IMET - International Conference on Interactive Media, Smart Systems and Emerging Technologies (2023)
N. Pelechano, F. Liarokapis, D. Rohmer, and A. Asadipour (Editors)

Exploring the Impact of Synthetic Data Generation on Texture-based Image Classification Tasks

B. Yordanov[1] C. Harvey[1] I. Williams[1] C. Ashley[2] and P. Fairbrass[2]

[1]DMTLab, School of Computing and Digital Technology, Birmingham City University, Birmingham, United Kingdom
borislav.yordanov@bcu.ac.uk, carlo.harvey@bcu.ac.uk, ian.williams@bcu.ac.uk
[2]SBS Insurance Services, Staffordshire, United Kingdom
c.ashley@sbs-claims.co.uk, p.fairbrass@sbs-claims.co.uk

Abstract

In this study, we introduce a novel pipeline for synthetic data generation of textured surfaces, motivated by the limitations of conventional methods such as Generative Adversarial Networks (GANs) and Computer-Aided Design (CAD) models in our specific context. We also investigate the pipeline's role in an image classification task. The primary objective is to determine the impact of synthetic data generated by our pipeline on classification performance. Using EfficientNetV2-S as our image classifier and a dataset of three texture classes, we find that synthetic data can significantly enhance classification performance when the amount of real data is scarce, corroborating previous research. However, we also observe that the balance between synthetic and real data is crucial, as excessive synthetic data can negatively impact performance when sufficient real data is available. We theorize that this might stem from imperfections in the synthetic data generation process that distort fine details essential for accurate classification, and propose possible improvements to the synthetic data generation pipeline. Furthermore, we acknowledge the potential limitations of our study and provide several promising avenues for future research. This work illuminates the advantages and potential drawbacks of synthetic data in image classification tasks, emphasizing the importance of high-quality, realistic synthetic data that complements, rather than undermines, the use of real data.

Keywords: Synthetic Data, Image Classification, Textured Surface

CCS Concepts

• ***Computing methodologies*** → ***Object recognition; Appearance and texture representations;*** *Image representations; Supervised learning by classification; Computer graphics;*

1. Introduction

Image classification and transformations are a crucial task in computer vision processes with numerous practical applications in fields such as autonomous driving, harmonisation, robotics, and medical imaging [ZVSL18, STT*19, DHW20]. The accuracy of image classification models depends heavily on the quality and quantity of data used for training [GBWC12]. However, collecting and annotating large datasets can be time-consuming, expensive, and sometimes even impossible, especially when dealing with specialised domains or rare events [AMSD21]. To overcome these limitations, synthetic data generation has emerged as a promising solution to augment real data or even replace it entirely in certain scenarios [MCL*18, ABB*20].

While conventional methods for synthetic data generation, like Generative Adversarial Networks (GANs) and Computer-Aided Design (CAD) models, have shown success in various domains such as medical imaging [PCQ*20,LHW*19b,CDLA16,HFR*20], stereo or optical flow estimation [MIF*18] and robotic 3-D object classification [WPV19, CCY*20], their inherent limitations have prompted the need for tailored solutions in certain cases. In this study, we propose a novel pipeline for synthetic data generation of textured surfaces and investigate its role in image classification tasks. Our primary objective is to determine the impact of synthetic data generated by our pipeline on classification performance, using EfficientNetV2-S as our image classifier and a dataset of three texture classes. We aim to explore the potential benefits and limitations of synthetic data in this context, including the balance between synthetic and real data, the impact of imperfections in synthetic data generation, and possible improvements to the pipeline.

Our study builds upon previous research that has shown the potential of synthetic data in enhancing classification performance, especially when real data is scarce [HNBKK15,WHH18,FSB*21]. However, there is still a need for more systematic evaluations of synthetic data quality and their impact on different types of image classifiers and distributions of datasets. Our work contributes to this effort by presenting a comprehensive analysis of the effect of synthetic data on classification performance, including a thorough investigation of the trade-offs between synthetic and real data across a range of data distributions in both these types of data.

The findings of this study have significant implications for the development of image classification models that can operate efficiently and accurately with limited real data. Moreover, our work sheds light on the importance of high-quality, realistic synthetic data that complements, rather than undermines, the use of real data. We believe that the insights gained from this study will pave the way for future research in synthetic data generation for image classification and related fields.

This study makes two primary contributions:

1. We introduce a pipeline for generating synthetic data of textured surfaces, and;
2. We perform an ablation study to assess the impact of synthetic data on the accuracy of an image classifier when used in conjunction with varying amounts of real training data.

2. Related Work

The use of synthetic data has gained increasing attention in recent years as a cost-effective and time-efficient solution to data scarcity and annotation challenges in various computer vision tasks [dCBV09, GRBH18, BCC*18]. In image classification, several studies have explored the impact of synthetic data on model performance, with mixed results depending on the specific task, dataset, and synthetic data generation methods [MQHY20, YZY*22].

One common approach to generating synthetic data is based on generative adversarial networks (GANs) [GPAM*20], which learn to generate realistic images that resemble the training data distribution. For instance, Zhu et al. proposed a GAN-based method for generating synthetic data, which demonstrated improved classification accuracy compared to using only real data [ZLQL17]. Similarly, Momeni et al. used a GAN-based approach to generate synthetic medical images for lesion classification, showing their GAN model can be applied on unseen datasets, with different MRI parameters and diseases, to generate synthetic lesions with high diversity and without needing laboriously marked ground truth data whilst enhancing model performance [MFL*21].

Other studies have focused on more specialised synthetic data generation methods, such as texture synthesis. For example, Mu et al. proposed a texture transfer method for generating synthetic data of animal images, which improved classification performance when combined with real data [MQHY20]. Similarly, Ding et al. introduced a texture synthesis method for generating synthetic data of skin lesions, which showed promising results in improving classification accuracy [DZL*21].

While previous research has shown the potential benefits of synthetic data, there are also limitations and challenges associated with its use. One common issue is the quality and realism of synthetic data, which can affect model performance and generalisation. For instance, Pereira et al. found that their results obtained on synthetic data can misestimate the actual model performance when it is deployed on real data, suggesting the importance of carefully evaluating and controlling the quality of synthetic data as well as defining proper testing protocols [PKM*21]. In addition, there are questions around the balance between synthetic and real data, as well as the optimal ratio of synthetic to real data for a given task and dataset [MTM12]. Mogelmose et al. found that even when the synthetic data covered a large part of the parameter space, it would still perform significantly worse than real-world data.

Our study addresses some of these limitations and challenges by proposing a pipeline for synthetic data generation of textured surfaces and thoroughly evaluating its impact on classification performance. We investigate the trade-offs between synthetic and real data. Our work contributes to the ongoing research on synthetic data generation for image classification and highlights the importance of high-quality, realistic synthetic data that complements real data in enhancing model performance in image classification tasks.

3. Methodology

Our experiments focus on training an image classifier to distinguish between three distinct categories of carpets, based on their material composition:

1. carpets made of 100% synthetic fibres
2. carpets made of 100% wool fibres
3. carpets consisting of a blend of synthetic and wool fibres

This classification task has practical implications in the flooring insurance claim handling industry, where companies must accurately determine the category of a damaged carpet to provide a replacement with comparable price and quality. Notably, our proposed methodology exhibits the potential to be generalised and utilised in other applications that involve textured surfaces.

3.1. Synthetic Pipeline

For the generation of synthetic data, we chose to construct our own pipeline rather than relying on conventional methods such as the use of Generative Adversarial Networks (GANs) or Computer-Aided Design (CAD) models. This decision stemmed from the inherent limitations of these methodologies within our specific context. GANs, despite their proven effectiveness in data generation, inherently operate within the confines of the distribution of the training data [GPAM*20]. They do not possess the capacity to generate data that lies outside of this distribution. Given our training dataset, which was tightly distributed and comprised almost exclusively of top-down carpet images provided by carpet manufacturers, this restriction posed by GANs was deemed detrimental. These images were characterised by their uniform and consistent lighting conditions, and as such, the use of GANs would have generated synthetic data with similar lighting and camera angles. Consequently, the resulting model would have lacked the capability to generalise effectively to the real-world task of identifying carpet materials from arbitrary photos, which are expected to display variations in lighting and viewpoint. CAD models, on the other hand, although offering a higher degree of control over the generation process [MQHY20], were not a suitable choice due to the scarcity of detailed carpet models available. The majority of existing CAD models fail to capture the intricacies and fine texture details that are crucial for the accurate categorisation of carpets based on their material composition.

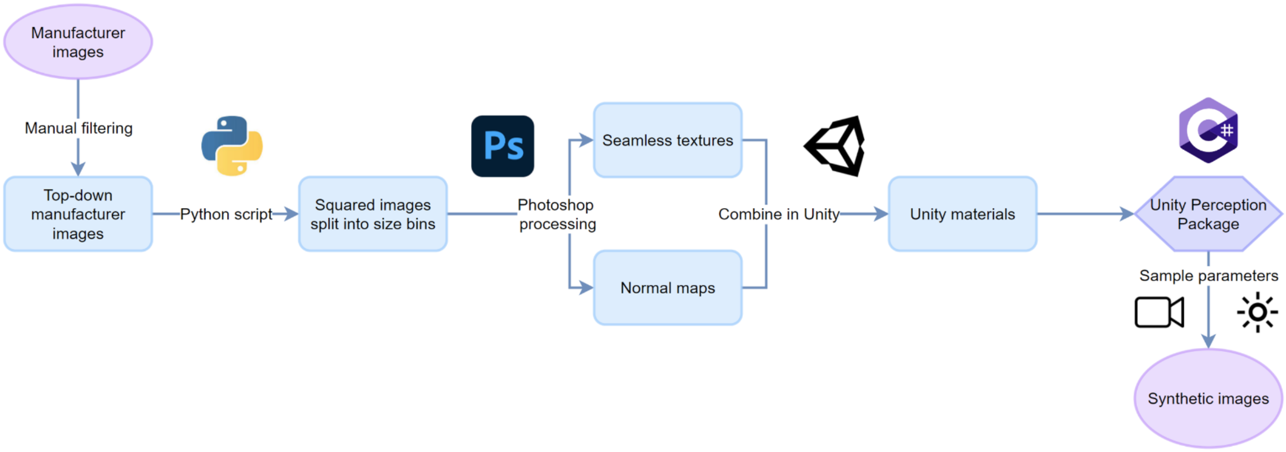

Figure 1: *A novel pipeline for generating synthetic data depicting textured surfaces. The pipeline combines manual filtering with automated image processing techniques to create seamless textures and normal maps using Python scripts and batch processing in Photoshop. The synthesized seamless textures and normal maps are then combined into materials within the Unity platform and utilised by the Unity Perception Package to generate realistic synthetic images.*

In light of the above constraints, we chose a tailored pipeline for synthetic data generation. Our pipeline aimed to generate diverse, high-quality synthetic images of carpets with varied lighting conditions and camera angles, thereby enhancing the robustness and generalisability of the machine learning model to real-world conditions. The synthetic data generation pipeline devised for this study is depicted in Figure 1. Initially, professional product photographs were obtained from various carpet manufacturers in the UK, resulting in a dataset of 3,355 manufacturer carpet images, comprising 1,356 synthetic fibre carpets, 602 wool fibre carpets, and 1,397 synthetic-wool blend carpets, this dataset is partially illustrated in Figure 2. The decision to source images from carpet manufacturers was primarily influenced by the necessity to obtain images of carpets with verified material compositions. In this context, carpet manufacturers serve as the most reliable source, offering specific information about the exact material composition of each carpet. This method ensured that we could confidently categorise the images in our initial dataset, providing a reliable foundation for our synthetic data generation pipeline.

To facilitate their integration into the synthetic data generation pipeline, images suitable for use as textures in a 3D environment were manually filtered, ensuring a top-down perspective, consistent lighting, and a tileable pattern. A total of 2,676 images met these criteria and were subsequently processed using a Python script, which cropped the largest possible square from each image and sorted them into one of four distinct size bins: 512 x 512, 1024 x 1024, 2048 x 2048, or 4096 x 4096. These size bins were chosen to align with the power-of-2-based efficiency of 3D software and contemporary graphics hardware. To allocate images to the appropriate bins, the script considered whether an image's size was within 10% of the higher bin's dimensions, in which case it was assigned to that bin; otherwise, it was assigned to the lower bin. This strategy aimed to optimise texture resolution while minimising artefacts introduced through upsampling.

The selected images were further processed in Adobe Photoshop (version 23.2.1) [Ado22] using the software's batch processing feature, converting them into seamless textures and generating corresponding normal maps. This procedure involved resizing images to match their respective bin dimensions, applying an offset with wrap-around to relocate the seam, and employing Photoshop's Content-Aware Fill [Fre11, KLW12, WC12] feature to eliminate that seam, yielding seamless textures. Subsequently, normal maps were automatically computed based on the texture's colour values using Photoshop's Generate Normal Map feature [QLSZ18, MCC22]. The entire process is illustrated in the first row of Figure 3. The default Photoshop parameters were used for all of these processing steps.

Within the Unity game engine (version 2020.3.27f1) [Uni22], individual materials were created for each texture and normal map pair. Utilising the Unity Perception Package [Uni20], synthetic data were generated by randomly selecting carpet materials, applying them to a flat surface, and sampling parameters such as the virtual camera's distance and angle, carpet rotation, and light intensity, colour temperature, and rotation. The Unity High Definition Render Pipeline (HDRP) was employed for this stage, offering enhanced realism through its physics-based lighting. Table 1 provides a comprehensive overview of the randomised parameters, and Figure 3 presents examples of synthetic images generated from a single carpet material.

The synthetic data generation process described here can be applied to other tasks requiring textured surface synthetic data generation. However, it is important to note that parameter values may need to be adjusted for the different processing steps, depending on the specific application.

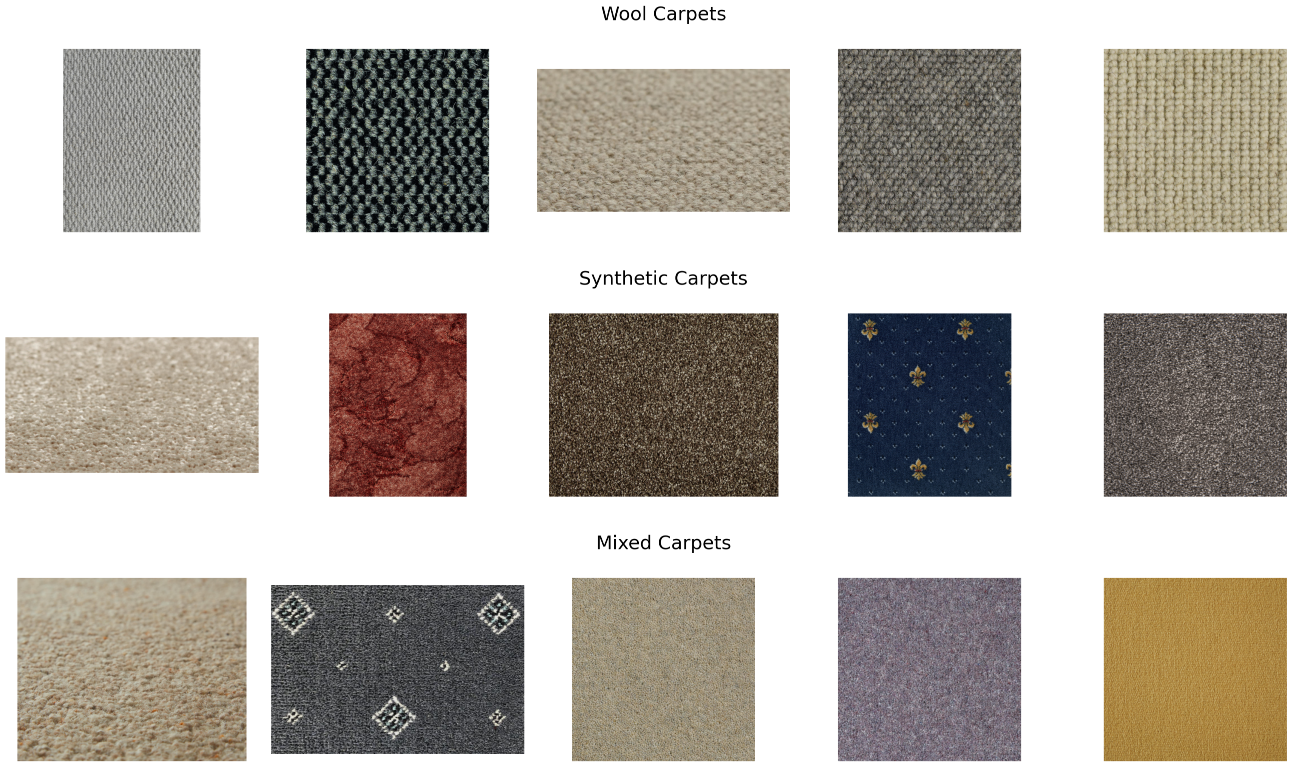

Figure 2: *A sample from the full population of the dataset of real images used in this study for classification tasks, illustrating the class types (synthetic, wool and mixed fibres).*

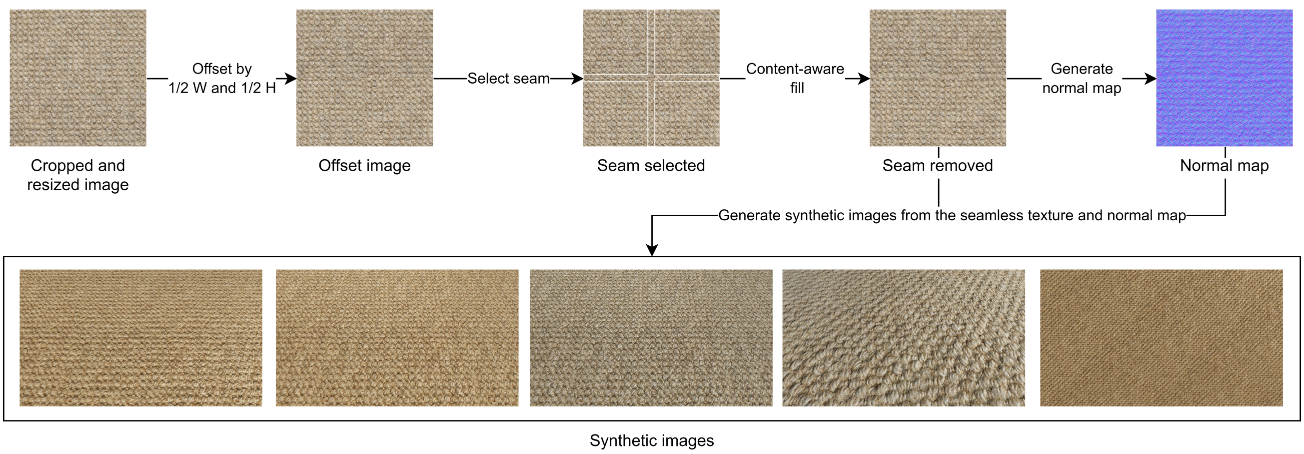

Figure 3: *The transformation of a single image as it undergoes the various stages of the processing pipeline, followed by examples of synthetic images generated from the processed image.*

3.2. Ablation Study

To evaluate the efficacy of the generated synthetic data, we conducted an ablation study centred around the task of training an image classifier to differentiate among the three aforementioned carpet categories. Our training experiments utilised an EfficientNetV2-S [TL21] image classification model, chosen for its near state-of-the-art performance on image classification benchmarks and relatively smaller size compared to other state-of-the-art

Table 1: *The randomised parameters in the simulation scenario created with the Unity Perception Package*

Sampled Parameter	Min Value	Max Value
Camera X rotation	40	90
Camera Y position	2	8
Surface Y rotation	0	360
Light intensity	40000	130000
Light temperature	5000	8000
Light X rotation	30	150
Light Y rotation	0	360

models. This allowed for expedited training and a higher number of experiments.

In our experiments, we applied transfer learning, a prevalent machine learning technique that takes advantage of pre-training to enhance a model's performance on a new task. Transfer learning has been demonstrated to improve training efficiency and accuracy, particularly when training data is scarce [YCBL14]. Foundational work on transfer learning includes that of Pan and Yang [PY10], as well as recent surveys by Weiss et al.. [WKW16] and Zhuang et al. [ZQD*20] that offer comprehensive overviews of transfer learning techniques and applications. We employed an EfficientNetV2-S pre-trained on the ImageNet-21K dataset [RDS*15] and fine-tuned only the final classification layer. To the best of our understanding, ImageNet-21K constitutes the largest freely accessible pre-training dataset suitable for transfer learning applications. Furthermore, empirical evidence suggests that a more expansive pre-training dataset invariably correlates with superior model performance, as corroborated by Ridnik et al. [RBBNZM21], which is why we decided to run our experiments with ImageNet-21K.

Regarding training hyperparameters, we largely adhered to those used by the EfficientNetV2 authors in their transfer learning experiments [TL21], with the exception of utilising a smaller batch size of 64 due to hardware limitations and a dropout rate of 0.2 in the final layer for enhanced regularization. A detailed breakdown of the hyperparameters can be found in Table 2. We ran each training experiment for 50 epochs, which proved sufficient for model convergence.

Table 2: *The hyperparameters employed in all training experiments.*

Hyperparameter	Value
Model architecture	EfficientNetV2-S
Pre-training dataset	ImageNet-21K
Batch size	64
Number of epochs	50
Learning rate	0.001
Optimizer	Adam (default parameters) [KB14]
Loss function	Categorical Cross-Entropy
Dropout rate (final layer)	0.2

We adopted a less conventional approach for fitting images to the model's input dimensions. Rather than resizing images to fit a square with the appropriate dimensions, we took a central crop of each image with the required dimensions and discarded the remaining image. This approach is based on the rationale that observing a larger area of the carpet is less crucial for material classification than observing finer details. A central crop provides increased detail at the expense of discarding parts of the image.

Moreover, we employed data augmentation during training, as it has been proven to enhance performance and robustness in machine learning models [LHW*19a, NT18, KNF*19]. Data augmentation was successfully utilised by the EfficientNetV2 authors as well [TL21], which reinforced our decision to employ this technique. It involves applying various transformations to existing images, such as rotations, translations, flips, and colour distortions, to generate new, similar images with slight variations. This helps prevent overfitting and improve generalization performance, especially when training data is limited [SK19], [PW17]. The specific augmentations applied in our training pipeline are detailed in Table 3.

Table 3: *The data augmentation applied in the training pipeline. The Python Albumentations library was utilised for these augmentations [BIK*20].*

Augmentation type	Min Value	Max Value	Probability
Rotation	-90	90	0.25
JPEG compression	80	100	0.25
Blur	3	5	0.25
Brightness	-0.1	0.1	0.25
Contrast	-0.1	0.1	0.25
Hue	-10	10	0.25
Saturation	-15	15	0.25
Value	-10	10	0.25
Flip horizontally and/or vertically			0.25

Our experimental data consisted of the dataset of 3,355 real carpet images obtained from manufacturers, used in our synthetic data generation pipeline, and 15,000 synthetic images generated by the pipeline. Of the real images, 3,166 were used for training and a class-balanced set of 189 images for testing. A validation set was deemed unnecessary, as no hyperparameter tuning was performed and all models were trained for an equal number of epochs. Synthetic images were utilised solely for training, not evaluation.

The ablation study comprised two parts. The first part investigated the accuracy of image classifiers on the testing dataset when trained on an increasing number of real images combined with a constant number of synthetic images. We conducted these experiments with 5,000 and 15,000 synthetic images, as well as with only real images and no synthetic data supplementation, to compare the accuracies. We employed 15 distinct increments of real image quantities, providing finer resolution at the lower end and more equal spacing thereafter. Table 4 displays the 15 increments along with the percentage of our full real image training dataset that each increment represents.

The second part of the ablation study involved utilising a training set capped at 3,166 examples with varying proportions of real to synthetic images to examine how the accuracy evolved as the real-synthetic ratio changed. The percentages from the same 15 steps depicted in Table 4 were used for this experiment. To achieve 5-fold cross-validation and yield more accurate results, all training experiments were executed five times, with the average result being reported.

4. Results

In this section, we showcase the outcomes of our ablation study, which was designed to evaluate the influence of synthetic data on the accuracy of an image classifier when combined with different quantities of real training data. Our investigation is divided into two segments, with the corresponding findings visualised in Figure 4 and Figure 5. The insights obtained from this study help us understand the trade-offs between real and synthetic data in the context of training image classifiers, with a focus on applications involving textured surfaces.

4.1. Impact of synthetic data on classifier performance with increasing numbers of real images

Figure 4 displays the 5-fold cross-validated results of the EfficientNetV2-S image classifier's performance when trained on increasing quantities of real images, both in combination with a constant number of synthetic images (5,000 and 15,000) and without synthetic data supplementation. Our findings reveal that the inclusion of synthetic data in the training dataset initially enhances the classifier's accuracy when the number of real images is limited. However, as the quantity of real images increases, the benefit of adding synthetic data diminishes. Eventually, at a certain threshold of real images, which for our experiments was between 316 and 474 real images, adding synthetic data begins to negatively affect the classifier's performance, and it becomes more effective to train on real data exclusively. In the comparative analysis between scenarios that incorporated either 5,000 or 15,000 synthetic images, discernible differences in performance emerged only in two specific instances: where no real data was incorporated and where the smallest subset of real data (comprising 79 examples) was used. In these particular cases, classifiers trained with the larger synthetic dataset of 15,000 images outperformed those trained with 5,000 synthetic images. For all other experimental conditions, the addition of either 5,000 or 15,000 synthetic images yielded statistically indistinguishable effects on classifier performance.

4.2. Classifier performance with varying ratios of real to synthetic images in a fixed-size dataset

Figure 5 presents the 5-fold cross-validated results of the EfficientNetV2-S image classifier's performance when trained on a dataset limited to 3,166 examples, featuring varying proportions of real to synthetic images. Our findings indicate that the classifier's accuracy exhibits a linear increase as the proportion of real data in the training dataset expands. This observation implies that, within the boundaries of a fixed dataset size, the incorporation of more real data consistently enhances the classifier's performance.

In summary, our results demonstrate that synthetic data generated through the pipeline we introduced can be advantageous in improving the accuracy of an image classifier when the amount of available real data is limited. However, as the real dataset grows, the contribution of synthetic data becomes less significant, and eventually detrimental to the classifier's performance. These findings highlight the importance of carefully considering the balance between real and synthetic data when training image classifiers for tasks involving textured surfaces, such as carpet material classification, and emphasise the need for further refinement of synthetic data generation techniques.

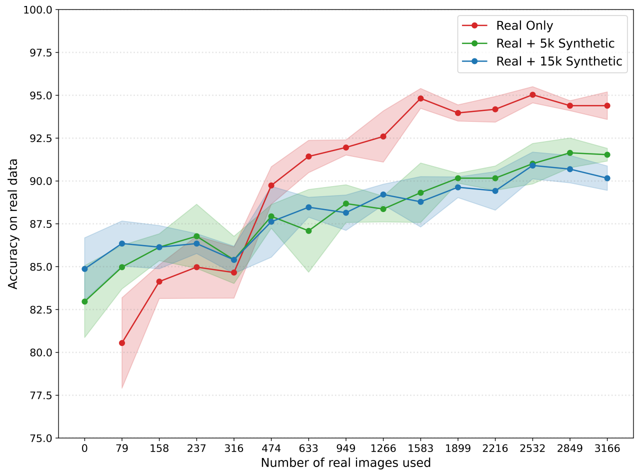

Figure 4: *Comparing the 5-fold cross-validated accuracy of the image classifier (EfficientNetV2-S) when trained on increasing numbers of real images with and without the addition of synthetic data. Note that the synthetic datasets have an additional data point because they can be run without any real data.*

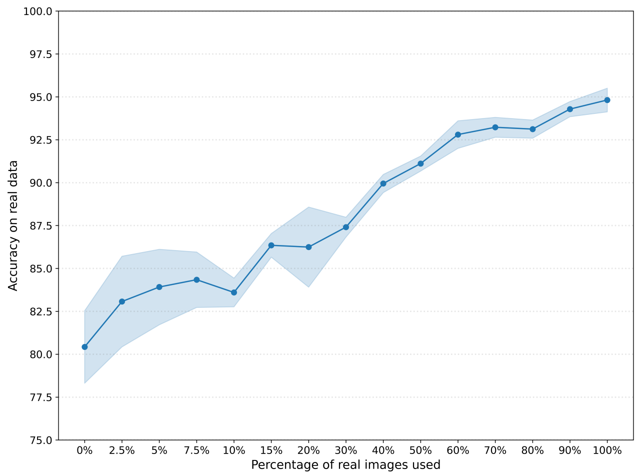

Figure 5: *Comparing the 5-fold cross-validated accuracy of the image classifier (EfficientNetV2-S) when trained on an increasing proportion of real to synthetic images in a dataset capped at 3,166 examples.*

Table 4: *The 15 distinct increments of real images utilised for the ablation study along with the percentage of the full real image training dataset that each increment represents.*

# Real images	0	79	158	237	316	474	633	949	1266	1583	1899	2216	2532	2849	3166
% Real images	0%	2.5%	5.0%	7.5%	10%	15%	20%	30%	40%	50%	60%	70%	80%	90%	100%

5. Discussion

The results of this study offer valuable insights into the role of synthetic data in image classification tasks, particularly in the context of textured surface classification. In line with the research question, we found that synthetic data is especially beneficial when there is a limited amount of real data available and further acquisition of real data is challenging or expensive. This finding aligns with previous research demonstrating the advantages of synthetic data when real data is scarce [TPA*18], [PSAS15]. However, we also found that the balance between synthetic and real data is crucial, as excessive synthetic data can hurt performance when sufficient real data is available. This is an interesting observation, and to the best of our knowledge, there are no studies that explicitly show the detrimental effects of synthetic data when ample real data is present. Our results support the hypothesis that, with an increasing number of real examples, the benefits of synthetic data may be overshadowed by imperfections in the generation pipeline.

It is essential to underline that our findings are specifically in the context of the synthetic data generated by our pipeline. One possible reason for the observed degradation in performance when using excessive synthetic data is the distortion of fine details during the synthetic data generation process. Particularly, the content-aware fill and automatically generated normal maps from Photoshop could be introducing artefacts and distortions to subtle details that affect the classifier's ability to discern between classes. For tasks where macro details are more important than micro details, the use of synthetic data generated from this pipeline might provide a greater improvement in accuracy and could even remain beneficial when a large amount of real data is available. Further research is needed to explore this relationship in greater depth.

The diminishing returns of synthetic data observed as the amount of real data increases suggest that improvements can be made to our synthetic data generation pipeline. Incorporating Generative Adversarial Networks (GANs) [GPAM*20] at various stages of the pipeline can enhance the realism and quality of the generated synthetic data. GANs could be employed for generating higher-quality textures, improving texture mapping, simulating realistic lighting conditions, and refining the final images. In addition to GANs, other advanced computer vision techniques, such as feature-based matching [BTVG06], can be used to automatically align and merge multiple 2D images of a textured surface into a single, high-quality 3D texture map. Such advancements could potentially lead to higher-quality synthetic data that captures the necessary details for accurate classification, thus mitigating the observed synthetic-data-induced performance degradation as real data quantity increases.

There are some limitations to our study that warrant discussion, as they could influence the generalisability of our results. Firstly, we employed a single image classifier (EfficientNetV2-S) and a single pretraining dataset (ImageNet-21K). Exploring alternative architectures, pretraining datasets, and fine-tuning strategies could potentially yield different results, providing a more comprehensive understanding of the trade-offs between real and synthetic data. Additionally, our study was limited to a single task with only three classes. It's conceivable that for tasks with more classes, the balance point at which adding synthetic data becomes detrimental, as shown in Figure 4, could shift. Moreover, the same data augmentation setup was utilised for all our training experiments, which is a factor that could potentially affect the balance between the benefits and drawbacks of using synthetic data. Future research should explore the interplay between data augmentation and synthetic data in image classification tasks.

6. Conclusion

In conclusion, this study presented a novel pipeline for synthetic data generation of textured surfaces and investigated its role in image classification tasks. Our findings suggest that synthetic data can significantly enhance classification performance when the amount of real data is limited. However, the balance between synthetic and real data is crucial, as excessive synthetic data can negatively impact performance when sufficient real data is available. This underscores the importance of developing better synthetic data generation pipelines that produce high-quality, realistic images that can complement real data effectively, even when the latter is abundant. While our study provides valuable insights, there are still several areas for improvement and exploration, which future research should address.

6.1. Future Work

One area to investigate is the effect of using different image classifiers, pretraining datasets, and fine-tuning strategies on the performance trade-off between real and synthetic data. This would involve examining how different classification models respond to synthetic data generated by our pipeline, and identifying which models are better suited for leveraging the benefits of synthetic data. Moreover, we plan to evaluate the impact of pretraining on the classification performance when using synthetic data, and identify the optimal fine-tuning strategy to maximise the benefits of synthetic data.

We also need to explore the relationship between the number of classes in the classification task and the balance point at which adding synthetic data becomes detrimental. We hypothesise that adding synthetic data to the training set may lead to diminishing returns when the number of classes increases beyond a certain threshold. We plan to conduct experiments with varying numbers of classes to investigate this relationship and determine the optimal balance point between real and synthetic data.

For further clarity in this area we wish to better understand the interaction and combined impact between data augmentation methods and synthetic data usage in image classification tasks. We will investigate how data augmentation and synthetic data generation can be used together to improve the performance of image classifiers, especially in scenarios where real data is limited. We anticipate that data augmentation methods can enhance the quality of synthetic data and lead to more robust image classifiers.

We should better understand the effect of enhancing the synthetic data generation pipeline with advanced techniques such as Generative Adversarial Networks (GANs) and feature-based matching to improve the quality and realism of synthetic data. This would involve exploring how GANs can be used to generate more diverse and realistic textures, and how feature-based matching can improve the alignment between synthetic and real data. We anticipate that these advanced techniques can further improve the quality of synthetic data and lead to more accurate image classifiers.

Another area to evaluate in the future is the performance of our synthetic data generation pipeline on other types of classification tasks, especially ones with a lower emphasis on small visual details, to determine its generalisability and effectiveness in a wider range of applications. This would involve testing the pipeline on datasets with different textures and objects to assess its performance in varied scenarios.

Finally, we plan to investigate the potential of the synthetic data pipeline in addressing dataset imbalance, particularly in scenarios where classes are significantly unequal in image quantity. This could help ensure fair representation and improved performance across all classes, and would involve exploring how synthetic data can be used to balance the distribution of classes in the training set.

7. Acknowledgements

GPUs used for this research were kindly donated by the NVIDIA Corporation.

References

[ABB*20] ANDREINI P., BONECHI S., BIANCHINI M., MECOCCI A., SCARSELLI F.: Image generation by gan and style transfer for agar plate image segmentation. *Computer methods and programs in biomedicine 184* (2020), 105268. 1

[Ado22] ADOBE INC.: Adobe photoshop. https://www.adobe.com/products/photoshop.html, 2022. 3

[AMSD21] ADAMS J., MURPHY E., SUTOR J., DODD A.: Assessing the qualities of synthetic visual data production. In *2021 9th International Conference on Information and Education Technology (ICIET)* (2021), IEEE, pp. 452–455. 1

[BCC*18] BARBOSA I. B., CRISTANI M., CAPUTO B., ROGNHAUGEN A., THEOHARIS T.: Looking beyond appearances: Synthetic training data for deep cnns in re-identification. *Computer Vision and Image Understanding 167* (2018), 50–62. 2

[BIK*20] BUSLAEV A., IGLOVIKOV V. I., KHVEDCHENYA E., PARINOV A., DRUZHININ M., KALININ A. A.: Albumentations: fast and flexible image augmentations. *Information 11*, 2 (2020), 125. 5

[BTVG06] BAY H., TUYTELAARS T., VAN GOOL L.: Surf: Speeded up robust features. *Lecture notes in computer science 3951* (2006), 404–417. 7

[CCY*20] CHEN X., CHEN Y., YOU B., XIE J., NAJJARAN H.: Detecting 6d poses of target objects from cluttered scenes by learning to align the point cloud patches with the cad models. *IEEE Access 8* (2020), 210640–210650. 1

[CDLA16] CORDIER N., DELINGETTE H., LÊ M., AYACHE N.: Extended modality propagation: image synthesis of pathological cases. *IEEE transactions on medical imaging 35*, 12 (2016), 2598–2608. 1

[dCBV09] DE CAMPOS T. E., BABU B. R., VARMA M.: Character recognition in natural images. In *International conference on computer vision theory and applications* (2009), vol. 1, SCITEPRESS, pp. 273–280. 2

[DHW20] DOLHASZ A., HARVEY C., WILLIAMS I.: Learning to observe: Approximating human perceptual thresholds for detection of suprathreshold image transformations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2020). 1

[DZL*21] DING S., ZHENG J., LIU Z., ZHENG Y., CHEN Y., XU X., LU J., XIE J.: High-resolution dermoscopy image synthesis with conditional generative adversarial networks. *Biomedical Signal Processing and Control 64* (2021), 102224. 2

[Fre11] FREEMAN W. T.: A perfect'match' technical perspective. *Communications of the ACM 54*, 11 (2011), 102–102. 3

[FSB*21] FINTZ M., SHOSHAN A., BHONKER N., KVIATKOVSKY I., MEDIONI G.: Synthetic data for model selection. *arXiv preprint arXiv:2105.00717* (2021). 1

[GBWC12] GE Y., BAI H., WANG J., CAO F.: Assessing the quality of training data in the supervised classification of remotely sensed imagery: A correlation analysis. *Journal of spatial science 57*, 2 (2012), 135–152. 1

[GPAM*20] GOODFELLOW I., POUGET-ABADIE J., MIRZA M., XU B., WARDE-FARLEY D., OZAIR S., COURVILLE A., BENGIO Y.: Generative adversarial networks. *Communications of the ACM 63*, 11 (2020), 139–144. 2, 7

[GRBH18] GOYAL M., RAJPURA P., BOJINOV H., HEGDE R.: Dataset augmentation with synthetic images improves semantic segmentation. In *Computer Vision, Pattern Recognition, Image Processing, and Graphics: 6th National Conference, NCVPRIPG 2017, Mandi, India, December 16-19, 2017, Revised Selected Papers 6* (2018), Springer, pp. 348–359. 2

[HFR*20] HE X., FANG L., RABBANI H., CHEN X., LIU Z.: Retinal optical coherence tomography image classification with label smoothing generative adversarial network. *Neurocomputing 405* (2020), 37–47. 1

[HNBKK15] HATTORI H., NARESH BODDETI V., KITANI K. M., KANADE T.: Learning scene-specific pedestrian detectors without real data. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2015), pp. 3819–3827. 1

[KB14] KINGMA D. P., BA J.: Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014). 5

[KLW12] KAUFMAN L., LISCHINSKI D., WERMAN M.: Content-aware automatic photo enhancement. In *Computer Graphics Forum* (2012), vol. 31, Wiley Online Library, pp. 2528–2540. 3

[KNF*19] KARIMI D., NIR G., FAZLI L., BLACK P. C., GOLDENBERG L., SALCUDEAN S. E.: Deep learning-based gleason grading of prostate cancer from histopathology images—role of multiscale decision aggregation and data augmentation. *IEEE journal of biomedical and health informatics 24*, 5 (2019), 1413–1426. 5

[LHW*19a] LEI C., HU B., WANG D., ZHANG S., CHEN Z.: A preliminary study on data augmentation of deep learning for image classification. In *Proceedings of the 11th Asia-Pacific Symposium on Internetware* (2019), pp. 1–6. 5

[LHW*19b] LEI Y., HARMS J., WANG T., LIU Y., SHU H.-K., JANI A. B., CURRAN W. J., MAO H., LIU T., YANG X.: Mri-only based synthetic ct generation using dense cycle consistent generative adversarial networks. *Medical physics 46*, 8 (2019), 3565–3581. 1

[MCC22] MOREIRA R. D., COUTINHO F., CHAIMOWICZ L.: Analysis and compilation of normal map generation techniques for pixel art. In *2022 21st Brazilian Symposium on Computer Games and Digital Entertainment (SBGames)* (2022), IEEE, pp. 1–6. 3

[MCL*18] MÜLLER M., CASSER V., LAHOUD J., SMITH N., GHANEM B.: Sim4cv: A photo-realistic simulator for computer vision applications. *International Journal of Computer Vision 126* (2018), 902–919. 1

[MFL*21] MOMENI S., FAZLOLLAHI A., LEBRAT L., YATES P., ROWE C., GAO Y., LIEW A. W.-C., SALVADO O.: Generative model of brain microbleeds for mri detection of vascular marker of neurodegenerative diseases. *Frontiers in Neuroscience 15* (2021), 778767. 2

[MIF*18] MAYER N., ILG E., FISCHER P., HAZIRBAS C., CREMERS D., DOSOVITSKIY A., BROX T.: What makes good synthetic training data for learning disparity and optical flow estimation? *International Journal of Computer Vision 126* (2018), 942–960. 1

[MQHY20] MU J., QIU W., HAGER G. D., YUILLE A. L.: Learning from synthetic animals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), pp. 12386–12395. 2

[MTM12] MØGELMOSE A., TRIVEDI M. M., MOESLUND T. B.: Learning to detect traffic signs: Comparative evaluation of synthetic and real-world datasets. In *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)* (2012), IEEE, pp. 3452–3455. 2

[NT18] NIEMI J., TANTTU J. T.: Deep learning case study for automatic bird identification. *Applied sciences 8*, 11 (2018), 2089. 5

[PCQ*20] PENG Y., CHEN S., QIN A., CHEN M., GAO X., LIU Y., MIAO J., GU H., ZHAO C., DENG X., ET AL.: Magnetic resonance-based synthetic computed tomography images generated using generative adversarial networks for nasopharyngeal carcinoma radiotherapy treatment planning. *Radiotherapy and Oncology 150* (2020), 217–224. 1

[PKM*21] PEREIRA M., KSHIRSAGAR M., MUKHERJEE S., DODHIA R., FERRES J. L.: An analysis of the deployment of models trained on private tabular synthetic data: Unexpected surprises. *arXiv preprint arXiv:2106.10241* (2021). 2

[PSAS15] PENG X., SUN B., ALI K., SAENKO K.: Learning deep object detectors from 3d models. In *Proceedings of the IEEE international conference on computer vision* (2015), pp. 1278–1286. 7

[PW17] PEREZ L., WANG J.: The effectiveness of data augmentation in image classification using deep learning. *arXiv preprint arXiv:1712.04621* (2017). 5

[PY10] PAN S. J., YANG Q.: A survey on transfer learning. *IEEE Transactions on knowledge and data engineering 22*, 10 (2010), 1345–1359. 5

[QLSZ18] QIAN K., LI Y., SU K., ZHANG J.: A measure-driven method for normal mapping and normal map design of 3d models. *Multimedia Tools and Applications 77* (2018), 31969–31989. 3

[RBBNZM21] RIDNIK T., BEN-BARUCH E., NOY A., ZELNIK-MANOR L.: Imagenet-21k pretraining for the masses. *arXiv preprint arXiv:2104.10972* (2021). 5

[RDS*15] RUSSAKOVSKY O., DENG J., SU H., KRAUSE J., SATHEESH S., MA S., HUANG Z., KARPATHY A., KHOSLA A., BERNSTEIN M., ET AL.: Imagenet large scale visual recognition challenge. *International journal of computer vision 115* (2015), 211–252. 5

[SK19] SHORTEN C., KHOSHGOFTAAR T. M.: A survey on image data augmentation for deep learning. *Journal of big data 6*, 1 (2019), 1–48. 5

[STT*19] SANTURKAR S., TSIPRAS D., TRAN B., ILYAS A., ENGSTROM L., MADRY A.: Computer vision with a single (robust) classifier. *arXiv preprint arXiv:1906.09453 4* (2019). 1

[TL21] TAN M., LE Q.: Efficientnetv2: Smaller models and faster training. In *International conference on machine learning* (2021), PMLR, pp. 10096–10106. 4, 5

[TPA*18] TREMBLAY J., PRAKASH A., ACUNA D., BROPHY M., JAMPANI V., ANIL C., TO T., CAMERACCI E., BOOCHOON S., BIRCHFIELD S.: Training deep networks with synthetic data: Bridging the reality gap by domain randomization. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops* (2018), pp. 969–977. 7

[Uni20] UNITY TECHNOLOGIES: Unity Perception package. https://github.com/Unity-Technologies/com.unity.perception, 2020. 3

[Uni22] UNITY SOFTWARE INC.: Unity. https://unity.com/, 2022. 3

[WC12] WU J.-L., CHOU Y.-Y.: An effective content-aware image inpainting method. *Journal of information science and engineering 28*, 4 (2012), 755–770. 3

[WHH18] WARD C. M., HARGUESS J., HILTON C.: Ship classification from overhead imagery using synthetic data and domain adaptation. In *OCEANS 2018 MTS/IEEE Charleston* (2018), IEEE, pp. 1–5. 1

[WKW16] WEISS K., KHOSHGOFTAAR T. M., WANG D.: A survey of transfer learning. *Journal of Big data 3*, 1 (2016), 1–40. 5

[WPV19] WEIBEL J.-B., PATTEN T., VINCZE M.: Addressing the sim2real gap in robotic 3-d object classification. *IEEE Robotics and Automation Letters 5*, 2 (2019), 407–413. 1

[YCBL14] YOSINSKI J., CLUNE J., BENGIO Y., LIPSON H.: How transferable are features in deep neural networks? *Advances in neural information processing systems 27* (2014). 5

[YZY*22] YAO Y., ZHENG L., YANG X., NAPTHADE M., GEDEON T.: Attribute descent: Simulating object-centric datasets on the content level and beyond. *arXiv preprint arXiv:2202.14034* (2022). 2

[ZLQL17] ZHU X., LIU Y., QIN Z., LI J.: Data augmentation in emotion classification using generative adversarial networks. *arXiv preprint arXiv:1711.00648* (2017). 2

[ZQD*20] ZHUANG F., QI Z., DUAN K., XI D., ZHU Y., ZHU H., XIONG H., HE Q.: A comprehensive survey on transfer learning. *Proceedings of the IEEE 109*, 1 (2020), 43–76. 5

[ZVSL18] ZOPH B., VASUDEVAN V., SHLENS J., LE Q. V.: Learning transferable architectures for scalable image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2018), pp. 8697–8710. 1