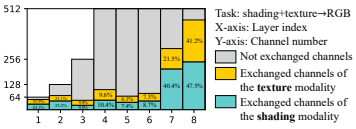
Table A: More ablation studies on dataset NYUDv2 with ResNet101. All experiments use unshared BNs. Ensemble indicates using ensemble to mix predictions of RGB&Depth.

Method	Convs	Ensemble	Mean IoU
Summing activation	Shared	✓	48.1 (%)
Summing activation + soft gate	Shared	✓	48.4 (%)
CBN	Unshared	✓	48.3 (%)
CBN	Shared	✓	48.9 (%)
Ours (concat-fusion block)	Shared	×	50.8 (%)
Ours	Shared	✓	**51.1** (%)

Figure A: Percentage of exchanged channels on dataset Taskonomy.



We thank all reviewers for their constructive comments and address the reviewers' concerns point by point below.

[To Reviewer #1:] Q1. More mid-fusion baselines: Thanks for the suggestions. We have conducted extra ablation studies raised by the reviewer in Table A. For the CBN strategy [DeVries et al. 2017] specifically, we modulate the BN of one modality conditional on the other. In this sense, CBN performs cross-modal message passing via BN modulation, which is clearly different from our method that directly exchanges channels between different modalities for fusion. As expected, all ablations (including CBN) are inferior to our method, which again verifies the validity of the proposed mechanism. Per the reviewer's concern, experiments on other datasets will also be added into the final version.

Q2. More explanations on fundamental parts: 1. We can **not** perform channel exchanging without ℓ_1, as it enables the discovery of unnecessary channels and comes as a pre-condition of Theorem 1. Naively exchanging, *e.g.*, 30% channels only gets IoU 47.2. **2.** We provide the comparison with the modulation approach in Table A. **3.** We can conduct channel exchanging upon unshared CNN, which has been evaluated in Table 6 (the 5th and 7th rows). Promisingly, its full-channel case getting 49.1 still outperforms baselines in Table 2 under the same setting. We believe our method is generally useful and channels are aligned to some extent given unshared CNN. **4.** We have discussed why regularizing half of the channels is beneficial in L155-161 and provided its empirical evidence in Table 1 (the 5th and 6th rows).

Q3. The rationality of our theorems: Theorem 1 is meaningful and crucial. Yes, ℓ_1 makes the parameters sparse, but it can not tell if each sparse parameter will keep small in training considering the gradient in Eq.(4). Conditional on BN, Theorem 1 proves that $\gamma = 0$ is attractive, which is nontrivial and novel. **Corollary 1** states that f' is more expressive than f when $\gamma = 0$, and thus the optimal f' always outputs no higher loss, which, yet, is not true for arbitrary f' (*e.g.* $f' = 10^6$). Besides, Corollary 1 holds upon unshared CNN (see L191) and is consistent with Table 6 in the unshared scenario (full-channel: 49.1 vs half-channel: 48.5), although full-channel exchanging is worse under the sharing setting.

Q4. On recommended references: We thank for the comment, will cite all the raised references (including the modulation-based and CBN methods), and will refresh the references related to early/late fusion and the fusion taxonomy. Our claims of "Introduction 3rd paragraph" are intuitively supported by the observations from Figure 1, namely, the aggregation-based fusion integrates multiple networks into one single branch and the alignment-based fusion conducts an indirect way of fusion by training. We will further clarify this in the final version.

Q5. Other comments: 1. In L201, "the standard setting" refers to train/test split, while "commonly/same with our setting" is on architecture design, as already described in L236-238. **2.** We agree our current version is inapplicable for heterogeneous fusion (*e.g.* CNN+RNN). That is why we, explicitly and honestly, point it out in L170-176. Given the effectiveness on homogeneous fusion, we conjecture that adding extra homogeneous MLP layers upon each CNN/RNN enables channel (or neural unit) exchanging for heterogeneous fusion, which will be left for future exploration.

[To Reviewer #2:] Q1. More experiments: We have carried out random exchanging like ShuffleNet or directly discarded unimportant channels without channel exchanging on NYUDv2, the IoUs of which are 46.8 and 47.5, respectively. Both are much worse than our method (51.1). We are willing to add these ablations into the final version.

Q2. Why use the mean: We have tried weighted summation of channels, but do not observe clear improvement (FID from 60.90 to 60.94 on the translation task Depth+Normal+Texture→RGB). We conjecture that γ of each modality in Eq. (6) already contributes to the weights, thus the simple and parameter-free mean delivers promising performance.

[To Reviewer #3:] We thanks for the positive comments, and are sorry for the unclear presentations on Theorem 1. We would like to highlight that our conclusion always holds regardless of the value of β. Our proof (in appendix) is interested in the gradient $dL/d\gamma$ only around $\gamma = 0$ (not other points) and finds that the probability of staying around $\gamma = 0$ is large and hits the Gaussian peak, even for the extreme case (β or $dL/d\mathbf{x}'$ is non-zero) raised by the reviewer. By mentioning $\beta = 0$ in L179, we initially tend to simplify the analysis of Corollary 1. Yet, our current proof of Corollary 1 does not require this assumption, and we will remove $\beta = 0$ to avoid confusion in the final version.

[To Reviewer #4:] Q1. Novelty and contributions compared to [38]: Our paper is inspired but differs from [38] in two aspects: **1.** while [38] aims at model compression, our paper attempts to conduct multimodal fusion. Thus upon [38], we employ channel exchanging for message passing, which is simple yet effective and novel compared to current fusion methods. **2.** Theorem 1 is related to [38], but [38] never confirms that unnecessary channels during training can hardly recover, and we rigorously prove it by exploring the behaviors of the ℓ_1 norm and BN layers.

Q2. More discussions: 1. Thanks for the reminding, and we have reported the percentage of exchanged channel for every layer in Figure A. It shows that upper layers are more active in exchanging. **2.** Please refer to Q5.2, R#1 for the discussion on heterogeneous fusion. **3.** Regarding the three-modal example (A, B, and C) raised by the reviewer, the ℓ_1 norm is conducted on disjoint channels of different modalities. It means in Eq.(6) that for each channel, only one modality will meet the criterion (scaling factor lower than the threshold). This setting will avoid unavailing exchanging as already described in L160-161. **4.** Not all experiments specify two modalities. We have tested the effectiveness of our method with modalities more than two in Table 5 (also Table 7, 8 in appendix), where our method still makes effect.

Q3. Other comments: We use V100 with GPU memory 32G, and the segmentation needs 20G. We will detail the GPU type in the final version. Applying the ℓ_1 norm alone delivers no improvement, but it is necessary for the discovery of unnecessary channels and channel exchanging (see also our response in Q2.1, R#1).