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

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

Vision, Learning and Control Research group
School of Electronics and Computer Science
University of Southampton
Southampton, UK

Abstract

Recent studies in material segmentation crop the image into patches to force the network to learn material features from local visual clues. This design is based on the expectation that the contextually invariant features can generalise the network to unseen images regardless of the object or scene in which the material appears. However, most approaches set a fixed patch resolution for all the images in a dataset, which does not consider the varying areas that materials cover within and across images due to the scene scale. As a consequence, the fixed patch resolution can limit the performance of networks. In consideration of this problem, this paper proposes a Dynamic Backward Attention Transformer (DBAT) to extract features from cross-resolution patches and dynamically aggregate these features based on per-pixel attention masks. Experiments show that DBAT achieves the best performance among state-of-the-art models (86.85% in average pixel accuracy, which is 2.15% higher than the second-best model) that can serve real-time inference. Moreover, we also illustrate the network behaviour through visualisation methods as well as descriptive statistics. The project code is available at https://github.com/heng-yuwen/Dynamic-Backward-Attention-Transformer.

1 Introduction

The dense material segmentation task aims at recognising the material for every pixel in a captured photo. This task is critical to various applications such as robot manipulation [42, 56] and spatial audio synthesis [10, 20, 29]. Since a specific material can have a variety of appearances, such as shape, colour and transparency [15, 37], identifying the materials from general RGB images remains challenging [3, 19, 40]. In order to improve the performance, recent material segmentation methods combine material features and contextual features [4, 19, 38, 39, 40]. Material features allow the network to identify the categories without a

massive dataset that covers all varied appearances, and contextual features can limit the possible categories of materials that appear in a given scene.

However, how to balance the material and contextual features has not been properly investigated. Schwartz *et al*. [38, 39, 40] proposed a multi-branch network architecture [54, 55]. They adopt one branch to extract material features from image patches and multiple pre-trained branches targeting contextually related tasks to extract contextual features. The material and contextual features are concatenated to predict the material labels. Heng *et al*. [19] also utilised this multi-branch architecture and employed the self-training strategy [12, 23, 60] to provide boundary information [5] as the contextual features. Their work shows that the material features extracted from image patches can generalise the network to unseen images and achieve state-of-the-art (SOTA) performance. However, the resolution of the image patches is fixed, which may not be the best choice to extract material features. As affected by the camera working distance d_w and field-of-view (FoV), the areas that materials cover vary within and across images. Ideally, small patch resolution should be applied to the boundary between materials, and large patch resolution can be used to cover as much information as possible for the region belonging to a single material.

In this paper, instead of searching for a fixed patch resolution, we devise a simple yet effective transformer architecture, DBAT, to aggregate cross-resolution features. Inspired by the hierarchical architecture of Swin transformer [28], which gradually merges image patches to get a global view, we propose a Dynamic Backward Attention (DBA) module to aggregate the intermediate features extracted from image patches with different resolutions. Concretely, a transformer feature map from a shallow layer contains features extracted from local patches [36], especially when using window-based self-attention. The proposed DBAT merges adjacent patches at each transformer stage to enlarge the patch resolution, and aggregates multiple intermediate feature maps so that it can identify the materials with cross-resolution patch features. To cope with the flexibility of d_w and FoV, a set of pixel-wise attention masks, which represent the dependency on each patch resolution, are applied in our DBA module to dynamically aggregate the feature maps. These masks are calculated from the deepest feature map (Map_4 in Figure 2) since it holds a relatively global perspective of the input image. Before feeding the aggregated feature into the decoder, we further propose a feature merging module which ensures the aggregated feature can improve the network performance through an attention-based residual connection.

The effectiveness of our proposed DBAT is examined through a comparison with SOTA segmentation networks that can achieve real-time performance (at least 24 frames per second). In addition to the numerical evaluation of segmentation performance, we also statistically and visually illustrate the behaviour of our DBAT, including the attention masks, the equivalent patch size of each attention head, and the Centered Kernel Alignment (CKA) [31, 36] heatmap. The statistical analysis is at the supplementary material.

We summarise our main contributions as follows:

- An effective dynamic backward attention transformer to enhance material features by dynamically adjusting the dependency on cross-resolution patch features.
- A feature merging module composed of the attention mechanism and residual connection to guide the aggregation of cross-resolution patch features.
- A batch of methods to illustrate the behaviour of our DBAT, including descriptive statistics as well as visualised images.

Our DBAT beats SOTA real-time models when evaluated on the sparsely labelled Local Material Database (LMD) [40] and OpenSurface Database [2]. In particular, with modern optimisation strategies such as learning rate warm-up [17] and polynomial decay [30], our DBAT reaches an average pixel accuracy (Pixel Acc) of 86.85% on the LMD, which is 21.21% higher than the most recent publication in [19], and outperforms the second-best model in this paper by 2.15%.

2 Background

Material Segmentation. Recent SOTA segmentation networks mainly employ the encoder-decoder architecture to extract features and recover the resolution [9, 21, 28, 50, 57, 60]. Although these networks are proved to work well out-of-the-box in many applications, for our material segmentation task, they still suffer from low accuracy [19, 39, 56]. Schwartz *et al*. [39, 40] pointed out that the challenges in annotating material labels put restrictions on the development of high-quality, large-scale datasets, which is one of the keys to achieving high accuracy with neural networks. To cope with the problem, the images are cropped into 48×48 patches so that the network can learn the generalisable material features hidden in local regions. Heng *et al*. [19] found that networks with fast pooling and atrous convolution [9, 32, 33, 51] may not be able to capture material features covering small regions of an image due to the absence of local features. Although their studies all emphasise the importance of local material features, the patch resolution is fixed at 48×48, which may not be the best choice for all images. Our DBAT, instead, learns from patches with multiple resolutions and lets the network adjust the dependency on each resolution dynamically.

Transformers in Vision Tasks. Transformers composed of self-attention [38] and Multi-layer Perceptron (MLP) [45] shows good performance in many vision tasks such as classification [8, 14, 34] and segmentation [28, 43, 59]. Depending on the scope to which the self-attention module applies, there are global or local transformers. The global transformers represented by ViT [14] and DeiT [46] employ the global self-attention to capture the correlation between each pair of embedded patch features. This design ensures that such transformers can have a global view from the first layer. However, the quadratic complexity in image size makes global transformers expensive to use. Moreover, a recent study [36] shows that global transformers can still have a local view at shallow layers. Their work states that learning from local regions at the beginning is important for good performance. In contrast, the local transformers such as Swin [27, 28] apply the self-attention module to windowed regions. This local design gradually increases the patch size through patch merging. As a result, features from multiple transformer stages are extracted from patches with different resolutions. By aggregating these features, our DBAT achieves the goal of learning from cross-resolution patches.

Network Interpretability. Studies of network interpretability aim at explaining how a network behaves through methods such as visualising the kernel weights [22, 48]. Another simple yet effective way is to plot the per-pixel attention masks on the input image to illustrate which part of the image contributes to the final decision [16, 26]. For transformers, however, the interpretability of the self-attention module remains challenging due to its high dimensionality and way of recursive connections. Carion *et al*. [6] proposed to reduce the dimensionality by visualising one self-attention layer for a single pixel at a time. Chefer *et*

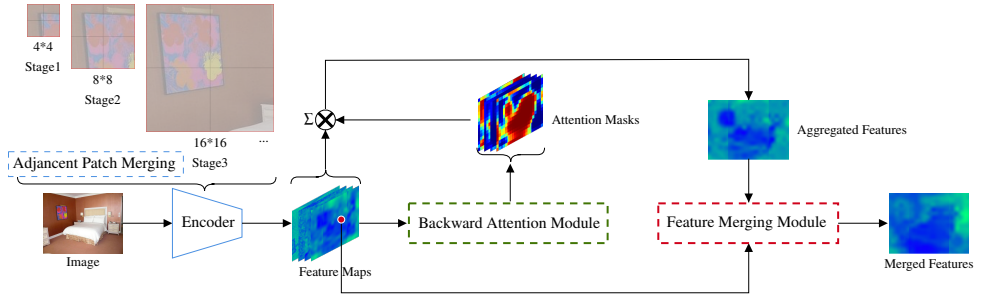


Figure 1: Dynamic Backward Attention Transformer architecture. It consists of an encoder backbone to provide cross-resolution features by merging adjacent patches at each transformer stage, a backward attention module to aggregate these features, and a feature merging module with a residual connection to ensure that the network learns complementary features.

al. [2] reassign a relevancy map to the input and propagate it through all the self-attention layers. However, these methods can only illustrate the transformer behaviour for a specific input image. To obtain a summarised explanation across the whole dataset, we plot the CKA matrix [31, 36] which measures the similarity between two layers with features evaluated by the same group of samples. As a model-independent method, the CKA matrix enables the quantitative comparison between two networks regardless of their architectures. In this paper, we illustrate the behaviour of our DBAT by computing the CKA matrix of itself and its backbone transformer. We show that DBAT learns new features from the aggregated cross-resolution patch features to improve its performance.

3 Dynamic Backward Attention Transformer

This section explains the methodologies that constitute our DBAT. As mentioned above, DBAT consists of three modules: backbone encoder, dynamic backward attention module, and feature merging module, as shown in Figure 1. The encoder extracts cross-resolution feature maps from the input image with multiple stages of transformer blocks. The dynamic backward attention module predicts attention masks to aggregate the feature maps extracted from transformer stages. The feature merging module combines the aggregated features and the last feature map using an attention-based residual connection. Finally, the merged features are passed into a segmentation decoder to produce the labels.

3.1 Dynamic Backward Attention

The DBA module relies on a backbone encoder to extract cross-resolution patch feature maps. Unlike previous research in material segmentation [4, 19, 38, 40], we choose to use a transformer instead of a convolutional neural network (CNN). We find the transformer a suitable encoder for the patch training strategy [19, 40] as it is designed to process image patches natively [14] in addition to its promising results for vision tasks [14, 27, 28, 35]. However, the global self-attention mechanism in ViT and its successors [46, 47, 52] can drop the local features, especially for small datasets [36]. To stay with local patch features, we choose the Swin transformer [28] as our encoder, which utilises the window-based self-

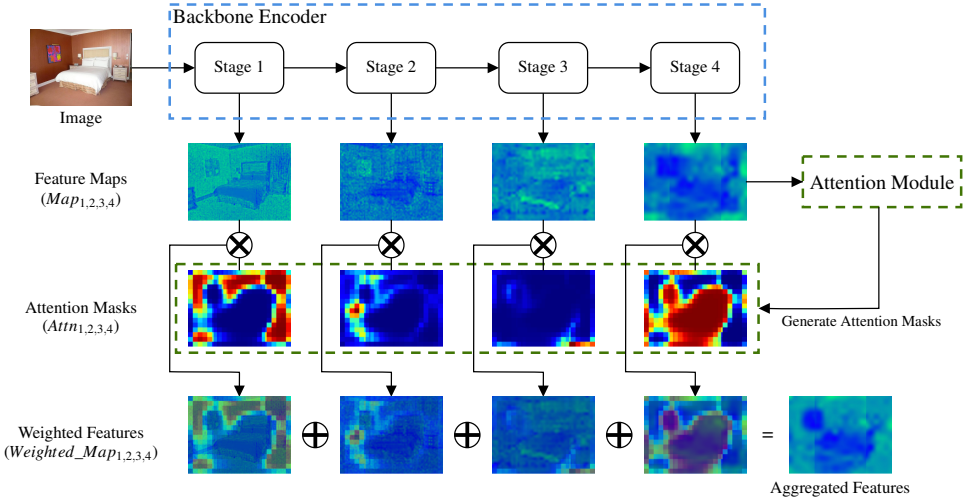


Figure 2: Structure of the DBA module. It performs a weighed sum across the feature maps, $Map_{1,2,3,4}$, to produce the aggregated feature. The weights are dynamically estimated based on the input image through the attention module, which takes the fourth feature map Map_4 as input.

attention. The transformer consists of four stages of transformer blocks. It learns from small image patches at the beginning and gradually increases the patch resolution at each stage by merging features extracted from adjacent patches. By gathering the outputs from the four stages, we get the cross-resolution patch features.

Figure 2 illustrates our dynamic backward attention module. Its objective is to construct a feature map from the cross-resolution patch features. Under the assumption that transformers can preserve spatial location information [36], we propose to perform a weighted sum across these features at each pixel position. The feature map spatial size of stage i can be represented as $(\frac{H}{2\times 2^i}, \frac{W}{2\times 2^i})$, where H and W are height and width of the input image, respectively. After gathering the cross-resolution patch feature maps Map_i, the attention module f_{attn} predicts the per-pixel attention masks, $Attn_i$, based on the fourth stage feature, Map_4. In detail, f_{attn} takes $Map_4 \in R^{c_4\times h\times w}$ as input, and processes Map_4 with three Conv-BatchNorm-ReLU blocks to get an unnormalised attention masks $f_{attn}(Map_4) \in R^{N\times h\times w}$. Here c_4 is the channel number of Map_4 and N is the number of stages in the encoder. The feature resolution (h,w) is the same between the input and output. In this paper, N is set to 4 and $(h,w) = (\frac{H}{32}, \frac{W}{32})$. To perform the aggregation operation, the attention masks are normalised with the SoftMax function so that the weights across the N channels of $Attn$ at each pixel position sum to 1, as shown by Equation 1. The i_{th} channel $Attn_i$ represents the weights of the feature features at $stage_i$. The weighted sum mechanism to aggregate the features for position (j,k) can be represented by Equation 2 and 3:

$$Attn_{i,jk} = \frac{e^{f_{attn}(Map_4)_{i,jk}}}{\sum_{i=1}^{i=N} e^{f_{attn}(Map_4)_{i,jk}}} \tag{1}$$

$$Weighted\_Map_{i,jk} = Attn_{i,jk} \times Map_{i,jk} \tag{2}$$

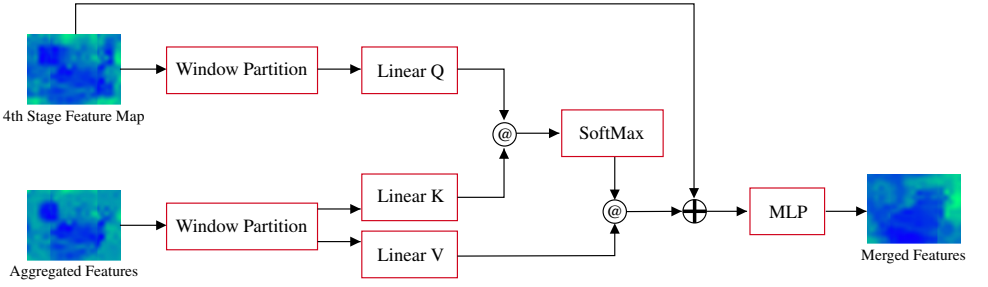


Figure 3: The feature merging module. It merges the relevant cross-resolution information from the aggregated patch feature into Map_4, through the window attention mechanism and residual connection.

$$Aggregated\ Feature_{jk} = \sum_{i=1}^{i=N} Weighted\_Map_{i,jk} \tag{3}$$

where (j,k) is the row and column indices of the aggregated feature. It is worth noting that their spatial shapes should be the same to perform the pixel-wise mapping between Map_i and $Attn_i$. Moreover, to normalise the per-pixel attention masks, the shapes of $Attn_i$ should all be the same as well. In this paper, we set $Attn_i$ to be the same size as Map_4, and down-sample $Map_{1,2,3}$ to $(\frac{H}{32}, \frac{W}{32})$ to minimise the computation and memory overhead. The down-sample method used in this paper is explained in the ablation study section.

3.2 Feature Merging Module

The feature merging module guides the DBA module to enhance the local features with complementary features. A straightforward method is the residual connection [18], *e.g.* $Merged\ Feature = Map_4 + Aggregated\ Feature$. However, this simple addition operation can over-emphasise the condition of Map_4, and break our DBA module that aggregates the features in a linear manner. Therefore, we propose to bring in more non-linear operations with an attention module [28, 41].

As shown in Figure 3, the feature merging module identifies the relevant information in the aggregated cross-resolution patch features based on the knowledge about the final stage feature map, and merges the complementary features into Map_4. The query matrix (Q) is predicted from from Map_4. The key (K) and value (V) matrices are predicted from the aggregated feature. The attention alignment scores (also known as similarity scores) are calculated through the matrix multiplication (represented by the @ symbol) between Q and K. The softmax operation then normalises the scores to generate relevant information from the value matrix (V). Notice that here we use the window attention in [28] as well. The partitioned features are normalised with the LayerNorm layer [1] before being fed into the fully-connected (FC) layers (Linear Q, K, V in Figure 3). The MLP is formed by two FC layers with ReLU activation layer in between. Together with the dynamic backward attention module, this process can be described as enhancing the material features by injecting cross-resolution patch features into Map_4.

4 Experiments

Datasets. We evaluate our proposed DBAT on the Local Material Database (LMD) [19, 39, 40], which contains 16 mutually exclusive labels for 5,845 low-resolution multi-source images, and the OpenSurfaces [2], which has 45 categories for 25,352 high-resolution images of indoor scenes. As annotating images with material labels is a challenging task [19, 39], these two datasets can only provide sparsely labelled segments. In particular, for the Opensurfaces, 27 of the 45 material classes have more than 60 samples. 39.44% of the samples are segmented as "wood" and "painted". The highly unbalanced category sizes make the evaluation on the OpenSurfaces less reliable compared with the LMD. As a consequence, the experiments primarily focus on the LMD, and record the performance evaluated on the OpenSurfaces as an additional piece of evidence. It is worth noting that recent dataset MCubeS [24] is not evaluated since this paper focuses on material segmentation with RGB images.

Evaluation metrics. In our experiments, we use three metrics: mean pixel accuracy (Pixel Acc), mean class accuracy (Mean Acc), and mean intersection over union (mIoU), to assess the model's ability to identify materials. For each image in the LMD, since the ground truth segment only labels a small area of the entire material, the mIoU numerator would be much smaller than it should be. Therefore, the LMD is not evaluated with mIoU. In addition, we also report the resources that each model requires, including the number of trainable parameters and the number of FLOPs. To set a selection criterion for SOTA models, we measure the frames per second (FPS) and choose the model variants that can support real-time inference.

Implementation details. The backbone encoders to be compared are pre-trained on ImageNet [13]. This pre-training step replaces the multi-branch multi-task configuration in [39] which ensures the models can have prior knowledge about the contextual information in the image, such as scene and objects. For the Swin backbone, the implementation follows the original paper and sets the window size as 7. The decoder is set as the Feature Pyramid Network (FPN) [25] since Heng *et al*. [19] shows that FPN can recognise the materials that cover a small image region well. Moreover, it is worth noting that our DBAT is different from FPN. The FPN is designed to learn from low-level and high-level features, while our DBAT aims to learn material features from cross-resolution patches. We use the AdamW optimiser to train the networks with batch size 16, coefficients β_1 0.9, β_2 0.999, and weight decay coefficient 0.01. Further, the learning rate is warmed-up linearly from 0 to 0.00006 with 1,500 training steps, and decreased second-order polynomially with 200 epochs for the LMD, 34 epochs for the OpenSurfaces. More discussion about the training configurations is in the supplementary materials.

Quantitative Analysis. Table 1 compares the segmentation performance of our DBAT against the baseline model ResNet-152 [18], and four SOTA models, ResNest-101 [53], EfficientNet-b5 [44], Swin-t [28], and CAM-SegNet [19]. The CAM-SegNet is chosen because it is the newest architecture for the material segmentation task, although it does not meet the real-time selection criterion. Its local branch is equipped with the DBA module, and its performance is reported as CAM-SegNet-DBA. Among all the models that can serve real-time inference, our DBAT achieves the highest accuracy across the two datasets in terms of Pixel Acc/Mean Acc/mIoU. Specifically, when evaluated on the LMD, DBAT

Datasets	LMD		OpenSurfaces			–		
Architecture	Pixel Acc	Mean Acc	Pixel Acc	Mean Acc	mIoU	#params (M)	#flops (G)	FPS
ResNet-152	80.68 ± 0.11	73.87 ± 0.25	83.80	63.56	52.09	60.75	70.27	31.35
ResNeSt-101	82.45 ± 0.20	75.31 ± 0.29	85.10	67.13	55.32	48.84	63.39	25.57
EfficientNet-b5	83.17 ± 0.06	76.91 ± 0.06	84.63	65.47	53.25	30.17	20.5	27.00
Swin-t	84.70 ± 0.26	79.06 ± 0.46	86.19	69.41	57.71	29.52	34.25	33.94
CAM-SegNet-DBA	86.12 ± 0.15	79.85 ± 0.28	**86.64**	69.92	**58.18**	68.58	60.83	17.79
DBAT	**86.85 ± 0.08**	**81.05 ± 0.28**	86.28	**70.68**	58.08	56.03	41.23	27.44

Table 1: Segmentation performance on the LMD and the OpenSurfaces. The FPS is calculated by processing 1000 images with one NVIDIA 3060ti. The uncertainty evaluation is reported across five runs.

achieves +0.73%/+1.2% higher than the second-best model, CAM-SegNet-DBA. It is also +2.15%/+1.99% higher than its backbone encoder, Swin-t. As for the OpenSurfaces, our DBAT beats the chosen four real-time models, and its performance is close to the multi-branch CAM-SegNet-DBA (-0.36%/+0.76%/-0.10%) with 9.65 more FPS and 19.6G fewer FLOPs. Compared with the original Pixel Acc (71.65%) of CAM-SegNet reported in [19], our DBAT achieves an improvement of 21.21%.

Qualitative Analysis. Figure 4 shows the predicted segmentation for one indoor scene. 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. Moreover, the DBAT segmented image has fewer flying pixels than the EfficientNet-b5 and Swin-t. This indicates that the generalisable material features that DBAT learns can distinguish different materials accurately. The supplementary material contains two more segmented images for bedroom and kitchen photos.

Network Behaviour Analysis. This section addresses the network behaviour through the CKA matrix [31, 36], which measures the similarity between the feature maps (X, Y) extracted from two arbitrary layers. More detail about the CKA matrix is included in the supplementary material. Figure 5 visualises the CKA heatmap of the DBAT in Figure 5

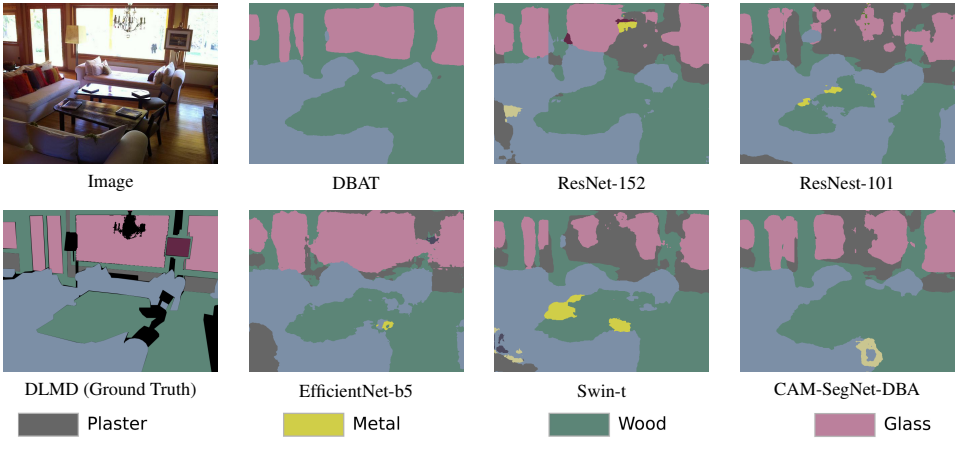
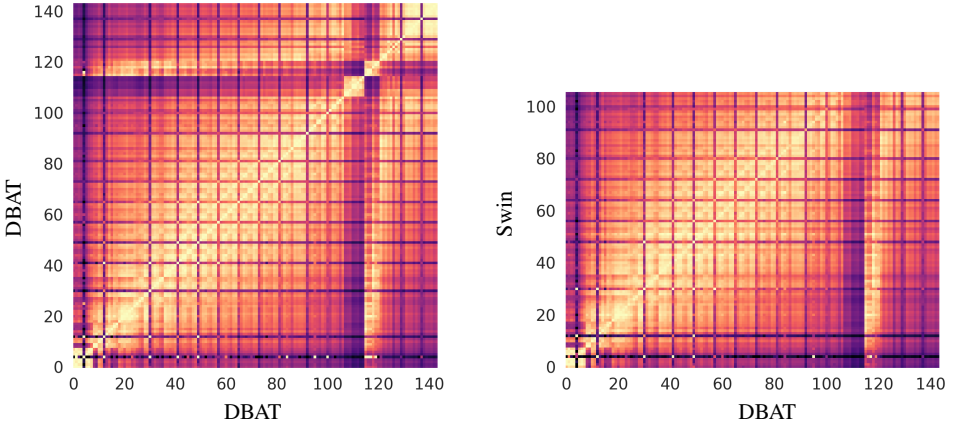


Figure 4: Predicted segmentation of one living room image.



(a) The CKA matrix of our DBAT

(b) The CKA matrix between DBAT and Swin

Figure 5: The CKA matrix where each position measures the similarity between the features extracted by two arbitrary layers. The brighter the colour is, the more similar features these two layers extract.

(a) and the cross model heatmap between DBAT and Swin in Figure 5 (b). The layers are indexed by the forward propagation order. Before layer 106, the DBAT shares the same network architecture as the Swin. The bright line in Figure 5 (b) connecting (0, 0) and (105, 105) gets darker when approaching point (105, 105). This indicates that the Swin backbone in the DBAT extracts similar features to when used alone at shallow layers and gradually learns something new when approaching deeper layers. The dark region in Figure 5 (a) from layer 106 to 113 reflects the attention masks predicted by the DBA module. By gathering the cross-resolution feature maps, the aggregated features contain information from both shallow and deep layers, illustrated by the bright region between layer 113 and 124. After layer 124, the feature merging module combines the relevant information from the aggregated features into Map_4, which is extracted from layer 106. This module produces a feature map that differs from the feature extracted by Swin, as shown by the points around (140, 100) in Figure 5 (b).

Ablation Study. This section analyses the effectiveness of each component in the DBAT by taking them off one by one. Table 2 shows that without the feature merging module, the Pixel Acc decreases by 1.61% and Mean Acc decreases by 2.04%. This indicates that the residual connection is important to guarantee an increase in performance. After removing the dynamic backward attention module, the performance drops by another 0.72% in Pixel Acc and 0.13% in Mean Acc. The per-category analysis in the supplementary material shows that the DBA module improves the performance in materials that usually have unique appearances, such as paper, stone, fabric and wood. This indicates that the cross-resolution features successfully learn from distinguishable material features.

In addition to analysing network components, this section also studies the implementation choices that fulfil the DBAT: 1) how to predict the attention masks; 2) how to downsample the feature maps; 3) how to merge the aggregated features with Map_4. Table 3 shows the performance differences by switching one of the implementations in DBAT to its alternative. DBAT originally adopts regular convolutional kernels to generate the attention masks

Architecture	Δ Pixel Acc	Δ Mean Acc
- Feature merging	-1.61	-2.04
- Dynamic backward attention	-2.33	-2.17

Table 2: The ablation study to analyse each component of our DBAT.

as in [11]. By replacing the kernel with its dilated version [49], the attention masks are predicted with an enlarged receptive field. However, the performance decreases significantly. This indicates that restricting its view to local regions is critical for the dynamic attention module. As mentioned in Section 3.1, the cross-resolution feature maps need to be down-sampled to cope with the fixed size attention masks. DBAT chooses to use the MLP, and the size is the same as the down-sample rate. Instead of using trainable kernels, a superficial non-parametric pooling layer can also do the job, which decreases the performance by (-0.88%/-1.58%). The slight decrease in Pixel Acc and the significant decrease in Mean Acc indicate that the trainable down-sampling method can help balance the performance across different material categories. As for the feature merging module, a simple residual connection only harms the performance by (-0.58%/0.64%). This highlights that the DBAT needs the Map_4 to guide the aggregation of cross-resolution features.

Implementation Choices		Δ Pixel Acc	Δ Mean Acc
Generate Attention Masks	CNN –> Dilated CNN	-2.15	-2.67
Down-sample	MLP –> Average Pooling	-0.88	-1.58
Feature Merging	Attention –> Residual Connection	-0.58	-0.64

Table 3: The study of different implementation choices in each component of our DBAT. The component on the left side of the arrow is replaced with the right side in the experiment.

5 Conclusion

This paper proposed a single branch network to enhance the material features dynamically by aggregating the cross-resolution patch features. Our DBAT beats all chosen models that can serve real-time applications on two datasets, and achieves comparable performance with fewer FLOPs than the multi-branch CAM-SegNet. We also illustrated its mechanism through visualising the attention masks as well as the CKA heatmaps. We show that about half of the information comes from layers that extract features from small patches which can help the network separate multiple overlapping materials. In addition, the CKA heatmap shows that the aggregated feature is highly similar to shallow layers. In the future, we plan to interpret the material features that our DBAT learns by comparing them with features extracted from different tasks, such as object segmentation.

ACKNOWLEDGMENT

This work was supported by the EPSRC Programme Grant Immersive Audio-Visual 3D Scene Reproduction Using a Single 360 Camera (EP/V03538X/1).

References

[1] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[2] Sean Bell, Paul Upchurch, Noah Snavely, and Kavita Bala. OpenSurfaces: A richly annotated catalog of surface appearance. *ACM Trans. on Graphics (SIGGRAPH)*, 32 (4), 2013.

[3] Sean Bell, Paul Upchurch, Noah Snavely, and Kavita Bala. Material recognition in the wild with the materials in context database. *Computer Vision and Pattern Recognition (CVPR)*, 2015.

[4] Sean Bell, Paul Upchurch, Noah Snavely, and Kavita Bala. Material recognition in the wild with the materials in context database. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3479–3487, 2015.

[5] Alexey Bokhovkin and Evgeny Burnaev. Boundary loss for remote sensing imagery semantic segmentation. In *International Symposium on Neural Networks*, pages 388–401. Springer, 2019.

[6] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.

[7] Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 782–791, 2021.

[8] Chun-Fu Richard Chen, Quanfu Fan, and Rameswar Panda. Crossvit: Cross-attention multi-scale vision transformer for image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 357–366, 2021.

[9] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018.

[10] Long Chen, Wen Tang, Nigel W John, Tao Ruan Wan, and Jian J Zhang. Context-aware mixed reality: A learning-based framework for semantic-level interaction. In *Computer Graphics Forum*, volume 39, pages 484–496. Wiley Online Library, 2020.

[11] Yinpeng Chen, Xiyang Dai, Mengchen Liu, Dongdong Chen, Lu Yuan, and Zicheng Liu. Dynamic convolution: Attention over convolution kernels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11030–11039, 2020.

[12] Hao Cheng, Chaochen Gu, and Kaijie Wu. Weakly-supervised semantic segmentation via self-training. In *Journal of Physics: Conference Series*, volume 1487, page 012001. IOP Publishing, 2020.

[13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020.

[15] Roland W Fleming. Visual perception of materials and their properties. *Vision Research*, 94:62–75, 2014. ISSN 0042-6989. doi: https://doi.org/10.1016/j.visres.2013.11.004.

[16] Hiroshi Fukui, Tsubasa Hirakawa, Takayoshi Yamashita, and Hironobu Fujiyoshi. Attention branch network: Learning of attention mechanism for visual explanation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10705–10714, 2019.

[17] Akhilesh Gotmare, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. A closer look at deep learning heuristics: Learning rate restarts, warmup and distillation. In *International Conference on Learning Representations*, 2018.

[18] 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.

[19] 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.

[20] Hansung Kim, Luca Remaggi, Philip JB Jackson, and Adrian Hilton. Immersive spatial audio reproduction for vr/ar using room acoustic modelling from 360 images. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pages 120–126. IEEE, 2019.

[21] Seung-Wook Kim, Hyong-Keun Kook, Jee-Young Sun, Mun-Cheon Kang, and Sung-Jea Ko. Parallel feature pyramid network for object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 234–250, 2018.

[22] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

[23] T Hoang Ngan Le, Khoa Luu, and Marios Savvides. Fast and robust self-training beard/moustache detection and segmentation. In *2015 international conference on biometrics (ICB)*, pages 507–512. IEEE, 2015.

[24] Yupeng Liang, Ryosuke Wakaki, Shohei Nobuhara, and Ko Nishino. Multimodal material segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19800–19808, 2022.

[25] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

[26] Nian Liu, Junwei Han, and Ming-Hsuan Yang. Picanet: Learning pixel-wise contextual attention for saliency detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3089–3098, 2018.

[27] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. Swin transformer v2: Scaling up capacity and resolution. *arXiv preprint arXiv:2111.09883*, 2021.

[28] 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.

[29] Aoife McDonagh, Joseph Lemley, Ryan Cassidy, and Peter Corcoran. Synthesizing game audio using deep neural networks. In *2018 IEEE Games, Entertainment, Media Conference (GEM)*, pages 1–9. IEEE, 2018.

[30] Purnendu Mishra and Kishor Sarawadekar. Polynomial learning rate policy with warm restart for deep neural network. In *TENCON 2019-2019 IEEE Region 10 Conference (TENCON)*, pages 2087–2092. IEEE, 2019.

[31] Thao Nguyen, Maithra Raghu, and Simon Kornblith. Do wide and deep networks learn the same things? uncovering how neural network representations vary with width and depth. *International Conference on Learning Representations*, 2020.

[32] Teerapong Panboonyuen, Kulsawasd Jitkajornwanich, Siam Lawawirojwong, Panu Srestasathiern, and Peerapon Vateekul. Semantic labeling in remote sensing corpora using feature fusion-based enhanced global convolutional network with high-resolution representations and depthwise atrous convolution. *Remote Sensing*, 12(8):1233, 2020.

[33] Siyuan Qiao, Liang-Chieh Chen, and Alan Yuille. Detectors: Detecting objects with recursive feature pyramid and switchable atrous convolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10213–10224, 2021.

[34] Yuhao Qing, Wenyi Liu, Liuyan Feng, and Wanjia Gao. Improved transformer net for hyperspectral image classification. *Remote Sensing*, 13(11):2216, 2021.

[35] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[36] 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.

[37] Shubhangi S Sapkale and Manoj P Patil. Material classification using color and texture features. In *International Conference on Recent Trends in Image Processing and Pattern Recognition*, pages 49–59. Springer, 2018.

[38] Gabriel Schwartz. *Visual Material Recognition*. Drexel University, 2018.

[39] Gabriel Schwartz and Ko Nishino. Material recognition from local appearance in global context. In *Biol. and Artificial Vision (Workshop held in conjunction with ECCV 2016)*, 2016.

[40] Gabriel Schwartz and Ko Nishino. Recognizing material properties from images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(8):1981–1995, 2020. doi: 10.1109/TPAMI.2019.2907850.

[41] Yuanyuan Shen, Edmund M-K Lai, and Mahsa Mohaghegh. Effects of similarity score functions in attention mechanisms on the performance of neural question answering systems. *Neural Processing Letters*, pages 1–20, 2022.

[42] Nithin Shrivatsav, Lakshmi Nair, and Sonia Chernova. Tool substitution with shape and material reasoning using dual neural networks. *arXiv preprint arXiv:1911.04521*, 2019.

[43] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7262–7272, 2021.

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

[45] Ilya O Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, et al. Mlp-mixer: An all-mlp architecture for vision. *Advances in Neural Information Processing Systems*, 34, 2021.

[46] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357. PMLR, 2021.

[47] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 568–578, 2021.

[48] Zijie J Wang, Robert Turko, Omar Shaikh, Haekyu Park, Nilaksh Das, Fred Hohman, Minsuk Kahng, and Duen Horng Polo Chau. Cnn explainer: Learning convolutional neural networks with interactive visualization. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):1396–1406, 2020.

[49] Yunchao Wei, Huaxin Xiao, Honghui Shi, Zequn Jie, Jiashi Feng, and Thomas S Huang. Revisiting dilated convolution: A simple approach for weakly-and semi-supervised semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7268–7277, 2018.

[50] Yongfeng Xing, Luo Zhong, and Xian Zhong. An encoder-decoder network based fcn architecture for semantic segmentation. *Wireless Communications and Mobile Computing*, 2020, 2020.

[51] Muzhou Xu, Shan Zhong, Chunping Liu, Shengrong Gong, Zhaohui Wang, and Yu Xia. Acclvos: Atrous convolution with spatial-temporal convlstm for video object segmentation. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 2089–2096. IEEE, 2021.

[52] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zi-Hang Jiang, Francis EH Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 558–567, 2021.

[53] 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.

[54] Hongyan Zhang, Yue Liao, Honghai Yang, Guangyi Yang, and Liangpei Zhang. A local-global dual-stream network for building extraction from very-high-resolution remote sensing images. *IEEE Transactions on Neural Networks and Learning Systems*, 2020.

[55] Hongyang Zhang, Junru Shao, and Ruslan Salakhutdinov. Deep neural networks with multi-branch architectures are intrinsically less non-convex. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1099–1109. PMLR, 2019.

[56] Cheng Zhao, Li Sun, and Rustam Stolkin. Simultaneous material segmentation and 3d reconstruction in industrial scenarios. *Frontiers in Robotics and AI*, 7:52, 2020.

[57] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017.

[58] Hengshuang Zhao, Jiaya Jia, and Vladlen Koltun. Exploring self-attention for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10076–10085, 2020.

[59] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6881–6890, 2021.

[60] Barret Zoph, Golnaz Ghiasi, Tsung-Yi Lin, Yin Cui, Hanxiao Liu, Ekin Dogus Cubuk, and Quoc Le. Rethinking pre-training and self-training. In *NeurIPS*, 2020.