National University of Singapore

Attention-based 2.5D segmentation with a generative adversarial network

shou zhang, linfeng li, yixuan zhou, yinjie zheng

Abstract

Compared with RGB information, RGB-D information introduces depth information, which can provide the corresponding geometric relationship for the RGB image. However, directly inputting depth data into the existing RGB semantic segmentation framework or simply fusing the two domain features lead to performance degradation. To fuse two domain features and extract useful information from them, we proposed a transfusion module, where a transformer block is designed to extract global information of the fused features.Besides, in a multimodal task, several loss functions should be created to supervise the training of each modal. While instead of a complicated loss function, we introduce a GAN architecture and utilize the discriminator to act as a powerful loss function to capture the distance between inference and ground truth.Experimental results show that our method brings an improvement over the baseline on NYUDv2 dataset.

Related work

- ACNet is a deep-convolutional based 2.5D segmentation model. To help the two modalities cooperate better, a channel-wise attention module was introduced to mask-out unimportant information.
- GAN is a framework where a generator tries to yield high quality result and a discriminator tries to distinguish ground truth from inferencing result. This zero-sum game serves as an indirect training for generator and has shown good performance in many tasks. In our work, we adopted SalGAN because of its ability to exaggerate visual saliency of an image.
- Transformer is a deep learning model that adopts self-attention mechanism and can differentially weight significance of each part of input data. The multi-head attention module is especially helpful in fusion of feature maps from RGB and depth branch.

Methodology

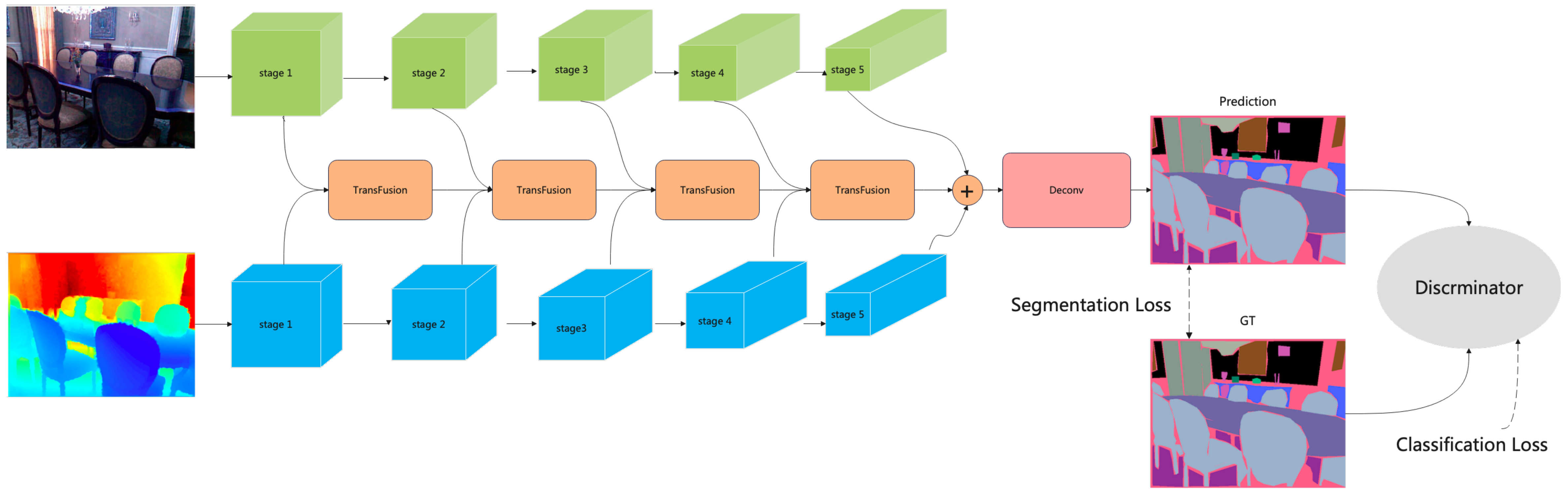
- GAN architecture

  Generator: we design the loss function with both cross-entropy loss between inference and ground truth and the output of discriminator.

$$L_G = CE + log(p_{fake|G})$$

$$CE = -\frac{1}{n}\sum_{i}\sum_{c=1}^{k} y^{ic} * log(p^{ic})$$

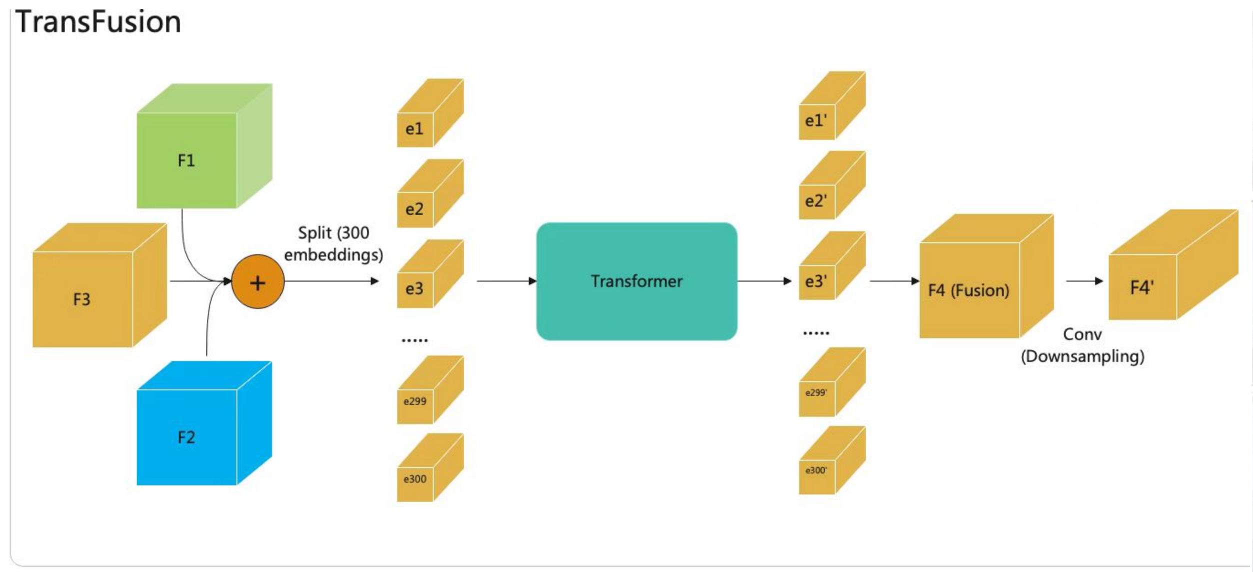
$$L_D = -log(p_{real|GT}) - log(p_{fake|G})$$



  Discriminator: we simply use several conv kernels followed by a fully connected layer to evaluate the similarity between inference and ground truth.

- transfusion

  Transfusion block consists of a transformer block and a convolution kernel. We fuse two domain features with an element-wise strategy, and then we split the fused feature into 300 patches and use the transformer block to extract global information. Finally, a 3 x 3 convolution kernel are used to downsample features and we will bring it to the next transfusion block as an input.



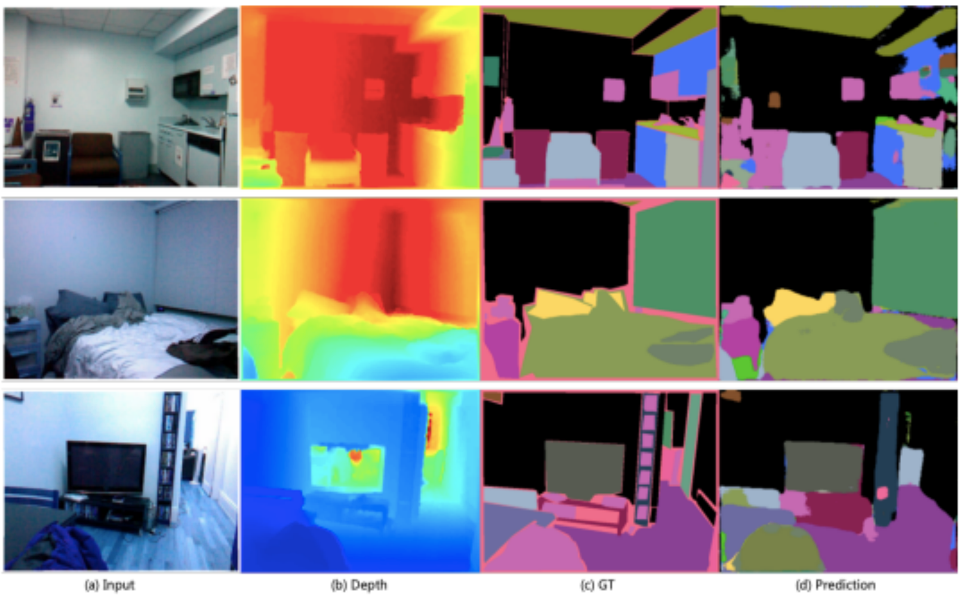


Figure 1: Some qualitative results on NYUDv2.

Experiments

NYUDv2 dataset is used as our training and test set. Same as the official setting, 795 image pairs are used for training and 654 are used for testing. Some qualitative results are show in Figure 1. Some quantitative results are shown in Table 1. Baseline refers to the basic Unet with multimodal inputs, where the decoder fuses the additions of the multi-level features from two encoders for two modalities. Conv fusion represents the fusion block without transformer block.

Table 1: Some results on NYUDv2.

Method	mIOU	PixAcc
baseline (Unet)	0.433	71.7%
+ conv fusion	0.431	71.4%
+ transformer	0.437	71.7%
+ adversarial	0.442	72.1%

Reference

[1] Yanhua Cheng, Rui Cai, Zhiwei Li, Xin Zhao, and Kaiqi Huang. Locality-sensitive deconvolution networks with gated fusion for rgb-d indoor semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 3029–3037, 2017.

[2] Xinxin Hu, Kailun Yang, Lei Fei, and Kaiwei Wang. Acnet: Attention based network to exploit complementary features for rgbd semantic segmentation. In 2019 IEEE International Conference on Image Processing (ICIP),pages 1440–1444. IEEE, 2019.

[3] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 3431–3440, 2015.

[4] Junting Pan, Cristian Canton Ferrer, Kevin McGuinness, Noel E. O'Connor, Jordi Torres, Elisa Sayrol, and Xavier Giro-i Nieto. Salgan: Visual saliency prediction with generative adversarial networks, 2017.

[5] Seong-Jin Park, Ki-Sang Hong, and Seungyong Lee. Rdfnet: Rgb-d multi-level residual feature fusion for indoor semantic segmentation. In Proceedings of the IEEE international conference on computer vision, pages 4980–4989,2017.