# Structure from Collision

## Supplementary Material

Takuhiro Kaneko
NTT Corporation

https://www.kecl.ntt.co.jp/people/kaneko.takuhiro/projects/sfc/

## Contents

# A. Detailed analyses and discussions

## A.1. Detailed ablation studies

Owing to space limitations in the main text, we conducted an ablation study focusing only on the selected key components. In this appendix, we present the detailed ablation studies to further assess the effectiveness of the proposed method from multiple perspectives. Specifically, we examine the effects of *each appearance-preserving loss* (Appendix A.1.1), *keyframe selection* (Appendix A.1.2), and *background loss* (Appendix A.1.3).

### A.1.1. Effect of each appearance-preserving loss

As explained in Section 3.3 regarding appearance-preserving constraints, we adopt two appearance-preserving losses (APL): the pixel-preserving loss $\mathcal{L}_{\text{pixel}_0}$ (Equation 9) and the depth-preserving loss $\mathcal{L}_{\text{depth}_0}$ (Equation 10). These losses help prevent the degradation of the external structure, which is effectively learned from the first frame of the video sequence, during the fitting process across the entire video sequence. In the ablation study presented in Sections 4.2 and 4.3, we ablated both losses simultaneously to examine the overall effect of APL. In a more detailed ablation study, we assessed the performance when each of the appearance-preserving losses was individually ablated.

| $\mathcal{L}_{\text{pixel}_0}$ | $\mathcal{L}_{\text{depth}_0}$ | 0 | $(\frac{1}{2})^3$ | $(\frac{2}{3})^3$ | $(\frac{3}{4})^3$ | Avg. |
|---|---|---|---|---|---|---|
| | | 0.106 | 0.423 | 0.898 | 1.326 | 0.688 |
| ✓ | | 0.105 | 0.142 | 0.334 | 0.342 | 0.231 |
| | ✓ | 0.079 | 0.313 | 0.314 | 0.287 | 0.248 |
| ✓ | ✓ | 0.081 | 0.122 | 0.195 | 0.262 | 0.165 |

Table 7. Results of the detailed ablation study of appearance-preserving losses when the cavity size $s_c$ is varied. The score indicates CD ($\times 10^3 \downarrow$). A checkmark ✓ indicates that the corresponding loss was used.

| $\mathcal{L}_{\text{pixel}_0}$ | $\mathcal{L}_{\text{depth}_0}$ | left | right | up | down | Avg. |
|---|---|---|---|---|---|---|
| | | 0.845 | 0.783 | 0.805 | 0.583 | 0.754 |
| ✓ | | 0.295 | 0.451 | 0.325 | 0.311 | 0.345 |
| | ✓ | 0.362 | 0.299 | 0.348 | 0.389 | 0.349 |
| ✓ | ✓ | 0.303 | 0.258 | 0.274 | 0.291 | 0.281 |

Table 8. Results of the detailed ablation study of appearance-preserving losses when the cavity location $l_c$ is varied. The score indicates CD ($\times 10^3 \downarrow$). A checkmark ✓ indicates that the corresponding loss was used.

**Results.** Table 7 summarizes the results when the cavity size $s_c$ is varied, and Table 8 summarizes the results when the cavity location $l_c$ is varied. Our findings are threefold.

*(1) No APL vs. either of* $\mathcal{L}_{pixel_0}$ *and* $\mathcal{L}_{depth_0}$. Both SfC-NeRF with only $\mathcal{L}_{\text{pixel}_0}$ and SFC-NeRF with only $\mathcal{L}_{\text{depth}_0}$ outperformed SfC-NeRF without APL in all cases. These results indicate that both $\mathcal{L}_{\text{pixel}_0}$ and $\mathcal{L}_{\text{depth}_0}$ effectively enhance the performance of *SfC*.

*(2) Full APL vs. either of* $\mathcal{L}_{pixel_0}$ *and* $\mathcal{L}_{depth_0}$. SfC-NeRF with both $\mathcal{L}_{\text{pixel}_0}$ and $\mathcal{L}_{\text{depth}_0}$ outperformed SfC-NeRF with only $\mathcal{L}_{\text{pixel}_0}$ and SFC-NeRF with only $\mathcal{L}_{\text{depth}_0}$ in most cases. These results indicate that $\mathcal{L}_{\text{pixel}_0}$ and $\mathcal{L}_{\text{depth}_0}$ contribute to improving the performance of *SfC* from different perspectives, and they are most effective when used together.

*(3)* $\mathcal{L}_{pixel_0}$ *vs* $\mathcal{L}_{depth_0}$. The superiority or inferiority of each loss depends on the cavity setting. This is related to the

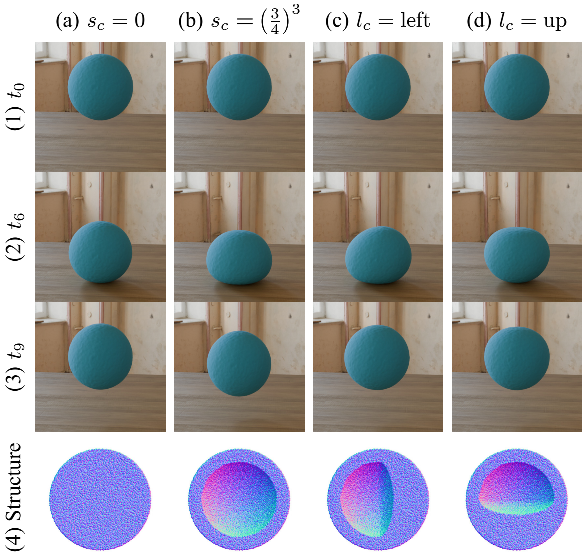

Figure 5. Comparison of appearances for objects having different internal structures when $t$ is varied within $\{t_0, t_6, t_9\}$.

learnability of the 3D appearance, and further detailed analyses will be an interesting direction for future research.

### A.1.2. Effect of keyframe selection

As discussed in Section 3.3 regarding the keyframe constraints, we employ a keyframe pixel loss $\mathcal{L}_{\text{pixel}_k}$ (Equation 11) to effectively capture shape changes caused by internal structures. Specifically, we selected the frame immediately after the collision as the keyframe ($k = 6$, where $k$ is the keyframe index) for the experiments described in the main text. An important question is whether this choice of $k$ is optimal. To investigate this, we evaluated the change in performance by varying the value of $k$, specifically within $\{6, 9\}$. Figure 5 shows a comparison of the appearances of objects with different internal structures in these keyframes. For reference, we also provided scores for the model without keyframe pixel loss (denoted as $k = None$).

**Results.** Table 9 summarizes the results when the cavity size $s_c$ is varied, and Table 10 summarizes the results when the cavity location $l_c$ is varied. Our findings are twofold.

*(1) $\mathcal{L}_{pixel_6}$ vs. $\mathcal{L}_{pixel_9}$.* SfC-NeRF with $\mathcal{L}_{\text{pixel}_6}$ outperformed that with $\mathcal{L}_{\text{pixel}_9}$ in most cases. As shown in Figure 5, immediately after the collision (at $t_6$ (2)), the difference in the shapes of the objects is noticeable. However, as time progressed after the collision (at $t_9$ (3)), the difference in the shapes of the objects diminished, whereas the difference in their positions became more pronounced. We consider this to be the main reason why SfC-NeRF with $\mathcal{L}_{\text{pixel}_6}$ performed better than that with $\mathcal{L}_{\text{pixel}_9}$.

*(2) $\mathcal{L}_{pixel_6}$/$\mathcal{L}_{pixel_9}$ vs. None.* We found that SfC-NeRF with $\mathcal{L}_{\text{pixel}_6}$ or $\mathcal{L}_{\text{pixel}_9}$ outperformed SfC-NeRF without the keyframe pixel loss in most cases. These results indicate that strategically weighing frames is more effective than treating all frames equally.

| $k$ | 0 | $(\frac{1}{2})^3$ | $(\frac{2}{3})^3$ | $(\frac{3}{4})^3$ | Avg. |
|---|---|---|---|---|---|
| None | 0.082 | 0.127 | 0.211 | 0.325 | 0.186 |
| 6 | 0.081 | 0.122 | 0.195 | 0.262 | 0.165 |
| 9 | 0.082 | 0.120 | 0.208 | 0.290 | 0.175 |

Table 9. Analysis of the effect of keyframe selection when the cavity size $s_c$ is varied. The score indicates CD ($\times 10^3 \downarrow$). When $k =$ None, the keyframe pixel loss $\mathcal{L}_{\text{pixel}_k}$ was not used. In contrast, when $k \in \{6, 9\}$, $\mathcal{L}_{\text{pixel}_k}$ was used.

| $k$ | left | right | up | down | Avg. |
|---|---|---|---|---|---|
| None | 0.308 | 0.296 | 0.307 | 0.313 | 0.306 |
| 6 | 0.303 | 0.258 | 0.274 | 0.291 | 0.281 |
| 9 | 0.296 | 0.296 | 0.313 | 0.303 | 0.302 |

Table 10. Analysis of the effect of keyframe selection when the cavity location $l_c$ is varied. The score indicates CD ($\times 10^3 \downarrow$). When $k =$ None, the keyframe pixel loss $\mathcal{L}_{\text{pixel}_k}$ was not used. In contrast, when $k \in \{6, 9\}$, $\mathcal{L}_{\text{pixel}_k}$ was used.

### A.1.3. Effect of background loss

As mentioned in the explanation of preprocessing in Section 4.1, we use a background loss $\mathcal{L}_{\text{bg}}$ by leveraging the fact that the background segmentation has been obtained. For example, when the background is excluded using a white color, this background loss helps distinguish whether the white area belongs to the background or a foreground object. This approach is not unrealistic because we use background segmentation that is not manually created but instead predicted from a given image using a DNN-based image matting model [6]. However, it is important and interesting to investigate the effectiveness of the background loss. To this end, we investigated the performance of *SfC-NeRF$_{-bg}$*, where the background loss ($\mathcal{L}_{\text{bg}}$) was ablated. In this setting, the performance of a model trained using only the first frame of the video sequence (Step (i) in Figure 2(a)) also changes because the background loss is also ablated in this step. We referred to this model as *Static$_{-bg}$*. We compared the scores of these models with those of the original models (i.e., *SfC-NeRF* and *Static*).

**Results.** Table 11 summarizes the results when cavity size $s_c$ is varied, and Table 12 summarizes the results when cavity location $l_c$ is varied. Our findings are twofold.

*(1) SfC-NeRF vs. SfC-NeRF$_{-bg}$.* SfC-NeRF outperformed SfC-NeRF$_{-\text{bg}}$ in most cases. As mentioned above, the background loss is useful for distinguishing background and foreground objects, allowing for more accurate capture of external structures. The movement of an object is affected by both its external and internal structures. Therefore, if the external structure can be estimated more accu-

| | 0 | $(\frac{1}{2})^3$ | $(\frac{2}{3})^3$ | $(\frac{3}{4})^3$ | Avg. |
|---|---|---|---|---|---|
| Static | 0.093 | 0.294 | 0.920 | 1.574 | 0.720 |
| SfC-NeRF | 0.081 | 0.122 | 0.195 | 0.262 | 0.165 |
| Static$_{-\text{bg}}$ | 0.093 | 0.290 | 0.906 | 1.545 | 0.708 |
| SfC-NeRF$_{-\text{bg}}$ | 0.101 | 0.149 | 0.222 | 0.279 | 0.188 |

Table 11. Results of the ablation study of background loss when the cavity size $s_c$ is varied. The score indicates CD ($\times 10^3 \downarrow$).

| | left | right | up | down | Avg. |
|---|---|---|---|---|---|
| Static | 0.841 | 0.842 | 0.815 | 0.813 | 0.828 |
| SfC-NeRF | 0.303 | 0.258 | 0.274 | 0.291 | 0.281 |
| Static$_{-\text{bg}}$ | 0.831 | 0.830 | 0.799 | 0.800 | 0.815 |
| SfC-NeRF$_{-\text{bg}}$ | 0.324 | 0.210 | 0.361 | 0.277 | 0.293 |

Table 12. Results of the ablation study of background loss when the cavity location $l_c$ is varied. The score indicates CD ($\times 10^3 \downarrow$).

rately, the internal structure can also be estimated more accurately.

*(2) SfC-NeRF$_{-bg}$ vs Static$_{-bg}$.* SfC-NeRF$_{-\text{bg}}$ outperformed Static $_{-\text{bg}}$ except when dealing with filled objects ($s_c = 0$ in Table 11).[5] These results indicate that the proposed method is effective in improving the performance of *SfC* even without the use of advanced techniques such as background loss.

## A.2. Extended experiments

### A.2.1. Experiment IV: Influence of collision angle

In the above experiments, the collision angle was fixed, as shown in Figures 6–13, regardless of the internal structure and physical properties, to focus on comparisons related to the internal structures and physical properties. For completeness, we investigated the influence of *collision angle* $\theta_c$ on the performance of *SfC*. Specifically, we selected objects with default settings ($s_c = (\frac{2}{3})^3$, $l_c$ = center, and elastic material defined by $\hat{E} = 1.0 \times 10^6$ and $\hat{\nu} = 0.3$) as the objects of investigation and examined their performance when only the collision angles were altered. The objects were rotated in the depth direction, as shown in Figure 14. The collision angle $\theta_c$ was chosen from $\{0°, 22.5°, 45°, 67.5°, 90°\}$. We compared the performance of *Static* and *SfC-NeRF*.

**Results.** Table 13 summarizes the quantitative results. Figure 14 presents the qualitative results. Our findings are twofold.

*(1) SfC-NeRF vs. Static.* SfC-NeRF outperformed the Static in all cases. These results indicate that optimizing the internal structure through a video sequence using the proposed method is beneficial, regardless of the collision angle.

| Sphere | 0° | 22.5° | 45° | 67.5° | 90° |
|---|---|---|---|---|---|
| Static | 1.164 | 1.163 | 1.163 | 1.162 | 1.160 |
| SfC-NeRF | 0.067 | 0.068 | 0.066 | 0.067 | 0.066 |
| Cube | 0° | 22.5° | 45° | 67.5° | 90° |
| Static | 0.775 | 0.776 | 0.848 | 0.768 | 0.776 |
| SfC-NeRF | 0.201 | 0.173 | 0.627 | 0.201 | 0.201 |
| Bicone | 0° | 22.5° | 45° | 67.5° | 90° |
| Static | 0.933 | 0.925 | 0.918 | 0.921 | 0.926 |
| SfC-NeRF | 0.144 | 0.194 | 0.187 | 0.146 | 0.154 |
| Cylinder | 0° | 22.5° | 45° | 67.5° | 90° |
| Static | 0.891 | 0.905 | 0.915 | 0.905 | 0.964 |
| SfC-NeRF | 0.342 | 0.288 | 0.311 | 0.209 | 0.639 |
| Diamond | 0° | 22.5° | 45° | 67.5° | 90° |
| Static | 0.837 | 0.830 | 0.833 | 0.819 | 0.838 |
| SfC-NeRF | 0.220 | 0.300 | 0.222 | 0.163 | 0.209 |

Table 13. Comparison of CD ($\times 10^3 \downarrow$) when collision angle $\theta_c$ is varied.

*(2) Effect of collision angle.* We found that the collision angle influenced the performance of *SfC*. The strength of this effect depends on the object shape. There are three possible reasons for this performance variation: *(i) Changes in the estimation accuracy of external structures.* The internal structure was optimized under the constraint that the external structure, learned from the first frame, should be maintained. Therefore, when the accuracy of the external structure estimation changed, the accuracy of the internal structure estimation also changed. *(ii) Difference in the amount of deformation.* The amount of deformation varied depending on the collision angle. This factor also affected the ease of estimating the internal structure. *(iii) Asymmetry.* When an object was not symmetrical relative to the collision angle, its behavior after the collision became asymmetrical. Consequently, the ease of estimating the internal structure also becomes asymmetric.

## A.3. Evaluation from multiple perspectives

### A.3.1. Evaluation through video sequences

In the main experiments, we evaluated the models using the chamfer distance between the ground-truth particles $\hat{\mathcal{P}}^P(t_0)$ and the estimated particles $\mathcal{P}^P(t_0)$ in the *first frame* of the video sequence, i.e., at $t = t_0$. For the multidimensional analysis, we investigated the chamfer distance between the ground-truth particles $\hat{\mathcal{P}}^P(t)$ and the estimated particles $\mathcal{P}^P(t)$, averaged over the *entire video sequence*, i.e., $t \in \{t_0, \dots, t_{N-1}\}$. For clarity, we refer to the former (chamfer distance for the first static frame) as *CD$_{static}$* and

[5] When dealing with a filled object, inaccurate estimation of external structure is problematic because it leads to a discrepancy between the actual mass and the estimated mass. In this situation, if the estimated mass is encouraged to approach the ground truth mass through mass loss while maintaining the external appearance with appearance-preserving losses, the internal structure may be altered unnecessarily. As a result, SfC-NeRF$_{-\text{bg}}$ degrades the performance of *SfC* when handling filled objects. An accurate estimation of the external structure, aided by background loss, effectively addresses this issue.

|  | 0 |  | $(\frac{1}{2})^3$ |  | $(\frac{2}{3})^3$ |  | $(\frac{3}{4})^3$ |  | Avg. |  |
|---|---|---|---|---|---|---|---|---|---|---|
| Static | 0.093 | 0.104 | 0.294 | 0.309 | 0.920 | 1.057 | 1.574 | 1.964 | 0.720 | 0.859 |
| GO | 0.091 | 0.092 | 0.301 | 0.301 | 0.941 | 0.944 | 1.586 | 1.612 | 0.730 | 0.737 |
| GO$_{\text{mass}}$ | 0.081 | 0.083 | 0.319 | 0.325 | 1.244 | 1.266 | 2.291 | 2.367 | 0.984 | 1.010 |
| LPO | 0.092 | 0.091 | 0.284 | 0.282 | 0.841 | 0.833 | 1.406 | 1.380 | 0.656 | 0.646 |
| LPO$_{\text{mass}}$ | 0.087 | 0.087 | 0.284 | 0.283 | 0.876 | 0.868 | 1.477 | 1.451 | 0.681 | 0.672 |
| SfC-NeRF$_{-\text{mass}}$ | 0.089 | 0.090 | 0.226 | 0.225 | 0.550 | 0.544 | 1.148 | 1.112 | 0.503 | 0.493 |
| SfC-NeRF$_{-\text{APL}}$ | 0.106 | 0.108 | 0.423 | 0.421 | 0.898 | 0.886 | 1.326 | 1.307 | 0.688 | 0.680 |
| SfC-NeRF$_{-\text{APT}}$ | 0.085 | 0.101 | 0.261 | 0.279 | 0.332 | 0.337 | 0.661 | 0.680 | 0.335 | 0.349 |
| SfC-NeRF$_{-\text{key}}$ | 0.082 | 0.086 | 0.127 | 0.131 | 0.211 | 0.213 | 0.325 | 0.325 | 0.186 | 0.189 |
| SfC-NeRF$_{-\text{VA}}$ | 0.146 | 0.269 | 0.293 | 0.338 | 0.370 | 0.407 | 0.456 | 0.485 | 0.316 | 0.375 |
| SfC-NeRF | 0.081 | 0.085 | 0.122 | 0.126 | 0.195 | 0.196 | 0.262 | 0.258 | 0.165 | 0.166 |

Table 14. Comparison of CD ($\times 10^3 \downarrow$) when the cavity size $s_c$ is varied. This is an extended table of Table 1. For each condition, the left score indicates CD$_{\text{static}}$, the chamfer distance between $\mathcal{P}(t_0)$ and $\hat{\mathcal{P}}(t_0)$ at the first frame, i.e., $t = t_0$, and the right score indicates CD$_{\text{video}}$, the chamfer distance between $\mathcal{P}(t)$ and $\hat{\mathcal{P}}(t)$ averaged over the entire video sequence, i.e., $t \in \{t_0, \dots, t_{N-1}\}$.

|  | left |  | right |  | up |  | down |  | Avg. |  |
|---|---|---|---|---|---|---|---|---|---|---|
| Static | 0.841<br>(0.841) | 1.159<br>(1.294) | 0.842<br>(0.843) | 1.306<br>(1.154) | 0.815<br>(0.814) | 1.731<br>(1.246) | 0.813<br>(0.813) | 1.241<br>(1.727) | 0.828<br>(0.828) | 1.359<br>(1.355) |
| GO | 0.874<br>(0.872) | 0.879<br>(2.606) | 0.853<br>(0.856) | 0.870<br>(2.549) | 0.878<br>(0.881) | 0.875<br>(1.471) | 0.870<br>(0.870) | 1.035<br>(1.673) | 0.869<br>(0.870) | 0.915<br>(2.075) |
| GO$_{\text{mass}}$ | 1.349<br>(1.340) | 1.386<br>(3.134) | 1.334<br>(1.344) | 1.375<br>(3.126) | 1.104<br>(1.127) | 1.141<br>(1.866) | 1.001<br>(1.004) | 1.370<br>(1.805) | 1.197<br>(1.204) | 1.318<br>(2.483) |
| LPO | 0.791<br>(0.802) | 0.789<br>(2.493) | 0.787<br>(0.800) | 0.787<br>(2.507) | 0.796<br>(0.819) | 0.776<br>(1.468) | 0.743<br>(0.737) | 0.721<br>(1.471) | 0.779<br>(0.790) | 0.768<br>(1.985) |
| LPO$_{\text{mass}}$ | 0.824<br>(0.833) | 0.822<br>(2.529) | 0.817<br>(0.832) | 0.818<br>(2.556) | 0.828<br>(0.847) | 0.806<br>(1.497) | 0.775<br>(0.771) | 0.753<br>(1.538) | 0.811<br>(0.821) | 0.800<br>(2.030) |
| SfC-NeRF$_{-\text{mass}}$ | 0.513<br>(0.858) | 0.520<br>(2.502) | 0.485<br>(0.878) | 0.491<br>(2.661) | 0.705<br>(0.747) | 0.689<br>(1.506) | 0.479<br>(0.956) | 0.457<br>(1.762) | 0.545<br>(0.860) | 0.539<br>(2.108) |
| SfC-NeRF$_{-\text{APL}}$ | 0.845<br>(1.069) | 0.840<br>(2.885) | 0.783<br>(1.083) | 0.788<br>(2.943) | 0.805<br>(0.934) | 0.786<br>(1.764) | 0.583<br>(0.883) | 0.580<br>(1.750) | 0.754<br>(0.992) | 0.749<br>(2.335) |
| SfC-NeRF$_{-\text{APT}}$ | 0.624<br>(0.588) | 0.631<br>(1.920) | 0.428<br>(0.586) | 0.604<br>(1.486) | 0.384<br>(0.579) | 0.461<br>(1.196) | 0.464<br>(0.646) | 0.514<br>(1.305) | 0.475<br>(0.600) | 0.553<br>(1.477) |
| SfC-NeRF$_{-\text{key}}$ | 0.308<br>(0.372) | 0.307<br>(1.854) | 0.296<br>(0.396) | 0.326<br>(1.746) | 0.307<br>(0.387) | 0.306<br>(1.291) | 0.313<br>(0.389) | 0.343<br>(1.105) | 0.306<br>(0.386) | 0.321<br>(1.499) |
| SfC-NeRF$_{-\text{VA}}$ | 0.542<br>(0.639) | 0.611<br>(2.304) | 0.596<br>(0.757) | 0.767<br>(2.265) | 0.333<br>(0.445) | 0.389<br>(1.338) | 0.385<br>(0.549) | 0.421<br>(1.339) | 0.464<br>(0.597) | 0.547<br>(1.811) |
| SfC-NeRF | 0.303<br>(0.367) | 0.308<br>(1.821) | 0.258<br>(0.431) | 0.313<br>(1.647) | 0.274<br>(0.448) | 0.273<br>(1.262) | 0.291<br>(0.417) | 0.307<br>(1.204) | 0.281<br>(0.416) | 0.300<br>(1.483) |

Table 15. Comparison of CD and ACD ($\times 10^3 \downarrow$) when the cavity location $l_c$ is varied. This is an extended table of Table 2. For each condition, the left score indicates CD$_{\text{static}}$, the chamfer distance between $\mathcal{P}(t_0)$ and $\hat{\mathcal{P}}(t_0)$ at the first frame, i.e., $t = t_0$, and the right score indicates CD$_{\text{video}}$, the chamfer distance between $\mathcal{P}(t)$ and $\hat{\mathcal{P}}(t)$ averaged over the entire video sequence, i.e., $t \in \{t_0, \dots, t_{N-1}\}$. The gray score in parenthesis indicates the ACD. For each condition, the left score indicates ACD$_{\text{static}}$, the anti-chamfer distance at the first frame, and the right score indicates ACD$_{\text{video}}$, the anti-chamfer distance averaged over the entire video sequence. It is expected that each original CD is smaller than the corresponding ACD.

the latter (chamfer distance for the entire video sequence) as *CD$_{video}$*. In the evaluation of the influence of cavity location (Section 4.3), we introduce anti-chamfer distance, which is the chamfer distance between the predicted particles $\mathcal{P}^P(t_0)$ and the ground truth particles $\tilde{\mathcal{P}}^P(t_0)$, where the cavity is placed on the opposite side, in the *first frame* of the video sequence to evaluate how well the cavity location is captured. For further analysis, we calculated and averaged similar scores for the *entire video sequence*. For clarity, we refer to the former (anti-chamfer distance for the first static frame) as *ACD$_{static}$* and the latter (anti-chamfer distance for the entire video sequence) as *ACD$_{video}$*.

**Results.** Table 14 summarizes the results when the cavity size $s_c$ is varied, and Table 15 summarizes the results when the cavity location $l_c$ is varied. Our findings are fourfold.

*(1) CD$_{static}$ vs. CD$_{video}$.* The relative values of CD$_{\text{static}}$ and CD$_{\text{video}}$ vary across different cases. When calculating CD$_{\text{static}}$ in the first frame, the locations of the ground-truth objects and those of the synthesized objects were well aligned, allowing for a focus on differences in shapes.

In contrast, when calculating CD$_{\text{video}}$ for the entire video sequence, we need to consider not only the differences in shapes but also the differences in absolute locations. Misalignments accumulate over time because the locations must vary within the allowance of the physical constraints via DiffMPM [2]. Since the objective of this study was to correctly predict the shape, rather than the location, CD$_{\text{static}}$ is a more valid evaluation than CD$_{\text{video}}$ for this purpose.

*(2) Comparison of CD$_{static}$ and CD$_{video}$ among models.* Although there is some variation in the superiority of the models depending on the metric used, the general trend remains consistent: SfC-NeRF achieves the best score in most cases. The two exceptions are CD$_{\text{video}}$ for $s_c = 0$ in Table 14 and CD$_{\text{video}}$ for $l_c =$ left in Table 15. However, the difference from the best score is small (less than 0.002). These results validate the effectiveness of the proposed method compared with the baseline and ablated models, according to both metrics.

*(3) ACD$_{static}$ vs. ACD$_{video}$.* Comparing ACD$_{\text{static}}$ with ACD$_{\text{video}}$, ACD$_{\text{static}}$ is smaller than ACD$_{\text{video}}$. This is because the difference in location gradually increased after the collision when the cavity was located on the opposite side. Since the objective of this study is to correctly predict the shape, rather than the location, ACD$_{\text{static}}$ is a more valid evaluation than ACD$_{\text{video}}$ for this purpose.

*(4) Comparison of CD$_{static}$ and ACD$_{static}$ among models.* When comparing the models, the baselines (i.e., GO- and LPO-based models) tended to obtain similar CD$_{\text{static}}$ and ACD$_{\text{static}}$ values because they struggled to find the optimization direction, as shown in Figures 6–10. In contrast, the proposed models (i.e., SfC-NeRF-based models, including the ablated models) tend to obtain CD$_{\text{static}}$, which is smaller than ACD$_{\text{static}}$. These results indicate that the proposed models effectively capture the positional bias of the cavity. Notably, a larger ACD$_{\text{static}}$ does not indicate better performance unless CD$_{\text{static}}$ is adequately small because it is possible to increase ACD$_{\text{static}}$ while sacrificing CD$_{\text{static}}$.

### A.3.2. Evaluation per external shape

In Experiments I (Section 4.2) and II (Section 4.3), we reported the scores averaged over external shapes (i.e., sphere, cube, bicone, cylinder, and diamond objects). To evaluate from a different perspective, this appendix presents the scores for each external shape, averaged over other conditions, i.e., either $s_c \in \{0, (\frac{1}{2})^3, (\frac{2}{3})^3, (\frac{3}{4})^3\}$ or $l_c \in \{\text{left}, \text{right}, \text{up}, \text{down}\}$.

**Results.** Table 16 summarizes the results when the cavity size $s_c$ is varied (related to the results in Table 1), and Table 17 summarizes the results when the cavity location $l_c$ is varied (related to the results in Table 2). We found that although the scores were affected by the external shape, the same trends observed previously regarding the superiority or inferiority of the models were maintained. Specifically, SfC-NeRF outperformed both the baseline and ablated models in most cases.

| | Sphere | Cube | Bicone | Cylinder | Diamond |
|---|---|---|---|---|---|
| Static | 0.897 | 0.612 | 0.724 | 0.697 | 0.671 |
| GO | 0.889 | 0.637 | 0.704 | 0.756 | 0.663 |
| GO$_{\text{mass}}$ | 0.934 | 1.345 | 0.760 | 1.218 | 0.663 |
| LPO | 0.774 | 0.564 | 0.639 | 0.678 | 0.622 |
| LPO$_{\text{mass}}$ | 0.796 | 0.605 | 0.656 | 0.726 | 0.622 |
| SfC-NeRF$_{-\text{mass}}$ | 0.561 | 0.500 | 0.455 | 0.447 | 0.553 |
| SfC-NeRF$_{-\text{APL}}$ | 0.303 | 1.082 | 0.579 | 0.885 | 0.591 |
| SfC-NeRF$_{-\text{APT}}$ | 0.178 | 0.375 | 0.286 | 0.502 | 0.331 |
| SfC-NeRF$_{-\text{key}}$ | 0.081 | 0.173 | 0.159 | 0.288 | 0.230 |
| SfC-NeRF$_{-\text{VA}}$ | 0.113 | 0.279 | 0.363 | 0.558 | 0.268 |
| SfC-NeRF | 0.067 | 0.163 | 0.138 | 0.264 | 0.193 |

Table 16. Comparison of CD ($\times 10^3 \downarrow$) when the cavity size $s_c$ is varied. The scores were averaged over cavity sizes.

| | Sphere | Cube | Bicone | Cylinder | Diamond |
|---|---|---|---|---|---|
| Static | 1.006 | 0.719 | 0.824 | 0.818 | 0.772 |
| GO | 0.991 | 0.809 | 0.847 | 0.898 | 0.799 |
| GO$_{\text{mass}}$ | 1.065 | 1.528 | 0.934 | 1.332 | 1.125 |
| LPO | 0.954 | 0.673 | 0.764 | 0.804 | 0.701 |
| LPO$_{\text{mass}}$ | 0.980 | 0.723 | 0.796 | 0.845 | 0.711 |
| SfC-NeRF$_{-\text{mass}}$ | 0.695 | 0.480 | 0.424 | 0.595 | 0.533 |
| SfC-NeRF$_{-\text{APL}}$ | 0.548 | 1.064 | 0.373 | 1.194 | 0.592 |
| SfC-NeRF$_{-\text{APT}}$ | 0.318 | 0.502 | 0.374 | 0.730 | 0.451 |
| SfC-NeRF$_{-\text{key}}$ | 0.189 | 0.371 | 0.235 | 0.448 | 0.286 |
| SfC-NeRF$_{-\text{VA}}$ | 0.240 | 0.418 | 0.790 | 0.534 | 0.338 |
| SfC-NeRF | 0.152<br>(0.417) | 0.342<br>(0.386) | 0.231<br>(0.365) | 0.393<br>(0.491) | 0.289<br>(0.420) |

Table 17. Comparison of CD ($\times 10^3 \downarrow$) when the cavity location $l_c$ is varied. The scores were averaged over cavity sizes. The gray score in parenthesis indicates ACD ($\times 10^3$). It is expected that the original CD is smaller than it.

## A.4. Possible challenges with real data

As discussed in Section 5, since SfC is a novel task, this study focused on evaluating its fundamental performance using simulation data, leaving the validation with real data as a challenge for future research. However, it is both feasible and important to discuss the potential challenges associated with real data, and we address these in this appendix. Three potential challenges are outlined below:

*(1) Difficulty in accurately estimating external structures.* While significant progress has been made in recent years regarding the estimation of 3D external structures, it is not yet possible to estimate them accurately for all objects in all situations. Our method assumes that the external structure learned in the first frame of the video sequence is accurate. Therefore, if this estimation fails, overall performance is degraded. We believe that incorporating the concept of a physics-informed model, particularly in challenging scenarios (e.g., sparse views), such as Lagrangian particle op-

timization [3], could provide a solution to this issue.

*(2) Gap between real physics and the physics used in simulation.* Despite recent advancements in physical simulation models, discrepancies between real-world physics and the physics underlying the simulation still persist. We believe that refining the proposed method alongside physics-informed models (e.g., those discussed in Section 2) could help alleviate this problem.

*(3) Difficulty in accurately estimating physical properties.* As mentioned in Section 3.1, we address *SfC* under the assumption that the ground truth physical properties are available in advance to mitigate the chicken-and-egg problem between the physical properties and internal structures. This assumption is reasonable if the material can be identified; however, obtaining perfectly accurate values for the physical properties in real-world scenarios is challenging. While the issue of solving the chicken-and-egg problem remains, an appearance-based physical property estimation method has already been proposed (e.g., PAC-NeRF [5]). Combining our approach with previous methods for simultaneous optimization of physical properties and internal structure would be an exciting direction for future research.

## B. Qualitative results

This appendix presents qualitative results. The corresponding demonstration videos are available at `https://www.kecl.ntt.co.jp/people/kaneko.takuhiro/projects/sfc/`.

### B.1. Qualitative results for Experiments I and II

We provide the qualitative results for Experiments I (Section 4.2) and II (Section 4.3) in Figures 6–10.

### B.2. Qualitative results for Experiment III

We provide the qualitative results for Experiments III (Section 4.4) in Figures 11–13.

### B.3. Qualitative results for Experiment IV

We provide the qualitative results for Experiments IV (Appendix A.2.1) in Figure 14.

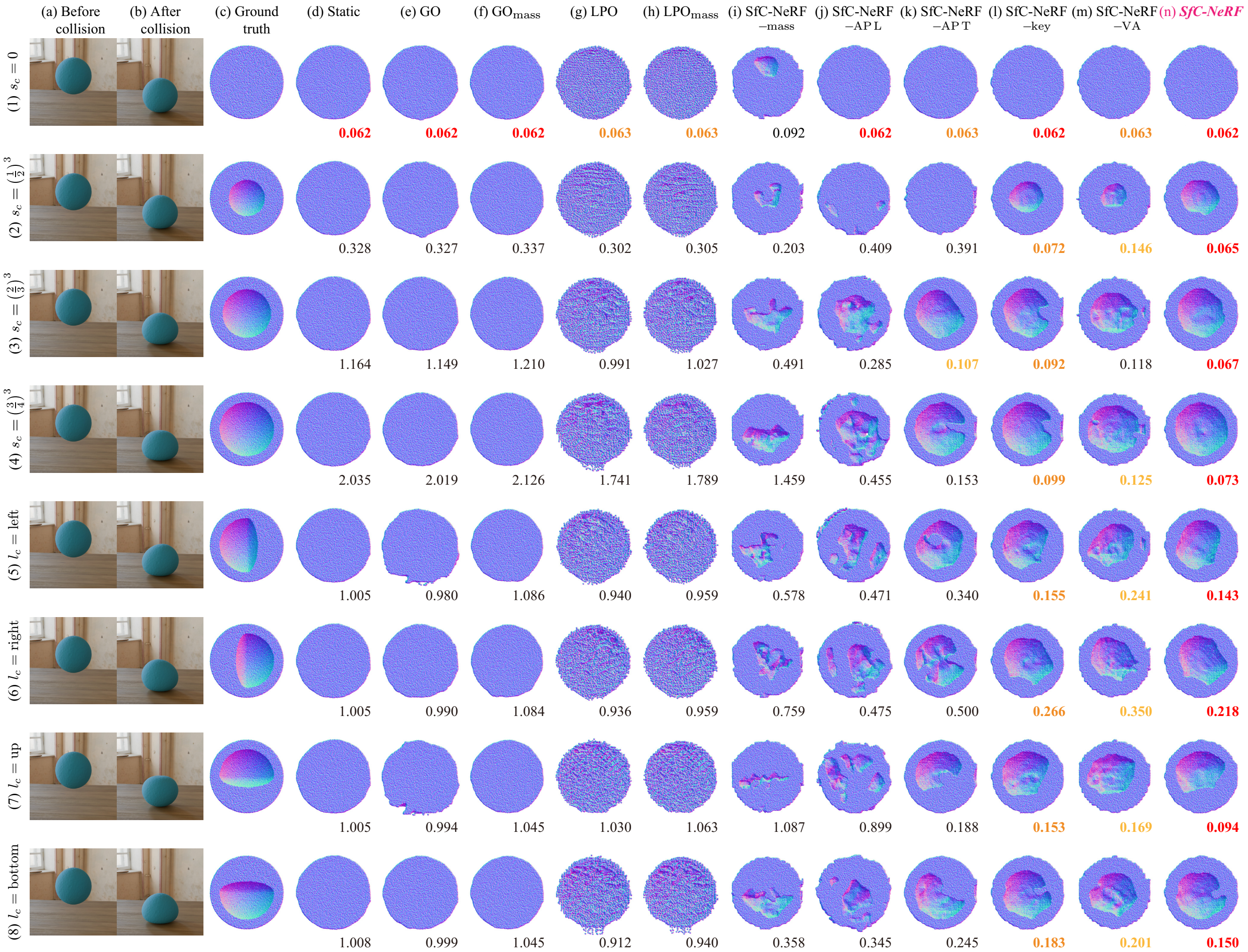

Figure 6. Comparison of learned internal structures for *sphere* objects. *(a) and (b) Examples of training images.* The images are zoomed in for easy viewing. *(a) Examples of training images* ***before*** *collision.* As shown in this column, the appearances of the objects are the same across all scenes (1)–(8). Consequently, it is difficult to distinguish the internal structures based solely on these appearances. *(b) Examples of training images* ***after*** *collision.* To overcome the difficulty mentioned above, we address *SfC*, in which we aim to identify the internal structures based on appearance changes before and after collision, as shown in (a) and (b). *(c)–(n) Internal structures visualized through cross-sectional views perpendicular to the ground.* In (d)–(n), the score below each image indicates CD ($\times 10^3 \downarrow$). *(c) Ground truth internal structures.* As shown in this column, although the external appearances are the same in (a), the internal structures are different. *(d) Internal structures learned from the first frames of the video sequences.* The same internal structures (i.e., the filled objects) were learned because the appearances were the same before the collision (a). *(e)–(h) Internal structures learned using the baselines (GO- and LPO-based models).* These models struggled to find optimal learning directions. *(i)–(m) Internal structures learned using the ablated models.* The ablated models are insufficient to prevent convergence to improper solutions. *(n) Internal structures learned using SfC-NeRF (full model).* The full model overcomes the above drawbacks and achieves the best CD.

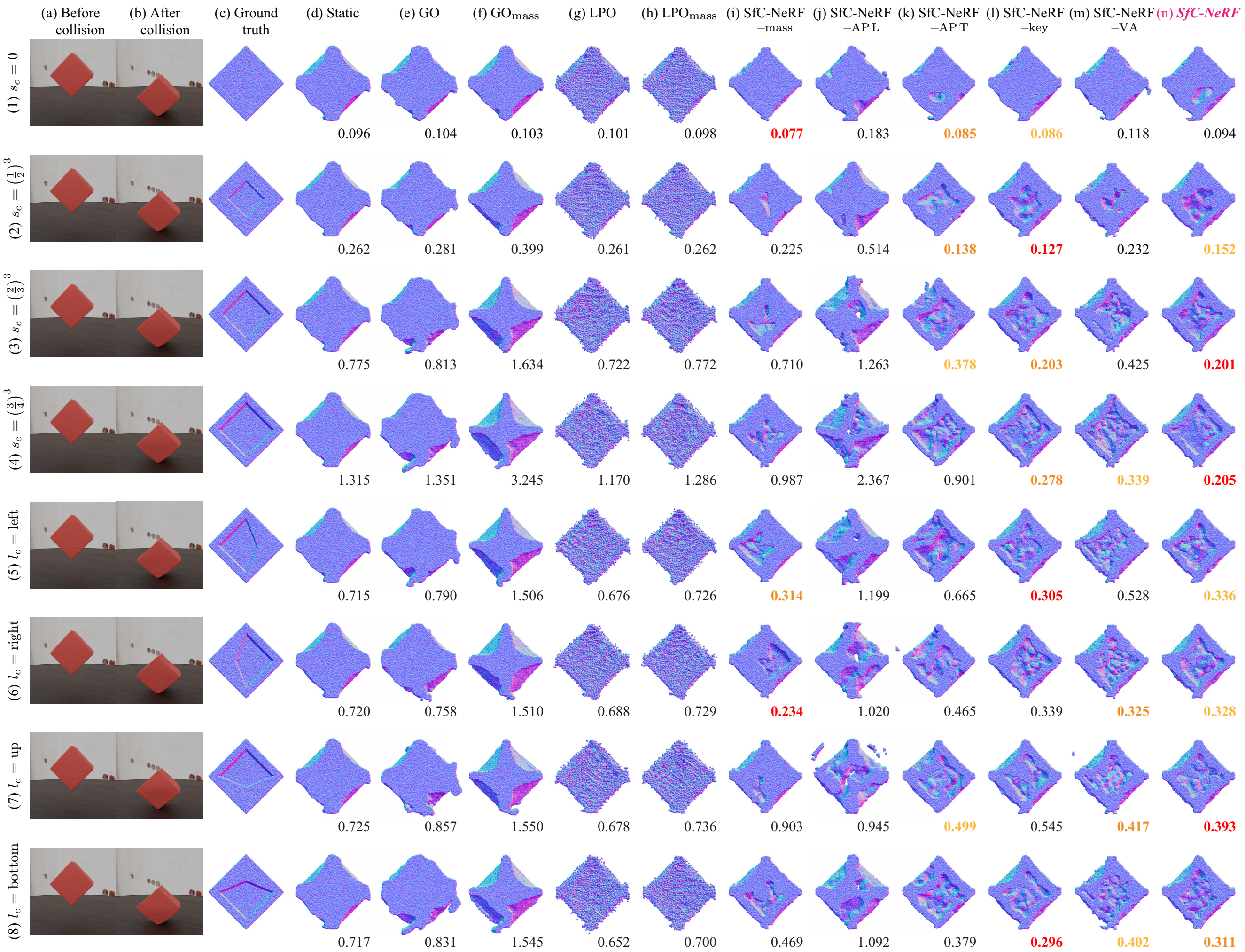

Figure 7. Comparison of learned internal structures for *cube* objects. The view of the figure is the same as that of Figure 6.

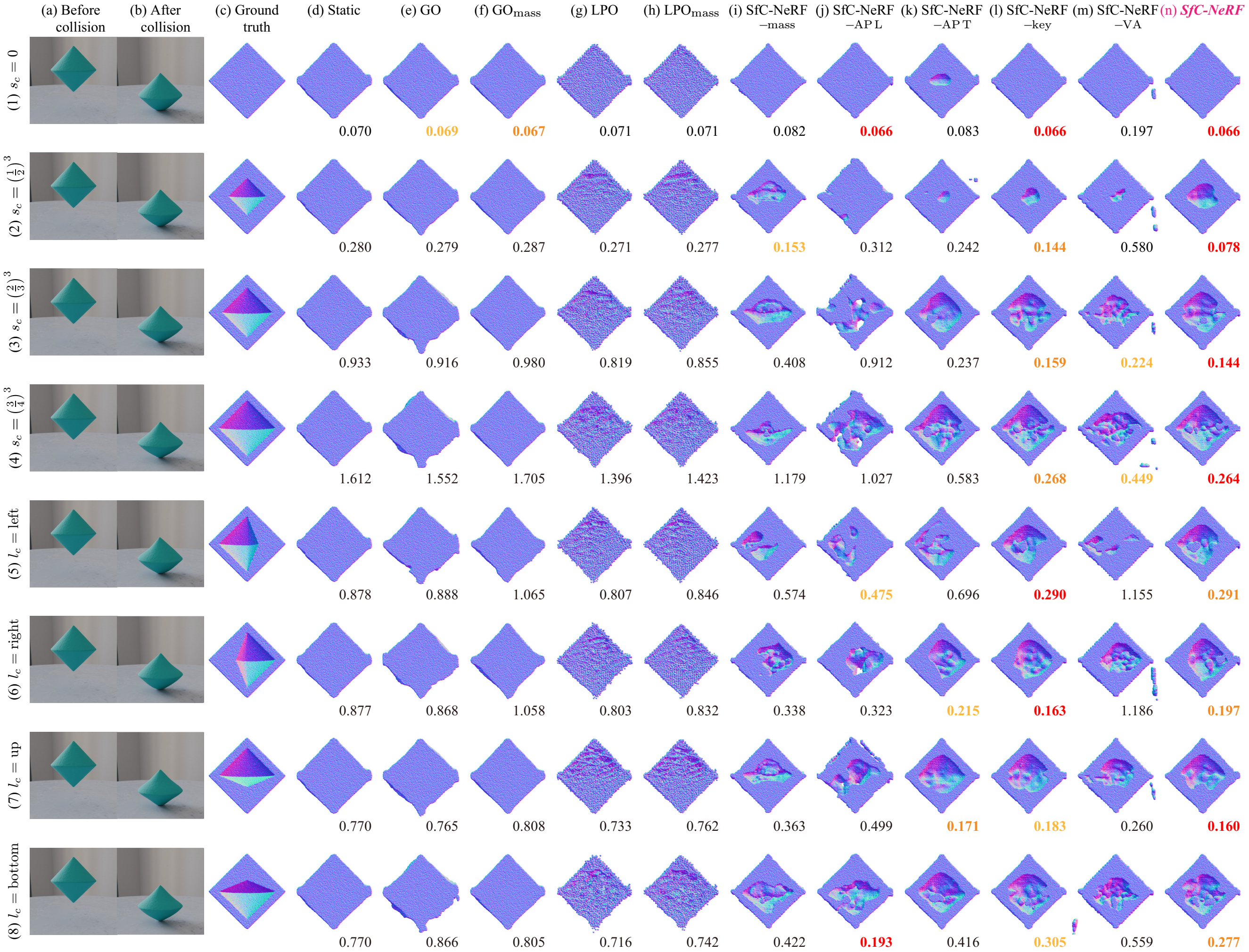

Figure 8. Comparison of learned internal structures for *bicone* objects. The view of the figure is the same as that of Figure 6.

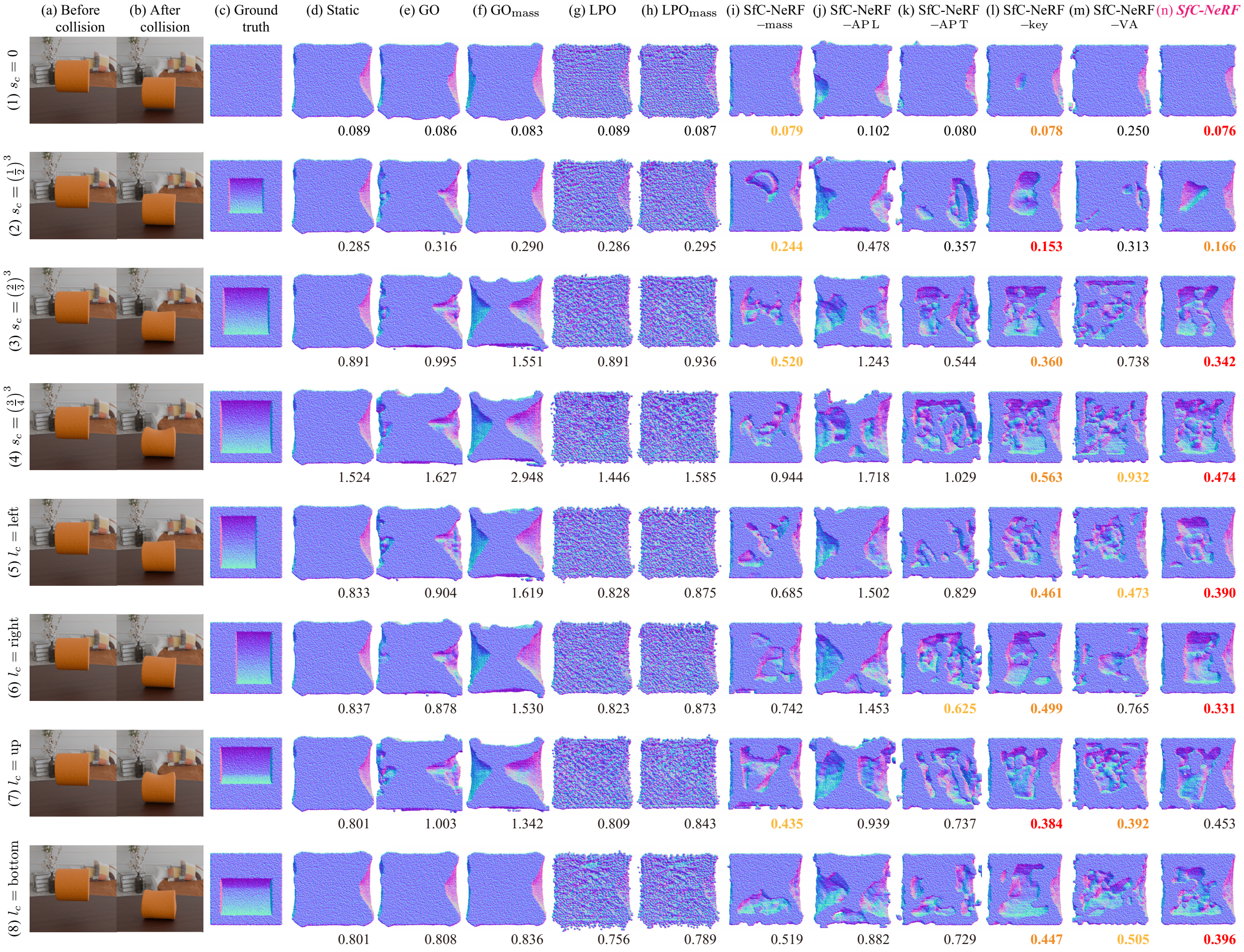

Figure 9. Comparison of learned internal structures for *cylinder* objects. The view of the figure is the same as that of Figure 6.

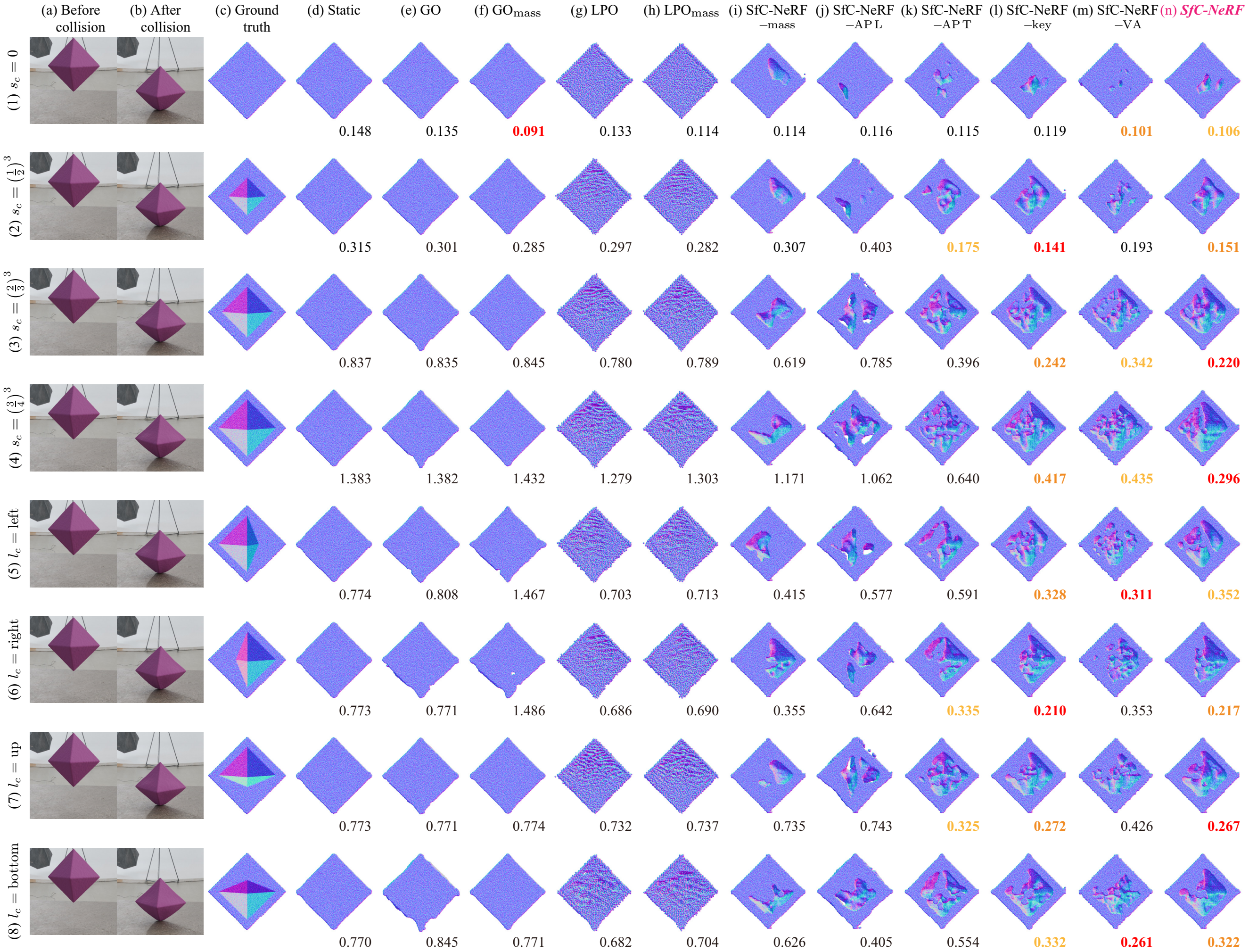

Figure 10. Comparison of learned internal structures for *diamond* objects. The view of the figure is the same as that of Figure 6.

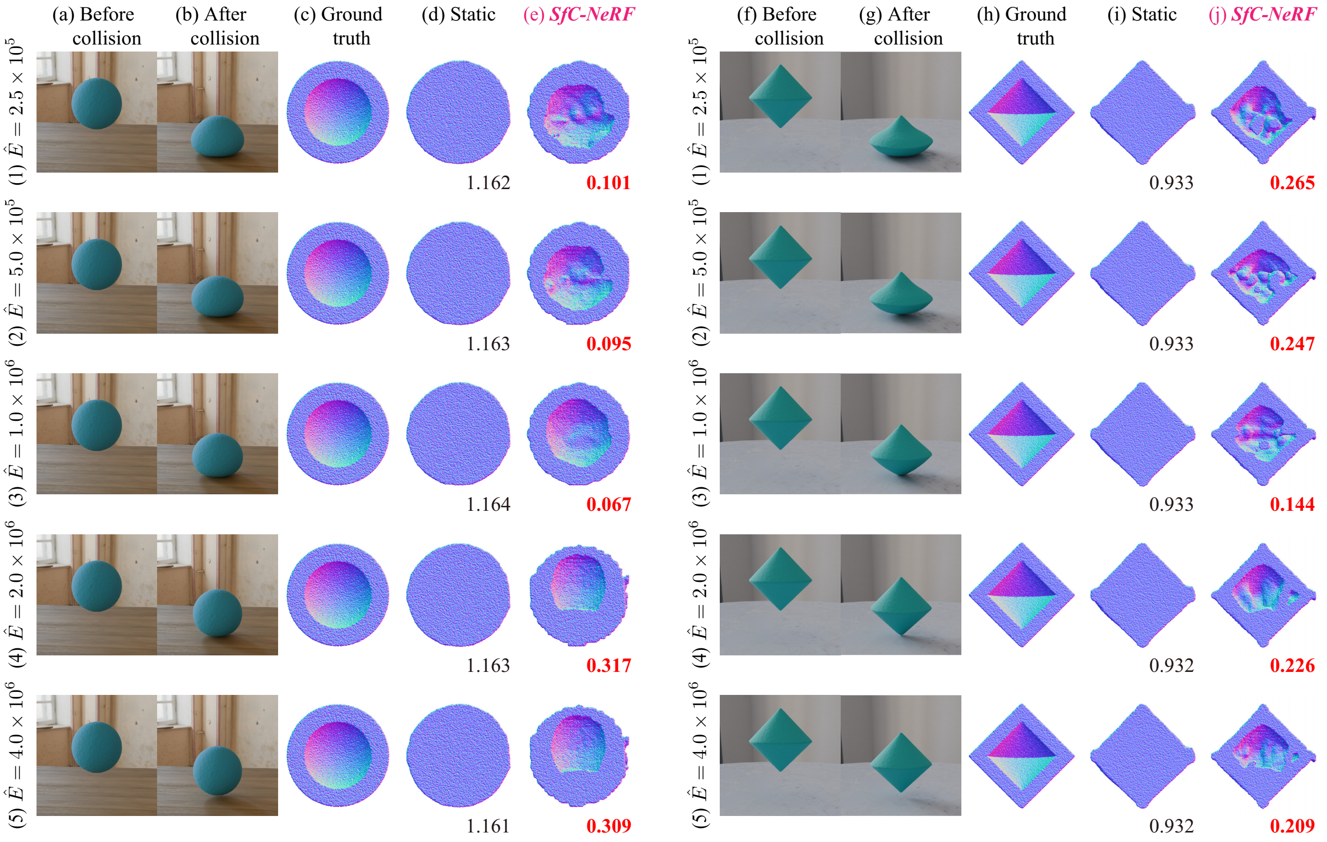

Figure 11. Comparison of learned internal structures for *sphere* objects (left) and *bicone* objects (right) when Young's modulus $\hat{E}$ is varied. Young's modulus is a measure of elasticity and quantifies tensile or compressive stiffness when force is applied. Here, we discuss the results for the sphere objects because the same tendencies are observed for the bicone objects. As shown in (a) and (c), the external appearances before collision (a) and the internal structures (c) are the same in all cases (1)–(5). However, as shown in (b), the shapes after collision differ because of variations in Young's modulus $\hat{E} \in \{2.5 \times 10^5, 5.0 \times 10^5, 1.0 \times 10^6, 2.0 \times 10^6, 4.0 \times 10^6\}$. In particular, as Young's modulus increases from top to bottom, the object becomes stiffer, and the amount of shape change decreases. In the Static model (b), the internal structure was learned from the first frame, which looks the same in all cases. As a result, the same internal structure was learned across all variations. In contrast, in SfC-NeRF (e), the internal structure was learned using video sequences with different appearances. In this example, the same internal structure is expected to be learned in all cases. However, the varying appearances after collision (b), which provide a clue for solving the problem, lead to different outcomes. As shown in (1)(b) and (2)(b), when the object is soft, it deforms significantly after collision. This makes it difficult to capture the internal structure consistently, as shown in (1)(e) and (2)(e). In contrast, as shown in (4)(b) and (5)(b), when the object is stiffer, the shape change is limited. This narrows the range within which internal structures can be estimated, as shown in (4)(e) and (5)(e). Since *SfC* is an ill-posed problem with multiple possible solutions, the obtained results are considered reasonable. However, further improvement remains a topic for future work.

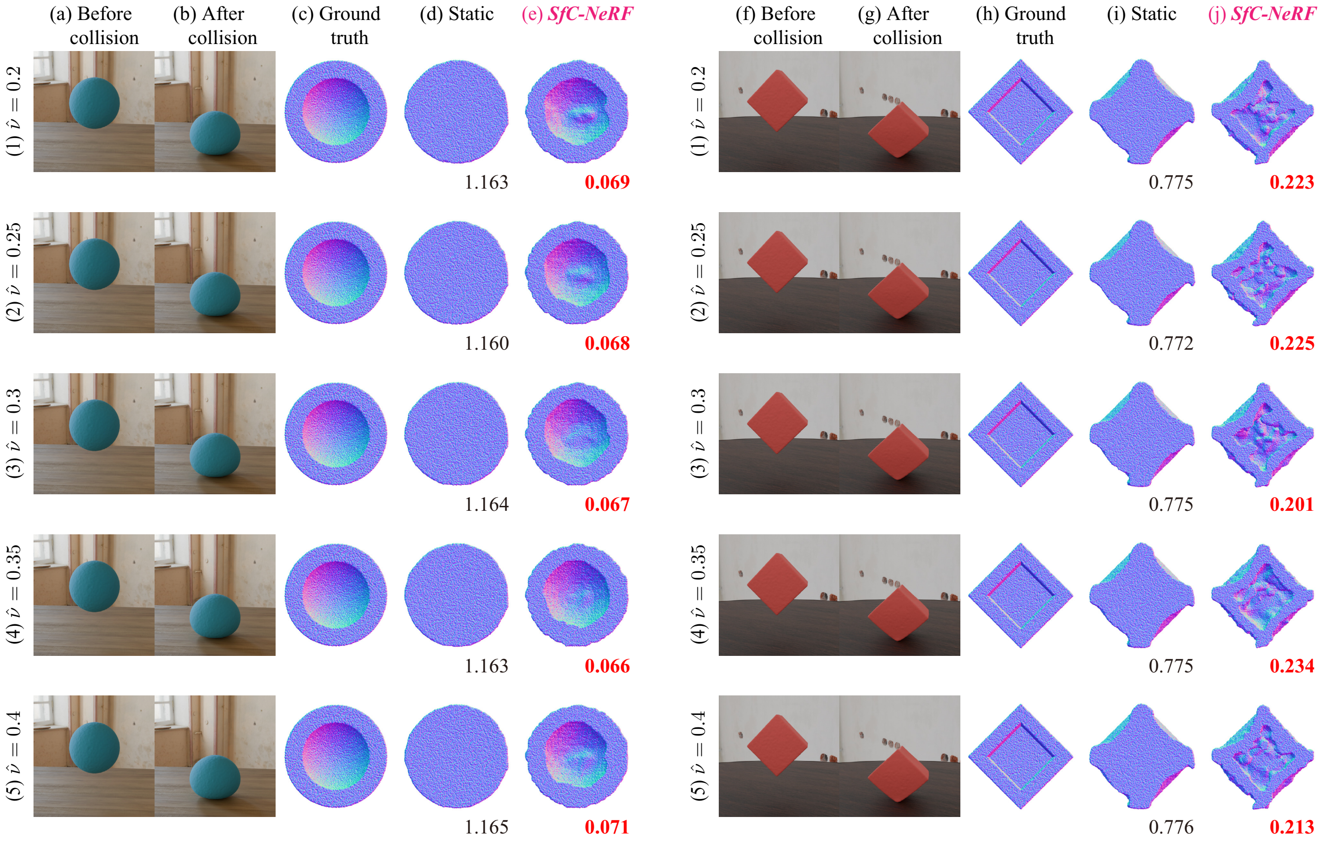

Figure 12. Comparison of learned internal structures for *sphere* objects (left) and *cube* objects (right) when Poisson's ratio $\hat{\nu}$ is varied. Poisson's ratio is a measure of Poisson effect and quantifies how much a material deforms in a direction perpendicular to the direction in which force is applied. We varied Poisson's ratio $\hat{\nu}$ within the range of values commonly observed in real materials, i.e., $\hat{\nu} \in \{0.2, 0.25, 0.3, 0.35, 0.4\}$. As shown in (b)(g), this physical property does not significantly affect the appearance after the collision, compared to the results when Young's modulus is varied (Figure 11). As a result, the learned internal structures are almost identical, as shown in (e)(j).

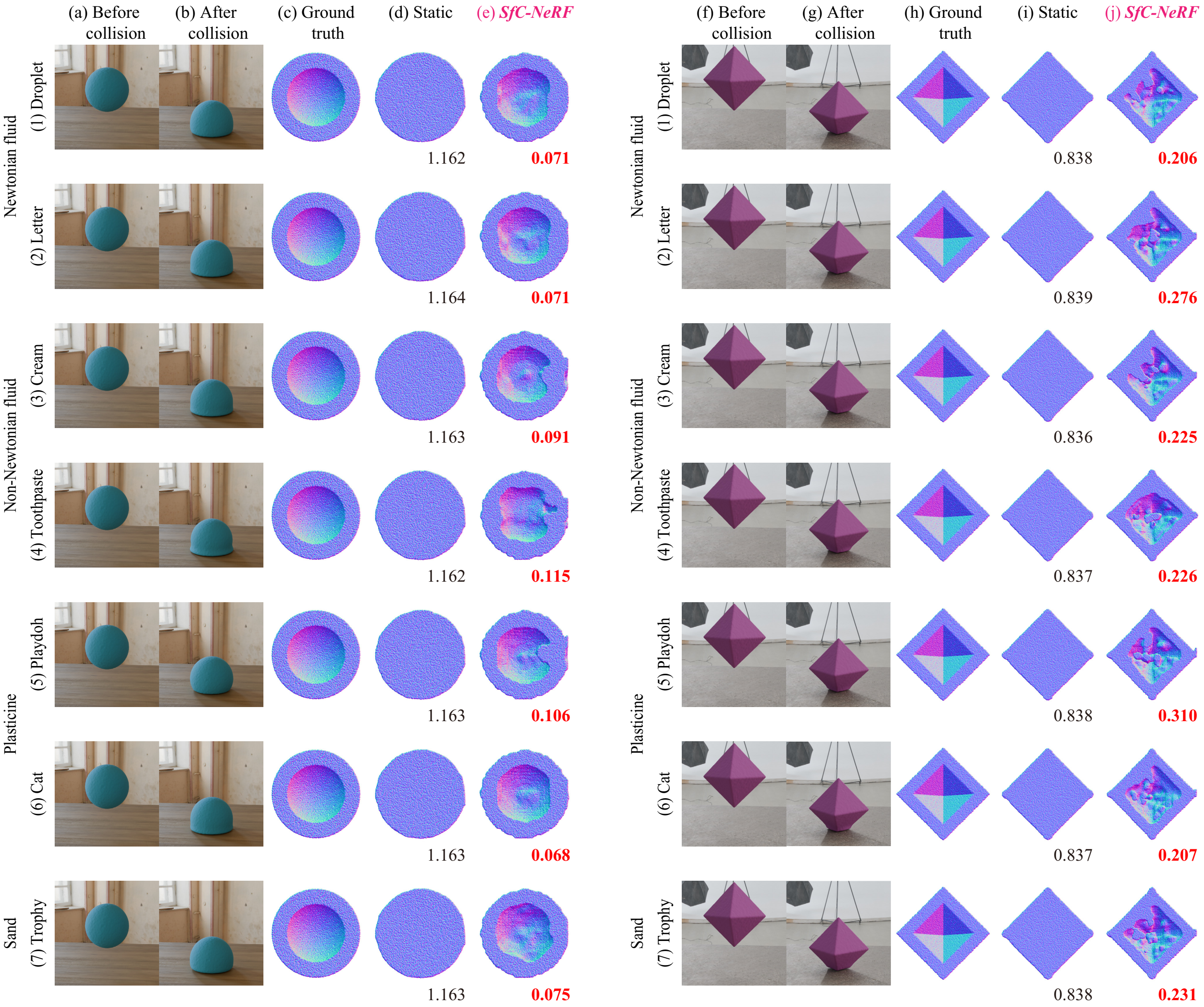

Figure 13. Comparison of learned internal structures for *sphere* objects (left) and *diamond* objects (right) with varying materials. The physical properties were based on the PAC-NeRF dataset [5]. Specifically: (1) Newtonian fluid with the "Droplet" setting (fluid viscosity $\hat{\mu} = 200$ and bulk modulus $\hat{\kappa} = 10^5$). (2) Newtonian fluid with the "Letter" setting ($\hat{\mu} = 100$ and $\hat{\kappa} = 10^5$). (3) Non-Newtonian fluid with the "Cream" setting (shear modulus $\hat{\mu} = 10^4$, bulk modulus $\hat{\kappa} = 10^6$, yield stress $\hat{\tau}_Y = 3 \times 10^3$, and plasticity viscosity $\hat{\eta} = 10$). (4) Non-Newtonian fluid with the "Toothpaste" setting ($\hat{\mu} = 5 \times 10^3$, $\hat{\kappa} = 10^5$, $\hat{\tau}_Y = 200$, and $\hat{\eta} = 10$). (5) Plasticine with the "Playdoh" setting (Young's modulus $\hat{E} = 2 \times 10^6$, Poisson's ratio $\hat{\nu} = 0.3$, and yield stress $\hat{\tau}_Y = 1.54 \times 10^4$). (6) Plasticine with the "Cat" setting ($\hat{E} = 2 \times 10^6$, $\hat{\nu} = 0.3$, and $\hat{\tau}_Y = 3.85 \times 10^3$). (7) Sand with the "Trophy" setting ($\hat{\theta}_{fric} = 40°$). These results demonstrate that *SfC-NeRF* (e)(j) improves structure estimation compared to Static (d)(i), regardless of the material. However, the improvement rate depends on the material. As an initial approach to address *SfC*, we proposed a general-purpose method in this study. However, it would be interesting to develop methods specifically tailored to individual materials in future work.

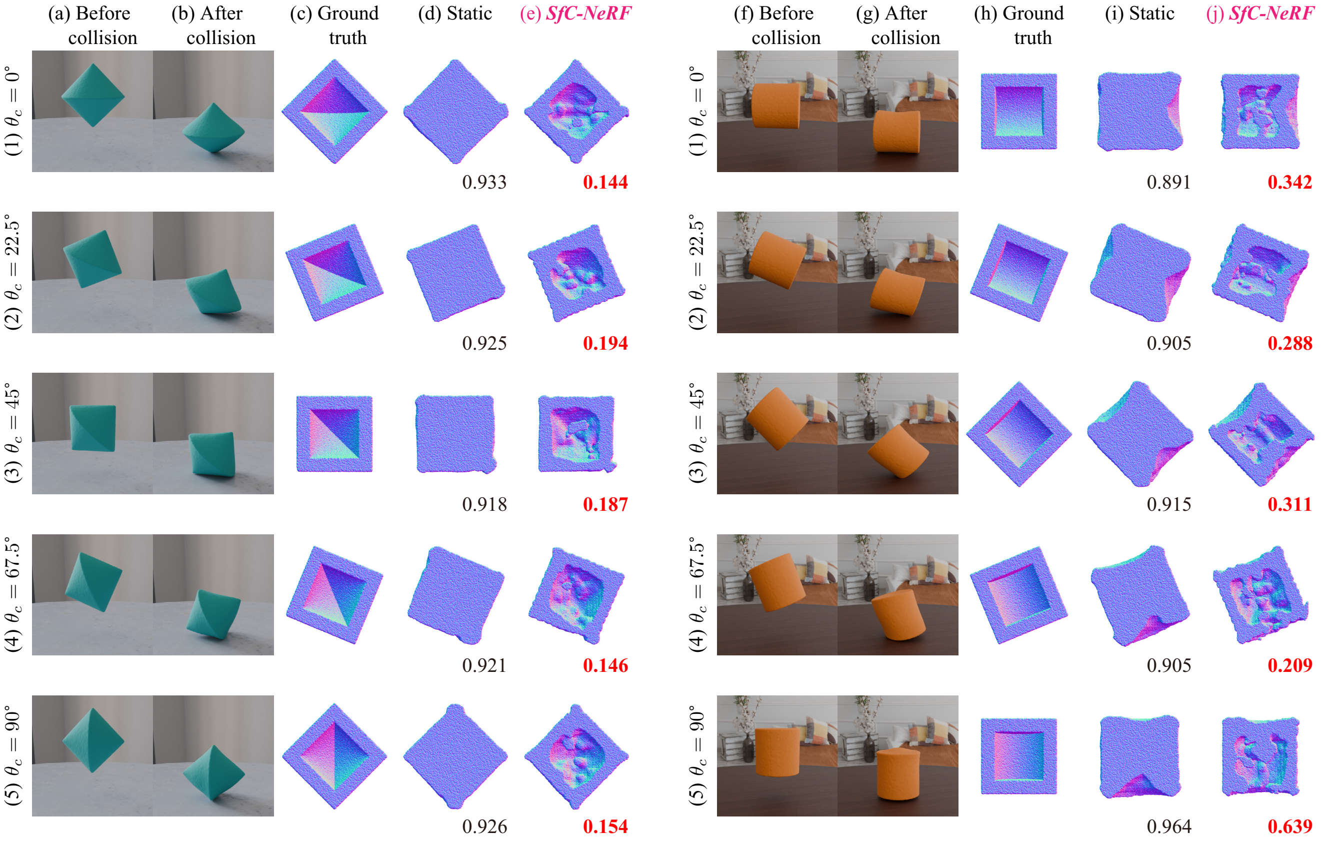

Figure 14. Comparison of learned internal structures for *bicone* objects (left) and *cylinder* objects (right) when collision angle $\theta_c$ is varied. We varied collision angle $\theta_c \in \{0°, 22.5°, 45°, 67.5°, 90°\}$. We found that the effect of collision angle on the estimation of the internal structure depends on the object shape. (a)–(e) In the case of an object, such as *bicone*, where the object is entirely visible regardless of the collision angle, the estimation performance remains relatively stable across different collision angles. (f)–(j) In contrast, in the case of an object, such as *cylinder*, where the visible area varies greatly depending on the collision angle, the estimation performance also changes with the collision angle. For example, in (5)(g), the bottom of the object is not visible when it collides with the ground. As a result, a hole is generated at the bottom of the object in (5)(j). This issue may be alleviated by improving camera placement. Other possible factors that affect estimation performance are discussed in Appendix A.2.1.

# C. Implementation details

## C.1. Dataset

Since *SfC* is a new task and no established dataset is available, we created a new dataset called the *SfC dataset* based on the protocol of PAC-NeRF [5], which is a pioneering study of geometry-agnostic system identification. In the main experiments presented in Section 4, we prepared a total of 115 objects by changing the external shape, internal structure, and material of the objects. Figure 3 shows examples of data in this dataset. First, we prepared five external shapes: *sphere*, *cube*, *bicone*, *cylinder*, and *diamond*. Regarding the internal structure and material, we set the default values as follows: the cavity size rate for the filled object, $s_c$, was set to $(\frac{2}{3})^3$, the cavity location, $l_c$, was set to the center, and the material was defined as an elastic material with Young's modulus $\hat{E} = 10^6$ and Poisson's ratio $\hat{\nu} = 0.3$. For this default properties, one of them was changed as follows.

*(a) Three different sized cavities*: $s_c \in \{0, (\frac{1}{2})^3, (\frac{3}{4})^3\}$.

*(b) Four different locations of cavities*: the center $l_c$ is moved in $\{\text{up}, \text{down}, \text{left}, \text{right}\}$.

*(c-1) Eight different elastic materials*: those with four different Young's moduli $\hat{E} \in \{2.5 \times 10^5, 5 \times 10^5, 2 \times 10^6, 4 \times 10^6\}$ and those with four different Poisson's ratios $\hat{\nu} \in \{0.2, 0.25, 0.35, 0.4\}$.

*(c-2) Seven different materials*: two Newtonian fluids, two non-Newtonian fluids, two plasticines, and one sand. The physical properties of these materials were based on the PAC-NeRF dataset [5]. Specifically, the two Newtonian fluids include one with the "Droplet" setting (fluid viscosity $\hat{\mu} = 200$ and bulk modulus $\hat{\kappa} = 10^5$) and one with the "Letter" setting ($\hat{\mu} = 100$ and $\hat{\kappa} = 10^5$). The two non-Newtonian fluids include one with the "Cream" setting (shear modulus $\hat{\mu} = 10^4$, bulk modulus $\hat{\kappa} = 10^6$, yield stress $\hat{\tau}_Y = 3 \times 10^3$, and plasticity viscosity $\hat{\eta} = 10$) and one with the "Toothpaste" setting ($\hat{\mu} = 5 \times 10^3$, $\hat{\kappa} = 10^5$, $\hat{\tau}_Y = 200$, and $\hat{\eta} = 10$). The two plasticines include one with the "Playdoh" setting (Young's modulus $\hat{E} = 2 \times 10^6$, Poisson's ratio $\hat{\nu} = 0.3$, and yield stress $\hat{\tau}_Y = 1.54 \times 10^4$) and one with the "Cat" setting ($\hat{E} = 2 \times 10^6$, $\hat{\nu} = 0.3$, and $\hat{\tau}_Y = 3.85 \times 10^3$). The sand is defined with the "Trophy" setting ($\hat{\theta}_{fric} = 40°$).

Thus, we created 5 external shapes × (1 default + 3 sizes + 4 locations + (8 + 7) materials) = 115 objects.

In this appendix, we also prepared 20 objects for the extended experiments described in Appendix A.2. Specifically, we consider four collision angles: $\theta_c \in \{22.5°, 45°, 67.5°, 90°\}$. Thus, in this appendix, we created 5 external shapes × 4 collision angles = 20 objects. The total number of objects created in the main text and this appendix is $115 + 20 = 135$.

Following the PAC-NeRF study [5], the ground truth data were generated using the MLS-MPM simulator [1], where each object fell freely under the influence of gravity and collided with the ground plane. The images were rendered under various environmental lighting conditions and ground textures using a photorealistic renderer. Each scene was captured from 11 viewpoints using cameras spaced in the upper hemisphere including an object.

## C.2. Model

We implemented the models based on the official PAC-NeRF code [5].[6] PAC-NeRF represents an Eulerian grid-based scene representation using voxel-based NeRF (specifically, direct voxel grid optimization (DVGO) [7]) and conducts a Lagrangian particle-based differentiable physical simulation using a differentiable MPM simulator (specifically, DiffTaichi [2]). More specifically, DVGO represents a volume density field $\sigma^{G'}$ using a 3D dense voxel grid and represents a color field $\mathbf{c}^{G'}$ using the combination of a 4D dense voxel grid and a 2-layer multi-layer perceptron (MLP) with a hidden dimension of 128. When this MLP is used, positional embedding in the viewing direction $\mathbf{d}$ is used as an additional input. We set the resolutions of $\sigma^{G'}$ and $\mathbf{c}^{G'}$ to match those in PAC-NeRF [5].

## C.3. Training settings

We performed static optimization (Figure 2(i)) using the same settings as those used for PAC-NeRF. Specifically, we trained the model for 6000 iterations using the Adam optimizer [4] with learning rates of 0.1 for the volume density grid and color grid, and a learning rate of 0.001 for the MLP. The momentum terms $\beta_1$ and $\beta_2$ were set to 0.9 and 0.999, respectively. In the dynamic optimization (Figure 2(ii)), we trained the model for 1000 iterations using the Adam optimizer [4] with a default learning rate of 6.4 for the volume density grid. The momentum terms $\beta_1$ and $\beta_2$ were set to 0.9 and 0.999, respectively. We found that a large learning rate is useful for efficiently reducing the volume density; however, it is not necessary when the estimated mass $m$ sufficiently approaches the ground truth mass $\hat{m}$. Therefore, we divided the learning rate by 2 (with a minimum of 0.1) as long as the estimated mass $m$ was below the ground truth mass $\hat{m}$. Conversely, we multiplied the learning rate by 2 (with a maximum of 6.4) as long as the estimated mass $m$ exceeded the ground truth mass $\hat{m}$.

We conducted volume annealing every 100 iteration during the dynamic optimization. When the estimated mass $m$ is significantly larger than the ground truth mass $\hat{m}$ (specifically, when the difference exceeds 10 in practice), the expansion process is skipped to prevent the estimated mass $m$ from deviating further from the ground truth mass $\hat{m}$.

[6] https://github.com/xuan-li/PAC-NeRF

In appearance-preserving training, static optimization was performed using settings similar to those mentioned above (i.e., static optimization in Step (i) (Figure 2(i))), but the number of iterations was reduced to 10.

We empirically set the hyperparameters for the total loss $\mathcal{L}_{\text{full}}$ (Equation 12) as $\lambda_{\text{mass}} = 1$, $\lambda_{\text{pres}} = 100$, $w_{\text{depth}} = 0.01$, and $\lambda_{\text{key}} = 10$. The hyperparameter for the background loss $\mathcal{L}_{\text{bg}}$ was set to $w_{\text{bg}} = 0.2$.

## C.4. Evaluation metrics

As mentioned in Section 3.1, we use particles $\mathcal{P}^P(t_0)$ to represent the structure (including the internal structure) of an object and estimate $\mathcal{P}^P(t_0)$ to match the ground truth $\hat{\mathcal{P}}^P(t_0)$. Therefore, we evaluated the model by measuring the distance between $\mathcal{P}^P(t_0)$ and $\hat{\mathcal{P}}^P(t_0)$ using the *chamfer distance (CD)*. The smaller the value, the higher the degree of matching. As mentioned in Section 4.3, we also used the *anti-chamfer distance (ACD)*, which is the chamfer distance between the predicted particles $\mathcal{P}^P(t_0)$ and the ground truth particles $\tilde{\mathcal{P}}^P(t_0)$, where the cavity was placed on the opposite side, to evaluate the capture of the cavity location.

## References

[1] Yuanming Hu, Yu Fang, Ziheng Ge, Ziyin Qu, Yixin Zhu, Andre Pradhana, and Chenfanfu Jiang. A moving least squares material point method with displacement discontinuity and two-way rigid body coupling. *ACM Trans. Graph.*, 37(4), 2018. 16

[2] Yuanming Hu, Luke Anderson, Tzu-Mao Li, Qi Sun, Nathan Carr, Jonathan Ragan-Kelley, and Frédo Durand. DiffTaichi: Differentiable programming for physical simulation. In *ICLR*, 2020. 5, 16

[3] Takuhiro Kaneko. Improving physics-augmented continuum neural radiance field-based geometry-agnostic system identification with Lagrangian particle optimization. In *CVPR*, 2024. 6

[4] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 16

[5] Xuan Li, Yi-Ling Qiao, Peter Yichen Chen, Krishna Murthy Jatavallabhula, Ming Lin, Chenfanfu Jiang, and Chuang Gan. PAC-NeRF: Physics augmented continuum neural radiance fields for geometry-agnostic system identification. In *ICLR*, 2023. 6, 14, 16

[6] Shanchuan Lin, Andrey Ryabtsev, Soumyadip Sengupta, Brian L. Curless, Steven M. Seitz, and Ira Kemelmacher-Shlizerman. Real-time high-resolution background matting. In *CVPR*, 2021. 2

[7] Cheng Sun, Min Sun, and Hwann-Tzong Chen. Direct voxel grid optimization: Super-fast convergence for radiance fields reconstruction. In *CVPR*, 2022. 16