UC San Diego

Lecture 11:

Deep Learning on Multi-view Representation

Instructor: Hao Su

Feb 15, 2018

Agenda

- **Deep Learning on Multi-view Representation**
  - **Classification**
  - Segmentation
  - Reconstruction

Task: 3D classification

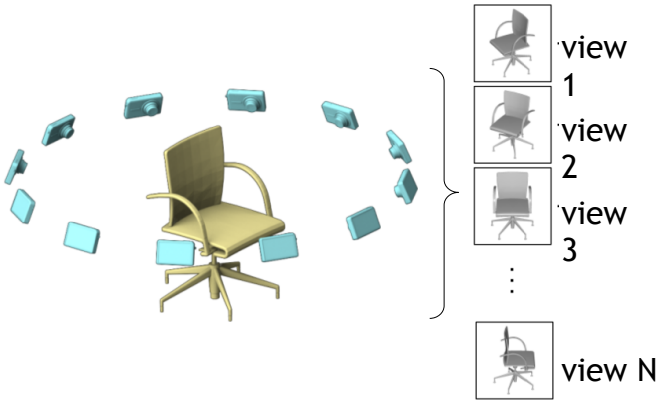
This is a chair!

Given an input shape

Hang Su, Subhransu Maji, Evangelos Kalogerakis, Erik Learned-Miller, "**Multi-view Convolutional Neural Networks for 3D Shape Recognition**", *Proceedings of ICCV 2015*

[credit: Hang S

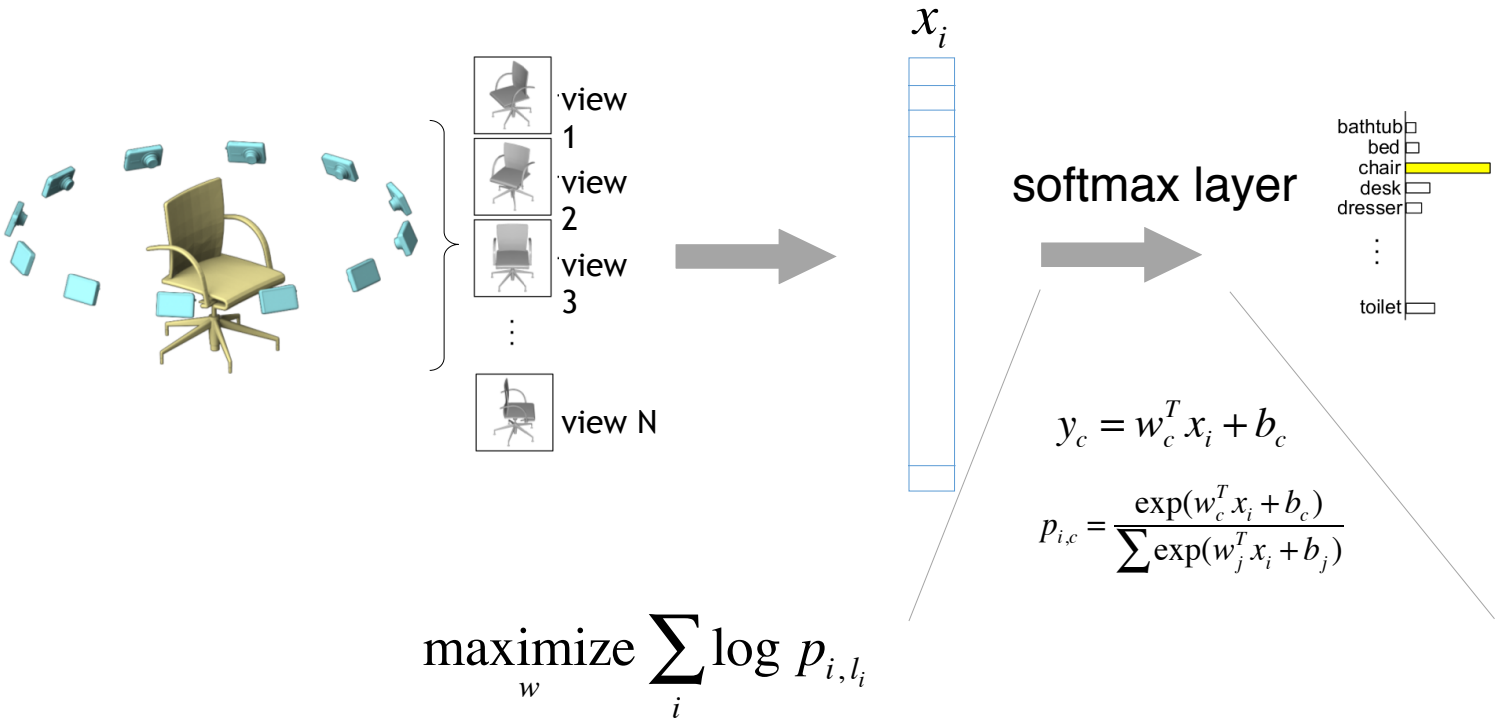
Render with multiple virtual cameras



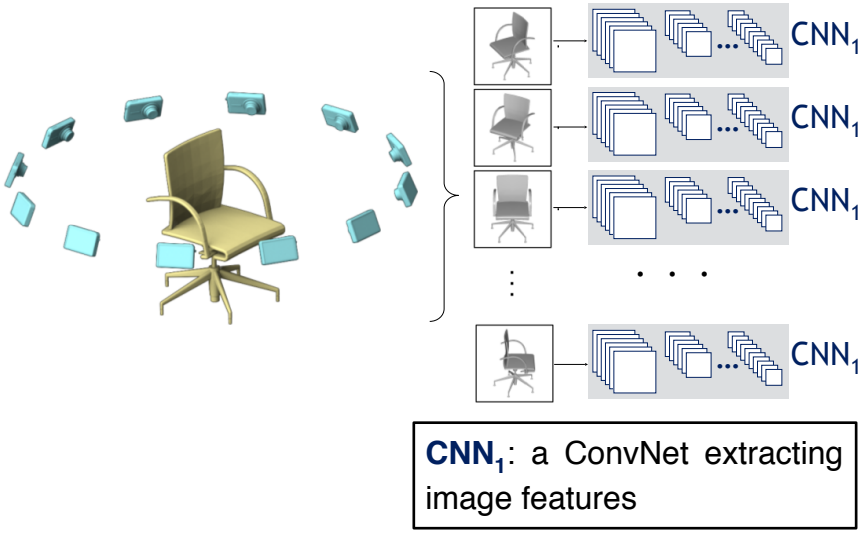
Hang Su, Subhransu Maji, Evangelos Kalogerakis, Erik Learned-Miller, "**Multi-view Convolutional Neural Networks for 3D Shape Recognition**", *Proceedings of ICCV 2015*

[credit: Hang S

Traditional approach: feature+linear classifier



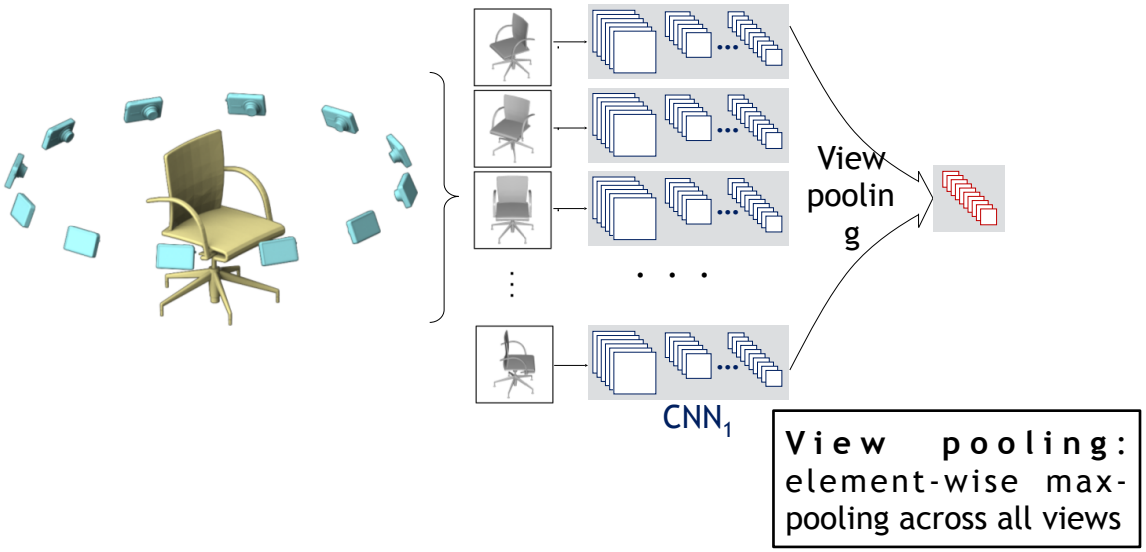
The rendered images are passed through CNN_1 for image features



Hang Su, Subhransu Maji, Evangelos Kalogerakis, Erik Learned-Miller, "**Multi-view Convolutional Neural Networks for 3D Shape Recognition**", *Proceedings of ICCV 2015*

[credit: Hang Su]

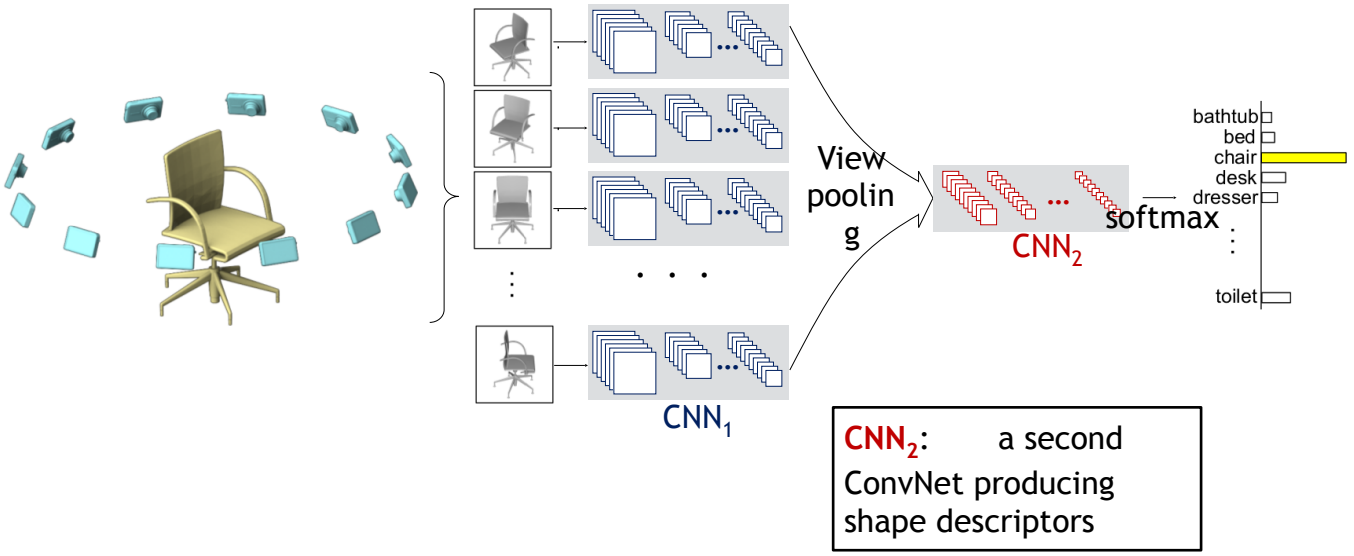
All image features are combined by view pooling ...



Hang Su, Subhransu Maji, Evangelos Kalogerakis, Erik Learned-Miller, "**Multi-view Convolutional Neural Networks for 3D Shape Recognition**", *Proceedings of ICCV 2015*

[credit: Hang Su]

… and then passed through CNN_2 and to generate final predictions

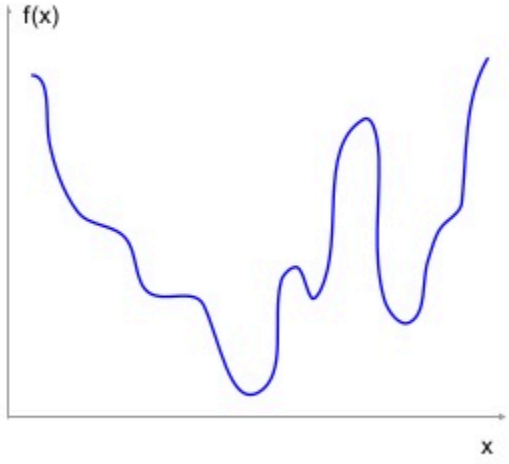


Hang Su, Subhransu Maji, Evangelos Kalogerakis, Erik Learned-Miller, "**Multi-view Convolutional Neural Networks for 3D Shape Recognition**", *Proceedings of ICCV 2015*

[credit: Hang Su]

Learning by fine-tuning

- Neural network optimization is non-convex



- In general, training from more data converges at a better local minima

- However, what if your training dataset D is not big?

Learning by fine-tuning (cont.)

Pre-training

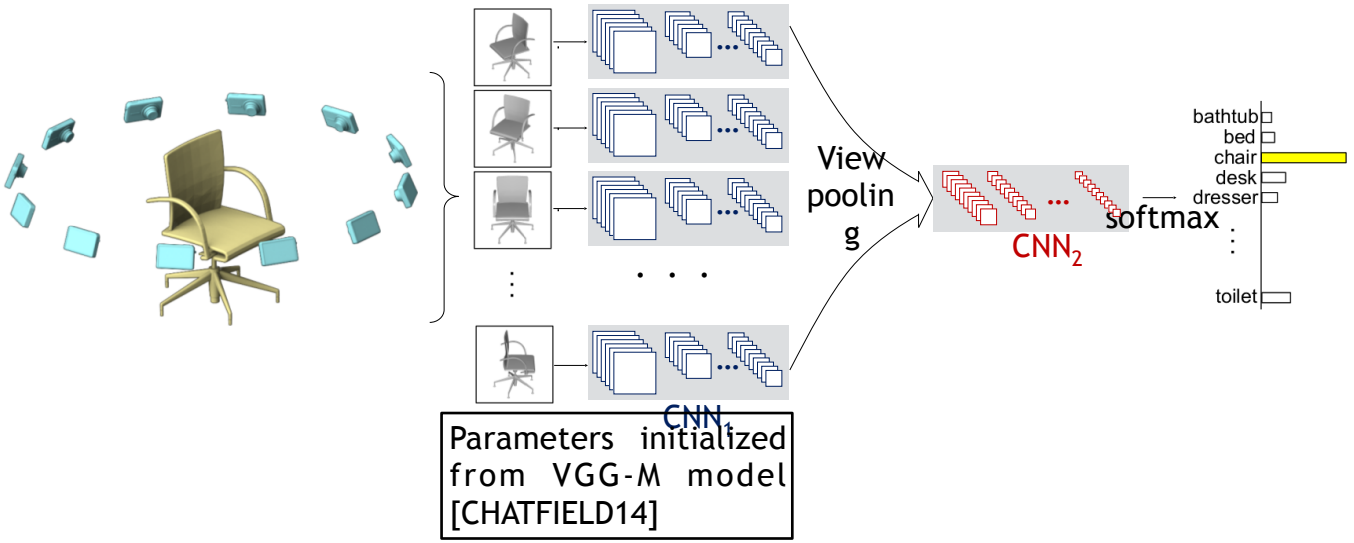
- Find a source of massive data D' with similar statistics
- Learn the network parameters from D'

Fine-tuning

- Starting from the learned parameters on D', minimize the network loss on D

A technique for *transfer learning,* quite effective in practice

Training: network parameters are pre-trained on image classification …

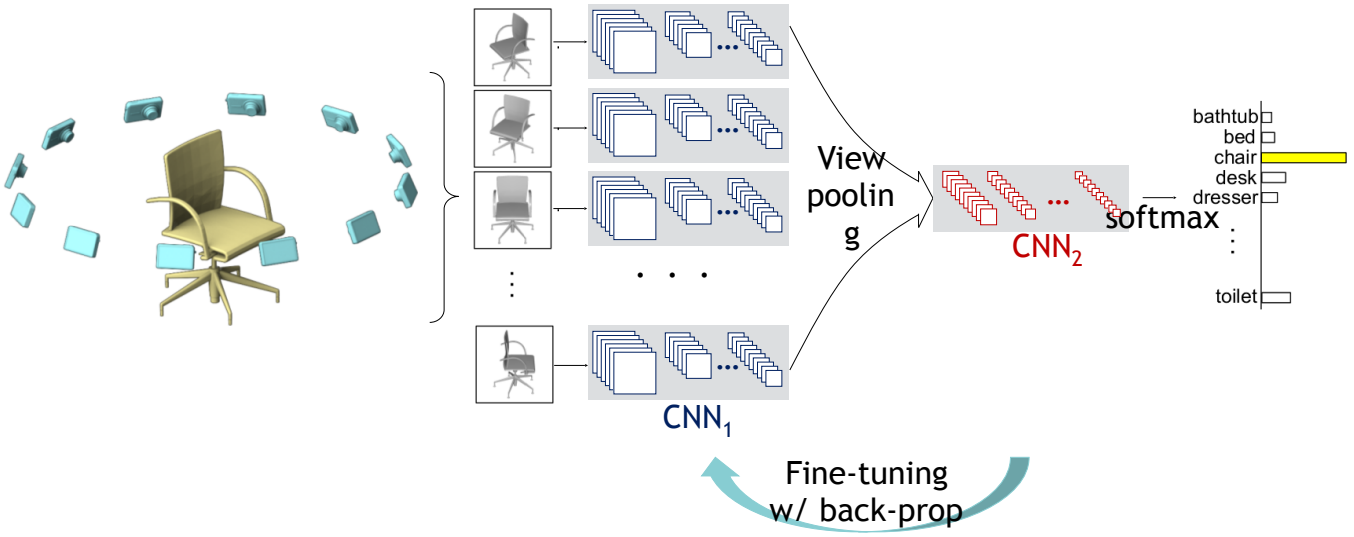


[CHATFIELD14] K. Chatfield et. al., "Return of the Devil in the Details: Delving Deep into Convolutional Nets", BMVC 2014

Hang Su, Subhransu Maji, Evangelos Kalogerakis, Erik Learned-Miller, "**Multi-view Convolutional Neural Networks for 3D Shape Recognition**", *Proceedings of ICCV 2015*

[credit: Hang Su]

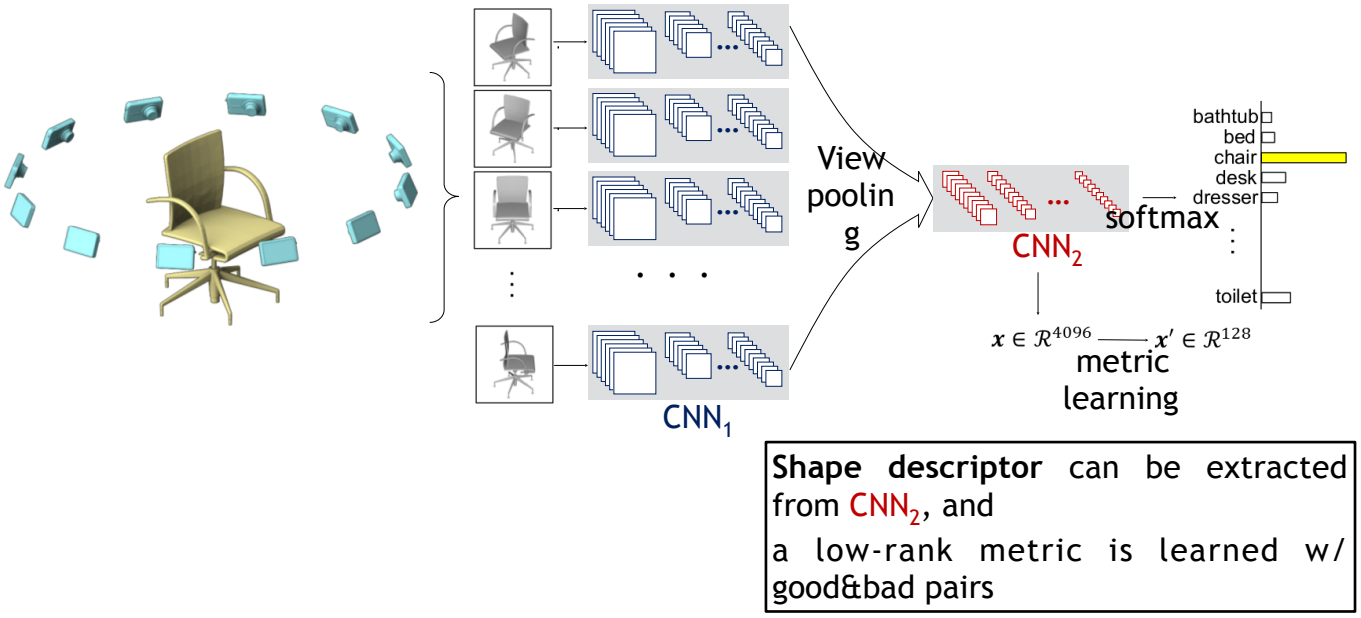
… and then fine-tuned on 3D datasets



Hang Su, Subhransu Maji, Evangelos Kalogerakis, Erik Learned-Miller, "**Multi-view Convolutional Neural Networks for 3D Shape Recognition**", *Proceedings of ICCV 2015*

[credit: Hang Su]

Extract compact shape descriptor for other applications



Shape descriptor can be extracted from CNN_2, and

a low-rank metric is learned w/ good&bad pairs

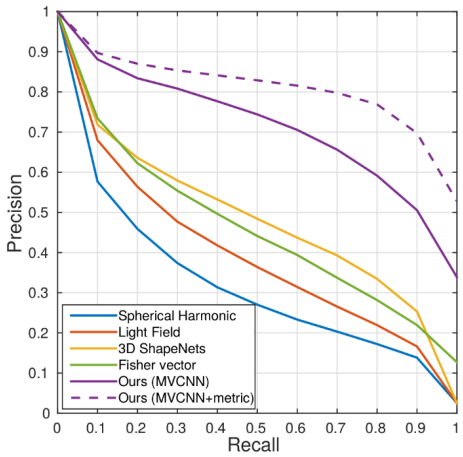
Hang Su, Subhransu Maji, Evangelos Kalogerakis, Erik Learned-Miller, "**Multi-view Convolutional Neural Networks for 3D Shape Recognition**", *Proceedings of ICCV 2015*

[credit: Hang Su]

Experiments – classification & retrieval

On **ModelNet40**, compared against:

- 3 existing methods:
  SPH, LFD, 3D ShapeNets
- 2 strong baselines:
  Fisher vectors, CNN

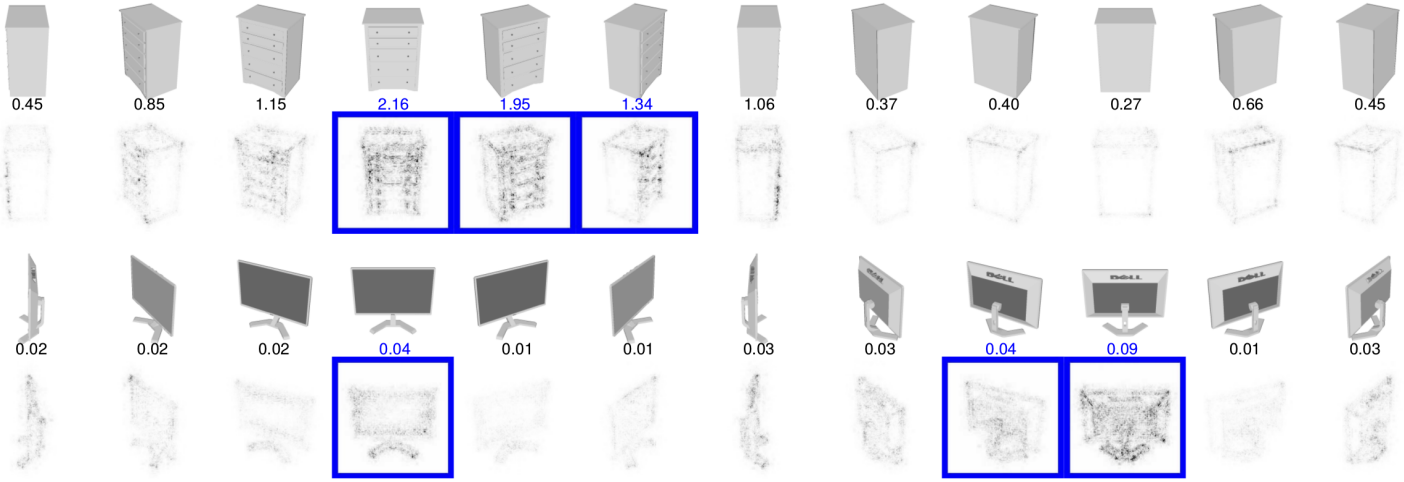


Method	Classification (Accuracy)	Retrieval (mAP)
SPH [16]	68.2%	33.3%
LFD [5]	75.5%	40.9%
3D ShapeNets [37]	77.3%	49.2%
FV, 12 views	84.8%	43.9%
CNN, 12 views	88.6%	62.8%
MVCNN, 12 views	**89.9%**	70.1%
MVCNN+metric, 12 views	89.5%	**80.2%**
MVCNN, 80 views	90.1%	70.4%
MVCNN+metric, 80 views	**90.1%**	**79.5%**

[credit: Hang Su]

Visualization of saliency across views

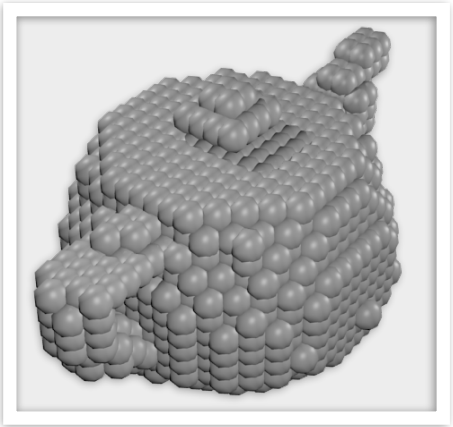
$$[\omega_1, \omega_2 \dots \omega_K] = \left[\left.\frac{\partial F_c}{\partial I_1}\right|_S, \left.\frac{\partial F_c}{\partial I_2}\right|_S \dots \left.\frac{\partial F_c}{\partial I_K}\right|_S \right]$$

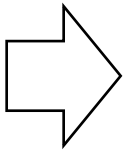


[credit: Hang Su]

How do you use multi-view approach for point cloud?

Sphere Rendering *Images*





Multi-View Image CNN

[credit: CVPR 2016 sp

Practical multi-view CNN

State-of-the-art performance for **3D mesh classification**

Issues:

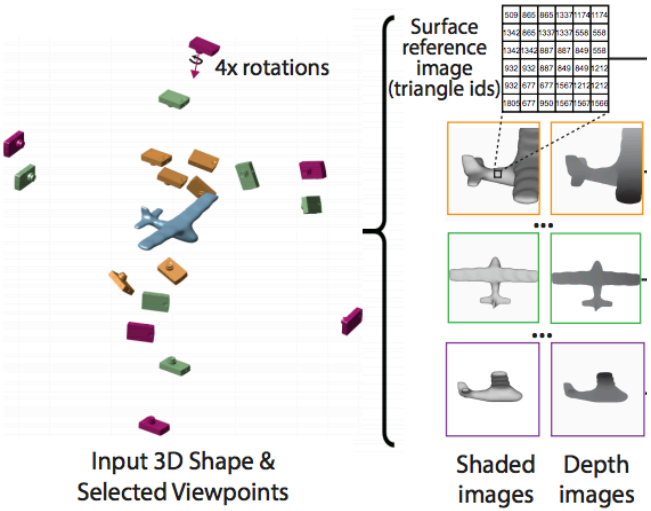
- What viewpoints to select? In particular, where shall we place the camera in a scene?
- What if the input is noisy and incomplete? e.g., point cloud

Agenda

- **Deep Learning on Multi-view Representation**
  - Classification
  - **Segmentation**
  - Reconstruction

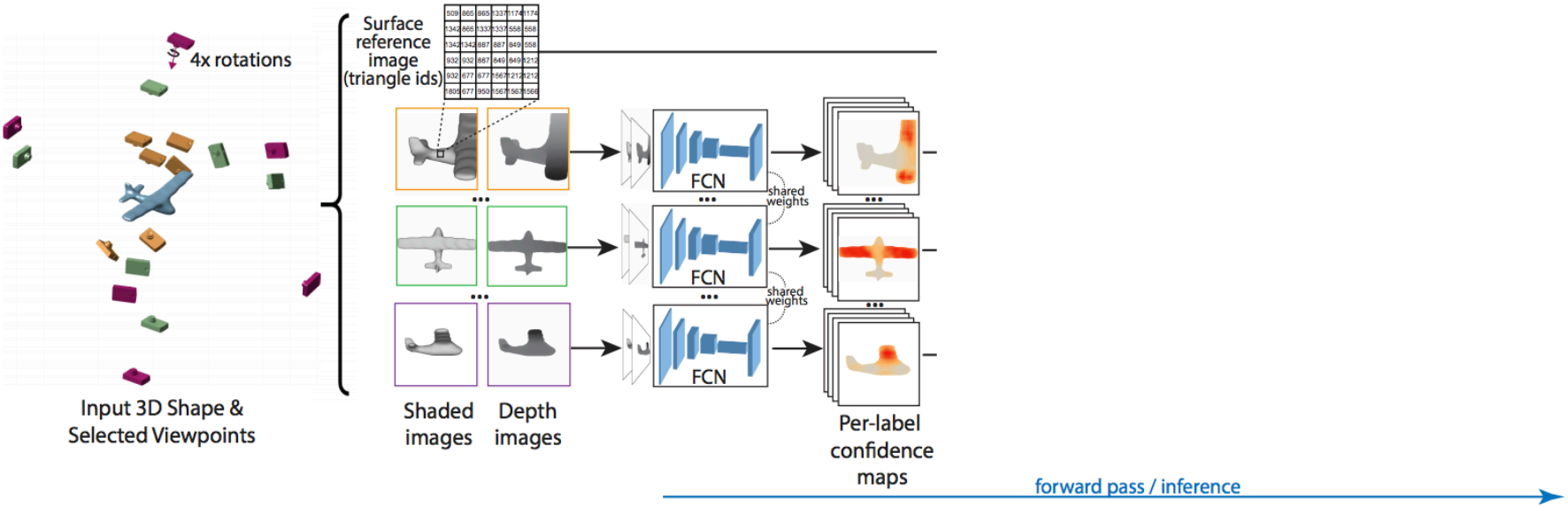
3D segmentation

Basic architecture



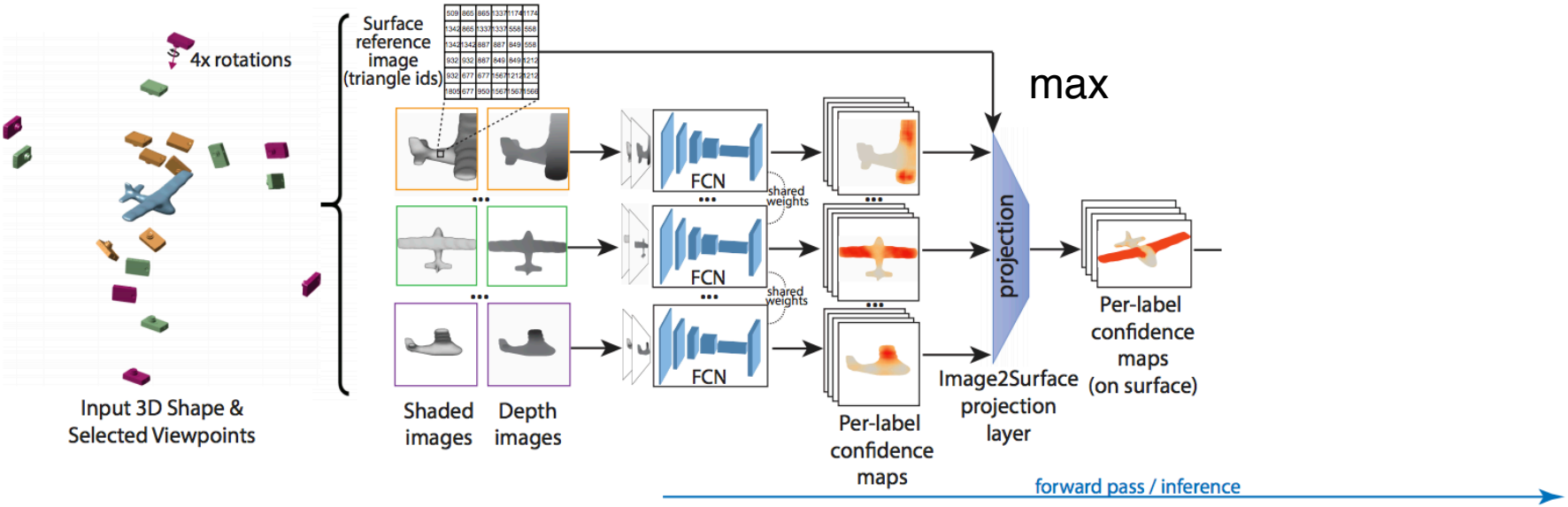
Evangelos Kalogerakis, Melinos Averkiou, Subhransu Maji, Siddhartha Chaudhuri,
"3D Shape Segmentation with Projective Convolutional Networks",
CVPR2017

Basic architecture



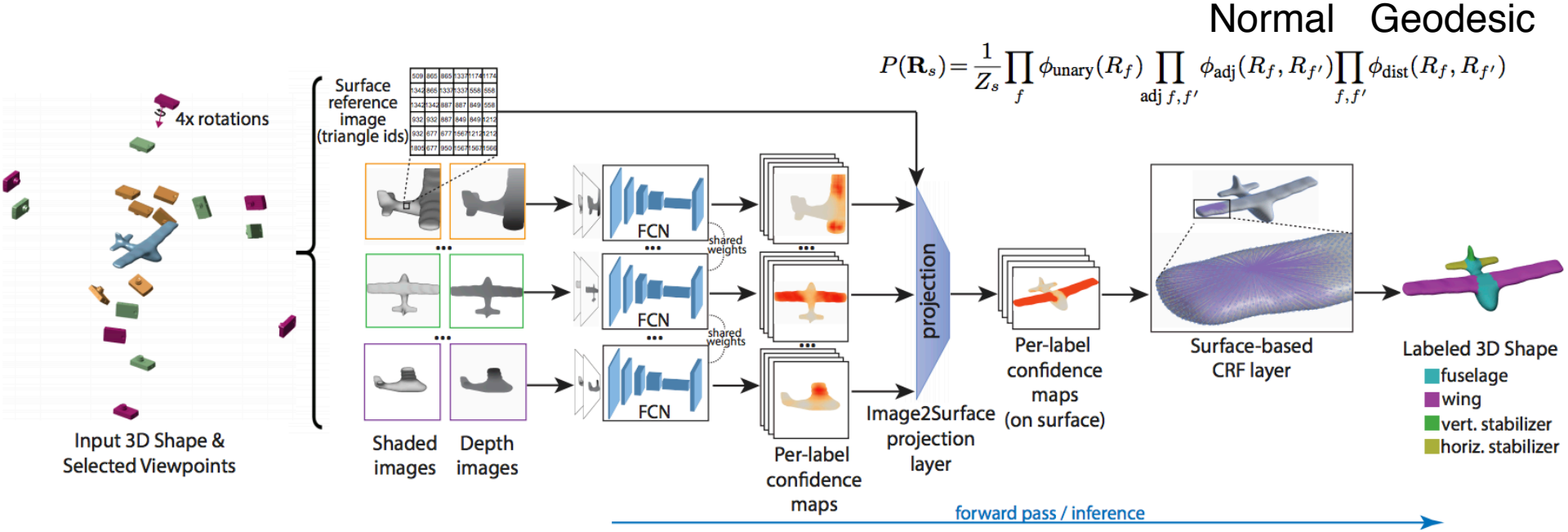
Evangelos Kalogerakis, Melinos Averkiou, Subhransu Maji, Siddhartha Chaudhuri,
"3D Shape Segmentation with Projective Convolutional Networks",
CVPR2017

Basic architecture



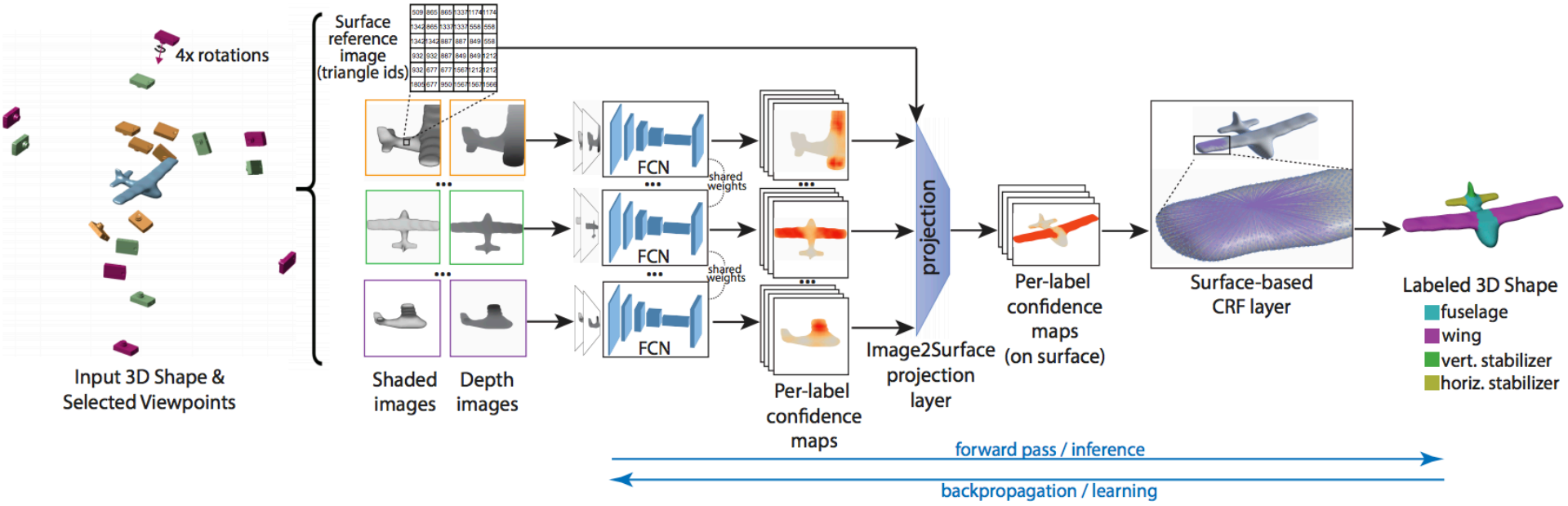
Evangelos Kalogerakis, Melinos Averkiou, Subhransu Maji, Siddhartha Chaudhuri,
"3D Shape Segmentation with Projective Convolutional Networks",
CVPR2017

Basic architecture



Evangelos Kalogerakis, Melinos Averkiou, Subhransu Maji, Siddhartha Chaudhuri,
"3D Shape Segmentation with Projective Convolutional Networks",
CVPR2017

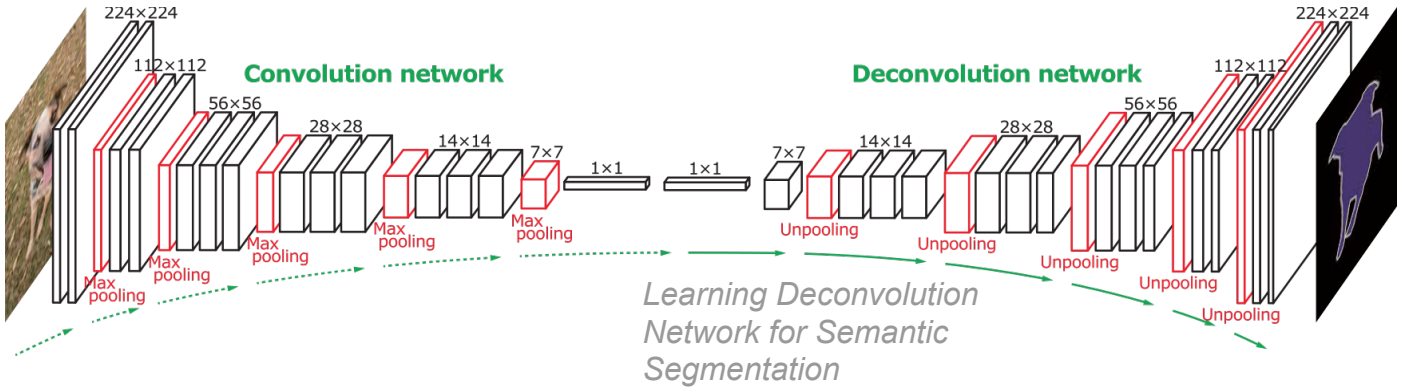
Basic architecture

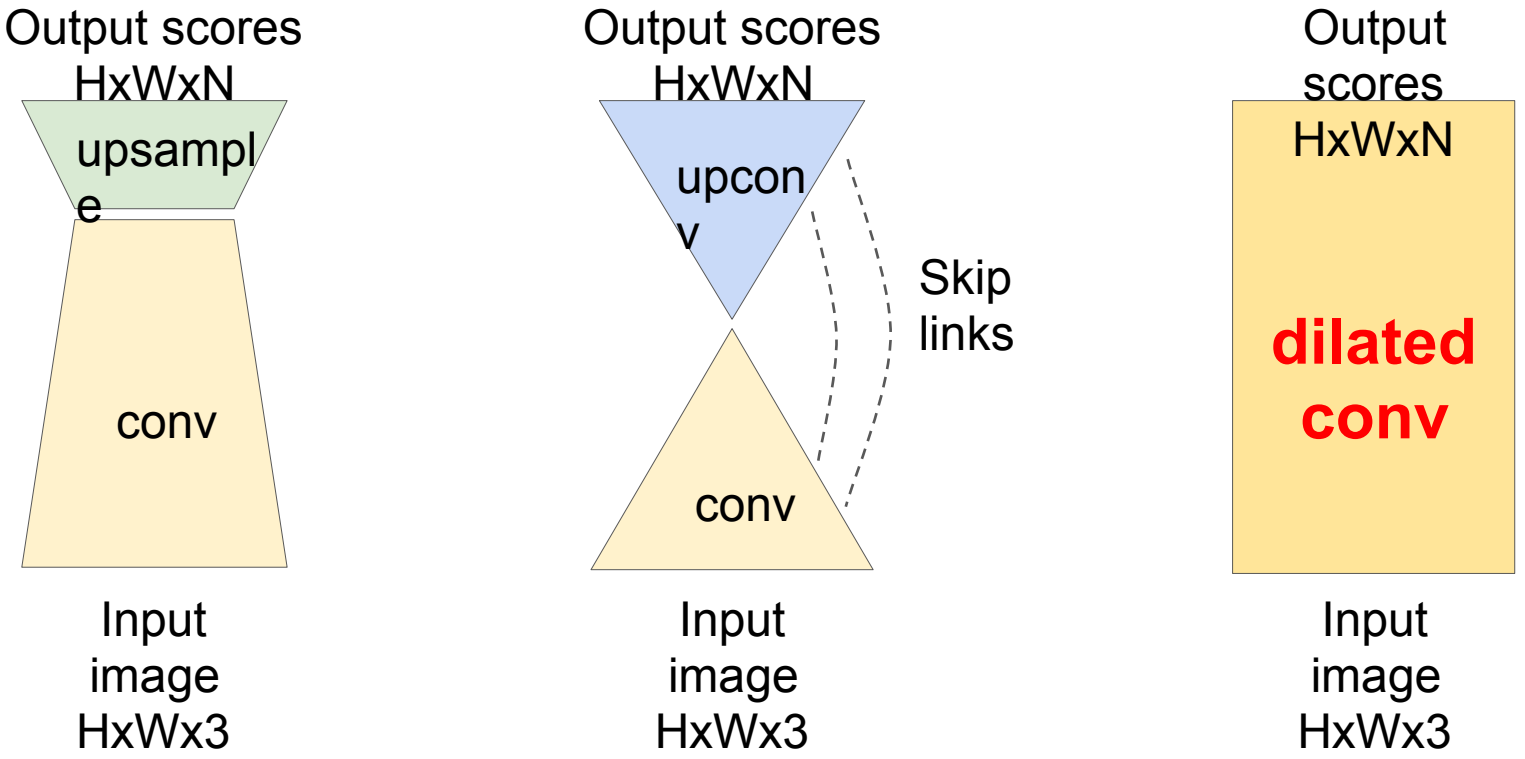


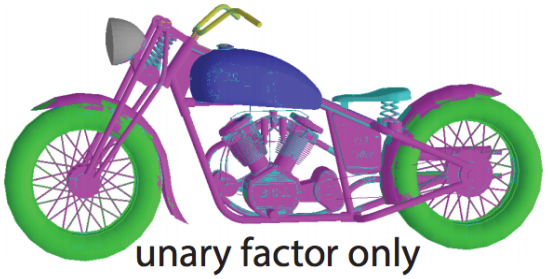
Evangelos Kalogerakis, Melinos Averkiou, Subhransu Maji, Siddhartha Chaudhuri,
"3D Shape Segmentation with Projective Convolutional Networks",
CVPR2017

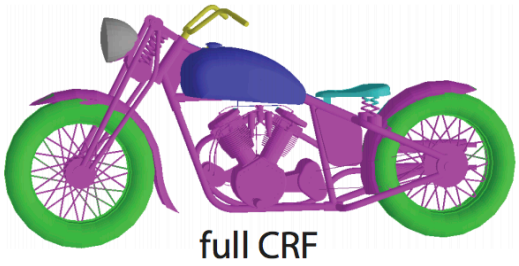
Fully Convolutional Network (FCN)

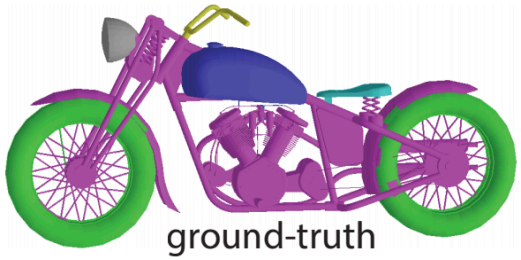
Segmentation:




Output scores HxWxN
upsampl e
conv
Input image HxWx3
Output scores HxWxN
upcon v
Skip links
conv
Input image HxWx3
Output scores HxWxN
dilated conv
Input image HxWx3


unary factor only


full CRF


ground-truth

Performance (cont.)

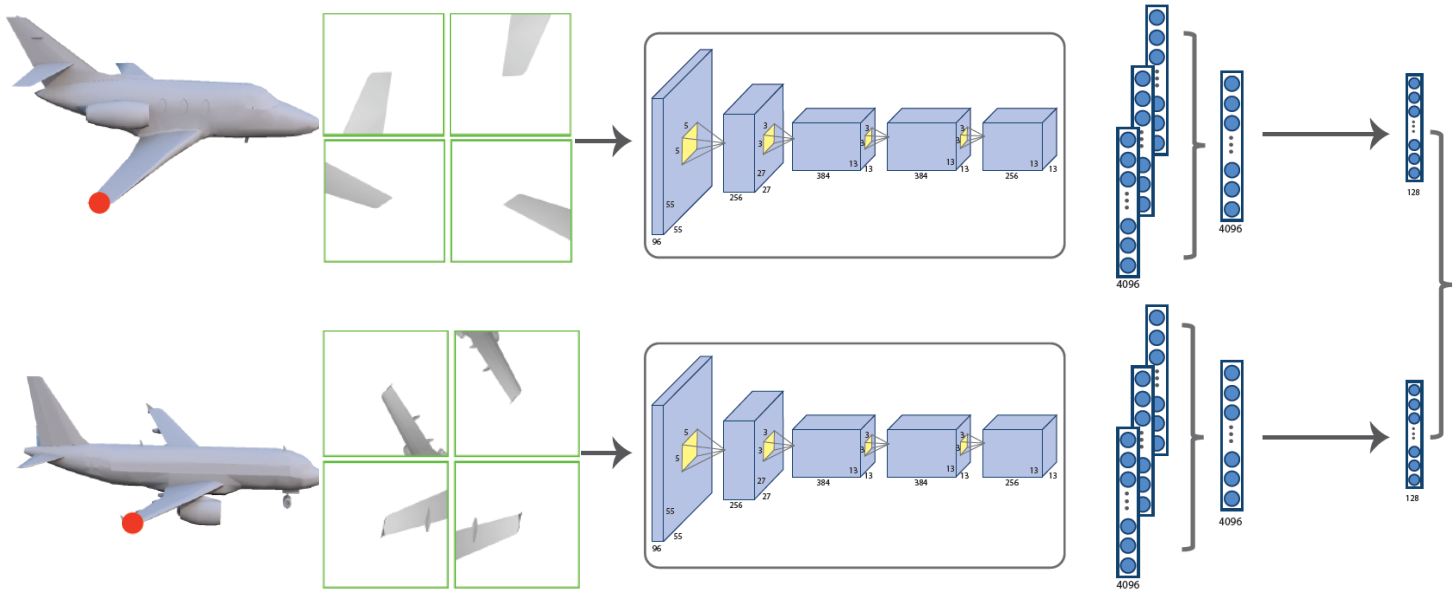
- Viewpoint selection to maximize surface coverage
- Combination of view-based network with surface-based graphical model
- ~88% labeling accuracy on ShapeNet
  (trained per category, 50%-50% split, max 250 shapes for training)

Challenges:

- View-based network does not process invisible points
- View-based representations have redundancy
- Slow to train (~week for a few hundreds of shapes)
- Aggregating view representations via max-pooling may lose information

Surface correspondences with multi-view convnets

Aggregates point-based descriptors across local views. Trained such that similar points have similar descriptors based on synthetically generated correspondences.



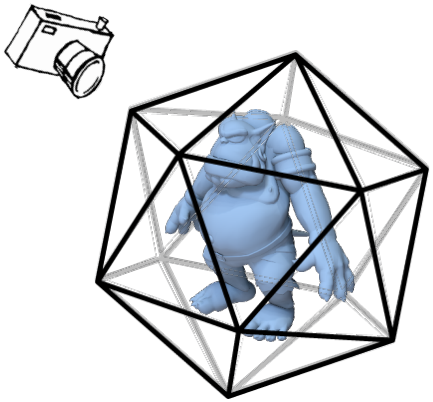
Haibin Huang, Evangelos Kalogerakis, Siddhartha Chaudhuri, Duygu Ceylan, Vladimir Kim, Ersin Yumer
Learning Local Shape Descriptors with View-Based Convolutional Neural Network, ACM TOG (to appear)

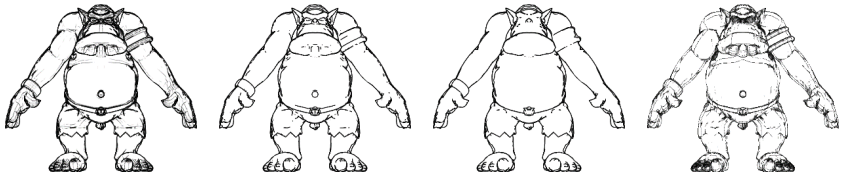
Scan-to-shape matching

shows some robustness to noise, better performance than volumetric net (3DMatch)

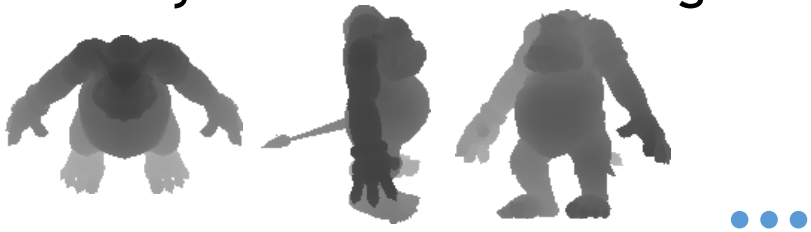
(similar colors correspond to points with similar descriptors)

Training data



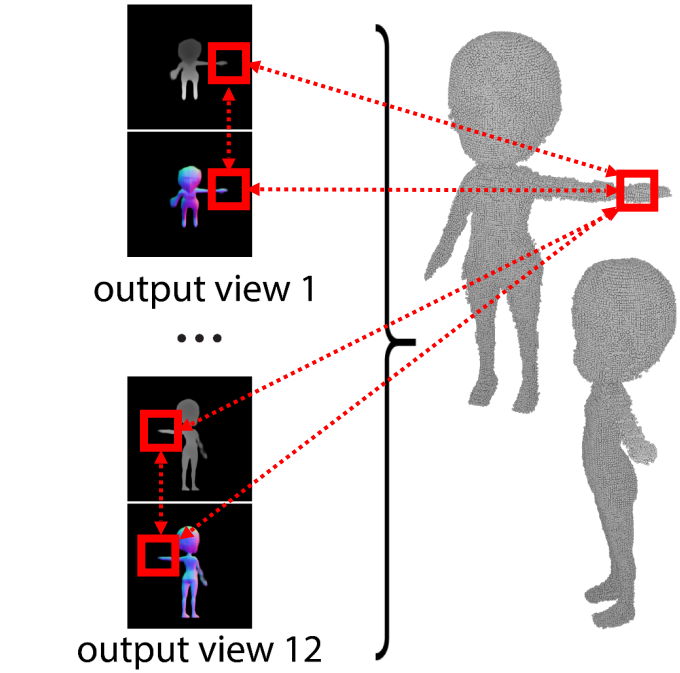


Synthetic line drawings



Training depth and normal maps

Consolidate multi-view depth and normal maps



multi-view depth & normal maps

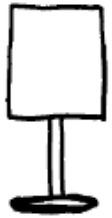
optimized 3D point cloud

Optimization for fusion

- Depth derivatives should be consistent with normals
- Corresponding depths and normals across different views should agree

reference shape

Line drawings



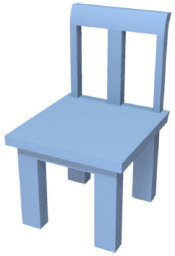
ShapeMVD

Tatarchenko et al. (same loss/fusion)

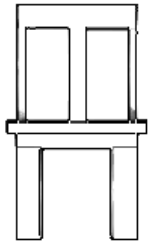
volumetric

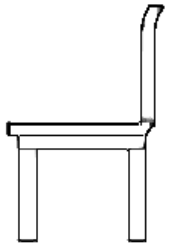
nearest retrieval

reference shape



Line drawings





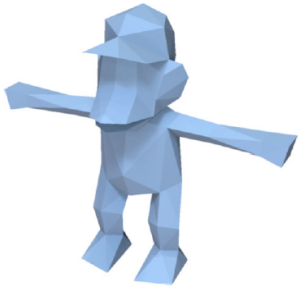
ShapeMVD

Tatarchenko et al. (same loss/fusion)

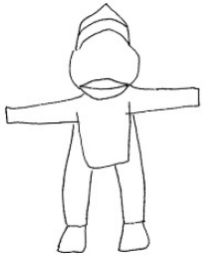
volumetric

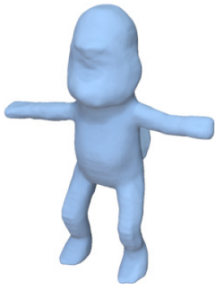
nearest retrieval

reference shape



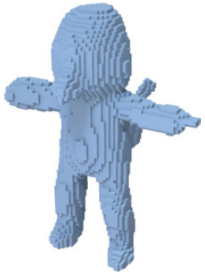
Line drawings



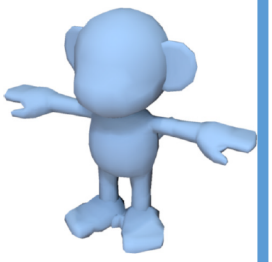


ShapeMVD

Tatarchenko et al. (same loss/fusion)



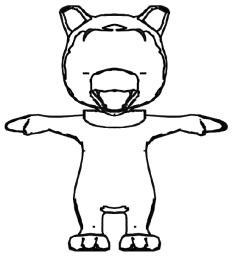
volumetric

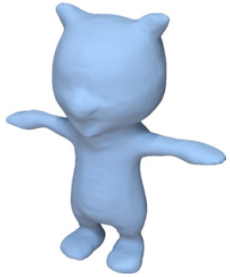


nearest retrieval

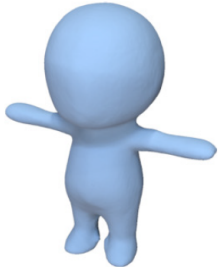
reference shape

Line drawings

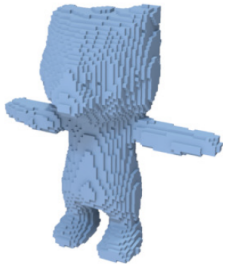




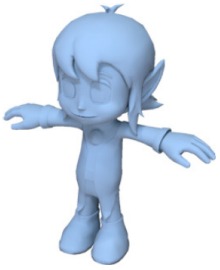
ShapeMVD



Tatarchenko et al. (same loss/fusion)

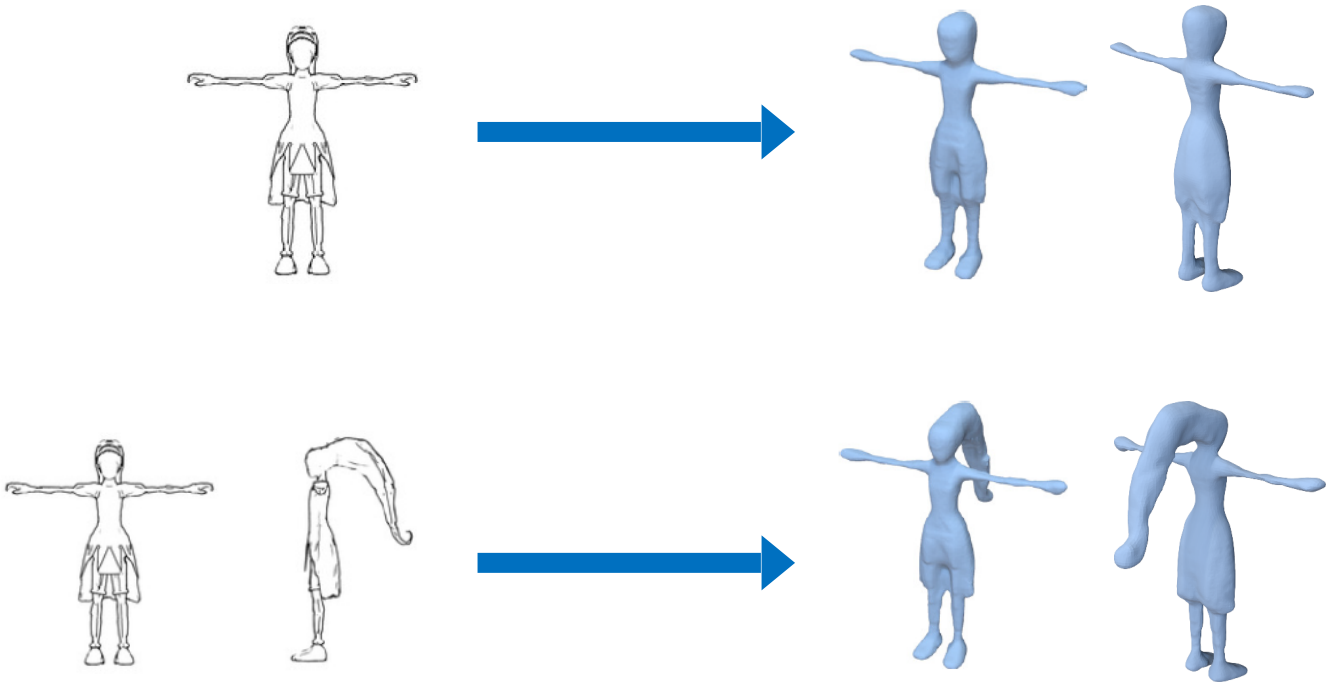


volumetric



nearest retrieval

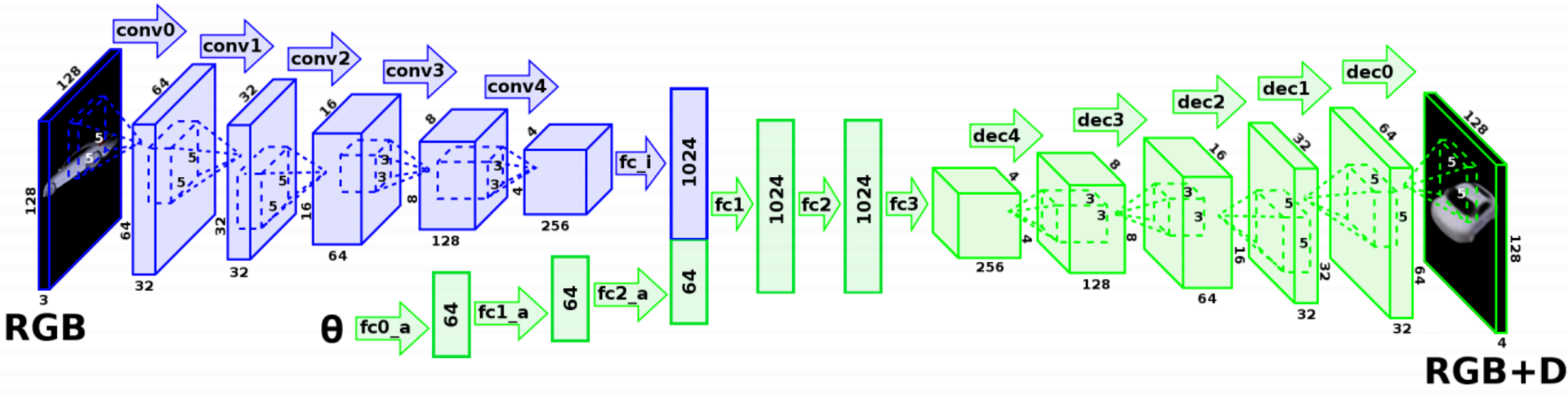
Single vs two input line drawings



Agenda

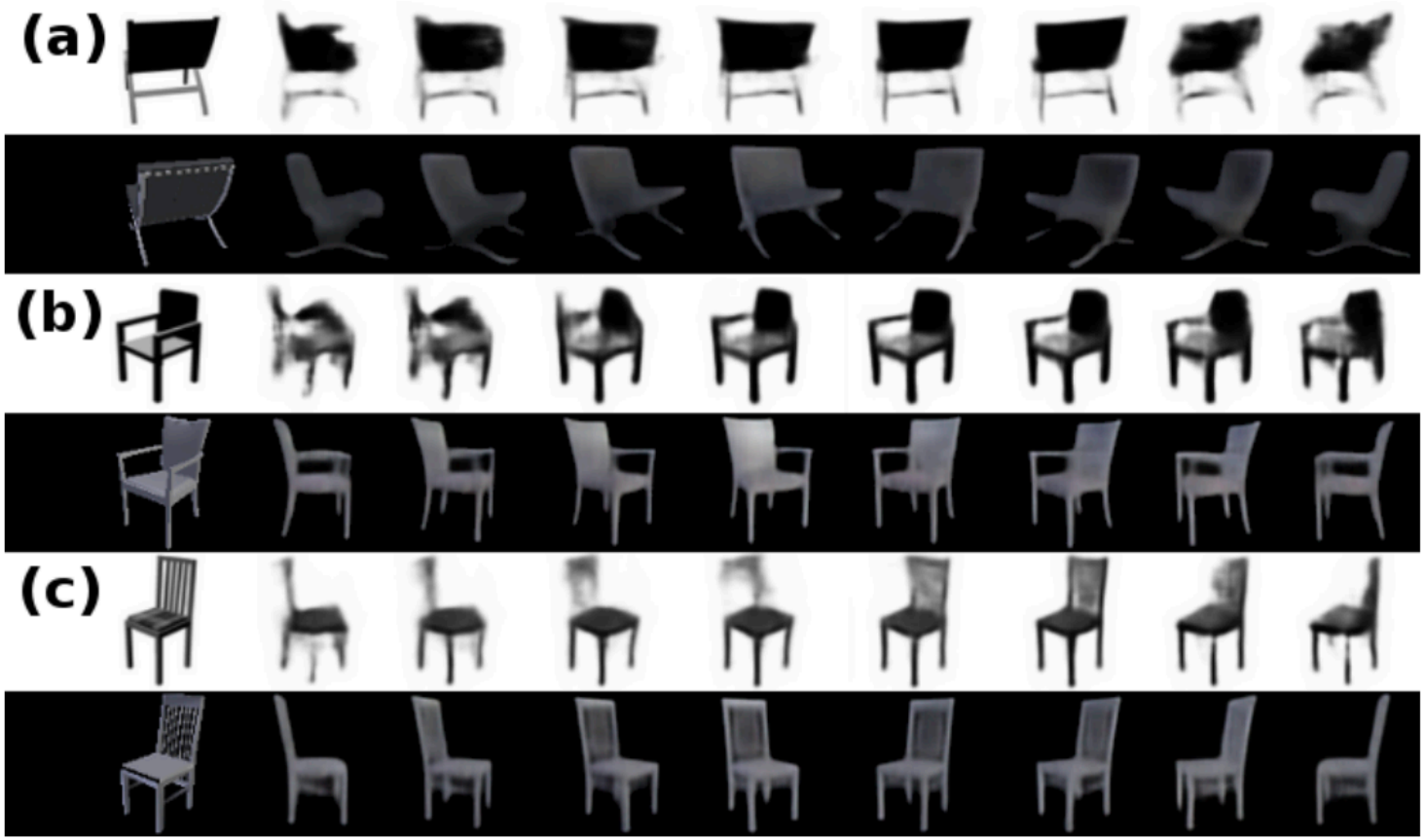
- **Deep Learning on Multi-view Representation**
  - Classification
  - Segmentation
  - **Novel-view synthesis**

Idea 1: Direct novel-view synthesis



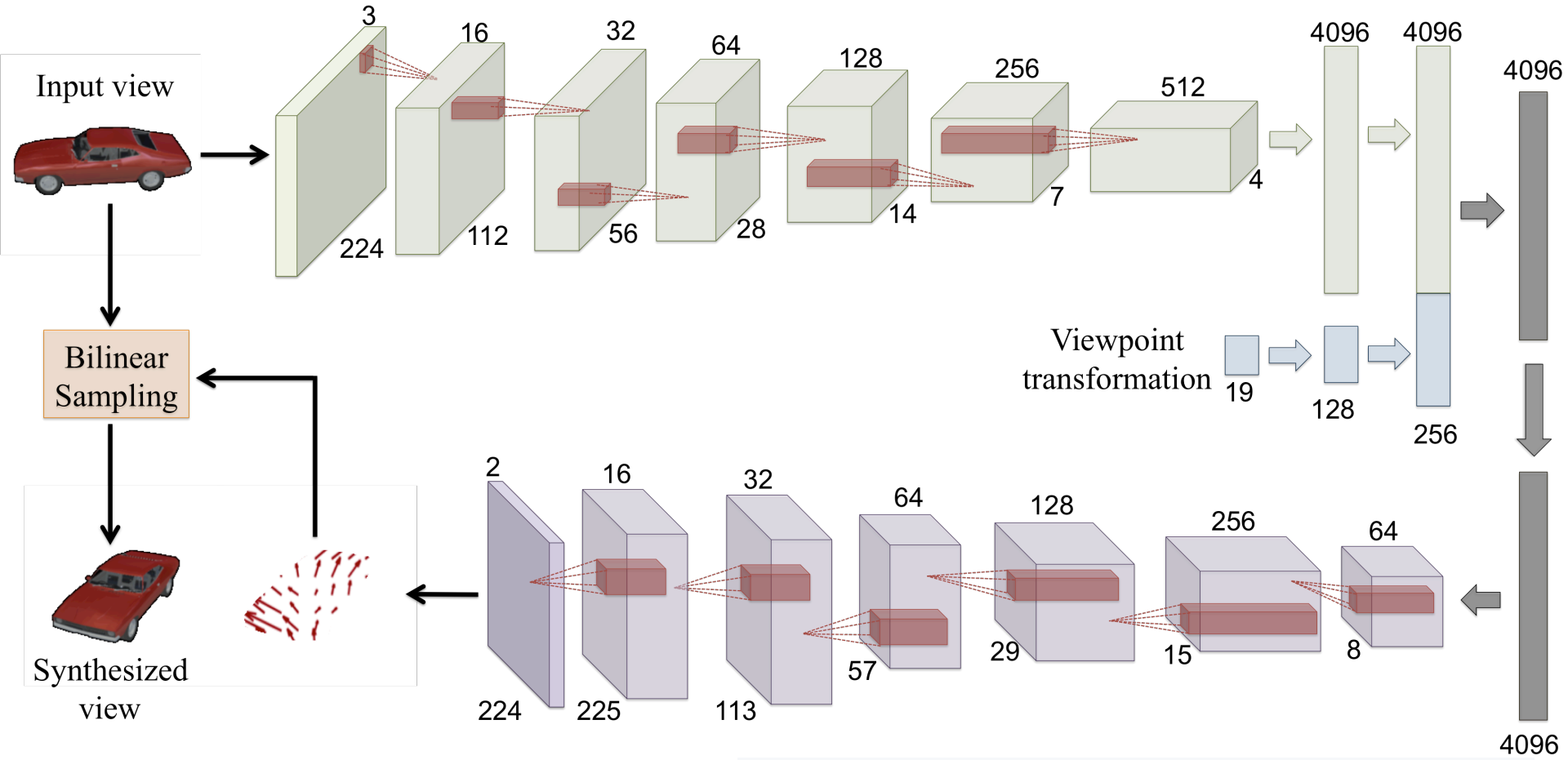
Maxim Tatarchenko, Alexey Dosovitskiy, Thomas Brox,
"Multi-view 3D Models from Single Images with a Convolutional Network",
ECCV2016

Results



Idea 2: Flow synthesis

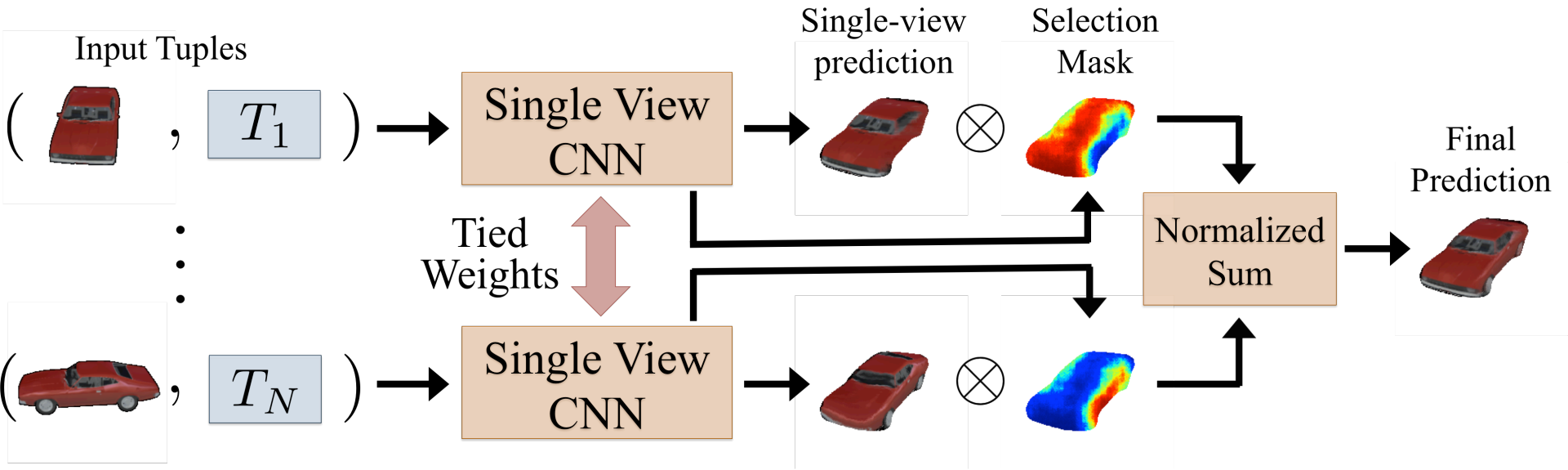
Single-view network architecture:



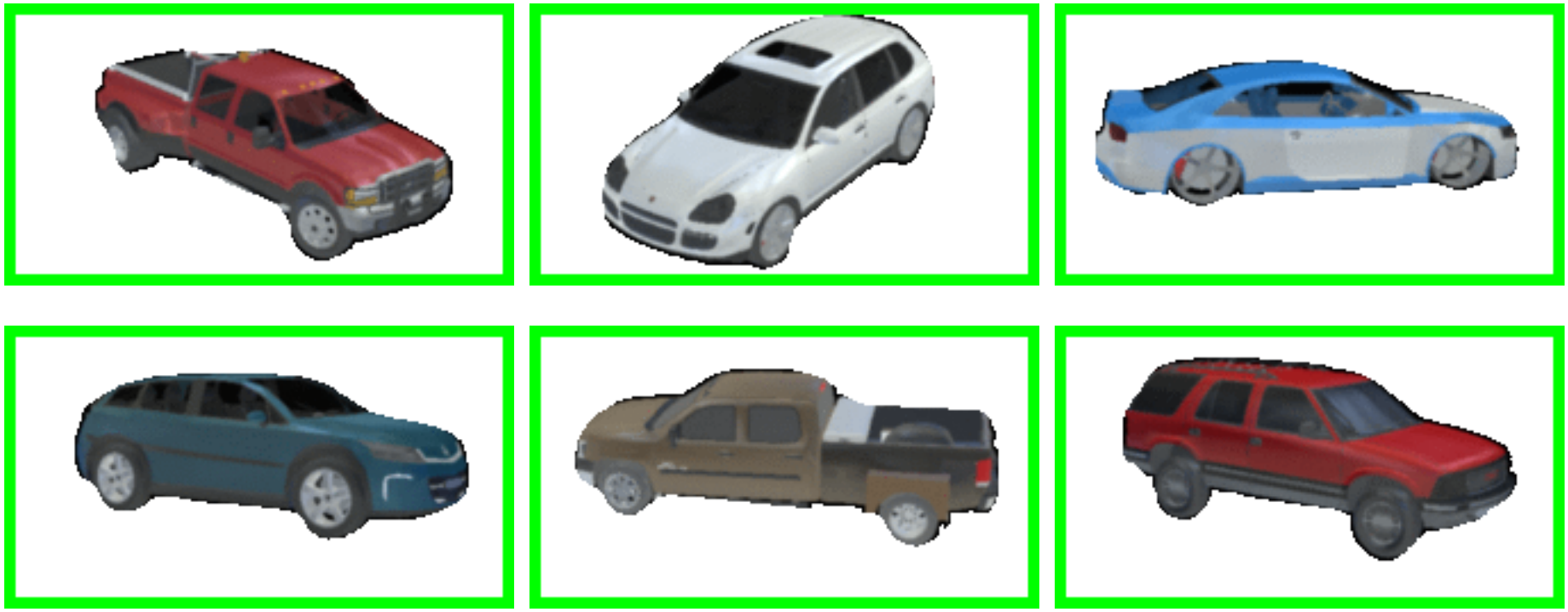
Zhou et al, View Synthesis by Appearance Flow, ECCV 2016

Idea 2: Flow synthesis

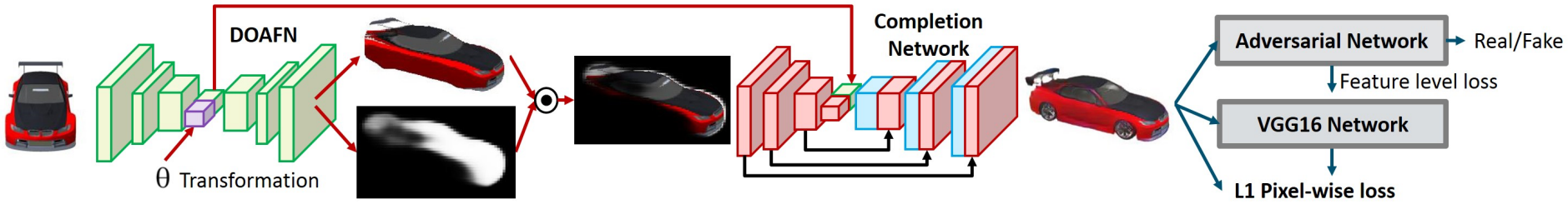
Multi-view network architecture:



Idea 2: Flow synthesis



Combine both ideas



- First, apply flow prediction
- Second, conduct invisible part hallucination

Park et al, Transformation-Grounded Image Generation Network for Novel 3D View Synthesis, CVPR 2017

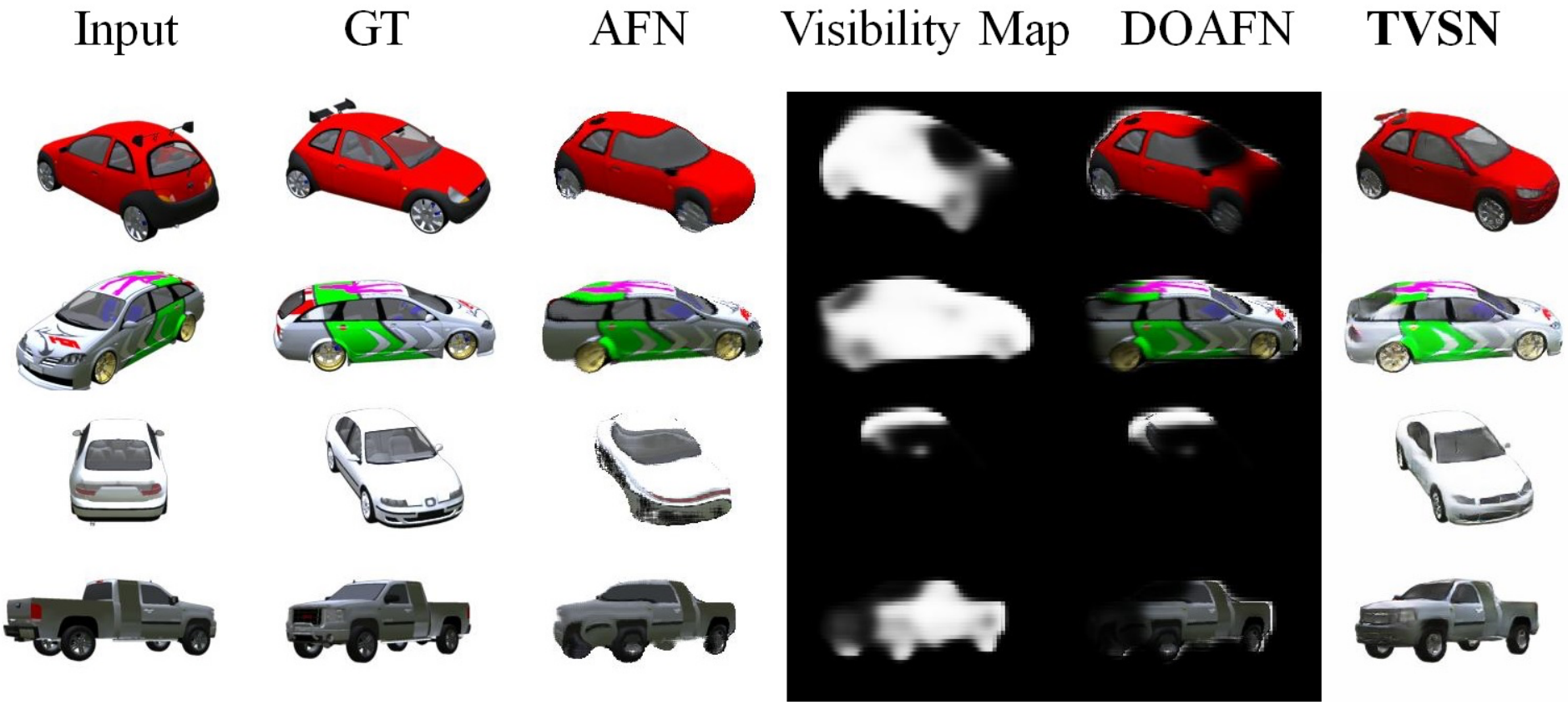
Combine both ideas



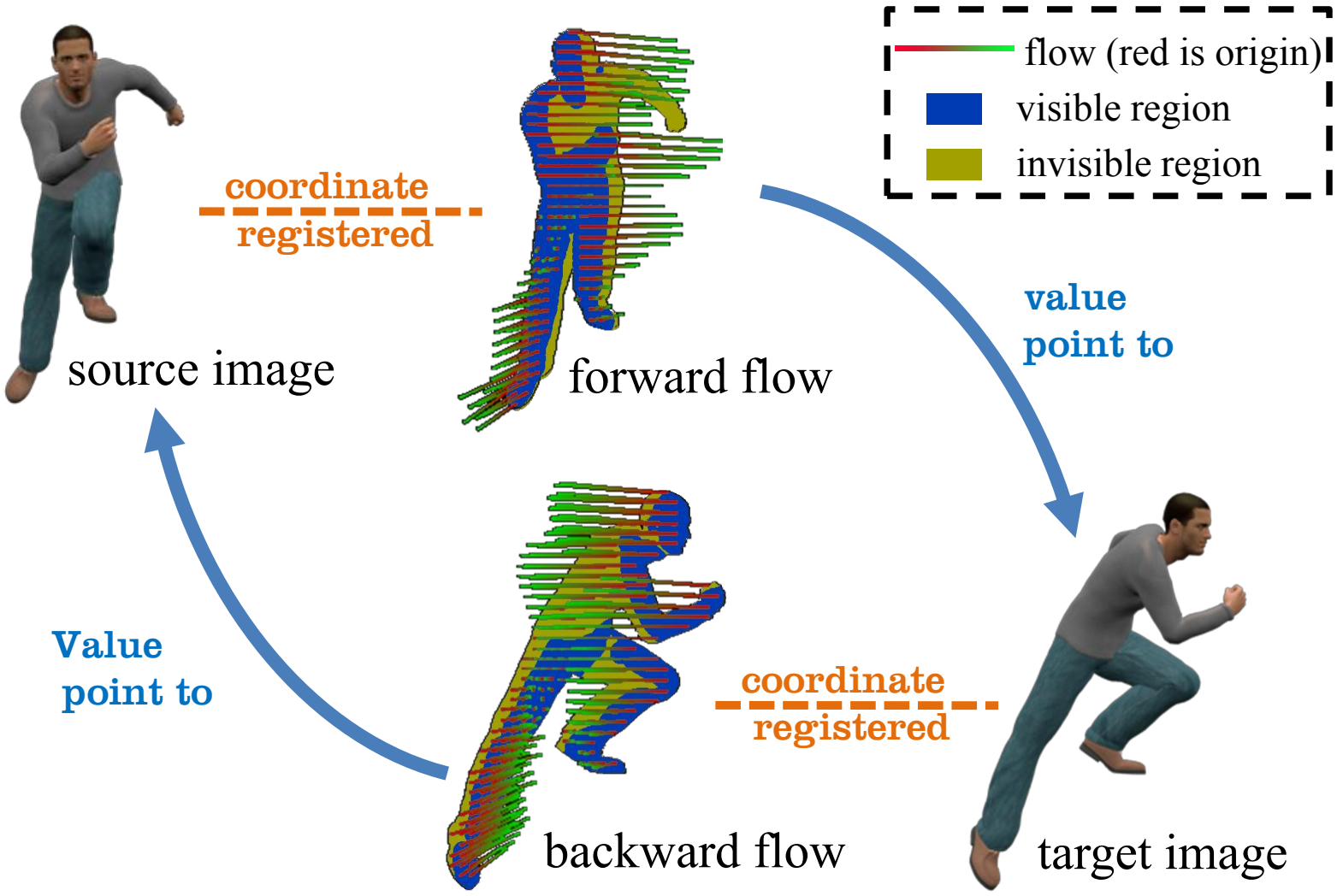





Combine both ideas

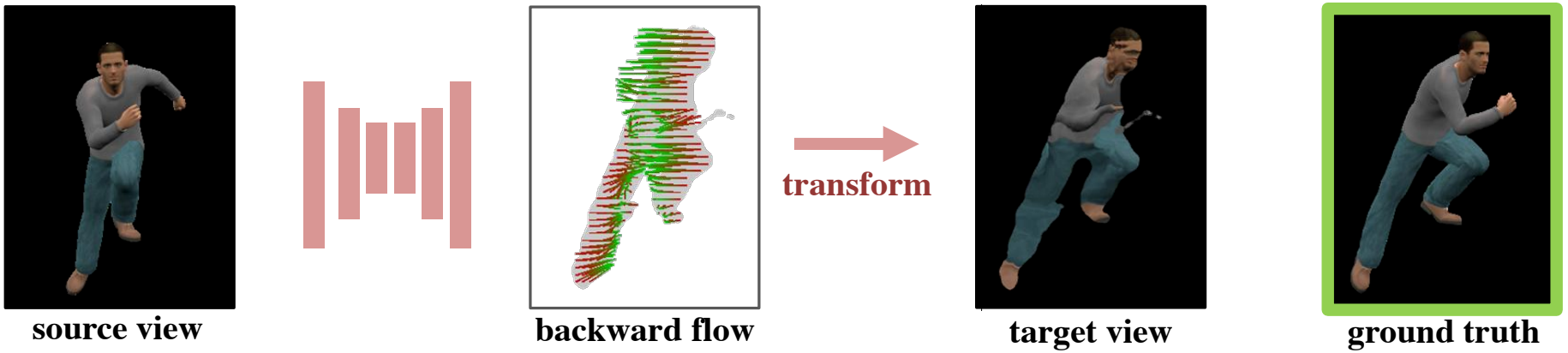


Further improvement: Assist flow synthesis by depth estimation



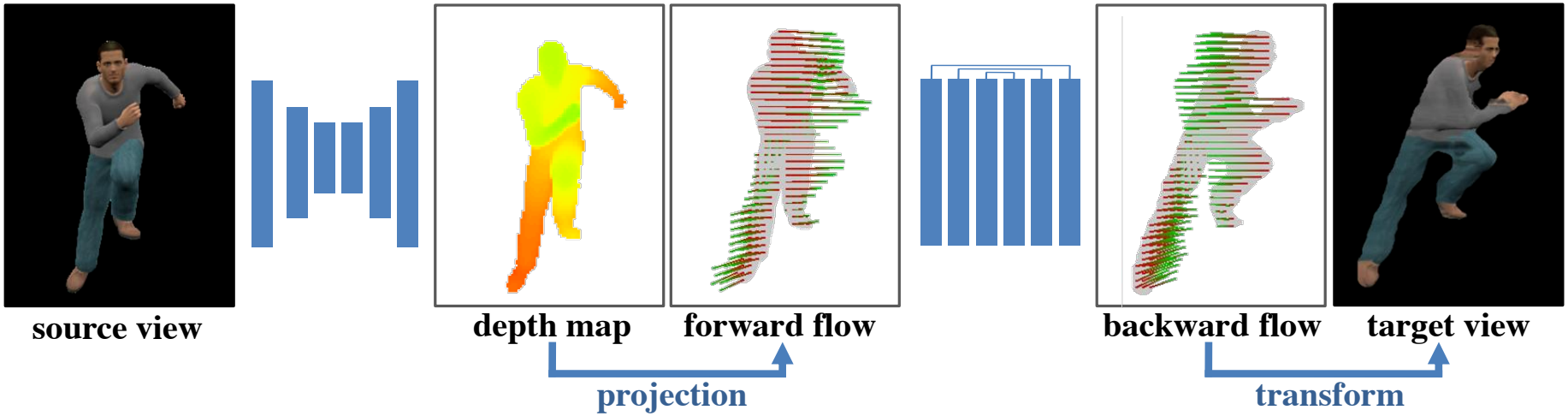
My submission to CVPR'18

Further improvement: Assist flow synthesis by depth estimation



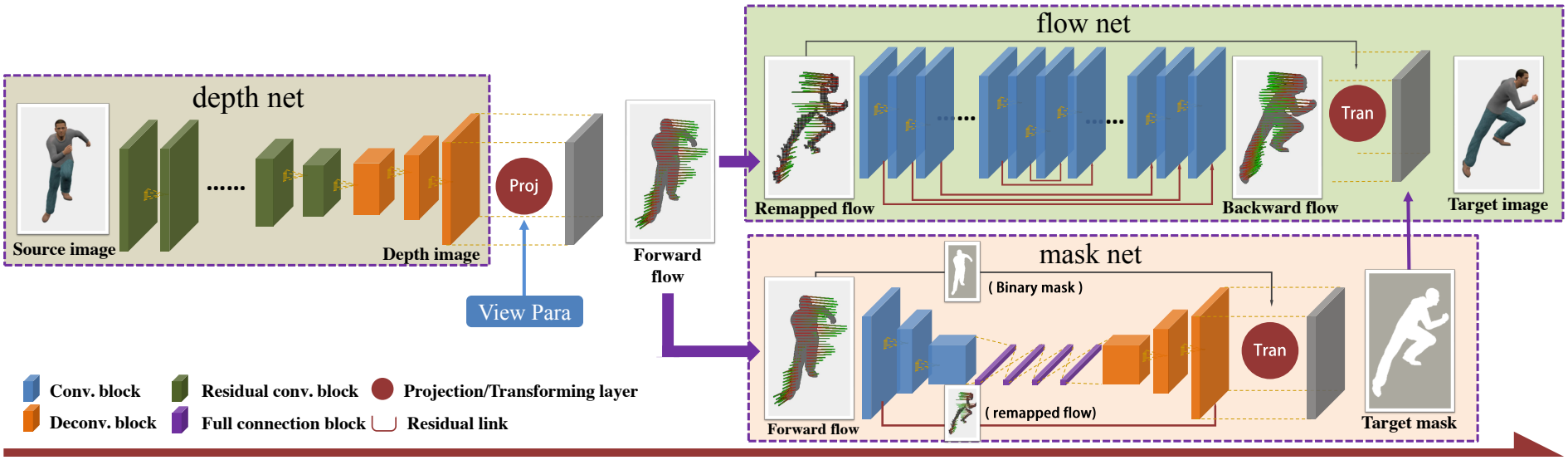
prior art

our work



My submission to CVPR'18

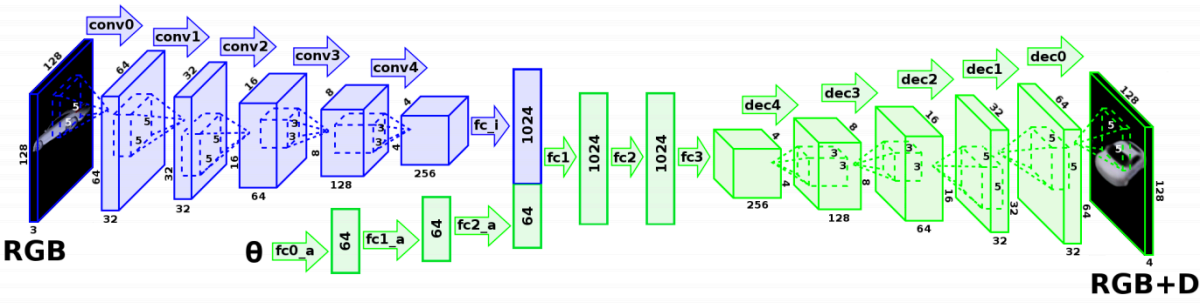
Further improvement: Assist flow synthesis by depth estimation

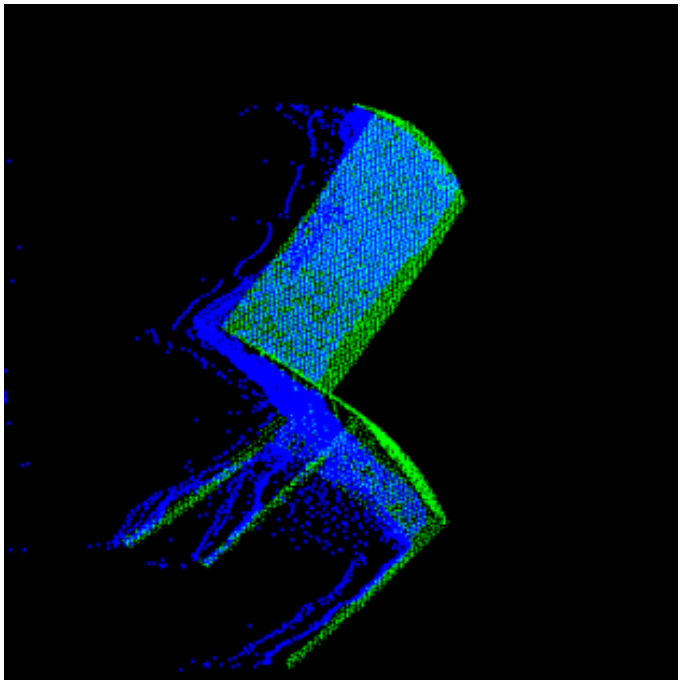


My submission to CVPR'18

Key challenges

- Each view only contains partial information





- Regular structures in 3D cannot be well captured
  - e.g., symmetry, straightness, roundish