Transformers-Based Few-Shot Learning for Scene Classification in Child Sexual Abuse Imagery

Thamiris Coelho[*,1], Leo S. F. Ribeiro[1], João Macedo[2,3], Jefersson A. dos Santos[2,4], Sandra Avila[1]
[1]Recod.ai Lab, Instituto de Computação, Universidade Estadual de Campinas (UNICAMP), Campinas, Brazil
[2]Departamento de Ciência da Computação, Universidade Federal de Minas Gerais (UFMG), Belo Horizonte, Brazil
[3]Departamento de Polícia Federal, Belo Horizonte, Brazil
[4]School of Computer Science, University of Sheffield, Sheffield, United Kingdom

***Abstract*—Sexual abuse affects many children globally, with over 36 million reports in the past year. The vast amount of multimedia content exceeds law enforcement's analysis capacity, necessitating reliable automated classification tools. While effective, deep learning methods require extensive data and costly annotations that are restricted to law enforcement. This Master's thesis addresses these challenges using Transformer-based models for classifying indoor scenes, where such content is often found. Utilizing few-shot learning, the study reduces the need for extensive annotations, comparing classic few-shot models with Transformer-based models and exploring different methods for feature vector aggregation. The findings show that aggregating vectors using the mean is most effective, achieving 73.50 ± 0.09% accuracy with just five annotated samples per class. Evaluated with the Brazilian Federal Police, the model achieved 63.38 ± 0.09% balanced accuracy on annotated child sexual abuse indoor scenes, indicating the technique's potential to aid preliminary screening efforts.**

I. Introduction

Child sexual abuse is a crime that affects about 9–19.7% of girls and 3–7.9% of boys [1]–[3], including indecent exposure, forced sex, and sex trafficking. According to the USA's National Center for Missing & Exploited Children (NCMEC)[1], in 2023, the number of reports of suspected child sexual exploitation was more than 36 million, making it a record year.

Due to legal and ethical restrictions, access to sensitive data related to child sexual abuse is limited to the police. Consequently, most popular child sexual abuse detection tools rely on hash comparison [4], [5]. Microsoft's PhotoDNA [6] is the most well-known tool used by major companies like Meta, X (formerly Twitter), and Google. However, hash-based methods struggle with minor alterations in the visual content, such as scaling or color changes [7], making them ineffective for new content. As a result, more robust methods, such as Deep Neural Networks, have been adopted to classify Child Sexual Abuse Imagery (CSAI) [8]–[10].

To deal with the inability to access the data, some methods use related problem-solving, such as nudity detection and age estimation, to help classify CSAI [9]. Scene classification, mainly indoor scenes where most abuse occurs [11]–[13], is another promising yet underexplored approach.

Deep Learning methods are state-of-the-art to solve many problems, particularly image and video classification problems [14]–[18]. However, the best methods demand massive amounts of annotated data for good results. Even if the methods to classify such sensitive data run inside the police-restricted environments, annotating this kind of data is challenging, and being exposed to it for an extended period can compromise mental well-being [19].

In this work, we use few-shot learning (FSL) for indoor scene classification since most visual material for child sexual abuse is recorded in those environments. Although there are several datasets for scene recognition, and this kind of data is easy to annotate, as far as we know, no FSL work is proposed for this task. In this scenario, only a few samples are annotated, especially using scene labels. This way, police agents will rapidly adjust the model using only a few annotated samples, considering their data.

In most works, classifications of outdoor and indoor scenes are considered together, despite their significantly different characteristics. Indoor scenes are more complex, as the type of room is usually determined by the objects present, such as a bed and desk indicating a bedroom, while the absence of a bed might suggest an office [20]. Conversely, global information is more crucial in outdoor scenes, while local and global information matters indoors. These substantial differences can cause models with good performance on outdoor scenes to perform poorly on indoor scenes [21].

With that in mind, the models must focus on the objects in the scene to classify them correctly. Transformers-based models [22] can give attention to the most essential parts of the data to solve some tasks. Fig. 1 shows some examples of attention maps focusing on objects or people in an image.

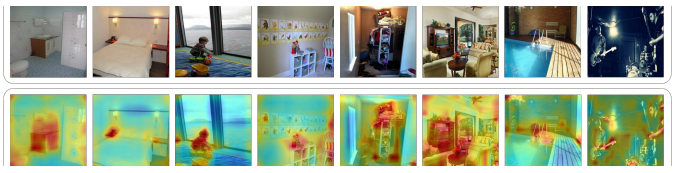


Fig. 1: Attention maps from our final model. The patches of the image with higher attention are represented in red, while the lower attention is represented in blue.

Contributions. This work is the first to use FSL methods to

[*]M.Sc. Dissertation
[1]https://www.missingkids.org/cybertiplinedata

classify indoor scenes. We compared few-shot Transformer-based and purely convolutional approaches and showed that Transformers-based approaches lead to better results for the target task. We also compared different aggregation methods to understand the best way to aggregate feature vectors of samples from the same class. The average — the most used method in FSL — is the best way to aggregate features for our target task. As far as we know, this is the first work to compare aggregation methods in FSL. Furthermore, we present the results of our final model on a CSAI dataset. For the classification of indoor scenes on the CSAI task, the model achieved a balanced accuracy of 63.38 ± 0.09% for the few-shot evaluation pipeline, using only five samples per class to classify the samples. Our results show that indoor scene features are relevant in the CSAI classification.

II. Few-Shot Learning Classification

Few-shot learning (FSL) aims to solve a task using a minimal number of labeled examples by leveraging prior knowledge. FSL typically involves three sets [23], [24]. The first set is called *base set*, a large dataset containing multiple classes to pre-train the model. The second set is the *support set* (or *novel set*), which contains samples from the FSL task; the classes in this set and the base set are disjointed. The last set is the *query set*, which contains the samples to be predicted. The support set comprises K samples per class and N different classes. Based on that, an FSL task can also be called N-way K-shot learning.

Most FSL methods follow the meta-learning paradigm, where the model learns to learn. The model gradually learns generic information from the base set during training. Then, at test time, the meta-learner for the FSL task is generalized using the support set. To learn gradually, the training process is made by episodes. Each episode is similar to the few-shot task, which contains K samples of N classes randomly chosen from the base set and the same number of the query set in the validation set [25], [26].

General machine learning methods in the testing stage classify unseen samples from the same classes seen during training based on what was learned from the training set. In contrast, in meta-testing, the objective is to classify classes never seen during meta-training. So, during meta-training, the model is trained using the base set, and then, on meta-testing, it uses the support set to classify the samples in the query set.

Besides the training approaches, FSL can be classified into three categories [27]: optimization-based, data-based, or metric-based. *Metric-based* represents data using a lower dimension (embedding), then uses simple models or distance functions to compare and classify the samples using their embedding [28]. In this work, we developed a metric-based few-shot learning method trained inductively using Transformers techniques, such as self-attention.

III. Related Work

A. Scene Classification

Scene classification, particularly for indoor environments, remains a challenge [29]. For Places365 dataset [30], InternImage [31] model reached 61.2% top-1 accuracy results (state-of-the-art). Seong et al. [32] reached 90.3% accuracy on MIT Indoor [33] and 77.3% on SUN-397 dataset, while Lopez et al. [34] achieved 74.0% on SUN-397 [35], all using supervised methods. Valois et al. [12] proposed a self-supervised approach, attaining 71.6% balanced accuracy on a derived dataset of indoor scenes from Places [30].

In this work, we evaluate our method in the same dataset proposed by Valois et al. [12]. This study is the closest we have from the perspective of the final task.

B. CSAI Classification

Vitorino et al. [8] proposed the first deep neural network, where they used a pre-trained network and performed a two-tiered transfer learning: initially for detecting pornography, then fine-tuning for CSAI detection.

Other works proposed combining adult detection with age estimation. Macedo et al. [9] combined Yahoo's open source pornography detector [36] with a network trained for age group and gender identification. Similar works also used neural networks to estimate age with adult content detection [37]–[40]. Dalins et al. [41] implemented an additional method to determine CSAI levels into ten categories, from *no sexual activity* to different *levels of child abuse*.

To bring insights on what could be done to help CSAI detection, Laranjeira et al. [13] proposed an analysis template to understand CSAI images without seeing them, using the Region-based annotated Child Pornography Dataset (RCPD) [9], highlighting the correlation between context information from objects and scenes and CSAI.

Unlike most of those works, we do not aim to classify CSAI directly in this work. We aim to help the CSAI investigation triage possible material candidates to be analyzed.

C. Embedding Learning

Embedding learning approaches aim to learn an embedding function so that the embedding for each sample is closer if the samples are similar. In Table I, we summarize the most popular embedding learning methods, highlighting with (*) the methods reproduced in this work.

All the metric-based works are relevant for indoor classification, but we select only a few to reproduce. One of our objectives is to compare purely convolutional with Transformers-based networks; for that reason, graph neural networks [43] are not the focus of this work. Finally, we reproduce ProtoNet [26], Relation Network [25], and Baseline++ [23] as CNNs-based few-shot learning. Those works are important for understanding how purely CNNs perform for indoor scene tasks. Even though optimization-based are parametric, which is not desirable for CSAI, we also reproduce them for indoor scene classification, as they are relevant to the FSL literature.

TABLE I: Embedding learning methods. Methods with (*) are the ones we could reproduce.

	Method	Dataset	Network Type	Backbone	Similarity Measure
CNNs-based	ProtoNet [26]*	Omniglot, *mini*ImageNet, CUB	CNN	Conv-4	Euclidean distance
	Relation Network [25]*	Omniglot, *mini*ImageNet, CUB, AWA	CNN	Conv-4	Learned distance
	TADAM [42]	*mini*ImageNet, FC100	Adaptative CNN	ResNet-12	Euclidean distance
	GNN [43]	Omniglot, *mini*ImageNet	CNN, GNN	Conv-4	Learned distance
	SNAIL [44]	Omniglot, *mini*ImageNet	CNN with Attention	–	Learned distance
	Baseline++ [23]*	*mini*ImageNet, CUB	CNN	ResNet18	Cosine similarity
Transformers-based	FEAT [45]*	*mini*ImageNet, *tiered*ImageNet, OfficeHome	CNN and Self-Attention	ResNet-18	Cosine similarity
	CrossTransformers [46]*	Meta-Dataset	CNN and Self-Attention	ResNet-34	Euclidean distance
	SSFormers [47]*	*mini*ImageNet, *tiered*ImageNet, CIFAR-FS, FC100	CNN and Self-Attention	ResNet-12	Custom
	P>M>F (ProtoNet) [48]*	*mini*ImageNet, CIFAR-FS, Meta-Dataset	Transformers	ViT small	Cosine similarity
	FewTURE [49]	miniImageNet, tieredImageNet, CIFAR-FS, FC100	Transformers	ViT small	Cosine similarity
	HCTransformers [50]	miniImageNet, tieredImageNet, CIFAR-FS, FC100	Transformers	ViT small	Linear classifier
	SUN [24]	miniImageNet, tieredImageNet, CIFAR-FS	Transformers	Visformer	Cosine similarity
	SP [51]	miniImageNet, tieredImageNet, CIFAR-FS, FC100	Transformers	Visformer	Cosine similarity
	SMKD [52]	miniImageNet, tieredImageNet, CIFAR-FS, FC100	Transformers	ViT small	Linear classifier

Transformers in Few-Shot Learning

Transformers [22], [53] in FSL started being used to adapt the features that were extracted using a CNN [45]–[47], [54]. More recently, Transformers started being explored as a feature extractor [24], [48]–[52], a challenge since Transformers tend to overfit when trained on a small dataset. The former approach relies on CNNs for the inductive bias and on self-attention to improve feature extraction; the latter makes use of only Transformers to extract features, and as these lack inductive bias, the training process needs to be adapted, or the pre-training stage needs to be done using a large dataset.

Except for SSFormers [47], which calculates similarity from all sparse attention patches, the other works [24], [48]–[52] generate a prototype to perform classification. That makes the model less computationally costly, allowing it to run in constrained environments like the ones available for CSAI. However, resource-intensive networks such as FewTURE [49], HCTransformers [50], SUN [25], and SMDK [52] remain impractical for such environments.

IV. Methodology

To solve indoor scene classification with FSL, we first compared existing embedding-learning few-shot methods applied to our task. Then, with the best model in hands, we thoroughly studied the method's hyperparameters and tested different backbone options aligned with our research questions and final indoor scene classification task.

A. Experimental Design

Fig. 2 illustrates the pipeline followed in our experiments. One of our objectives is to compare two FSL approaches for indoor scene classification; these approaches are purely convolutional networks and models based on Transformers. Our selection of methods to reproduce all follow the proposed pipeline with differences in specific training protocol, network backbone, and the chosen definition for vector similarity. For each method, we followed the original proposed parametrization and training protocols.

The studies reproduced also have a pre-training stage in common (referred to as *episodic meta-learning* on a base set). Most were pre-trained on *mini*ImageNet except for

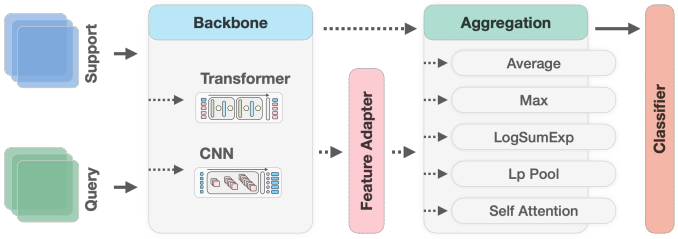


Fig. 2: Experimental pipeline followed in this work. The dotted line arrows represent the possibilities for the experiments, and the solid line arrows represent fixed steps in the experiments.

P>M>F [48], where the authors' goal was to present a simple pre-training > meta-training > fine-tuning pipeline that could reap the benefits of episodic meta-training using a large-scale dataset. Then, they used ImageNet-1K [55] for this stage. For each method, we also consider an extra meta-training stage where training is resumed with the Places600 dataset, yielding models fine-tuned for scenes.

This first experiment across the methods gave us answers for research question Q1 (**Do Transformers-based methods outperform convolutional neural network methods for indoor classification?**), comparing convolutional methods with Transformer-based ones. Then, within the Transformer-based models, we observe that the new architecture is used in two ways: as the backbone of the model to extract the features [48], or to adapt feature extracted from a purely convolutional network to a new task [45]–[47]. So, comparing those methods, we can answer the research question Q2 (**What is the more accurate way to use Transformers for indoor classification? As a feature adapter or a feature extractor?**).

With the answers to these questions and a selected model at hand, we consider one specific aspect of designing with Transformers; these models output a multi-vector representation that needs to be aggregated to be used for classification. This procedure raises the specific question Q3 (**What are the impacts of feature vector aggregators on few-shot models for indoor classification?**).

Finally, given a selected model for the indoor scene classification method and an aggregation approach attached, our next

step is to evaluate our model in a real CSAI dataset. This is done through our partnership with law-enforcement agents and yields an answer to our final question Q4 (**How to develop an indoor classification model that generalizes for the CSAI environment?**).

B. Evaluation

We defined two pipelines for evaluation. In both, we evaluate the methods through episodes composed of a support and a query set. For the *Few-Shot Evaluation*, as the name implies, we consider the pipeline used in the FSL literature. Episodes are randomly sampled from the test split of a dataset. This is adequate for comparing FSL methods as it considers the test set a collection of small tasks, each with a small (K-shots) "training set" (the support). However, it is easy to see that this approach is inadequate compared to the traditional classification literature, as using test set samples for training would be considered a leak in that setup.

To allow for such comparisons, we then consider the *General Evaluation* pipeline; each episode's support set is sampled from the validation set, which shares classes but not samples with the test set. Queries are then composed of the entire test set, making for metrics comparable to a general classification task. This pipeline remains true to the FSL protocol, as few samples are used for the support set.

V. Implementation Details

The target task is classifying indoor scene images from the classes in Places8 [12]. To cover all classes in each episode, we considered an 8-way 5-shot protocol. Nine FSL methods were compared, and the original hyperparameters were employed for each. Most studies use *mini*ImageNet for pre-training except for P>M>F [48], pre-trained on ImageNet-1K [55]. For fine-tuning, Places600 was used, a dataset sampled from Places395 that excludes classes considered in Places8, and randomly sampled 600 samples per class; the procedure for fine-tuning followed the original study protocol for 100 epochs with 2000 episodes for each epoch.

Each model was evaluated following the *Few-Shot Evaluation* pipeline; 10,000 8-way 5-shot episodes were randomly sampled for testing, with 15 queries per class per episode. The mean of top-1 accuracy is reported.

VI. Results and Discussion

We conducted a series of experiments using pre-trained backbones to extract feature vectors, and we also fine-tuned those networks on the Places600 dataset. We compared those models and chose the best one for the research testing aggregator methods. After that, we tested the model on the evaluation datasets. First, we evaluate the model on the Places8 test set. Then, to understand the model's generalization ability, we evaluate it on the OOD Scenes dataset. Lastly, through agents from the Federal Police, we evaluated our model on CSAI datasets.

A. Comparison of Few-Shot Methods

First, we apply existing models to our dataset, evaluating the pre-trained model and fine-tuning the model for indoor scenes using Places600. We classified the methods into purely convolutional and Transformers-based. Table II reports the results of FSL methods pre-trained and fine-tuned.

TABLE II: Results of few-shot methods on Places8 validation set. We report the results of the model without fine-tuning and fine-tuning. We report top-1 accuracy and a 95% confidence interval.

	Model	Pre-training Dataset	Pre-trained Accuracy (%)	Fine-tuning Accuracy (%)
Pure CNN	Baseline++ [23]	–	–	38.36 ± 0.09
	ProtoNet [26]	*mini*ImageNet	37.48 ± 0.09	43.13 ± 0.10
	RelationNet [25]	*mini*ImageNet	30.49 ± 0.09	38.69 ± 0.09
	MAML [56]	*mini*ImageNet	34.42 ± 0.09	36.42 ± 0.09
	LEO [57]	*mini*ImageNet	31.09 ± 0.09	32.66 ± 0.91
Transformers	FEAT [45]	*mini*ImageNet	45.43 ± 0.09	45.34 ± 0.10
	SSFormers [47]	*mini*ImageNet	41.20 ± 0.11	46.27 ± 0.12
	CrossTransformers [46]	*mini*ImageNet	46.67 ± 0.09	45.07 ± 0.22
	ProtoNet (P>M) [48]	ImageNet-1K	68.76 ± 0.09	**71.86 ± 0.10**
		*mini*ImageNet	45.49 ± 0.10	52.86 ± 0.09

Comparing convolutional methods with those based on Transformers, the CNNs underperformed. The best CNN model was ProtoNet, fine-tuned using ResNet as the backbone, with an accuracy of 43.13 ± 0.10%, while the worst pre-trained Transformers-based model showed an accuracy of 45.07 ± 0.22%. Comparing the best results, the CNN model performed 9 percentage points worse than the Transformer-based one. We believe this result is because of the self-attention property that gives weights to the patches that are more important to the classification. This experiment answers our first research question Q1, showing that *Transformer-based methods outperform purely convolutional ones*.

Now, comparing Transformer-based methods, we can answer our Q2 by comparing Transformers as adapters or full backbones. We consider only versions pre-trained on *mini*ImageNet for a fair comparison. We can see that P>M — full backbone — reaches an accuracy of 52.86 ± 0.09%, 6 percentage points superior to second best CrossTransformers — a feature adapter. Q2 is answered then: *Transformers are better employed as substitutes for CNNs instead of as adapters*.

One extra exciting observation can be made: our comparison between P>M pre-trained on ImageNet-1K and *mini*ImageNet corroborates the findings of Hu et al. that using large-scale training data even from a distinct domain (objects vs. scenes) is too advantageous to be ignored. We will consider this method for our following experiments.

B. Comparison of Aggregation Methods

Most FSL works that follow the ProtoNet approach — that is, generating a prototype for each class in the support set — use the average to aggregate the feature vectors of the classes' samples. To our knowledge, this design choice is yet unexplored in the literature.

To amend this gap, we investigated five aggregators commonly used in pooling operations: Average pooling, Max pooling, LogSumExp, Lp pooling, and Self-attention. Experiments are performed using a ProtoNet with ViT Small and ResNet-18 backbone, both pre-trained with ImageNet-1K. Results are in Table III.

TABLE III: Results for the aggregation methods on Places8 validation set. We report top-1 accuracy and 95% confidence.

Backbone	Aggregation Method (%)				
	Average	Max	LogSumExp	Lp Pooling	Self Attn.
ViT Small	**72.50** ± 0.09	65.80 ± 0.09	64.05 ± 0.10	55.21 ± 0.10	52.90 ± 0.13
ResNet-18	59.05 ± 0.10	52.61 ± 0.10	51.63 ± 0.10	50.50 ± 0.10	36.24 ± 0.12

Using average as the aggregator method led to our task's best result; it beat the second-best result by 6 percentage points. We observed the same behavior for both backbones; The best was average, followed by max, LogSumExp, Lp pooling, and self-attention. This answers our Q3: *using the average is the best way to aggregate feature vectors for FSL.*

With all the decisions made, we have the best model in our hands. Our final model is a P>M, pre-trained in ImageNet-1K, with average aggregation. Its final accuracy on the validation set was 72.50 ± 0.09%.

C. Final Results on Places8

For the Places8 test set, the model was tested under both protocols presented in Sec. IV-B: *Few-shot Evaluation* and *General Evaluation*. Fig. 3 shows the confusion matrix for the *Few-shot Evaluation* protocol. We observe that the confusion matrix pattern matches the results on the validation set, showing that the model is consistent in classifying indoor scenes. It is possible to see the model presents uncertainty when comparing *bedroom*, *living room* and *child's room*; there is also misclassification between *child's room*, *bedroom* and *classroom*. The average accuracy with 96% confidence was 73.50 ± 0.09%.

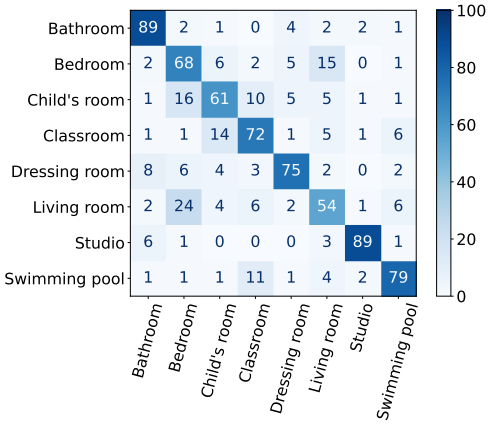


Fig. 3: Confusion matrix with the results of the model on the Places8 test set. The confusion matrix presents the accuracy (%) of the predicted classes following the few-shot evaluation protocol.

When following the *General Evaluation* protocol, the average balanced accuracy was 69.25 ± 0.05%. Because this protocol is comparable with traditional classification methods, we look at the work of Valois et al. [12] — that proposes and evaluates on Places8. Their reported result is 71.6% of balanced accuracy. This is promising for our model as we achieved our result using only 5 samples per class.

D. OOD Scenes

Due to the small dataset size, we only followed the *General Evaluation* pipeline, using the Places600 validation set to sample the support set. Our model achieved 65.42 ± 0.09% in accuracy. The work of Valois et al. reported 77.5% for this dataset, likely showing that more than 5 samples per class are necessary to achieve good out-of-distribution generalization.

Fig. 4 contains the predictions through all the episodes. It shows the number of episodes in which each image was correctly or incorrectly classified. We can observe that most episodes result in 8/10 images in the *child's room* subset being misclassified. In the two images classified correctly, one has a child in a bedroom, and the other has toys on the floor. This is well aligned with the observable distribution of images in the Places8 validation set, where images from this class often have either or both elements.

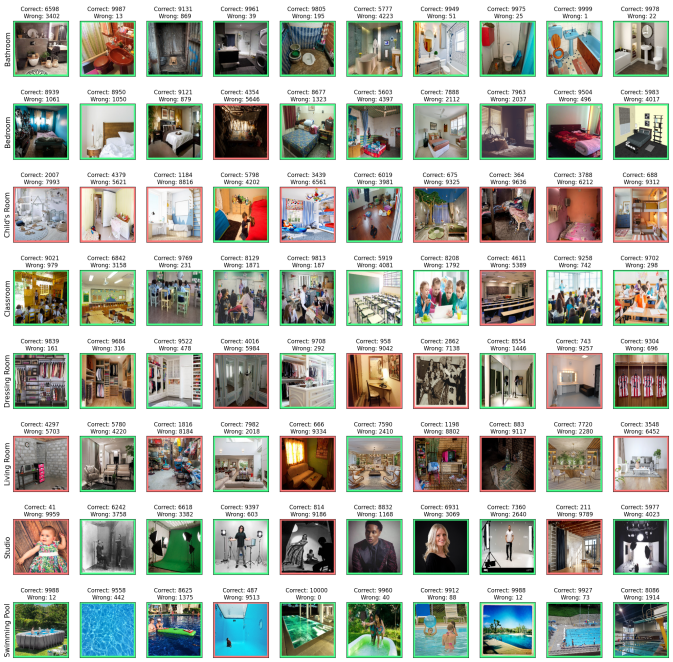


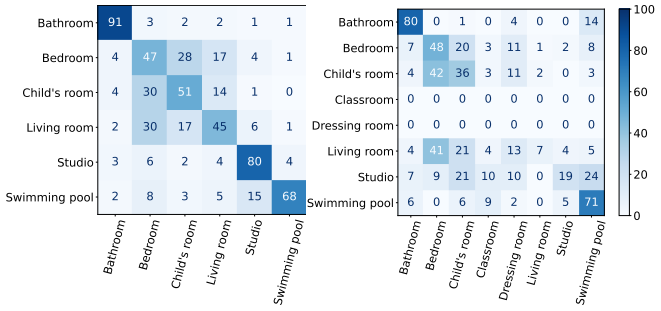
Fig. 4: Prediction on OOD Scenes [12] dataset through 10,000 episodes. The frame color emphasizes whether most predictions were correct (green) or wrong (red).

E. CSAI Final Tests

For the final tests on CSAI datasets, we collaborated with a Brazilian Federal Police agent. We only performed evaluation in this step, and the training stage was performed only with non-CSAI-related datasets. Three datasets were considered:

CSAI, RCPD [9], and CSAI indoor. The first two are labeled for CSAI classification, while CSAI indoor is a subset of the CSAI dataset and is also labeled for indoor scene recognition.

CSAI Indoor: For the CSAI Indoor dataset, both protocols presented in Sec. IV-B were considered. Fig. 5 shows the confusion matrix for both experiments.



(a) Few-shot protocol. (b) General protocol.

Fig. 5: Confusion matrix with the model's accuracy (%) on the CSAI indoor dataset.

For the *Few-shot Evaluation*, we have only six classes in the confusion matrix (Fig. 5a) because only those classes were found on the dataset. The average accuracy is 63.38 ± 0.09%. We can observe that the confused predictions align with what was observed on the public sets. While this is not a perfect result, mistakes are aligned, and one can say in answer to Q4 that *it is indeed possible to generalize this kind of classification model to CSAI with only a few samples.*

For the *General Evaluation*, we present the confusion matrix in Fig. 5b. Here, similar behavior for *bedroom* and *child's room* can be observed, but more interestingly, this model has mistaken more *living room* samples for *bedroom* and *child's room*. The study of Valois et al. reported similar misclassifications from their model, leading us to hypothesize that this is due to domain differences between CSAI and public images when regarding these classes. Our model presents better robustness to the domain shift in CSAI, achieving 43.43 ± 0.09% in balanced accuracy against their 36.7%.

CSAI: The CSAI dataset consists of six classes: *CSAI*, *Suspected CSAI*, *Porn*, *People*, *Drawing*, and *Other*; none of those classes are present in Places8. Therefore, we only performed the *Few-shot Evaluation* protocol on CSAI dataset. Fig. 6 shows the confusion matrix of the experiment.

From the confusion matrix, we can observe misclassification between *CSAI*, *Suspected CSAI*, and *Porn*. This is expected, given the visual similarities across these classes. The method similarly wrongly classified samples from *People* into these three categories; we hypothesize this is due to the *People* class containing photographs of people not completely dressed yet without sexual connotations. The average accuracy for the CSAI dataset was 50.82 ± 0.11%.

RCPD: For RCPD, we also performed only the *Few-shot Evaluation* protocol since the dataset only has a binary annotation for CSAI. For this experiment, the model achieved

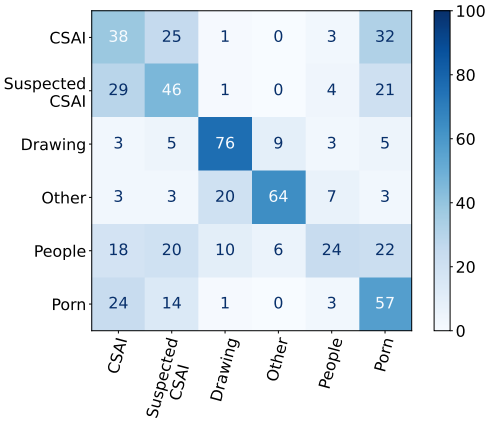


Fig. 6: Confusion matrix with the model's results on the CSAI dataset. Values reported are the accuracy (%) for prediction per class.

an average accuracy of 65.40 ± 0.16%. Individual precision for each class was 91% for CSAI and 41% for not-CSAI; We hypothesize the low rate for not-CSAI is due to the set containing porn images and images containing nudity or seminudity.

The results obtained from the experiments in CSAI and RCPD datasets showed that even with little data in the support set, the model could classify the samples quite well, given that it was trained on samples of scenes. That indicates the usefulness of indoor scene classification features for CSAI investigation. Combined with other complementary features, it could compose a robust CSAI classifier.

VII. Research Accomplishments

We summarize our main achievements as follows:

- Human Rights Academic Recognition Award: "Prêmio de Reconhecimento Acadêmico em Direitos Humanos Unicamp–Instituto Vladimir Herzog 2023".
- Dissemination of our research findings on "Jornal da Unicamp", "Algoritmo detecta cenas de abuso sexual infantil", https://tinyurl.com/2v7kmerc.

Acknowledgments

This work is partially funded by FAPESP 2023/12086-9, and the Serrapilheira Institute R-2011-37776. T. Coelho is also funded by Becas Santander and Instituto de Computação da Unicamp (Alumni Grant). L. S. F. Ribeiro is also funded by FAPESP 2022/14690-8, S. Avila is also funded by FAPESP 2020/09838-0, 2013/08293-7, H.IAAC 01245.003479/2024-10, and CNPq 316489/2023-9.

References

[1] L. Leopold and H. Engelhartdt, "Education and physical health trajectories in old age. evidence from the survey of health, ageing and retirement in europe (share)," *International Journal of Public Health*, 2013.

[2] N. Pereda, G. Guilera, M. Forns, and J. Gómez-Benito, "The prevalence of child sexual abuse in community and student samples: A meta-analysis," *Clinical psychology review*, vol. 29, no. 4, pp. 328–338, 2009.

[3] M. Stoltenborgh, M. H. Van Ijzendoorn, E. M. Euser, and M. J. Bakermans-Kranenburg, "A global perspective on child sexual abuse: Meta-analysis of prevalence around the world," *Child maltreatment*, vol. 16, no. 2, pp. 79–101, 2011.

[4] M. de Castro Polastro and P. M. da Silva Eleuterio, "Nudetective: A forensic tool to help combat child pornography through automatic nudity detection," in *Workshops on Database and Expert Systems Applications*, 2010, pp. 349–353.

[5] C. Peersman, C. Schulze, A. Rashid, M. Brennan, and C. Fischer, "icop: Live forensics to reveal previously unknown criminal media on p2p networks," *Digital Investigation*, vol. 18, pp. 50–64, 2016.

[6] M. Inc., "Photodna cloud services," https://www.microsoft.com/en-us/PhotoDNA, 2020.

[7] C. Schulze, D. Henter, D. Borth, and A. Dengel, "Automatic detection of csa media by multi-modal feature fusion for law enforcement support," in *International conference on multimedia retrieval*, 2014, pp. 353–360.

[8] P. Vitorino, S. Avila, M. Perez, and A. Rocha, "Leveraging deep neural networks to fight child pornography in the age of social media," *Journal of Visual Communication and Image Representation*, 2018.

[9] J. Macedo, F. Costa, and J. A. dos Santos, "A benchmark methodology for child pornography detection," in *Conference on Graphics, Patterns and Images (SIBGRAPI)*, 2018.

[10] R. Brewer, B. Westlake, T. Swearingen, S. Patterson, D. Bright, A. Ross, K. Logos, and D. Michalski, "Advancing child sexual abuse investigations using biometrics and social network analysis," *Trends and Issues in Crime and Criminal Justice*, no. 668, pp. 1–16, 2023.

[11] E. Bursztein, E. Clarke, M. DeLaune, D. M. Elifff, N. Hsu, L. Olson, J. Shehan, M. Thakur, K. Thomas, and T. Bright, "Rethinking the detection of child sexual abuse imagery on the internet," in *The World Wide Web Conference*, 2019, pp. 2601–2607.

[12] P. H. V. Valois, J. Macedo, L. S. F. Ribeiro, J. A. dos Santos, and S. Avila, "Leveraging self-supervised learning for scene recognition in child sexual abuse imagery," *arXiv preprint arXiv:2403.01183*, 2024.

[13] C. Laranjeira da Silva, J. Macedo, S. Avila, and J. dos Santos, "Seeing without looking: Analysis pipeline for child sexual abuse datasets," in *ACM Conference on Fairness, Accountability, and Transparency*, 2022.

[14] M. Perez, S. Avila, D. Moreira, D. Moraes, V. Testoni, E. Valle *et al.*, "Video pornography detection through deep learning techniques and motion information," *Neurocomputing*, 2017.

[15] M. V. Adão Teixeira and S. Avila, "What should we pay attention to when classifying violent videos?" in *ARES*, 2021.

[16] D. Moreira, S. Avila, M. Perez, D. Moraes, V. Testoni, E. Valle, S. Goldenstein, and A. Rocha, "Multimodal data fusion for sensitive scene localization," *Information Fusion*, 2019.

[17] ——, "Pornography classification: The hidden clues in video space-time," *Forensic Science International*, 2016.

[18] A. Ishikawa, E. Bollis, and S. Avila, "Combating the elsagate phenomenon: Deep learning architectures for disturbing cartoons," in *IEEE International Workshop on Biometrics and Forensics*, 2019, pp. 1–6.

[19] J. A. Kloess, J. Woodhams, H. Whittle, T. Grant, and C. E. Hamilton-Giachritsis, "The challenges of identifying and classifying child sexual abuse material," *Sexual Abuse*, vol. 31, no. 2, pp. 173–196, 2019.

[20] J. Qiu, Y. Yang, X. Wang, and D. Tao, "Scene essence," in *CVPR*, 2021.

[21] Z. Yu, L. Jin, and S. Gao, "P^2Net: Patch-match and plane-regularization for unsupervised indoor depth estimation," in *ECCV*, 2020, pp. 206–222.

[22] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *NeurIPS*, 2017, pp. 5998–6008.

[23] W.-Y. Chen, Y.-C. Liu, Z. Kira, Y.-C. Wang, and J.-B. Huang, "A closer look at few-shot classification," in *ICLR*, 2019.

[24] B. Dong, P. Zhou, S. Yan, and W. Zuo, "Self-promoted supervision for few-shot transformer," in *ECCV*, 2022.

[25] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. Torr, and T. M. Hospedales, "Learning to compare: Relation network for few-shot learning," in *CVPR*, 2018.

[26] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," in *NeurIPS*, 2017.

[27] N. Bendre, H. T. Marín, and P. Najafirad, "Learning from few samples: A survey," *arXiv preprint arXiv:2007.15484*, 2020.

[28] Y. Wang, Q. Yao, J. T. Kwok, and L. M. Ni, "Generalizing from a few examples: A survey on few-shot learning," *ACM Computing Surveys*, vol. 53, no. 3, pp. 1–34, 2020.

[29] T. A. Patel, V. K. Dabhi, and H. B. Prajapati, "Survey on scene classification techniques," in *ICACCS*, 2020, pp. 452–458.

[30] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 6, 2017.

[31] W. Wang, J. Dai, Z. Chen, Z. Huang, Z. Li, X. Zhu, X. Hu, T. Lu, L. Lu, H. Li, X. Wang, and Y. Qiao, "Internimage: Exploring large-scale vision foundation models with deformable convolutions," in *CVPR*, 2023.

[32] H. Seong, J. Hyun, and E. Kim, "Fosnet: An end-to-end trainable deep neural network for scene recognition," *IEEE Access*, 2020.

[33] A. Quattoni and A. Torralba, "Recognizing indoor scenes," in *CVPR*, 2009, pp. 413–420.

[34] A. López-Cifuentes, M. Escudero-Viñolo, J. Bescós, and Á. García-Martín, "Semantic-aware scene recognition," *Pattern Recognition*, 2020.

[35] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba, "Sun database: Large-scale scene recognition from abbey to zoo," in *CVPR*, 2010, pp. 3485–3492.

[36] J. Mahadeokar and G. Pesavento, "Open sourcing a deep learning solution for detecting nsfw images," *Retrieved August*, 2016.

[37] J. Rondeau, *Deep Learning of Human Apparent Age for the Detection of Sexually Exploitative Imagery of Children*. University of Rhode Island, 2019.

[38] F. Anda, N.-A. Le-Khac, and M. Scanlon, "Deepuage: improving underage age estimation accuracy to aid csem investigation," *Forensic Science International: Digital Investigation*, vol. 32, p. 300921, 2020.

[39] A. Gangwar, V. González-Castro, E. Alegre, and E. Fidalgo, "Attm-cnn: Attention and metric learning based cnn for pornography, age and child sexual abuse (csa) detection in images," *Neurocomputing*, 2021.

[40] J. Rondeau, D. Deslauriers, T. Howard III, and M. Alvarez, "A deep learning framework for finding illicit images/videos of children," *Machine Vision and Applications*, vol. 33, no. 5, p. 66, 2022.

[41] J. Dalins, Y. Tyshetskiy, C. Wilson, M. J. Carman, and D. Boudry, "Laying foundations for effective machine learning in law enforcement. majura–a labelling schema for child exploitation materials," *Digital Investigation*, vol. 26, pp. 40–54, 2018.

[42] B. Oreshkin, P. Rodríguez López, and A. Lacoste, "Tadam: Task dependent adaptive metric for improved few-shot learning," in *NeurIPS*, vol. 31, 2018.

[43] V. G. Satorras and J. B. Estrach, "Few-shot learning with graph neural networks," in *ICLR*, 2018.

[44] N. Mishra, M. Rohaninejad, X. Chen, and P. Abbeel, "A simple neural attentive meta-learner," in *ICLR*, 2018.

[45] H.-J. Ye, H. Hu, D.-C. Zhan, and F. Sha, "Few-shot learning via embedding adaptation with set-to-set functions," in *CVPR*, 2020.

[46] C. Doersch, A. Gupta, and A. Zisserman, "Crosstransformers: spatially-aware few-shot transfer," in *NeurIPS*, vol. 33, 2020, pp. 21 981–21 993.

[47] H. Chen, H. Li, Y. Li, and C. Chen, "Sparse spatial transformers for few-shot learning," *Sci. China Inf. Sci.*, 2023.

[48] S. X. Hu, D. Li, J. Stühmer, M. Kim, and T. M. Hospedales, "Pushing the limits of simple pipelines for few-shot learning: External data and fine-tuning make a difference," in *CVPR*, 2022.

[49] M. Hiller, R. Ma, M. Harandi, and T. Drummond, "Rethinking generalization in few-shot classification," in *NeurIPS*, 2022.

[50] Y. He, W. Liang, D. Zhao, H.-Y. Zhou, W. Ge, Y. Yu, and W. Zhang, "Attribute surrogates learning and spectral tokens pooling in transformers for few-shot learning," in *CVPR*, 2022, pp. 9119–9129.

[51] W. Chen, C. Si, Z. Zhang, L. Wang, Z. Wang, and T. Tan, "Semantic prompt for few-shot image recognition," in *CVPR*, 2023.

[52] H. Lin, G. Han, J. Ma, S. Huang, X. Lin, and S.-F. Chang, "Supervised masked knowledge distillation for few-shot transformers," in *CVPR*, 2023, pp. 19 649–19 659.

[53] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," *ICLR*, 2021.

[54] L. Liu, W. L. Hamilton, G. Long, J. Jiang, and H. Larochelle, "A universal representation transformer layer for few-shot image classification," in *ICLR*, 2021.

[55] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *CVPR*, 2009, pp. 248–255.

[56] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *ICML*, 2017, pp. 1126–1135.

[57] A. A. Rusu, D. Rao, J. Sygnowski, O. Vinyals, R. Pascanu, S. Osindero, and R. Hadsell, "Meta-learning with latent embedding optimization," in *ICLR*, 2019.