NeurIPS | 2020

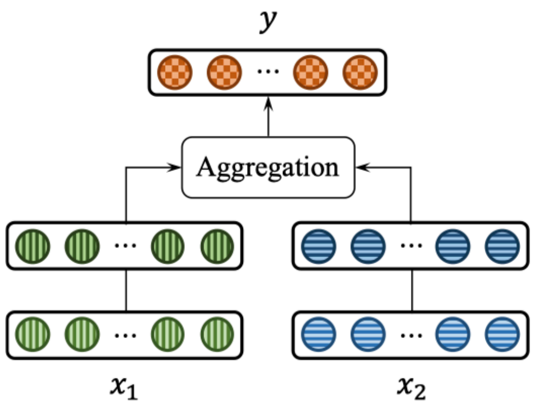
# Deep Multimodal Fusion by Channel Exchanging

Yikai Wang[1], Wenbing Huang[1], Fuchun Sun[1], Tingyang Xu[2], Yu Rong[2], Junzhou Huang[2]

[1] Tsinghua University [2] Tencent AI Lab

# NeurIPS | 2020

➢ **A sketched comparison between existing fusion methods and ours :**



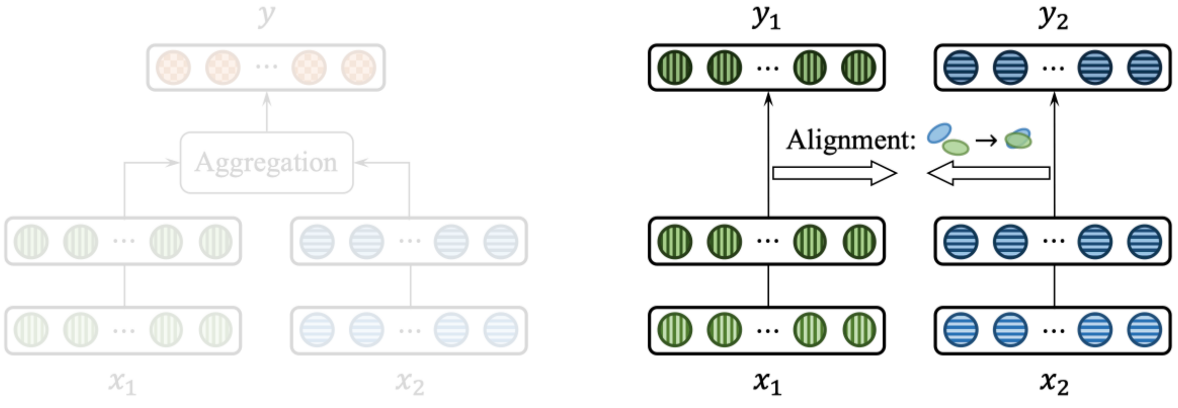

**(a)** Aggregation-based fusion

- The **aggregation-based fusion** processes each modality with a separate sub-network and then combine all their outputs via an aggregation operation, e.g. averaging, concatenation, or adding with attention.


[1] Seungyong Lee et al. “RDFNet: RGB-D Multi-level Residual Feature Fusion for Indoor Semantic Segmentation”. In: ICCV. 2017.
[2] Caner Hazirbas et al. “FuseNet: Incorporating Depth into Semantic Segmentation via Fusion- Based CNN Architecture”. In: ACCV. 2016.
[3] Abhinav Valada et al. “Self-Supervised Model Adaptation for Multimodal Semantic Segmentation”. In: IJCV. 2020.

# NeurIPS | 2020

➢ **A sketched comparison between existing fusion methods and ours :**



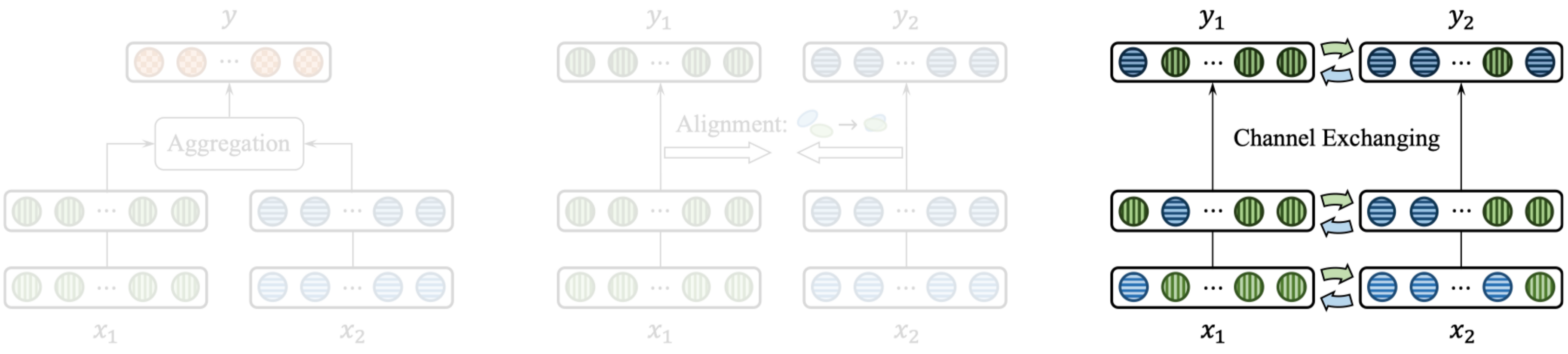

(a) Aggregation-based fusion (b) Alignment-based fusion

- The **alignment-based fusion** leverages an alignment loss $\mathrm{Alig}_{f_{1:M}}(\boldsymbol{x}^{(i)})$ (usually specified as MMD) for capturing the inter-modal concordance while keeping the outputs of all sub-networks.


[4] Jinghua Wang et al. "Learning Common and Specific Features for RGB-D Semantic Segmen- tation with Deconvolutional Networks". In: ECCV. 2016.
[5] Yanhua Cheng et al. "Locality-Sensitive Deconvolution Networks with Gated Fusion for RGB-D Indoor Semantic Segmentation". In: CVPR. 2017.
[6] Sijie Song et al. "Modality Compensation Network: Cross-Modal Adaptation for Action Recognition". In: IEEE Trans. Image Process. 2020.

# NeurIPS | 2020

➢ **A sketched comparison between existing fusion methods and ours :**



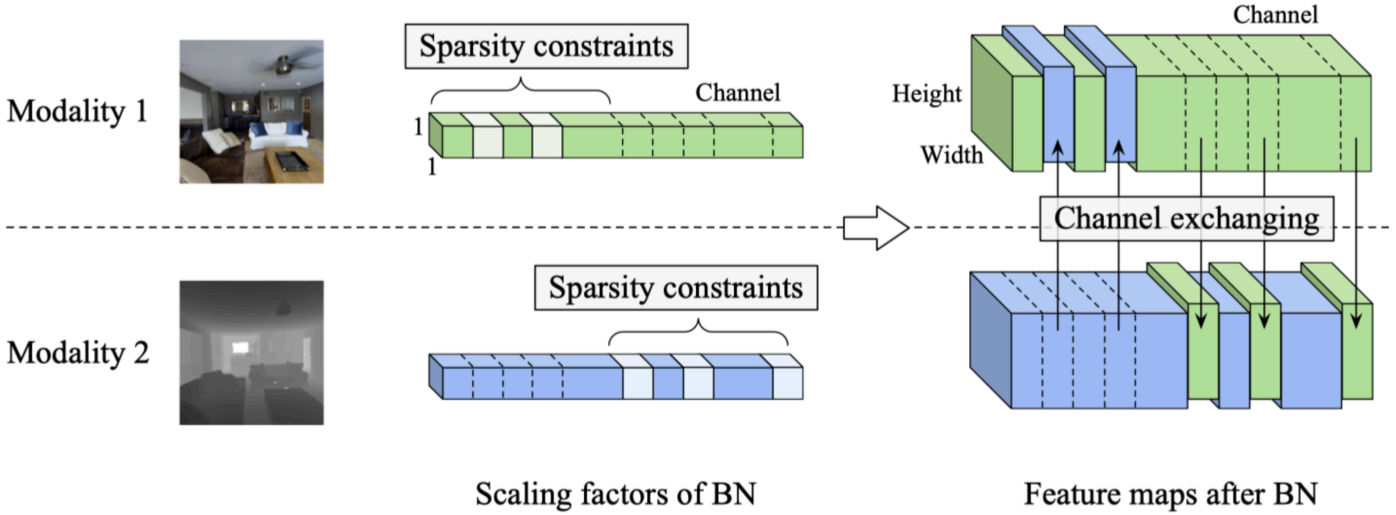

(a) Aggregation-based fusion (b) Alignment-based fusion **(c)** Ours

- This work proposes **Channel-Exchanging-Network (CEN)** for multimodal fusion, in which:
  - ✓ A **global criterion** is applied as a **self-guidance** during training for adaptive feature fusion;
  - ✓ Fusion can take place at **every layer throughout encoder**, instead of several pre-designed fusion positions like existing methods;
  - ✓ The multimodal architecture is almost as **compact** as a unimodal network, **with zero fusion parameter**.

# NeurIPS | 2020

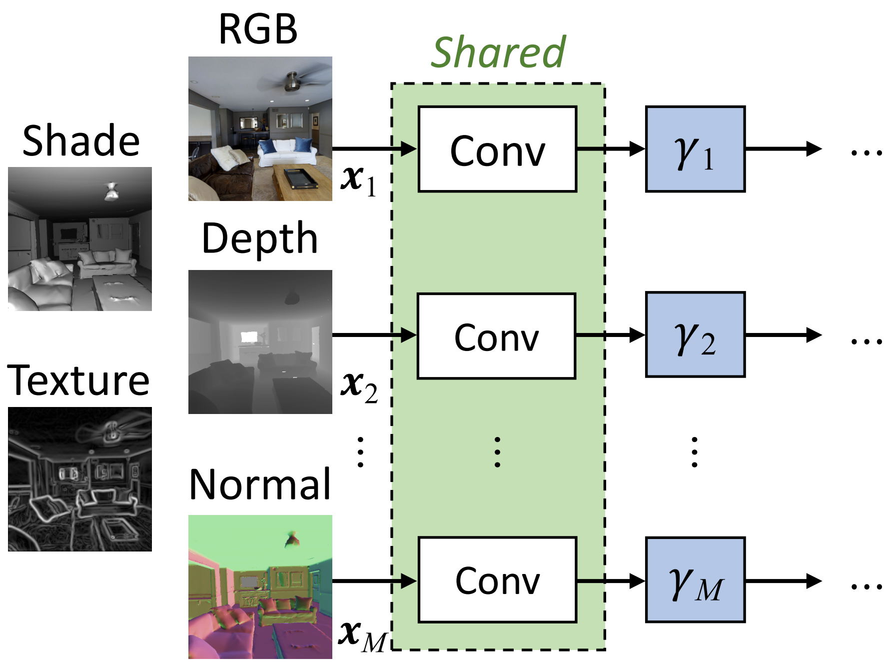
- **Summary of the overall method: channel exchanging by comparing BN scaling factor**
  - Create **sparse activations** by using a L1 norm over the **BN scaling factors**;
  - **Exchange an activation** when its BN scaling factor is **lower than a threshold**.
- **Details, the whole optimization objective of our method is:**

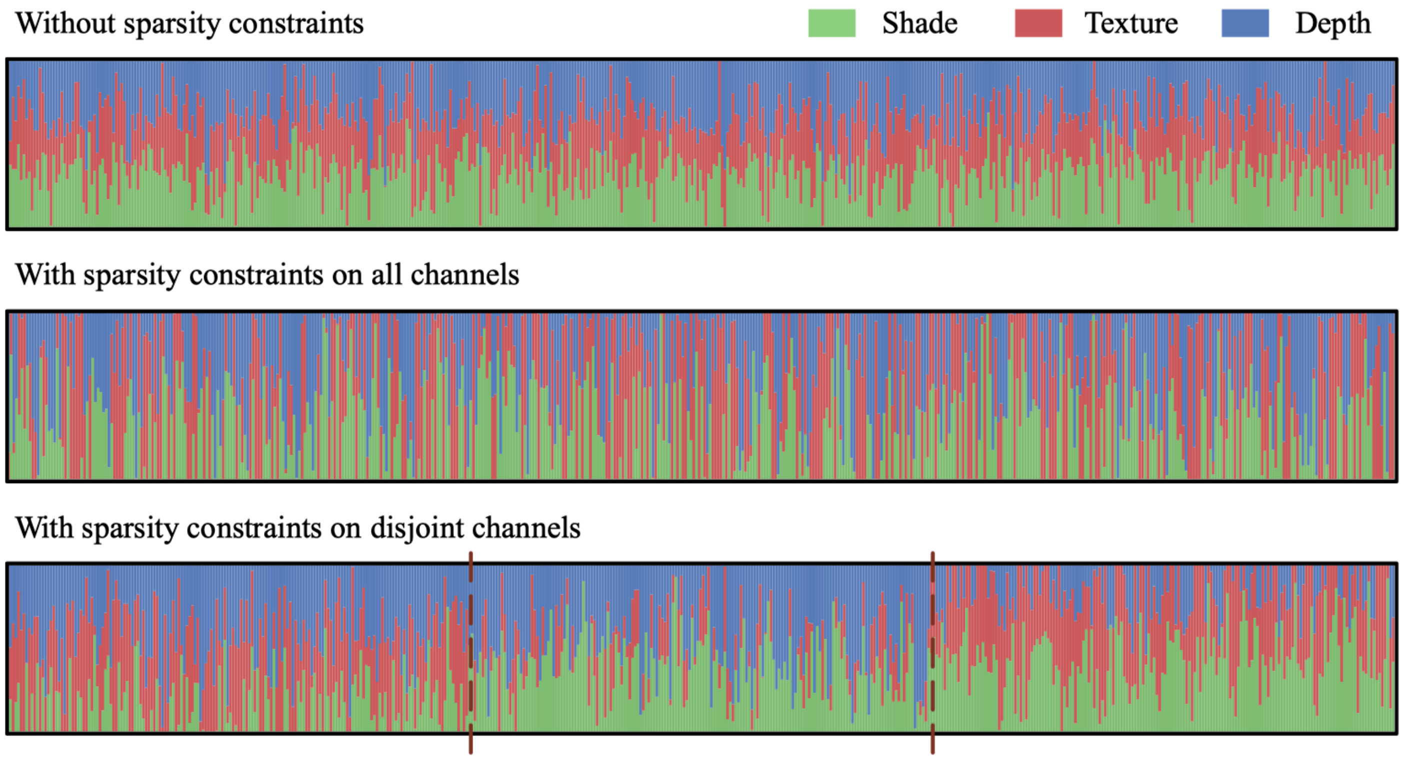
$$\min_{f_{1:M}} \frac{1}{N}\sum_{i=1}^{N}\mathcal{L}\left(\sum_{m=1}^{M}\alpha_m f_m(\boldsymbol{x}^{(i)}), \boldsymbol{y}^{(i)}\right) + \boxed{\lambda\sum_{m=1}^{M}\sum_{l=1}^{L}|\hat{\gamma}_{m,l}|} \quad s.t. \sum_{m=1}^{M}\alpha_m = 1$$



## NeurIPS | 2020

- **Additionally, we use sub-network sharing with independent BNs:**
  - Better for channel alignment, and capture the common patterns in different modalities;
  - Decoupled scaling factors can evaluate the importance of the channels of different modalities.



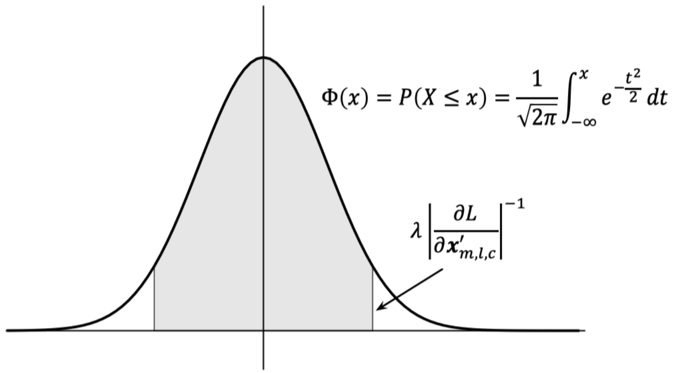



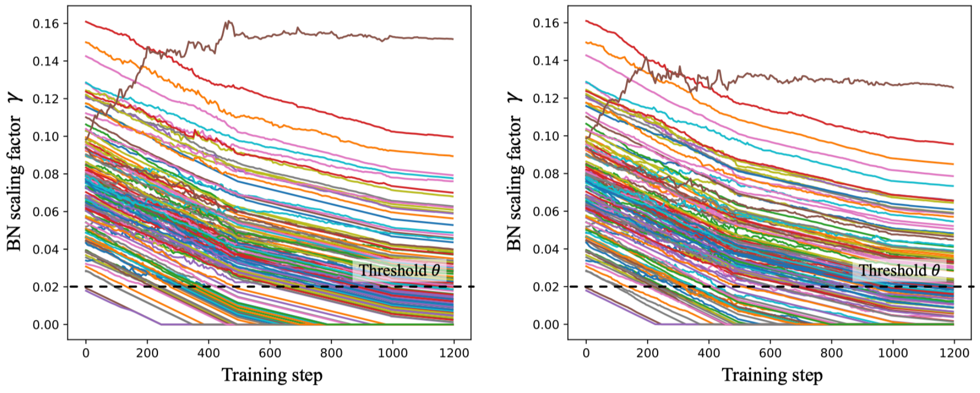

Proportion of scaling factors for each modality: $\gamma_c^{m,l,c}/(\gamma_c^{1,l,c}+\gamma_c^{2,l,c}+\gamma_c^{3,l,c})$

# NeurIPS | 2020

$$\bigstar: \quad \min_{f_{1:M}} \frac{1}{N}\sum_{i=1}^{N}\mathcal{L}\left(\sum_{m=1}^{M}\alpha_m f_m(\boldsymbol{x}^{(i)}), \boldsymbol{y}^{(i)}\right) + \boxed{\lambda\sum_{m=1}^{M}\sum_{l=1}^{L}|\hat{\gamma}_{m,l}|} \quad s.t. \sum_{m=1}^{M}\alpha_m = 1$$

- **Analysis:**
  - ✓ **Theorem 1.** Suppose $\{\gamma_{m,l,c}\}_{m,l,c}$ are the BN scaling factors of any multimodal fusion network (without channel exchanging) optimized by Equation ★. The probability of $\gamma_{m,l,c}$ being attracted to $\gamma_{m,l,c} = 0$ during training (a.k.a. $\gamma_{m,l,c} = 0$ is the local minimum) is equal to $2\Phi\left(\lambda\left|\frac{\partial L}{\partial x'_{m,l,c}}\right|^{-1}\right) - 1$, where $\Phi$ derives the cumulative probability of standard Gaussian.
  - ✓ **Corollary 1.** If the minimal of Equation ★ implies $\gamma_{m,l,c} = 0$, then the channel exchanging (assumed no crossmodal parameter sharing) will only decrease the training loss, *i.e.* $min_{f'_{1:M}} L \le min_{f_{1:M}} L$, given the sufficiently expressive $f'_{1:M}$ and $f_{1:M}$ which denote the cases with and without channel exchanging, respectively.



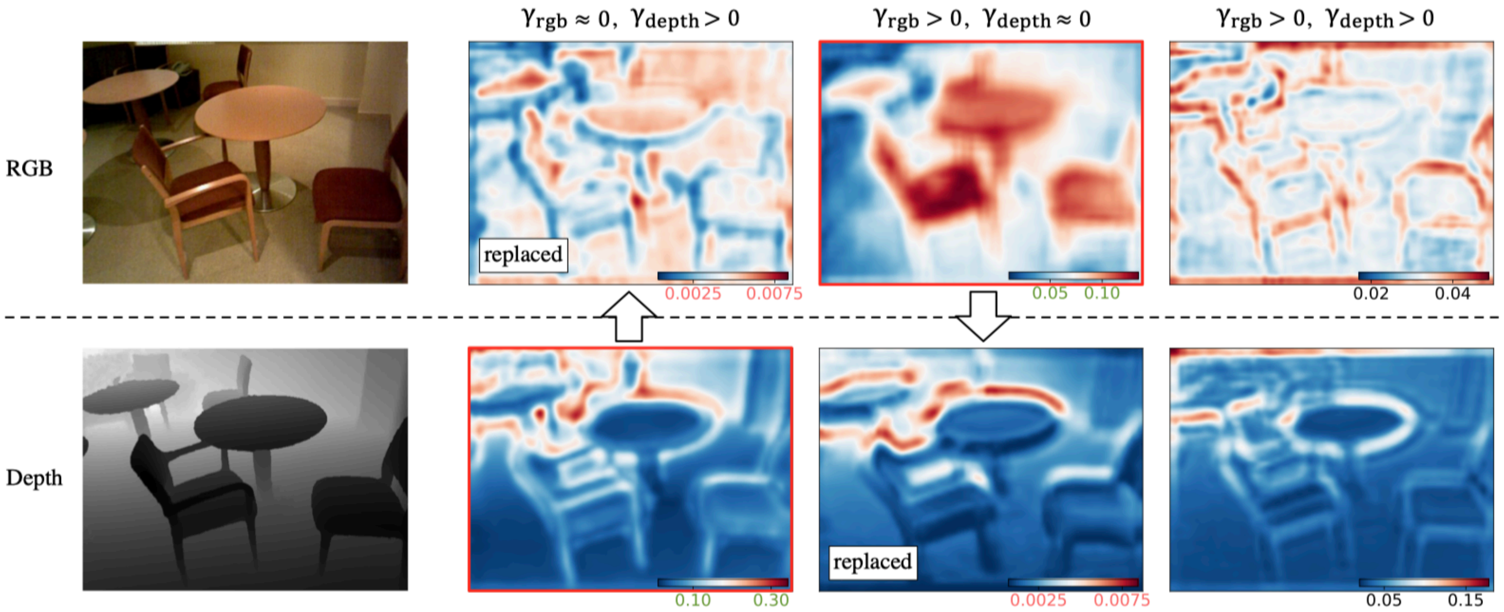




(a) Scaling factors of the first 128 channels (with sparsity constraints) when channel exchanging is applied

(b) Scaling factors of the first 128 channels (with sparsity constraints) when channel exchanging is **NOT** applied

# NeurIPS | 2020

➢ **Experiments: semantic segmentation and image-to-image translation**




Visualization of the averaged feature maps for RGB and Depth. From left to right: the input images, the channels of $(\gamma_{rgb} \approx 0, \gamma_{depth} > 0)$, $(\gamma_{rgb} > 0, \gamma_{depth} \approx 0)$, and $(\gamma_{rgb} \approx 0, \gamma_{depth} > 0)$.

# NeurIPS | 2020

➤ **Experiments: semantic segmentation and image-to-image translation**

| Convs | BNs | $\ell_1$ Regulation | Exchange | Mean IoU (%) RGB | Depth | Ensemble |
|---|---|---|---|---|---|---|
| Unshared | Unshared | × | × | 45.5 | 35.8 | 47.6 |
| Shared | Shared | × | × | 43.7 | 35.5 | 45.2 |
| Shared | Unshared | × | × | 46.2 | 38.4 | 48.0 |
| Shared | Unshared | Half-channel | × | 46.0 | 38.1 | 47.7 |
| Shared | Unshared | Half-channel | ✓ | **49.7** | **45.1** | **51.1** |
| Shared | Unshared | All-channel | ✓ | 48.6 | 39.0 | 49.8 |

Detailed results for different versions of our CEN on NYUDv2. All results are obtained with the backbone RefineNet (ResNet101) of single-scale evaluation for test.

# NeurIPS | 2020

➢ **Experiments: semantic segmentation and image-to-image translation**

| Modality | Approach | Commonly-used setting Mean IoU (%) | Commonly-used setting Params in total (M) | Same with our setting Mean IoU (%) RGB / Depth / Ensemble | Same with our setting Params in total (M) | Params used for fusion (M) |
|---|---|---|---|---|---|---|
| RGB | Uni-modal | 45.5 | 118.1 | 45.5 / - / - | 118.1 | - |
| Depth | Uni-modal | 35.8 | 118.1 | - / 35.8 / - | 118.1 | - |
| RGB-D | Concat (early) | 47.2 | 120.1 | 47.0 / 37.5 / 47.6 | 118.8 | 0.6 |
| | Concat (middle) | 46.7 | 147.7 | 46.6 / 37.0 / 47.4 | 120.3 | 2.1 |
| | Concat (late) | 46.3 | 169.0 | 46.3 / 37.2 / 46.9 | 126.6 | 8.4 |
| | Concat (all-stage) | 47.5 | 171.7 | 47.8 / 36.9 / 48.3 | 129.4 | 11.2 |
| | Align (early) | 46.4 | 238.8 | 46.3 / 35.8 / 46.7 | 120.8 | 2.6 |
| | Align (middle) | 47.9 | 246.7 | 47.7 / 36.0 / 48.1 | 128.7 | 10.5 |
| | Align (late) | 47.6 | 278.1 | 47.3 / 35.4 / 47.6 | 160.1 | 41.9 |
| | Align (all-stage) | 46.8 | 291.9 | 46.6 / 35.5 / 47.0 | 173.9 | 55.7 |
| | Self-att. (early) | 47.8 | 124.9 | 47.7 / 38.3 / 48.2 | 123.6 | 5.4 |
| | Self-att. (middle) | 48.3 | 166.9 | 48.0 / 38.1 / 48.7 | 139.4 | 21.2 |
| | Self-att. (late) | 47.5 | 245.5 | 47.6 / 38.1 / 48.3 | 203.2 | 84.9 |
| | Self-att. (all-stage) | 48.7 | 272.3 | 48.5 / 37.7 / 49.1 | 231.0 | 112.8 |
| | Ours | - | - | **49.7 / 45.1 / 51.1** | **118.2** | **0.0** |

Comparison with three typical fusion methods including concatenation (concat), fusion by alignment (align), and self-attention (self-att.) on NYUDv2.

# NeurIPS | 2020

➢ **Experiments: semantic segmentation and image-to-image translation**

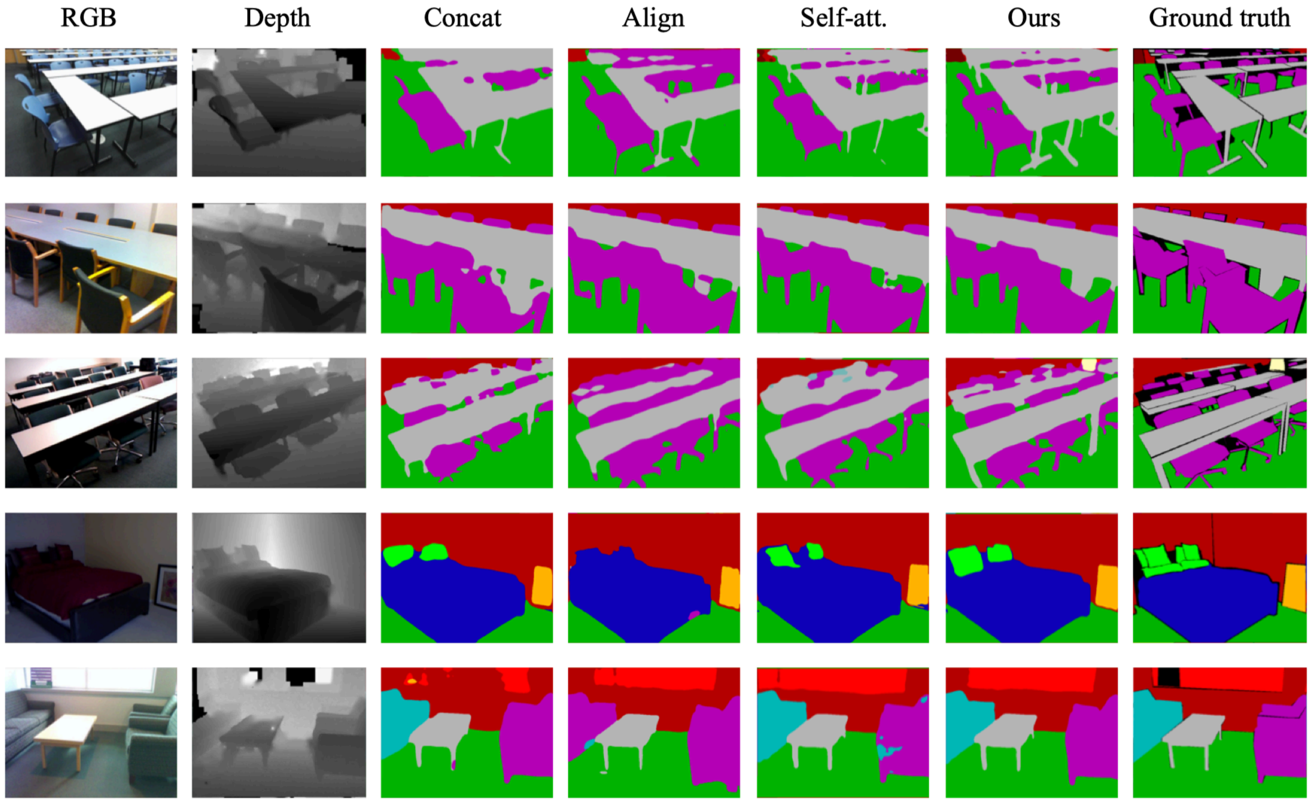
| Modality | Approach | Backbone Network | NYUDv2 Pixel Acc. (%) | NYUDv2 Mean Acc. (%) | NYUDv2 Mean IoU (%) | SUN RGB-D Pixel Acc. (%) | SUN RGB-D Mean Acc. (%) | SUN RGB-D Mean IoU (%) |
|---|---|---|---|---|---|---|---|---|
| RGB | FCN-32s [34] | VGG16 | 60.0 | 42.2 | 29.2 | 68.4 | 41.1 | 29.0 |
| | RefineNet [32] | ResNet101 | 73.8 | 58.8 | 46.4 | 80.8 | 57.3 | 46.3 |
| | RefineNet [32] | ResNet152 | 74.4 | 59.6 | 47.6 | 81.1 | 57.7 | 47.0 |
| RGB-D | FuseNet [19] | VGG16 | 68.1 | 50.4 | 37.9 | 76.3 | 48.3 | 37.3 |
| | ACNet [22] | ResNet50 | - | - | 48.3 | - | - | 48.1 |
| | SSMA [45] | ResNet50 | 75.2 | 60.5 | 48.7 | 81.0 | 58.1 | 45.7 |
| | SSMA [45] † | ResNet101 | 75.8 | 62.3 | 49.6 | 81.6 | 60.4 | 47.9 |
| | CBN [46] † | ResNet101 | 75.5 | 61.2 | 48.9 | 81.5 | 59.8 | 47.4 |
| | 3DGNN [37] | ResNet101 | - | - | - | - | 57.0 | 45.9 |
| | SCN [31] | ResNet152 | - | - | 49.6 | - | - | 50.7 |
| | CFN [30] | ResNet152 | - | - | 47.7 | - | - | 48.1 |
| | RDFNet [29] | ResNet101 | 75.6 | 62.2 | 49.1 | 80.9 | 59.6 | 47.2 |
| | RDFNet [29] | ResNet152 | 76.0 | 62.8 | 50.1 | 81.5 | 60.1 | 47.7 |
| | Ours-RefineNet (single-scale) | ResNet101 | 76.2 | 62.8 | 51.1 | 82.0 | 60.9 | 49.6 |
| | Ours-RefineNet | ResNet101 | 77.2 | 63.7 | 51.7 | 82.8 | 61.9 | 50.2 |
| | Ours-RefineNet | ResNet152 | 77.4 | 64.8 | 52.2 | 83.2 | 62.5 | 50.8 |
| | Ours-PSPNet | ResNet152 | **77.7** | **65.0** | **52.5** | **83.5** | **63.2** | **51.1** |

† indicates our implemented results.

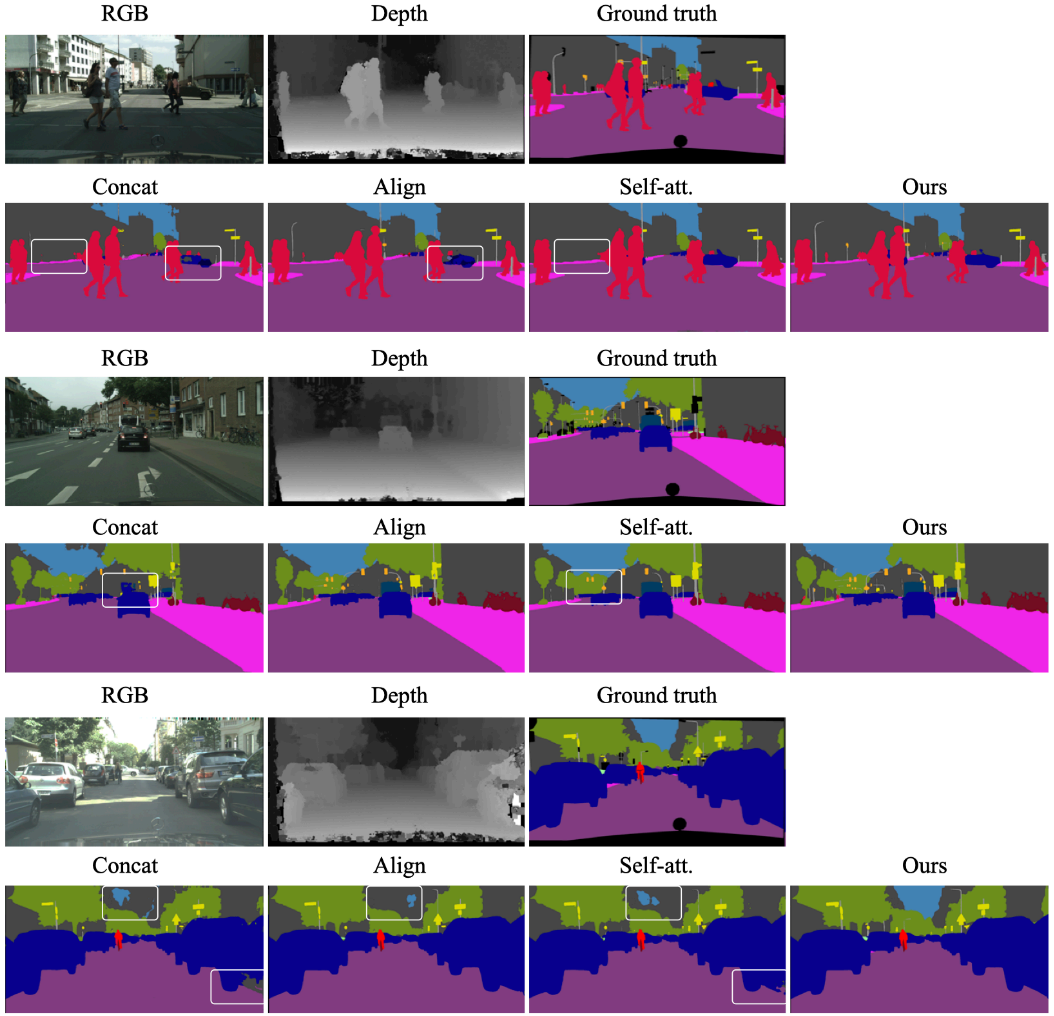
Comparison with SOTA methods on semantic segmentation.

# NeurIPS | 2020

- ➢ **Experiments:** **semantic segmentation** **and image-to-image translation**




On NYUDv2 and SUN RGB-D datasets




On Cityscapes dataset

# NeurIPS | 2020

➢ **Experiments: semantic segmentation and image-to-image translation**

| Modality | Ours | Baseline | Early | Middle | Late | All-layer |
|---|---|---|---|---|---|---|
| Shade+Texture →RGB | **62.63 / 1.65** | Concat | 87.46 / 3.64 | 95.16 / 4.67 | 122.47 / 6.56 | 78.82 / 3.13 |
| | | Average | 93.72 / 4.22 | 93.91 / 4.27 | 126.74 / 7.10 | 80.64 / 3.24 |
| | | Align | 99.68 / 4.93 | 95.52 / 4.75 | 98.33 / 4.70 | 92.30 / 4.20 |
| | | Self-att. | 83.60 / 3.38 | 90.79 / 3.92 | 105.62 / 5.42 | 73.87 / 2.46 |
| Depth+Normal →RGB | **84.33 / 2.70** | Concat | 105.17 / 5.15 | 100.29 / 3.37 | 116.51 / 5.74 | 99.08 / 4.28 |
| | | Average | 109.25 / 5.50 | 104.95 / 4.98 | 122.42 / 6.76 | 99.63 / 4.41 |
| | | Align | 111.65 / 5.53 | 108.92 / 5.26 | 105.85 / 4.98 | 105.03 / 4.91 |
| | | Self-att. | 100.70 / 4.47 | 98.63 / 4.35 | 108.02 / 5.09 | 96.73 / 3.95 |

Comparison on image-to-image translation. Evaluation metrics are FID/KID ($\times 10^{-2}$). Lower values indicate better performance.

| Modality | Depth | Normal | Texture | Shade | Depth+Normal | Depth+Normal +Texture | Depth+Normal +Texture+Shade |
|---|---|---|---|---|---|---|---|
| FID | 113.91 | 108.20 | 97.51 | 100.96 | 84.33 | 60.90 | 57.19 |
| KID ($\times 10^{-2}$) | 5.68 | 5.42 | 4.82 | 5.17 | 2.70 | 1.56 | 1.33 |

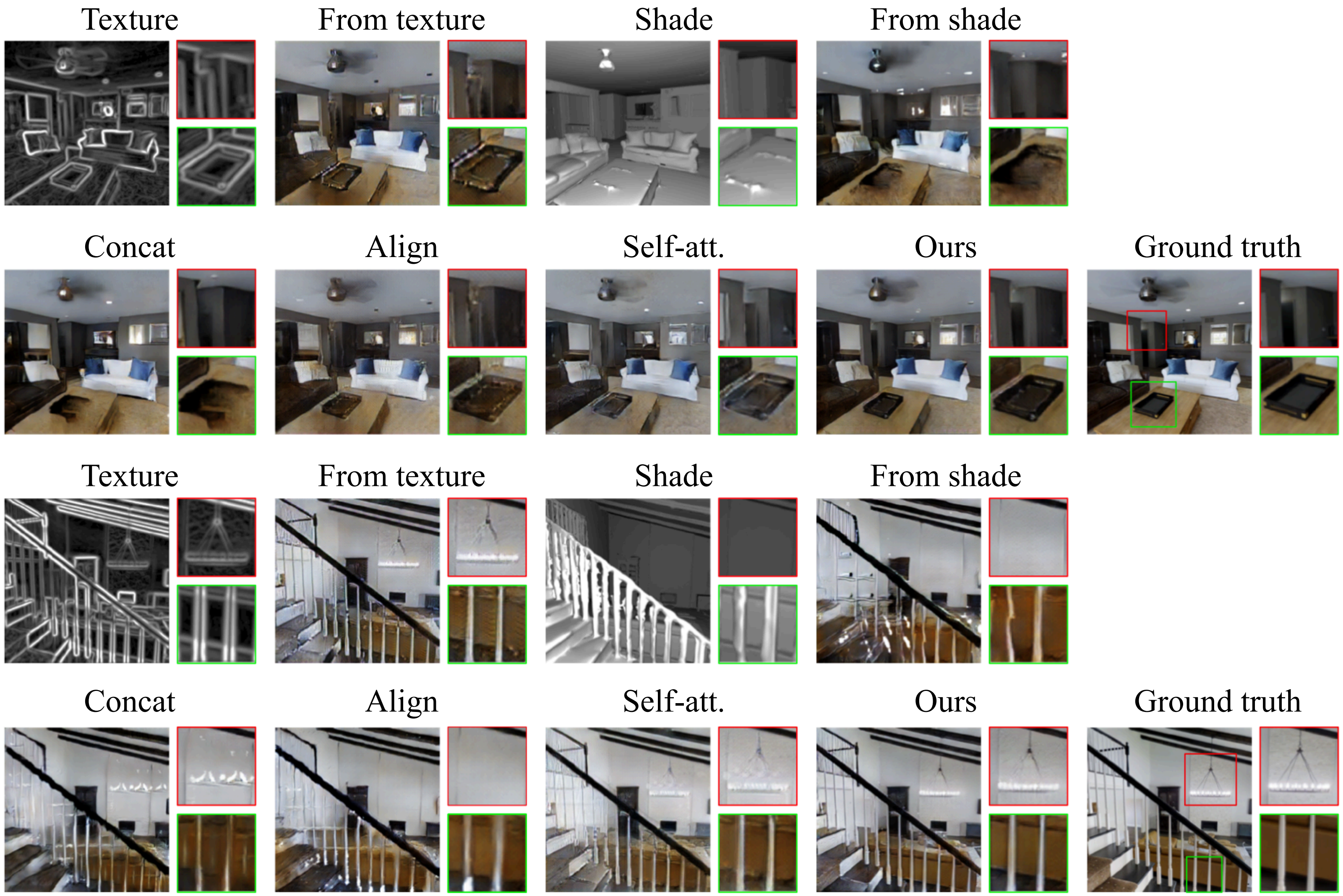
Multimodal fusion on image translation (to RGB) with modalities from 1 to 4.

# NeurIPS | 2020

➢ **Experiments: semantic segmentation and image-to-image translation**

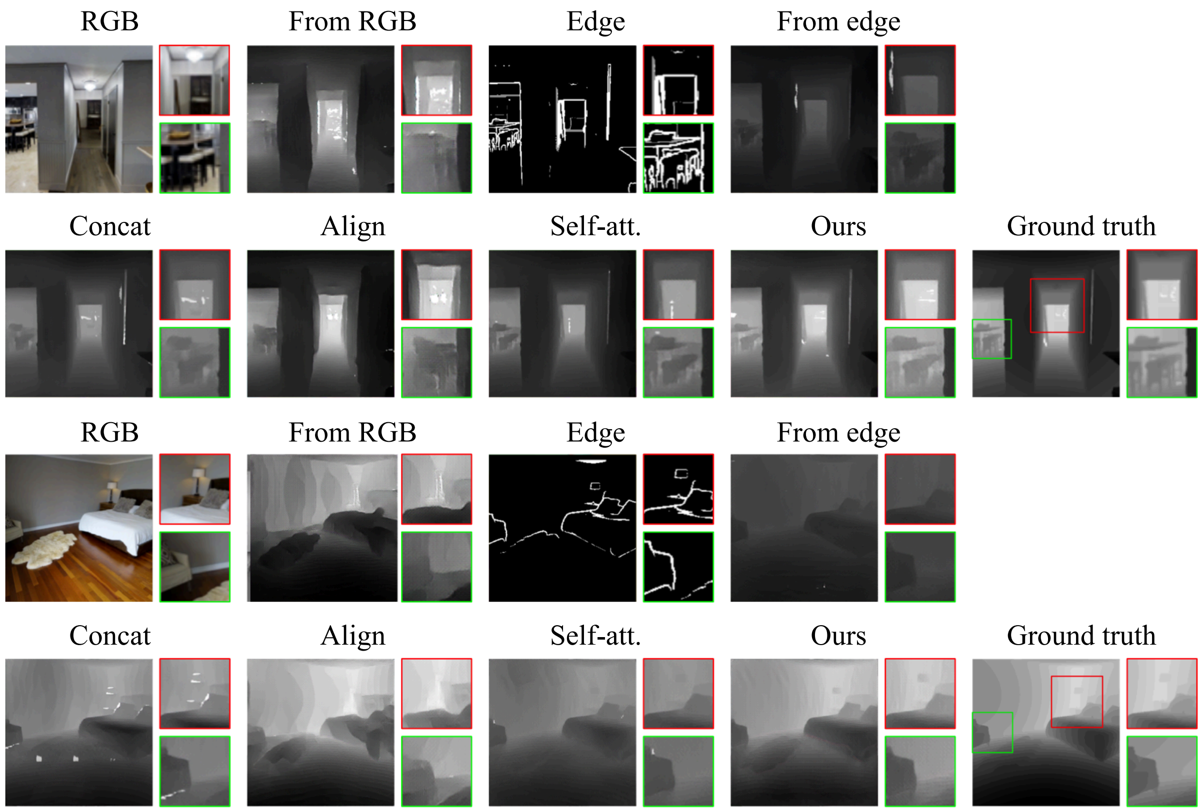
On Taskonomy dataset

Texture + Shade → RGB

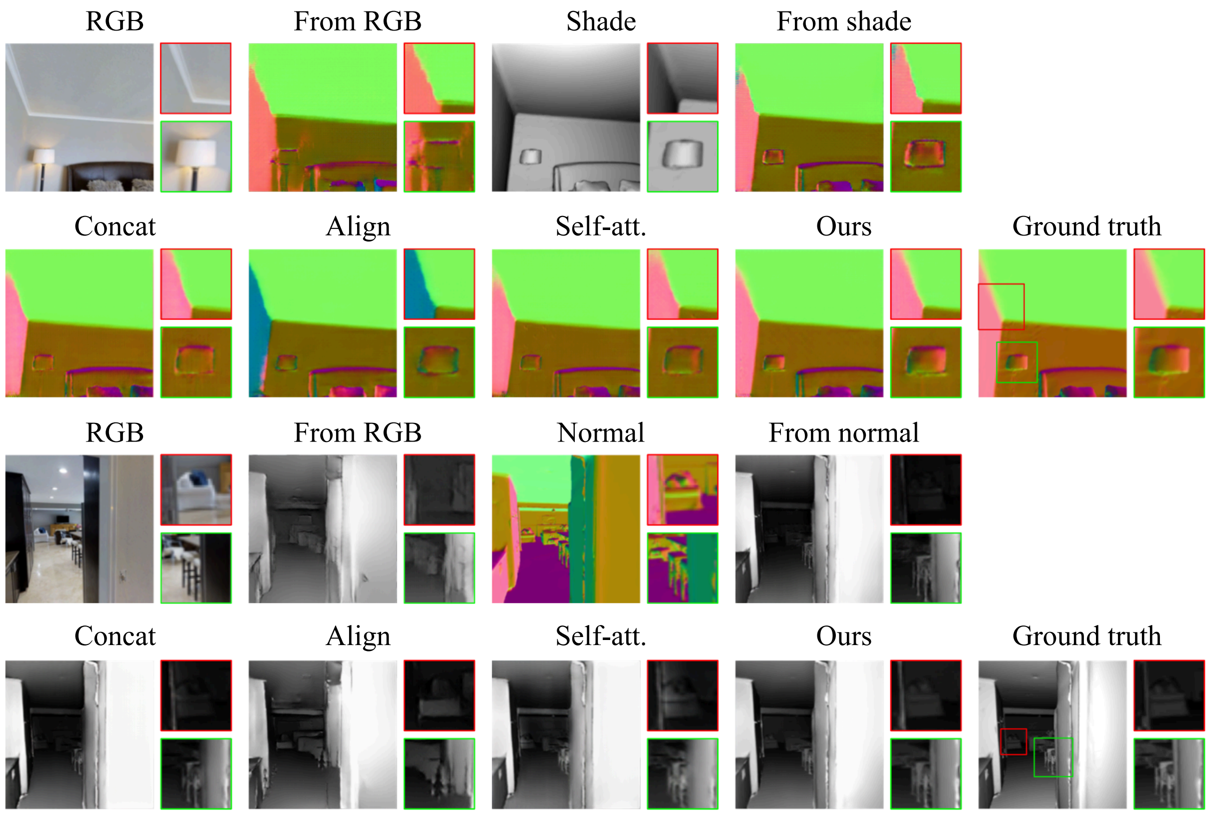


# NeurIPS | 2020

- **Experiments: semantic segmentation and image-to-image translation**



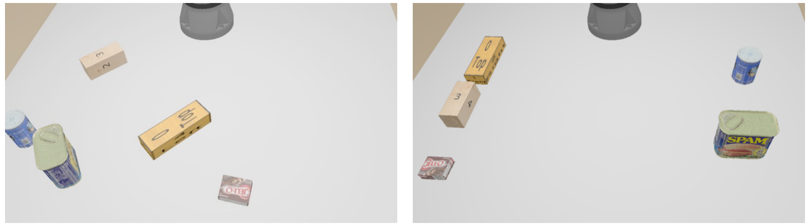

RGB + Edge → Depth

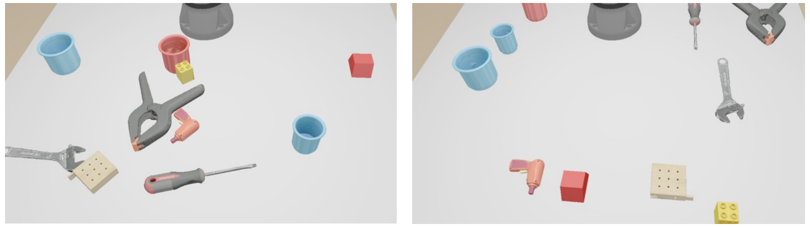



RGB + Shade → Normal RGB + Normal → Shade

# NeurIPS | 2020

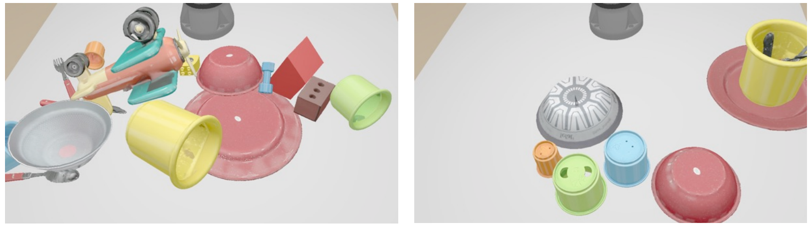
- After the paper submission, we verify the effectiveness of our multimodal channel exchanging in the IROS2020 Robotic Grasping Competition, OCRTOC.





Input Output

Input Output



Input Output

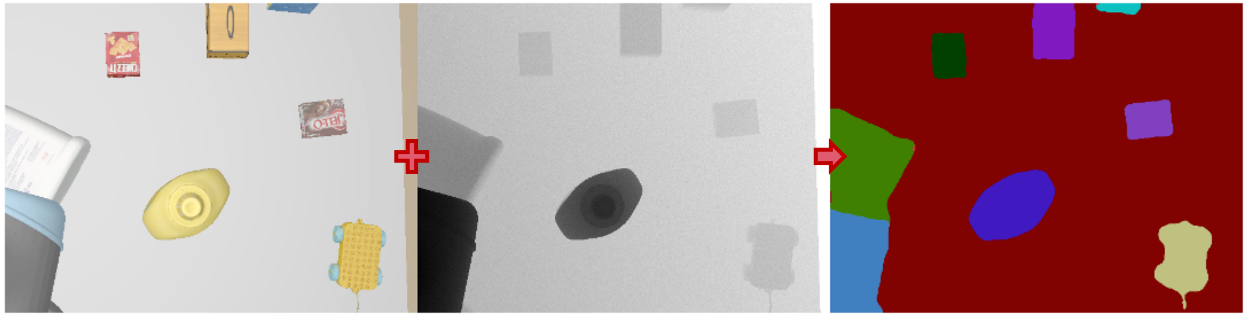
**Table Organization**

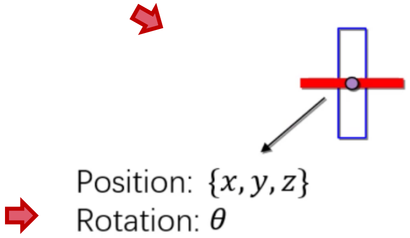
# NeurIPS | 2020

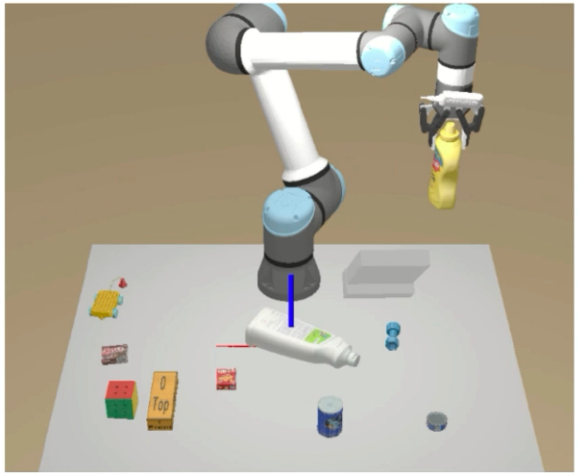
➢ We achieve the 1st place among the 17 teams for the simulation track, and 3rd place for the real robot track.

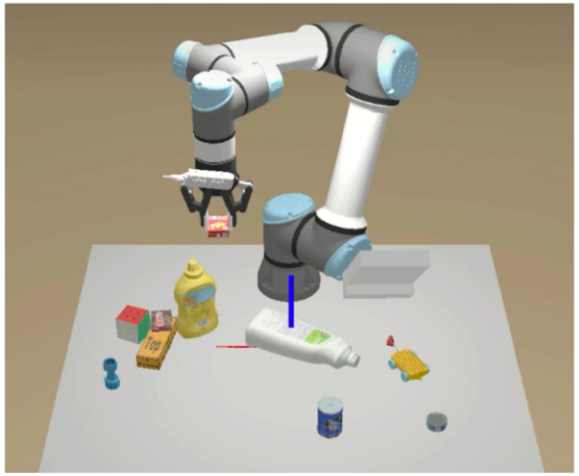
Front camera

Top camera







The channel exchanging method is used to predict the semantic masks of objects.

# Deep Multimodal Fusion by Channel Exchanging

Yikai Wang, Wenbing Huang, Fuchun Sun, Tingyang Xu, Yu Rong, Junzhou Huang

Code and models at: https://github.com/yikaiw/CEN

**Thank you for your listening!**