Achieving RGB-D Level Segmentation Performance from a Single ToF Camera

Pranav Sharma[3], Jigyasa Singh Katrolia[1], Jason Rambach[1], Bruno Mirbach[1] and Didier Stricker[1,2]
[1]*German Research Center for Artificial Intelligence DFKI, Kaiserslautern, Germany*
[2]*RPTU Kaiserslautern, Germany*
[3]*FAU Erlangen, Germany*

Keywords: Multi-Modal Image Segmentation, Depth Image, Infrared Image, Machine Learning, Time-of-Flight, Deep Learning.

Abstract: Depth is a very important modality in computer vision, typically used as complementary information to RGB, provided by RGB-D cameras. In this work, we show that it is possible to obtain the same level of accuracy as RGB-D cameras on a semantic segmentation task using infrared (IR) and depth images from a single Time-of-Flight (ToF) camera. In order to fuse the IR and depth modalities of the ToF camera, we introduce a method utilizing depth-specific convolutions in a multi-task learning framework. In our evaluation on an in-car segmentation dataset, we demonstrate the competitiveness of our method against the more costly RGB-D approaches.

1 INTRODUCTION

The research field of semantic segmentation is dominated by RGB images. Only recently it shifted in the direction of RGB-D semantic segmentation (Hazirbas et al., 2017; Wang and Neumann, 2018; Cao et al., 2021; Cheng et al., 2017). However, RGB images may not always be available due to practical, logistical and financial reasons. RGB-D cameras incur higher cost and more effort to calibrate the two cameras. Their larger package size often limits their place in real-world applications. Indeed, Time-of-Flight (ToF) depth cameras are often deployed without an accompanying RGB camera for applications like gesture control, in-car monitoring, industry automation and building management (Schneider et al., 2022a; Katrolia et al., 2021a). Infrared images (IR) on the other hand are a by-product of ToF depth cameras (no additional sensor needed), but have not been explored sufficiently, specifically in combination with depth data (Agresti et al., 2017; Su et al., 2016).

Infrared images from ToF cameras provide the magnitude of the modulated light reflected from the scene and contain shape and semantic features in a different spectral range (Hahne, 2012). Due to the similarities between RGB and IR images, it is natural to attempt to adapt existing RGB-D fusion approaches to combine IR and depth images (Schneider et al., 2022b; Katrolia et al., 2021b). However, in most RGB-D methods depth information is only an accessory to the color information and is consumed by the same type of neural network layers despite their differences. Some recent works proposed depth-specific operations like depth-aware (Wang and Neumann, 2018) and shape-aware (Cao et al., 2021) convolutions. We observe that both IR and depth outputs from a ToF camera are related in many ways and therefore these depth-specific operations can be applied to IR images as well. For example, the intensity of light reflected from an object decreases as distance to the object increases. Object surfaces closer to the camera will reflect more light implying that pixel intensities in an infrared image varies with the shape of that object. We can see this by comparing the same

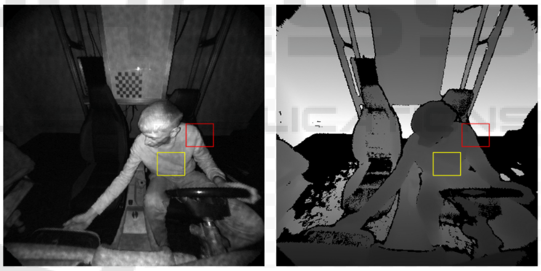

Figure 1: Patch similarities of IR and Depth modalities of a ToF Camera.

Sharma, P., Katrolia, J., Rambach, J., Mirbach, B. and Stricker, D.
Achieving RGB-D Level Segmentation Performance from a Single ToF Camera.
DOI: 10.5220/0012265100003654

In *Proceedings of the 13th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2024)*, pages 171-178
ISBN: 978-989-758-684-2; ISSN: 2184-4313

colored patches in Figure 1 and note how both image patches have the same relative changes in pixel values. We use this observation to leverage the shape-aware convolution operation for both IR and depth images to learn more meaningful features from both modalities.

We aim to use the available modalities from a single ToF camera to achieve semantic segmentation performance comparable to RGB-D methods using an architecture that is tailored to IR-Depth (IR-D) input. We take inspiration from (Cao et al., 2021; Wang and Neumann, 2018) and design a depth-aware shape convolution operation that consumes IR-D input in a multi-task learning (MTL) architecture with depth completion as an auxiliary task. Our proposed method surpasses the baseline RGB-D based methods using only a ToF camera. To summarize, the main contributions of our paper are:

- We show that recently introduced depth-aware and shape convolution operations are suitable for IR images and can achieve same performance on IR-D images as compared to RGB-D data. We propose a new convolution operation combining the two and show that it is better than using either of them alone for IR-D data.
- We show that multi-task learning with depth completion as an auxiliary task can be beneficial for the depth and IR segmentation task.
- In our experimental evaluation on the TICaM dataset, we show that with our proposed approach it is possible to surpass RGB-D segmentation performance using only depth and infrared (IR-D) images from a single ToF camera.

2 RELATED WORK

RGB-D Semantic Segmentation. A wide range of approaches have been proposed for incorporating the depth information in RGB-D semantic segmentation and can be broadly classified into three categories as outlined by (Barchid et al., 2021): (1) Using the depth channel as an additional input and performing fusion at different levels (early (Song et al., 2017), feature-level (Hazirbas et al., 2017) or late (Cheng et al., 2017) fusion); (2) Depth as supervision signal for auxiliary tasks such as depth estimation or completion (Jiao et al., 2019) and (3) Depth-specific operations. Depth-Aware CNN (Wang and Neumann, 2018) showed that object boundaries correlate with depth gradients and created *depth-aware convolution* and *depth-aware pooling* functions. ShapeConv (Cao et al., 2021) used a *shape convolution* kernel to ensure convolution kernels give consistent responses to object classes at different locations in the scene. (Chen et al., 2019) used depth information to adjust the neighbourhood size of a 3D convolution filter.

Depth and IR for Semantic Segmentation. Depth-only methods for semantic segmentation are typically applied to solve very specific tasks, like object manipulation via mechanical arm (Zhou et al., 2018), hand segmentation or hand and object segmentation (Rezaei et al., 2021; Lim et al., 2019). Even fewer methods have explored the combination of IR and depth. (Su et al., 2016) used IR to classify materials since the infrared response depends on the material of the object. (Agresti et al., 2017) used IR image as an additional input channel for improving the depth predicted from a setup containing both stereo and ToF cameras and (Lorenti et al., 2018) used semi-supervised learning for image segmentation via region merging. (Katrolia et al., 2021b) compared segmentation of depth maps using image-based methods against the use of point clouds inputs derived from the depth maps.

Multi-Task Learning (MTL). The survey by (Crawshaw, 2020) describes many examples where simultaneous learning of two or more related tasks can boost the performance on either task. RGB-D segmentation is enhanced using auxiliary tasks like depth and surface normal prediction (Wang et al., 2022). MTL methods typically employ a single encoder to learn features from the available input modalities and two separate task-specific decoders to perform prediction (Wang et al., 2022). Cross-Stitch network (Misra et al., 2016) replace total parameter sharing with controlled sharing between the two tasks using learned sharing weights.

3 BACKGROUND AND NOTATION

In this section, we briefly introduce our notation and describe the depth-specific convolution operations introduced by ShapeConv (Cao et al., 2021) and Depth-Aware CNN (Wang and Neumann, 2018).

3.1 ShapeConv: Shape-Aware Convolutional Layer

In order to design convolutions that are invariant to different depth values (*base*) when the underlying relative difference in depth in a local patch (*shape*) remains same, ShapeConv (Cao et al., 2021) suggested decomposing an image patch $\mathbb{P} \in R^{K_h \mathrm{x} K_w \mathrm{x} C_{in}}$ into a *shape* component $\mathbb{P}_s$ and a *base* component $\mathbb{P}_b$. These two components are operated on separately by a corresponding shape $\mathbb{W}_S$ and base kernel $\mathbb{W}_B$ before being passed to the standard convolutional kernel $\mathbb{K}$. Instead of decomposing the image patch, the convolutional kernel itself can be decomposed to the respective components as shown in equation (1). Here $m(\mathbb{K})$ refers to mean of the kernel.

$$\begin{aligned}\mathbb{K}_B &= m(\mathbb{K})\\ \mathbb{K}_S &= \mathbb{K} - m(\mathbb{K})\end{aligned} \tag{1}$$

Shape convolution is then written with *ShapeConv* as:

$$\begin{aligned}\mathbb{F} &= \mathit{ShapeConv}(\mathbb{K}, \mathbb{W}_B, \mathbb{W}_S, \mathbb{P}_i)\\ &= \mathit{Conv}(\mathbb{W}_B \diamond m(\mathbb{K}) + \mathbb{W}_S * (\mathbb{K} - m(\mathbb{K})), \mathbb{P})\\ &= \mathit{Conv}(\mathbb{W}_B \diamond \mathbb{K}_B + \mathbb{W}_S * \mathbb{K}_S, \mathbb{P})\\ &= \mathit{Conv}(\mathbf{K_B} + \mathbf{K_S}, \mathbb{P})\\ &= \mathit{Conv}(\mathbf{K_{BS}}, \mathbb{P})\end{aligned} \tag{2}$$

Here, $\diamond$ and $*$ represents the base and shape product respectively. $\mathbb{W}_B$ and $\mathbb{W}_S$ are learnable weights corresponding to *base* and *shape* components respectively.

3.2 Depth-Aware CNN and Depth Similarity

In depth-aware convolution DCNN (Wang and Neumann, 2018), pixels with similar depth values to the centre pixel are weighted more than other pixels. This property is named depth similarity. The depth similarity function $\mathrm{F}_D(p_i, p_j)$ calculates the difference of depth values $D(p_i), D(p_j)$ between two pixels p_i and p_j respectively.

$$\mathrm{F}_D(p_i, p_j) = \exp(-\alpha|D(p_i) - D(p_j)|) \tag{3}$$

Depth-aware convolution is written with the depth similarity function F_D as:

$$y(p_0) = \sum_{p_n \in R} w(p_n)\mathrm{F}_D(p_0, p_0 + p_n)x(p_0 + p_n) \tag{4}$$

In Equation (4), the depth similarity term (F_D) is introduced with the convolution operation. The convolved features are weighted by F_D . The parameter α weighs the influence of the depth similarity function F_D on the convolution operation.

4 METHOD

We propose a depth-aware shape convolution operation applied within in a multi-task learning network. Our primary task is semantic segmentation using concatenated infrared and depth images from a ToF camera and our auxiliary task is depth completion for missing pixels in raw depth images.

4.1 Depth-Aware Shape Convolution

We design a depth-aware shape convolution, where the shape kernel in ShapeConv is supplemented with the depth similarity measure F_D as computed in equation (3). Formally, this integration can be written in two steps. First the kernel is decomposed into *shape* and *base* kernel as shown in equation (2). After the calculation of the weights $\mathbf{K_{BS}}$, the term $w(p_n)$ in equation (4) is replaced with $\mathbf{K_{BS}}$ calculated from shape kernel. In this way, kernel weights calculated using the shape kernel are integrated with the depth similarity of DCNN. Equation (3) can thus be rewritten for our Depth-aware ShapeConv as:

$$y(p_0) = \sum_{p_n \in R} \mathbf{K_{BS}}_n \mathrm{F}_D(p_0, p_0 + p_n)x(p_0 + p_n) \tag{5}$$

4.2 Infrared and Depth-Aware Multi-Task Network

We realize a hard parameter sharing-based multi-task network with semantic segmentation as the main task and depth completion as the auxiliary task (Figure 2). We use ResNet-101 as the backbone feature extractor to encode features from concatenated infrared and depth (IR-D) images. The convolution layers in the ResNet encoder are replaced with depth-aware shape convolutions presented in section 4.1. The extracted features are passed to two task-specific decoders that generate final segmentation masks and depth values for missing pixels. For training the depth filling branch, the ground truth is prepared as described in section 5.1. We follow the training strategy from (Mao et al., 2020) and use predicted depth values only for missing pixels to calculate the error between ground truth and predicted depth. The dense depth map predicted by the network is then multiplied with the missing pixels mask (1 for missing pixels, 0 otherwise) to keep predicted depth values only for the pixels that are missing in raw image. The remaining values are then replaced by the corresponding depth values in input image.

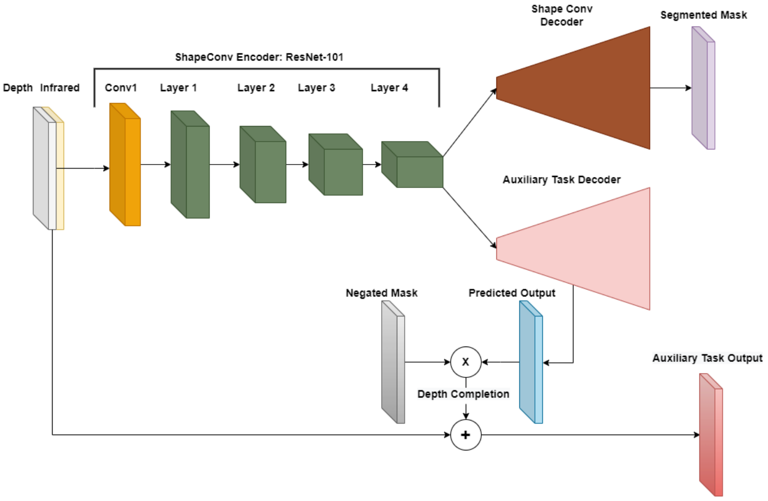

Figure 2: MTL architecture for segmentation task and dense depth prediction using depth-aware shape convolutions.

Table 1: Comparison of our proposed method to segmentation baselines for RGB-D and IR-D data. The best result in the IR-D category is marked bold and second best is underlined. Our MTL-DA-ShapeConv method achieves state-of-the-art results in the IR-D category and even outperforms RGB-D methods by a large margin on the class accuracy and mean IoU metrics.

Input	Baselines	Pixel Acc.	Class Acc.	Mean IoU	f.w.IoU
RGB-D	ShapeConv (Cao et al., 2021)	**97.86**	**81.25**	**77.39**	**95.92**
	Depth-Aware CNN (Wang and Neumann, 2018)	94.63	66.88	54.14	90.78
	FuseNet (Hazirbas et al., 2017)	95.35	56.89	42.46	92.43
IR-D	ShapeConv (Cao et al., 2021)	**97.75**	81.31	74.61	**95.76**
	Depth-Aware CNN (Wang and Neumann, 2018)	93.52	60.52	50.03	88.45
	FuseNet (Hazirbas et al., 2017)	93.18	55.73	39.61	88.84
	Ours DA-ShapeConv	97.57	85.08	78.39	95.42
	Ours MTL-DA-ShapeConv	97.73	**85.98**	**79.73**	95.68

Table 2: Number of model parameters (in millions) and per image inference time (in milliseconds) for baselines and proposed MTL architecture. (Here, ch represents channels).

	Method	Input	Params (million)	Time (ms)
	ShapeConv (Cao et al., 2021)	4-ch	60.55	35.04
	ShapeConv (Cao et al., 2021)	2-ch	60.54	32.20
Ours	DA-ShapeConv	2-ch	60.54	37.12
	MTL-DA-ShapeConv	2-ch	78.12	37.16

5 EXPERIMENTS

5.1 Dataset

We evaluate our approach on the in-car cabin dataset TICaM (Katrolia et al., 2021a) that provides RGB, depth and infrared images recorded with a single ToF camera and corresponding ground-truth segmentation masks. TICaM is the only dataset that fulfills our experimental requirements. Surprisingly we could not find any other dataset that provided segmentation masks for all three image modalities: RGB, depth and infrared.

We used the real image-set of TICaM with the suggested split of 4666 training images and 2012 test images for our experiments. Following (Katrolia et al., 2021a) we combine different object classes into a single ’object’ class to have 6 object classes in total. The RGB images have different resolution and FoV to depth and infrared images. To align the images, the RGB images are first mapped to the pinhole model of the depth images. Subsequently all the images are centre-cropped to size 230 × 418. Normalization of the infrared image is implemented by first removing the outliers by calculating the 99th percentile of the image and then scaling the image to the range of 0-255. To further enrich the information, histogram equalization and gamma filtering are applied to the normalized infrared image.

Ground-Truth Preparation. To train for the auxiliary task of dense depth prediction, completed depth images are required as ground truth. As the TICaM dataset only provides depth images with holes, filled versions of these depth maps are artificially created in this work using the ”Colorization using Optimization” scheme (Levin et al., 2004). The depth images are filled by enforcing similar depth values to the neighbouring pixels with similar intensities. The information on pixel intensities is provided by infrared images. During training, the error is calculated for the missing pixels following the training strategy of (Mao et al., 2020).

5.2 State-of-the-Art Comparison

Table 1 provides an evaluation of segmentation accuracy of our proposed method. We include results of state-of-the-art methods on RGB-D as well as on our target modality, IR-D ToF data. We choose three existing methods for RGB-D segmentation and train them on both RGB-D and IR-D images to establish our baselines and better evaluate the difference between RGB-D and IR-D inputs. We choose FuseNet (Hazirbas et al., 2017) since it is an established and well-tested network on many benchmark datasets for RGB-D segmentation, however it has not been tested yet on TICaM dataset. ShapeConv (Cao et al., 2021) and Depth-aware CNN (Wang and Neumann, 2018) on the other hand are more recent methods that use novel convolution operations unlike FuseNet. FuseNet and Depth-aware CNN use a VGG-16 backbone, while ShapeConv uses ResNet-101. All networks use SGD optimizer with a momentum of 0.9 and weight decay of 5×10^{-4} to update the weights. By default, Deeplab v3+ and ShapeConv use pre-trained weights while FuseNet and Depth-aware CNN are initialized using the kaiming initialization (He et al., 2015). For training with IR-D images we replicate infrared images to form 3-channel images and concatenate them with single channel depth images. We report pixel accuracy, class accuracy, mean Intersection-over-Union (IoU) and frequency weighted IoU (f.w.IoU) in Table 1 for all results.

We can observe that the architectures that incorporate depth in an informed manner outperform FuseNet which simply concatenates depth with other modalities. Also, the combination of infrared and depth can be used instead of RGB-D input while achieving almost the same performance on segmentation but the disparity between the achieved performance is least when using depth-aware architectures with ShapeConv outperforming the other two methods. We can see also from Figure 3 that ShapeConv with both 3-channel and 1-channel infrared images have similar mask predictions.

Our proposed methods applied on IR-D images are presented in Table 1 as well. DA-ShapeConv is the integration of depth aware convolutions into the ShapeConv architecture as described in Section 4.2. We can note that combining both depth-aware and shape convolutions gives significant improvement on class accuracy as well as mean IoU, compared to the ShapeConv and Depth-Aware CNN baselines and even their RGB-D versions. MTL-DA-ShapeConv denotes our Multi-Task network, also incorporating the proposed auxiliary task of depth completion. We can see that the MTL architecture improves over the best performing RGB-D ShapeConv baseline with significant improvement in class accuracy and mean IoU. Also, the improvement over our method without the auxiliary task (DA-ShapeConv) validates the use of multi-task learning with depth completion for this task. Overall, the results prove our main hypothesis, that by applying our method we can outperform state-of-the-art RGB-D methods using only IR-D images provided by a single ToF camera.

The results in Table 1 are also directly comparable with results on the TICAM dataset using only depth information as reported in (Katrolia et al., 2021b). This clearly indicates the advantage of the IR-D combination of modalities over the use of the D channel only for segmentation.

Additionally, we provide qualitative results of the segmentation output from all evaluated network variants on images from the TICAM dataset in Figure 3. We see that our methods provide smoother masks around the edges, especially for the smaller objects on the passenger seat as well as better class predictions in some cases.

Finally, a runtime comparison for the evaluated methods is provided in Table 2. The DA-ShapeConv method does not have an impact on the number of parameters and a limited increase in runtime per frame compared to ShapeConv. The MTL-DA-ShapeConv method leads to an increase in parameters for the network due to the additional branch, however this does not impact the inference runtime.

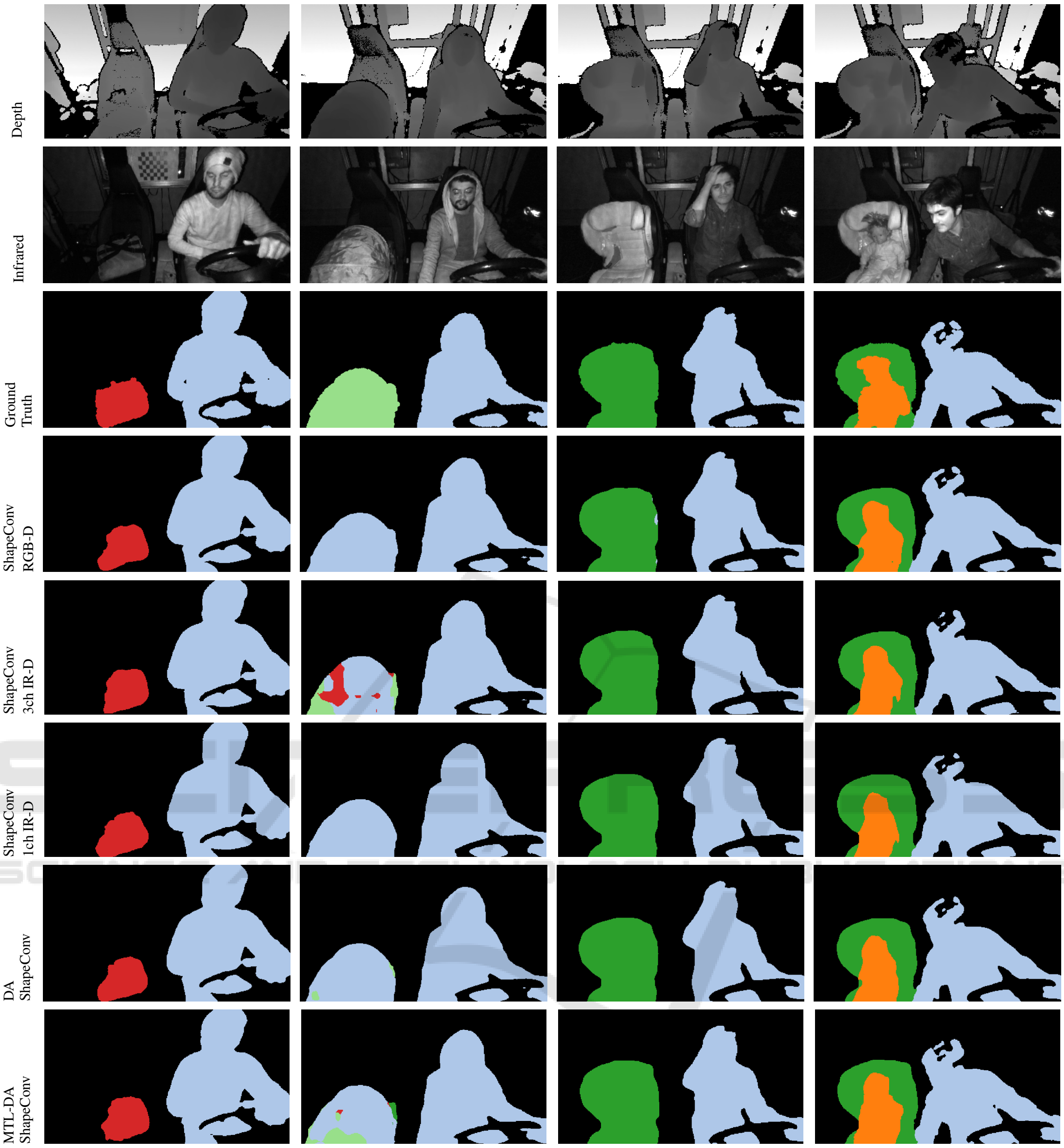

Figure 3: Predictions from RGB-D and proposed IR-D methods.

6 CONCLUSION

We designed a network for IR-D segmentation that performs equally well as RGB-D segmentation so that inconvenient and expensive RGB-D cameras can be replaced with single Time-of-Flight (ToF) cameras. We showed that existing fusion approaches for RGB-D segmentation can be used with IR-D input if standard convolutions are replaced with depth-specific convolutions. We then presented a combination of depth-aware and shape-aware convolutions, and designed a multi-task learning (MTL) architecture with this new convolution operation. We employ hard parameter sharing between our main and auxiliary tasks of segmentation and depth filling respectively. Through progressive modifications to the input, the convolution operation, and the network architecture we showed that we can outperform all baseline methods. We conclude that using images from a single ToF camera, it is possible to surpass RGB-D segmentation performance with our designed MTL architecture.

ACKNOWLEDGEMENTS

This work was partially funded within the Electronic Components and Systems for European Leadership (ECSEL) joint undertaking in collaboration with the European Union's H2020 Framework Program and the Federal Ministry of Education and Research of the Federal Republic of Germany (BMBF) under grant agreement 16ESE0424/GA826600 (VIZTA), and partially funded by the German Ministry of Education and Research (BMBF) under Grant Agreement 01IW20002 (SocialWear).

REFERENCES

Agresti, G., Minto, L., Marin, G., and Zanuttigh, P. (2017). Deep learning for confidence information in stereo and tof data fusion. In *IEEE International Conference on Computer Vision Workshops (CVPR-W)*, pages 697–705.

Barchid, S., Mennesson, J., and Djéraba, C. (2021). Review on indoor rgb-d semantic segmentation with deep convolutional neural networks. In *International Conference on Content-Based Multimedia Indexing (CBMI)*, pages 1–4. IEEE.

Cao, J., Leng, H., Lischinski, D., Cohen-Or, D., Tu, C., and Li, Y. (2021). Shapeconv: Shape-aware convolutional layer for indoor rgb-d semantic segmentation. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7088–7097.

Chen, Y., Mensink, T., and Gavves, E. (2019). 3d neighborhood convolution: learning depth-aware features for rgb-d and rgb semantic segmentation. In *International Conference on 3D Vision (3DV)*, pages 173–182. IEEE.

Cheng, Y., Cai, R., Li, Z., Zhao, X., and Huang, K. (2017). Locality-sensitive deconvolution networks with gated fusion for rgb-d indoor semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3029–3037.

Crawshaw, M. (2020). Multi-task learning with deep neural networks: A survey. *arXiv preprint arXiv:2009.09796*.

Hahne, U. (2012). *Real-time depth imaging*. PhD thesis, Berlin Institute of Technology.

Hazirbas, C., Ma, L., Domokos, C., and Cremers, D. (2017). Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture. In *Asian Conference on Computer Vision, 2016*, pages 213–228. Springer.

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034.

Jiao, J., Wei, Y., Jie, Z., Shi, H., Lau, R. W., and Huang, T. S. (2019). Geometry-aware distillation for indoor semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Katrolia, J. S., El-Sherif, A., Feld, H., Mirbach, B., Rambach, J. R., and Stricker, D. (2021a). Ticam: A time-of-flight in-car cabin monitoring dataset. In *British Machine Vision Conference (BMVC)*, page 277.

Katrolia, J. S., Krämer, L., Rambach, J., Mirbach, B., and Stricker, D. (2021b). Semantic segmentation in depth data: A comparative evaluation of image and point cloud based methods. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 649–653. IEEE.

Levin, A., Lischinski, D., and Weiss, Y. (2004). Colorization using optimization. *ACM Transactions on Graphics*, 23.

Lim, G. M., Jatesiktat, P., Kuah, C. W. K., and Ang, W. T. (2019). Hand and object segmentation from depth image using fully convolutional network. In *International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 2082–2086. IEEE.

Lorenti, L., Giacomantone, J., and Bria, O. N. (2018). Unsupervised tof image segmentation through spectral clustering and region merging. *Journal of Computer Science & Technology*, 18.

Mao, J., Li, J., Li, F., and Wan, C. (2020). Depth image inpainting via single depth features learning. In *International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, pages 116–120. IEEE.

Misra, I., Shrivastava, A., Gupta, A., and Hebert, M. (2016). Cross-stitch networks for multi-task learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3994–4003.

Rezaei, M., Farahanipad, F., Dillhoff, A., Elmasri, R., and Athitsos, V. (2021). Weakly-supervised hand part segmentation from depth images. In *PErvasive Technologies Related to Assistive Environments*, pages 218–225.

Schneider, P., Anisimov, Y., Islam, R., Mirbach, B., Rambach, J., Stricker, D., and Grandidier, F. (2022a). Timo—a dataset for indoor building monitoring with a time-of-flight camera. *Sensors*, 22(11):3992.

Schneider, P., Rambach, J., Mirbach, B., and Stricker, D. (2022b). Unsupervised anomaly detection from time-of-flight depth images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 231–240.

Song, H., Liu, Z., Du, H., Sun, G., Le Meur, O., and Ren, T. (2017). Depth-aware salient object detection and segmentation via multiscale discriminative saliency fusion and bootstrap learning. *IEEE Transactions on Image Processing*, 26(9):4204–4216.

Su, S., Heide, F., Swanson, R., Klein, J., Callenberg, C., Hullin, M., and Heidrich, W. (2016). Material classification using raw time-of-flight measurements. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3503–3511.

Wang, W. and Neumann, U. (2018). Depth-aware cnn for rgb-d segmentation. In *European Conference on Computer Vision (ECCV)*, pages 135–150.

Wang, Y., Tsai, Y.-H., Hung, W.-C., Ding, W., Liu, S., and Yang, M.-H. (2022). Semi-supervised multi-task learning for semantics and depth. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2505–2514.

Zhou, M., Song, W., Shen, L., and Zhang, Y. (2018). Stacked objects segmentation based on depth image. In *International Conference on Optical and Photonic Engineering (icOPEN)*, volume 10827, pages 363–368. SPIE.