# The Devil is in the Labels: Noisy Label Correction for Robust Scene Graph Generation
## ********** Supplementary Document **********

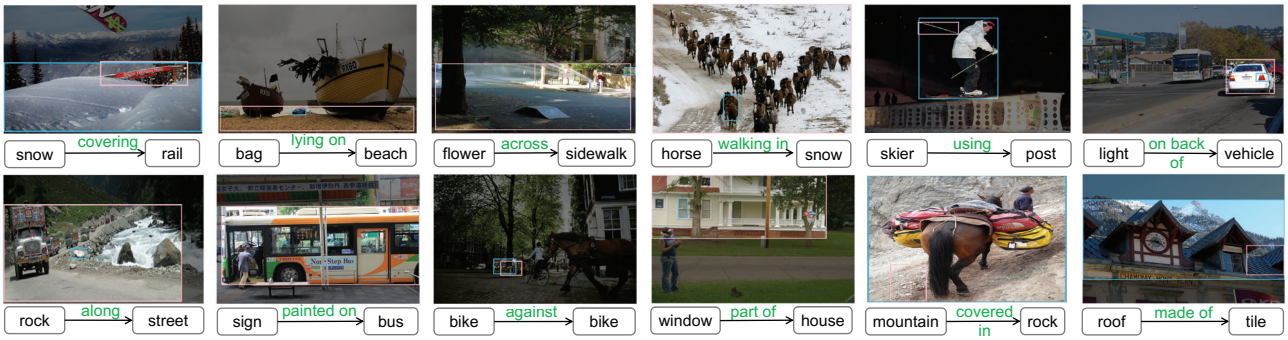


**Figure 1:** The examples of the triplets detected by Negative Noisy Sample Detection that do not appear in the training set.

## Appendix

This supplementary document is organized as follows:

- Implementation Details in Sec. A.
- Qualitative results of Neg-NSD in Sec. B.
- Influence of the cutoff distance in Pos-NSD in Sec. C.
- Different sample statistics in each component of NICE in Sec. D.
- Potential Negative Societal Impact in Sec. E.

## A. Implementation Details

**NICE Training Details.** In Neg-NSD, we used model Motifs [4] as OOD detection model $\mathrm{F}^{n}_{\mathrm{sgg}}$. The training settings (*e.g.*, learning rate and batch size) follow the same settings of [2] under PredCls task, except that it was trained with only foreground samples. In Pos-NSD, we used a pre-trained Motifs [4] provided by [2] as $\mathrm{F}^{p}_{\mathrm{sgg}}$ to extract triplet features (cf. $\boldsymbol{h}_i^k$ in Eq. 4) under PredCls task. The number of divided subsets was set to 4. In NSC, the $a$, $b$ and $c$ were set to 1, 0, and 10, respectively. Note that although we used two Motifs models (one is an off-the-shelf model) in NICE, we only need to try NICE for one time, and then we can use the obtained cleaner annotations for any SGG models.

**SGG Training Details.** Since NICE is a model-agnostic strategy, thus, for different baselines (*e.g.*, Motifs [4] and VCTree [3]), we followed their respective configurations[1].

## B. Qualitative Results of Neg-NSD

In Figure 1, we visualized some "unseen" visual relation triplet categories mined by Neg-NSD that never appear in the original VG dataset. Some of these triplets that are easily overlooked by the annotators, such as the relation `against` between `bike` and `bike`, or the relation `along` between `rock` and `street`. These harvested new visual relation triplet categories increase both the number and diversity of samples in tail categories.

## C. Influence of the Cutoff Distance in Pos-NSD

In the Pos-NSD module, we followed a previous clustering algorithm [1][2] to divide all positive samples into multiple sets based on the local density. This clustering algorithm is based on the assumption that *cluster centers are characterized by a higher density than their neighbors and by a relatively large distance from points with higher densities*.

[1] We utilized the SGG benchmark provided by [2] for all baselines.

[2] The clustering algorithm does not divide the samples into subsets from noisy to clean like us in Pos-NSD. It only makes the features of samples of the same cluster be close to each other, and be far away from other clusters.

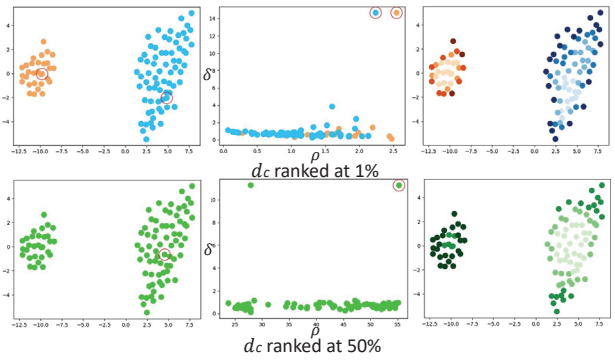


**Figure 2:** A toy example of the clustering results of hundreds of randomly generated samples. The clustering results for cutoff distance $d_c$ ranked at 50% and 1% , respectively. 1) Left: Sample distribution. Different colors correspond to different clusters, *i.e.*, two clusters for $d_c$ at 50% and one cluster for $d_c$ at 1%. 2) Middle: Local density $\rho$ and distance $\delta$ figure for clustering center decision. 3) Right: Local density distribution. The color from dark to light represents the local density from small to large. The cluster centers are circled in **red**.

To clearly understand the influence of cutoff distance $d_c$ on the final clustering results, there are two important concepts: local density $\rho$ and distance $\delta$. Specifically, for each sample $i$, its local density $\rho_i$ is the same as Eq. (5) in Pos-NSD, and its $\delta_i$ is defined as:

$$\delta_i = \begin{cases} \min_{j:\rho_j>\rho_i}(d_{ij}), & \text{if } \exists j \text{ s.t. } \rho_j > \rho_i, \\ \max(d_{ij}), & \text{otherwise .} \end{cases} \quad (1)$$

Thus, if sample $i$ has the largest local density over the whole set (*i.e.*, all samples with the same predicate category), $\delta_i$ represents the farthest distance between two samples in the whole set. Otherwise, $\delta_i$ represents the closest distance between sample $i$ and the sample $j$, and sample $j$ is a sample with a higher local density (*i.e.*, $\rho_j > \rho_i$).

Based on the two concepts ($\rho$ and $\delta$), we can refer to Algorithm 1 to cluster all samples. Firstly, we select the clustering centers according to both local density $\rho$ and distance $\delta$, and then cluster the remaining samples according to distance $\delta$. Specifically, for the selection of clustering centers, they all have large local density $\rho$ and distance large $\delta$, as mentioned in [1]. Smaller cutoff distance $d_c$ is more likely to lead to multiple samples with large local density $\rho$ and distance large $\delta$, resulting in multiple cluster centers. As shown in Figure 2, when the cutoff distance $d_c$ is small (ranked at 1%), there are two samples with large $\rho$ and $\delta$, so two clusters are generated; while when the $d_c$ is large (ranked at 50%), there is only one sample with $\rho$ and $\delta$, so only one cluster is generated. For the other samples, they are arranged according to local density $\rho$ and then assigned to cluster of sample nearest it with a higher local density *i.e.*, the cluster of the sample with which $\delta$ is calculated, and the results are displayed in Figure 2(right).

In this paper, our purpose is only to detect all noisy samples (in different clusters), rather than dividing these samples into multiple semantic clusters. Considering the distribution of local density $\rho$ and distance $\delta$ in Figure 2, $\rho$ has better discrimination of noisy samples, while $\delta$ does not. Hence, in Pos-NSD, we directly adopt K-Means algorithm to divide different subsets according to the degrees of noise (local density $\rho$). To prove that small cutoff distance can better detect the samples at margin (noisy samples) of each cluster, we show the distribution of local density in Figure 2. As shown in Figure 2, when the cutoff distance $d_c$ is small, local density is arranged from small to large in two centers, and the local density intervals of the two clusters are consistent. However, when the cutoff distance $d_c$ is large, almost one cluster has a smaller local density interval than the other, leading the entire cluster with a relatively small local density interval to be considered as noisy. Therefore, setting a small cutoff distance $d_c$ and judging the noise degree according to local density $\rho$ can distinguish noise samples located in different semantic clusters.

## D. Different Sample Statistics in Each Component of NICE

In this section, we made a statistics on the number of training samples of the cleaner version of training dataset after the NICE training. Specifically, the number of different categories after applying each component are reported in Table 1 and Figure 2. Based on these results, we have the following observations:

1. **Overview**: As each module is stacked, the sample number of head categories with less informative predicates decreases significantly, while the sample number of tail categories with more informative predicates increases, which can alleviate the long-tail distribution to some extent.

2. **Neg-NSD**: After performing Neg-NSD on the original dataset, the number of samples in some tail categories (*e.g.*, `covering`, `covered in` and `painted on`) increases in Table 1 (# 2) (92.3k vs.5.6k).

3. **Pos-NSD**: As reported in Table 1 (# 3), only the number of clean positive samples (not in the noisiest subset) is reported. Compared with the body and tail categories, more noisy samples in head categories have been screened out (95.5k, 15.0k and 15.9k decreases in head, body and tail categories). This is because the predicates in head categories always have multiple semantic clusters, and the total number of samples at the margin of each semantic cluster is greater than that of

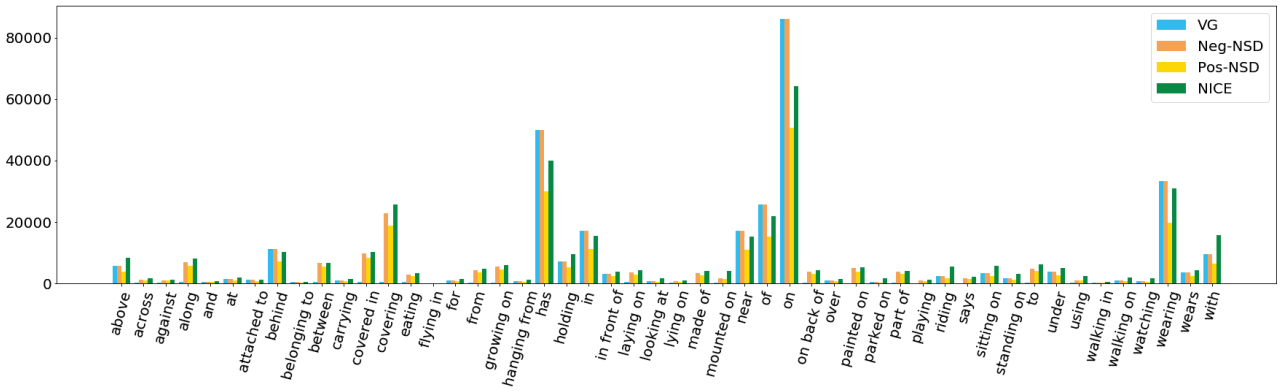


**Figure 3:** VG dataset statistics in different modules of NICE.

**Algorithm 1:** Clustering

**Input:** Distance matrix $\boldsymbol{D}^k = (d_{ij}^k)_{N\times N} \in \mathbb{R}^{N\times N}$, cutoff distance $d_c$, local density threshold $\rho_c$ and distance threshold $\delta_c$ for screening cluster centers.

**Output:** Clustering Category for each sample $\{c_i\}_{i=1}^N$.

// Calculate local density $\rho$
**for** $i = 2, 3, \cdots, N$ **do**
  $\rho_i^k = \sum_j \mathbf{1}(d_c^k - d_{ij}^k)$
Arrange $\{\rho_i\}_{i=1}^N$ in descending order and get a new index $\{q_i\}_{i=1}^N$ that satisfies $\rho_{q_1} \geq \rho_{q_2} \geq \cdots \geq \rho_{q_N}$
// Calculate distance $\delta$
**for** $i = 2, 3, \cdots, N$ **do**
  $\delta_{q_i} = d_{max}$
  **for** $j = 1, 2, \cdots, i-1$ **do**
    **if** $d_{q_i q_j} < \delta_{q_i}$ **then**
      $\delta_{q_i} = d_{q_i q_j}, n_{q_i} = q_j$
$\delta_{q_1} = d_{max}, n_{q_1} = -1$
// Select the cluster centers
$k = 0$
**for** $i = 1, 2, \cdots, N$ **do**
  **if** $\rho_{q_i} \geq \rho_c$ *and* $\delta_{q_i} \geq \delta_c$ **then**
    $c_{q_i} = k, k = k + 1$
  **else**
    $c_{q_i} = -1$
// Clustering the other samples
**for** $i = 1, 2, \cdots, N$ **do**
  **if** $c_{q_i} = -1$ **then**
    $c_{q_i} = c_{n_{q_i}}$

predicates in the tail categories with only single one semantic cluster.

| # | Components | | | Sample Number | | | |
|---|---|---|---|---|---|---|---|
| | N-NSD | P-NSD | NSC | head | body | tail | total |
| 1 | ✗ | ✗ | ✗ | 240,672 | 51,095 | 5,551 | 297,318 |
| 2 | ✓ | ✗ | ✗ | 240,672 | 51,095 | 92,256 | 384,023 |
| 3 | ✓ | ✓ | ✗ | 145,142 | 36,072 | 76,286 | 257,500 |
| 4 | ✓ | ✓ | ✓ | 198,326 | 77,872 | 107,825 | 384,023 |

**Table 1:** The positive sample number of head, body and tail predicates after superimposing each component of NICE in VG dataset. The third row (#3) shows only the number of clean positive samples.

4. **NSC**: After applying the NSC module, the number of some samples in tail and body predicates is further increased, as shown in Table 1 (# 4 vs. #2) (77.9k vs. 51.1k and 107.8k vs. 92.3k). The reasons may come from that the semantics of predicates in the head categories often overlap with multiple predicates in tail categories, which increases the likelihood that the noisy samples of each semantic cluster are assigned with cleaner labels in tail categories.

## E. Potential Negative Societal Impact

There are two possible potential negative societal impacts of our NICE: 1) Since OOD detection model in Neg-NSD will assign new triplet categories to all detected noisy negative samples, the assigned triplet categories may be unreasonable, such as ⟨woman-eating-plate⟩, ⟨person-eating-person⟩, and ⟨child-eating-wire⟩. After equipping SGG models (trained with these unreasonable annotations) into some downstream tasks (*e.g.*, image captioning or question answering), these unreasonable triplets may have a misleading impact on human cognition of living habits. Of course, this situation can be easily avoided, as long as we add some constraints about the generated triplet categories, such as filtering out all impossible triplets categories. 2) Since the basic target of Neg-NSD is mining

missing annotated triplets, it can harvest much more training samples without any threshold constraints (*e.g.*, predicate confidence). Accordingly, a larger amount of training samples will be generated. If all these low-quality training samples are used for model training, it may lead to a huge waste of computing resources. Similarly, we can manually set some thresholds to control the number of generated samples and avoid this negative impact.

## References


[1] Alex Rodriguez and Alessandro Laio. Clustering by fast search and find of density peaks. *Science*, pages 1492–1496, 2014. 1, 2

[2] Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. Unbiased scene graph generation from biased training. In *CVPR*, pages 3716–3725, 2020. 1

[3] Kaihua Tang, Hanwang Zhang, Baoyuan Wu, Wenhan Luo, and Wei Liu. Learning to compose dynamic tree structures for visual contexts. In *CVPR*, pages 6619–6628, 2019. 1

[4] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. Neural motifs: Scene graph parsing with global context. In *CVPR*, pages 5831–5840, 2018. 1