Deep Multimodal Feature Representation with Asymmetric Multi-layer Fusion

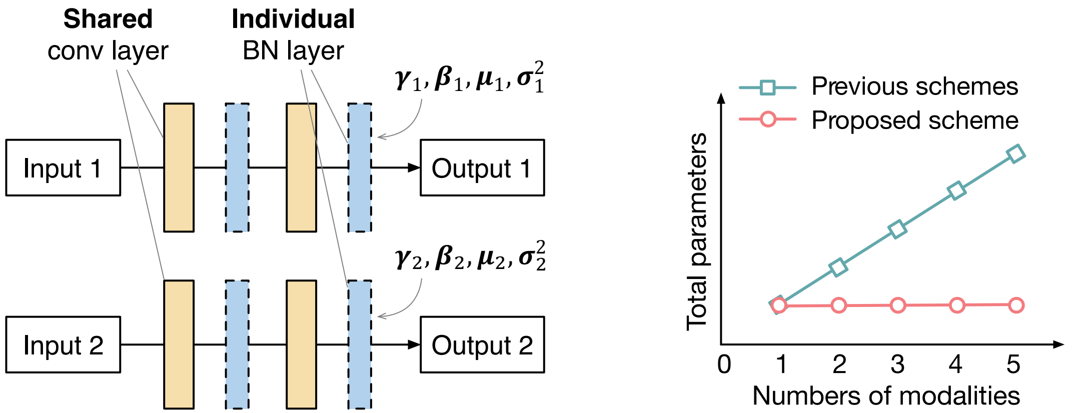
Yikai Wang[1], Fuchun Sun[1], Ming Lu[2], Anbang Yao[2]

[1] Tsinghua University [2] Intel Labs China

Deep Multimodal Feature Representation with Asymmetric Multi-layer Fusion

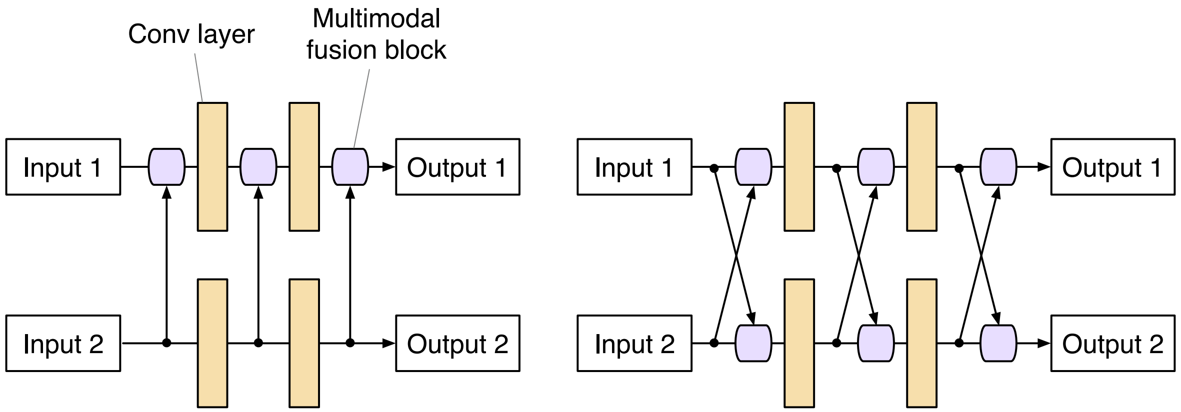
- **Summary:** This work tactfully bridges three interdependent yet parameter-free components, i.e., **Parameter Sharing Scheme**, **Cross-Modality Channel Shuffle** and **Modality-Specific Pixel Shift**, into a bidirectional compact scheme for fusing multimodal features, in the perspective of promoting feature representation learning.

- **Two architectural designs:**



(a) Our parameter-sharing scheme for multimodal training

(b) Comparison of total parameters w.r.t. numbers of modalities



(a) Unidirectional fusion

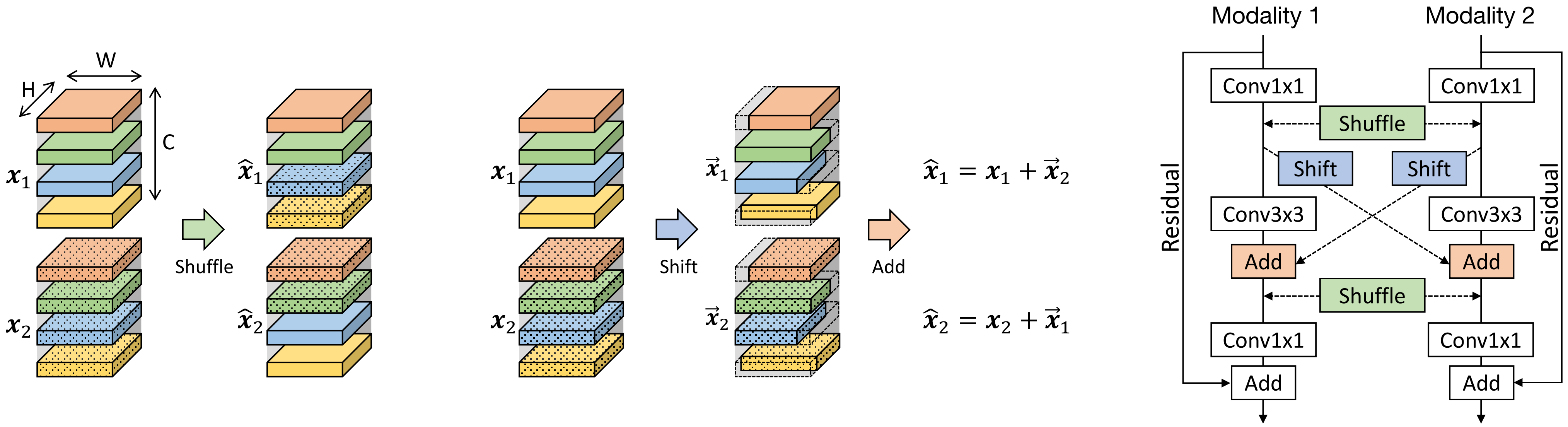
(b) Bidirectional fusion (with asymmetric fusion blocks)

Parameter-sharing Scheme: A compact multimodal fusion scheme, with shared Convs and individual BNs.

Bidirectional Fusion Scheme: A multi-layer fusion scheme, enabling each branch to exploit multimodal features.

Deep Multimodal Feature Representation with Asymmetric Multi-layer Fusion

- **Bidirectional Fusion Scheme, with two designed asymmetric fusion operations.**
  - Channel Shuffle: To strengthen the interaction of multimodal information flow across channels.
  - Pixel Shift: To improve spatial information communication of multimodal features.



(a) Channel shuffle operation

(b) Pixel shift operation

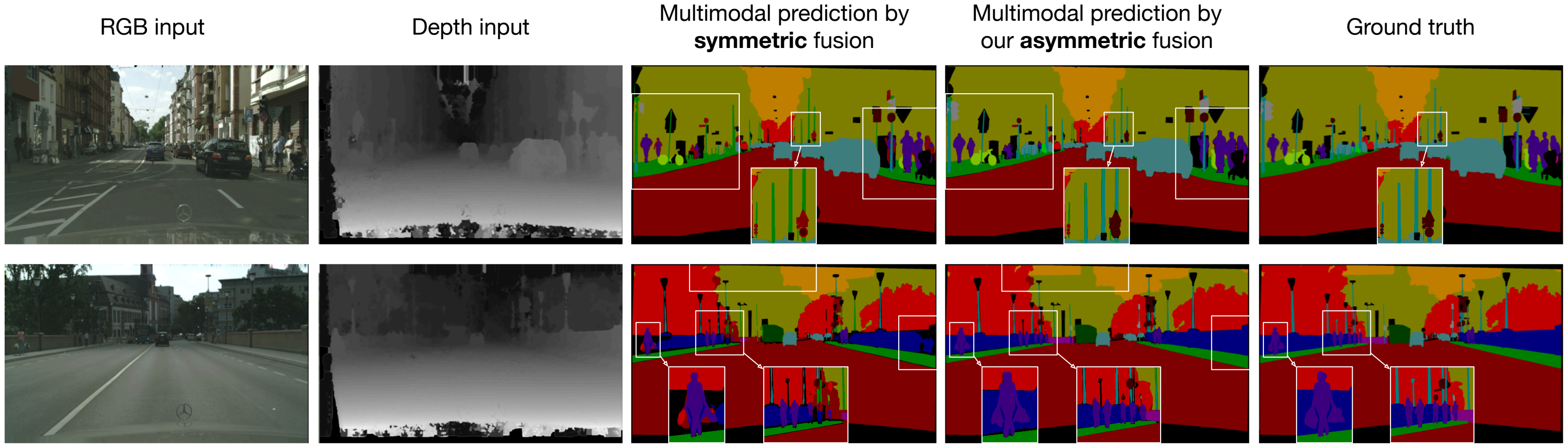
(c) Integration into residual blocks

$$\mathcal{F}(\boldsymbol{x}_1, \boldsymbol{x}_2) = \boldsymbol{x}_1[1, \cdots, T] \,||\, \boldsymbol{x}_2[T+1, \cdots, C],$$
$$\mathcal{F}(\boldsymbol{x}_2, \boldsymbol{x}_1) = \boldsymbol{x}_2[1, \cdots, T] \,||\, \boldsymbol{x}_1[T+1, \cdots, C],$$

$$\vec{\boldsymbol{x}}_1[c, h, w] = \mathcal{O}(\boldsymbol{x}_1)[c, h + \alpha_c + 1, w + \beta_c + 1], \qquad \mathcal{F}(\boldsymbol{x}_1, \boldsymbol{x}_2) = \boldsymbol{x}_1 + \vec{\boldsymbol{x}}_2,$$
$$\vec{\boldsymbol{x}}_2[c, h, w] = \mathcal{O}(\boldsymbol{x}_2)[c, h + \alpha_c + 1, w + \beta_c + 1], \qquad \mathcal{F}(\boldsymbol{x}_2, \boldsymbol{x}_1) = \boldsymbol{x}_2 + \vec{\boldsymbol{x}}_1,$$

Deep Multimodal Feature Representation with Asymmetric Multi-layer Fusion

- **Experiments:** We consider two tasks, including semantic segmentation and image translation.



Method	Data modality	Backbone	Pixel acc.	Mean acc.	IoU	#Params.
RefineNet [21]	RGB	ResNet101	73.8	58.8	46.4	118.10M
RefineNet [21]	RGB	ResNet152	74.4	59.6	47.6	133.74M
CFN [19]	RGB-D	ResNet152	-	-	47.7	-
SCN [20]	RGB-D	ResNet152	-	-	49.6	-
RDFNet [17]	RGB-D	ResNet101	75.6	62.2	49.1	366.71M
RDFNet [17]	RGB-D	ResNet152	76.0	62.8	50.1	398.00M
RefineNet †	RGB	ResNet101	73.8	59.0	46.5	118.10M
RefineNet †	Depth	ResNet101	64.0	45.6	34.3	118.10M
AsymFusion	RGB-D	ResNet101	76.6	63.5	50.8	**118.20M**
AsymFusion	RGB-D	ResNet152	**77.0**	**64.0**	**51.2**	**133.89M**

NYUDv2

Method	Data modality	Extra data	Backbone	IoU	#Params.
PSPNet [37]	RGB	×	ResNet101	80.9	56.27M
DeepLabv3 [3]	RGB	×	ResNet101	79.3	58.16M
Mapilary [1]	RGB	×	WideResNet38	78.3	135.86M
DeepLabv3+ [4]	RGB	×	Xecption65	78.8	43.48M
DPC [2]	RGB	×	Xecption65	80.9	41.82M
DRN [38]	RGB	×	WideResNet38	79.7	129.16M
AdapNet++ [28]	RGB	√	ResNet50	81.2	30.20M
SSMA [28]	RGB-D	√	ResNet50	82.2	56.44M
DeepLabv3+ †	RGB	×	Xecption65	79.4	43.48M
DeepLabv3+ †	Depth	×	Xecption65	62.3	43.48M
AsymFusion	RGB-D	×	Xecption65	**82.1**	**43.52M**

Cityscapes

Deep Multimodal Feature Representation with Asymmetric Multi-layer Fusion

➢ **Experiments:** We consider two tasks, including semantic segmentation and image translation.

This part involves a wide range of modalities including depth, normal, shade, texture and edge, and aims to translate these data to RGB.

Data modality	Concat	Average	Attention
Shade,Depth	96.5	101.3	87.3
Normal,Texture	88.9	93.0	83.3
Depth,Texture,Normal	86.4	90.2	81.5
Shade,Normal,Edge	92.8	94.4	85.6

Data modality	MMF	**AsymFusion**
Shade,Depth	92.0	**82.5**
Normal,Texture	85.9	**77.8**
Depth,Texture,Normal	82.1	**75.1**
Shade,Normal,Edge	88.6	**79.4**