Material Recognition for Immersive Interactions in Virtual/Augmented Reality

Yuwen Heng* Srinandan Dasmahapatra† Hansung Kim‡

School of Electronics and Computer Science
University of Southampton

Abstract

To provide an immersive experience in a mirrored virtual world such as spatially synchronised audio, visualisation of reproduced real-world scenes and haptic sensing, it is necessary to know the materials of the object surface which provides the optical and acoustic properties for the rendering engine. We focus on identifying materials from real-world images to reproduce more realistic and plausible virtual environments. To cope with considerable variation in material, we propose the DPT architecture which dynamically decides the dependency on different patch resolutions. We evaluate the benefits of learning from multiple patch resolutions on LMD and OpenSurfaces datasets.

Index Terms: Human-centered computing—Human computer interaction (HCI)—Interaction paradigms—Virtual reality; Computing methodologies—Artificial intelligence—Computer vision—Scene understanding

1 Introduction

Mirror World is a digital twin of the real (physical) environment in the virtual environment [5, 6]. It reproduces real-world structures in a virtual world in visual, geographical, and attributional senses. In virtual and augmented reality applications of the mirror world, interactions with the surrounding environment make immersive experience possible. To enhance the immersiveness of users, the render engine needs material properties, which can be used to infer how the lightwave or soundwave should be reflected by the object surfaces. For example, in virtual applications, plausible scenes can be created with physically-based rendering techniques, which trace the lightwave from the light source to the virtual camera, and calculates how the spectra change based on materials when interacting with the object surfaces. In augmented applications, the immersive audio can be synthesised based on the surrounding environments including material and scene structure. In this paper, we focus on the scenario that the applications are created from real-world scenes and propose to recognise material categories from images.

The material segmentation task aims at assigning material categories such as metal and plastic to each pixel of the image. Material segmentation is still a challenging task considering the variations in the appearance and shape of a single material. Although it is possible to train generalisable networks by learning material features from cropped patches, recent material segmentation methods fail to account the area that a material region can cover [2], and chooses a fixed patch resolution for the whole dataset. Ideally, small patch resolution should be applied to the boundaries between materials, while large patch resolution can be used to cover as much information as possible in areas belonging to a single material.

*e-mail: y.heng@soton.ac.uk
†e-mail:sd@ecs.soton.ac.uk
‡e-mail:h.kim@soton.ac.uk

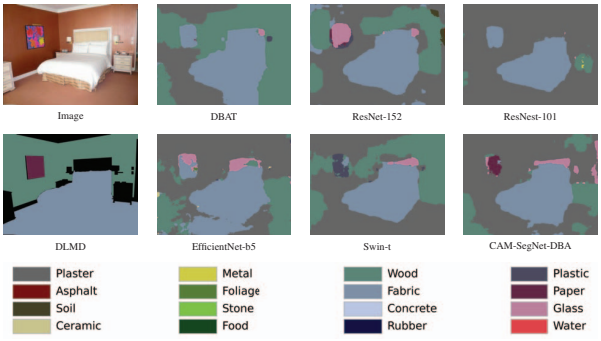


Figure 1: Predicted material segmentation of one bedroom image.

Instead of searching for a fixed patch resolution, we propose the Dynamic Patch Training (DPT) architecture. The DPT architecture consists of three modules, the cross-resolution feature extractor, the dynamic backward attention module, and the feature merging module. The cross-resolution feature extractor employs non-overlapping network kernels to process images as cropped patches. The adjacent patch features are then merged to increase the patch size at each network stage. In this paper, we evaluated two non-overlapping kernel choices, the transformer and the convolutional layer whose step is equal to its kernel size. The transformer-based variant is from [3]. The backward attention module predicts per-pixel attention masks to aggregate the intermediate cross-resolution patch features. Finally, the feature merging module is built upon the residual connection to regularise the DPT architecture to learn complementary features compared with the basic backbone encoder.

2 Proposed Method

The hypothesis behind the DPT training architecture is that the cross-resolution patch features can improve the material segmentation quality both qualitatively and quantitatively. This section will introduce the architecture components briefly.

Cross-Resolution Feature Extractor The cross-resolution feature extractor is designed to learn from patches cropped by multiple resolutions in a single path. Instead of cropping the images in advance, we first propose to enhance the CAM-SegNet [2] with DPT architecture. The local branch of CAM-SegNet is replaced with non-overlapping convolutional kernels whose stride is the same as the kernel size. In detail, we use 4×4 kernel followed by a multi-layer perceptron (MLP). The 4×4 kernel merges adjacent features to increase the patch size and downsample the feature map. The MLP makes the network deeper and learns features within the patch. Inspired by the MLP architecture, a transformer named Dynamic Backward Attention Transformer (DBAT) was proposed [3]. The DBAT replaces the convolutional kernel with window-based self-attention which learns material features within the window size.

979-8-3503-4839-2/23/$31.00 ©2023 IEEE
DOI 10.1109/VRW58643.2023.00131

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	$\mathbf{86.85 \pm 0.08}$	$\mathbf{81.05 \pm 0.28}$	86.28	**70.68**	58.08	56.03	41.23	27.44

Table 1: Material recognition performance on the LMD and the OpenSurfaces. The FPS is calculated by processing 1,000 images with one NVIDIA 3060ti. The uncertainty evaluation is reported across five runs. The Pixel Acc is the averaged per-pixel accuracy and the mean Acc is the accuracy averaged across each category. The #params is the number of trainable parameters of the models.

Dynamic Backward Attention Module The backward attention module predicts per-pixel attention masks to combine the cross-resolution features. As shown by the following equations, each pixel j,k of the image is processed by i different patch resolutions. Assume the encoder consists of four stages. The attention masks $Attn$ are predicted from the last feature map Map_4 and normalised by the softmax operation. These masks represent the dependency on each resolution for each pixel. The weighted sum operation gives the aggregated feature at pixel j,k.

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

$$Aggregated\ Feature_{jk} = \sum_{i=1}^{i=N} Attn_{i,jk} \times Map_{i,jk} \tag{2}$$

Feature Merging Module The feature merging module is used to regularise the DPT architecture to learn complementary features compared with its backbone encoder. We apply an attention module to identify the relevant information in the aggregated cross-resolution patch features based on the knowledge about the final stage feature map Map_4. The relevant features are then merged into Map_4 with a residual connection.

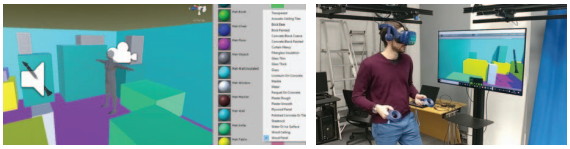
3 Experimental Evaluation

As shown in Table 1 [3], the CAM-SegNet-DBA and DBAT achieve the best performance compared with the other four chosen networks. For the real-time models, Although CAM-SegNet-DBA achieves 0.36%/0.10% improvement in Pixel Acc/mIoU when evaluated on OpenSurfaces, the DBAT performs better on LMD and runs 9.65 more frames per second with 19.6G fewer FLOPs.

Figure 1 shows the segmented images for the chosen models. The CAM-SegNet-DBA successfully identify the drawing on the wall as paper and other models tend to recognise it as glass or fabric. For the DBAT, it can segment the images with more adequate boundaries compared with the outputs of other models. These figures indicate that learning from image patches can help the network extract robust and generalisable material features.

4 Application to VR Reproduction

In order to provide better user experiences adapted to the human perceptual system in the mirror world, the rendering engine requires both reconstructed 3D objects and their surface material information. A particular example is the vision-based immersive sound synthesis [1,4]. 3D room geometry and object labels are reconstructed from a single image using the semantic scene reconstruction method. This 3D model can be loaded on a 3D rendering engine like Unity or Unreal for real-time interaction in VR. As shown in Figure 2, the estimated material information can be imported to Unity and used for real-time 3D sound generation using a spatial audio plug-in like Google Resonance or Steam Audio.



(a) Unity material assignment interface (b) Immersive sound VR experience

Figure 2: Material assignment for spatial sound rendering in Unity.

ACKNOWLEDGMENT

This work was partially supported by the EPSRC Programme Grant Immersive Audio-Visual 3D Scene Reproduction (EP/V03538X/1) and partially by the Korea Institute of Science and Technology (KIST) Institutional Program (Project No. 2E31591).

5 Conclusion

We proposed a dynamic patch training architecture to enhance material features by learning from cross-resolution patches. In addition, we provide an example which embeds a material segmentation network into the creation pipeline of a mirror world with accelerated efficiency. In the future, we plan to extend our DPT architecture with material property estimation and apply it to VR/AR applications such as remote collaborations, robot interaction and haptic interaction, etc.

References

[1] M. Alawadh, Y. Wu, Y. Heng, L. Remaggi, M. Niranjan, and H. Kim. Room acoustic properties estimation from a single 360° photo. In *2022 30th European Signal Processing Conference (EUSIPCO)*, pp. 857–861, 2022.

[2] Y. Heng, Y. Wu, S. Dasmahapatra, and H. Kim. Cam-segnet: A context-aware dense material segmentation network for sparsely labelled datasets. In *International Conference on Computer Vision Theory and Applications (06/02/22 - 08/02/22)*, February 2022.

[3] Y. Heng, Y. Wu, S. Dasmahapatra, and H. Kim. Enhancing material features using dynamic backward attention on cross-resolution patches. In *33rd British Machine Vision Conference 2022, BMVC 2022, London, UK, November 21-24, 2022*. BMVA Press, 2022.

[4] H. Kim, L. Remaggi, P. J. Jackson, and A. 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)*, pp. 120–126, 2019. doi: 10.1109/VR.2019.8798247

[5] A. Ricci, M. Piunti, L. Tummolini, and C. Castelfranchi. The mirror world: Preparing for mixed-reality living. *IEEE Pervasive Computing*, 14(2):60–63, 2015. doi: 10.1109/MPRV.2015.44

[6] Z. Zhang, B. Cao, J. Guo, D. Weng, Y. Liu, and Y. Wang. Inverse virtual reality: Intelligence-driven mutually mirrored world. In *2018 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 735–736, 2018. doi: 10.1109/VR.2018.8446260