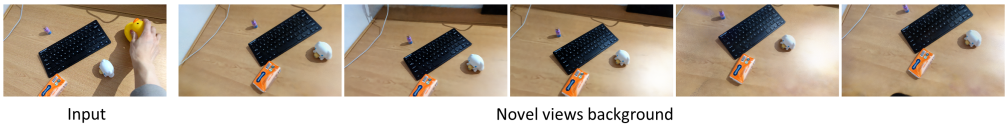

Figure 13: **Background novel view –** Our method learns the decoupled static background and can render it from unseen views, with the dynamic occluders cleanly removed.

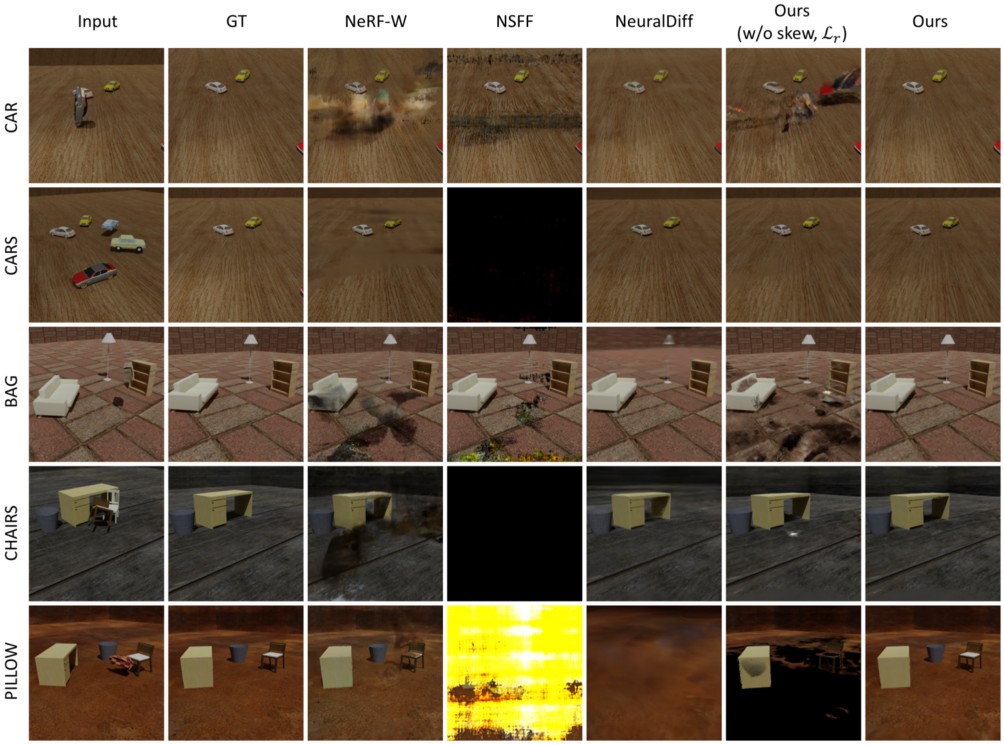

Figure 14: **Scene decoupling and novel view background recovery on synthetic scenes –** We train each method with videos containing various dynamic occluders and shadows, decouple the scene and render the background from unseen views to compare with the ground truth. Quantitative evaluation on corresponding scenes can be found at Table 1.

	Car		Cars		Bag		Chairs		Pillow		Mean	
	$\mathcal{J}\uparrow$	$\mathcal{F}\uparrow$	$\mathcal{J}\uparrow$	$\mathcal{F}\uparrow$	$\mathcal{J}\uparrow$	$\mathcal{F}\uparrow$	$\mathcal{J}\uparrow$	$\mathcal{F}\uparrow$	$\mathcal{J}\uparrow$	$\mathcal{F}\uparrow$	$\mathcal{J}\uparrow$	$\mathcal{F}\uparrow$
MG [64]	.603	.743	.363	.474	.629	.738	.484	.613	.044	.080	.424	.529
NeRF-W [31]	.072	.132	.098	.162	.027	.052	.154	.254	.194	.314	.109	.183
NSFF [26]	.083	.152	.058	.104	.102	.182	.046	.087	.104	.188	.079	.143
NeuralDiff [54]	.806	.891	.508	.578	.080	.144	.368	.513	.097	.177	.372	.461
Ours (w/o skew)	.814	.896	**.807**	**.883**	.342	.483	.114	.198	.347	.511	.485	.594
Ours (w/o L_r)	.076	.139	.174	.261	.048	.089	.237	.367	.040	.078	.115	.187
Ours (w/o skew, L_r)	.077	.142	.376	.464	.043	.081	.315	.453	.027	.053	.168	.238
Ours	**.848**	**.917**	.790	.874	**.703**	**.818**	**.551**	**.687**	**.693**	**.818**	**.717**	**.822**

Table 4: **Video segmentation –** We report Jaccard index $\mathcal{J}$ and boundary measure $\mathcal{F}$ on training views. Our method performs well on "car" and "cars" scenes without the use of skewed entropy because the background is clearly distinguishable from the moving object.

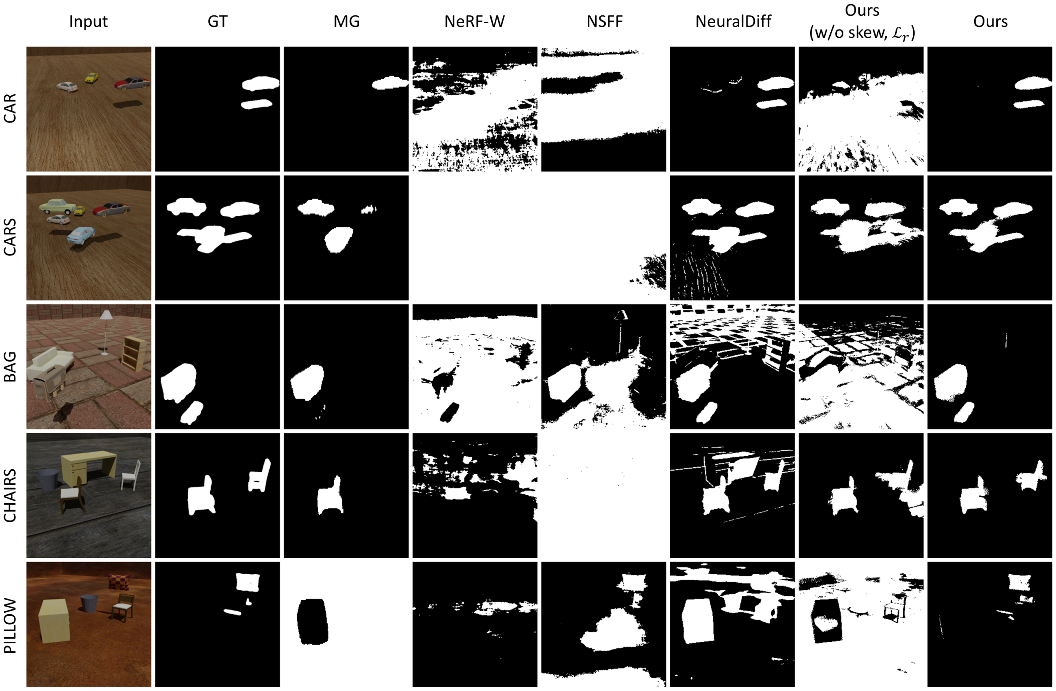

Figure 15: **Video segmentation (qualitative) –** MG fails to identify the moving pillow and segments everything out except for the table due to its Hungarian setting. NeRF-W learns the transient component with severe cloud-like effects. While our method achieves best segmentation in all the scenes.

pre-trained MG model on each scene for 5k iterations. For other NeRF-based methods, we used the same settings as in Section 4.4 and produced the alpha masks as the normalized radiance weights of the time-varying component, and then applied a threshold of 0.1 to obtain the binary masks. See Table 4, Figure 15 for the results.

E Novel View Synthesis – Table 5, Figure 16

Although the aim of our method is not to improve the quality of time-varying scene reconstruction, as a by-product, we find that by introducing a static component to fully utilize the network capacity, our method achieves more robust reconstruction for both the dynamic objects and background. We therefore also evaluate our method on the ability to synthesize the whole scene from novel views. We compare several approaches for dynamic scene reconstruction, including NeuralDiff [54] and a baseline version of HyperNeRF [40], which has the same architecture as our dynamic component. Since our method uses an additional static component and naturally has more network parameters and capacity, we also compare with a fair version of HyperNeRF with roughly the same number of total parameters by extending the NeRF MLP width to 375. Unlike novel view experiments in [40], we do not interleave between two cameras, but use only the right camera as training view and the left

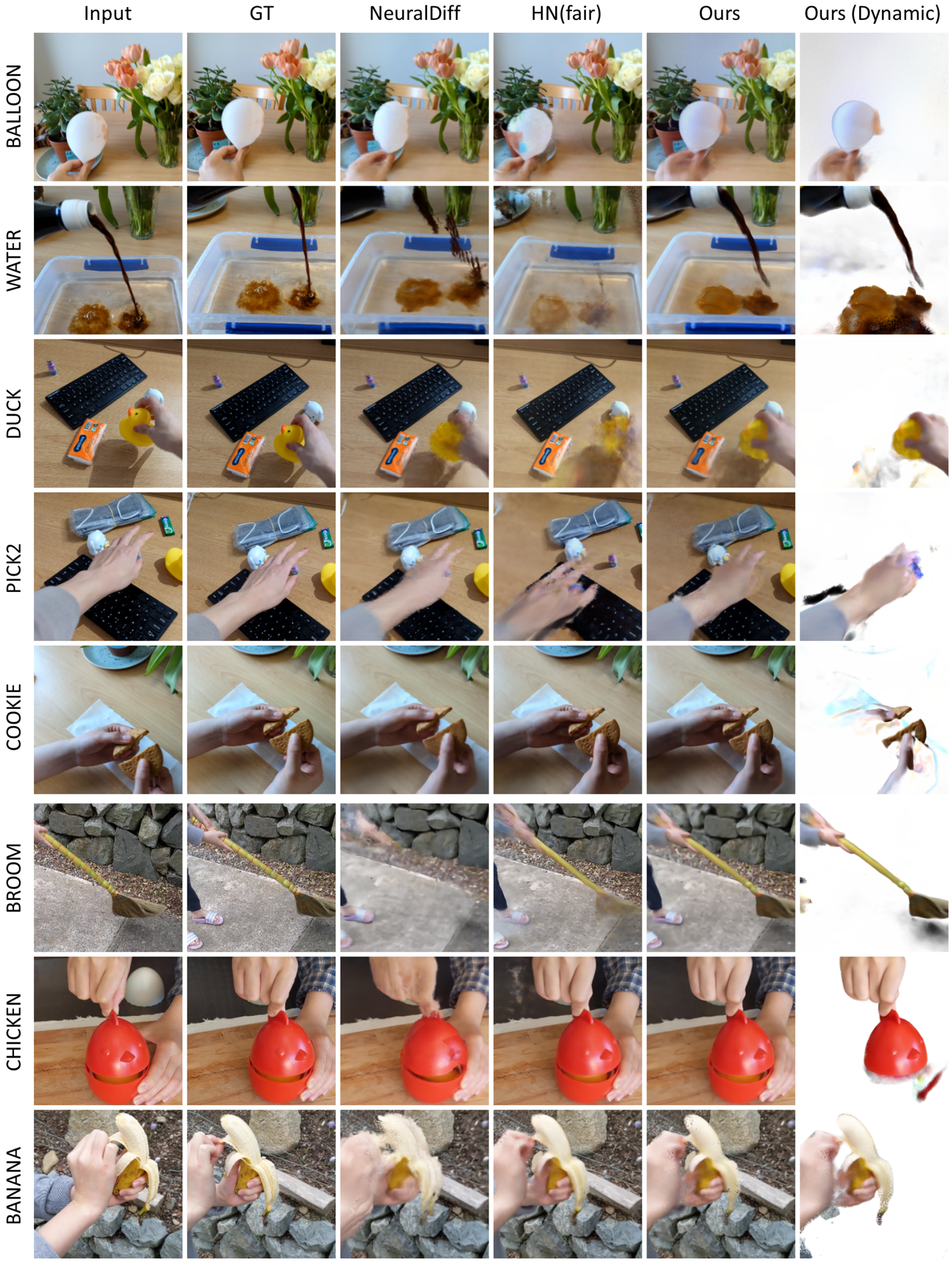

Figure 16: **Novel view synthesis (qualitative) –** For challenging scenes such as "water" and "duck" where the dynamic object moves rapidly or training/validation views differ largely, HyperNeRF [40] fails to reconstruct a reasonable shape for the dynamic object, while ours might potentially predict a shifted object pose, but can still render the view with high fidelity. We additionally show the decoupled dynamic object from our method. The quality is slightly degraded compared to the decoupling results in Figure 7 and 11 as we are rendering from the more challenging novel views.

camera as validation view. This presents greater challenges to all the methods. See Table 5, Figure 16 for the results.

	Pick2			Duck			Balloon			Water			Cookie		
	LPIPS↓	MS-SSIM↑	PNSR↑	LPIPS↓	MS-SSIM↑	PNSR↑	LPIPS↓	MS-SSIM↑	PNSR↑	LPIPS↓	MS-SSIM↑	PNSR↑	LPIPS↓	MS-SSIM↑	PNSR↑
NeuralDiff	**.208**	**.853**	**21.81**	.222	**.862**	21.92	.167	.836	20.13	.172	.811	18.36	.159	.875	**20.53**
HN (base)	.496	.413	13.06	.251	.830	20.64	.195	.803	17.81	.360	.483	15.06	.161	.836	19.93
HN (fair)	.486	.409	13.14	.253	.818	20.32	.187	.804	17.92	.361	.465	14.80	.162	.801	19.75
Ours	.253	.825	20.32	**.214**	.856	**22.07**	**.153**	**.858**	**20.92**	**.153**	**.849**	**21.63**	**.156**	**.877**	19.93

	Broom			Chicken			Banana			Mean		
	LPIPS↓	MS-SSIM↑	PNSR↑	LPIPS↓	MS-SSIM↑	PNSR↑	LPIPS↓	MS-SSIM↑	PNSR↑	LPIPS↓	MS-SSIM↑	PNSR↑
NeuralDiff	.631	.468	17.75	.249	.822	21.17	.303	.748	19.43	.264	.784	20.14
HN (base)	.524	.636	19.65	.222	.878	23.94	.223	.818	21.20	.304	.712	18.91
HN (fair)	**.503**	.624	19.38	**.180**	.881	23.68	**.194**	**.832**	**21.52**	.291	.704	18.82
Ours	.565	**.712**	**20.66**	.204	**.890**	**24.27**	.260	.820	21.35	**.245**	**.836**	**21.39**

Table 5: **Novel view synthesis (quantitative) –** We compare with NeuralDiff [54], a baseline version of HyperNeRF [40], denoted HN (base), and a fair version with extended network width to match the total number of parameters in our method, denoted HN (fair). Three scenes displayed in the bottom row are from HyperNeRF[40]

F Ambiguity between Dynamic Component and Shadow

The shadow field network represents the density-less shadows in a more physically realistic way, and resolves the ambiguity in their motion. However, the aim of our method is to achieve decoupling of dynamic occluders from the static environment, and we do not over-extend to consider further decoupling between objects and shadows, which would require more priors related to environmental lighting conditions and background texture.

We empirically found that shadow field is not necessary for scenes with strong and fast-moving shadows, where they can be directly learned by the dynamic component as thin layers on top of the static geometry; see Figure 17. On the other hand, there exists unsolvable ambiguity between the shadow and dynamic object, especially for which with a similar or darker color to the background, and hence could be potentially explained as a moving shadow instead of an actual 3D shape due to our monocular camera setting; see Figure 18. As the later case causes failed dynamic geometry reconstruction, leading to a severe decrease in novel view synthesis performance, we deliberately choose a large value of λ_ρ to suppress the shadow field for scenes containing a mixture of dynamic objects and shadows. Although this potentially favors the former case and causes more shadows to be interpreted as thin-layers, we found that such setting has minimal impact on performance of both scene decoupling and reconstruction, and is still sufficient to achieve correct shadow decoupling, as the shadow field resolves the ambiguity in shadow in very early stage of the training.

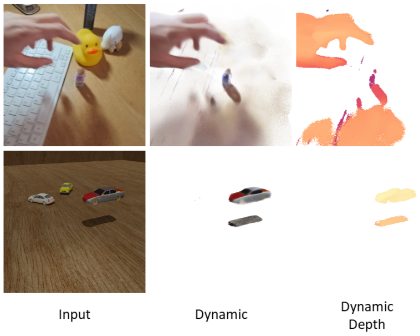

Figure 17: **Shadow as thin layer –** Although such representation could potentially represent more than just shadows, it still tends to exclude unnecessary texture from static background and learns only the darkening effects.

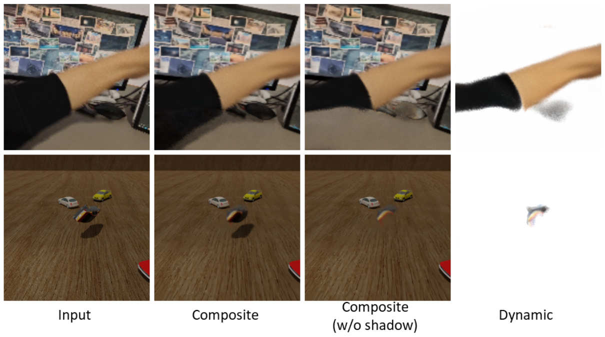

Figure 18: **Incorrect shadow –** The shadow field is incorrectly used to explain the black sleeve as well as the gray top of the moving car.