EdgeNet: Semantic Scene Completion from a Single RGB-D Image

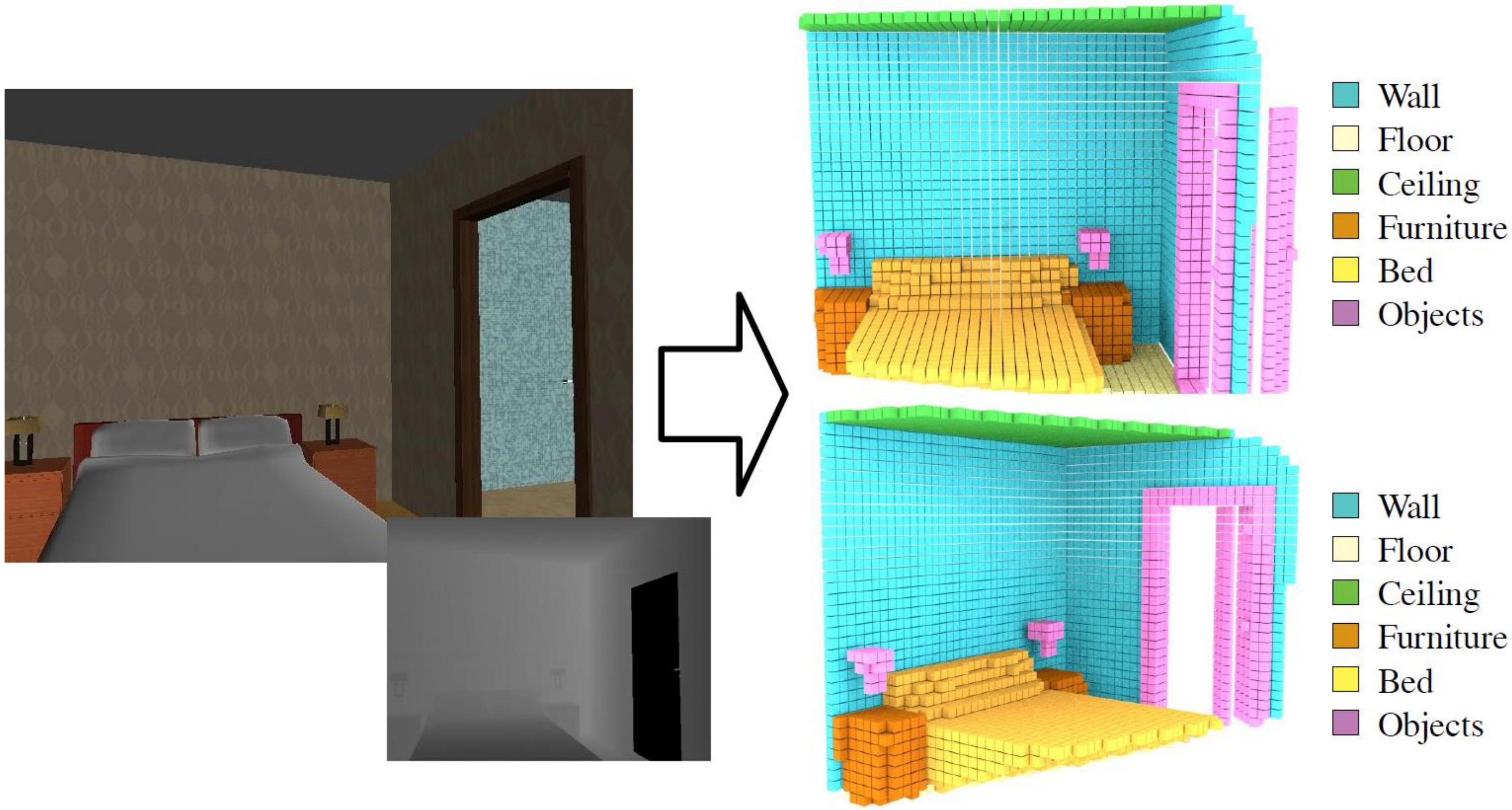
Aloisio Dourado, Teofilo Emidio de Campos
University of Brasilia
Brasilia, Brazil
aloisio.dourado.bh@gmail.com, t.decampos@st-annes.oxon.org

Hansung Kim
University of Southampton
Southampton, UK
H.Kim@soton.ac.uk

Adrian Hilton
University of Surrey
Surrey, UK
a.hilton@surrey.ac.uk

ICPR 2020
Milan, January 2021

Semantic Scene Completion



Introduced by Song *et al*.[1] in 2017

Trained a 3D CNN that jointly deals with completion and semantic segmentation

[1] Song, S., Yu, F., Zeng, A., Chang, A.X., Savva, M., and Funkhouser, T.: Semantic Scene Completion from a Single Depth Image. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, Hawaii, July 21-26, pp. 190–198, Piscataway, NJ, July 2017. IEEE.

Previous Works

Depth maps only

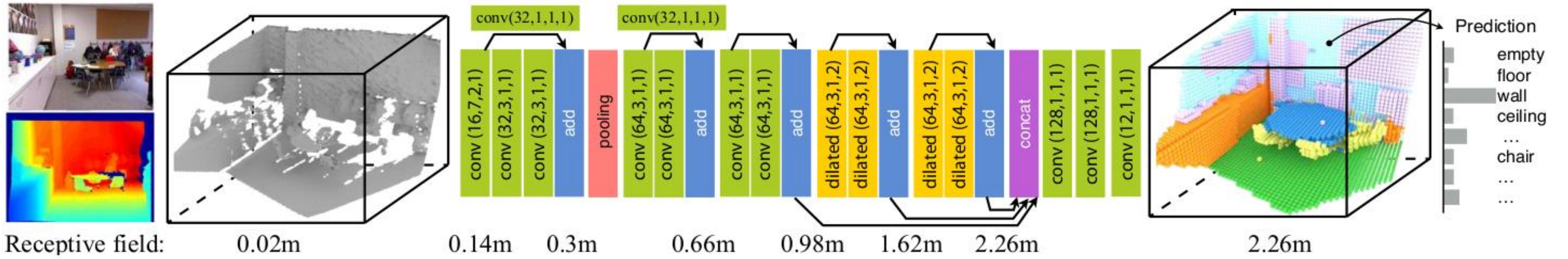
- SSCNET: Song et al. [1]

[1] Song, S., Yu, F., Zeng, A., Chang, A.X., Savva, M., and Funkhouser, T.: Semantic Scene Completion from a Single Depth Image. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, Hawaii, July 21-26, pp. 190–198, Piscataway, NJ, July 2017. IEEE.

Previous Works

Depth maps only

- SSCNET: Song et al. [1]
    - Encoder-decoder network architecture

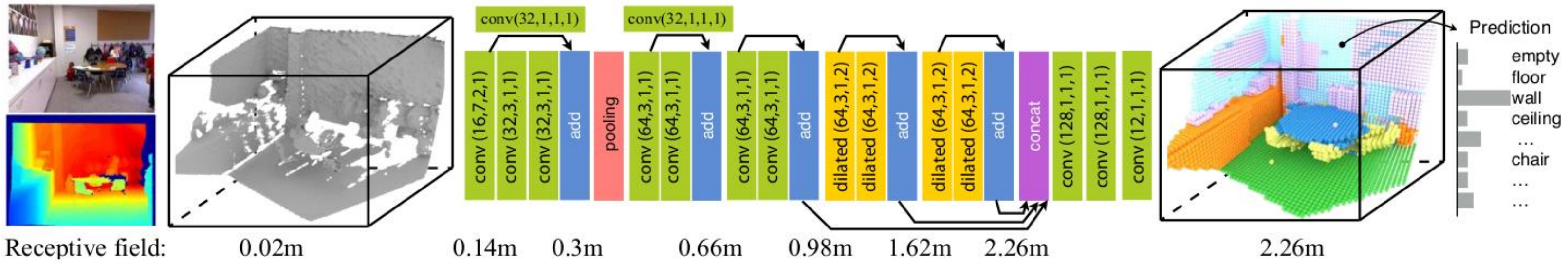


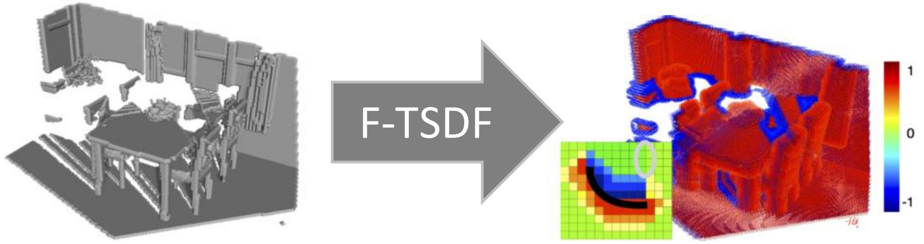
[1] Song, S., Yu, F., Zeng, A., Chang, A.X., Savva, M., and Funkhouser, T.: Semantic Scene Completion from a Single Depth Image. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, Hawaii, July 21-26, pp. 190–198, Piscataway, NJ, July 2017. IEEE.

Previous Works

Depth maps only

- SSCNET: Song et al. [1]
  - Encoder-decoder network architecture
  - Proposed F-TSDF encoding



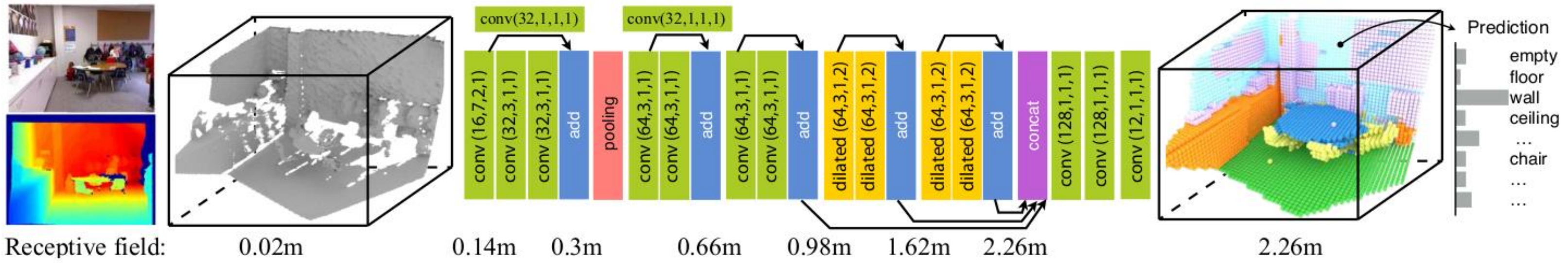


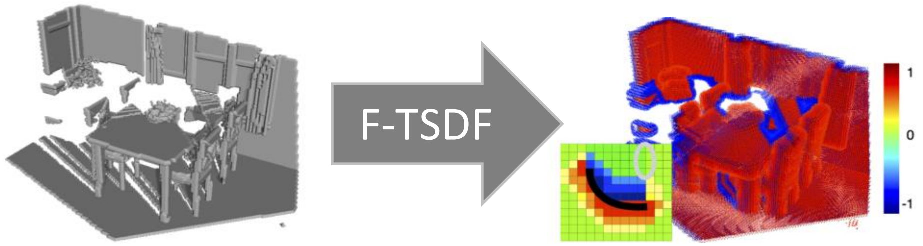
[1] Song, S., Yu, F., Zeng, A., Chang, A.X., Savva, M., and Funkhouser, T.: Semantic Scene Completion from a Single Depth Image. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, Hawaii, July 21-26, pp. 190–198, Piscataway, NJ, July 2017. IEEE.

Previous Works

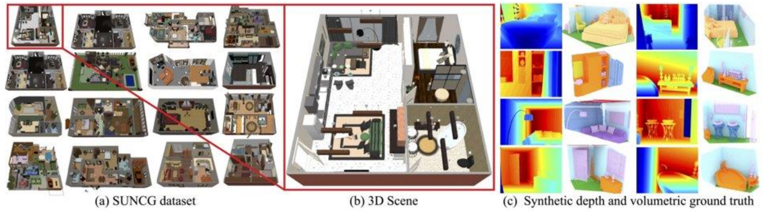
Depth maps only

- SSCNET: Song et al. [1]
  - Encoder-decoder network architecture
  - Proposed F-TSDF encoding
  - Introduced SUNCG Dataset




F-TSDF

SUNCG



[1] Song, S., Yu, F., Zeng, A., Chang, A.X., Savva, M., and Funkhouser, T.: Semantic Scene Completion from a Single Depth Image. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, Hawaii, July 21-26, pp. 190–198, Piscataway, NJ, July 2017. IEEE.

Previous Works

Depth maps only

- SSCNET: Song et al. [1]
- Guo and Tong [2]:
  - 2D features projected to 3D

[1] Song, S., Yu, F., Zeng, A., Chang, A.X., Savva, M., and Funkhouser, T.: Semantic Scene Completion from a Single Depth Image. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, Hawaii, July 21-26, pp. 190–198, Piscataway, NJ, July 2017. IEEE.

[2] Guo, Y. and Tong, X.: View-Volume Network for Semantic Scene Completion from a Single Depth Image. In Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, pp. 726–732, Stockholm, Sweden, July 2018. International Joint Conferences on Artificial Intelligence Organization, ISBN 978-0-9992411-2-7. https://doi.org/10.24963/ijcai.2018/101.

Previous Works

Depth maps only

- SSCNET: Song et al. [1]
- Guo and Tong [2]:
  - 2D features projected to 3D

Neglect the RGB channels from the input data

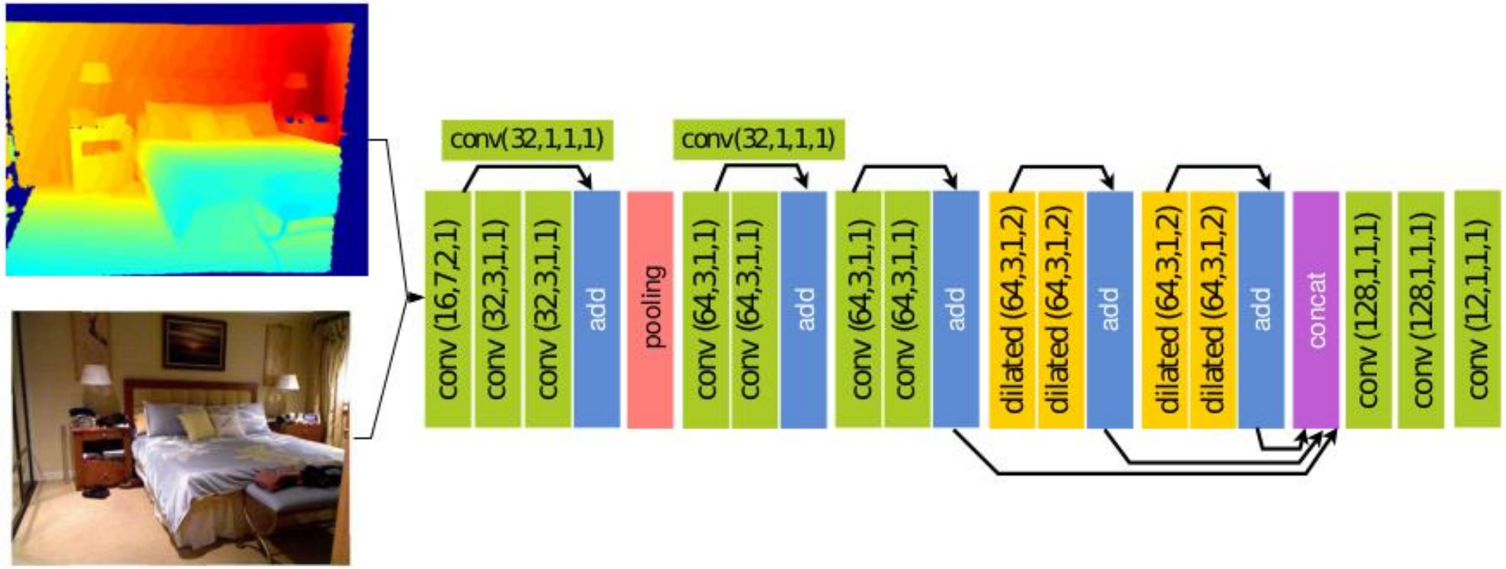
[1] Song, S., Yu, F., Zeng, A., Chang, A.X., Savva, M., and Funkhouser, T.: Semantic Scene Completion from a Single Depth Image. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, Hawaii, July 21-26, pp. 190–198, Piscataway, NJ, July 2017. IEEE.

[2] Guo, Y. and Tong, X.: View-Volume Network for Semantic Scene Completion from a Single Depth Image. In Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, pp. 726–732, Stockholm, Sweden, July 2018. International Joint Conferences on Artificial Intelligence Organization, ISBN 978-0-9992411-2-7. https://doi.org/10.24963/ijcai.2018/101.

Previous Works

Depth maps plus RGB

- Guedes *et al*.[3]

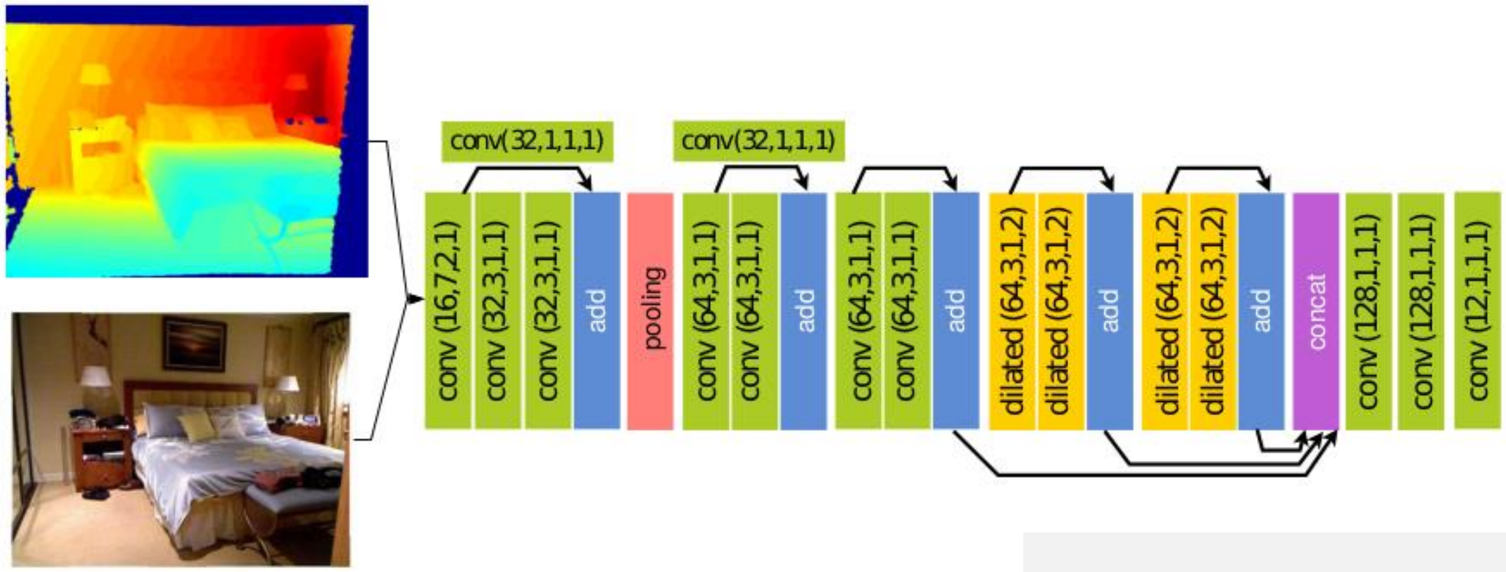


[3] Guedes, A.B.S., de Campos, T.E., and Hilton, A.: Semantic scene completion combining colour and depth: preliminary experiments. In ICCV workshop on 3D Reconstruction Meets Semantics (3DRMS), Venice, Italy, October 2017. Event webpage: http://trimbot2020.webhosting.rug.nl/events/events-2017/3drms/. Also published at arXiv:1802.04735.

Previous Works

Depth maps plus RGB

- Guedes *et al.*[3]



Suffers from RGB data sparsity after projection to 3D

[3] Guedes, A.B.S., de Campos, T.E., and Hilton, A.: Semantic scene completion combining colour and depth: preliminary experiments. In ICCV workshop on 3D Reconstruction Meets Semantics (3DRMS), Venice, Italy, October 2017. Event webpage: http://trimbot2020.webhosting.rug.nl/events/events-2017/3drms/. Also published at arXiv:1802.04735.

Previous Works

Depth map plus 2D segmentation

- Two stream 3D semantic scene completion: Garbade *et al*.[4]

[4] Garbade, M., Sawatzky, J., Richard, A., and Gall, J.: Two stream 3D semantic scene completion. Tech. Rep. arXiv:1804.03550, Cornell University Library, 2018. http://arxiv.org/abs/1804.03550.

Previous Works

Depth map plus 2D segmentation

- Two stream 3D semantic scene completion: Garbade *et al*.[4]
- TNetFusion: Liu *et al.*[5]

[4] Garbade, M., Sawatzky, J., Richard, A., and Gall, J.: Two stream 3D semantic scene completion. Tech. Rep. arXiv:1804.03550, Cornell University Library, 2018. http://arxiv.org/abs/1804.03550.

[5] Liu, S., HU, Y., Zeng, Y., Tang, Q., Jin, B., Han, Y., and Li, X.: See and think: Disentangling semantic scene completion. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.): Procedings of Conference on Neural Information Processing Systems 31 (NIPS), pp. 263–274, Reed Hook, NY, 2018. Curran Associates, Inc. http://papers.nips.cc/paper/7310-see-and-think-disentangling-semantic-scene-completion.

Previous Works

Depth map plus 2D segmentation

- Two stream 3D semantic scene completion: Garbade *et al*.[4]
- TNetFusion: Liu *et al.*[5]

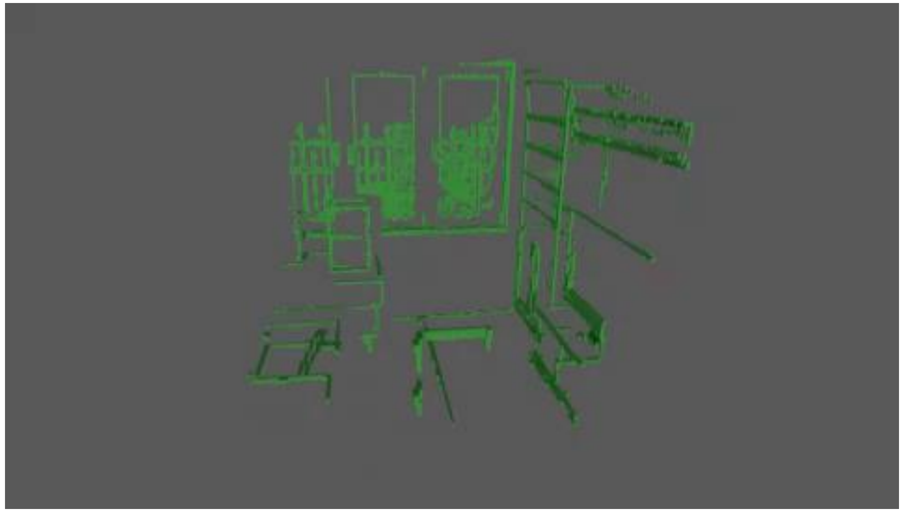
Requires a complex two step training procedure

[4] Garbade, M., Sawatzky, J., Richard, A., and Gall, J.: Two stream 3D semantic scene completion. Tech. Rep. arXiv:1804.03550, Cornell University Library, 2018. http://arxiv.org/abs/1804.03550.

[5] Liu, S., HU, Y., Zeng, Y., Tang, Q., Jin, B., Han, Y., and Li, X.: See and think: Disentangling semantic scene completion. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.): Procedings of Conference on Neural Information Processing Systems 31 (NIPS), pp. 263–274, Reed Hook, NY, 2018. Curran Associates, Inc. http://papers.nips.cc/paper/7310-see-and-think-disentangling-semantic-scene-completion.

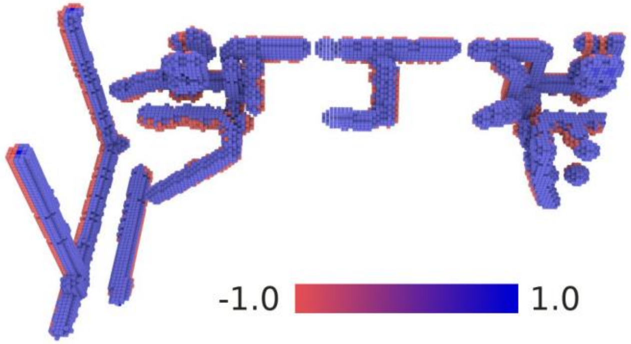
Our Approach: EdgeNet

- We extract boundary information from RGB data and project to 3D...

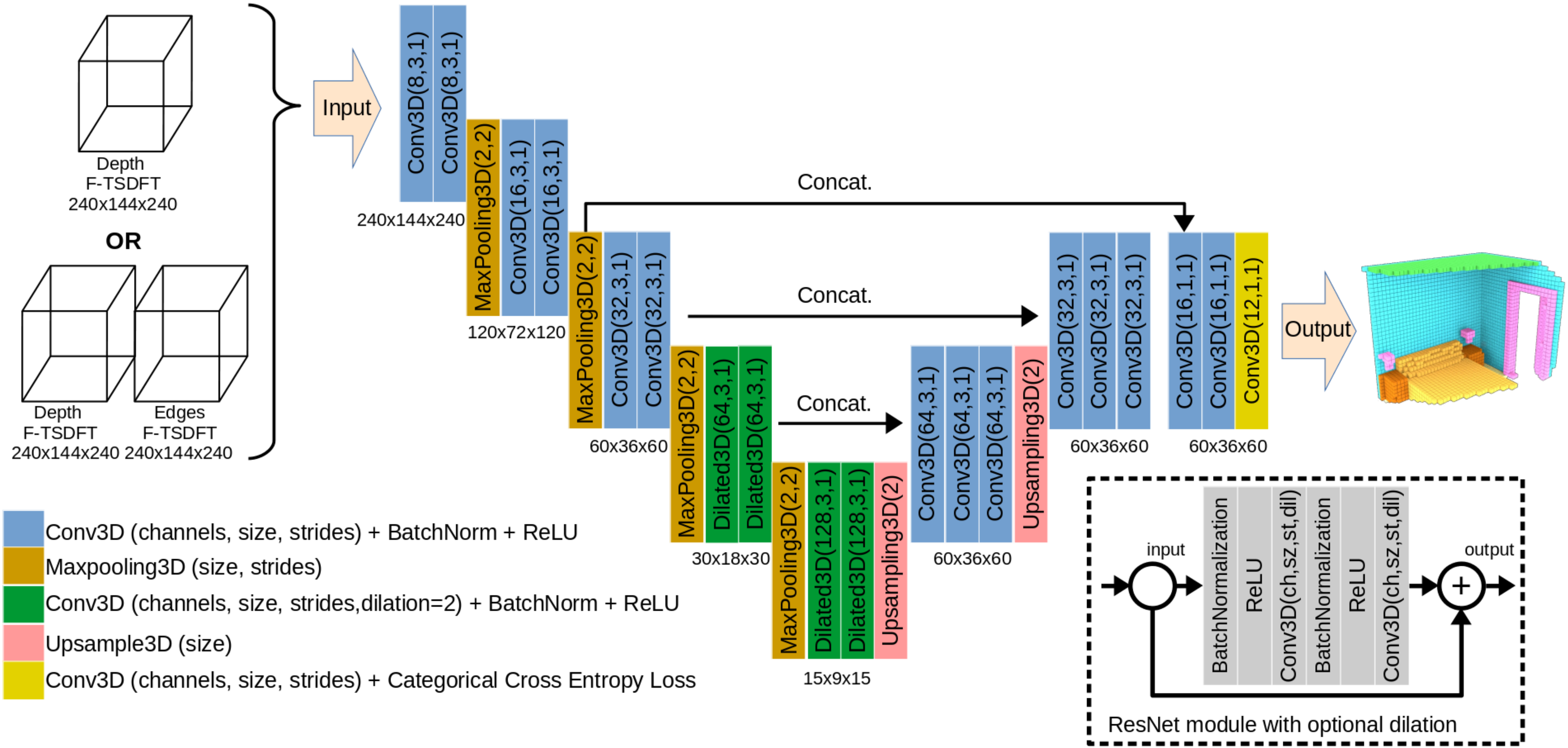


Our Approach: EdgeNet

- ...then, we apply F-TSDF to the projected edge volume



Network Architecture



🟦 Conv3D (channels, size, strides) + BatchNorm + ReLU

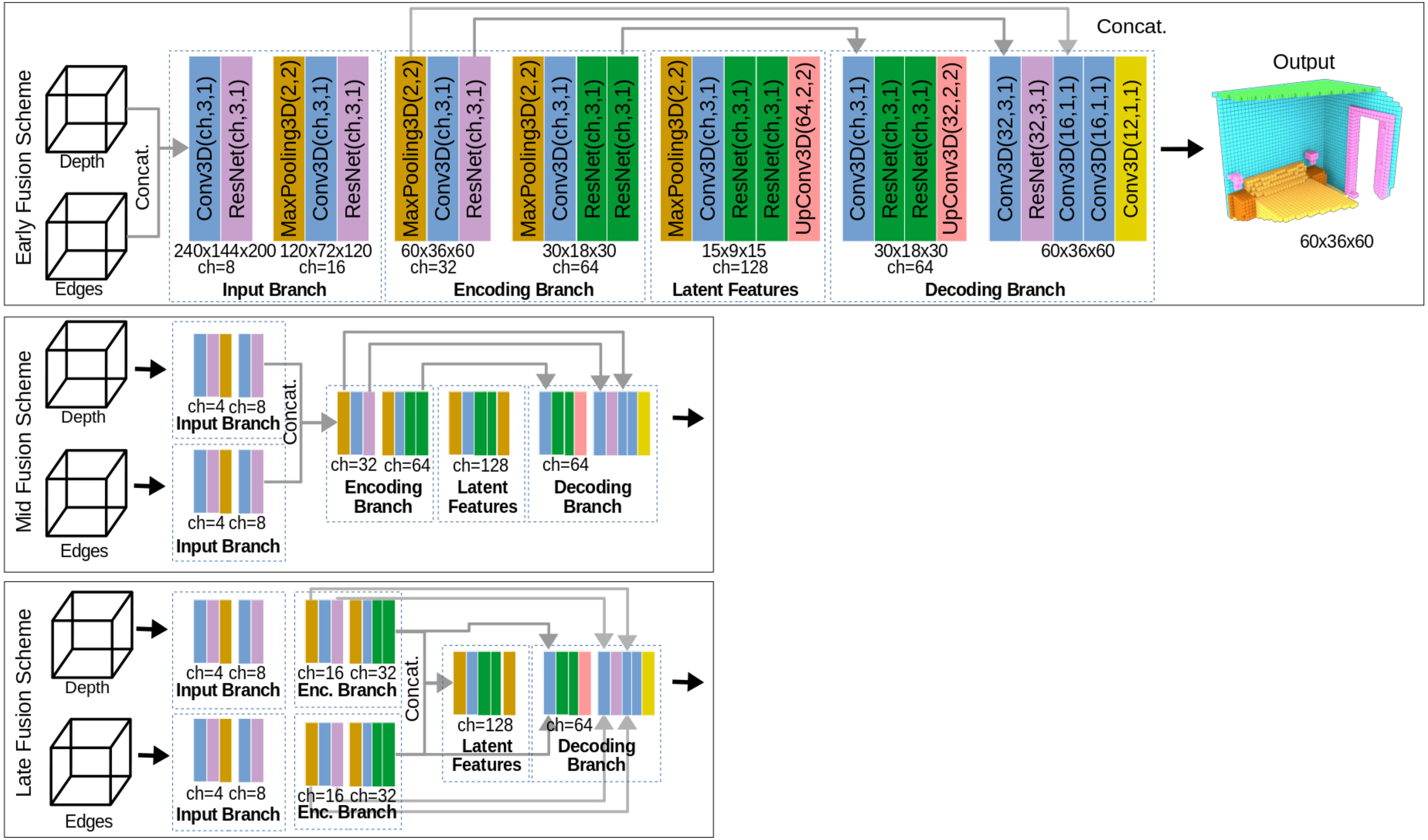
🟫 Maxpooling3D (size, strides)

🟩 Conv3D (channels, size, strides,dilation=2) + BatchNorm + ReLU

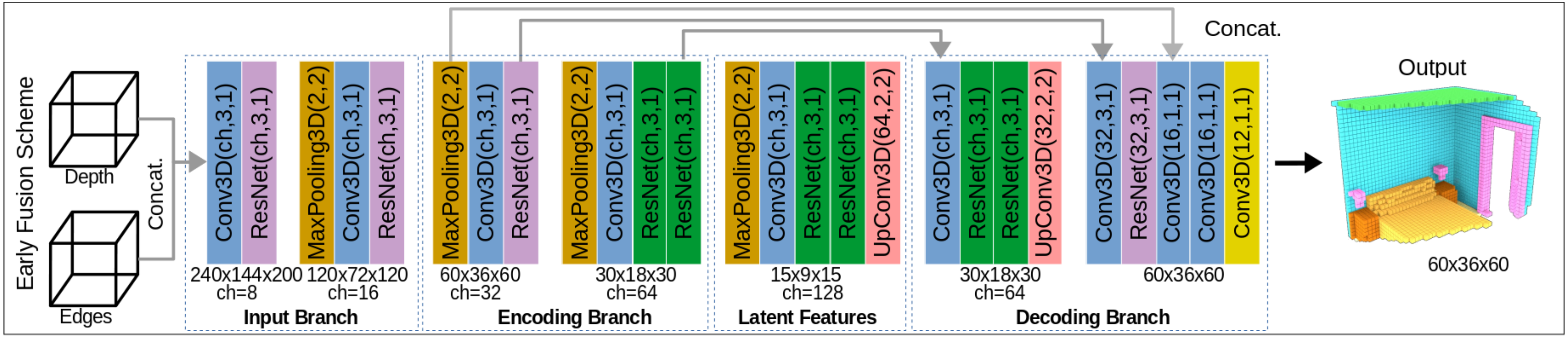
🟥 Upsample3D (size)

🟨 Conv3D (channels, size, strides) + Categorical Cross Entropy Loss

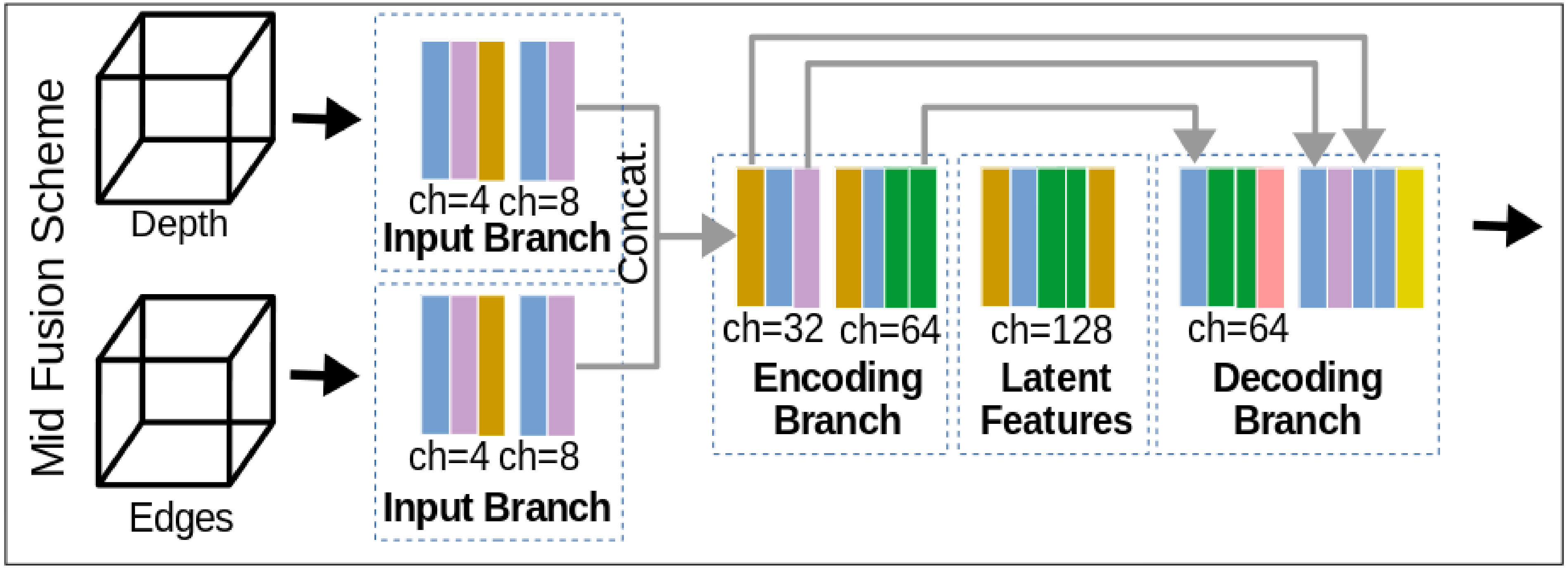
Network Architecture - Fusion Schemes



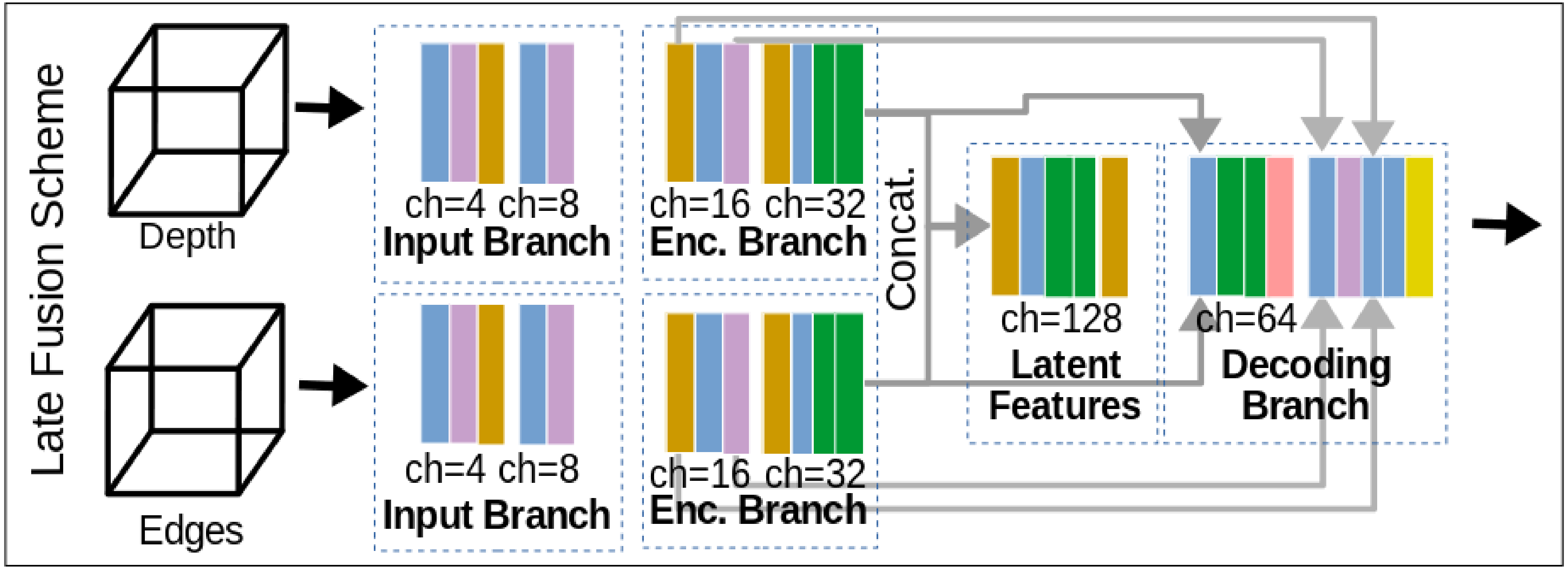
Network Architecture - Fusion Schemes



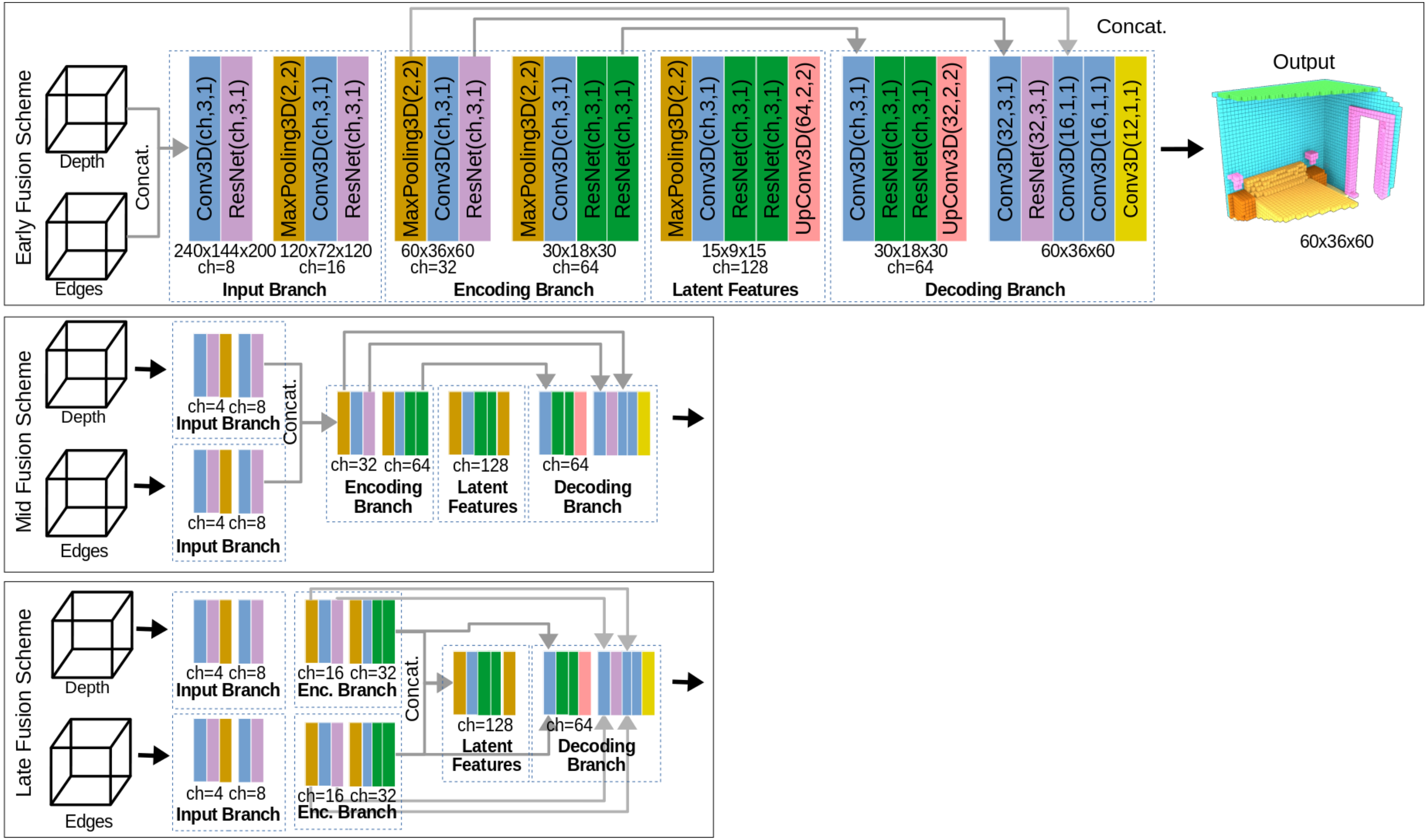
Network Architecture - Fusion Schemes



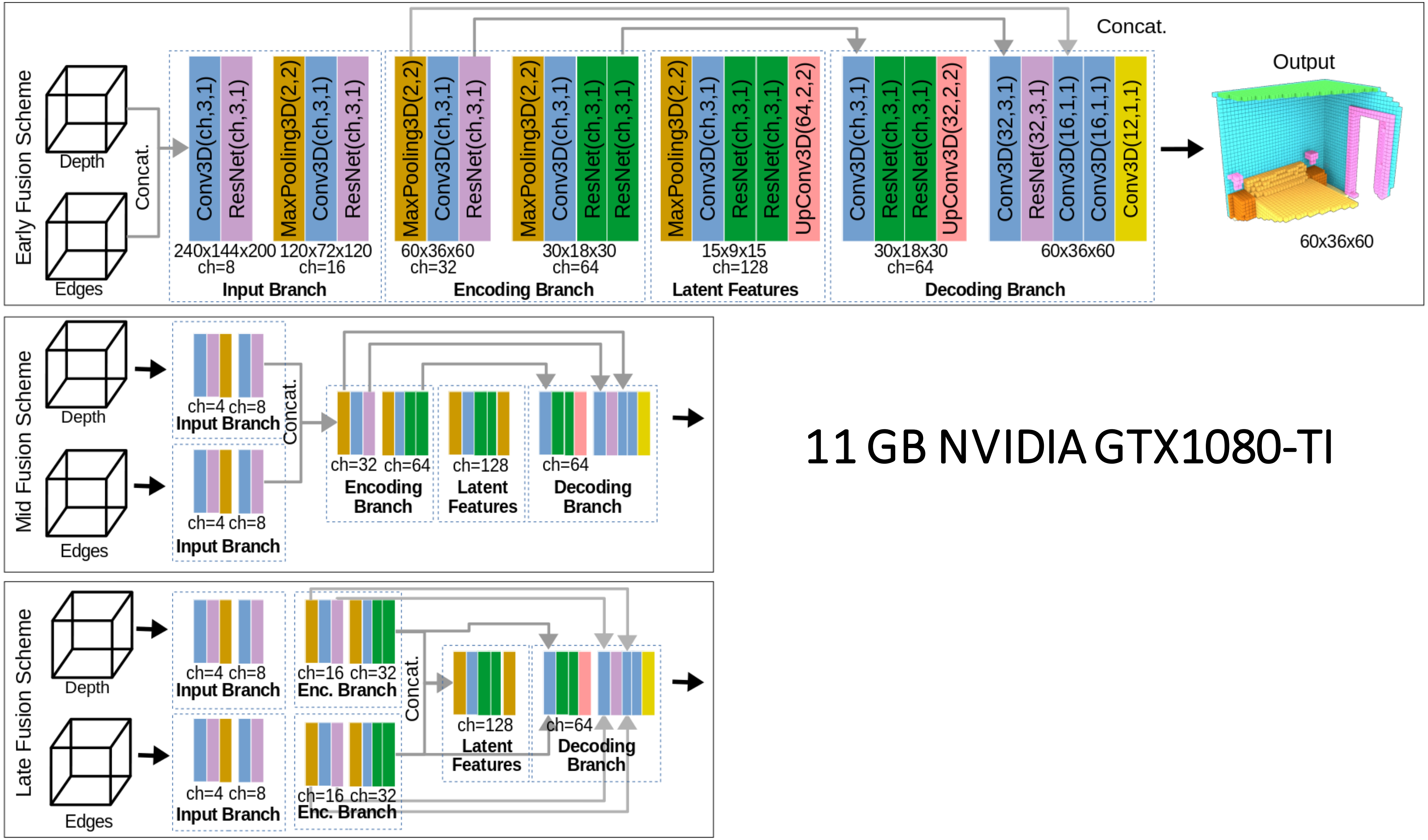
Network Architecture - Fusion Schemes



Network Architecture - Fusion Schemes



Network Architecture - Fusion Schemes



11 GB NVIDIA GTX1080-TI

Datasets

- SUNCG*

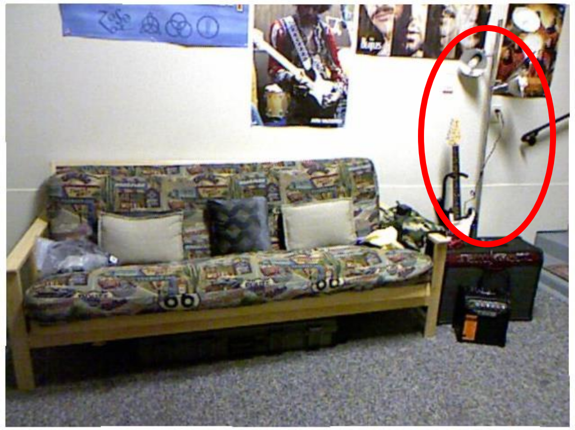
(a) SUNCG dataset (b) 3D Scene (c) Synthetic depth and volumetric ground truth

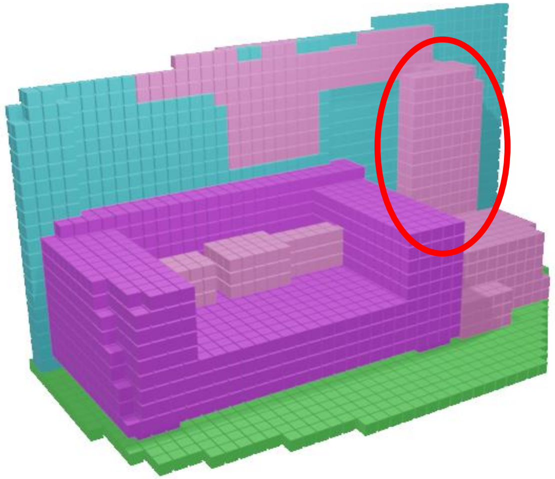
- NYUDv2**

Quantitative Results

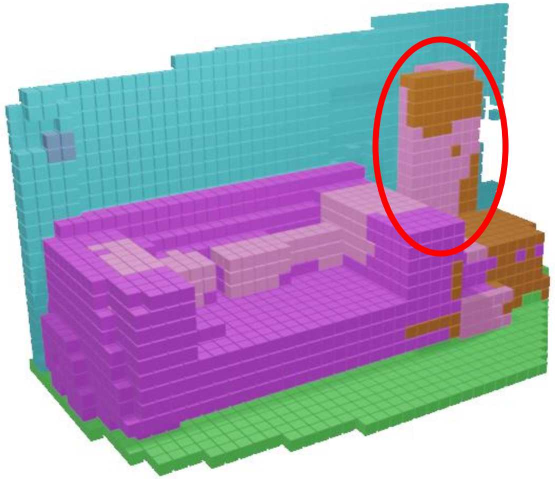
- SUNCG
  - New state-of-the-art result (70.3% avg. IoU)
- NYUD-V2
  - Our solution surpassed previous end-to-end approaches (33.7% avg. IoU)
  - EdgeNet's results are similar to non end-to-end solutions, with a much simpler training pipeline.
- Best Fusion Scheme: Mid Fusion (EdgeNet-MF)

Qualitative Results

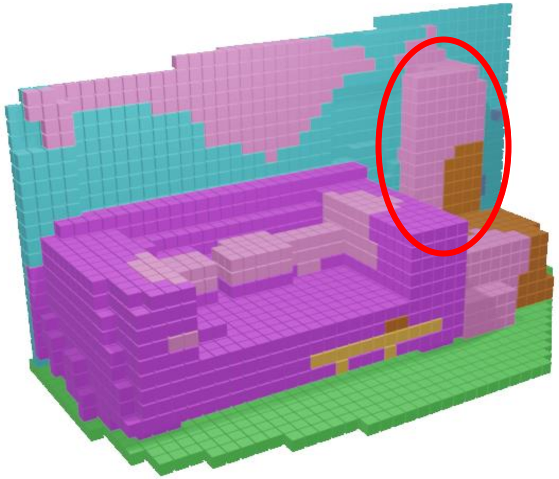




Ground Truth



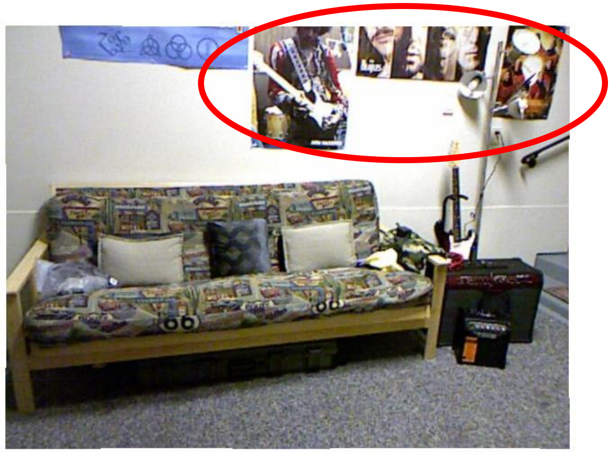
SSCNet

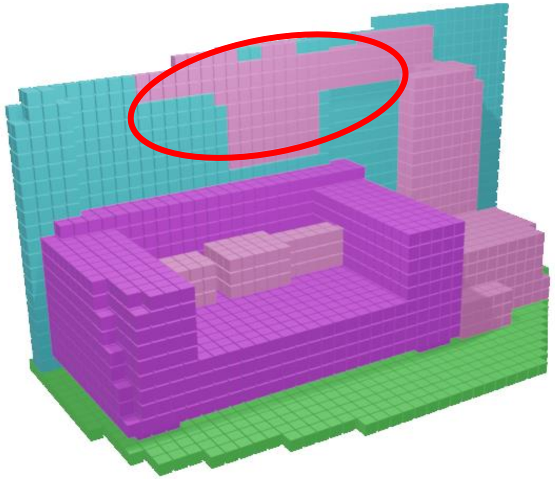


EdgeNet-MF

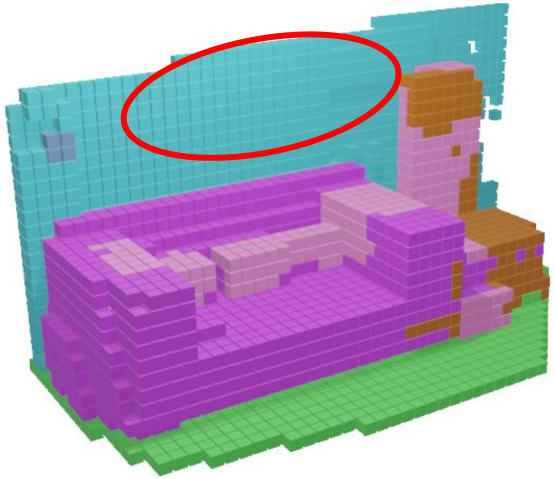
Higher overall accuracy

Qualitative Results

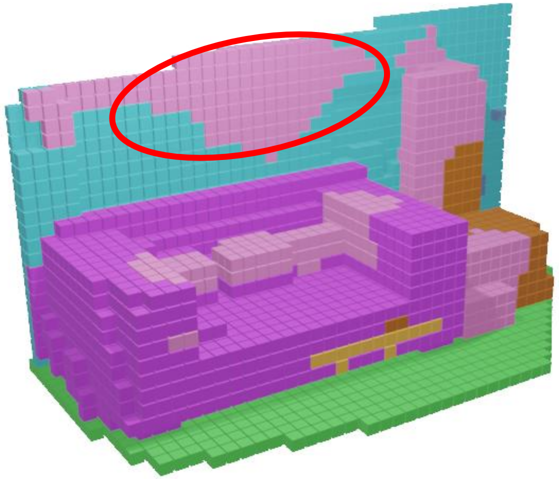




Ground Truth



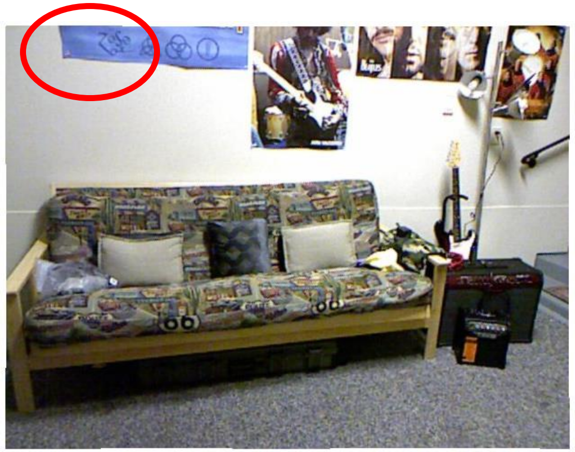
SSCNet

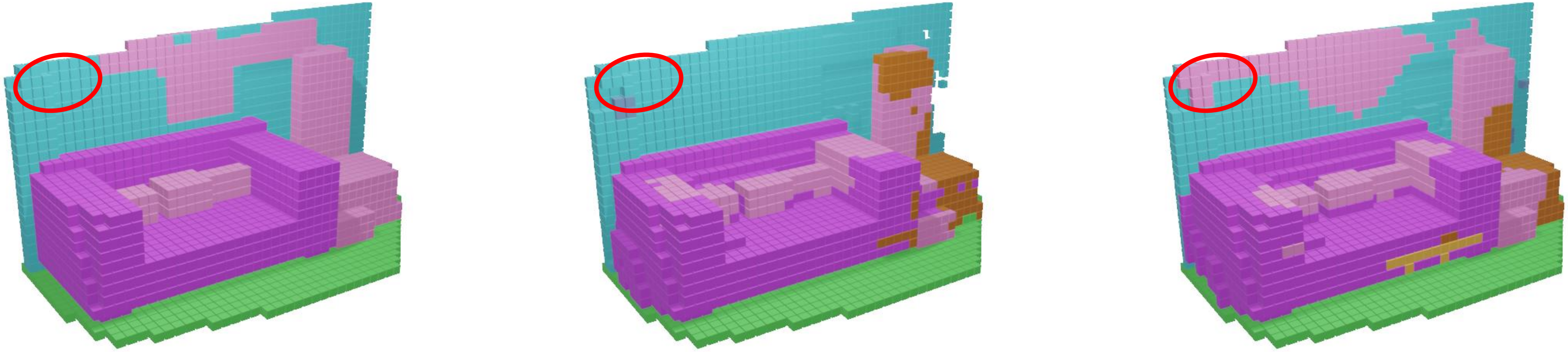


EdgeNet-MF

Hard-to-detect classes

Qualitative Results





Ground Truth

SSCNet

EdgeNet-MF

NYU Ground Truth errors

Conclusions

- A new end-to-end network architecture
- A new strategy to encode data from RGB channels
- Visually perceptible improvements in 3D
- Improvement over the state-of-the-art result on SUNCG
- We surpassed other end-to-end approaches on NYUDv2

Thank you!