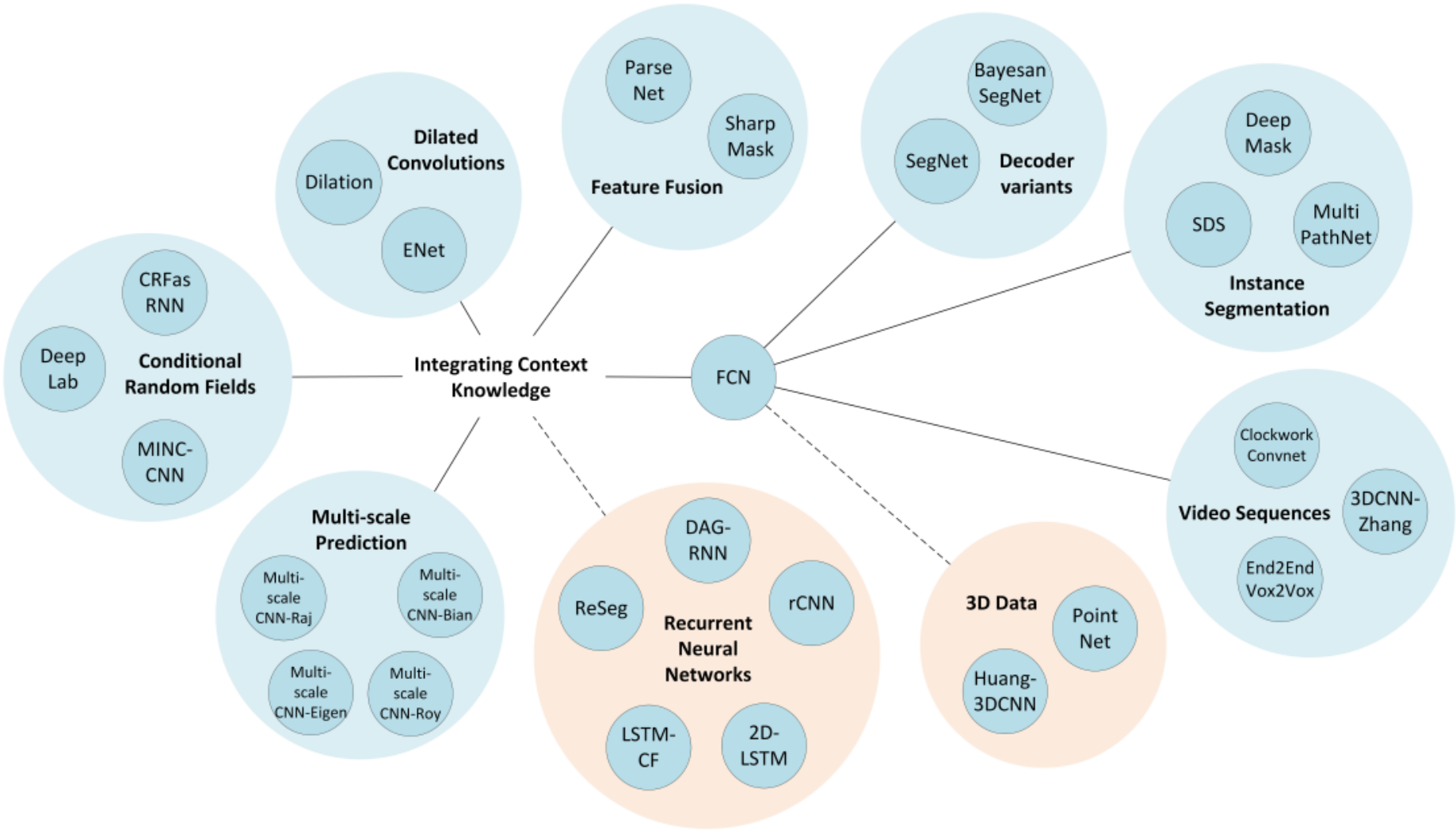
Deep Learning for RGB-D Image Segmentation

Jianping Fan
Department of Computer Science
UNC-Charlotte

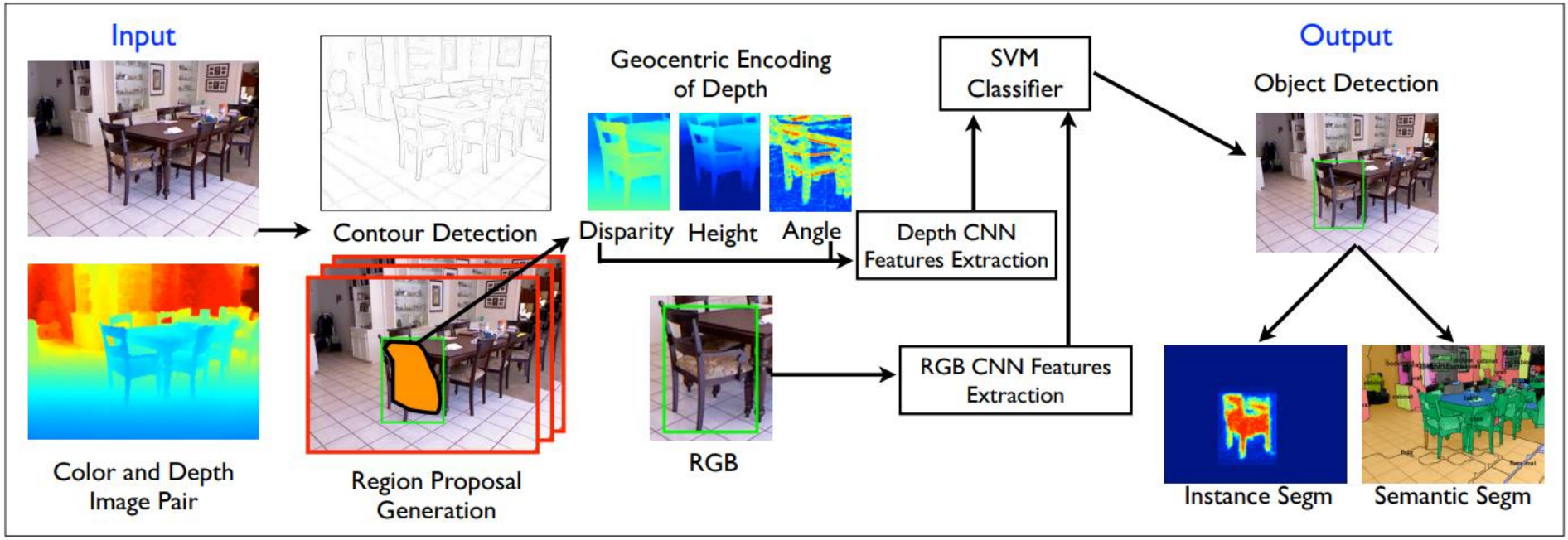
Course Website:
http://webpages.uncc.edu/jfan/itcs5152.html

Fully convolutional network (FCN) has been successfully applied in semantic segmentation of scenes represented with RGB images. Images augmented with depth channel provide more understanding of the geometric information of the scene in the image. The question is how to best exploit this additional information to improve the segmentation performance.

There are three key approaches for RGB-D image segmentation:
(a)Traditional Approaches
(b)Encoder-decoder Approaches
(c)Multi-scale Networks


FCN
Integrating Context Knowledge
Dilated Convolutions
Dilation
ENet
Feature Fusion
Parse Net
Sharp Mask
Decoder variants
Bayesan SegNet
SegNet
Instance Segmentation
Deep Mask
SDS
Multi PathNet
Conditional Random Fields
CRFas RNN
Deep Lab
MINC-CNN
Multi-scale Prediction
Multi-scale CNN-Raj
Multi-scale CNN-Bian
Multi-scale CNN-Eigen
Multi-scale CNN-Roy
Recurrent Neural Networks
DAG-RNN
ReSeg
rCNN
LSTM-CF
2D-LSTM
3D Data
Point Net
Huang-3DCNN
Video Sequences
Clockwork Convnet
3DCNN-Zhang
End2End Vox2Vox

Learning Rich Features from RGB-D Images for Object Detection and Segmentation, ECCV 2014



Overview: from an RGB and depth image pair, our system detects contours, generates 2.5D region proposals, classifies them into object categories, and then infers segmentation masks for instances of “thing”-like objects, as well as labels for pixels belonging to “stuff”-like categories.

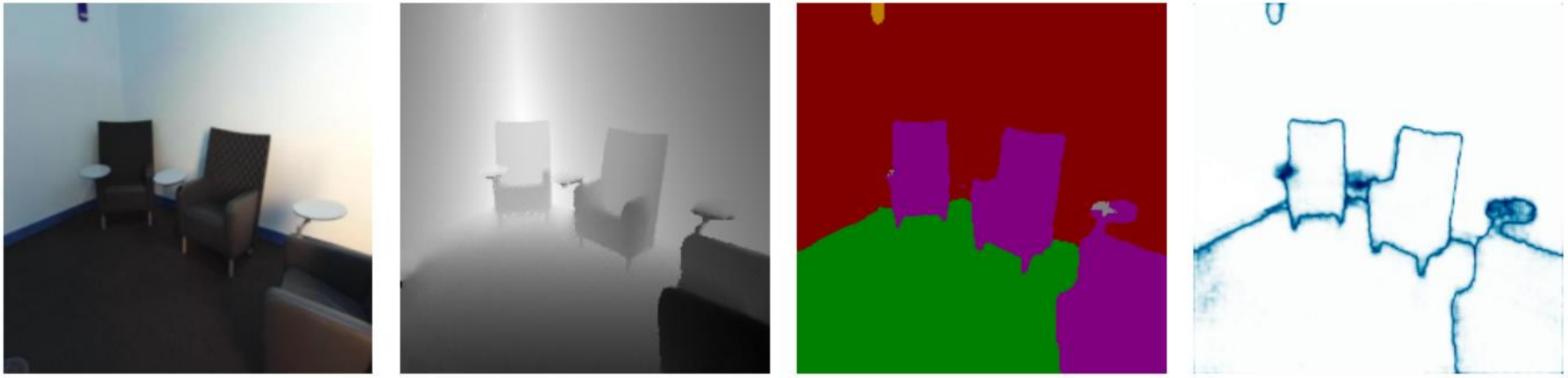
Learning Rich Features from RGB-D Images for Object Detection and Segmentation, ECCV 2014

We propose to encode the **depth image** with three channels at each pixel: **horizontal disparity**, **height above ground**, and the **angle** the pixel's local surface normal makes with the inferred gravity direction. We refer to this encoding as **HHA**. All three channels are linearly scaled to map observed values across the training dataset to the 0 to 255 range.

The HHA representation encodes properties of geocentric pose that emphasize complementary discontinuities in the image (depth, surface normal and height). Furthermore, it is unlikely that a CNN would automatically learn to compute these properties directly from a depth image, especially when very limited training data is available.
Our hypothesis is that a network designed for RGB images can also learn a suitable representation for HHA images.

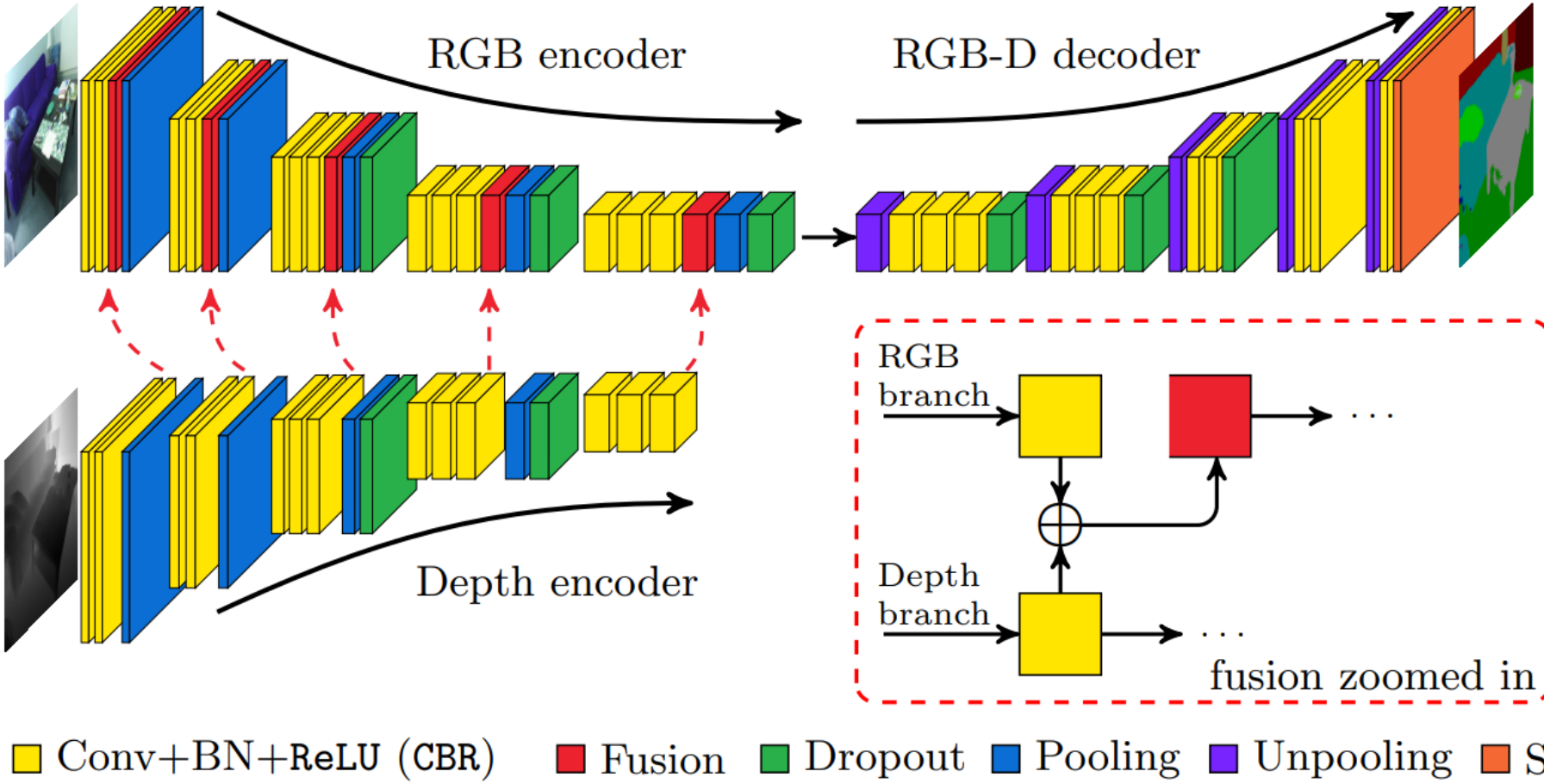
FuseNet: Incorporating Depth into Semantic Segmentation via Fusion-based CNN Architecture

With the availability of RGB-D cameras, it is expected that additional depth measurement will improve the accuracy. Here we investigate a solution how to incorporate complementary depth information into a semantic segmentation framework by making use of convolutional neural networks (CNNs). Recently encoder-decoder type fully convolutional CNN architectures have achieved a great success in the field of semantic segmentation. Motivated by this observation we propose an encoder-decoder type network, where the encoder part is composed of two branches of networks that simultaneously extract features from RGB and depth images and fuse depth features into the RGB feature maps as the network goes deeper.



An exemplar output of FuseNet. From left to right: input RGB and depth images, the predicted semantic labeling and the probability of the corresponding labels, where white and blue denote high and low probability, respectively

FuseNet: Incorporating Depth into Semantic Segmentation via Fusion-based CNN Architecture



The architecture of the proposed FuseNet. Colors indicate the layer type. The network contains two branches to extract features from RGB and depth images, and the feature maps from depth is constantly fused into the RGB branch, denoted with the red arrows. In our architecture, the fusion layer is implemented as an element-wise summation, demonstrated in the dashed box.

FuseNet: Incorporating Depth into Semantic Segmentation via Fusion-based CNN Architecture

FuseNet consists of an encoder-decoder type network architecture. Such network has two major parts:
(1) the encoder part extracts features;
(2) the decoder part up-samples the feature maps back to the original input resolution.

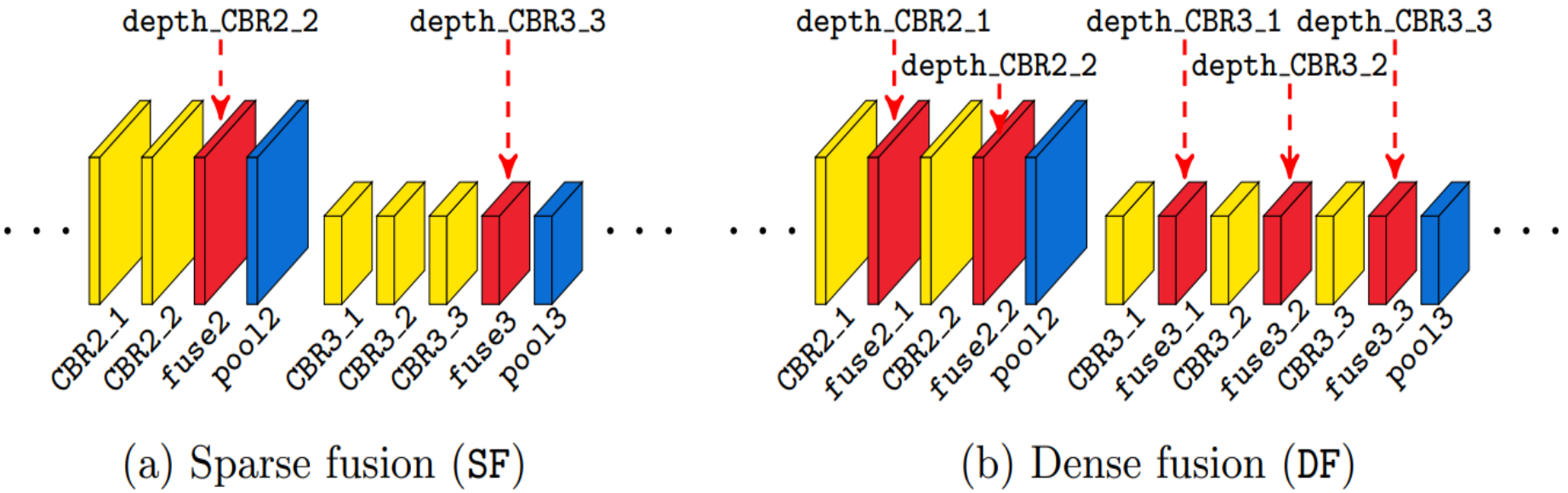
This encoder-decoder style has been already introduced in several previous works such as DeconvNet and SegNet and has achieved good segmentation performance. In addition, two branches extract features from RGB and depth images. We note that the depth image is normalized to have the same value range as color images, i.e. into the interval of [0,255]. In order to combine information from both input modules, we fuse the feature maps from the depth branch into the feature maps of the RGB branch.

FuseNet: Incorporating Depth into Semantic Segmentation via Fusion-based CNN Architecture

The encoder part of FuseNet resembles the 16-layer VGG net, except of the fully connected layers fc6, fc7 and fc8, since the fully connected layers reduce the resolution with a factor of 49, which increases the difficulty of the up-sampling part. In our network, we always use batch normalization (BN) after convolution (Conv) and before rectified linear unit1 (ReLU) to reduce the internal covariate shift. We refer to the combination of convolution, batch normalization and ReLU as CBR block, respectively. The BN layer first normalizes the feature maps to have zero-mean and unit-variance, and then scales and shifts them afterwards. In particular, the scale and shift parameters are learned during training. As a result, color features are not overwritten by depth features, but the network learns how to combine them in an optimal way.

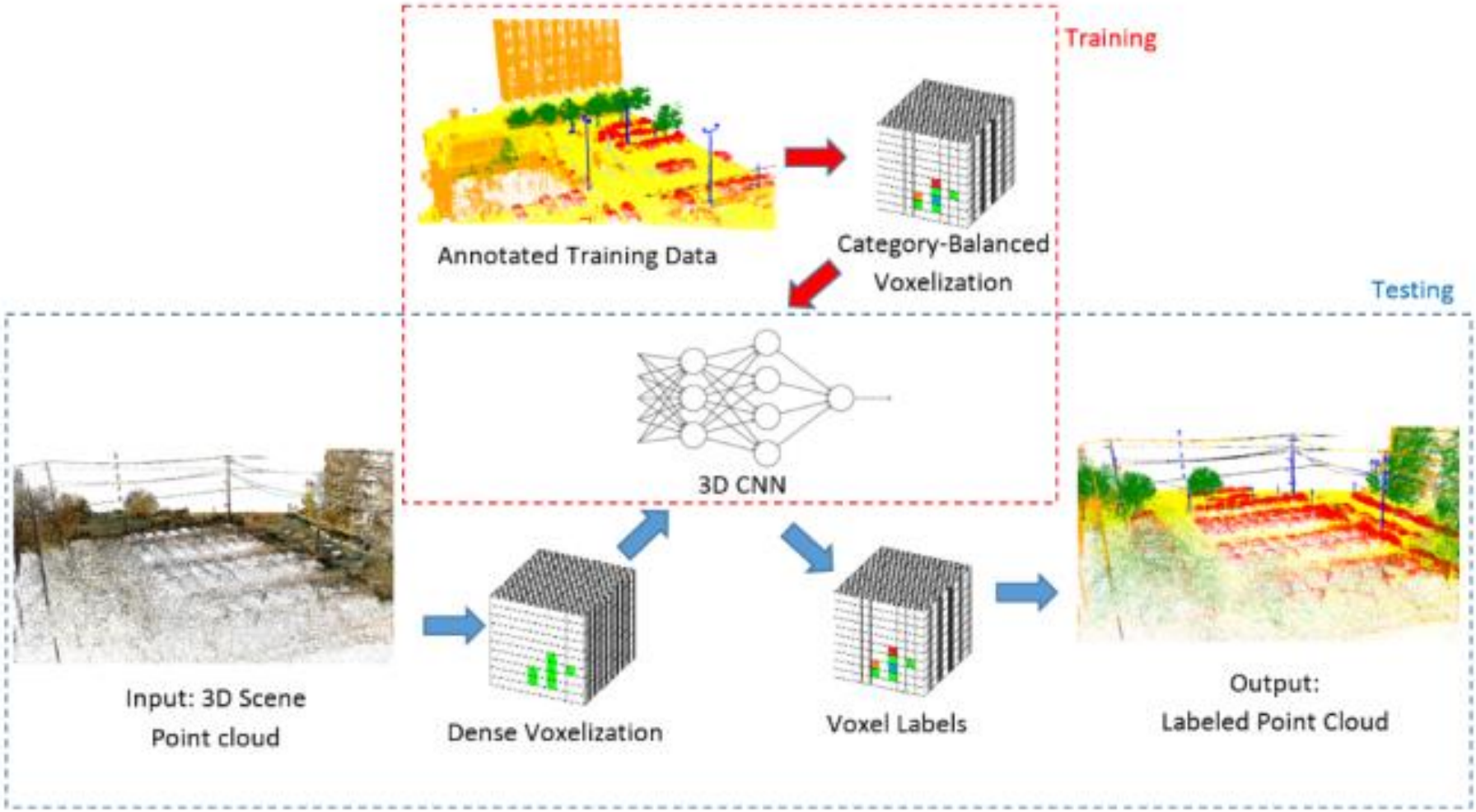
The decoder part is a counterpart of the encoder part, where memorized un-pooling is applied to up-sample the feature maps. In the decoder part, we again use the CBR blocks. We also did experiments with deconvolution instead of convolution, and observed very similar performance.

FuseNet: Incorporating Depth into Semantic Segmentation via Fusion-based CNN Architecture



(a) Sparse fusion (SF) (b) Dense fusion (DF)

Illustration of different fusion strategies at the second (CBR2) and third (CBR3) convolution blocks of VGG 16-layer net. (a) The fusion layer is only inserted before each pooling layer. (b) The fusion layer is inserted after each CBR block.



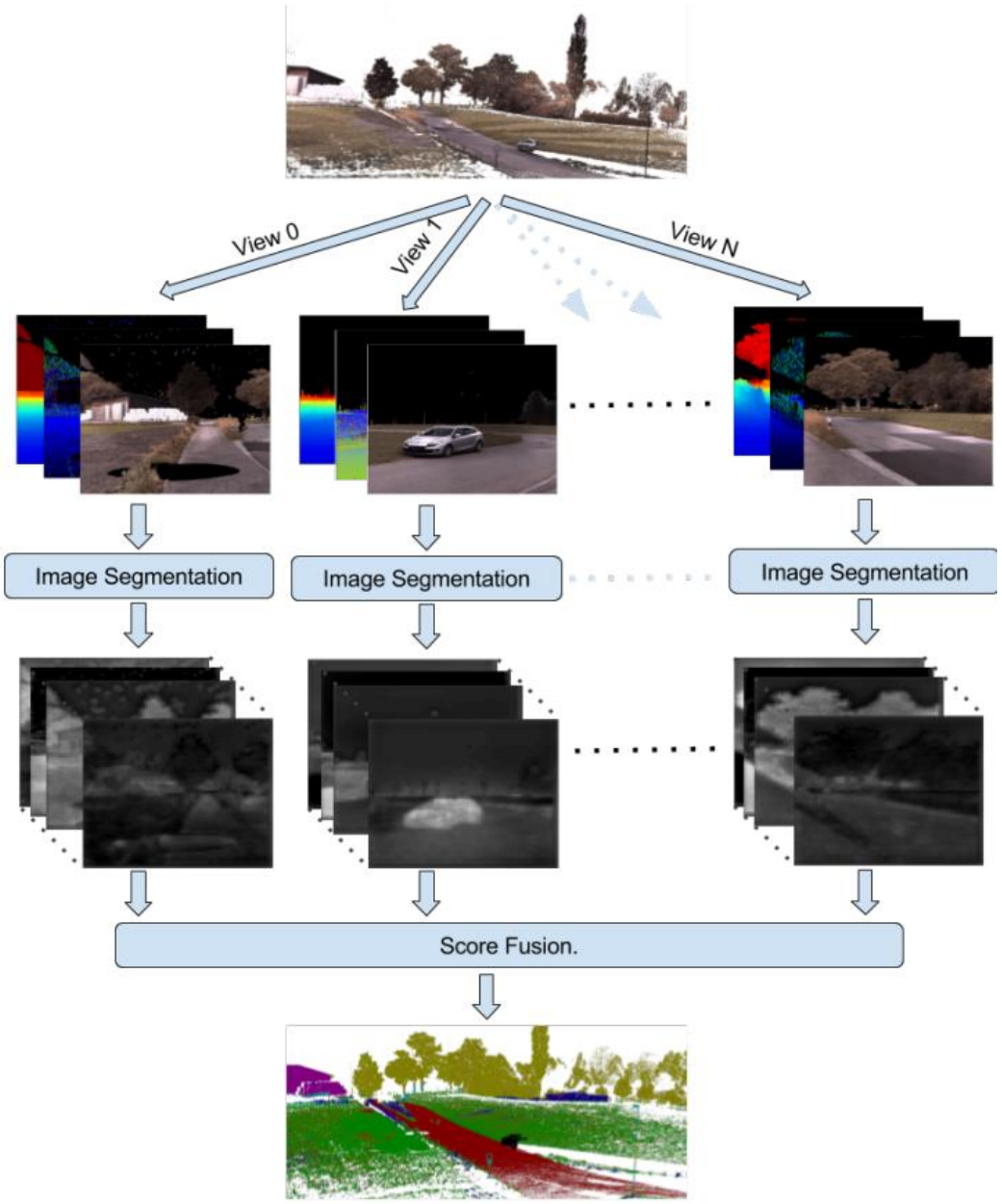
3DCNN based system presented by Huang et al. for semantic labeling of point clouds. Clouds undergo a dense voxelization process and the CNN produces pervoxel labels that are then mapped back to the point cloud.

Deep Projective 3D Semantic Segmentation, 2017

Semantic segmentation of 3D point clouds is a challenging problem with numerous real-world applications. While deep learning has revolutionized the field of image semantic segmentation, its impact on point cloud data has been limited so far. Recent attempts, based on 3D deep learning approaches (3DCNNs), have achieved below-expected results. Such methods require voxelizations of the underlying point cloud data, leading to decreased spatial resolution and increased memory consumption. Additionally, 3D-CNNs greatly suffer from the limited availability of annotated datasets.

In this paper, we propose an alternative framework that avoids the limitations of 3D-CNNs. Instead of directly solving the problem in 3D, we first project the point cloud onto a set of synthetic 2D-images. These images are then used as input to a 2D-CNN, designed for semantic segmentation. Finally, the obtained prediction scores are re-projected to the point cloud to obtain the segmentation results. We further investigate the impact of multiple modalities, such as color, depth and surface normals, in a multi-stream network architecture.

Deep Projective 3D Semantic Segmentation, 2017



An overview of the proposed method. The input point cloud is projected into multiple virtual camera views, generating 2D color, depth and surface normal images. The images for each view are processed by a multi-stream CNN for semantic segmentation. The output prediction scores from all views are fused into a single prediction for each point, resulting in a 3D semantic segmentation of the point cloud.

Deep Projective 3D Semantic Segmentation, 2017

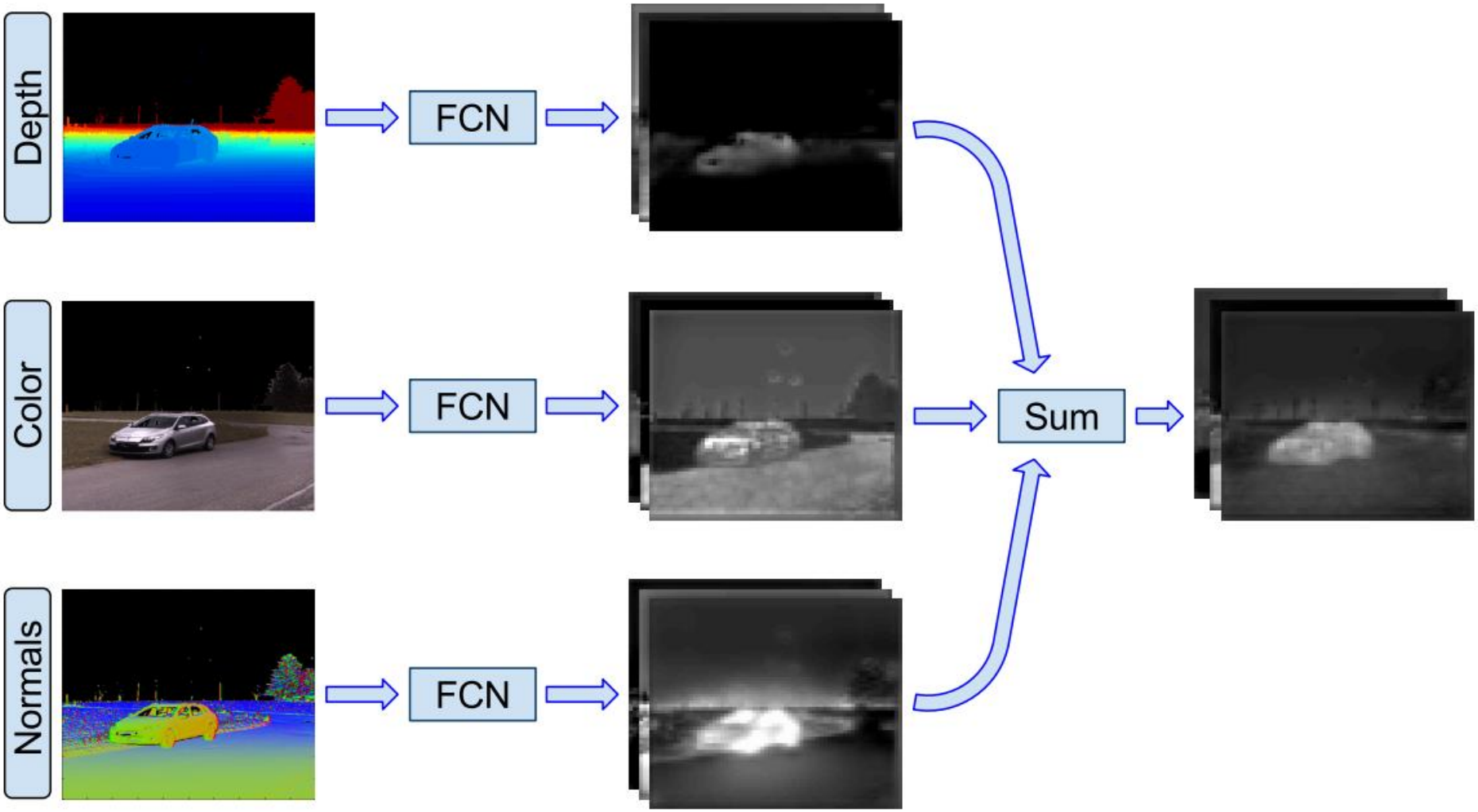


Illustration of the proposed multi-stream architecture for 2D semantic segmentation. Each input stream is processed by a Fully Convolutional Network. The prediction scores from each stream are summed to get the final prediction.

Depth-aware CNN for RGB-D Segmentation, 2018

Convolutional neural networks (CNN) are limited by the lack of capability to handle geometric information due to the fixed grid kernel structure. The availability of depth data enables progress in RGB-D semantic segmentation with CNNs. State-of-the-art methods either use depth as additional images or process spatial information in 3D volumes or point clouds. These methods suffer from **high computation and memory cost**.

To address these issues, we present Depth-aware CNN by introducing two intuitive, flexible and effective operations: **depth-aware convolution and depth-aware average pooling**. By leveraging depth similarity between pixels in the process of information propagation, geometry is seamlessly incorporated into CNN. Without introducing any additional parameters, both operators can be easily integrated into existing CNNs.

Depth-aware CNN for RGB-D Segmentation, 2018

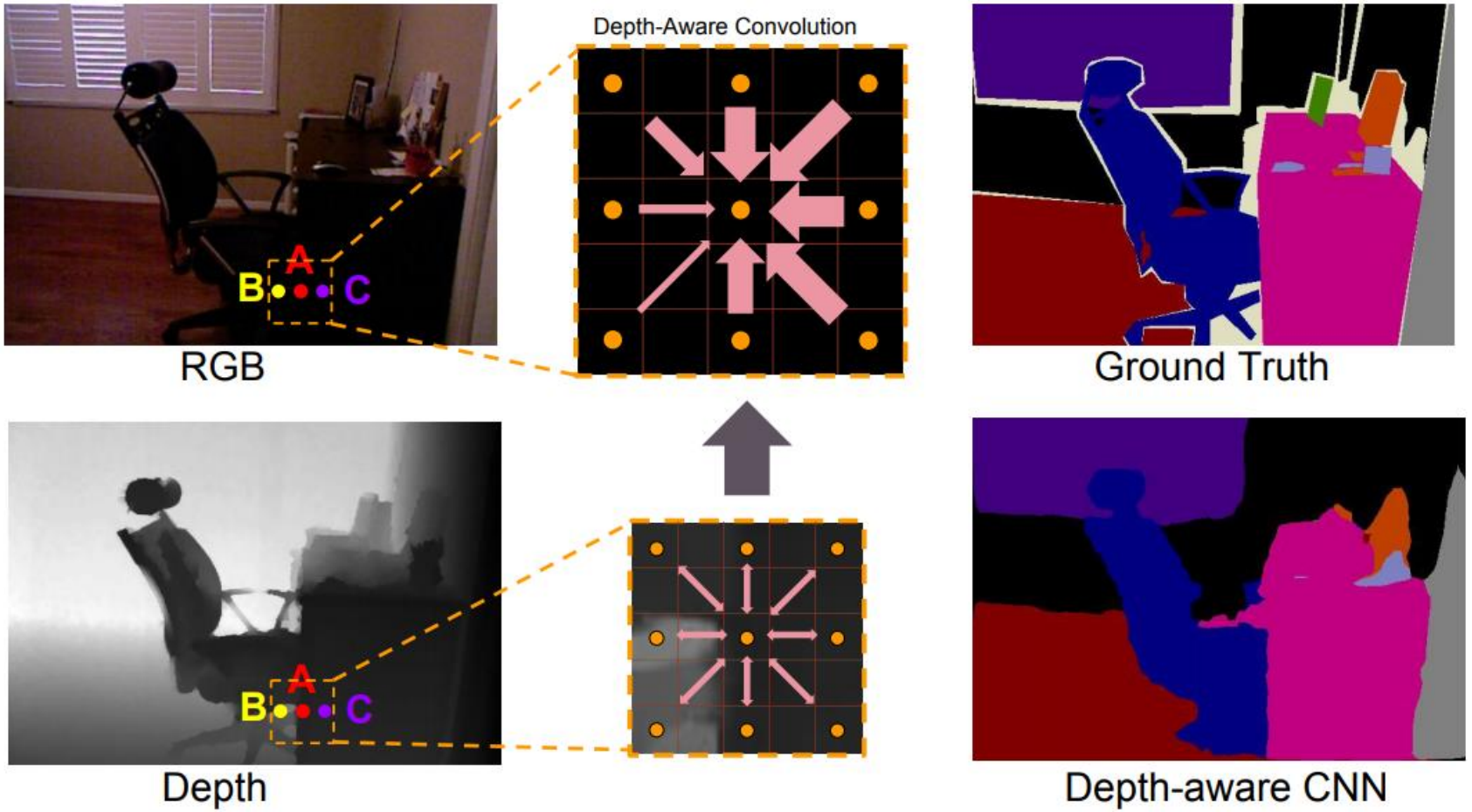


Illustration of Depth-aware CNN. A and C are labeled as table and B is labeled as chair. They all have similar visual features in the RGB image, while they are separable in depth. Depth-aware CNN incorporate the geometric relations of pixels in both convolution and pooling. When A is the center of the receptive field, C then has more contribution to the output unit than B. Figures in the rightmost column shows the RGB-D semantic segmentation result of Depth-aware CNN.

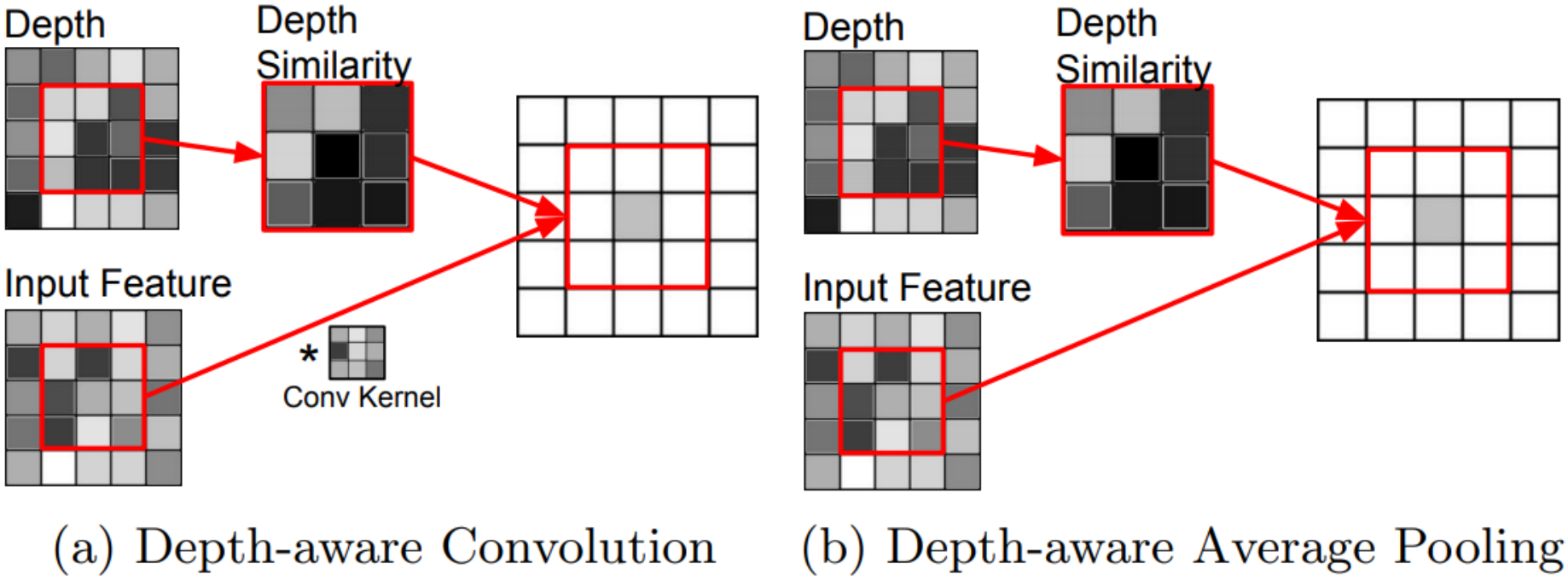
Depth-aware CNN for RGB-D Segmentation, 2018

Two new operators are introduced: (a) depth-aware convolution and (b) depth-aware average pooling.

Depth-aware convolution augments the standard convolution with a depth similarity term. We force pixels with similar depths with the center of the kernel to have more contribution to the output than others. This simple depth similarity term efficiently incorporates geometry in a convolution kernel and helps build a depth-aware receptive field, where convolution is not constrained to the fixed grid geometric structure.

The second introduced operator is **depth-ware average pooling**. Similarly, when a filter is applied on a local region of the feature map, the pairwise relations in depth between neighboring pixels are considered in computing mean of the local region. Visual features are able to propagate along with the geometric structure given in depth images. Such geometry-aware operation enables the localization of object boundaries with depth images.

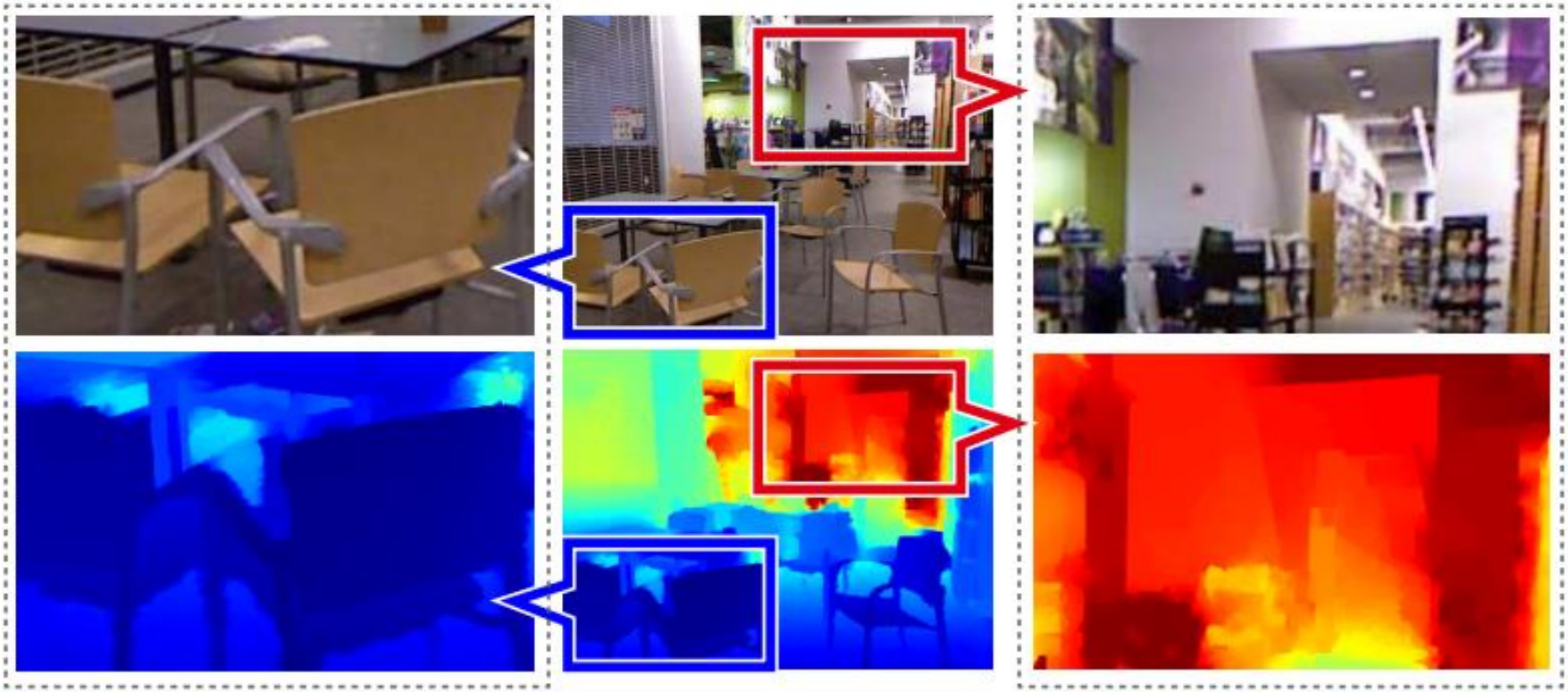
Depth-aware CNN for RGB-D Segmentation, 2018



(a) Depth-aware Convolution (b) Depth-aware Average Pooling

Illustration of information propagation in Depth-aware CNN. Without loss of generality, we only show one filter window with kernel size 3 × 3. In depth similarity shown in figure, darker color indicates higher similarity while lighter color represents that two pixels are less similar in depth. In (a), the output activation of depth-aware convolution is the multiplication of depth similarity window and the convolved window on input feature map. Similarly in (b), the output of depth-aware average pooling is the average value of the input window weighted by the depth similarity.

Cascaded Feature Network for Semantic Segmentation of RGB-D Images, ICCV 2017



There is correlation between depth and scene-resolution: the near field (highlighted in blue rectangle) consists of high scene-resolution, while the far field (highlighted in red rectangle) has low scene resolution.

Cascaded Feature Network for Semantic Segmentation of RGB-D Images, ICCV 2017

The key idea of this work is to use the depth to split the image into layers representing similar visual characteristic, or the “scene-resolution”, e.g., the resolution of the objects and scenes in generals.

There is correlation between depth and scene-resolution: lower scene-resolution appears in regions that have higher depth, and higher scene-resolution appears in the near field. In lower scene-resolution regions, objects and scenes densely co-exist, forming more complex correlation between objects/scenes relative to higher scene-resolution regions.

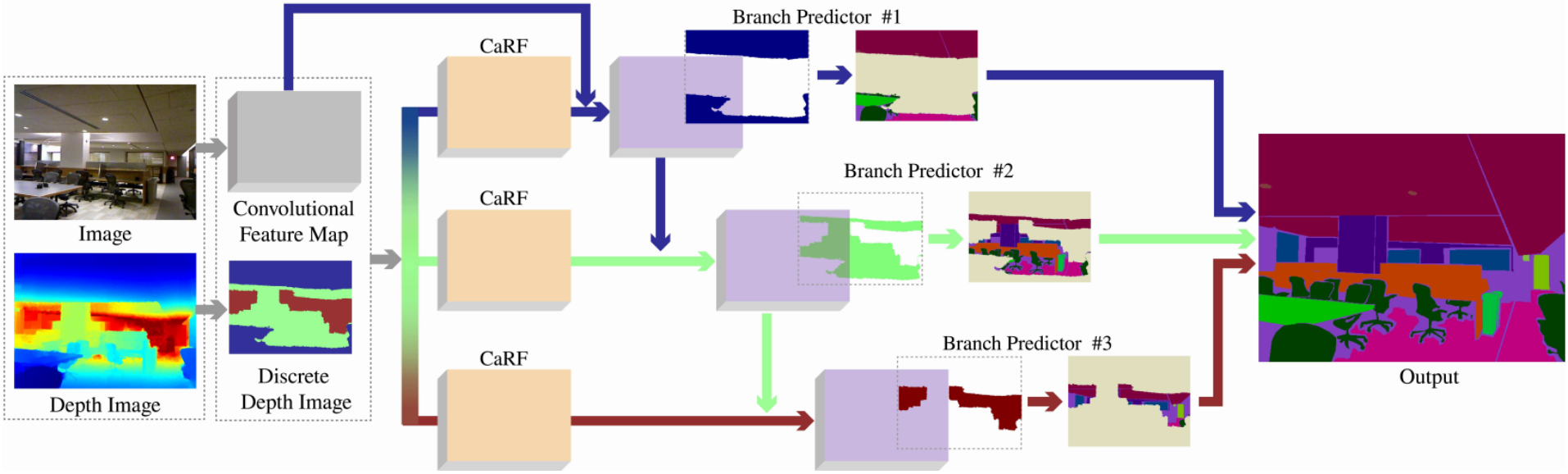
To better represent and learn the variant object/scene relationships, appropriate features should be constructed for different scene-resolutions.

Cascaded Feature Network for Semantic Segmentation of RGB-D Images, ICCV 2017

First, to make the feature more focused on the common visual characteristic of the observed scene, a context-aware receptive field (CaRF) is introduced. The CaRF provides a better control on the relevant contextual information of the learned features. The CaRFs are computed based on super-pixels, which are defined by the underlying scene structures. Thus, the contextual information provided by CaRF can alleviate negative effect of mixing the features of overly small or large regions.

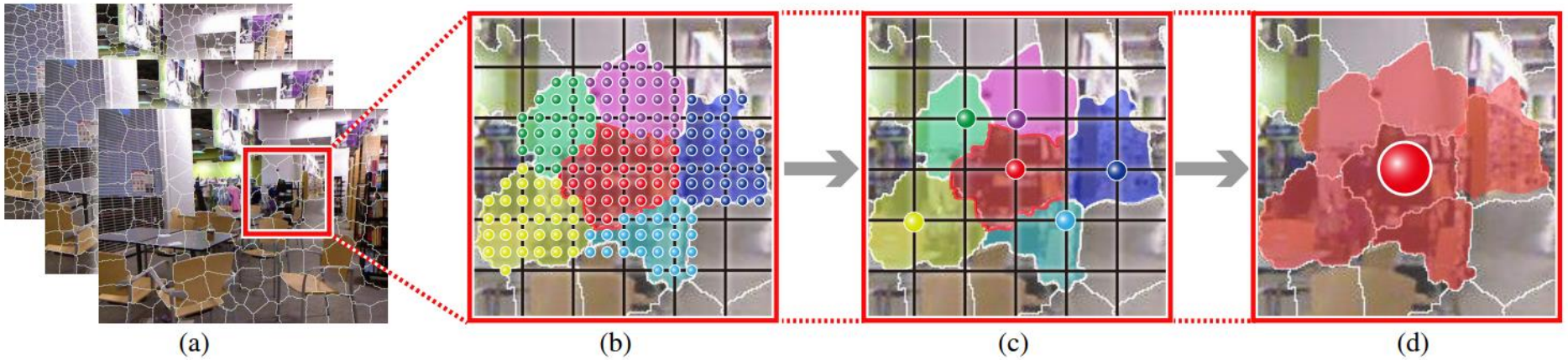
Second, a cascaded feature network (CFN) with parallel branches is developed, each of which focuses on semantic segmentation of regions of certain scene-resolution. Each branch is equipped with a CaRF. The combination of CaRF and cascaded network, enables regions in different scene-resolutions to communicate each other so as to wisely update shared convolutional features.

Cascaded Feature Network for Semantic Segmentation of RGB-D Images, ICCV 2017



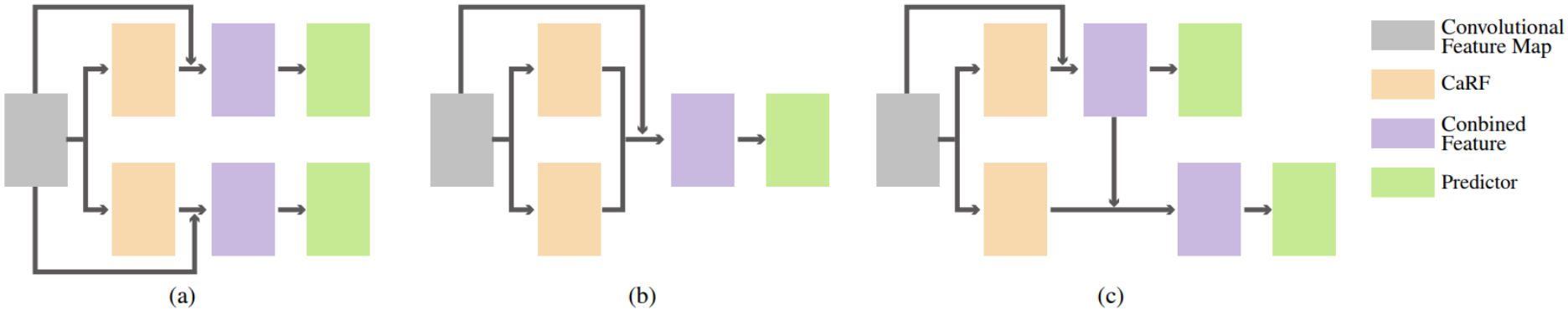
The overview of our cascaded feature network (CFN). Given the color image, we use CNN to compute the convolutional feature map. The discrete depth image is layered, where each layer represents a scene-resolution and is used to match the image regions to corresponding network branches that share the same convolutional feature map. Each branch has context-aware receptive field (CaRF), which produces contextual representation to combine with the feature from adjacent branch. The predictions of all branches are combined to achieve the eventual segmentation result.

Cascaded Feature Network for Semantic Segmentation of RGB-D Images, ICCV 2017



The two-level Context-aware Receptive Field (CaRF): (a) the image partitioned into super-pixels with different sizes; (b) at each node of the coarse grid we aggregate the features that reside in the same super-pixel; (c) the content of adjacent super-pixels is aggregated; (d) the aggregated content in a feature map represents a CaRF. The two-level CaRF is repeatedly applied to the images partitioned by super-pixels with diverse sizes. Note that the feature map has smaller resolution than the image due to down-sampling of network.

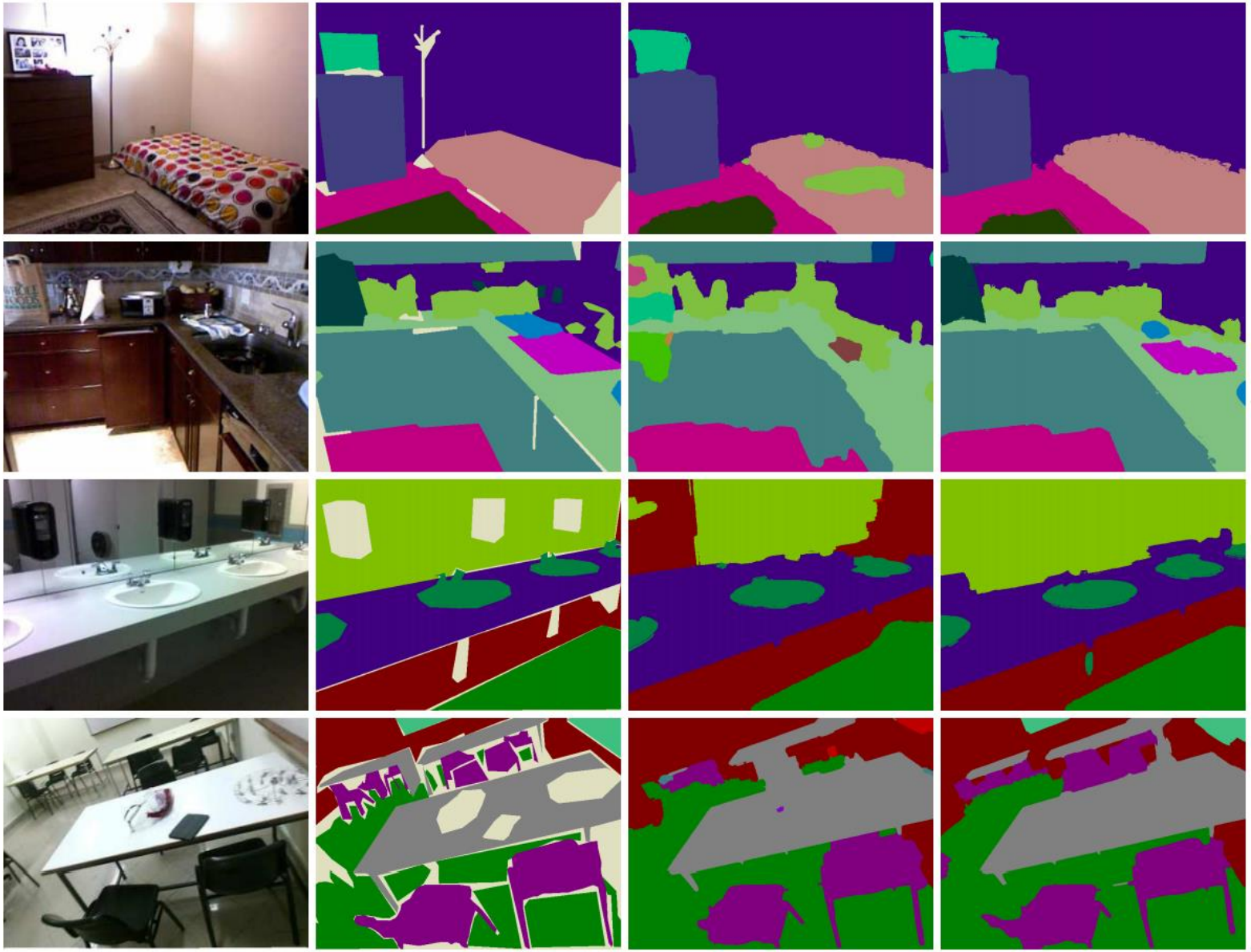
Cascaded Feature Network for Semantic Segmentation of RGB-D Images, ICCV 2017



The network can have separate branches (a), combined branches (b) or cascaded branches (c). For clarity, we illustrate it with two branches only. Each network can be extended to have more branches.

The idea of Context-aware Receptive Field (CaRF) is to aggregate convolutional features of local context into richer features that learn better the relevant content, where the receptive field is spatially-variant and defined its extent according the local context.

Cascaded Feature Network for Semantic Segmentation of RGB-D Images, ICCV 2017



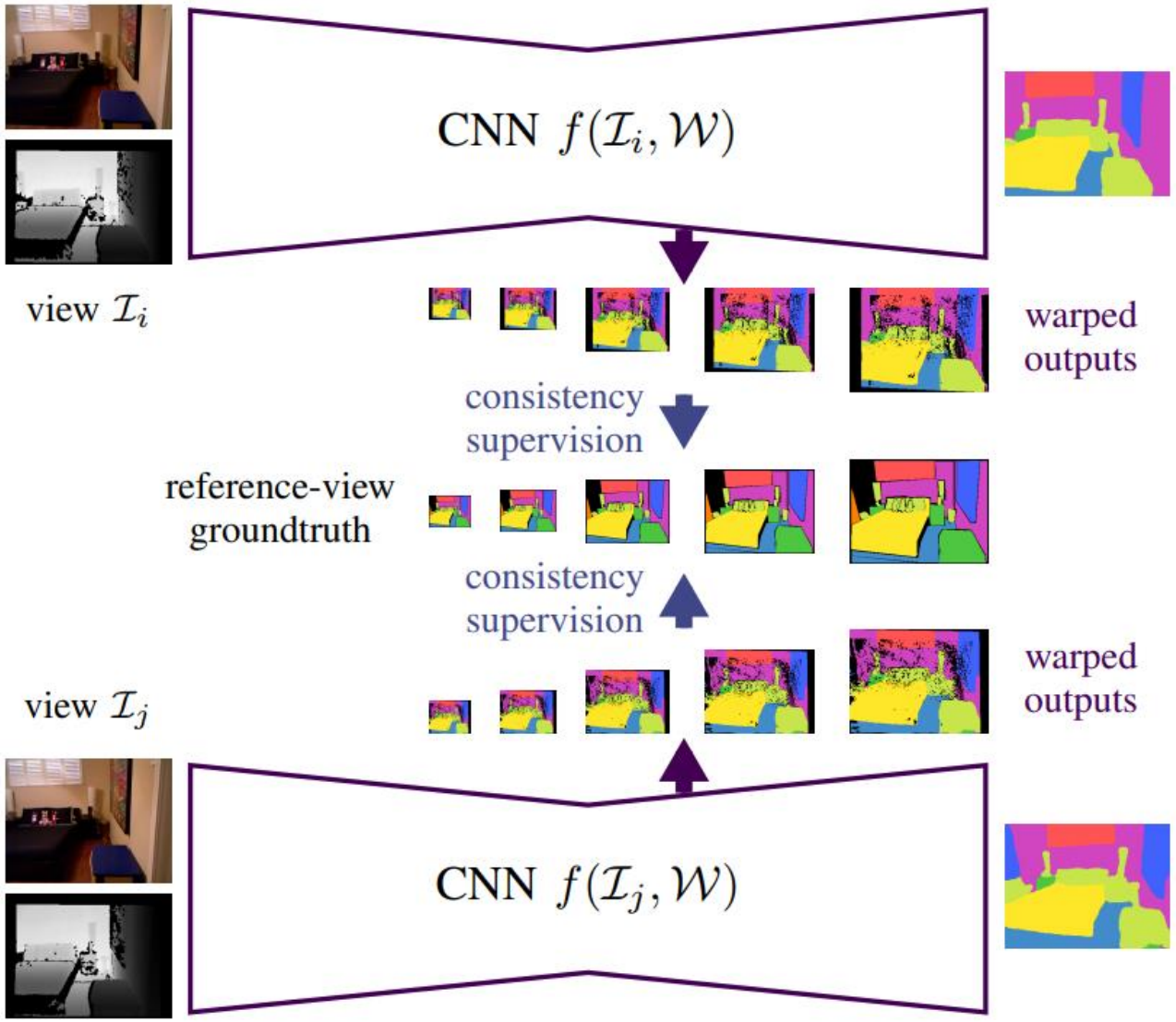
(a) Image (b) Ground-truth (c) Baseline (d) CFN

Multi-View Deep Learning for Consistent Semantic Mapping with RGB-D Cameras, 2017

A novel deep learning approach is developed for semantic segmentation of RGB-D images with multi-view context, where RGB and depth fusion are seamlessly fused via multi-scale deep supervision and multi-view consistency constraints.

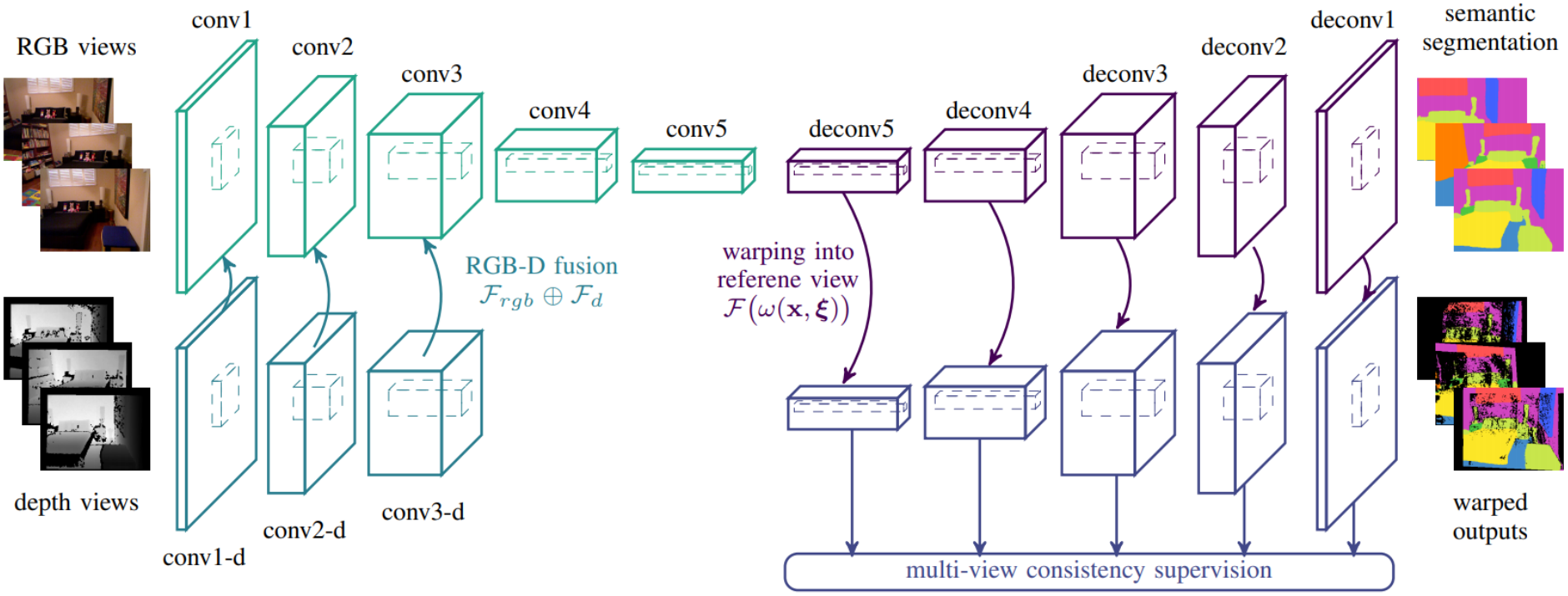
A shared principle is using the SLAM trajectory estimate to warp network outputs of multiple frames into the reference view with ground-truth annotation. By this, the network can learn features that are invariant under view-point change.

Multi-View Deep Learning for Consistent Semantic Mapping with RGB-D Cameras, 2017



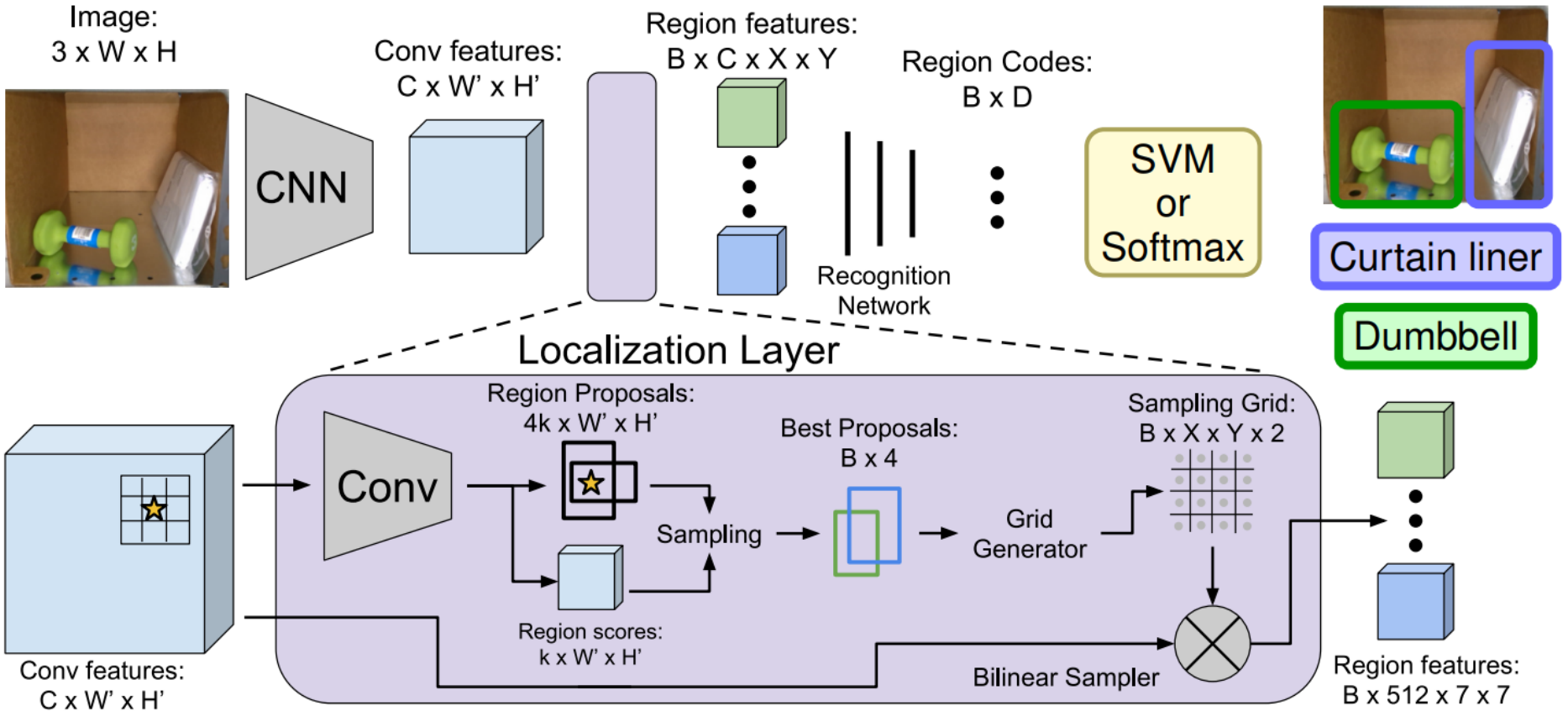
The key innovation is to enforce consistency by warping CNN feature maps from multiple views into a common reference view using the SLAM trajectory and to supervise training at multiple scales. Our approach improves performance for single-view segmentation and is specifically beneficial for multi-view fused segmentation.

Multi-View Deep Learning for Consistent Semantic Mapping with RGB-D Cameras, 2017

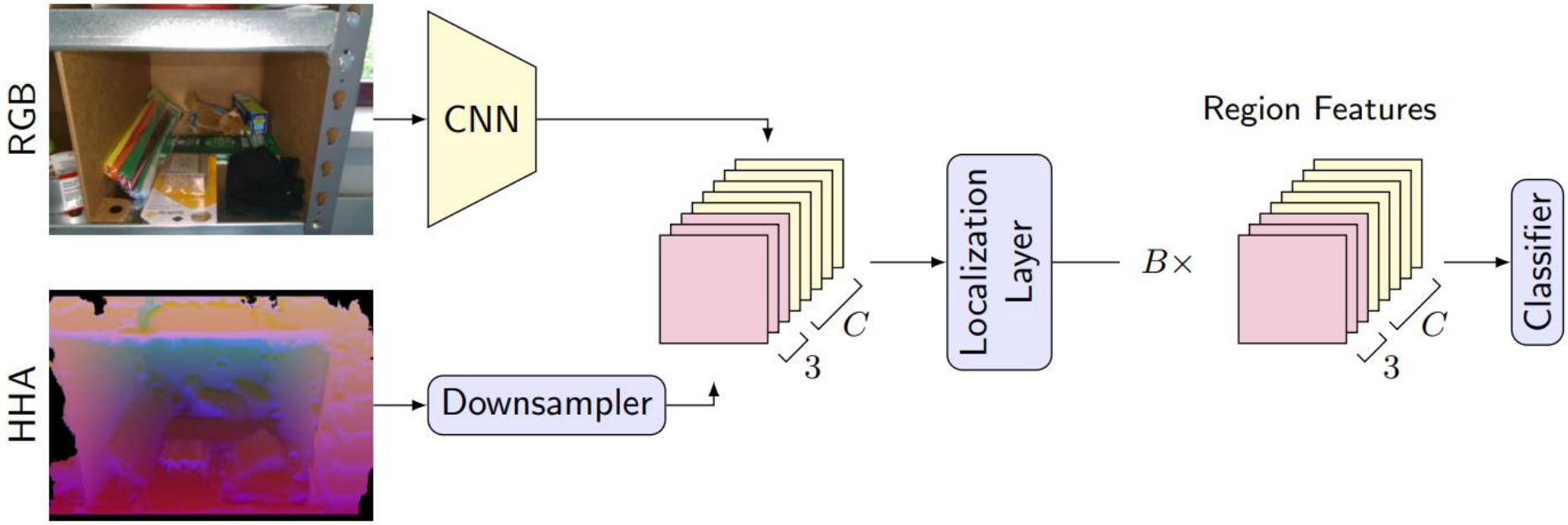


The CNN encoder-decoder architecture used in our approach. Input to the network are RGB-D sequences with corresponding poses from SLAM trajectory. The encoder contains two branches to learn features from RGB-D data as inspired by FuseNet. The obtained low-resolution high-dimension feature maps are successively refined through deconvolutions in the decoder. We warp feature maps into a common reference view and enforce multi-view consistency with various constraints. The network is trained in a deeply-supervised manner where loss is computed at all scales of the decoder

RGB-D Object Detection and Semantic Segmentation for Autonomous Manipulation in Clutter, 2017

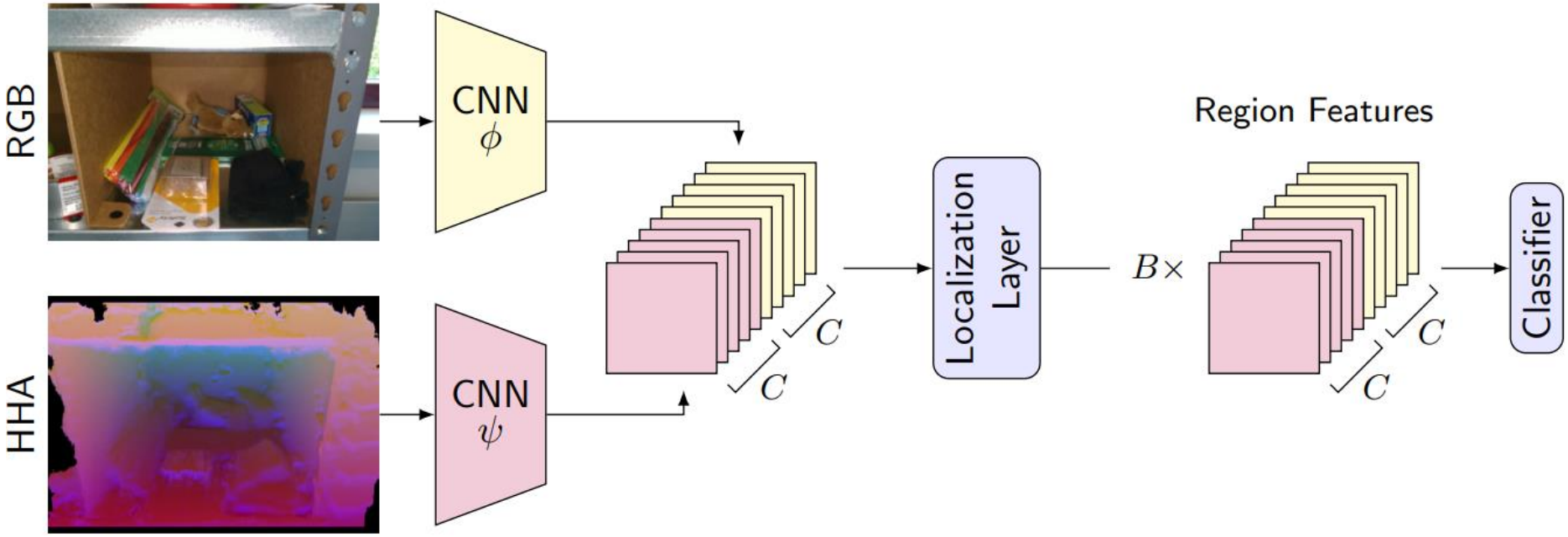


RGB-D Object Detection and Semantic Segmentation for Autonomous Manipulation in Clutter, 2017



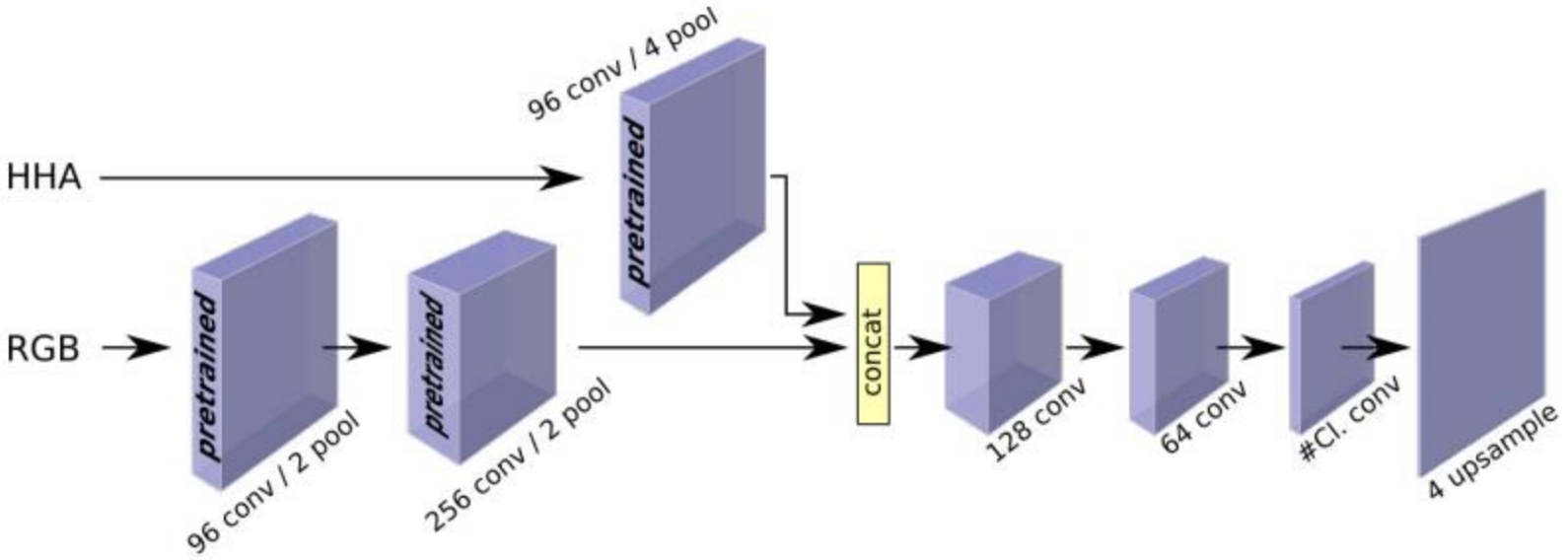
Detection pipeline with CNN features from RGB and downsampled HHA-encoded depth. C denotes the number of CNN feature maps after the last convolutional layer (512 for VGG-16). The internal proposal generator produces B proposals (1000).

RGB-D Object Detection and Semantic Segmentation for Autonomous Manipulation in Clutter, 2017

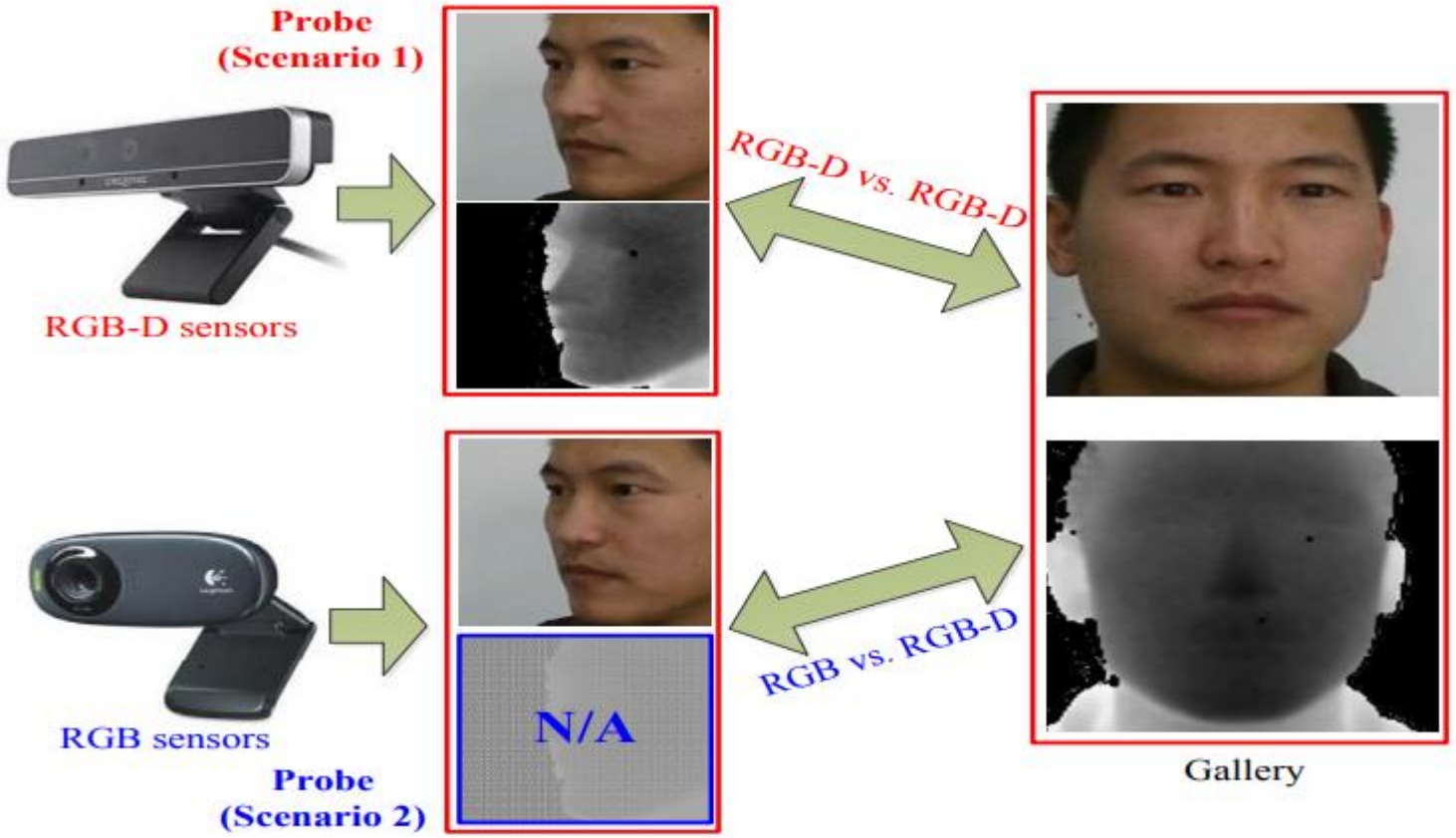


Detection pipeline with concatenated CNN features from RGB and HHA-encoded depth. C denotes the number of CNN feature maps after the last convolutional layer (512 for VGG-16). The internal proposal generator produces B proposals (1000). For the Cross Modal Distillation approach, CNN ψ is trained to imitate the pretrained CNN φ.

RGB-D Object Detection and Semantic Segmentation for Autonomous Manipulation in Clutter, 2017

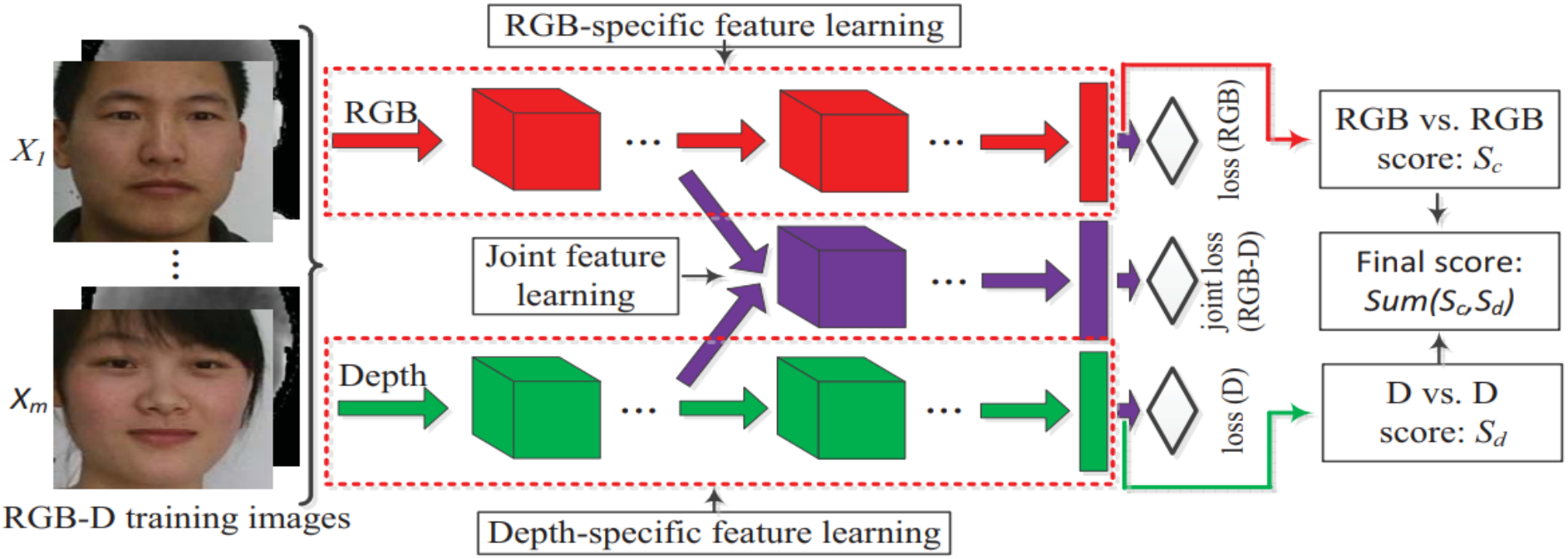


RGB-D Face Recognition via Deep Complementary and Common Feature Learning, 2018



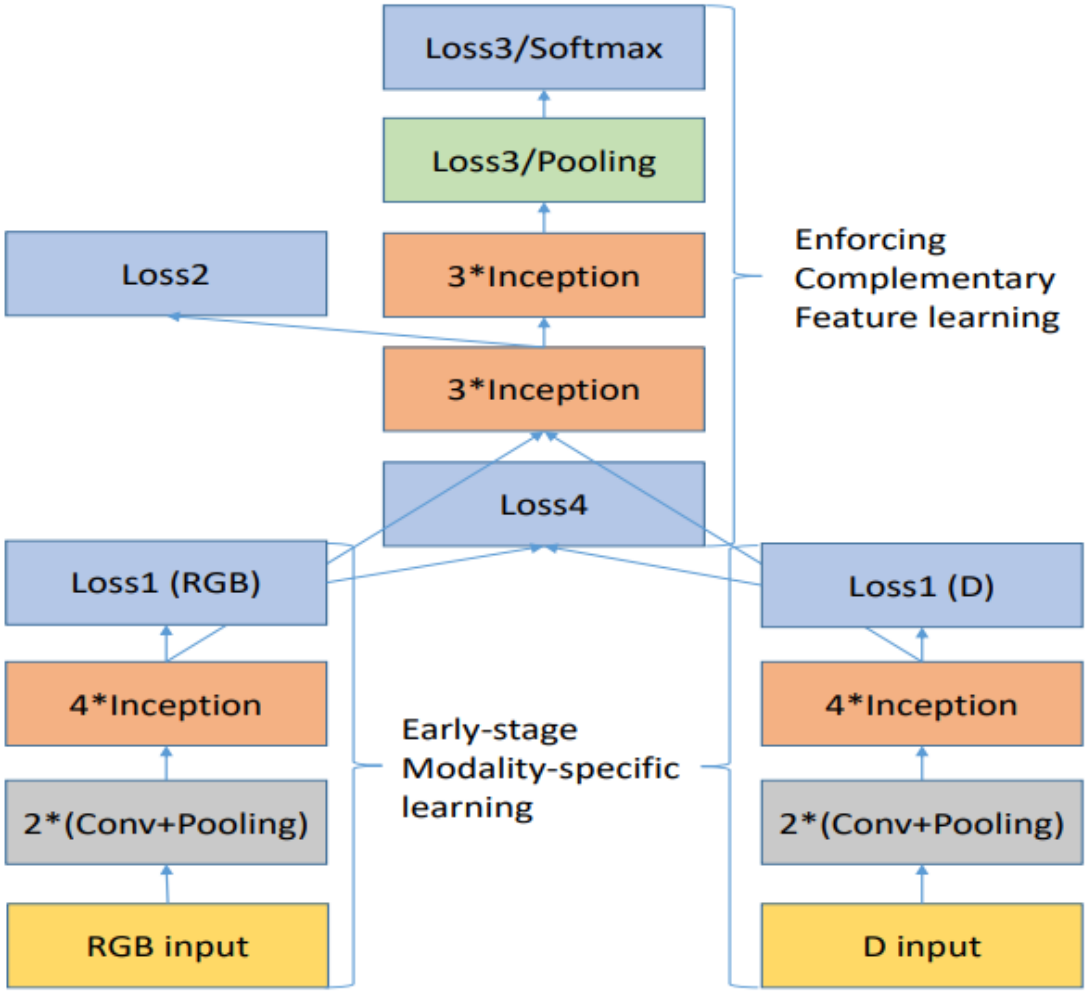
RGB-D face recognition (FR) consists of two typical senarios: (1) multi-modality matching, e.g., RGB-D probe vs. RGB-D gallery, where both the enrolled gallery and the probe images are captured using RGBD sensors, and (2) cross-modality matching, e.g., RGB probe vs. RGB-D gallery, where the gallery images remain RGB-D, but the probe images are captured by RGB sensors. The proposed approach addresses the two problems by learning complementary and common features.

RGB-D Face Recognition via Deep Complementary and Common Feature Learning, 2018



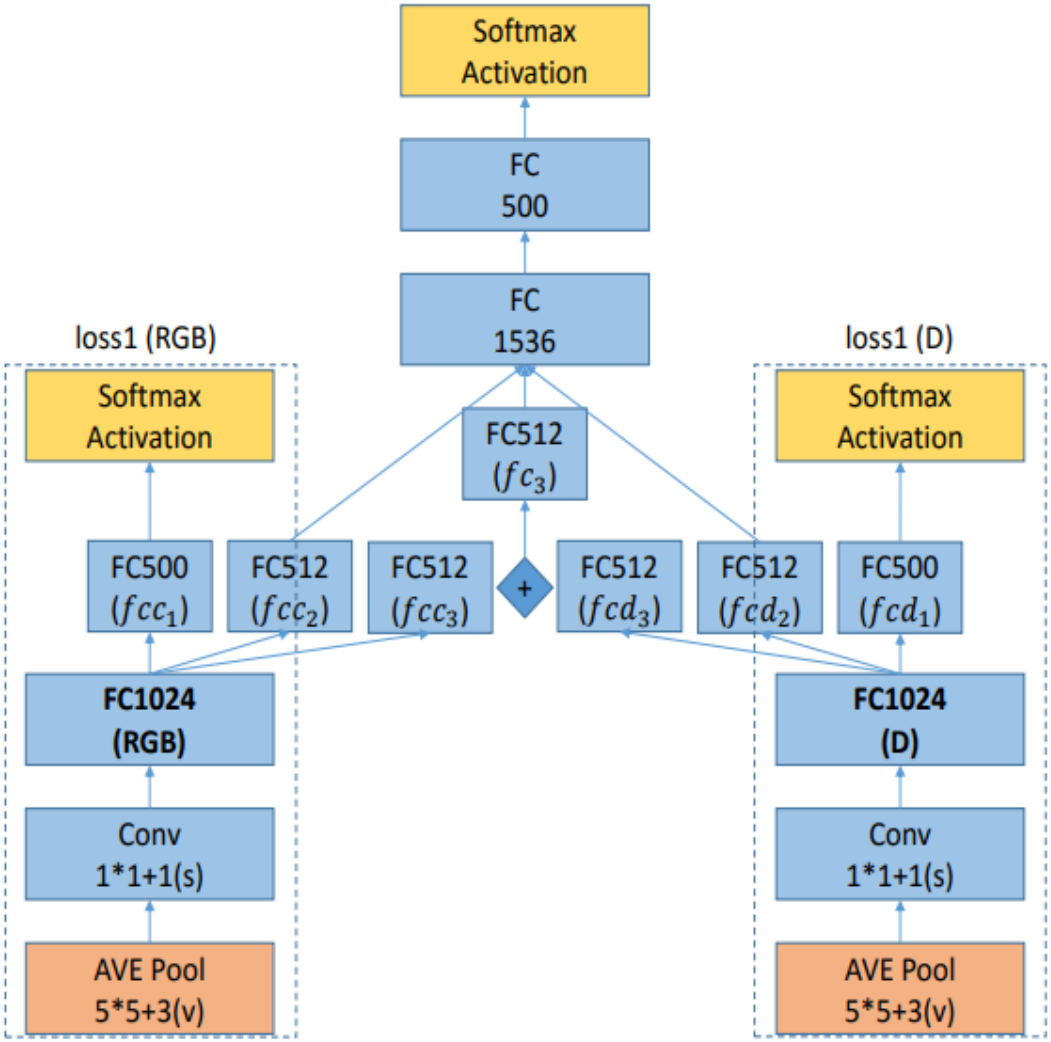
Overview of the proposed complementary feature learning approach from RGB and depth modalities, which handles multi-modality FR scenario such as RGB-D probe vs. RGB-D gallery.

RGB-D Face Recognition via Deep Complementary and Common Feature Learning, 2018



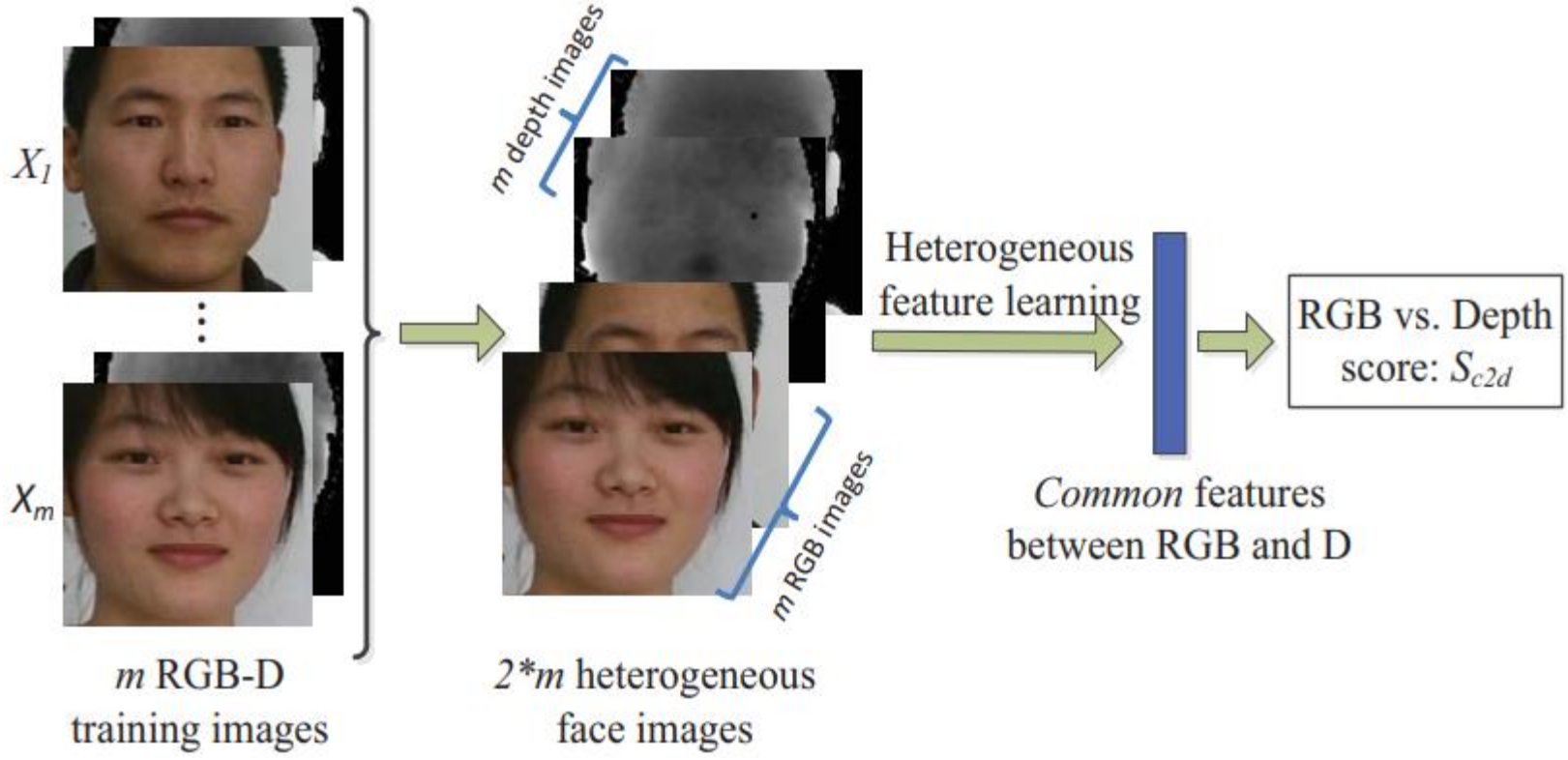
The network architecture of our complementary feature learning.

RGB-D Face Recognition via Deep Complementary and Common Feature Learning, 2018



The details of the loss4 unit introduced to enforce complementary feature learning. The plus symbol denotes the element-wise average.

RGB-D Face Recognition via Deep Complementary and Common Feature Learning, 2018



Overview of the proposed common feature learning approach between RGB and depth modalities, which handles cross-modality FR scenario such as RGB probe vs. depth gallery.