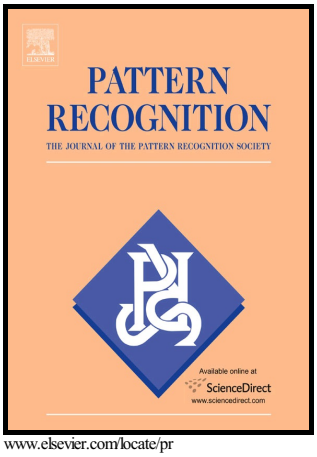
# Hierarchically Supervised Deconvolutional Network for Semantic Video Segmentation

Yuhang Wang, Jing Liu, Yong Li, Jun Fu, Min Xu, Hanqing Lu




www.elsevier.com/locate/pr

PII: S0031-3203(16)30307-7
DOI: http://dx.doi.org/10.1016/j.patcog.2016.09.046
Reference: PR5908

To appear in: *Pattern Recognition*

Received date: 16 March 2016
Revised date: 26 September 2016
Accepted date: 28 September 2016

Cite this article as: Yuhang Wang, Jing Liu, Yong Li, Jun Fu, Min Xu and Hanqing Lu, Hierarchically Supervised Deconvolutional Network for Semanti Video Segmentation, *Pattern Recognitio* http://dx.doi.org/10.1016/j.patcog.2016.09.046

# Hierarchically Supervised Deconvolutional Network for Semantic Video Segmentation

Yuhang Wang[a,b], Jing Liu[a], Yong Li[a,b], Jun Fu[a,b], Min Xu[c], Hanqing Lu[a]

[a] *National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China*
[b] *University of Chinese Academy of Sciences, Beijing, China*
[c] *University of Technology, Sydney, NSW, Australia*



**Abstract**

Semantic video segmentation is a challenging task of fine-grained semantic understanding of video data. In this paper, we present a jointly trained deep learning framework to make the best use of spatial and temporal information for semantic video segmentation. Along the spatial dimension, a hierarchically supervised deconvolutional neural network (HDCNN) is proposed to conduct pixel-wise semantic interpretation for single video frames. HDCNN is constructed with convolutional layers in VGG-net and their mirrored deconvolutional structure, where all fully connected layers are removed. And hierarchical classification layers are added to multi-scale deconvolutional features to introduce more contextual information for pixel-wise semantic interpretation. Besides, a coarse-to-fine training strategy is adopted to enhance the performance of foreground object segmentation in videos. Along the temporal dimension, we introduce Transition Layers upon the structure of HDCNN to make the pixel-wise label prediction consist with adjacent pixels across space and time domains. The learning process of the Transition Layers can be implemented as a set of extra convolutional calculations connected with HDCNN. These two parts are jointly trained as a unified deep network in our approach. Thorough evaluations are performed on two challenging video datasets, i.e., CamVid and GATECH. Our approach achieves state-of-the-art performance on both of the two datasets.

*Keywords:*
Semantic video segmentation, Deconvolutional neural network,
Coarse-to-fine training, Spatio-temporal consistence