UNIVERSITY OF AMSTERDAM

UvA-DARE (Digital Academic Repository)

HaarNet: Large-Scale Linear-Morphological Hybrid Network for RGB-D Semantic Segmentation

Groenendijk, R.; Dorst, L.; Gevers, T.

DOI
10.1007/978-3-031-57793-2_19

Publication date
2024

Document Version
Final published version

Published in
Discrete Geometry and Mathematical Morphology

License
Article 25fa Dutch Copyright Act (https://www.openaccess.nl/en/policies/open-access-in-dutch-copyright-law-taverne-amendment)

Link to publication

Citation for published version (APA):
Groenendijk, R., Dorst, L., & Gevers, T. (2024). HaarNet: Large-Scale Linear-Morphological Hybrid Network for RGB-D Semantic Segmentation. In S. Brunetti, A. Frosini, & S. Rinaldi (Eds.), *Discrete Geometry and Mathematical Morphology: Third International Joint Conference, DGMM 2024, Florence, Italy, April 15–18, 2024 : proceedings* (pp. 242–254). (Lecture Notes in Computer Science; Vol. 14605). Springer. https://doi.org/10.1007/978-3-031-57793-2_19

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, P.O. Box 19185, 1000 GD Amsterdam, The Netherlands. You will be contacted as soon as possible.
UvA-DARE is a service provided by the library of the University of Amsterdam (https://dare.uva.nl)

Download date:14 Nov 2026

HaarNet: Large-Scale Linear-Morphological Hybrid Network for RGB-D Semantic Segmentation

Rick Groenendijk(✉), Leo Dorst, and Theo Gevers

University of Amsterdam, Science Park 900, 1098XH Amsterdam, The Netherlands
{r.w.groenendijk,l.dorst,th.gevers}@uva.nl

Abstract. Signals from different modalities each have their own combination algebra which affects their sampling processing. RGB is mostly linear; depth is a geometric signal following the operations of mathematical morphology. If a network obtaining RGB-D input has both kinds of operators available in its layers, it should be able to give effective output with fewer parameters. In this paper, morphological elements in conjunction with more familiar linear modules are used to construct a mixed linear-morphological network called **HaarNet**. This is the first large-scale linear-morphological hybrid, evaluated on a set of sizeable real-world datasets. In the network, morphological Haar sampling is applied to both feature channels in several layers, which splits extreme values and high-frequency information such that both can be processed to improve both modalities. Moreover, morphologically parameterised ReLU is used, and morphologically-sound up-sampling is applied to obtain a full-resolution output. Experiments show that HaarNet is competitive with a state-of-the-art CNN, implying that morphological networks are a promising research direction for geometry-based learning tasks.

Keywords: Machine learning · Morphological neural networks · Multi-modal learning

1 Introduction

Semantic segmentation is a challenging pixel-level prediction task in computer vision. It has become common practice [15,31] to use multi-modal *colour* and *depth* (RGD-D) data since failure modes arising from one modality (*e.g.* homogeneity, noise, sparsity, etc.) can often be avoided when relying on multiple modalities. Datasets such as [1,8,31] facilitate research that delves into combining features from different signals by providing data from different sensors. However, optimal fusion strategies for the colour and depth data modalities are still an open area of research.

Through the works of Serra [28], the computer vision community is familiarised with the underlying algebraic structure of data that is acquired using

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2024
S. Brunetti et al. (Eds.): DGMM 2024, LNCS 14605, pp. 242–254, 2024.
https://doi.org/10.1007/978-3-031-57793-2_19

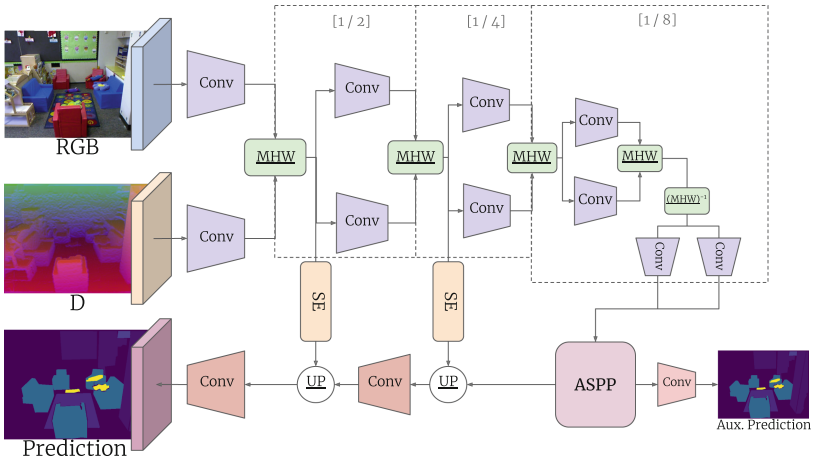

Fig. 1. The Morphological HaarNet. Depth and RGB modalities are separately encoded, but during down-sampling the Morphological Haar Wavelet (MHW, see Fig. 2 for more details) blocks ensure that the down-sampled signal is improved using high-frequency details from both modality feature signals. Up-sampling happens through morphological up-sampling and skip connections (UP). Both morphological types are underlined in this overview. The skip connections receive channel-wise attention from (multi-modal) SE [20] blocks. Multi-resolution features are combined using ASPP [5]. All non-linearities are replaced by the morphological dilation.

probing contact (*e.g.* LiDAR and radar) [26,29,32]. This field is called *mathematical morphology*, and it differs from the algebra of linear diffusion that is used to build convolutional neural networks (CNNs). Recently, it was shown that morphological operations are suitable for processing specific types of data, such as depth data [13]. Besides the benefits for encoding input signals, morphological sampling and interpolation may help to predict sharply delineated semantic boundaries. While there exist examples of methods that explore morphological networks [23], or embed morphological operations in few-layer linear networks [10,11,19,27,30,35], large-scale linear-morphological hybrids have not received much research attention.

This paper proposes the construction of a mixed linear-morphological network, coined HaarNet, to exploit linear and morphological operations jointly. The proposed morphological modules are: a down-sampling operation based on the morphological Haar transform; non-linearities formulated as parameterised dilations; and an up-sampling procedure that can innately delineate semantic boundaries.

2 Related Work

Fusion of RGB and depth images in neural networks may be done at the input or output stage, although the most common method is “middle” fusion. Middle

fusion can be either operator-based or attention-based. Attention-based fusion ensures that long-range dependencies between modalities can be modelled, rather than simply fusing on local modality-based features as operator-based fusion would do. While there are many articles addressing this type of fusion, of note is [7], who proposes to re-calibrate both feature encodings using Separation-and-Aggregation (SAGate) at each down-sampling step in the encoder. Moreover, a propagation module is introduced that normalises information flow, allowing the authors to make better use of pre-trained weights. In [6], the authors observe that alignment of features from RGB and depth is not often addressed, and propose the Global-Local Propagation network that makes use of attention blocks to warp and align feature volumes. CANet [40] uses a combination of channel-wise and spatial attention mechanisms to fuse depth and RGB using separate pathways.

The method proposed in the current paper, too, considers the attention-by-sampling, but in a more structured way through the morphological Haar transform. Most similar to our proposed method is the MSFNet introduced by [22]. In MSFNet, a *linear* Haar Wavelet transform is used to decompose high resolution signals into constituents, ensuring high frequency information is preserved throughout down-sampling. However, instead of using the wavelets as a means of down-sampling, the inverse transform is used to reconstruct the feature signal at the original resolution. In addition, the depth features are only used to attend the RGB features, so no joint optimisation of feature representations is being performed. HaarNet, on the other hand, allows to freely optimise the set of features from both modalities.

3 Method

To specify the composition of a Morphological Haar Wavelet Network, a brief overview of the fundamental morphological operations is required: the morphological *dilation* is defined on the semi-ring $\{\mathbb{R}_{-\infty}, \bigvee, +\}$ where $\bigvee$ denotes the supremum operation and $+$ is addition. This algebraic system extends the set of real numbers $\mathbb{R}$ with minus infinity: $\mathbb{R}_{-\infty} \equiv \mathbb{R} \cup -\infty$ [26]. The morphological *erosion* is the dual of the morphological dilation, and is constructed with infimum $\bigwedge$ over the extended set of reals $\{\mathbb{R} \cup \infty\}$. The dilation and erosion can be parameterised by means of a structuring element h. The dilation is defined as

$$g(\mathbf{x}) = [f \boxplus h](\mathbf{x}) = \bigvee_{\mathbf{z}} f(\mathbf{x} - \mathbf{z}) + h(\mathbf{z}) \ , \tag{1}$$

where f, g are the inputs and outputs of a layer in a neural network respectively, and $\mathbf{x}, \mathbf{z}$ are taken from indicator sets. Unlike convolutions in CNNs, morphological operations are often only applied in a channel-wise manner [13]; back-propagation over the maximum element results in slow and complicated learning [14]. Note that dilation is algebraically similar to convolution, though in the *logarithmic* sense [2]. As a consequence, it can serve as a signal processing framework with an essentially analogous structure to linear signal processing.

There are four novel elements on which the morphological HaarNet is built:

- The morphological Haar Wavelet module, which improves down-sampling compared to common strategies like strided convolutions or pooling.
- The ReLU non-linearity that is shown to be a special case of the morphological dilation; it can be parameterised to be a learnable threshold-filter.
- An improved sampling procedure relative to [13], following seminal papers on sampling theory in morphology [17,18] rather than being provenance-based.
- The use of additional pre-training to compensate for the reduced effectiveness of transfer learning when compared to baselines similar to a pre-trained ResNet [16].

The full network overview is shown in Fig. 1, which details at which stages multi-modal features are fused. Note that the network uses Squeeze-and-Excitation blocks [20] for attending channel-wise information during skip connections, although they are modified to attend separately to each modality feature volume before feature fusion. Another module that is used, and that is prevalent in pixel-level semantic prediction networks, is the multi-scale feature pyramid ASPP [5] to improve predictive performance for long-range semantic dependencies. These blocks are naturally augmented by morphological non-linearities (see Subsect. 3.2), to truly make this network a linear-morphological hybrid.

3.1 Morphological Haar Wavelets

The morphological Haar wavelet transform decomposes signal $f^\sigma\colon \mathbb{R}^{H\times W} \to \mathbb{R}$ at a particular scale σ into four subband signals: Down-sampled $f^{\sigma/2}\colon \mathbb{R}^{H/2\times W/2} \to \mathbb{R}$ and vertical, horizontal, and diagonal details $\phi_f^v, \phi_f^h, \phi_f^d\colon \mathbb{R}^{H/2\times W/2} \to \mathbb{R}$ respectively. A particularly helpful property of morphological wavelets over linear wavelets is that they retain local maxima of the signal over multiple scales; similar to pooling, this implies the network achieves invariance to particular kinds of noise [21]. The morphological Haar wavelet transform [18] is defined

$$f^{\sigma/2} = f^\sigma \boxplus \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}, \tag{2a}$$

$$\phi_f \equiv \left\{\phi_f^v, \phi_f^h, \phi_f^d\right\} = f^\sigma * \left\{ \begin{bmatrix} -1 & -1 \\ 1 & 1 \end{bmatrix}, \begin{bmatrix} -1 & 1 \\ -1 & 1 \end{bmatrix}, \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \right\}, \tag{2b}$$

where both morphological dilation $\boxplus$ and convolution $*$ are applied at a stride of 2 to get non-overlapping windows, and to ensure that the spatial dimensions of the signal f^σ are reduced by a factor 2. Clearly, $f^{\sigma/2}$ is similar to the application of max pooling which computes the local maximum value in a 2-by-2 neighbourhood. Max pool and $f^{\sigma/2}$ both reduce spatial feature size in addition to obtaining maximum-amplitude coefficients [4,36]. On top of this, detail signals ϕ encode all directional high-frequency information that was lost during down-sampling. A property of the wavelet transform is that it is possible to fully restore f by the inverse transform using subbands $f^{\sigma/2}$ and ϕ_f.

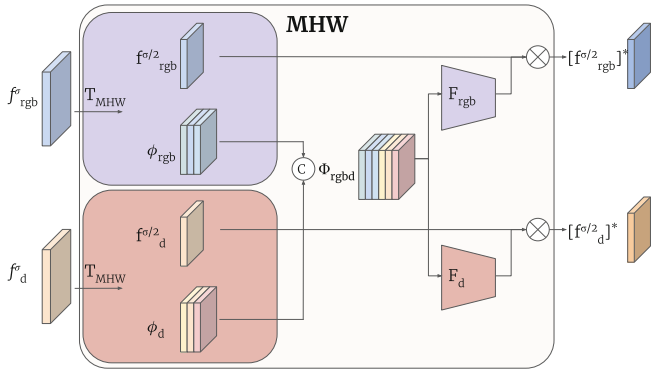

Fig. 2. The Morphological Haar Wavelet (MHW) block, which down-samples the signal f^{σ} by decomposing into subbands $f^{\sigma/2}$ and ϕ. A 2-layer MLP F is used to attend to channel-wise information based on combined high-frequency details Φ_{rgbd} for both modalities. This yields updated feature representations $\left[f^{\sigma/2}_{rgb}\right]^*$ and $\left[f^{\sigma/2}_{d}\right]^*$.

The idea of the Morphological Haar Wavelet (MHW) block is to use combined directional details of RGB and depth modalities $\Phi_{rgbd} \equiv \{\phi_{rgb}, \phi_d\}$ to simultaneously improve both the down-sampled RGB feature signal $f^{\sigma/2}_{rgb}$ and the depth feature signal $f^{\sigma/2}_{d}$. To this end, MHW uses two distinct 2-layer CNNs F_{rgb}, F_d. Each F takes as input the combined high-frequency signal Φ_{rgbd} to obtain the updated and improved down-sampled signals $[f^{\sigma/2}]^*$. For RGB features this can be expressed as

$$\left[f^{\sigma/2}_{rgb}\right]^* (\mathbf{x}) = f^{\sigma/2}_{rgb}(\mathbf{x}) \otimes \sigma\left(F_{rgb}\left(\Phi_{rgbd}\left(\mathbf{x}\right)\right)\right) , \tag{3}$$

where $\sigma(\cdot)$ is a sigmoid, $\otimes$ is element-wise multiplication, and $\mathbf{x}$ taken from the indicator set over the signal. Similarly, for the improved depth features $[f^{\sigma/2}_{d}]^*$. The combined operation ensures not only that high frequency information is used explicitly during down-sampling, but also forces the resulting information to be dependent on both modalities through their common Φ_{rgbd}. The full procedure is depicted schematically in Fig. 2.

3.2 Formulating ReLU as Dilation

The ReLU activation layer [12] is a special case of the morphological dilation with a flat structuring element that is the morphological delta-function, and a morphological bias term. The morphological delta-function $\delta(\cdot)$ is 0 only at location 0, and $-\infty$ elsewhere. To demonstrate the equivalence between ReLU and morphological dilation, first consider the definition of the ReLU:

$$\xi(\mathbf{y}) = \begin{cases} \mathbf{y} & \text{if} \quad \mathbf{y} \geq 0 \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

Similar to the convolutional bias, the dilation can be written to include a morphological bias term [3], denoted h_0. First, consider the a specific instance of the morphological dilation:

$$\xi\left(f\left(\mathbf{x}\right)\right) = 0 \vee f\left(\mathbf{x}\right) = 0 \vee \left[f \boxplus \delta\right]\left(\mathbf{x}\right), \tag{5}$$

using the morphological delta-function $\delta(\cdot)$. The above can be generalised by setting $h_0 = 0$ and allowing an arbitrary structuring element h:

$$\xi\left(f\left(\mathbf{x}\right)\right) = h_0 \vee \left[f \boxplus h\right]\left(\mathbf{x}\right). \tag{6}$$

Thus, a form is reached that suggests a natural extension of ReLU by morphological kernel h and bias h_0. In HaarNet, all activation layers are replaced by the morphological dilation in which the morphological bias term h_0 is learnable. Such a learnable ReLU was coined Flexible ReLU (FReLU) in [25], but the connection to the morphological dilation was not recognised in this work. A strong belief is held by the authors that understanding of networks improves by recognising that nearly all modern CNNs exploit inherently morphological layers through ReLU [34] and pooling [13].

3.3 Morphological Up-Sampling

In encoder-decoder architectures, up-sampling from low-resolution features is vital to obtain high-quality predictions. In [13], it was shown that non-linear up-sampling can aid semantic networks, since it delineate boundaries better than linear up-sampling. The procedure also did not introduce sparsity the way standard unpooling does [38,39]. However, morphological up-sampling can be improved using insights from [18], where it is proven by the theory of adjunctions that the down-sampling dilation has an up-sampling dual that is an erosion with the same structuring element. In addition, up-sampling is of equidistant nature: intervals at which to place back values are kept constant. As a consequence, the total reconstruction operator ρ is a closing: $\rho[f] = [f \boxplus h] \boxminus h$. Note, however, that in prediction tasks *reconstruction* of the same signal is not the goal, rather the input signal must be *processed* for analysis. Unlike [18], this paper is concerned with obtaining the most expressive sampling processing operator.

Based on experimentation, the proposed morphological up-sampling operator for prediction networks is as follows: First, equidistant sampling is used, and not provenance mapping as in [13]. This means low-resolution features are placed back at a regular interval. Second, all sampling operators are dilations. Third, structuring elements are not constrained to be the same across down- and up-sampling operators. That is, they are all parameterised individually. Lastly, in HaarNet, the up-sampling procedure is combined with a skip connection at the same resolution.

4 Experiments

In this section, HaarNet is applied to an RGB-D semantic segmentation task on three different datasets, both indoor and outdoor. HaarNet is compared to

most notably SAGate [7] as it is encoder-decoder-based, reaches state-of-the-art consistently, and public code is available for replication of the results and for a fair comparison. Large-scale linear-morphological hybrids have not received research attention; the goal of the experiments is to show that even when network design principles from linear networks are used, linear-morphological hybrids still perform at least on par with linear counterparts.

Implementation. All modules are implemented in PyTorch [24], except for morphological operations, which are absent in PyTorch and potentially intractable when naively done. These operations are implemented directly in CUDA/C++. Source code is made available at https://github.com/rickgroen/haarnet.

Dataset and Training. The experiments are performed on three datasets: NYUv2 ($N_{\text{train}} = 795$, $N_{\text{test}} = 654$, $C = 40$), 2D-3D-S ($N_{\text{train}} = 52903$, $N_{\text{test}} = 17593$, $C = 13$), and CityScapes ($N_{\text{train}} = 2975$, $N_{\text{test}} = 500$, $C_{train} = 30$, $C_{test} = 19$). The first two datasets are indoor datasets, the last is an outdoor autonomous driving dataset. RGB images are normalised with training set statistics to follow a zero-mean Gaussian. According to convention, depth images are converted to the HHA encoding [15] and normalised in the range $[0, 1]$. Based on initial experimentation, all networks are trained using SGD with Nesterov Momentum [33] with a learning rate λ_0 of 5e−3. The learning rate is updated at each epoch by a polynomial scheduler according to $\lambda_{epoch} = \lambda_0 \left(1 - \frac{epoch}{500}\right)^{0.9}$. The networks are trained for 500, 12, and 200 epochs for the NYU, 2D-3D-S, and CityScapes datasets respectively. Models are evaluated using mean Intersection over Union (mIoU), pixel accuracy, and a boundary F1-score [9] to measure performance at semantic edges.

4.1 The Necessity of Pre-training

It is hypothesised that many contemporary networks are architecturally biased to the pre-trained ResNet weights on ImageNet, since pre-training is such a powerful means to obtain state-of-the-art performance [37]. Specifically, when the architecture starts to significantly differ from that of the ResNet, exploiting the pre-trained weights becomes less valuable. When a fundamentally different network, like HaarNet, is introduced, it should undergo pre-training for fair assessment of its novel capabilities.

Pre-training new encoders on ImageNet, however, is not computationally feasible in most scenarios because of the scale of this dataset. To adapt, this paper joins the three datasets (*i.e.* NYUv2, 2D-3D-S, and CityScapes) and maps all semantic labels to a joint set of 67 labels, including semantic classes from both indoor and outdoor scenes. Pre-training is performed in two phases: first, the decoder is frozen and the encoder is trained using the joint dataset. Second, the decoder is trained on a specific dataset with a frozen encoder.

Consider Table 1, in which the pre-training hypothesis is tested. All networks use a ResNet50 encoder, although both the SAGate baseline and HaarNet

Table 1. Effects of Pre-trained Weights. ResNet50 encoders are initialised either randomly or use weights pre-trained from ImageNet. HaarNet performs better on two out of three metrics when randomly initialised. When ImageNet weights are used, a basic ResNet benefits most.

	mIoU	pixel accuracy	Boundary F1
Random Initialisation			
Middle Fusion Dual ResNet	0.193	0.582	0.220
SA-Gate [7]	**0.266**	0.603	0.213
HaarNet	0.252	**0.624**	**0.231**
Naive Transfer of ImageNet Weights			
Middle Fusion Dual ResNet	0.382	0.687	**0.314**
SA-Gate [7]	**0.408**	**0.697**	0.267
HaarNet	0.360	0.687	0.292

Table 2. Indoor Segmentation Results. When trained using the same set-up, HaarNet performs at least on par with the baselines using the least amount of additional parameters, as reported in the second column.

Network		NYUv2			2D-3D-S		
	+#param	mIoU	pixel accuracy	Boundary F1	mIoU	pixel accuracy	Boundary F1
Dual ResNet Baseline	+25.8M	0.474	0.749	0.373	0.565	0.787	0.564
SAGate [7]	+21.8M	0.504	0.768	0.381	0.561	**0.806**	**0.577**
HaarNet	+13.1M	**0.507**	**0.770**	**0.392**	**0.568**	0.794	0.567

make significant changes to their encoder. SAGate compensates for architectural changes by the normalising the outputs of their fusion blocks; HaarNet has no such inherent rectification. It is shown that without any pre-training, HaarNet performs best on two out of three metrics. However, when ImageNet weights for ResNet are used as a pre-training procedure, HaarNet benefits least; a basic ResNet benefits most.

4.2 Segmentation Performance

Indoor Semantic Segmentation. The predictive performance of HaarNet on indoor segmentation datasets is shown in Table 2. For NYUv2, HaarNet outperforms the general baseline and state-of-the-art SAGate on all metrics, provided the networks are trained on the same system, using the same pre-training, and with the same hyper-parameters. HaarNet is also more efficient, given that it only uses 13.1M additional parameters on top of the encoder, versus 20-25M for the other baselines. For 2D-3D-S, performance is on par with SAGate. The main difference between NYUv2 and 2D-3D-S, apart from size, is the quality of the depth signals. To see so, review the qualitative results, as depicted in Fig. 3, in which the relative best and worst predictions of HaarNet versus SAGate are shown. From the centre rows, it can be seen that the depth signals of 2D-3D-S contain many reconstruction artefacts. Since morphological operations are more

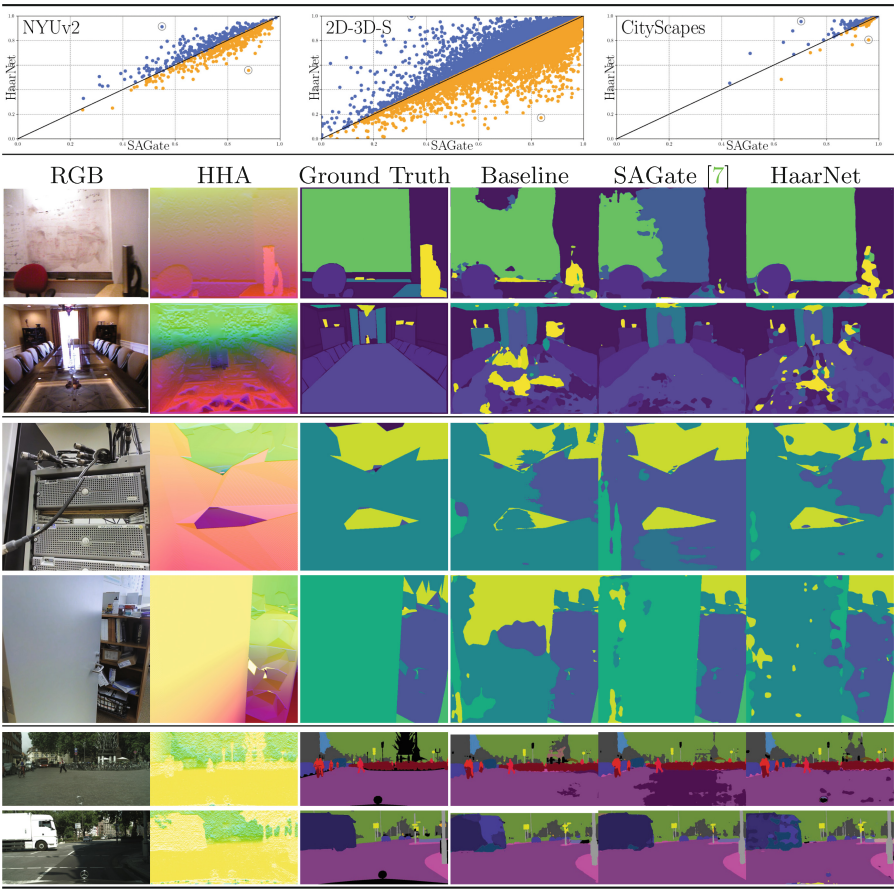

Fig. 3. Qualitative Results of HaarNet. The top three scatter plots show accuracy per image for HaarNet versus SAGate, and select the relative best and worst sample. These samples from NYUv2, 2D-3D-S and Cityscapes are shown in the rows with images, interweaving relative best (top, HaarNet outperforms SAGate) and worst (bottom, vice versa) samples. In the second row (NYUv2, worst), a reflective surface results in a low-quality depth signal. The morphological HaarNet misclassifies at those depth artefacts, although it does recover the lamp situated on the table, which is not labelled as lamp. The third and fourth row show that 2D-3D-S has relatively coarse depth and semantics, complicating learning for all networks. Here, the third best and tenth worst images are shown, since the other failure modes are completely homogeneous images. Finally, CityScapes images show that HaarNet is able to classify the road mostly correctly. However, the monochromatic truck poses a challenge. Best viewed in colour.

sensitive to geometric modalities, it is possible the depth artefacts complicate learning for HaarNet.

Table 3. Outdoor Segmentation Results. Unlike results from indoor datasets, HaarNet significantly outperforms the baselines. It could be this is due to the high amount of structure present in the scenes, which allows absolute operators like morphological ones to make more rigid predictions.

Network	mIoU	pixel accuracy	Boundary F1
Dual ResNet Baseline	0.709	0.952	0.714
SAGate [7]	0.713	0.953	0.728
HaarNet	**0.762**	**0.961**	**0.780**

Table 4. Ablation Study. Rows show performance due to the morphological up-sampling layers (M-UP), morphological non-linearities (M-ReLU), and the Haar module (MHW). Modules are replaced with simple (linear) counterparts when possible. Results indicate that, while networks adapt to the available modules, a network that is a full linear-morphological hybrid performs best.

M-UP	M-ReLU	MHW	mIoU	pixel accuracy	Boundary F1
			0.469	0.751	0.373
✓			0.475	0.752	0.367
✓	✓		0.469	0.753	0.371
✓		✓	0.481	0.755	0.363
✓	✓	✓	**0.510**	**0.771**	**0.385**

Outdoor Semantic Segmentation. HaarNet is also evaluated on the outdoor CityScapes dataset; results are shown in Table 3. Here, HaarNet significantly outperforms the baselines. This may be because autonomous driving datasets are much less varied than indoor datasets since traffic is structured to be predictable. Morphological operators are more calibrated than linear operators, since they act absolutely rather than relatively.

The performance due to three core modules of HaarNet are evaluated: morphological up-sampling layers, morphological non-linearities (Section 3.2), and the morphological Haar Wavelet (MHW) module (Section 3.1). Results are shown in Table 4. In the ablation study, the base class without any of the introduced modules is architecturally equivalent to a DeeplabV3 [5]. All networks exploit ResNet-101 backbones, and are trained for 100 epochs on NYUv2 40-class setting. Results indicate that the Haar module is important for achieving best performance, since it enables modality fusion at the same time as informed down-sampling. The ablation also shows that networks are flexible: when modules are replaced by simple (linear) counterparts, the network can come to depend on the reliable DeepLab architecture.

5 Conclusions

It is the authors' strong belief that the set of operators made available to a network should reflect the admissible transformations on the input modalities. Morphological operations are more suitable to analysis of geometry-based modalities

than convolutions are. In this paper, all non-linearities and pooling are replaced by their parameterised morphological equivalents, which entails natural and structural generalisation of core network modules. Leveraging these generalised modules, resulting from the field of mathematical morphology, yields promising results, even when core network design principles are kept similar to those of CNNs. This was achieved by utilising three core MM concepts and introducing a new pre-training regime, resulting in the large-scale linear-morphological hybrid network called **HaarNet**. Further research should address attention mechanisms that enable networks to focus on specific features using modality-specific operators. This should be explored both across scales and along feature depth; the latter has yet to be investigated in the literature.

References

1. Armeni, I., Sax, S., Zamir, A.R., Savarese, S.: Joint 2d-3d-semantic data for indoor scene understanding. arXiv preprint arXiv:1702.01105 (2017)
2. Burgeth, B., Weickert, J.: An explanation for the logarithmic connection between linear and morphological system theory. IJCV **64**, 157–169 (09 2005)
3. Charisopoulos, V., Maragos, P.: Morphological perceptrons: geometry and training algorithms. In: ISMM. pp. 3–15. Springer (2017)
4. Chen, B., Polatkan, G., Sapiro, G., Blei, D., Dunson, D., Carin, L.: Deep learning with hierarchical convolutional factor analysis. TPAMI **35**(8), 1887–1901 (2013)
5. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. TPAMI **40**(4), 834–848 (2017)
6. Chen, S., Zhu, X., Liu, W., He, X., Liu, J.: Global-local propagation network for rgb-d semantic segmentation. arXiv preprint arXiv:2101.10801 (2021)
7. Chen, X., Lin, K.Y., Wang, J., Wu, W., Qian, C., Li, H., Zeng, G.: Bi-directional cross-modality feature propagation with separation-and-aggregation gate for rgb-d semantic segmentation. In: ECCV. pp. 561–577. Springer (2020)
8. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: CVPR (2016)
9. Csurka, G., Larlus, D., Perronnin, F., Meylan, F.: What is a good evaluation measure for semantic segmentation? TPAMI **26**(1) (2004)
10. Dimitriadis, N., Maragos, P.: Advances in the training, pruning and enforcement of shape constraints of morphological neural networks using tropical algebra. arXiv preprint arXiv:2011.07643 (2020)
11. Franchi, G., Fehri, A., Yao, A.: Deep morphological networks. Pattern Recogn. **102**, 107246 (2020)
12. Fukushima, K.: Cognitron: A self-organizing multilayered neural network. Biol. Cybern. **20**(3–4), 121–136 (1975)
13. Groenendijk, R., Dorst, L., Gevers, T.: Morphpool: Efficient non-linear pooling & unpooling in cnns. In: 33rd British Machine Vision Conference 2022, BMVC 2022, London, UK, November 21-24, 2022. BMVA Press (2022)
14. Groenendijk, R., Dorst, L., Gevers, T.: Geometric back-propagation in morphological neural networks. TPAMI pp. 1–8 (2023)

15. Gupta, S., Girshick, R., Arbeláez, P., Malik, J.: Learning rich features from rgb-d images for object detection and segmentation. In: ECCV. pp. 345–360. Springer (2014)
16. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016)
17. Heijmans, H.J., Goutsias, J.: Nonlinear multiresolution signal decomposition schemes. ii. morphological wavelets. TIP **9**(11), 1897–1913 (2000)
18. Heijmans, H.J., Toet, A.: Morphological sampling. CVGIP: Image understanding **54**(3), 384–400 (1991)
19. Hernández, G., Zamora, E., Sossa, H., Téllez, G., Furlán, F.: Hybrid neural networks for big data classification. Neurocomputing **390**, 327–340 (2020)
20. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: CVPR. pp. 7132–7141 (2018)
21. Jarrett, K., Kavukcuoglu, K., Ranzato, M., LeCun, Y.: What is the best multi-stage architecture for object recognition? In: ICCV. pp. 2146–2153. IEEE (2009)
22. Jiang, S., Xu, Y., Li, D., Fan, R.: Multi-scale fusion for rgb-d indoor semantic segmentation. Sci. Rep. **12**(1), 20305 (2022)
23. Mondal, R., Santra, S., Mukherjee, S.S., Chanda, B.: Morphological network: How far can we go with morphological neurons? In: BMVC. BMVA Press (2022)
24. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: Pytorch: An imperative style, high-performance deep learning library. In: NeuRIPS, pp. 8024–8035. Curran Associates, Inc. (2019)
25. Qiu, S., Xu, X., Cai, B.: Frelu: flexible rectified linear units for improving convolutional neural networks. In: ICPR. pp. 1223–1228. IEEE (2018)
26. Ritter, G.X., Sussner, P.: An introduction to morphological neural networks. In: ICPR. vol. 4, pp. 709–717. IEEE (1996)
27. Roy, S.K., Mondal, R., Paoletti, M.E., Haut, J.M., Plaza, A.: Morphological convolutional neural networks for hyperspectral image classification. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing **14**, 8689–8702 (2021)
28. Serra, J.: Image analysis and mathematical morphology (1983)
29. Serra, J., Vincent, L.: An overview of morphological filtering. Circuits Systems Signal Process. **11**(1), 47–108 (1992)
30. Shen, Y., Shih, F.Y., Zhong, X., Chang, I.C.: Deep morphological neural networks. PRAI **36**(12), 2252023 (2022)
31. Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor segmentation and support inference from RGBD images. ECCV **7576**, 746–760 (2012)
32. Sussner, P.: Morphological perceptron learning. In: Proceedings of the 1998 ISIC/CIRA. pp. 477–482. IEEE (1998)
33. Sutskever, I., Martens, J., Dahl, G., Hinton, G.: On the importance of initialization and momentum in deep learning. In: ICML. pp. 1139–1147. PMLR (2013)
34. Velasco-Forero, S., Angulo, J.: Morphoactivation: Generalizing relu activation function by mathematical morphology. In: DGMM. pp. 449–461. Springer (2022)
35. Velasco-Forero, S., Rhim, A., Angulo, J.: Fixed point layers for geodesic morphological operations. In: BMVC. BMVA Press (2022)
36. Xie, L., Tian, Q., Wang, M., Zhang, B.: Spatial pooling of heterogeneous features for image classification. TIP **23**(5), 1994–2008 (2014)
37. Yosinski, J., Clune, J., Bengio, Y., Lipson, H.: How transferable are features in deep neural networks? NeuRIPS **27** (2014)

38. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: ECCV. pp. 818–833. Springer (2014)
39. Zeiler, M.D., Taylor, G.W., Fergus, R.: Adaptive deconvolutional networks for mid and high level feature learning. In: ICCV. pp. 2018–2025. IEEE (2011)
40. Zhou, H., Qi, L., Huang, H., Yang, X., Wan, Z., Wen, X.: Canet: Co-attention network for rgb-d semantic segmentation. Pattern Recogn. **124**, 108468 (2022)