Enhancing Material Features Using Dynamic Backward Attention on Cross-Resolution Patches (Supplementary Material)

Yuwen Heng
https://www.ecs.soton.ac.uk/people/yh1n19
Yihong Wu
https://www.ecs.soton.ac.uk/people/yw10y19
Srinandan Dasmahapatra
https://www.ecs.soton.ac.uk/people/srinanda
Hansung Kim
https://www.ecs.soton.ac.uk/people/hk1f20

The VLC Group
University of Southampton
Southampton, UK

1 Training Configuration

As for the training process, we notice that networks may not perform well on the LMD [3]. One possible reason is the Adam optimiser, which is proved to impair the generalisation ability of networks [12]. To improve the network performance, three advanced technologies were adopted: the AdamW optimiser [6], the linear learning rate warm-up [7], and the polynomial learning rate decay [1]. Furthermore, as the LMD images have various sizes, we crop the image into 512×512 patches to facilitate batch normalisation [9]. Equipped with these training methods, we successfully boost the Pixel Acc to above 85% from below 70% [3].

2 Per-Category Performance Analysis

This section describes the performance analysis of models on the LMD for each material category. Our DBAT achieves the highest accuracy (averaged across five runs) for four materials, fabric, soil, stone, and wood. Compared with the second best accuracy for these materials, DBAT increases the accuracy by 2.39%, 4.28%, 13.63% and 2.36% respectively. For foliage, metal, plastic, rubber, and water, the performance of DBAT is close to the second best accuracy, with decrements of -0.22%, -1.6%, -1.04%, -0.57%, and -0.34%.

3 Centered Kernel Alignment

The entity of the CKA matrix is calculated with Equation 1, where HSIC_1 stands for the unbiased estimator of the Hilbert-Schmidt independence criterion [10]. Since $\text{HSIC}_1 = 0$ only when X and Y are independent, and CKA is invariant to isotropic scaling, they together

© 2021. The copyright of this document resides with its authors.
It may be distributed unchanged freely in print or electronic forms.

Model	DBAT	ResNet-152	ResNeSt-101	EfficientNet-B5	Swin-t	CAM-SegNet-DBA
Asphalt	88.66 ± 0.72	88.66 ± 0.17	94.35 ± 0.27	82.17 ± 2.80	91.83 ± 1.09	89.87 ± 1.94
Ceramic	68.31 ± 1.31	65.29 ± 3.19	62.86 ± 0.67	73.34 ± 0.42	75.35 ± 0.42	75.01 ± 0.64
Concrete	66.90 ± 1.07	50.89 ± 1.67	60.53 ± 2.00	59.36 ± 2.98	57.42 ± 4.88	69.20 ± 2.81
Fabric	93.14 ± 0.16	85.53 ± 0.22	86.420 ± 0.92	85.33 ± 0.20	88.71 ± 0.50	90.79 ± 0.43
Foliage	95.35 ±0.12	93.55 ± 0.33	91.25 ± 1.16	88.21 ± 0.32	95.57 ± 0.45	94.04 ± 0.79
Food	93.27 ± 0.22	90.27 ± 0.22	94.96 ± 0.34	95.84 ± 0.14	92.51 ± 0.83	95.19 ± 0.24
Glass	73.27 ± 0.67	72.58 ± 2.50	68.33 ± 0.34	77.83 ± 0.94	77.95 ± 0.99	84.88 ± 1.11
Metal	79.99 ± 0.51	75.35 ± 0.94	80.66 ± 0.34	76.67 ± 0.28	81.54 ± 1.36	81.83 ± 0.48
Paper	73.83 ± 0.67	64.52 ± 2.87	71.14 ± 1.99	77.21 ± 0.13	63.05 ± 1.90	66.48 ± 1.43
Plaster	71.43 ± 0.71	68.01 ± 0.53	78.76 ± 0.62	73.11 ± 0.64	78.12 ± 1.90	72.37 ± 1.03
Plastic	50.62 ± 1.45	34.87 ± 1.21	36.07 ± 3.42	39.59 ± 0.64	51.64 ± 1.31	52.07 ± 2.28
Rubber	82.61 ± 1.01	77.08 ± 3.61	79.57 ± 1.62	69.73 ± 0.29	83.48 ± 0.67	81.63 ± 1.79
Soil	84.25 ± 0.50	73.27 ± 1.63	73.15 ± 2.67	79.73 ± 0.55	76.89 ± 1.11	80.39 ± 1.73
Stone	86.94 ± 0.95	69.66 ± 1.42	52.12 ± 0.93	70.07 ± 0.76	73.05 ± 1.92	60.73 ± 2.76
Water	97.12 ± 0.10	95.49 ± 0.33	97.54 ± 0.28	95.30 ± 0.32	95.78 ± 0.70	94.95 ± 0.69
Wood	90.53 ± 0.37	76.05 ± 1.08	76.71 ± 1.23	86.69 ± 0.24	82.03 ± 1.11	87.63 ± 0.98
PixelAcc	86.85 ± 0.08	80.68 ± 0.11	82.45 ± 0.20	83.17 ± 0.06	84.71 ± 0.26	86.12 ± 0.15
MeanAcc	81.05 ± 0.28	73.87 ± 0.25	75.31 ± 0.29	76.91 ± 0.06	79.06 ± 0.46	79.85 ± 0.28

Table 1: Per-category performance analysis. The networks are trained five times to report the uncertainty.

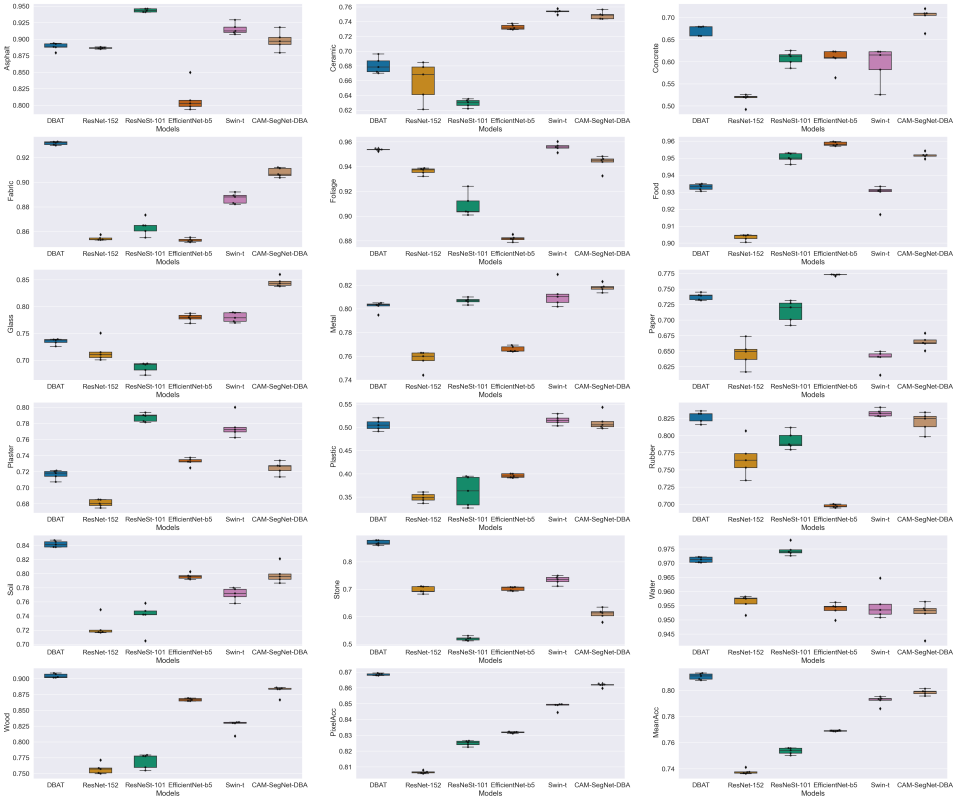
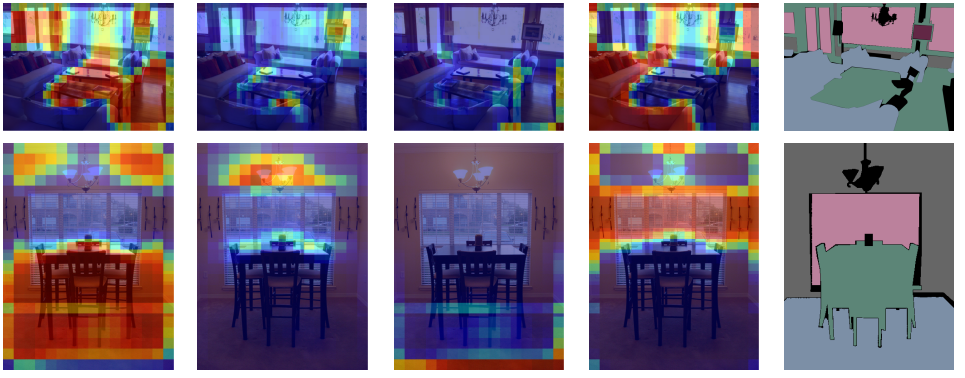


Figure 1: Boxplot of the performance on the LMD across five runs.



$Attn_1$ $Attn_2$ $Attn_3$ $Attn_4$ GT

Figure 2: Attention mask visualisation. GT: ground truth. The densely labelled ground truth images are collected from [3].

give a meaningful comparison of two networks with different architectures [4]. More detail about the CKA matrix is included in the supplementary material.

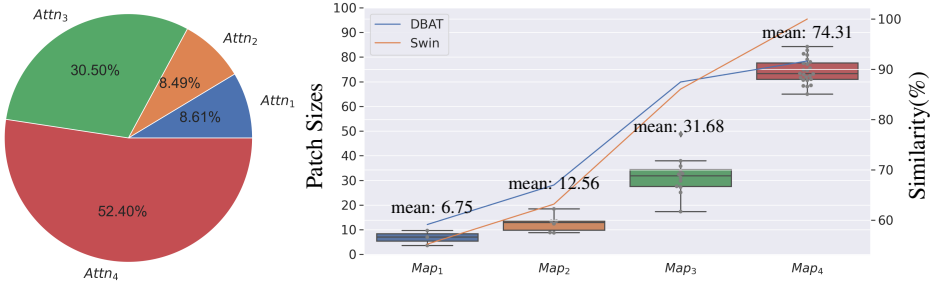
$$\mathrm{CKA}(X,Y) = \frac{\mathrm{HSIC}_1(XX^T, YY^T)}{\sqrt{\mathrm{HSIC}_1(XX^T)\mathrm{HSIC}_1(YY^T)}} \tag{1}$$

4 Attention Analysis

This section analyses the dynamic attention module by visualising the attention masks and calculating their descriptive statistics, including the average attention weights as well as the equivalent attention patch resolutions. Figure 2 shows the attention masks for images in the LMD test set. From $Attn_1$ to $Attn_4$, the patch resolution increases. It is discovered that the material covering multiple small objects or a small area tends to depend on the features extracted from small patch resolution. For example, the first column images in Figure 2 highlight the regions that the network concentrates on. The wooden area in the first row covers both the desk and the floor. The wooden chairs and the fabric floor are mutually overlapping in the second row. These objects or material regions are mutually overlapping, and small patches can isolate them and learn features at the boundaries.

Figure 3 shows the average attention weight in (a) and the equivalent patch size in (b, the box plot). The equivalent patch size, which represents the attention distance of each attention head, is calculated by transforming the attention diagonal distance in [8] to the side length of a square. As expected, the aggregated features mostly (52.40%) depend on $Attn_4$, which is thoroughly processed by the whole backbone encoder, with an average patch size of 74.31. Apart from $Attn_4$, the aggregated features also depend on feature maps extracted from small patch sizes. For example, $Attn_3$ is extracted from an average patch size of 31.68, and it contributes 30.50% to the aggregated features. Although $Attn_1$ is gathered from a shallow stage of the network, the aggregated features still depend on it to handle the overlapping material regions with a patch size of 6.75 on average.

To further illustrate the effect of the dynamic backward attention module, the similarity



(a) Averaged Attention Weights

(b) Equivalent Attention Patch Sizes

Figure 3: The descriptive analysis of attention weights.

scores comparing one layer with $Map_{1,2,3,4}$ from the CKA matrix is reported in Figure 3. (b, the line plot). The blue line compares Map_4 from the Swin encoder, and the orange line compares the aggregated features from the DBAT. The increased similarity scores against $Map_{1,2,3}$ clearly show that the aggregated feature depends on information from shallow layers.

5 Qualitative Analysis

This section shows the predicted material segmentation for three images. In Figure 4, ResNet-152 [7], ResNest-101 [13], Swin-t [5], and CAM-SegNet-DBA [3] segment the bed as fabric. However, the floor appears to be covered with a carpet whose material looks like fabric. The proposed DBAT and the EfficientNet-b5 [11] break the object boundary and segment part of the floor as fabric. Moreover, the DBAT segmented image have less flying pixels compared with the EfficientNet-b5. This indicates that the DBAT segments the images based on generalisable material features that are not related to contextual information such as objects.

The segmented materials in Figure 5 and Figure 6 provide more evidence that the DBAT can segment the images well with features extracted from cross-resolution patches. In Figure 5, the boundary between the wooden window frame and glass made windows in DBAT segmented image is more adequate than the segments predicted by other networks. In Figure 6, the segmented fabric aligns well with the ground truth with no noisy predictions. Considering the training of DBAT takes sparsely labelled segments, it is reasonable to say that DBAT learns the difference between materials from cross-solution patches.

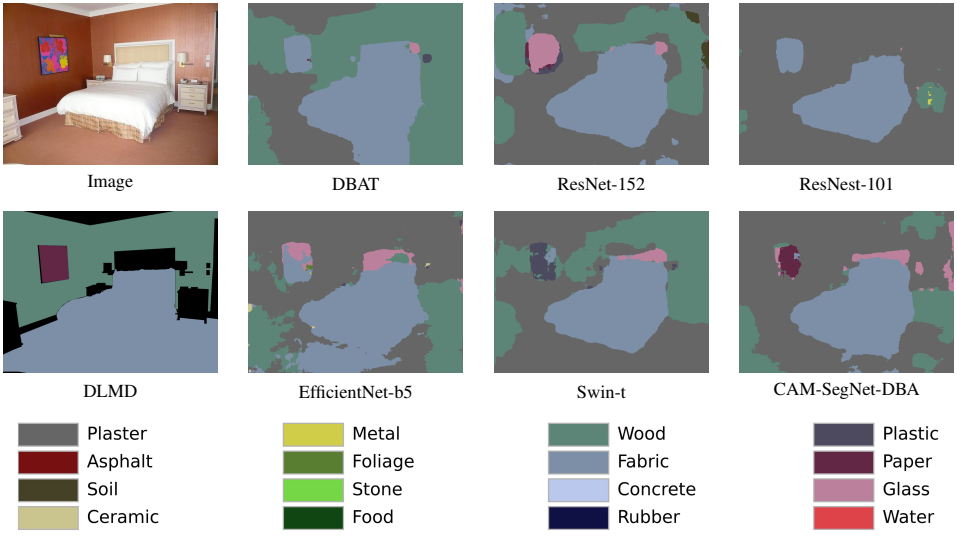


Figure 4: Predicted segmentation of one bedroom image.

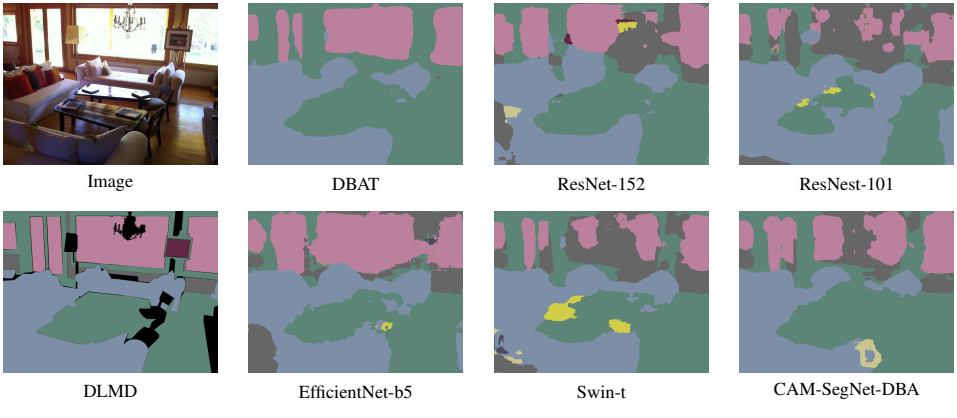


Figure 5: Predicted segmentation of one living-room image.

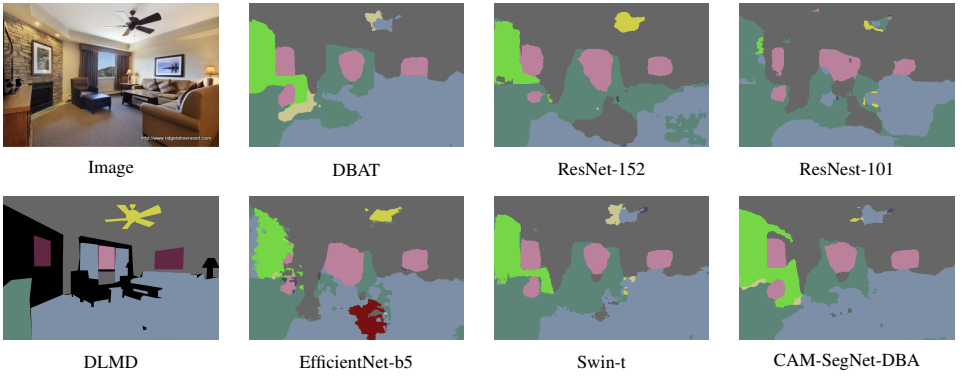


Figure 6: Predicted segmentation of another living-room image.

References

[1] Rong Ge, Sham M Kakade, Rahul Kidambi, and Praneeth Netrapalli. The step decay schedule: A near optimal, geometrically decaying learning rate procedure for least squares. *Advances in Neural Information Processing Systems*, 32, 2019.

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[3] Yuwen Heng, Yihong Wu, Hansung Kim, and Srinandan Dasmahapatra. Cam-segnet: A context-aware dense material segmentation network for sparsely labelled datasets. In *17th International Conference on Computer Vision Theory and Applications (VISAPP)*, volume 5, pages 190–201, 2022.

[4] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International Conference on Machine Learning*, pages 3519–3529. PMLR, 2019.

[5] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021.

[6] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2018.

[7] Jerry Ma and Denis Yarats. On the adequacy of untuned warmup for adaptive optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 8828–8836, 2021.

[8] Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? *Advances in Neural Information Processing Systems*, 34, 2021.

[9] Shibani Santurkar, Dimitris Tsipras, Andrew Ilyas, and Aleksander Madry. How does batch normalization help optimization? *Advances in neural information processing systems*, 31, 2018.

[10] Le Song, Alex Smola, Arthur Gretton, Justin Bedo, and Karsten Borgwardt. Feature selection via dependence maximization. *Journal of Machine Learning Research*, 13 (5), 2012.

[11] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.

[12] Ashia C Wilson, Rebecca Roelofs, Mitchell Stern, Nati Srebro, and Benjamin Recht. The marginal value of adaptive gradient methods in machine learning. *Advances in neural information processing systems*, 30, 2017.

[13] Hang Zhang, Chongruo Wu, Zhongyue Zhang, Yi Zhu, Haibin Lin, Zhi Zhang, Yue Sun, Tong He, Jonas Mueller, R Manmatha, et al. Resnest: Split-attention networks. *arXiv preprint arXiv:2004.08955*, 2020.