One-sample Guided Object Representation Disassembling (Paper ID 2171)

We sincerely thank all reviewers for the constructive comments. We provide short responses here due to the page limit.

To Reviewer #1

Q1: a) How does the representation length influence disassembling? b) Is there a common set of parameters?
A1: Smaller length yields better representation disassembling, while larger length leads to better image reconstruction. For images with complex scenes, the representation length should be set larger. b) Yes. They are given in Line 199.
Q2: Will the method still work on multiple object image? Q3: What is the exact goal of the dual-swap step?
A2: In this case, features of multiple objects will be disassembled into the same part. Visual results are shown in Fig. I.
A3: The dual-swap step is devised for generating self-supervised information for the unannotated samples.
Q4: a) Number of supervised samples in DSD. b) Why are the results of MONet [6] and IODINE [12] different?
A4: a) Around $5,000$ samples augmented with annotated one-samples are used. b) MONet [6] and IODINE [12] learn object representations and corresponding masks. The fixed masks lead to the non-swappable defects of [6,12].
Q5: a) Why choosing two-layer MLP as classifier? b) Does the classifier affect the disassembling effect?
A5: a) It is a simple classifier and trainable by gradient descent. b) Yes. Simpler but not too simple classifiers better reflect the properties of the representation, which favors the disassembling.

To Reviewer #2

Q1: Extension to multi-object setting. Q2: Disentangling features of scene objects seems a goal of the method.
A1: Thanks for the comment. The method is indeed able to handle images with multiple objects, in which case features of multiple objects will be extracted together and denoted by the same part of the representation. Fig. I below shows the visual results of multi-object samples in COCO and miniImageNet (a newly added dataset). More results will be added to the revision. **A2**: In fact, disentangling features of objects is not our goal. Rather, our method aims to disassemble features between one-sample guided objects and the background. Features of guided objects are denoted by one part of the representation, while features of unguided objects, which are treated as the background, are denoted using the remaining part of the representation. We will clarify this in the revision.
Q3: Writing issues: unclarity of the introduction and conclusion; non-standard or unintuitive terminologies.
A3: Thanks for the nice suggestions. We will clearly list the main goals in introduction and amend terminologies.
Q4: Improving related works. Q5: Meaning of prior methods not working on complicated backgrounds.
A4: Thanks. We will revise this part and provide clearer distinctions between our method and prior ones. **A5:** We meant that, prior methods like MONet [6] work on synthetic images only but not real-world ones. We will clarify this.

Figure I: Visual results reconstructed with original or object-swapped representations on multi-object images.

To Reviewer #3

Q1: How robust is the method to the parameters? Q2: Will the method still work for multi-object images?
A1: Please refer to **A1(b)** of **R#4**. **A2:** Yes. Fig. I shows the results on multi-object images.
Q3: Why choosing only ten categories from COCO? Q4: The blurred results in Fig. 3. Other architectures?
A3: The sample numbers in these ten categories are balanced. **A4:** Small representation length favors disassembling but harms image reconstruction, leading to the blurred results. We tested the method with other architectures, yet results showed that changing architecture barely helps. More labeled samples help improve the blurred results.
Q5: Connection with 'fuzzy logic'? Q6: Image size? Q7: How to find the best representation length?
A5: No. **A6:** SVHN:32×32; Others: 64×64. **A7:** Setting larger representation length for a more complicated image.

To Reviewer #4

Q1: Are samples of COCO cropped to patches and resized? Q2: The sensitivity of hyper-parameters.
A1: Thanks. We only resized samples to 32×32 (SVHN) or 64×64 (others), but did not crop them. **A2**: The vital parameters (β, ρ) and δ respectively control the disassembling on annotated and unannotated samples. The (AMS, AIS) for setting (β, ρ, δ) to be $(1, 100, 5)$, $(10, 1000, 5)$, and $(10, 1000, 50)$ are $(14.32, 7.83)$, $(13.69, 6.51)$, and $(13.52, 6.31)$, respectively. Reasonably large β and ρ favor disassembling. More results will be added to the revision.
Q3: Showing results on miniImageNet. Q4: Classification equations are expressed in an unfamiliar way.
A3: Fig. I(b) shows visual results; the AMS and AIS scores for (S-AE, DSD, MONet, IODINE, Ours) are respectively $(14.45, 13.36, 10.43, 17.49, 6.98)$ and $(7.81, 5.76, 9.82, 18.93, 4.03)$. **Q4**: Our loss requires computing each value in the predicted class vector. To be consistent, therefore, equations are written in an expanded form. We will clarify this.
Q5: Explaining why supervised methods (S-AE) are worse than semi-supervised methods (One-GORD).
A5: Thanks. With only labels and a two-layer MLP classifier, S-AE is not able to effectively extract the entire discriminant features. However, dual-swap and object reconstruction of our method favor representation disassembling.
Q6: Try ResNet as the decoder of MoNet or IODINE to deal with complex backgrounds. Q7: λ_{cla}^{d}? and typos.
A6: Thanks. Still, the modified MoNet and IODINE do not work. One reason is that the Gaussian-distribution assumption facilitates decomposition but limits the representation's capacity. **A7:** It should be λ_{rec}^{d}. We will fix typos.