Thank you to the reviewers for providing helpful feedback which will improve the paper. We appreciate that the reviewers found our method novel, simple and generic yet effective, and compelling for its potential impact of training sound separation on real-world mixtures. We hope that the specific points below address outstanding issues.

			FUSS			YFCC100M
Supervised	p_0	Unsupervised	MSi	SS	MoMi	MoMi
FUSS	0.2	YFCC100M	13.6	34.6	**6.8**	9.0
FUSS	0.0	YFCC100M	12.5	11.1	6.6	9.0
FUSS	0.2	–	**13.7**	**35.6**	6.5	4.8
FUSS	0.0	–	12.4	11.9	6.4	4.9
–	–	FUSS	11.9	3.6	5.2	8.4
–	–	YFCC100M	10.8	6.9	4.8	**10.6**

1. *R3, R4: Concern about evaluation, that we train on synthetic mixtures, while the ultimate goal is to use real mixtures.* We agree this is a limitation for the results on some tasks. However, the paper does show results with training on real in-the-wild Freesound mixtures for universal separation (see Sec. 4.3.). Evaluation is another matter: we used synthetic data with source-level references so that we can measure objective source-level SNR. Such evaluations are not possible with real-world mixtures, but we plan on adding some surrogate evaluations in the final revision. We plan to separate mixtures of 2 in-the-wild mixtures (MoMs) into their constituent sources, and measuring SI-SNRi of the RemixPIT-reconstructed *mixtures*, called MoMi, which correlates with source-level MSi (see table).

2. *R1: Section 4 is main weakness*. We revised section 4 to be more clear, in line with other specific responses here.

3. *R1 1) Meaning of matched vs mismatched?*. "Matched" means e.g. train on anechoic, test on anechoic, and "mismatched" means e.g. train on anechoic, test on reverberant, and vice versa. We will clarify this in the final revision.

4. *R1 2) Section 4.1.: The 1+1 source mixture case exactly corresponds to the supervised case.* $p\%$ of each training batch are unsupervised mixtures drawn from a $p\%$ subset of a dataset of 1-or-2-source mixtures. The remainder of the batch are supervised mixtures with corresponding reference sources drawn from the subset complement. The model has to infer the number of sources for unsupervised mixtures. Although easier than 2-source MoMs, we think that 1-or-2 sources is more realistic, because single-speaker audio is typical in the real world.

5. *R1 4) Data availability*. We intend to release data recipes for all non-released data, including reverberant version of Libri2Mix, Librivox+Freesound speech enhancement data, and clips used from Freesound and YFCC100M.

6. *R1 5) How was it ensured that FUSS and freesound do not intersect?* There may be some overlap. To avoid this, we reran experiments on independent YFCC100M audio as unsupervised in-the-wild data (see table), with similar results.

7. *R1 6) L274: Meaning of "randomly zeroing out one of the supervised mixtures with probability p0"?* With probability p_0, one of the mixtures in the supervised MoM is set to zero, along with its sources, equivalent to using a single mixture.

8. *R1 7) L278: MSi and SS not properly explained.* Sorry for the lack of explanation, we've clarified the text. MSi is SI-SNRi of sources from mixtures containing 2 or more sources, and SS is absolute SI-SNR for single-source mixtures.

9. *R1 10) L326: "remains challenging because of the lack of ground truth"* The reviewer points out a useful method for collecting more realistic supervised data. Unfortunately, it is not ideal as background noises in each recording will be added together in the resulting mixtures. Note that RemixPIT performs similar remixing of recordings except that it also handles separation of background noise as well as overlapping speech in each mixture. So RemixPIT could be applied to CHiME-5, for example, which contains significant non-stationary background noise (the other CHiME datasets are synthetic). However, SNR evaluation without signal-level ground truth would still be a problem with CHiME-5.

10. *R2: Can RemixPIT be used in other domains e.g. images?* Yes, and such experiments are interesting future work. Though visual objects are more commonly simply occluded in reality, transparency is an analogue of audio mixing.

11. *R3: Brute-force RemixPIT implementation won't scale to more mixtures*. As noted in the paper, we intend to defer this to future work. We believe exact solutions have exponential complexity, but good approximations can suffice under certain assumptions. Also note that using more mixtures may increase the mismatch with the unmixed data.

12. *R3: How are rightmost models in section 4.1 plots trained?* From scratch with RemixPIT on unsupervised data only.

13. *R3: From Table 1, the improvement over the pure-supervised learning method is minor.* Good point, we'll be more precise about claims of improvement in revision. Our new evaluations (table above) that show semi-supervised models achieve comparable performance to purely supervised models for FUSS, but also improve MoMi on YFCC100M.

14. *R4: Using just up to 2 sources in speech separation seems a bit limiting again.* We agree, but for the speech separation experiments we chose to focus on standard tasks with ground-truth source references to measure performance.

15. *R4: Statistically significant change in SI-SNRi?* In the final revision, we plan to report this for all experiments.

16. *R4: Performance of fully-supervised speech separation models is sometimes exceeded by fully unsupervised models*. The supervised models in our initial submission were 4-output models trained on supervised MoMs with 2-4 sources, and thus is still a bit mismatched to 2-source test mixtures. However, using MoMs implicitly remixes sources, acting as training data augmentation. This helps on WSJ0-2mix, which is small. Since submission, we reduced this mismatch and improved matched supervised performance by training on individual mixtures instead of MoMs while also using explicit source remixing augmentation. We'll include these results in the final revision.

17. *R4: Questions about speech enhancement*. Training from just mixtures is possible, but we would need another method to determine which estimated source is speech. Light supervision of speech presence and the RemixPIT constraints force the model to always output speech as source 1. Noise is matched (i.e. always drawn from Freesound).