Combining Models from Multiple Sources for RGB-D Scene Recognition

Xinhang Song*, Shuqiang Jiang, Luis Herranz

Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (CAS),
Institute of Computing Technology, CAS, Beijing, 100190, China

{xinhang.song, shuqiang.jiang}@vipl.ict.ac.cn, lherranz@cvc.uab.es

中国科学院计算技术研究所
Institute of Computing Technology,Chinese Academy of Sciences

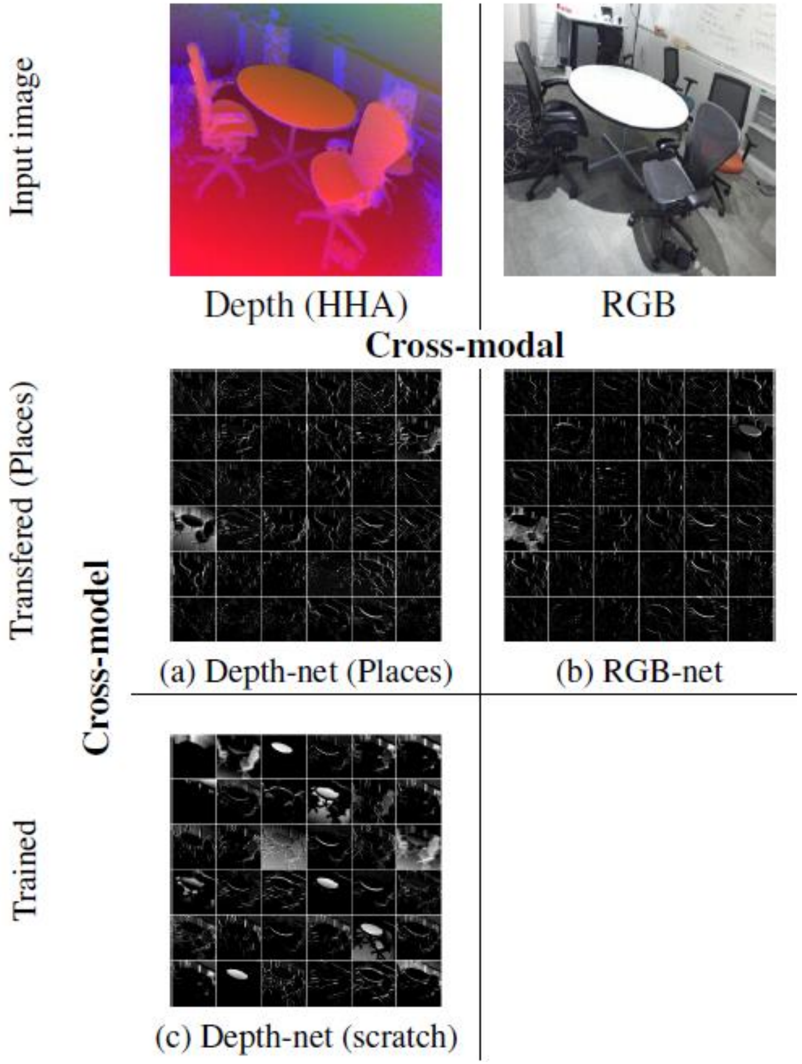
Outline

- Introduction
- Related works
- Multi-source depth network
- Multi-source Multi-modal RGB-D network
- Experiments
- Conclusion

Introduction

➢ RGB-D data

- Captured by the devices with both RGB and depth cameras, such as Microsoft Kinect.
- Depth images represent the distance between camera and objects.
- Depth images are not affected by the light.
- Combining RGB and depth data for image recognition works better than only using RGB or depth data.



Introduction

➢ Limitations of current RGB-D scene recognition

- RGB-D data is **expensive** to collect (leading to the **lack** of training data).
- RGB and depth images are two different modalities.
- Only use low-level filters learned from RGB data cannot exploit properly **depth-specific** patterns.
- RGB and depth features are only combined at high-levels but rarely at **lower-levels**.

(a) RGB-D data collection robot

(b) Data collection robot for Image CLEF competition (2013)

(c) Data collection of NYU RGB-D database

Figure 1: Machines for RGB-D data collection.

Introduction

Motivation

- Learning better multi-modal features that capture modality-specific patterns, and also leveraging the vast knowledge already available in other source models such as Places-CNN .

Contributions

- A multi-source depth CNN combining complementary models to capture modality-specific features.
- A layer selection algorithm to find discriminative combinations of layers.

Related works

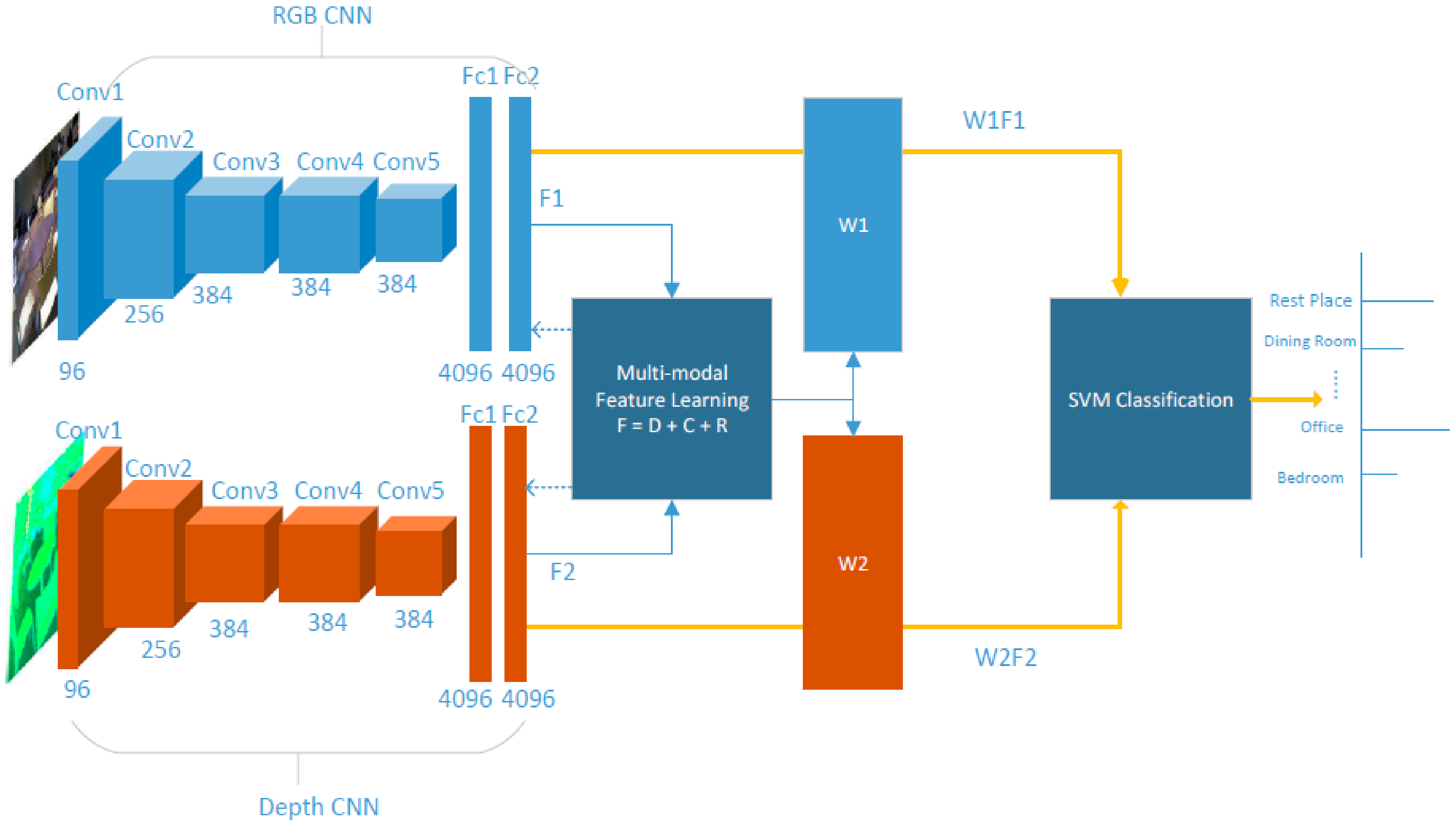
- Fine tune the pretrained CNN model to depth data
  - SUN RGB-D database [Song et al 2015]
  - Guide depth CNN training with RGB CNN [Zhu et al 2016]
  - Include R-CNN to extract local features [Wang et al 2016]
- Train CNN model from scratch with depth data
  - Weakly supervised training with patches [Song et al 2017]

Related works

- Propose SUN RGB-D database
  - 19 indoor scene categories
  - 10335 RGB-D images
- Train CNN models with large scale RGB database, i.e., Places.
- Fine tune the pretrained CNN models with RGB and Depth data of SUN RGB-D.
- Concatenate the fc7 activation of RGB and Depth CNN model.

[Song et al., 2015] Shuran Song, Samuel Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on, pages 567–576, June 2015.

Related works

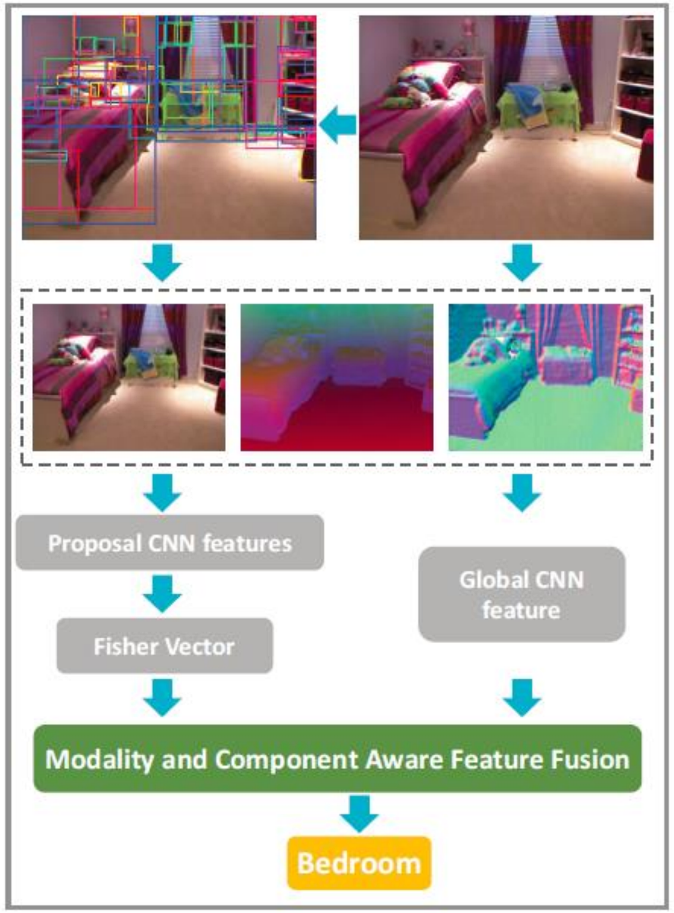


1.RGB CNN guides Depth CNN using the distances between fc activation as the supervision;
2.The concatenation of RGB and Depth CNN fc activation are fed to SVM.

[Zhu et al 2016] H. Zhu, J.-B.Weibel, and S. Lu. Discriminative multi-modal feature fusion for rgbd indoor scene recognition. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2016.

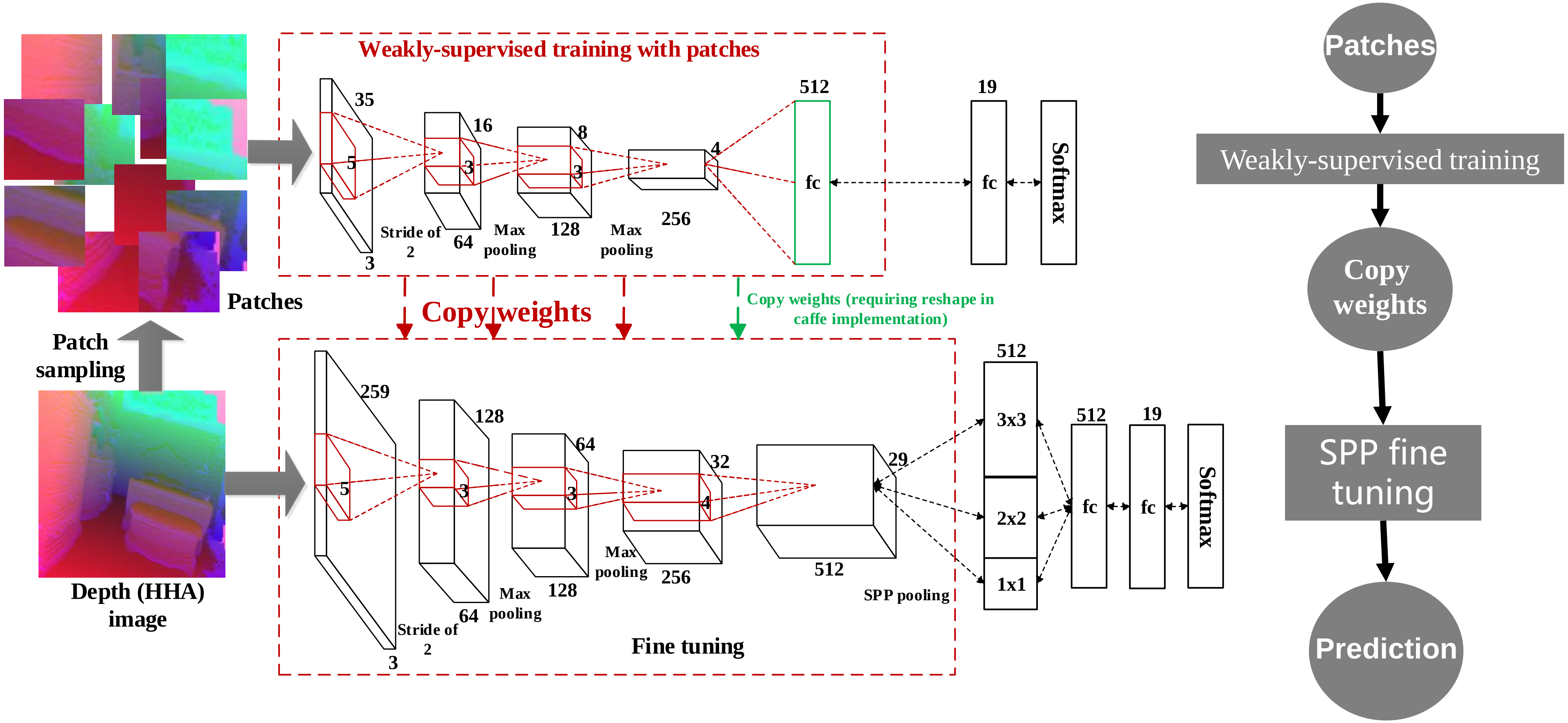
Related works

- Fine tune Place-CNN to RGB and Depth CNN (global feature);
- Fine tune Object detection framework R-CNN [2] to RGB-D data (local feature);
- Combine global and local feature for prediction[1].



[1] A. Wang, J. Cai, J. Lu, and T.-J. Cham. Modality and component aware feature fusion for rgb-d scene classification. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2016.
[2] S. Gupta, R. Girshick, P. Arbelaez, and J. Malik. Learning rich features from rgb-d images for object detection and segmentation. In ECCV, 2014.

Related works



[Song et al] Xinhang Song, Luis Herranz , Shuqiang Jiang. “Depth CNNs for RGB-D scene recognition: learning from scratch better than transferring from RGB-CNNs” Thirty-First AAAI Conference on Artificial Intelligence （AAAI） 2017

Multi-source depth network

- **Base Models**
  - RGB-net
    - Transferring from Places-CNN to RGB data
  - **Depth-net (Places)**
    - Transferring from Places-CNN to depth data
  - **Depth-net (scratch)**
    - Training directly from scratch with depth data

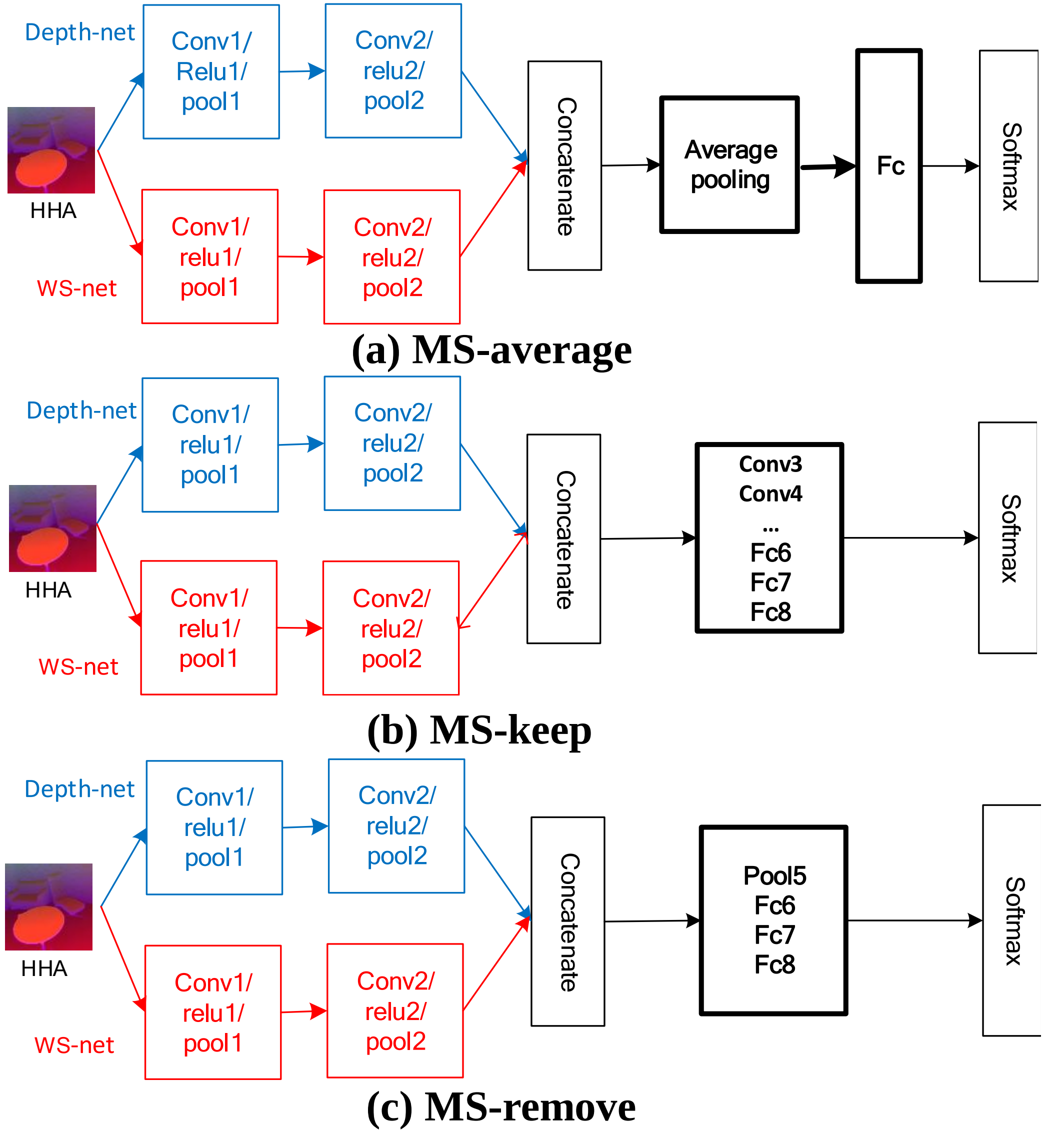
Multi-source depth network

Multi-source depth network: Depth-net (scratch) and Depth-net (Places)

(a)MS-average: *average* pooling after concatenation

(b)MS-keep: *keep* the subsequential convolutional layers after concatenation

(c)MS-remove: *remove* convolutional layers (but keep pooling layers after concatenation)



Multi-source Multi-modal RGB-D network

Layer selection algorithm

Algorithm: beam search
Initialization: L = {}
Search iteration:

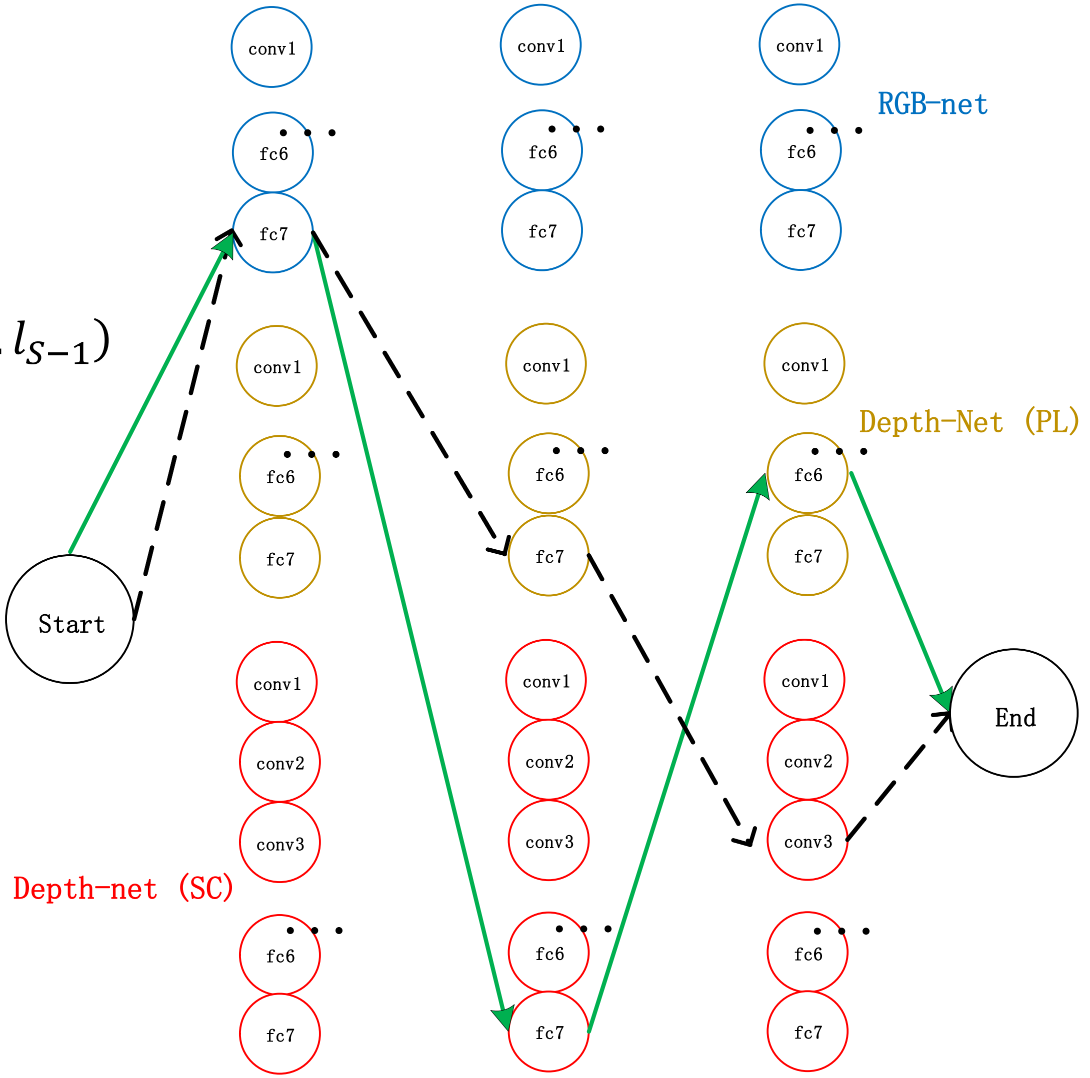
$$L = arg \min_{\{l_1, l_2, \dots l_S\}} (1-\lambda)\,\mathrm{P}(l_1, l_2, \dots l_S) + \lambda C(l_S | l_1, l_2, \dots l_{S-1})$$

P: probability of error (POE)
C: average correlation coefficient (ACC)

Layer concatenation

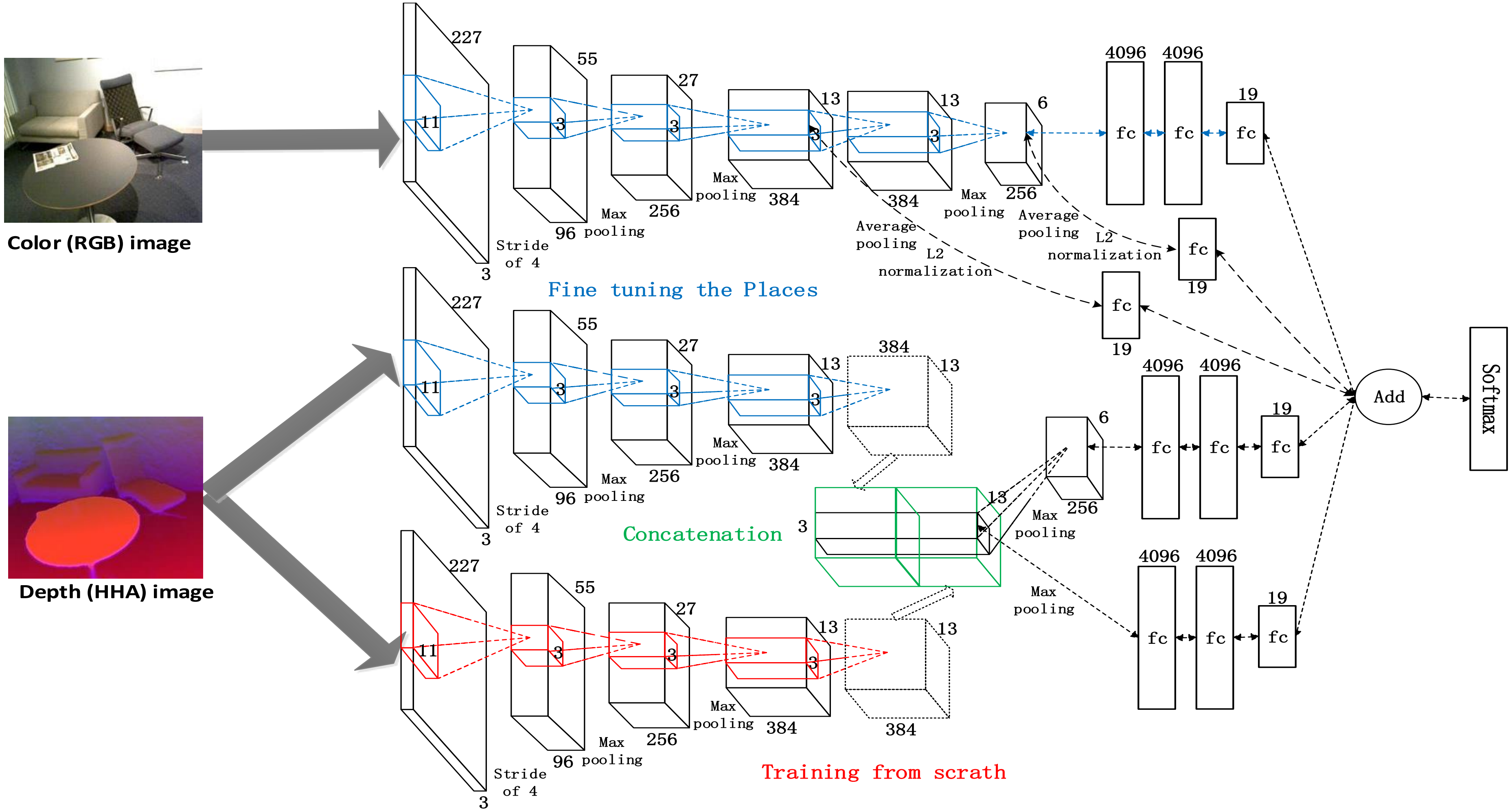
$$F_{MS} = [W_{SC} \quad W_{depth}] \begin{bmatrix} F_{SC} \\ F_{depth} \end{bmatrix} = \overline{W} \begin{bmatrix} F_{SC} \\ F_{depth} \end{bmatrix}$$

The sum of the output is equivalent to the concatenation of input (DAG-CNN) [Yang et al 2015]



Institute of Computing Technology, Chinese Academy of Sciences
中科院计算所

Multi-source Multi-modal RGB-D network



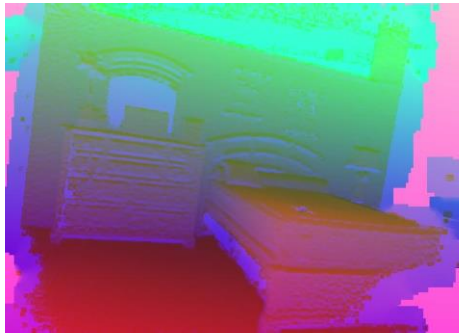
The proposed multi-modal multi-source framework combining RGB-net, MS-depth-net after layer selection. The add operation is the element-wise sum

Experiments

- Database: SUN RGB-D, NYU D2
  - SUN RGB-D
    - 19 indoor scene categories
    - 10335 RGB-D images
  - NYU D2
    - 10 indoor scene categories
    - 1349 RGB-D images
- HHA encoding:
  - Depth images are encoded to HHA

RGB

Depth



HHA

Experiments

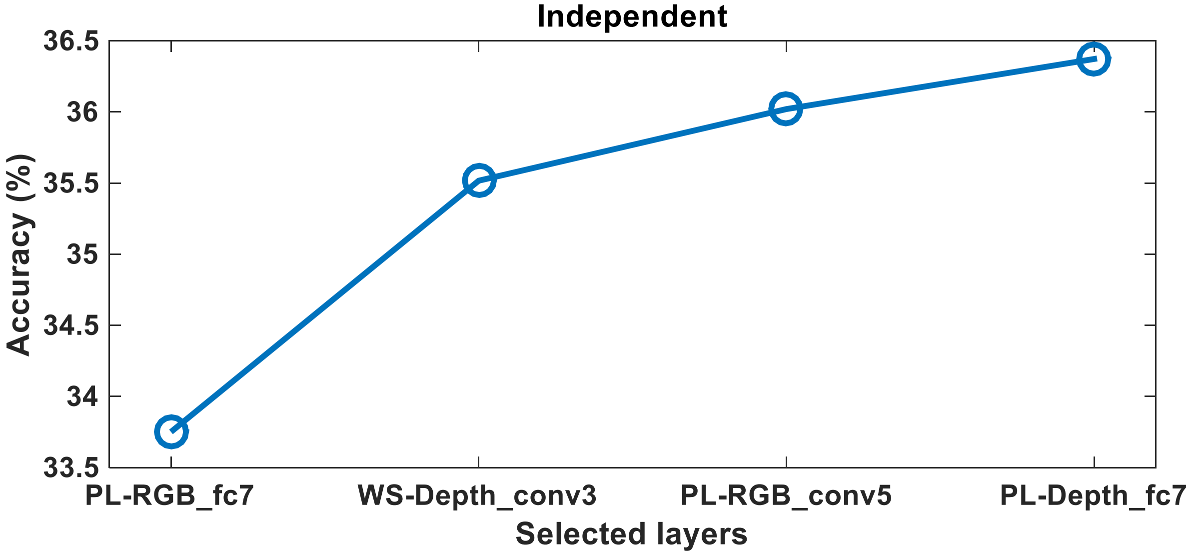
- Multi-source depth networks

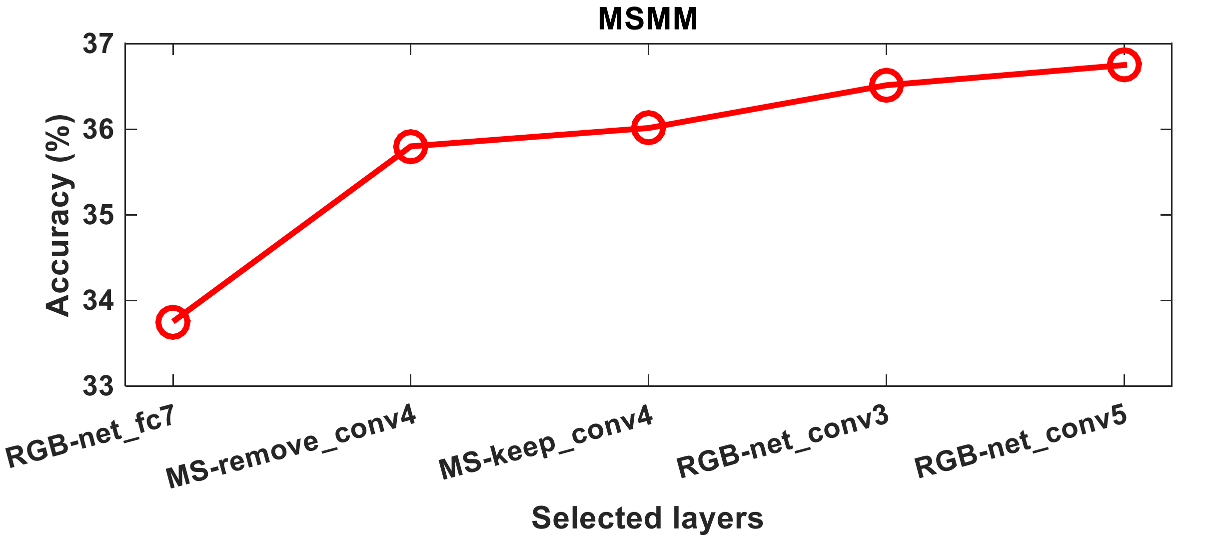
	conv1	conv2	conv3	conv4	conv5	fc6	fc7
RGB-net	20.4	25.8	27.1	**30.1**	**31.5**	**31.5**	**32.7**
Depth-net(PL)	22.5	24.2	25.3	25.3	25.7	26.2	26.3
Depth-net(SC)	21.3	25.5	26.7	26.0	26.3	26.4	26.5
MS-average	**25.0**	**27.7**	27.1	27.0	27.1	28.0	27.9
MS-keep	21.0	24.8	27.4	28.0	29.5	29.8	29.4
MS-remove	22.2	**27.7**	**28.5**	29.4	29.5	29.8	29.4

- Best performances: RGB-net@fc7, Depth-net (Places)@fc7 and Depth-net (scratch)@fc3
- MS-keep and MS-remove perform better than MS-average, except for conv1 (very shallow model)

Experiments

- Layer selection
  - Selection of layers (each iteration includes a new layer) with B=3
  - Only the best path is shown
- Independent
  - Layers are selected only from the three base models
- MSMM
  - Layers are selected from RGB-net and MS-net





Experiments

- SUN RGB-D

Method	CNN model			Accuracy (%)		
	RGB	Depth		RGB	Depth	RGB-D
Baseline (concatenate fc7)	RGB-net	Depth-net (PL)	-	35.4	30.9	39.1
	RGB-net	Depth-net (PL)	-	41.5	37.5	45.4
Two CNNs (concat fc7+FT)	RGB-net	Depth-net (PL)		41.5	37.5	48.5
	RGB-net		Depth-net (SC)	41.5	37.2	48.3
Multi-source depth		Depth-net (PL)	Depth-net (SC)	-	**40.1**	-
Three CNNs (concat. fc7+FT)	RGB-net	Depth-net (PL)	Depth-net (SC)	-	-	49.8
Layer selection (independent)	RGB-net	Depth-net (PL)	Depth-net (SC)	-	-	50.6
Layer selection (MSMM)	RGB-net	Depth-net (PL) + (SC)		-	-	**51.2**
Layer selection (MSMM+wSVM)	RGB-net	Depth-net (PL) + (SC)		-	-	**52.3**
State-of-the-art	Gupta et al.[Gupta et al., 2014a]			37.0	-	41.5
	Wang et al.[Wang et al., 2016]			40.4	36.3	48.1
	Song et al.[Song et al., 2017]			-	-	52.4

Experiments

中科院计算所
Institute of Computing Technology, Chinese Academy of Sciences

- NYU D2

Method	CNN model			Accuracy (%)
	RGB	Depth		
Baseline (concatenate fc7)	RGB-net	Depth-net (PL)	-	45.4
Three CNNs (concat. fc7+FT)	RGB-net	Depth-net (PL)	Depth-net (SC)	63.9
Layer selection (independent)	RGB-net	Depth-net (PL)	Depth-net (SC)	64.3
Layer selection (MSMM)	RGB-net	Depth-net (PL) + (SC)		65.1
Layer selection (MSMM+wSVM)	RGB-net	Depth-net (PL) + (SC)		**66.7**
State-of-the-art	Gupta et al.[Gupta et al., 2014a]			45.4
	Wang et al.[Wang et al., 2016]			63.9
	Song et al.[Song et al., 2017]			65.8

Conclusion

- A multi-source depth CNN combining complementary models to capture modality-specific features.
- A layer selection algorithm to select discriminative combinations of layers.
- A multi-source multi-modal CNN to obtain good performances on different SUN RGB-D databases.

Thank you!

Q&A