Impact of Advertial Data Augmentation on Robust Indoor Scene Recognition: An Evaluation Study

ANDREA ONOFREI, University of Twente, The Netherlands

Indoor scene recognition is an emerging technology with significant potential for smart homes, robotics, and virtual/augmented reality applications. However, the robustness of indoor scene recognition algorithms against adversarial attacks is a significant concern for their practical deployment. This research project aims to investigate the impact of adversarial data augmentation on the robustness of an indoor scene recognition algorithm. We develop an indoor scene recognition model based on image captioning and comprehensively analyse its accuracy in classifying indoor scenes. Subsequently, we employ data augmentation techniques, specifically image superimposition on both the test and training datasets. By superimposing five images featuring diverse objects to our dataset, such as a Christmas tree, airplane, monkey, train, and palm tree, we aim to evaluate the model's reaction to noisy input and assess its ability to generalize in the presence of unexpected objects within indoor scenes. Moreover, by training a model on superimposed and standard images, we aim to evaluate whether the new model has enhanced regularisation, transfer learning capabilities, and noise tolerance.

Additional Key Words and Phrases: Indoor scene recognition, Robustness, Image captioning, Image superimposition

1 INTRODUCTION

Indoor scene recognition is a field of study that involves developing algorithms to recognise different indoor environments. In this study, we propose a novel approach for indoor scene recognition by adopting the high-level context representation method proposed in the paper [5]. Specifically, our algorithm for indoor scene recognition is based on captioning the image and subsequently utilising the high-level context representation for classification purposes.

Goal 1 To implement an indoor scene recognition algorithm based on image captioning, using a subset of the **Places365 dataset** of indoor scene images by adopting the high-level context representation approach proposed in the paper [5]

Goal 2 To augment the test dataset by **superimposing** to the test pictures images with objects that are not typically found indoors in order to test the robustness and accuracy of the indoor scene recognition algorithm

Goal 3 To compare the performance of the initial model, trained exclusively on standard data, with that of an alternative model trained on a blended dataset comprising 80% standard data and 20% superimposed data.

Hence, the following research questions (RQs) emerge:

RQ1 How effective is the implementation of an indoor scene recognition algorithm based on image captioning using a subset of the Places365 dataset, considering the adoption of the high-level context representation approach proposed in [5]?

RQ2 What is the impact of augmenting the test dataset by superimposing images with objects not typically found indoors on the robustness and accuracy of the indoor scene recognition algorithm?

RQ3 How does the performance of the initial model, trained exclusively on standard data, compare to that of an alternative model trained on a blended dataset comprising 80% standard data and 20% superimposed data?

The subsequent sections of this paper are organized as follows: **Section 2** provides an overview of the Related Work related to our research, highlighting prior studies and developments in the field of indoor scene recognition, image captioning, and image superimposition. **Section 3** focuses on dataset creation. **Section 4** details the methodology adopted for this study, elucidating the steps taken to implement the indoor scene recognition algorithm based on image captioning. **Section 5** of the paper presents the experimental details, highlighting our specific choices to optimize the training process and ensure the reproducibility of the experiments. **Section 6** presents the obtained results and offers an in-depth discussion of the findings. The performance of the implemented algorithm, including its robustness and accuracy in classifying indoor scenes, is thoroughly analyzed and interpreted. Furthermore, the impact of augmenting the test dataset with superimposed images is examined and discussed. Finally, **Section 7** summarizes the key conclusions drawn from this research. It highlights the implications of the findings in relation to the goals outlined earlier. **Section 8** outlines future work that can be done, suggesting areas where further exploration and refinement can contribute to the advancement of indoor scene recognition algorithms.

2 RELATED WORK

To gather relevant literature related to the research domain, the databases Scopus, Google Scholar, and IEEE were utilized. By employing search terms such as "Indoor scene recognition", "Image Captioning" and "Data Augmentation Techniques", numerous scholarly documents were identified, which had previously explored these areas of interest. There are several approaches to indoor scene recognition, including utilizing object representations, Bayesian object relation models, and deep learning-based algorithms [14][18][11][10]. These algorithms can be used for various applications, such as mobile robot navigation and indoor localization[18][10][1]. Some studies have proposed using a combination of different algorithms to improve recognition accuracy[11]. The mentioned studies present an interesting opportunity to perform a comparative analysis of their algorithms with the proposed implementation in terms of accuracy and performance. During the literature review process focusing on image captioning for indoor scene recognition, a limited number of relevant studies were found. Only one paper [13] was identified.

TScIT 39, July 7, 2023, Enschede, The Netherlands

The study proposes a framework for indoor scene captioning from streaming video. As for Data Augmentation, several papers propose useful techniques. In paper [4], the authors investigate techniques to improve the robustness of the solution to motion blur using training data augmentation at each or both stages of the solution, i.e., object detection and captioning, and observe improved results. Paper [16] does quantitative analysis of various basic image manipulation techniques, suggesting ways of adding noise and objects to images. The paper [3] proposes a technique that can be used to improve model accuracy and make networks more robust to adversarial attacks: it creates new image data based on image/label pairs, where a patch from one of the two images in the pair is superimposed on to the other image, creating a new augmented sample. We would use a similar mechanism, but instead of superimposing only a patch of the image, we will overlay the entire image while re-scaling it to a reduced size.

Fig. 1. **Raw Caption**: A woman taking a picture of herself in a bathroom mirror.
Processed Caption: take picture bathroom mirror

Fig. 2. **Raw Caption**: A bar with a bunch of bottles of wine.
Processed Caption: bar bunch bottle wine

3 DATASET CREATION

3.1 Standard train and test dataset

A dataset comprising indoor scene images was compiled. The dataset was specifically narrowed down to a subset of seven distinct scene categories, namely bar, bathroom, bedroom, classroom, kitchen, living room, and movie theatre. These categories were extracted from the Places365 dataset, each consisting of 5000 images. For training the baseline model, we divided the dataset into train (80% of the images) and test (20% of the images).

3.2 Superimposed train and test dataset

To evaluate the robustness of our models, we augment the standard test dataset using image superimposition techniques. We conduct testing on nine distinct test datasets, which consist of the original images from the standard test datasets superimposed with additional images. These additional images represent either one, three, or five pictures of varying sizes (10x10 pixels, 20x20 pixels, and 40x40 pixels), depicting various objects such as a Christmas tree, airplane, monkey, train, and palm tree. For a detailed view of the images that were utilized for superimposition, please refer to the Appendix - Table 3. We are training the alternative model on a dataset comprising 80% standard data and 20% superimposed data. The superimposed training data were divided into nine groups of 889 images(in total, for training, we use 40000 images; for the superimposed data, we use only 20% of it, so 8000 images, which split into nine groups gives approximately 889 images), on which we superimposed either one, three, or five pictures of varying sizes (10x10 pixels, 20x20 pixels, and 40x40 pixels). Figures 3,4,5 depict what pictures with superimposed images look like and what captions our model returns. Fig 8 represents the original image from the dataset and its caption. In the appendix, we present Word Clouds of the simple and augmented test datasets to show the difference between the frequency of words.

Fig. 3. **Raw Caption**: A woman standing in a room with her hand.
Processed Caption: stand room hand

Fig. 4. **Raw Caption**: A woman holding her face in front of a crowd.
Processed Caption: face crowd

Fig. 5. **Raw Caption**: A woman holding a green sign in a room
Processed Caption:hold green sign room

Fig. 6. **Raw Caption**: A woman standing in front of a crowd of people.
Processed Caption: stand crowd

Fig.3. represents the original picture from the Places365 dataset, category - bar. Fig.4. has three images of size 20x20 added. We can notice that the augmented objects are visible, but the caption does not notice them. Instead, the fact that the Christmas tree was added to the woman's face made the model interpret it as the woman holding her face.Fig.5. has one image of 40x40 pixels added to it. The model notices the object but does not recognize it, mentioning only its colour. In Fig.6. five images of size 10x10 pixels were added. We can notice that even though the objects are not really visible, it still influences the caption.

4 RESEARCH METHODOLOGY

This section describes the methodology employed in this research project to investigate the impact of adversarial data augmentation on the robustness of an indoor scene recognition algorithm. We adopt the approach proposed in the paper [5] adapting it to the task of indoor room recognition. In figure 4 we represent the adapted architecture for indoor room recognition.

4.1 Extraction of high-level descriptions

For the image captioning task, we are using ExpansionNet-v2, an image captioning model based on the Swin-Transformer architecture [8]. This model was utilized to generate captions for the images. To enhance the quality of the captions, a refinement process was applied to the initial raw captions. This involved the elimination of stop words, which are commonly used words in a language (such as articles, prepositions, and pronouns) that lack semantic significance. Including these words in the caption would introduce unnecessary complexity due to their frequent occurrence in the English language. Therefore, their removal resulted in a more representative corpus. To achieve this, we used spaCy, a publicly available library for natural language processing.

Furthermore, we excluded common nouns such as "man," "woman,", "people", etc. from the refined captions as these terms do not contribute to the scene recognition process. Additionally, lemmatization was applied, a linguistic technique that reduces words to their base or root forms. The remaining words, referred to as valid words, were employed in subsequent steps for generating data representations.

To provide a visual illustration of the captioning process, Figures 1 and 2 showcase examples of images along with their original captions and the corresponding processed captions.

4.2 Co-occurrence mining

The second step in our methodology involves generating co-occurrence matrices to capture patterns between the labels within the dataset, which will be utilized for future conditional probability calculations. After pre-processing the captions in the dataset, we store this information and count the occurrences of each indoor room scene and the valid words within each caption. This process results in a matrix $M_r \in \mathbb{N}^{W \times R}$, where W represents the number of valid words extracted from the corpus and R represents the number of indoor rooms in the dataset. Thus, the element $M_{r_{ij}}$ in this matrix denotes the number of instances where indoor class R_j occurs in conjunction with the valid word W_i. We refer to this matrix as the labels co-occurrence matrix.

Similarly, we generate another co-occurrence matrix based on the co-occurrence of valid words themselves. Using a sliding window of size 3, we capture the co-occurrence patterns among the valid words, resulting in a matrix $M_w \in \mathbb{N}^{W \times W}$. In this matrix, the element $M_{w_{ij}}$ represents the number of times the valid word W_i appears together with the valid word W_j.

To facilitate a clear understanding, we present in Fig. 7 a heatmap illustrating the occurrence matrix of labels and some selected words. These co-occurrence matrices provide valuable insights into the relationships and associations between indoor scenes and valid words, as well as the co-occurrence patterns among the valid words themselves. They serve as essential resources for graph generation.

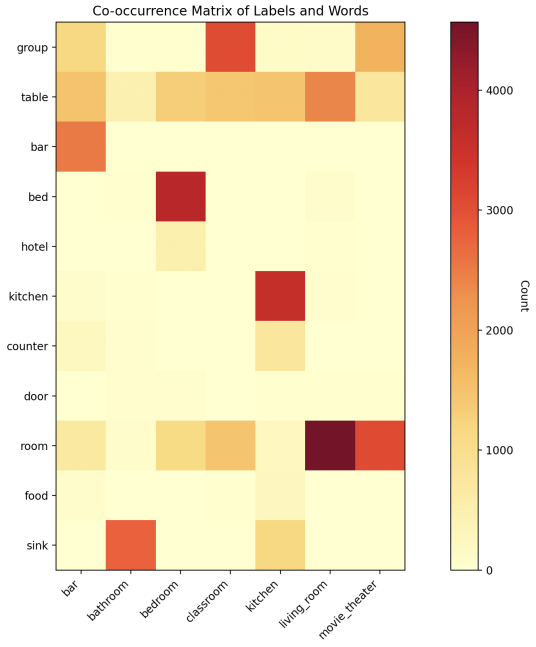

Fig. 7. Heatmap of labels co-occurrence matrix for the training dataset. The X-axis represents the indoor scene categories, and the Y-axis represents popular words in the training dataset and their count in the captions of each of the categories. For example, in the class movie-theatre, there are more than 3000 captions that contain the word "room".

4.3 Graph generation

After obtaining the co-occurrence matrices, we were able to encapsulate the relationships between valid words and indoor scenes, as well as among the words themselves. Recognizing the efficacy of graphs in representing structured data, we choose to model the representations that we have using graphs. Similar to the paper [5] we use the Deep Graph Library, a framework-agnostic library for generating and manipulating graphs.

We begin by constructing an empty graph $G = (V, E)$, where V is a set of nodes and E is a set of edges. In this case, $V = E = \{\emptyset\}$. For each valid word W, we add a new node V_{W_i} to the set of nodes V of the graph. We employ GloVe [15] to obtain the valid word embedding, which is used as the feature $X \in \mathbb{R}^{50}$ for node V_{W_i}. In the event that the valid word is not present in GloVe, we randomly sample the embedding from a uniform distribution $[-0.01, 0.01]$. We store this representation for future use in case the valid word reappears.

Subsequently, we add a node V_C for each indoor room location. As there are seven possible rooms, we introduce seven nodes and establish edges $e = (V_W, V_{C_i})$ between the valid word and each class. We define the weight w_e according to the equation below:

$$w_e = P(C_i|W) = \frac{M_{C_{W,i}}}{\sum M_{C_W}}$$

Here, the edge weight w_e between the valid word node V_W and the i_{th} indoor category C_i is denoted by $P(C_i|W)$. This value is computed by dividing the co-occurrence between the valid word W and the indoor category C_i by the sum of the co-occurrence between the valid word W and all possible categories C extracted from the co-occurrence matrix M.

4.4 Deep GCN for Indoor Scene Recognition

For Indoor Scene Recognition, we employ a deep graph convolutional neural network that classifies the graphs and, consequently, the images that they represent.

Let $G_1, \ldots, G_N$ represent a set of graphs and $C \in \mathbb{R}^7$ denote the set of indoor room categories. The goal is to classify each graph according to its corresponding indoor room category. To achieve this, we use the adapted Graph Isomorphism Network (GIN), proposed in paper [5] due to its simple architecture. In the proposed approach, the input graph's features are directly stored in the hidden representations stack as h_0. Subsequently, we iterate through a GIN convolutional block that consists of a Graph Isomorphism Network layer (2), batch normalization (3), and Rectified Linear Unit activation (4). This block is repeated five times, generating representations h_1 to h_5 (5). We perform average pooling on these hidden representations to reduce their dimensionality (6). Finally, after passing through single-layer feed-forward networks (7) with a selected dropout = 0.5 (8) the model outputs the classification labels (9). The numbers in brackets represent the steps presented in Figure 8. During training, we learn the categorical labels, and our loss function is defined as a weighted combination of individual losses from each output. Considering a prediction $\hat{y}$cat where $\hat{y}$cat $\in \mathbb{R}^C$, the loss for this prediction is denoted as $L = \lambda_{\text{cat}} L_{\text{cat}}$, where L_{cat} represents the loss of each individual prediction. For L_{cat}, we employ a weighted Euclidean loss, defined as follows:

$L_{2_{\text{cat}}}(\hat{y}_{\text{cat}}) = \sum_{i=1}^{7} w_i(\hat{y}_{\text{cat}_i} - y_{\text{cat}_i})^2$

where $\hat{y}_{\text{cat}i}$ represents the prediction for the i-th category and $y\text{cat}_i$ is the corresponding ground-truth label. The loss weights are assigned an equal weight of 1/7.

5 EXPERIMENTS

5.1 Validation Metrics

For analyzing the performance and robustness of the classification models we use metrics of Accuracy, Precision, Recall, F1-Score and confusion matrices. In the appendix section, we describe in more detail what each metric represents and present the confusion matrix for all the tests.

5.2 Implementation details

To get the best outcome for the training, we did an empirical comparison of different optimizers: RMSprop [7], Adam [9], Adadelta [17], AdaGrad [6]. Also, we experimented with several batch sizes for the data loaders. We obtained the best results for the model trained with the optimizer RMSprop with a learning rate of 0.001, weight decay of 0.0003 and batch size set to 8. Regarding the experimentation environment, we are using version 2.0.1 of the PyTorch framework in a conda environment, and we train and test our model on a Macbook M1 Pro with 16 GB RAM.

Dataset	Accuracy	Precision	Recall	F1-Score
Test dataset	0.88	0.883	0.88	0.88
Test dataset repeated	0.88	0.884	0.88	0.88
Test + 1x10	0.586	0.832	0.586	0.55
Test + 1x20	0.582	0.824	0.582	0.55
Test + 1x40	0.583	0.83	0.583	0.55
Test + 3x10	0.589	0.829	0.589	0.557
Test + 3x20	0.585	0.828	0.585	0.554
Test + 3x40	0.574	0.828	0.574	0.548
Test + 5x10	0.59	0.829	0.59	0.561
Test + 5x20	0.587	0.827	0.587	0.557
Test + 5x40	0.574	0.826	0.574	0.552

Table 1. Performance Metrics for the baseline model

Dataset	Accuracy	Precision	Recall	F1-Score
Test dataset	0.282	0.69	0.282	0.18
Test + 1x10	0.368	0.82	0.36	0.29
Test + 1x20	0.36	0.82	0.36	0.29
Test + 1x40	0.36	0.82	0.36	0.29
Test + 3x10	0.37	0.82	0.37	0.29
Test + 3x20	0.37	0.78	0.37	0.29
Test + 3x40	0.36	0.78	0.36	0.29
Test + 5x10	0.37	0.70	0.37	0.29
Test + 5x20	0.36	0.70	0.36	0.29
Test + 5x40	0.36	0.78	0.36	0.29

Table 2. Performance Metrics for the model trained on superimposed images

In Tables one and two, the inclusion of a "+ digit 'x' number" indicates the number of images that we are going to augment to the initial image(1,3 or 5 images) 'x' the size of the image that we are going to superimpose (10x10,20x20 or 40x40 pixels)

6 RESULTS AND DISCUSSIONS

6.1 ANSWER TO RQ 1

The baseline model, trained on standard images, exhibited a good performance on unaltered images, achieving high levels of accuracy (0.88). The precision, recall and F1-score values are also consistently high for the test dataset, indicating a good balance between correctly identifying positive and negative samples. The experiment was repeated, yielding similar results(Test dataset repeated in Table 1). Due to the absence of state-of-the-art models based on image captioning for indoor scene recognition, direct performance comparisons are limited. Nonetheless, we can compare the results with the paper [18], which presents a novel model BORM (Bayesian Object Relation Model), that utilizes meaningful object representations for indoor scene recognition. The proposed model has an accuracy of 83.1% on a subset of 7 classes of the Places365 dataset; however, combined with the PlacesCNN model, it achieves an accuracy of 90.1%.

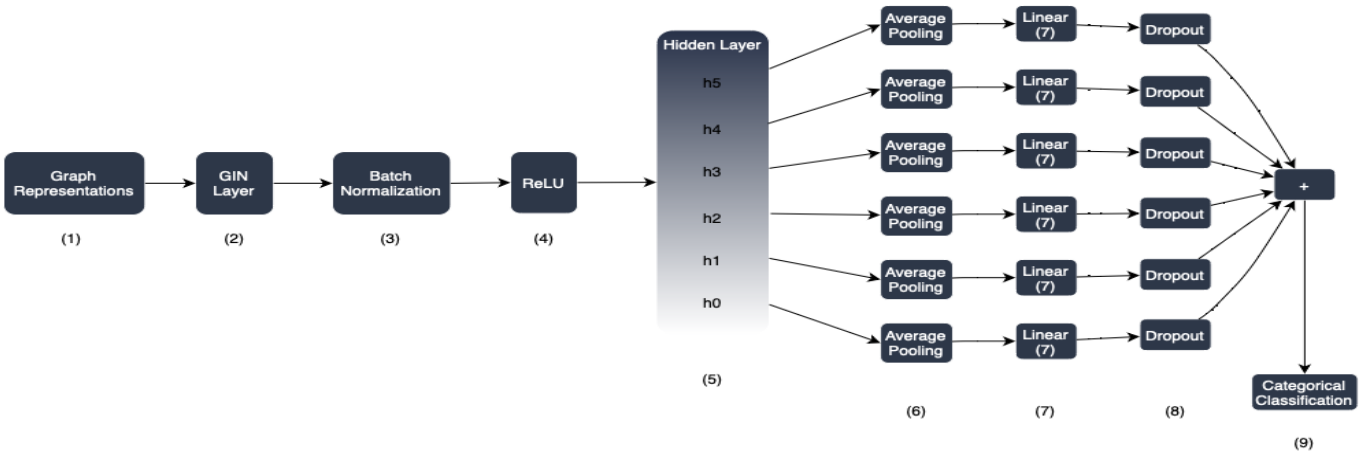

Fig. 8. Adapted architecture for indoor room recognition from the paper [5].

Moreover, from this paper, we can find out the accuracies of other state-of-the-art models on the subset of 7 classes of the Places365 dataset. According to [18], the accuracy of the PlacesCNN model with the base architecture ResNet18 as a backbone network is 80.4%, 82.7% for ResNet50 and 87.3 for DEDUCE. Accuracy reported in the paper [2] for a CNN-based indoor scene recognition model tested on 24,000 indoor home scene images was 97.14%. The comparison with state-of-the-art models indicates that our model performs on par with them.

6.2 ANSWER TO RQ 2

The impact of augmenting the test dataset by superimposing images with objects not typically found indoors on the robustness and accuracy of the baseline indoor scene recognition algorithm is significant. When superimposed images are introduced during testing, the model's performance deteriorates notably. The accuracy drops to around 0.58-0.59 for different configurations of superimposed images, and the precision, recall, and F1-score values also decrease substantially. This suggests that the model struggles to handle the variations and complexities introduced by the superimposed images.

6.3 ANSWER TO RQ 3

The performance of the initial model, trained exclusively on standard data, is significantly better than that of an alternative model trained on a blended dataset comprising 80% standard data and 20% superimposed data. The alternative model trained on superimposed images encounters challenges in accurately recognizing and classifying indoor scenes. It achieves an accuracy of 0.28% for the test dataset and around 0.365% for the augmented test dataset.

The Appendix-C contains a comprehensive collection of confusion matrices for all the combinations of datasets and models that we considered in our research. It also contains insightful observations regarding the outcomes derived from these matrices. Notably, the baseline model exhibits a pronounced bias towards the first class, namely "bar," when subjected to augmented test datasets. Conversely, the alternative model showcases a distinctive bias towards the "kitchen" class in its response patterns.

6.4 POSSIBLE REASONS FOR THE PERFORMANCE OF THE BASELINE MODEL

6.4.1 ***Captioning model exhibits a peculiar response to data superimposition***. Upon analyzing the captions generated for the augmented test dataset, it was observed that the presence of superimposed images **did not lead to substantial changes in the frequency of appearance for objects depicted in those images** (e.g., monkey, train, airplane, palm tree, etc.). However, **notable variations were observed in the content and context of the generated captions**. For example, an image of a tree superimposed with images of people dancing resulted in a caption such as "A woman holding a green sign in a room" (Fig. 5). The generated caption does not explicitly acknowledge the presence of the dancing individuals but instead responds to the superimposed objects by associating the concept of a woman holding a sign in a room. Given the model's reliance on captions as the basis for co-occurrence matrices and graphs, it is possible to discern a potential explanation for the observed decline in performance metrics. In Appendix B, we investigate the word clouds and do not notice any drastic change in the frequency of the words; only as the pictures become larger, some limited number of words appear. However, as we can notice in Table 1, this cannot be the reason for the drastic decrease in performance - from 0.88 % to 0.57%, as these words, e.g. teddy, Christmas do not appear in the Word cloud for the augmented test datasets 1x10.

6.4.2 ***Limited generalisation***. The model trained on the baseline dataset might not have learned robust representations that can effectively generalize to the variations present in the superimposed images. This lack of generalization can result in decreased performance when encountering new, augmented data.

6.4.3 ***Limited training with superimposed images***. The model might not have been exposed to sufficient training examples with

superimposed images. Training the model on a more diverse dataset that includes standard and superimposed data could improve its ability to handle such variations. However, as we can see from the results in Table 2, this does not seem to be the case.

6.5 Possible reasons for the performance of the model trained on superimposed images

6.5.1 ***Lack of contextual understanding of the captioning model.*** The model's performance may be hindered by the limited ability of the captioning model to understand the context and relationships between objects in the superimposed images. Superimposition creates complex visual scenes where objects interact or occlude with each other, making it challenging for the model to interpret and classify the individual components accurately. Without a comprehensive understanding of the context, the model may struggle to make precise predictions.

6.5.2 ***Inadequate training strategy.*** The training strategy employed for the indoor recognition model trained on superimposed images may not have been optimal. The model may require specific techniques or adjustments to effectively learn from and adapt to the complexities introduced by superimposed images. The current training approach may not have adequately addressed the challenges posed by the combination of multiple overlapping objects.

7 CONCLUSIONS

In conclusion, this study has provided a comprehensive investigation into the impact of superimposed images on the performance of indoor scene recognition models. Through rigorous experimentation and thorough analysis, we have obtained valuable insights into the inherent challenges associated with the presence of superimposed objects and their ramifications for model performance.

The baseline model, trained on standard images, exhibited commendable performance on unaltered images, achieving high levels of accuracy, on par with state-of-the-art models and maintaining a well-balanced trade-off between precision and recall.

Nevertheless, when subjected to superimposed images, the baseline model's performance experienced a significant decline, highlighting its difficulty in effectively handling the intricate variations and complexities introduced by the superimposed objects. Several factors have been identified as possible explanations for this decline, including the limited responsiveness of the captioning model to superimposed data and its limited generalization capability.

Furthermore, the model trained on superimposed images encountered challenges in accurately recognizing and classifying indoor scenes. The model's constrained contextual understanding and deficiencies in the training strategy are plausible factors contributing to its suboptimal performance.

Several avenues for future research can be explored to further improve the performance of indoor scene recognition models in the context of superimposed images. In the following section, we discuss potential directions for future work.

8 FUTURE WORK

Although this research project provided valuable insights into the impact of adversarial data augmentation on indoor scene recognition, there are several avenues for future exploration and refinement:

Exploration of different data augmentation techniques: While this study focused on superimposing images with non-conventional indoor objects, there are other data augmentation techniques that can be investigated. Techniques such as adding noise to the image(e.g salt and pepper), rotation, scaling, and translation could be explored to introduce additional variations in the dataset and further enhance the model's robustness.

Investigation of different image captioning models: The implementation of indoor scene recognition based on image captioning utilized the ExpansionNet-v2 model. However, it proved susceptible to image superimposition. So for future work other state-of-the-art image captioning models could be considered, such as those based on transformer architectures [12]. Comparing the performance of different models could provide insights into the most effective approach for this task.

Integration of other modalities: In addition to visual information, indoor scene recognition can benefit from the integration of other modalities, such as audio and depth data. Exploring the fusion of multiple modalities could enhance the model's performance and enable more comprehensive scene understanding. By addressing these future research directions, we can further advance the field of indoor scene recognition and contribute to the development of robust and practical algorithms for applications in smart homes, robotics, and virtual/augmented reality.

9 ACKNOWLEDGEMENT

I would like to express my sincere gratitude to Estefanía Talavera Martínez for her invaluable guidance and mentorship as my supervisor throughout this project. Additionally, I extend my appreciation to Willams de Lima Costa for his valuable assistance and contributions during the course of this project. His insights and assistance have been instrumental in the successful completion of this work.

REFERENCES

[1] B Anbarasu and G Anitha. 2021. Vision-based Position estimation and Indoor scene recognition algorithm for Quadrotor Navigation. *Journal of Physics: Conference Series* 1969, 1 (July 2021), 012001. https://doi.org/10.1088/1742-6596/1969/1/012001

[2] Amlan Basu, Keerati Kaewrak, Lykourgos Petropoulakis, Gaetano Di Caterina, and John J. Soraghan. 2022. Indoor Home Scene Recognition through Instance Segmentation Using a Combination of Neural Networks. In *2022 IEEE World Conference on Applied Intelligence and Computing (AIC)*. IEEE, Sonbhadra, India, 167–173. https://doi.org/10.1109/AIC55036.2022.9848982

[3] Marcus D. Bloice, Peter M. Roth, and Andreas Holzinger. 2019. Patch augmentation: Towards efficient decision boundaries for neural networks. http://arxiv.org/abs/1911.07922 arXiv:1911.07922 [cs, stat].

[4] Shashank Bujimalla, Mahesh Subedar, and Omesh Tickoo. 2021. Data augmentation to improve robustness of image captioning solutions. https://doi.org/10.48550/arXiv.2106.05437 arXiv:2106.05437 [cs].

[5] Willams de Lima Costa, Estefania Talavera Martınez, Veronica Teichrieb, and Lucas Silva Figueiredo. 2023. High-level context representation for emotion recognition in images. In *Google Docs*. https://drive.google.com/file/d/11-KTQA9W104lV3ZtgZOf5FBKSmBIRn_8/view?usp=embed_facebook

[6] John Duchi, Elad Hazan, and Yoram Singer. 2011. Adaptive Subgradient Methods for Online Learning and Stochastic Optimization. *Journal of Machine Learning Research* 12, 61 (2011), 2121–2159. http://jmlr.org/papers/v12/duchi11a.html

[7] Alex Graves. 2014. Generating Sequences With Recurrent Neural Networks. http://arxiv.org/abs/1308.0850 arXiv:1308.0850 [cs].

[8] Jia Cheng Hu, Roberto Cavicchioli, and Alessandro Capotondi. 2022. ExpansionNet v2: Block Static Expansion in fast end to end training for Image Captioning. http://arxiv.org/abs/2208.06551 arXiv:2208.06551 [cs].
[9] Diederik P. Kingma and Jimmy Ba. 2017. Adam: A Method for Stochastic Optimization. http://arxiv.org/abs/1412.6980 arXiv:1412.6980 [cs].
[10] Boney Labinghisa and Dong Myung Lee. 2021. A Deep Learning based Scene Recognition Algorithm for Indoor Localization. *2021 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)* (April 2021), 167–170. https://doi.org/10.1109/ICAIIC51459.2021.9415278 Conference Name: 2021 International Conference on Artificial Intelligence in Information and Communication (ICAIIC) ISBN: 9781728176383 Place: Jeju Island, Korea (South) Publisher: IEEE.
[11] Camila Laranjeira, Anisio Lacerda, and Erickson R. Nascimento. 2019. On Modeling Context from Objects with a Long Short-Term Memory for Indoor Scene Recognition. *2019 32nd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)* (Oct. 2019), 249–256. https://doi.org/10.1109/SIBGRAPI.2019.00041 Conference Name: 2019 32nd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI) ISBN: 9781728152271 Place: Rio de Janeiro, Brazil Publisher: IEEE.
[12] Chenliang Li, Haiyang Xu, Junfeng Tian, Wei Wang, Ming Yan, Bin Bi, Jiabo Ye, He Chen, Guohai Xu, Zheng Cao, Ji Zhang, Songfang Huang, Fei Huang, Jingren Zhou, and Luo Si. 2022. mPLUG: Effective and Efficient Vision-Language Learning by Cross-modal Skip-connections. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 7241–7259. https://aclanthology.org/2022.emnlp-main.488
[13] Xinghang Li, Di Guo, Huaping Liu, and Fuchun Sun. 2021. Robotic Indoor Scene Captioning from Streaming Video. *2021 IEEE International Conference on Robotics and Automation (ICRA)* (May 2021), 6109–6115. https://doi.org/10.1109/ICRA48506.2021.9560904 Conference Name: 2021 IEEE International Conference on Robotics and Automation (ICRA) ISBN: 9781728190778 Place: Xi'an, China Publisher: IEEE.
[14] Hirokazu Madokoro, Hanwool Woo, Stephanie Nix, and Kazuhito Sato. 2020. Benchmark Dataset Based on Category Maps with Indoor–Outdoor Mixed Features for Positional Scene Recognition by a Mobile Robot. *Robotics* 9, 2 (May 2020), 40. https://doi.org/10.3390/robotics9020040
[15] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, Qatar, 1532–1543. https://doi.org/10.3115/v1/D14-1162
[16] Mohammed Ehsan Ur Rahman, Hrudheeshta Anishetty, Arjun Kumar Kollpaka, Aishwarya Yelishetty, and Swetha Reddy Ganta. 2021. A Quantitative Analysis of Basic vs. Deep Learning-based Image Data Augmentation Techniques. *2021 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES)* (Sept. 2021), 1–9. https://doi.org/10.1109/ICSES52305.2021.9633781 Conference Name: 2021 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES) ISBN: 9781665435215 Place: Chennai, India Publisher: IEEE.
[17] Matthew D. Zeiler. 2012. ADADELTA: An Adaptive Learning Rate Method. http://arxiv.org/abs/1212.5701 arXiv:1212.5701 [cs].
[18] Liguang Zhou, Jun Cen, Xingchao Wang, Zhenglong Sun, Tin Lun Lam, and Yangsheng Xu. 2021. BORM: Bayesian Object Relation Model for Indoor Scene Recognition. https://doi.org/10.48550/arXiv.2108.00397 arXiv:2108.00397 [cs].

A METRICS

Accuracy: Accuracy measures the overall correctness of the model's predictions. It calculates the ratio of correct predictions to the total number of predictions made. This metric provides a general assessment of the model's performance regarding correctly classified instances. **Precision**: Precision focuses on the proportion of correctly predicted positive instances out of all instances predicted as positive. It helps evaluate the model's ability to minimize false positives, which is especially important in scenarios where misclassifying positive instances can have significant consequences. **Recall**: Recall, also known as sensitivity or true positive rate, measures the proportion of correctly predicted positive instances out of all actual positive instances. It assesses the model's ability to identify all relevant positive instances, without missing any. **F1-Score**: The F1-Score combines precision and recall into a single metric, providing a balanced evaluation of the model's performance. It calculates the harmonic mean of precision and recall, which gives equal weight to both metrics. The F1-Score is useful when there is an imbalance between the number of positive and negative instances in the dataset. By considering these metrics together, we gain a comprehensive understanding of the model's robustness in terms of its overall accuracy, ability to minimize false positives, ability to capture all relevant positives, and the trade-off between precision and recall.

Christmas Tree	Airplane	Monkey	Train	Palm Tree

Table 3. Images used for augmentation of the standard dataset

B OBSERVATIONS ON THE WORD CLOUDS

By comparing the word clouds for the 10 test datasets we can notice that the words and their frequency do not change dramatically in any of the datasets. However as the augmented images become bigger we start noticing them in the dataset: e.g in augmented test dataset 1x40 we start noticing a small number of the word "teddy", and as the number of the augmented images increases, it's frequency increases.Moreover in the sets with images of size 40x40 the words:"green" and "tree" appear.In the set 3x40, and 5x40 the word "christmas" appears in a high frequency, even though it did not appear in any of the other datasets.

Fig. 9. Word Cloud for the initial test dataset

Fig. 10. Word cloud for the augmented test dataset 1x10

Fig. 11. Word cloud for the augmented test dataset 3x10

Fig. 12. Word cloud for the augmented test dataset 5x10

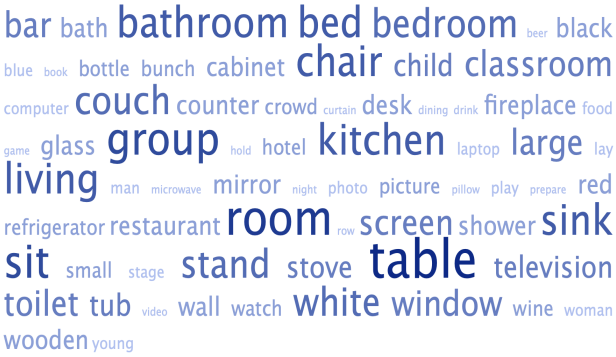

Fig. 13. Word cloud for the augmented test dataset 1x20

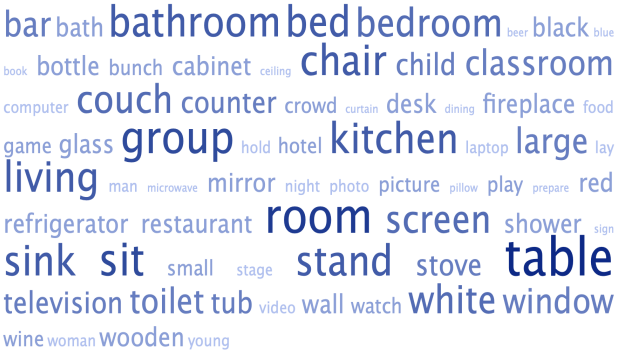

Fig. 14. Word cloud for the augmented test dataset 3x20

Fig. 15. Word cloud for the augmented test dataset 5x20

Fig. 16. Word cloud for the augmented test dataset 1x40

Fig. 17. Word cloud for the augmented test dataset 3x40

Fig. 18. Word cloud for the augmented test datasets 5x40

C OBSERVATIONS ON THE CONFUSION MATRICES

As we can notice in the confusion matrices for the baseline model with the augmented test dataset, the model is mainly misclasifying the class living room,kitchen and classroom(majority of the images representing living-room were classified as bar, as well as a big number of kitchen and classroom).We can see that the model has a predictive bias for the class bar, but only when noise is added to the dataset. We tried to mitigate this bias for the bar class, but it resulted in worse accuracy's overall.The alternative model has a bias for the kitchen class.

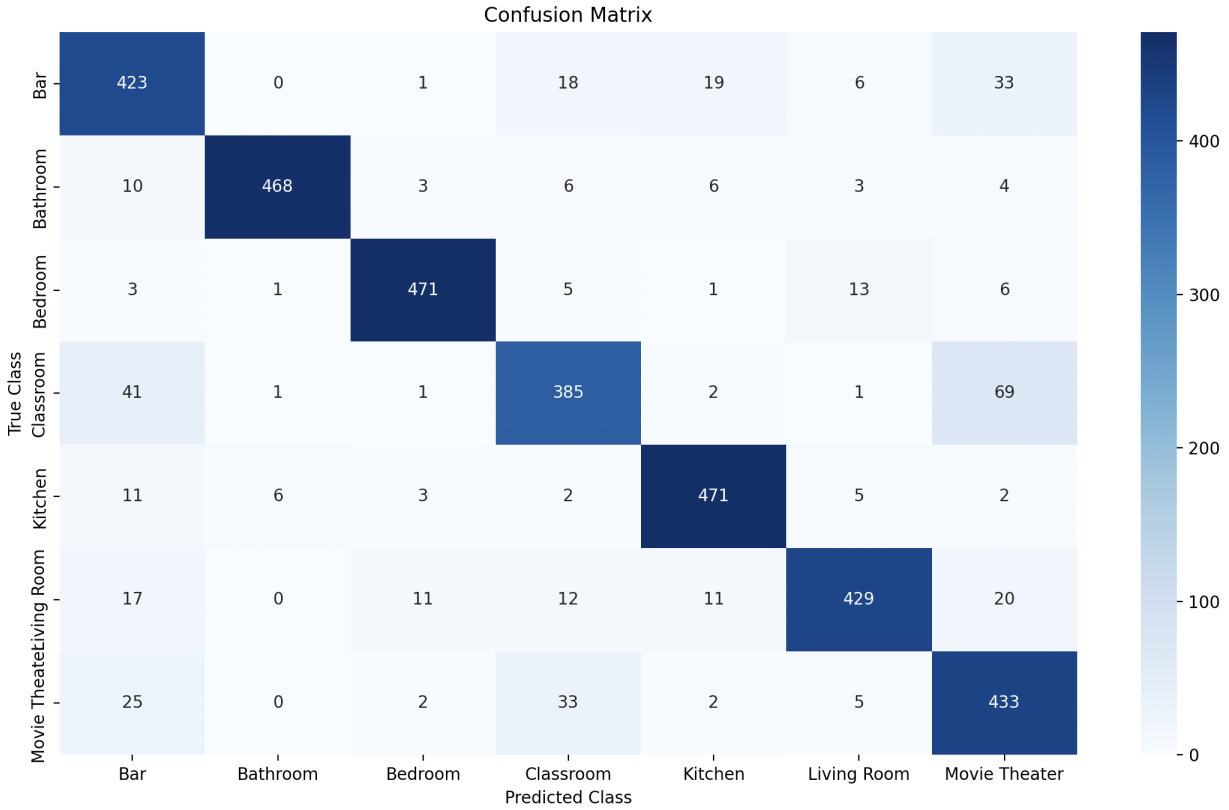

Fig. 19. Confusion matrix for the baseline model with the initial test dataset

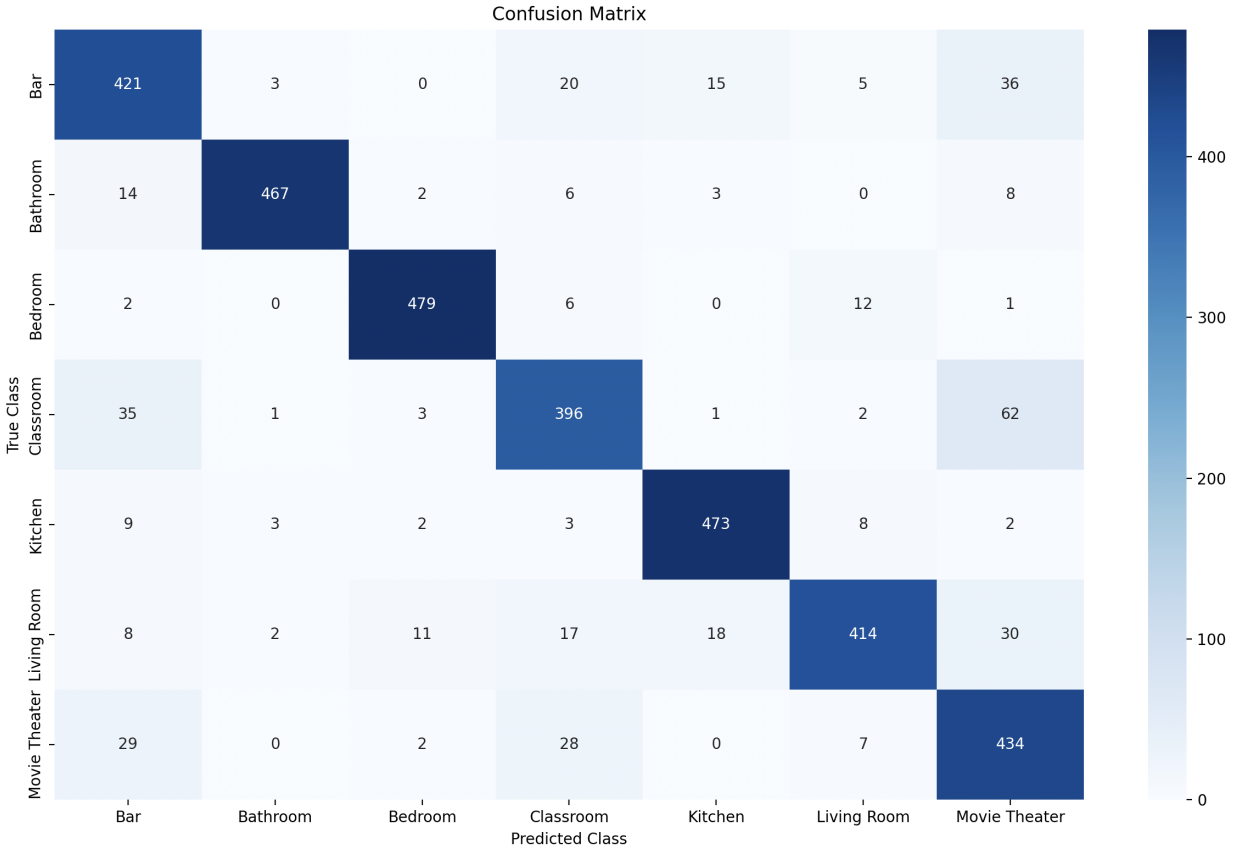

Fig. 20. Confusion matrix for the baseline model with the a repeated version of the test dataset

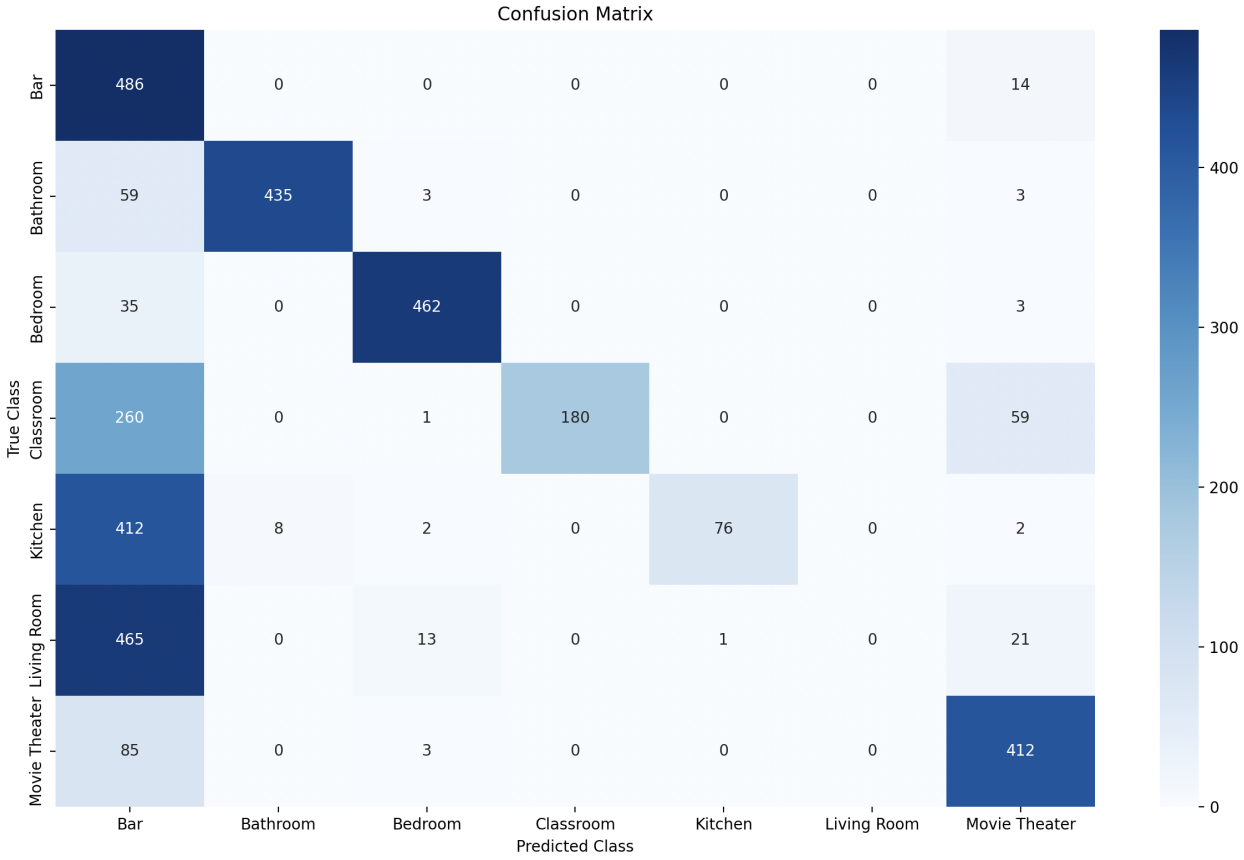

Fig. 21. Confusion matrix for the baseline model with the augmented test dataset 1x10

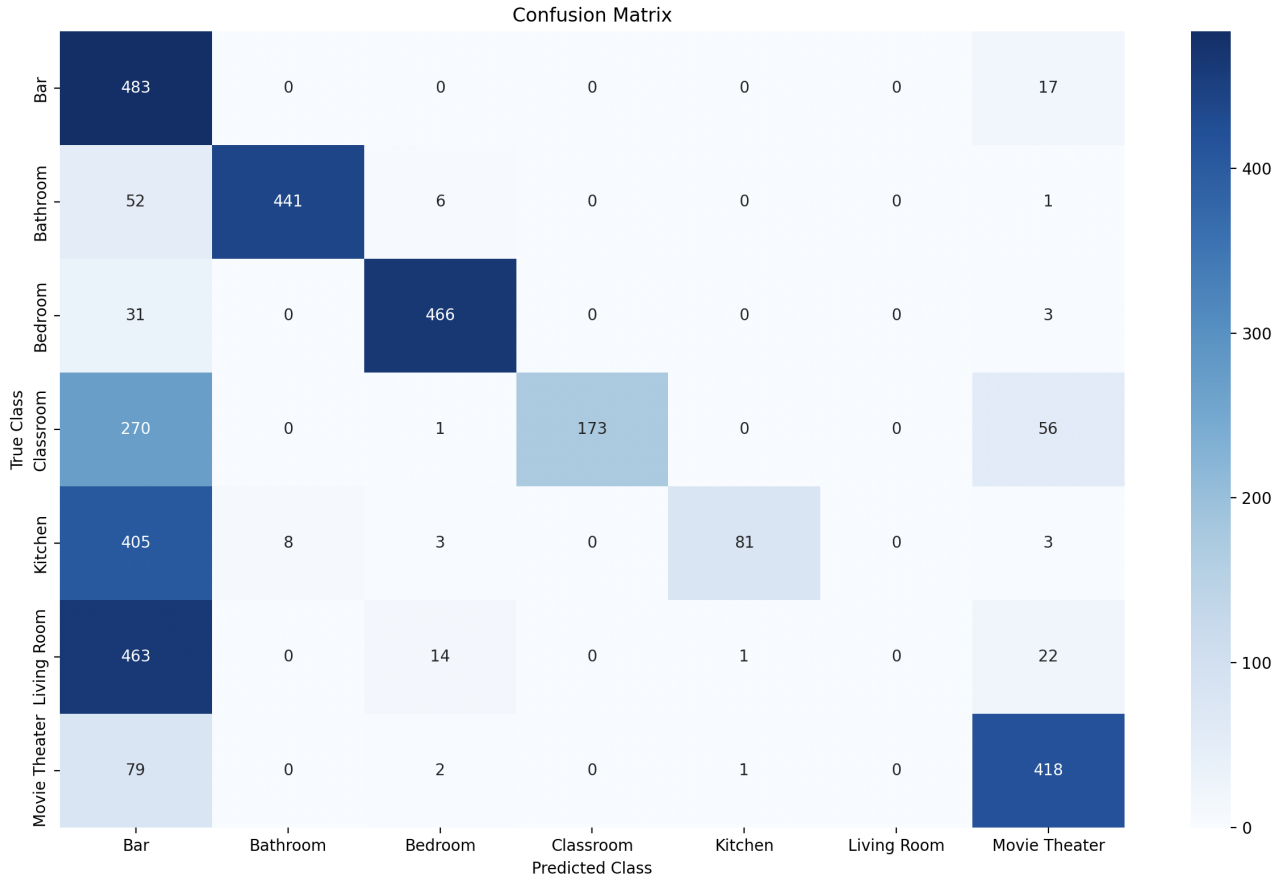

Fig. 22. Confusion matrix for the baseline model with the augmented test dataset 3x10

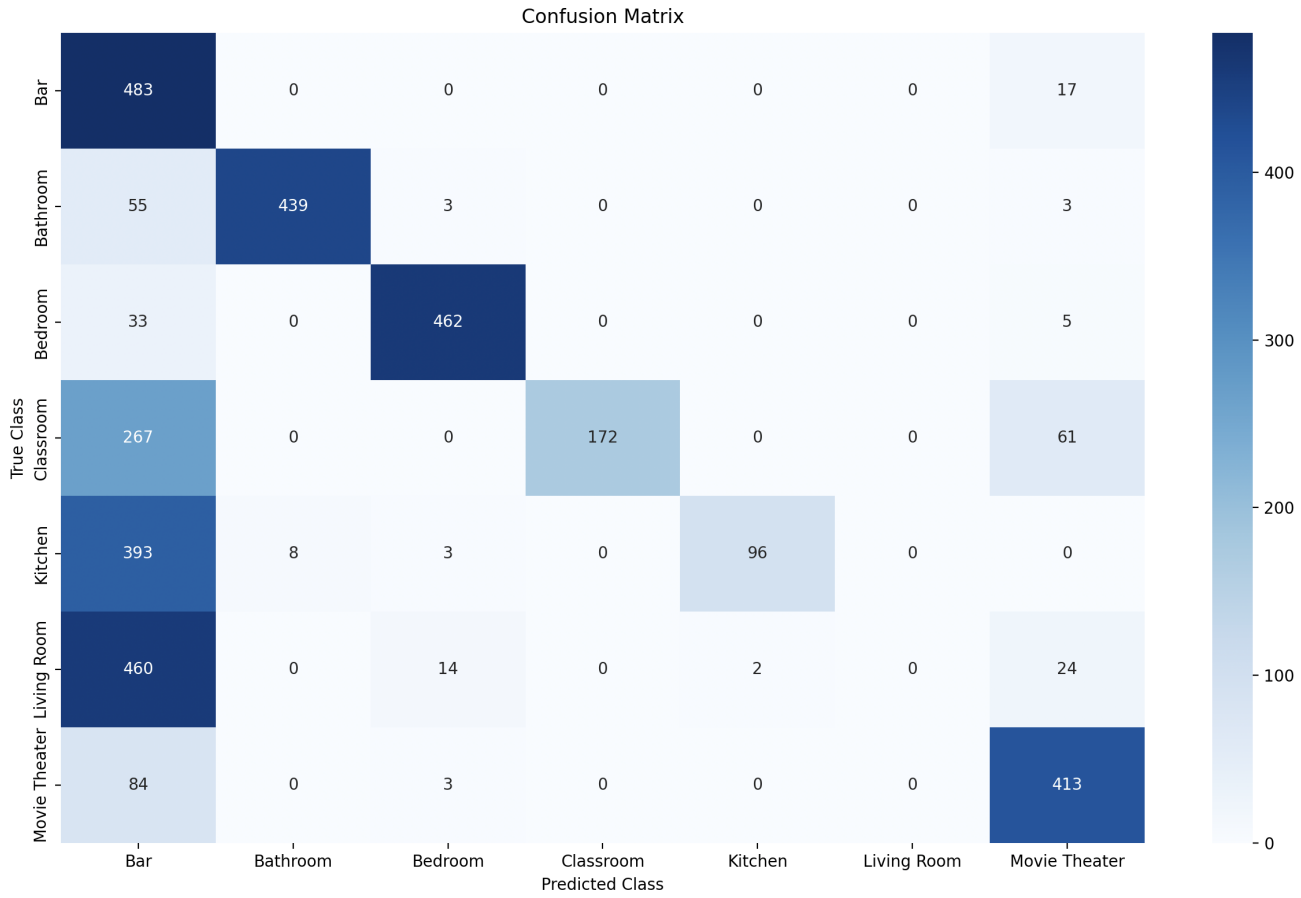

Fig. 23. Confusion matrix for the baseline model with the augmented test dataset 5x10

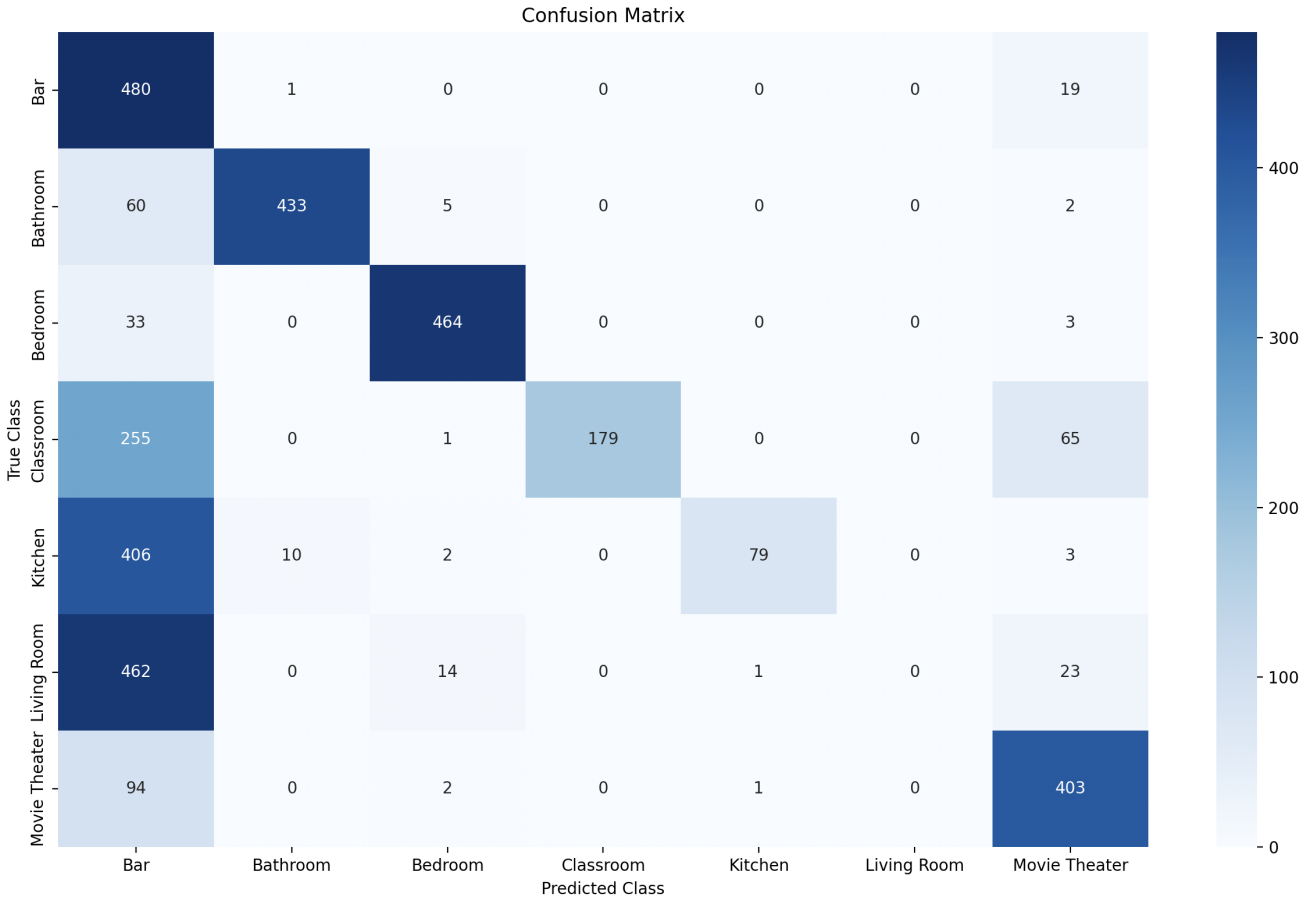

Fig. 24. Confusion matrix for the baseline model with the augmented test dataset 1x20

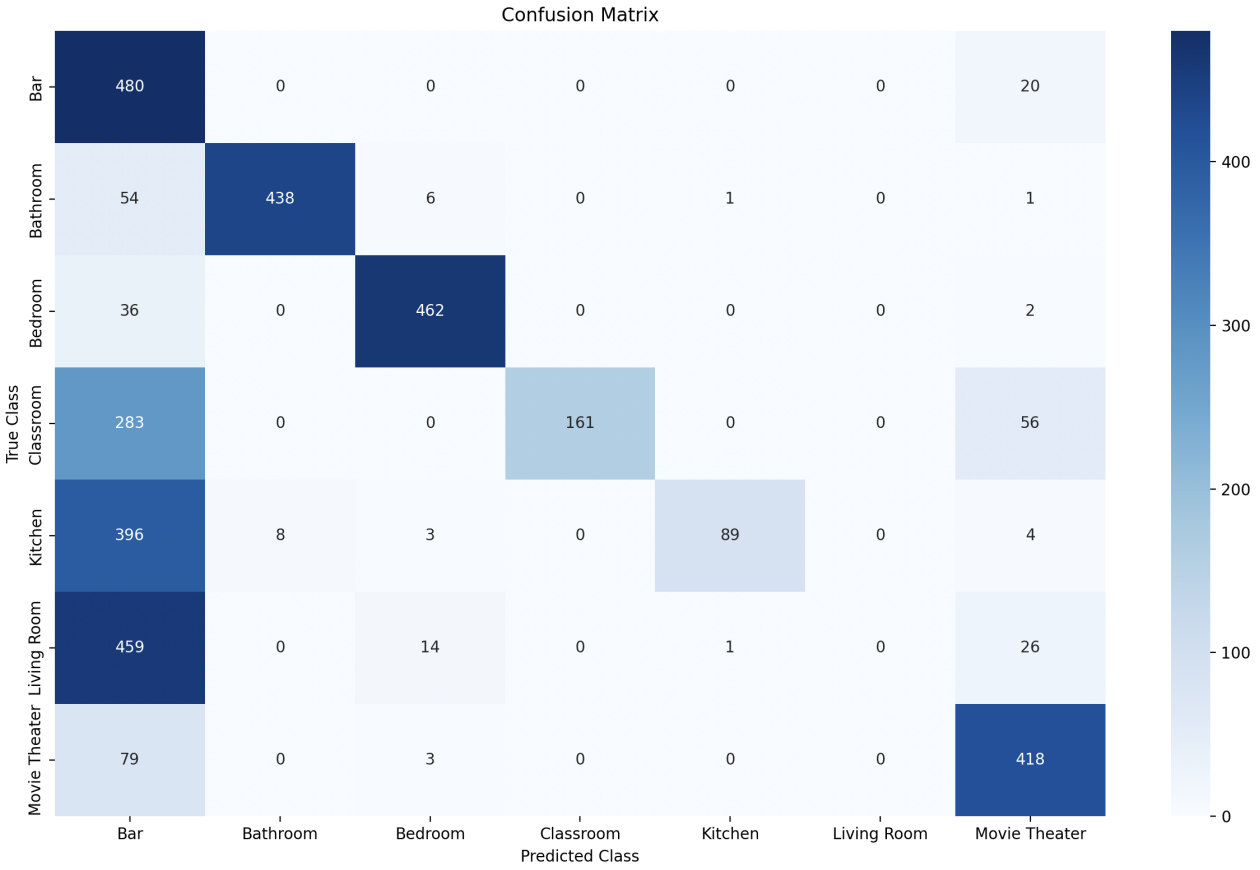

Fig. 25. Confusion matrix for the baseline model with the augmented test dataset 3x20

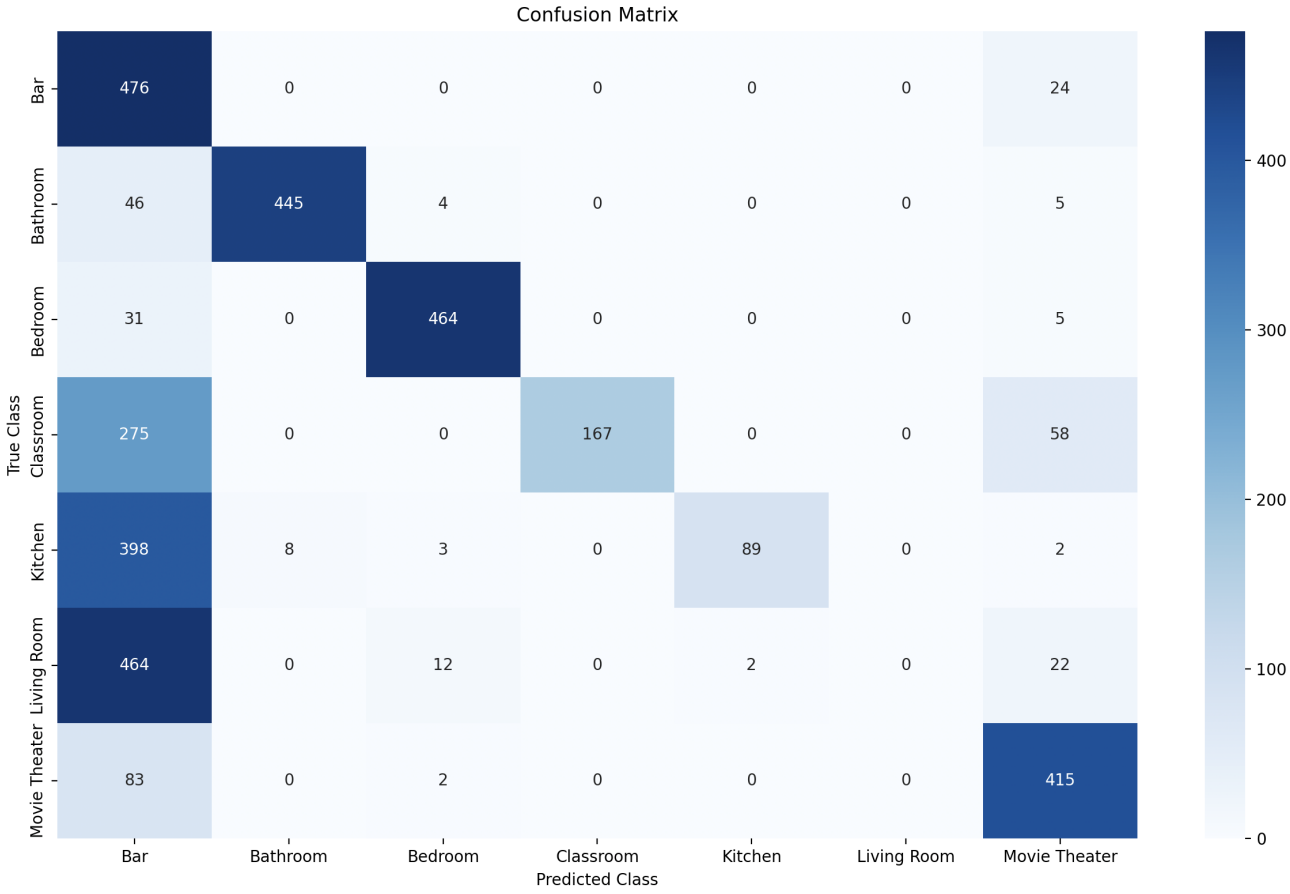

Fig. 26. Confusion matrix for the baseline model with the augmented test dataset 5x20

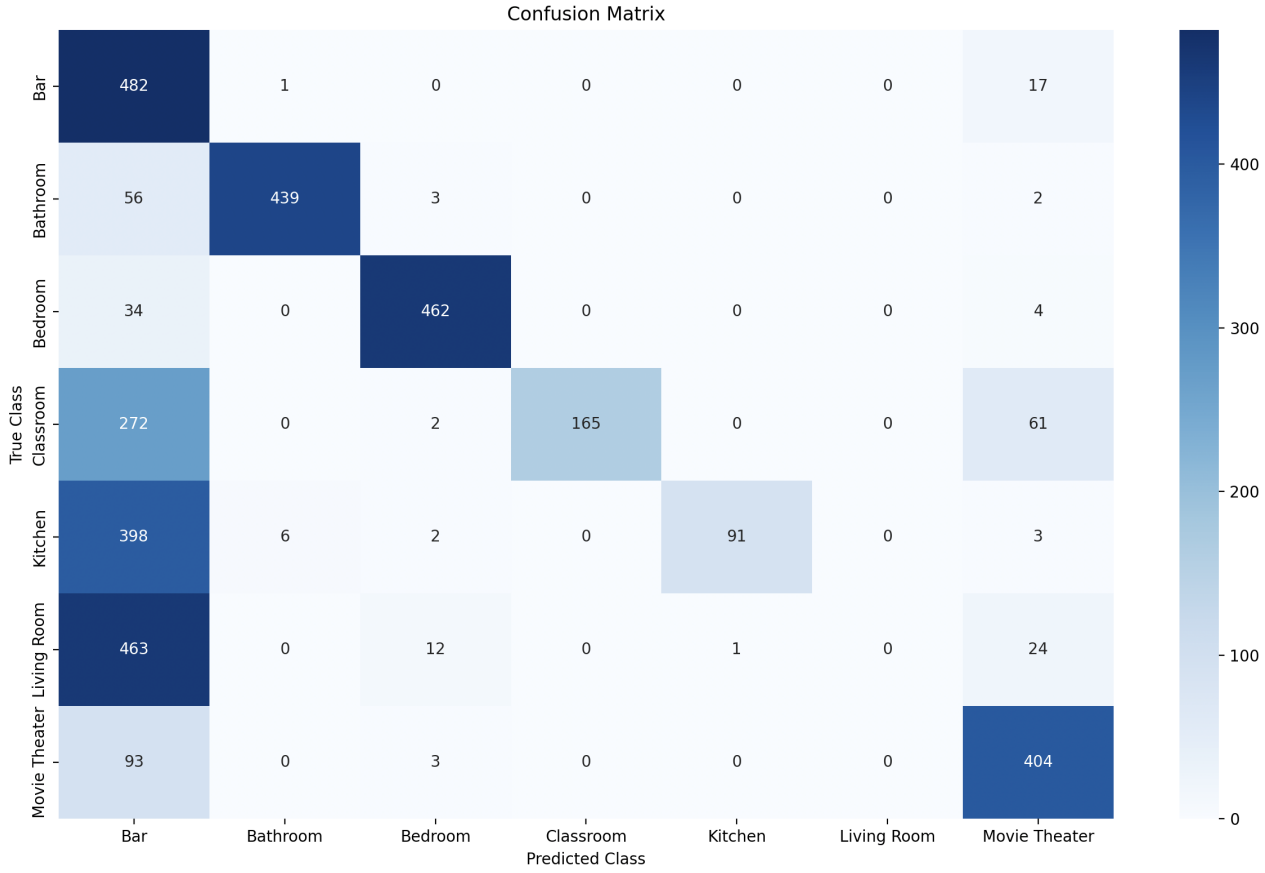

Fig. 27. Confusion matrix for the baseline model with the augmented test dataset 1x40

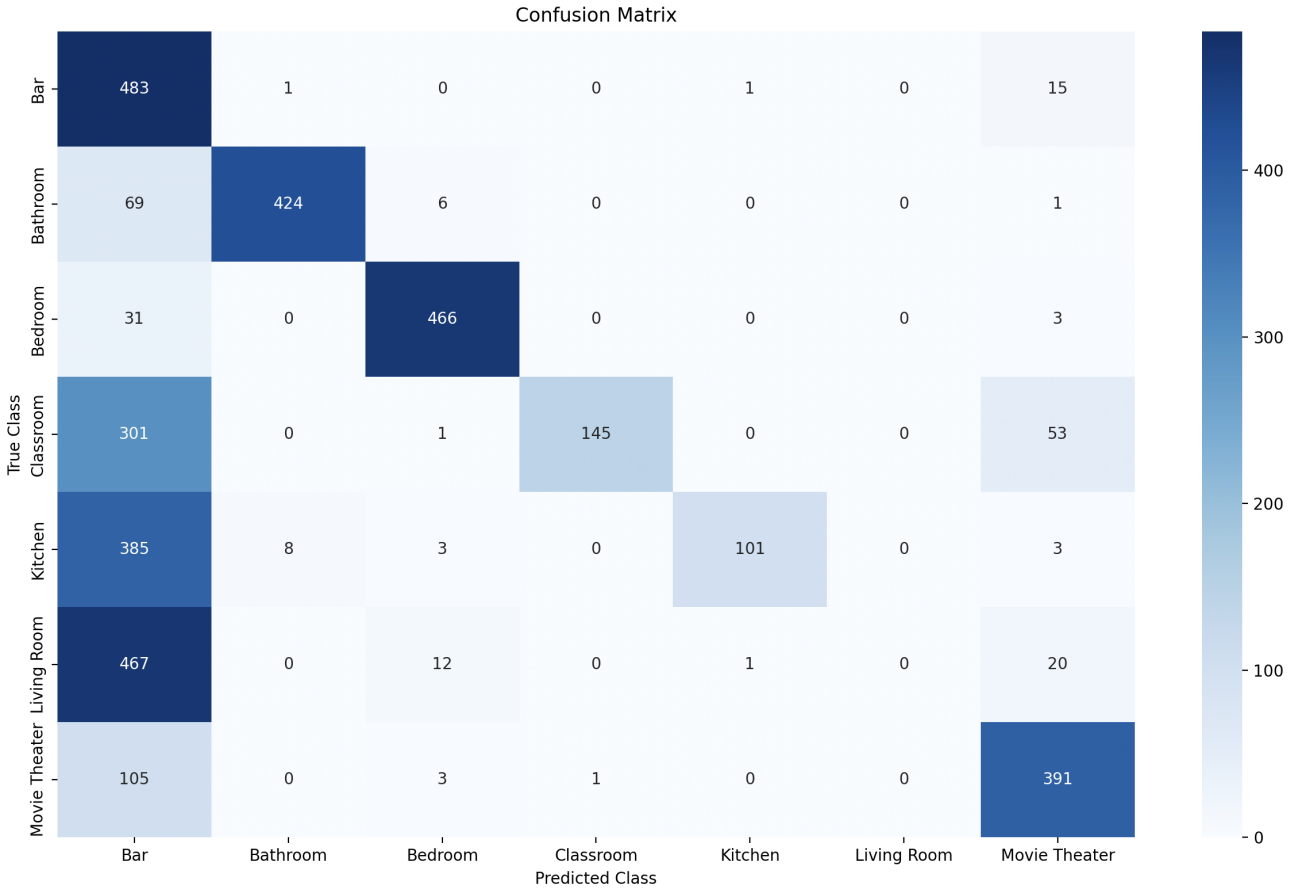

Fig. 28. Confusion matrix for the baseline model with the augmented test dataset 3x40

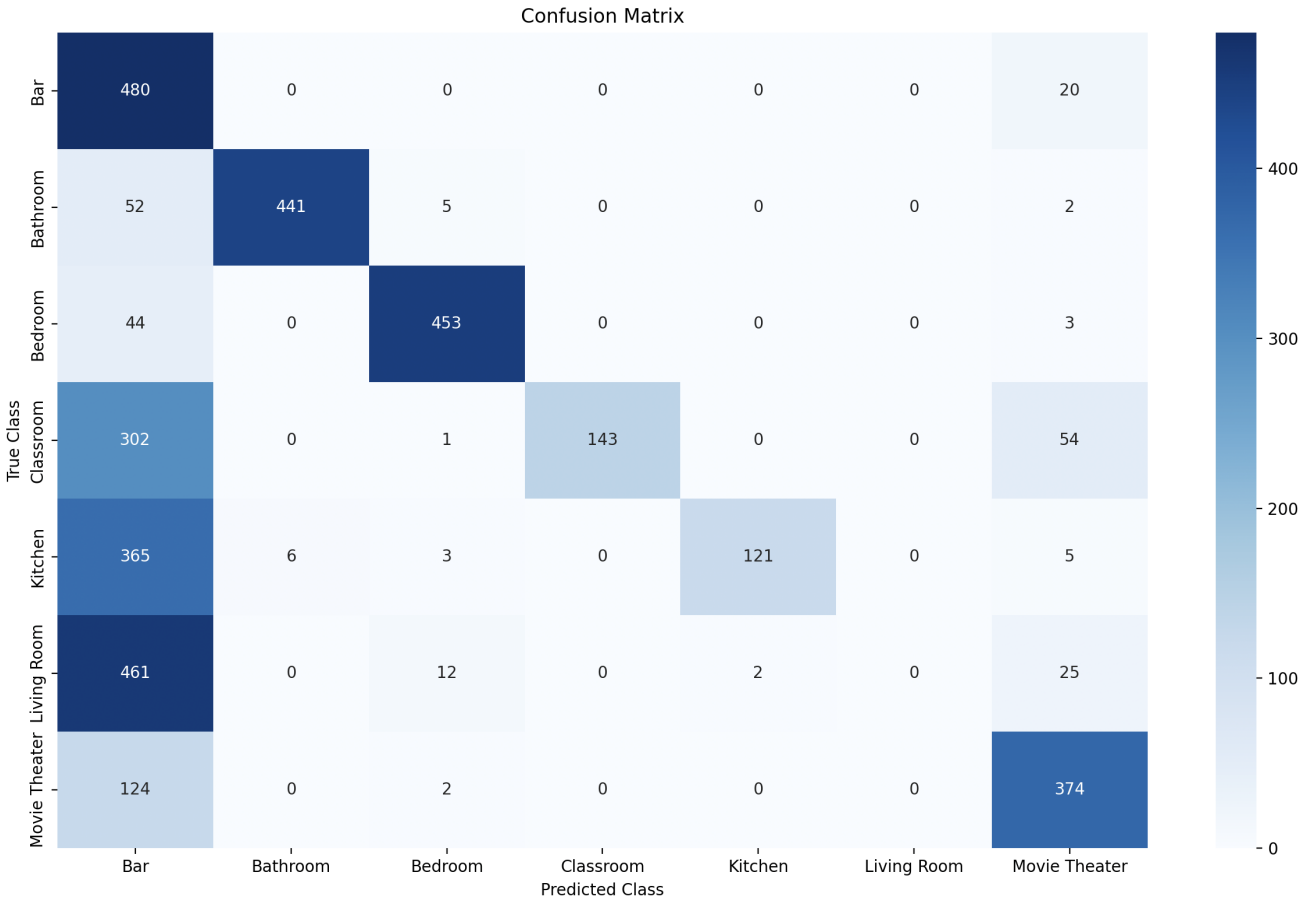

Fig. 29. Confusion matrix for the baseline model with the augmented test dataset 5x40

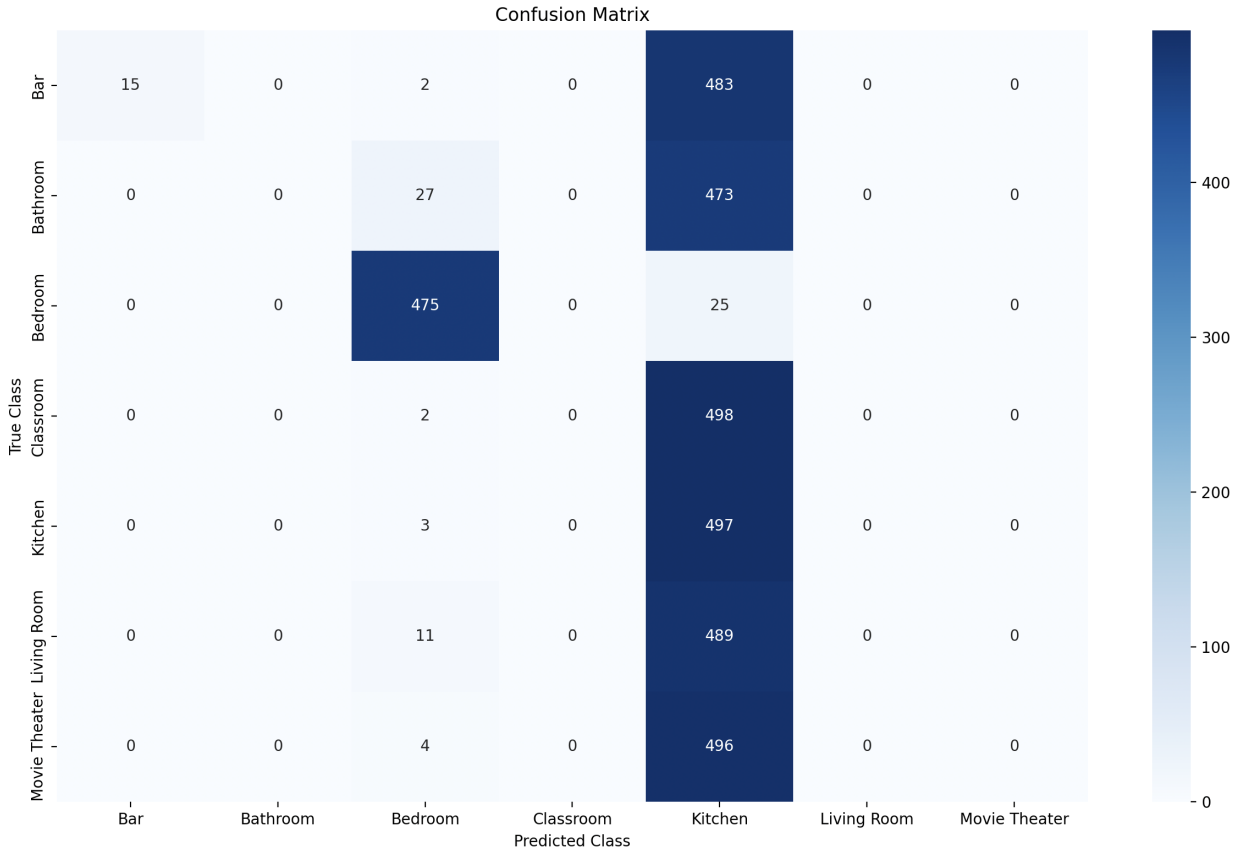

Fig. 30. Confusion matrix for the alternative model with the initial test dataset

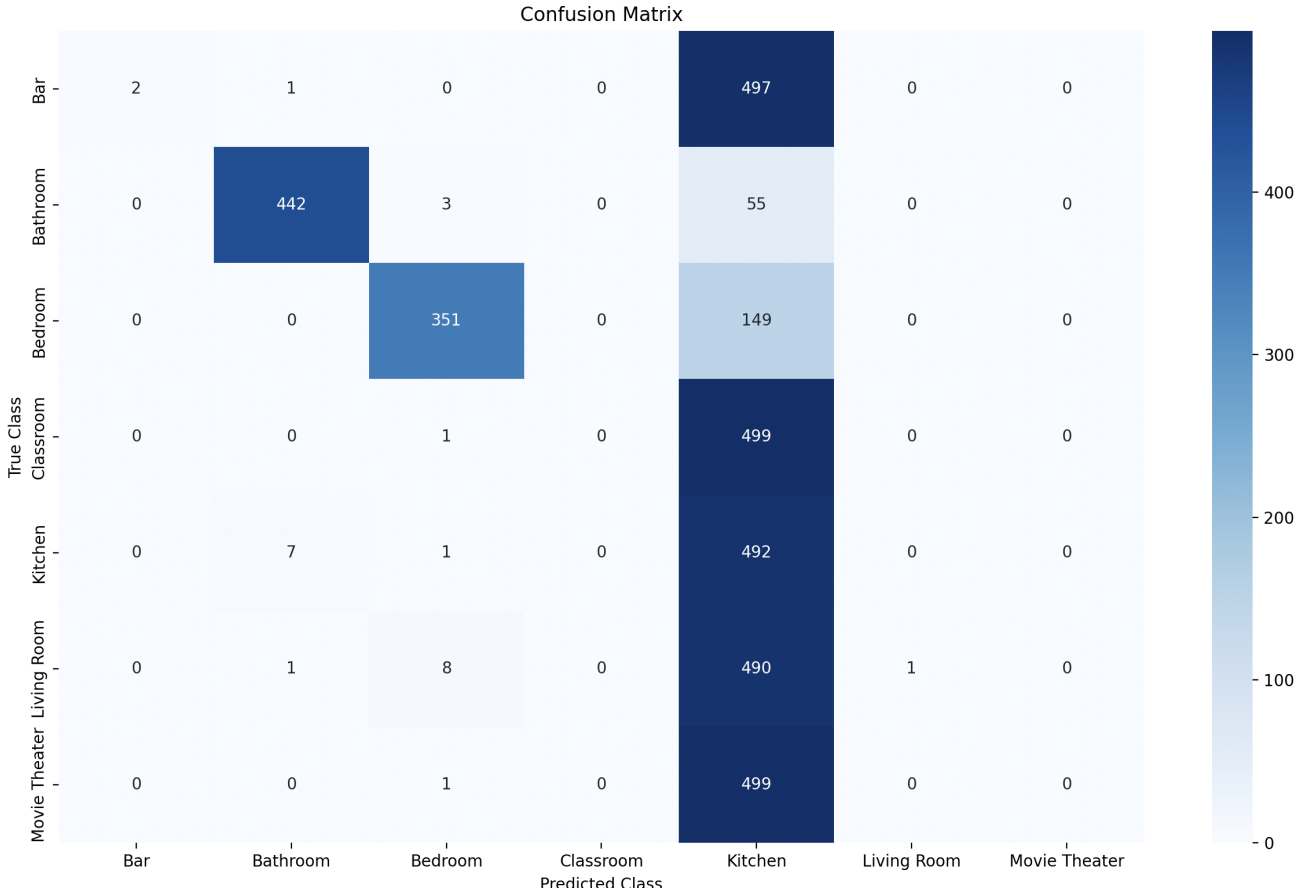

Fig. 31. Confusion matrix for the alternative model with the augmented test dataset 1x10

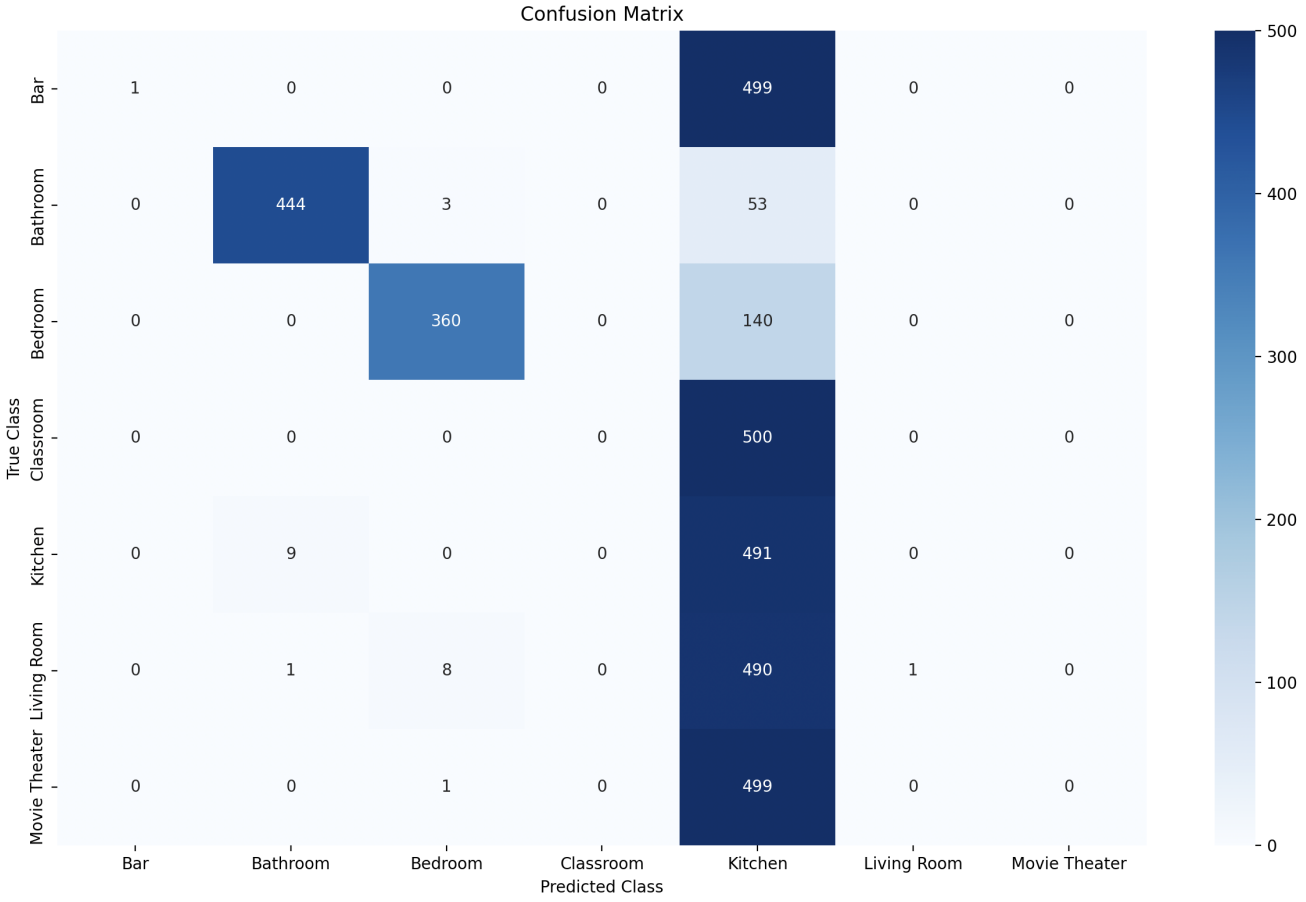

Fig. 32. Confusion matrix for the alternative model with the augmented test dataset 3x10

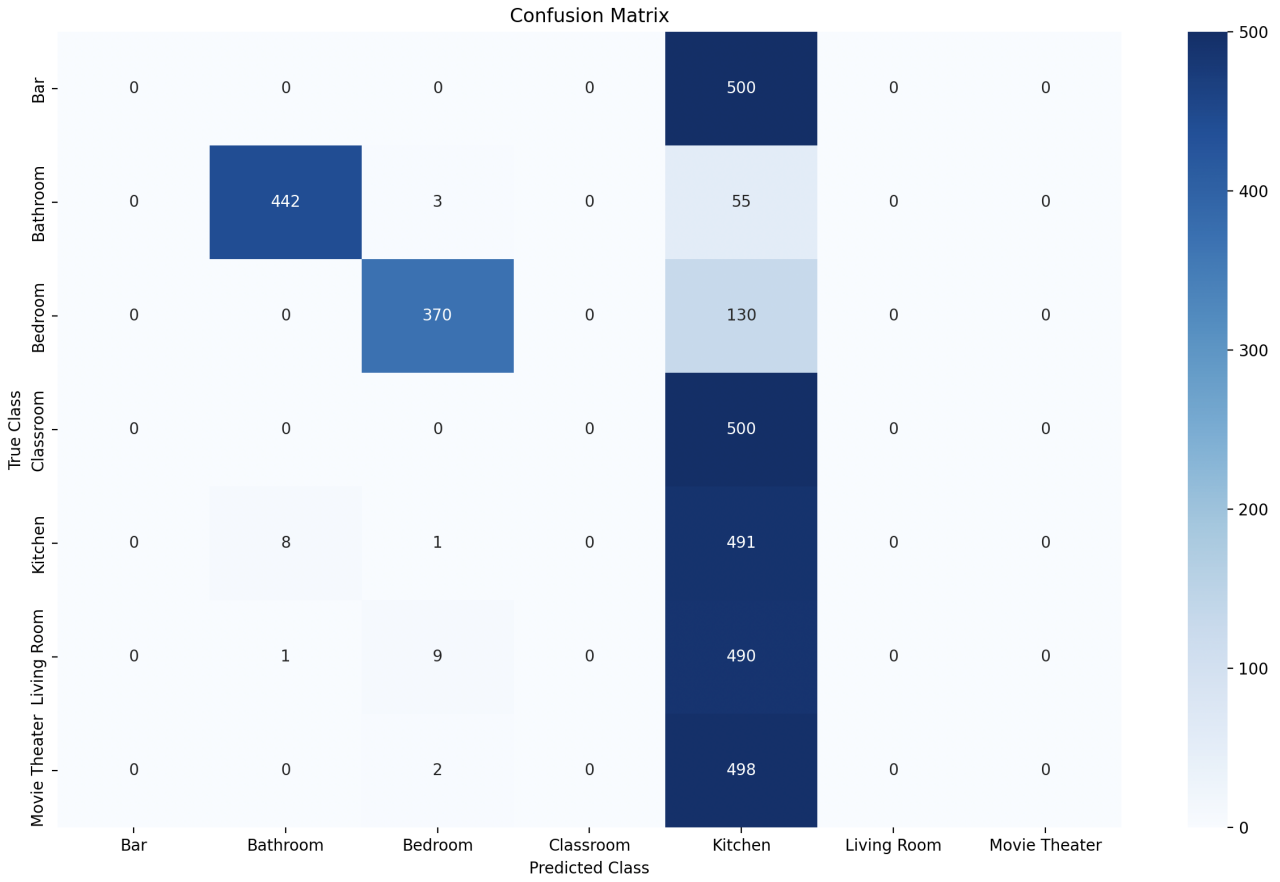

Fig. 33. Confusion matrix for the alternative model with the augmented test dataset 5x10

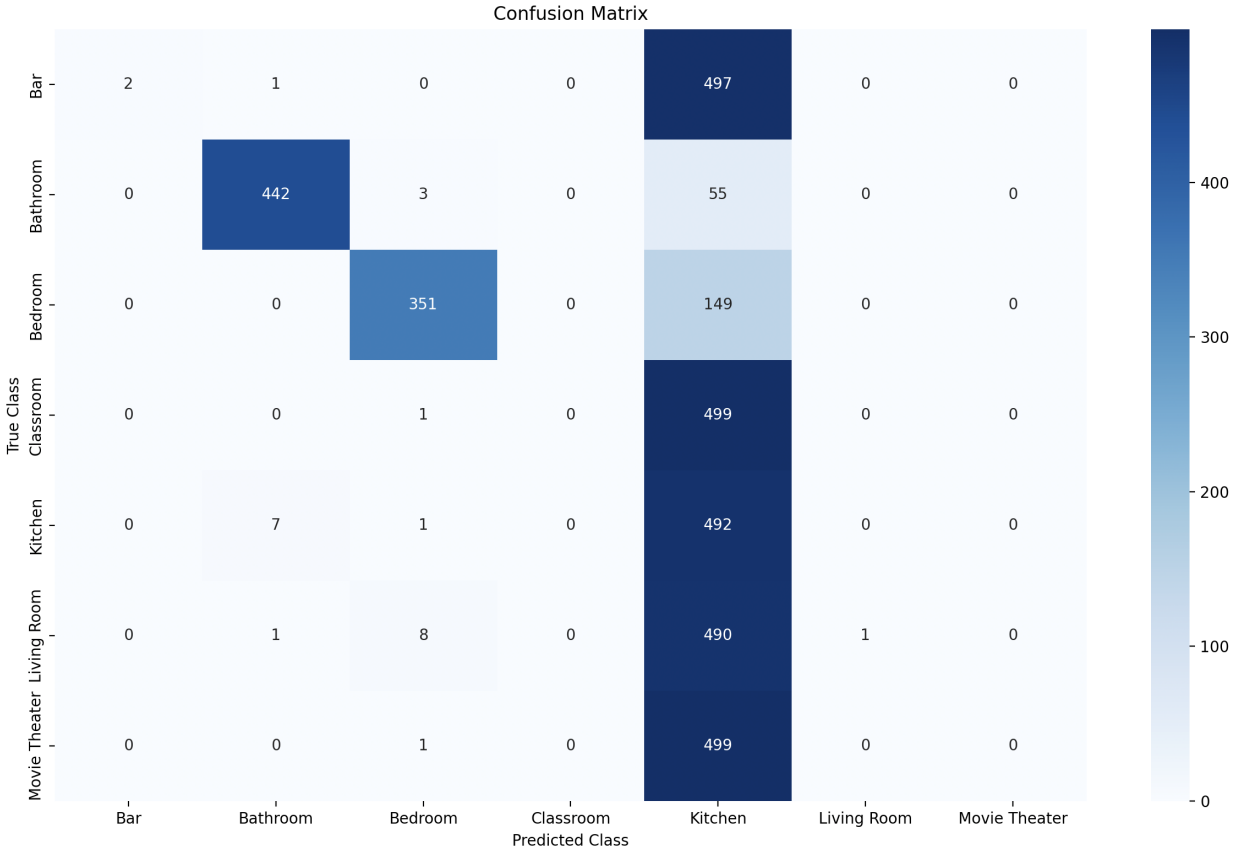

Fig. 34. Confusion matrix for the alternative model with the augmented test dataset 1x20

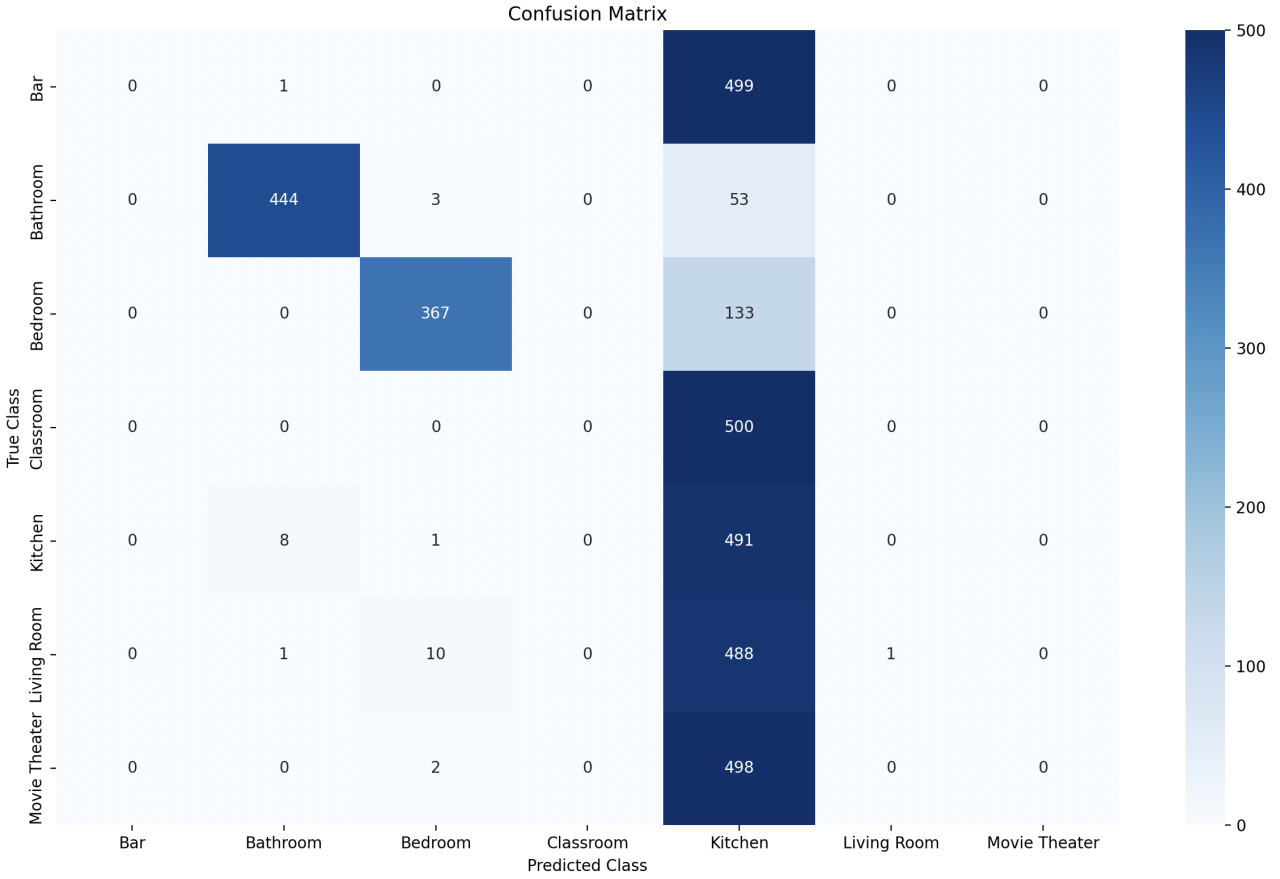

Fig. 35. Confusion matrix for the alternative model with the augmented test dataset 3x20

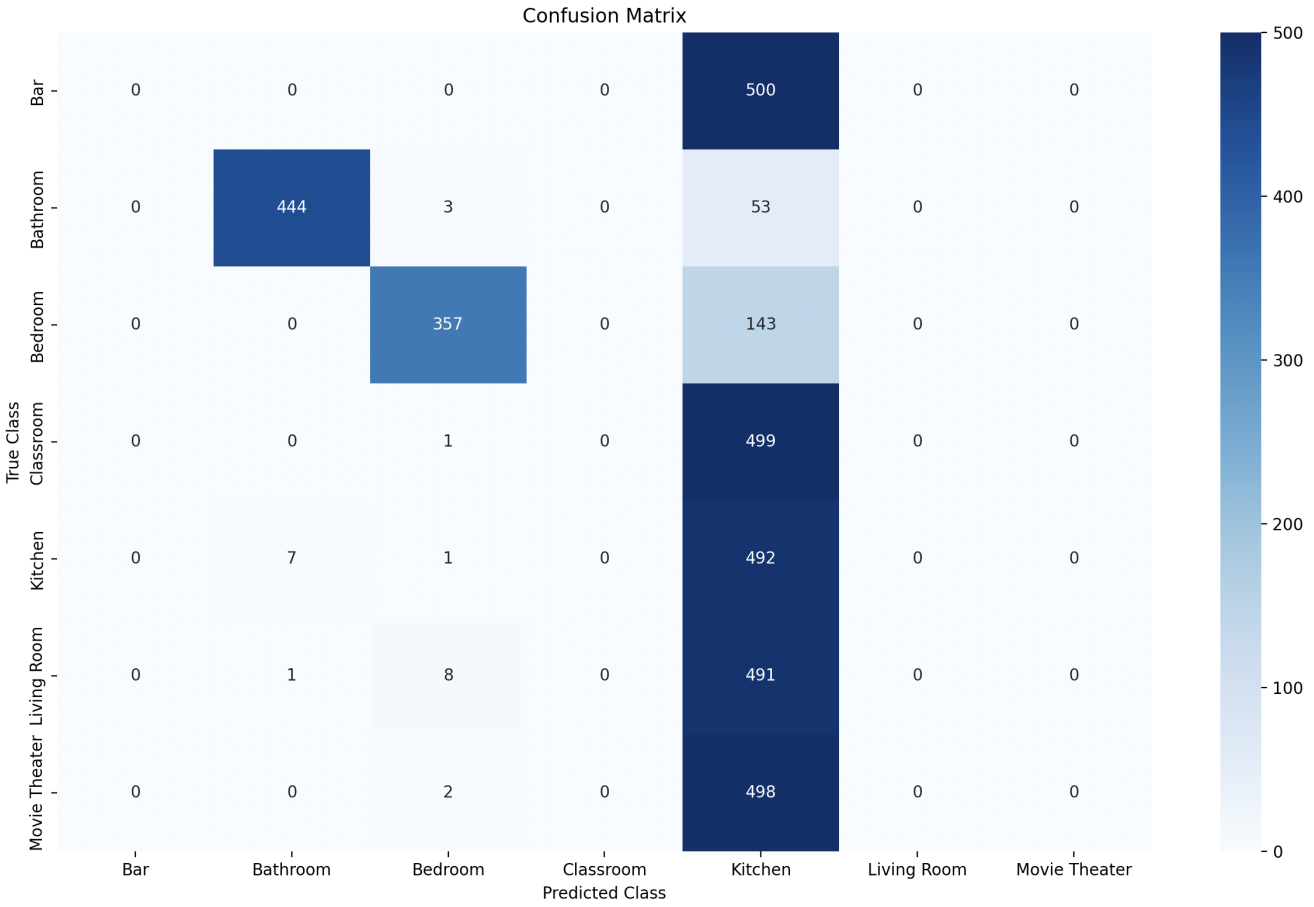

Fig. 36. Confusion matrix for the alternative model with the augmented test dataset 5x20

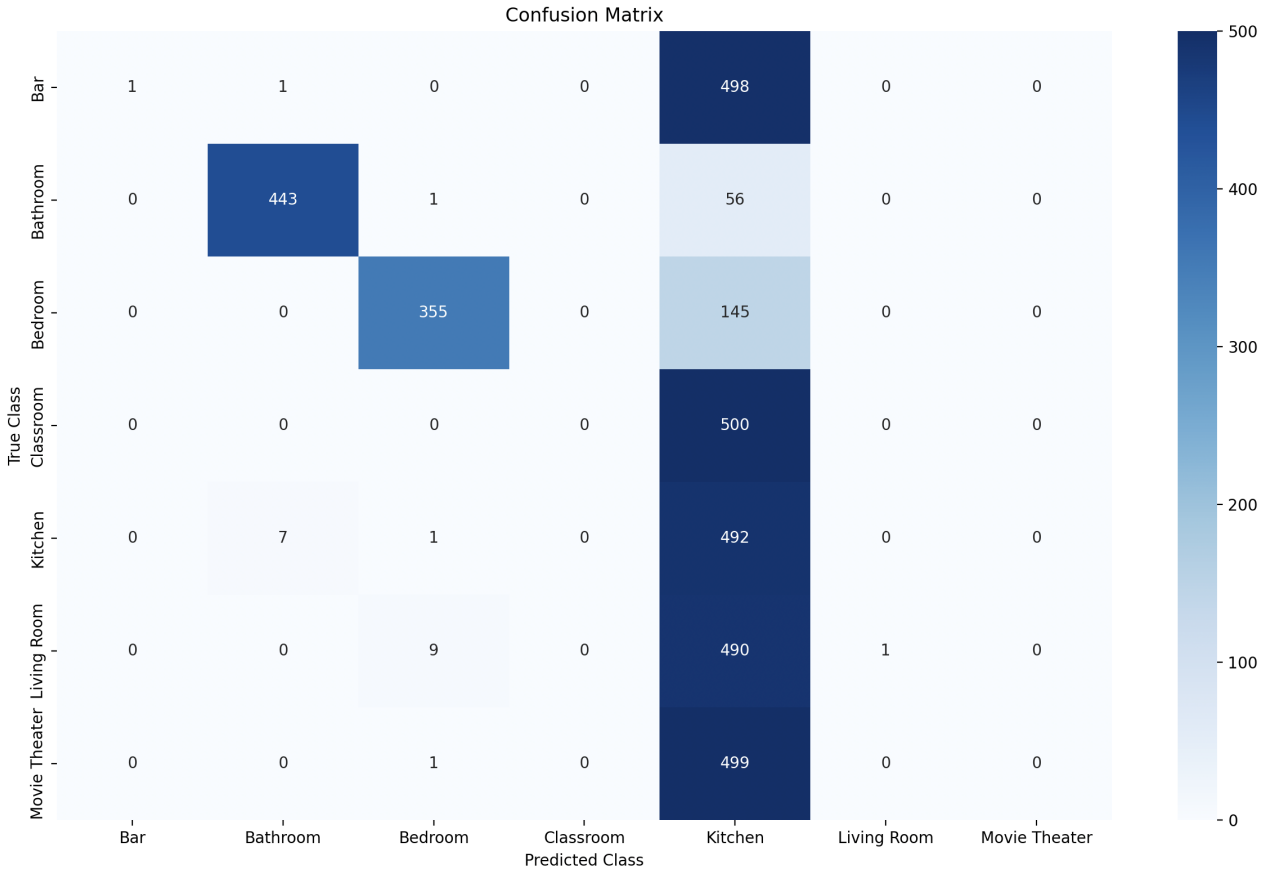

Fig. 37. Confusion matrix for the alternative model with the augmented test dataset 1x40

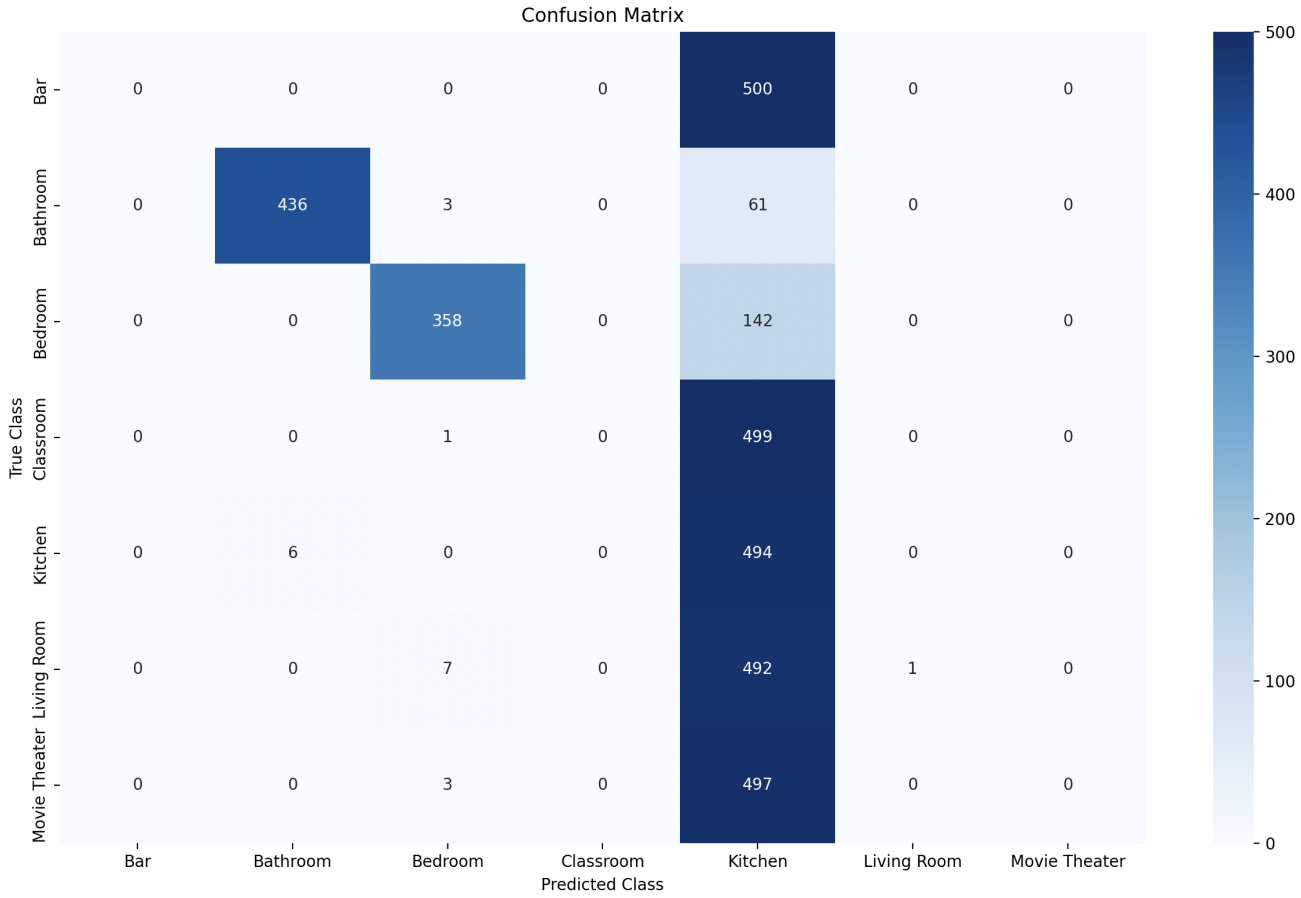

Fig. 38. Confusion matrix for the alternative model with the augmented test dataset 3x40

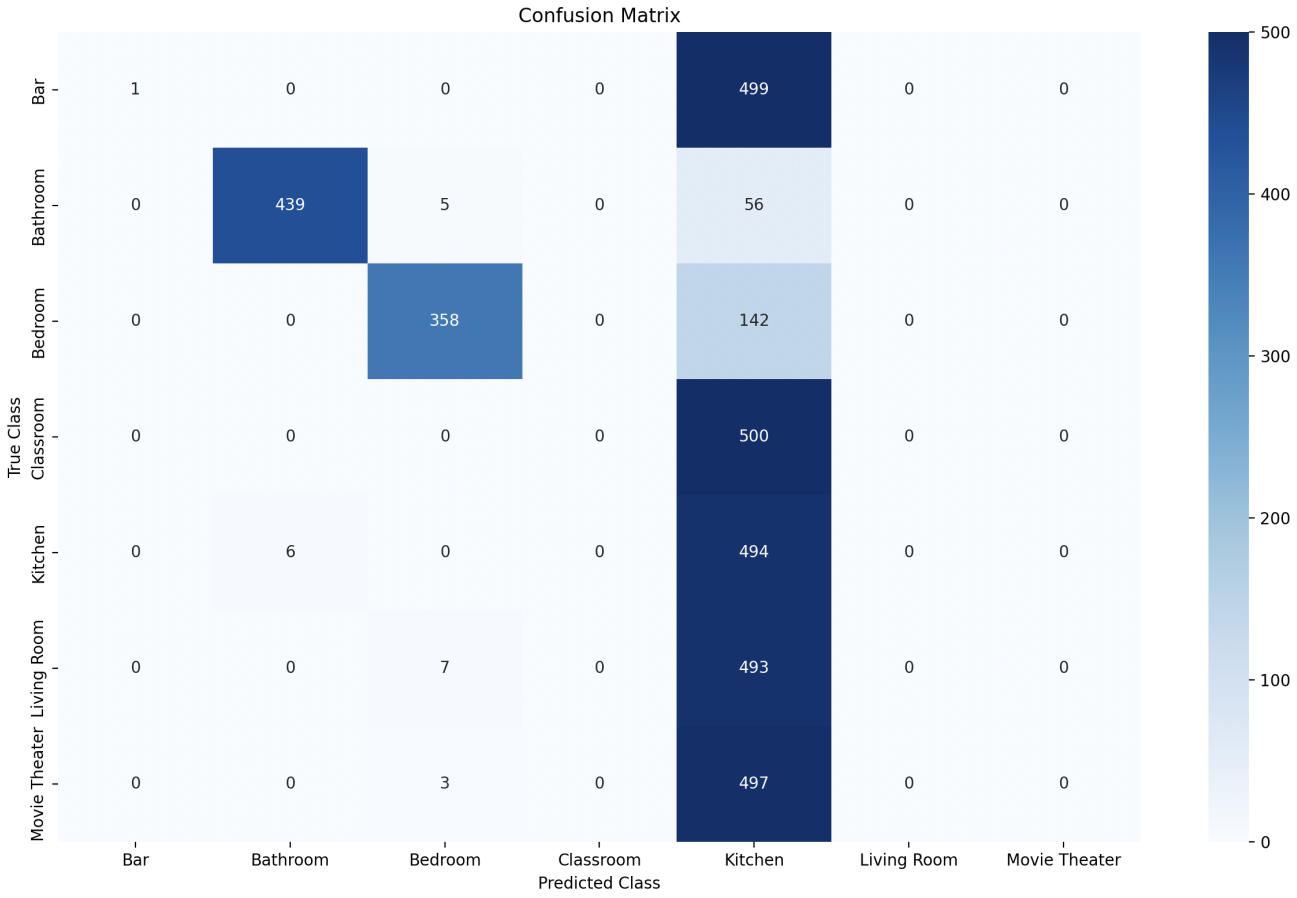

Fig. 39. Confusion matrix for the alternative model with the augmented test dataset 5x40