# Enforcing View-Consistency in Class-Agnostic 3D Segmentation Fields

Corentin Dumery, Aoxiang Fan, Ren Li, Nicolas Talabot, Pascal Fua

EPFL

## 1  Problem Statement

→ Given as input a set of class-agnostic 2D masks with little consistency across views, we aim to learn a meaningful **3D object field** that segments the different instances in the scene

→ Existing methods **either assume predefined classes** or only learn **feature fields that require clustering** as post-processing
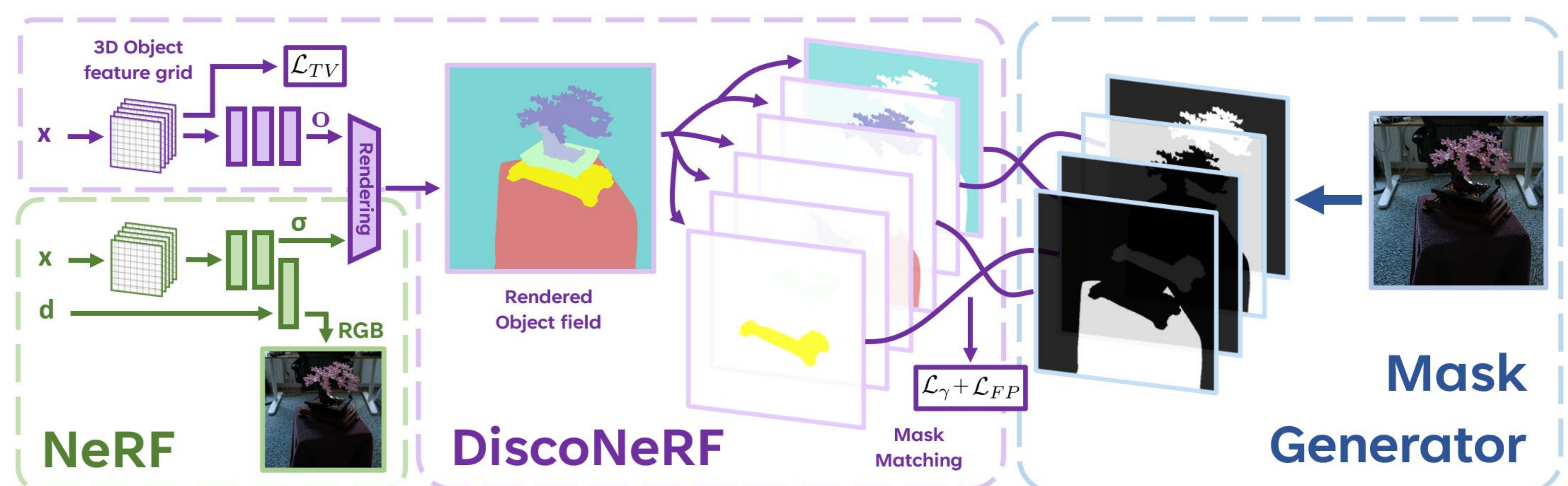


## 2  Method

- Augment NeRF with a **high-dimensional** object field, which is rendered similarly to color

- Each channel learns a **single** object with a per-pixel probability

$$\alpha(O_n, M_m) = \frac{\sum\limits_{i,j} \min(O_{n,i,j}, M_{m,i,j})}{\sum\limits_{i,j} \max(O_{n,i,j}, M_{m,i,j})}$$

$$\mathcal{L}_\gamma = \frac{1}{K} \sum_{m=1}^{K} \|M_m - O_{\gamma(m)}\|_2^2 ,$$

$$\mathcal{L}_{FP} = \frac{1}{K} \sum_{m=1}^{K} \sum_{n \neq \gamma(m)} \|M_m * O_n\|_2^2 ,$$

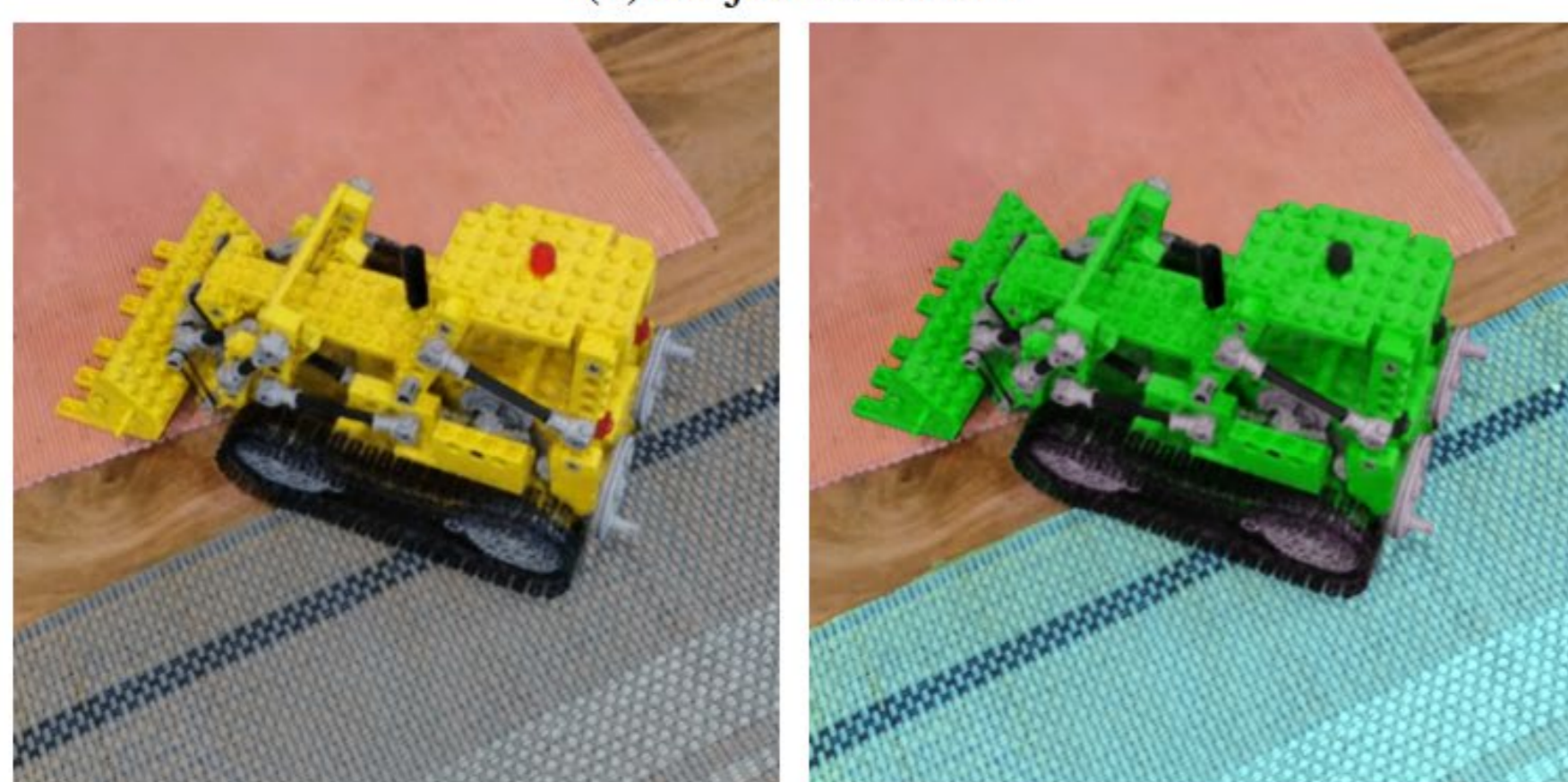$$\mathcal{L}_{TV} = \sum_{i,j \in \mathcal{N}} \|h_i - h_j\|_2^2$$



- **Find the best matching** γ that minimizes a generalized IoU α between 2D masks and object field channels

- **Supervise** object field with a matching loss and a false positives loss

- **Regularize** the object field to enforce spatial consistency with *Total Variation l*oss on the object field hash grid

## 3  Results


(a) Object removal


(b) Color editing


(c) Composition

- Class-agnostic segmentations on **Mip-NeRF360**

- Object field can be used to **condition rendering**

- This allows a wide range of **editing operations**

|           | DFFv2 | Instance-NeRF | Panoptic Lifting | PL+SAM | SAM (2D) | Ours   |
|-----------|-------|---------------|------------------|--------|----------|--------|
| IoU (%)   | 65.60 | 31.71         | 29.33            | 64.12  | 76.14    | **79.24** |
| BD (%)    | 74.57 | 53.75         | 55.41            | 78.32  | 87.12    | **87.51** |
| SBD (%)   | 73.82 | 44.19         | 41.73            | 72.90  | 83.39    | **85.34** |

→ The final segmentation does not require additional clustering

→ It natively provides a confidence estimates