---

**Proceedings Paper:**
Coelho, T. orcid.org/0009-0001-4894-0347, Ribeiro, L.S.F. orcid.org/0000-0003-1781-2630, Macedo, J. orcid.org/0000-0001-5558-3088 et al. (2 more authors) (2025) Minimizing risk through minimizing model-data interaction: A protocol for relying on proxy tasks when designing child sexual abuse imagery detection models. In: FAccT '25: Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency. FAccT '25: The 2025 ACM Conference on Fairness, Accountability, and Transparency, 23-26 Jun 2025, Athens, Greece. ACM , pp. 1543-1553. ISBN 9798400714825

https://doi.org/10.1145/3715275.3732102

---

# Minimizing Risk Through Minimizing Model-Data Interaction: A Protocol For Relying on Proxy Tasks When Designing Child Sexual Abuse Imagery Detection Models

Thamiris Coelho
Instituto de Computação, Universidade Estadual de Campinas (UNICAMP)
Campinas, Brazil

Leo S. F. Ribeiro
Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo (USP)
São Carlos, Brazil

João Macedo
Departamento de Ciência da Computação, Universidade Federal de Minas Gerais (UFMG)
Departamento de Polícia Federal (DPF)
Belo Horizonte, Brazil

Jefersson A. dos Santos
Department of Computer Science, University of Sheffield
Sheffield, United Kingdom

Sandra Avila
Instituto de Computação, Universidade Estadual de Campinas (UNICAMP)
Campinas, Brazil



## Abstract

The distribution of child sexual abuse imagery (CSAI) is an ever-growing concern of our modern world; children who suffered from this heinous crime are revictimized, and the growing amount of illegal imagery distributed overwhelms law enforcement agents (LEAs) with the manual labor of categorization. To ease this burden researchers have explored methods for automating data triage and detection of CSAI, but the sensitive nature of the data imposes restricted access and minimal interaction between real data and learning algorithms, avoiding leaks at all costs. In observing how these restrictions have shaped the literature we formalize a definition of "Proxy Tasks", i.e., the substitute tasks used for training models for CSAI without making use of CSA data. Under this new terminology we review current literature and present a protocol for making conscious use of Proxy Tasks together with consistent input from LEAs to design better automation in this field. Finally, we apply this protocol to study — for the first time — the task of Few-shot Indoor Scene Classification on CSAI, showing a final model that achieves promising results on a real-world CSAI dataset whilst having no weights actually trained on sensitive data.

## CCS Concepts

• **Applied computing** → **Evidence collection, storage and analysis**; • **Security and privacy** → **Privacy protections**; • **Social and professional topics** → *Computer crime.*

## Keywords

Child Sexual Abuse Recognition, Few-shot Learning, Scene Classification

**ACM Reference Format:**

Thamiris Coelho, Leo S. F. Ribeiro, João Macedo, Jefersson A. dos Santos, and Sandra Avila. 2025. Minimizing Risk Through Minimizing Model-Data Interaction: A Protocol For Relying on Proxy Tasks When Designing Child Sexual Abuse Imagery Detection Models. In *Proceedings of ACM Conference XX XXXXXXXX, XXXXXXXXXXXXX, XXX XXXXXXXXXXX (ACM XXXXX).* ACM, New York, NY, USA, 11 pages. https://doi.org/XXXXXXX.XXXXXXX


*ACM XXXXX, XXXX XX–XX, 2025, XXXXXX, XXXXXX*

ACM ISBN XXX-X-XXXX-XXXX-X/XX/XX
https://doi.org/XXXXXXX.XXXXXXX

## 1 Introduction

Child sexual abuse is a global problem with severe and lasting consequences for victims. 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. These reports included more than 100 million image/video files containing child sexual abuse. From 2021 to 2023, the number of child sexual abuse materials increased by more than 20%, verified by NCMEC's CyberTipline (for the online exploitation of children). In 2023, NCMEC scaled more than 63 thousand reports to law enforcement from urgent cases or the child was in imminent danger, a 140% increase compared to 2021.

The proliferation of online platforms has facilitated the spread of child sexual abuse imagery (CSAI), making the development of effective detection tools crucial [6]. Unfortunately, on a day-to-day basis, it turns out that the volume of material produced is much greater than the capacity for the visual analysis carried out by law enforcement professionals. In this context, the automatic classification of CSAI is paramount.

Due to legal and ethical barriers — necessary barriers — such sensitive data cannot be accessed by anyone besides trained law enforcement agents (LEAs). For this reason, traditional methods

[1] https://www.missingkids.org/cybertiplinedata

use hash comparison to detect CSAI [15, 22, 55, 61]. While helpful, they are sensitive to minor changes to the file contents, a flaw that allows criminals to keep sharing the same content as long as they are not byte-wise identical to the originals flagged by LEAs [46]. Recent efforts have been made to make hash-based comparison more "perceptual", but these deep-learning-based comparisons come with other concerns when applied to such sensitive data, as recent methods such as NeuralHash [3] were shown to be susceptible image manipulations that fool detection and attacks that can surface image details from the produced hashes [52].

Given the limitations of hash-based methods, attention turned then to the exploration of more robust content-based techniques, from classic bag-of-visual-words matching [4, 48] to current deep representation learning [18] techniques. The challenges posed by ethical and legal requirements remain of course, but access to CSAI to evaluate content-based models is possible through partnerships with law enforcement. These partnerships make for more trustworthy studies that were tested on real-world CSAI but do not imply such data is available for training models; there is a severe concern regarding data leak through models given that models can often be reversed to reveal their training data [7] — to this end, the UK government explicitly prohibits this use of their Child Abuse Image Database even if handled by LEAs [55]); even models trained with Privacy Preserving techniques are prone to leaking general statistics of datasets [66].

The question of minimizing model-data interactions then remains and this unassailable restriction has driven the use of a common pipeline in automatic CSAI recognition: before evaluating on CSAI, researchers often go through the usual CRISP-DM-like [45] pipeline (of task understanding, data understanding, data preparation, modeling, and evaluation) while working under an accessible, different — but related to CSAI — task; In this manuscript we formalize and name these as "Proxy Tasks" and further disassemble them as a (i) *related sub-task of CSAI recognition* (e.g., age estimation), which is often combined with (ii) *restrictions common in CSAI* (e.g., few labeled samples). Only after a model is produced under the Proxy Task it is evaluated with CSAI to perform either direct recognition or to use the Proxy Task as a data triage step that accelerates the work of LEAs.

This manuscript then presents the literature on automatic CSAI recognition while specifically framing each study through their use of one or more Proxy Tasks. We then observe gaps in the use of these tasks and present a protocol for developing new systems for automating recognition and triage of CSAI. Our proposal gives weight to consistent input from LEAs, specialists that best know the real data, and explicitly states the role of Proxy Tasks in the design of these new systems and how studies can consider new Proxy Tasks.

We show the effectiveness of this protocol with a case study. Through input from LEAs — from direct partnerships and the literature — we identify an under-explored aspect of CSAI with regards to scene composition and locale (a *related sub-task of CSAI recognition*) and study this aspect through the lens of few-shot Learning (given few labeled samples is a *restriction common in CSAI*), introducing for the first time the Proxy Task of few-shot indoor scene classification. Through thorough experimentation under the Proxy Task and final evaluation on CSAI we have found that (i) the state-of-the-art in few-shot learning can be successfully applied to classify indoor scenes in CSAI and that (ii) the features learned under this task are indeed relevant to direct CSAI classification.

The contributions of this manuscript are then that (1) we introduce the terminology of Proxy Tasks and apply it to current literature; (2) through this new frame we introduce a protocol for choosing new Proxy Tasks and designing new automated systems to aid in CSAI recognition and (3) we demonstrate the effectiveness of this protocol with the introduction of the novel Proxy Task of few-shot indoor scene classification, showing potential in the study of scene features for CSAI and promise in the exploration of Few-shot methods that forgo having learned weights depend on CSAI and contribute to minimizing model-data interactions.

## 2 Proxy Tasks and the Literature on CSAI Recognition

As previously discussed, much of the CSAI recognition literature is dedicated to improving upon the shortcomings of hash matching methods [5, 12, 22, 35] that are widely used by organizations around the world, including social media moderation practices. The biggest shortcoming of these methods is of course that they are not robust to visual changes and not designed to work with completely new materials being uploaded daily. The major challenge in developing robust and semantic methods for detecting CSAI is how access to real CSAI data is (and should be) restricted. Researchers have then relied on tasks that represent individual challenges but have public data for open evaluation. We call these tasks "Proxy Tasks".

We formalize these Proxy Tasks as a combination of two aspects, first and foremost, they are a (i) *related sub-task of CSAI recognition*. Identifying what these related sub-tasks are requires thorough communication and partnership with LEAs, the domain experts and stakeholders of CSAI Recognition. As an example, when discussing aspects used by LEAs for manual recognition, Laranjeira da Silva et al. noted that features like age, gender, face detection, ethnicity, scenes and objects can all offer valuable insights [27] for a final decision on whether a photograph depicts CSA or not. In addition to a sub-task, we believe it is relevant to consider (ii) *restrictions common in CSAI*; while this is not an aspect that has always been taken into account in the literature, we deem to highlight it given that these restrictions can be a roadblock to implementation and use by LEAs. A non-exhaustive list of such restrictions are: it is not possible to guarantee LEAs will always have the best hardware; model-CSAI interaction should be kept to a minimum; there will always be significant distribution shifts between CSAI and public data. We summarize these two aspects that compose Proxy Tasks in Fig. 1.

We turn our attention now to the literature of robust and semantic CSAI Recognition, identifying within each study the Proxy Task employed. Within the last decade, most research effort have been dedicated to using nudity detection, age estimators and combinations of both. In the studies of Polastro and Eleuterio [36, 37], the NuDetective tool was introduced, a nudity detection tool that was effective in detecting nakedness also on CSA images, making for a useful triage model in CSAI recognition. The work of Sae-Bae et al. [43] then combined facial and skin region features to detect nakedness and estimate age, while Schulze et al. [46] also

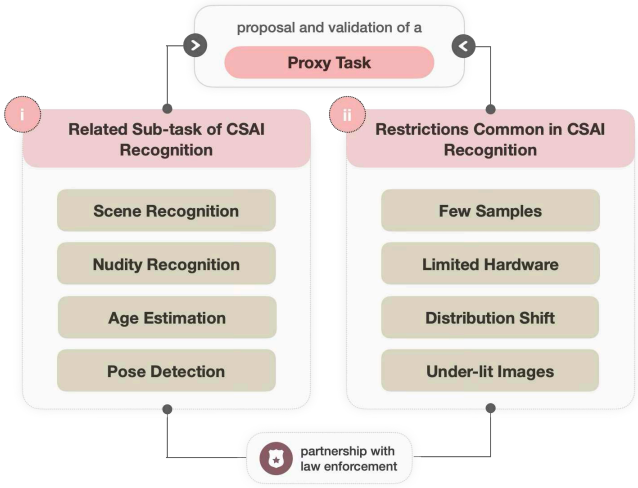



**Figure 1: The sensitive nature of CSAI requires that most research be done under a related substitute task before being applied to real data. Proxy Tasks combine a related sub-task of CSAI recognition with restrictions common in CSAI data; the figure shows examples of both aspects.**

employed nudity detection but combined it with a new Proxy Task, sentiment analysis. This latter study argued that by using nudity detection through skin region features, CSA images were easily confused with legal pornography, images that have very different contexts and image compositions. The use of sentiment analysis however made up for this and showed good potential for detection and explainability.

While introducing the first deep learning model, Castrillón-Santana et al. [8] relied on the age estimation Proxy Task for their model and fused feature vectors from CNNs with classic "shallow" descriptors. The study of Macedo et al. [29] made use of a pre-trained Yahoo OpenNSFW model [30] — a general-purpose pornography detection model — for nudity detection while also designing an age estimation model that takes face crops as input, again combining these common Proxy Tasks. The study of Vitorino et al. [59, 60] followed suit, but also showed through a two-tiered fine-tuning pipeline that there is significant benefit in using images from the nudity detection task for finetuning off-the-shelf pre-trained models before use with CSAI. Their hypothesis is that the improvement comes from softening the significant distribution shift between public and sensitive data; this consideration makes it the first study to comment on restrictions common in CSAI recognition, that we argue should be one of the core aspects of a well designed Proxy Task.

Used either for triage or for composing a full pipeline, the age estimation task remained popular for applications on CSAI, with multiple studies being dedicated exclusively to differentiating children from adults; The studies of Anda et al. [2] and Castrillón-Santana et al. [8] have introduced their own datasets, respectively VisAGe and AgeMega, while others have collected images of children from several datasets already available [9, 17]. The practice of using images of children raises concern of course, the UNICEF's Responsible Data for Children report [64] strongly recommends against images (or any other related materials) of children to be collected from online sources without proper consent and more specifically without "principles, practices, and tools that can enable the responsible handling" of such materials. While this discussion is not the focus of our manuscript, we find it is always appropriate to mention these shortcomings of the research we aim to extend.

When regarding the other popular Proxy Task of nudity detection, recent studies from Al-Nabki et al. [1] and Tabone et al. [54] have gone for a more specific definition of the task, specifically targeting reproductive organ detection for data triage purposes. The recent study of Valois et al. [56] introduced the related sub-task of indoor scene classification and studied self-supervised models as they are known to yield general representations that can better adapt to distinct downstream tasks with large distribution shifts. Our case study considers the same related sub-task of indoor scene classification but looks into the restriction of few samples being available for CSAI and studies few-shot methods for indoor scene classification.

In this section then we have shown how the pattern of using Proxy Tasks was present in the literature long before we wrote this manuscript and named them as such. We believe that by highlighting the studies through their Proxy Tasks future work can strive to have results that are more comparable across common tasks even if real CSAI data is not (and should not be) share-able for comparison. We furthermore hope that more Proxy Tasks are proposed, because as Lee et al. [28] pointed out, solutions that achieve the best results — even when doing manual detection — are often made with a combination of multiple methods (and therefore, multiple Proxy Tasks). In the next section we present our own protocol for proposing, designing methods for and evaluating Proxy Tasks.

## 3 Protocol for Developing CSAI Detection Solutions

While our protocol in general does not significantly differ from the aforementioned common pipeline of developing models under Proxy Tasks, we use this space to both formalize it and advocate for a few changes in how one should approach CSAI detection.

At the forefront of these changes, we believe that to further progress with automation when dealing with CSAI we must consider the nascent field of Data-Centric AI. Given the growing presence of AI-driven applications in our daily lives, it is unsurprising that desire has grown for a renewed focus on how these ever-larger models use data and how this data is analyzed, collected and used. A renewed focus then on a data-driven, FAccT-checked (Fairness, Accountability, and Transparency) approach to model design [31, 38, 65]. When considering CSAI, we see that only through consistent input and partnerships with LEAs we are able to have a better understanding of the available and real-world CSAI data to the interest of designing Proxy Tasks and better benchmarks.

The first step on our protocol is then to discuss with partner LEAs and look to CSAI literature beyond computer vision approaches to better understand the data, the restrictions and possible Proxy Tasks. Very few studies in the literature match this need for a Data-Centric in CSAI. In the their recent study, Laranjeira da Silva et al. [27] designed an analysis framework for child sexual abuse materials

(CSAM) data focused on providing valuable insights from the data to researchers without actually giving direct access. Their framework was applied to the Region-based Annotated Child Pornography Dataset (RCPD) [29] and showed how there are many facets to CSAM identification, including nudity and age (as has been explored) but also scene composition and illumination. They have also shown how current deep learning models correlate with real labels regarding object detection and people counting, allowing for automated analysis of these datasets. In the pair of studies from Kloess et al. [23, 24], LEAs were interviewed to highlight the hardest challenges in identifying CSA materials and their findings showed that the environment and objects present in a scene could provide clues as to whether it is inappropriate or not. They also explained that nudity is not an immediate flag for indecency; the posing of subjects and the presence of other images of a victim found together can be better clues of intention. We took these examples together with our own talks with partner LEAs at heart when considering a Proxy Task for our case study presented in Section 4.

A Data-Centric approach to Proxy Task design is of course only the starting point of our protocol, we better detail all the steps below and organize them in a diagram presented in Fig. 2.

(1) **Data-Centric Analysis**: We believe that only through consistent input from LEAs — the stakeholders of these systems — and thorough consideration of CSAI data we are able to better understand what are the features and their related Proxy Tasks best suited for improved automation of CSAI recognition. The resulting benchmarks, datasets, and proposed Proxy Tasks from these studies should then drive research independent from these sensitive materials, minimizing model-CSAI interaction.
(2) **Formalization of the Proxy Task**: With each Proxy Task that has never been attempted before (e.g., our case study, few-shot indoor scene classification), training and evaluation protocols and datasets to benchmark on must be defined.
(3) **Experiments with SoTA under Proxy Task**: Together with the formal definition of Proxy Tasks, practitioners should define where the current literature is placed in relation to the chosen task and its constraints. A good start is through creating baselines by benchmarking the best models from the related literature (e.g., SoTA models from general few-shot learning literature).
(4) **Design of New Methods under Proxy Task**: Having established benchmarks and baseline methods, the next step is to to design new methods specifically for the Proxy Task at hand, taking into account both the CSAI-related sub-task and the restrictions considered.
(5) **Evaluation with CSAM Dataset**: With a new method for the Proxy Task at hand, practitioners should make the necessary adaptations and evaluate the model either/both for automated CSAM recognition and data triage under real-world data. This step is of course done in partnership with LEAs, and its results feed back into the Data-Centric Analysis, closing the loop.

With this protocol at hand we show in our case study how it serves as a guide to the evaluation of a new Proxy Task, few-shot learning for indoor scene classification. We believe that the development and improvement of models through the lens of Proxy Tasks will allow for a future where LEAs have a library of Proxy models to choose from when evaluating a case and such library will make for better data triage, improved ensemble/combined models and better explainability and transparency on (semi-)automated decisions.

## 4 Case Study: Few-shot Learning for Indoor Scene Classification

As discussed in Section 3, one of the key intuitions from our initial analysis on Data-Centric studies of CSAI data and our own conversations with LEAs is the importance of features beyond nudity detection, such as the scene composition, placement and position of objects and people in a scene [23, 24, 27]. To that end we follow Valois et al. [56] in basing our Proxy Task on the related sub-task of indoor scene classification. The specificity of only looking at "indoor" scenes stems from CSAI being seldom depicted in public spaces and the task of scene classification being a useful way of evaluating aspects of composition, placement and positioning of entities in a scene while also being a task that is well documented and is present in public benchmarks.

In addition to the related sub-task, we believe a good Proxy Task should take into consideration restrictions common when working with CSAI. We know from the literature and our talks with LEAs that large datasets are not available for CSAI (at least not for training models, and definitely not publicly) and it is furthermore unadvisable to incur LEAs with the cost of labeling more data to that end; after all, the mental health crisis in CSA investigation units is a significant concern and a common challenge faced by professionals in this field [51] and we do not wish to increase that burden. The restriction we consider is then that there are and there will often be very few labeled samples for CSAI. We turn our attention to a particular learning protocol from the general computer vision literature and compose our Proxy Task as *few-shot* indoor scene classification.

More objectively and following the steps suggested by our protocol, our aim is then **to evaluate the state-of-the-art in few-shot learning using public data for indoor scene classification**. The best-performing model under this task will then be **tested with CSAI data through our partnership with LEAs**.

### 4.1 The Literature as it Relates to the Proxy Task

*Scene classification* aims to classify indoor and outdoor environments. While global spacial properties are enough in outdoor environments, both global and local properties are essential in indoor scenes, making the classification a challenge [34]. In those environments, the objects in the scene can be crucial to the classification [39].

Classic supervised methods for scene classification have achieved notable success on benchmarks. FOSNet [47] achieved an accuracy of 90.3% for MIT Indoor [40] and 77.3% for SUN-397 [62]; while the accuracy for Places365, a more robust dataset, drops significantly to 60.1%. OmniVec2 [50], the state-of-the-art for Places365, achieved 65.1%. These methods are however designed to be trained with

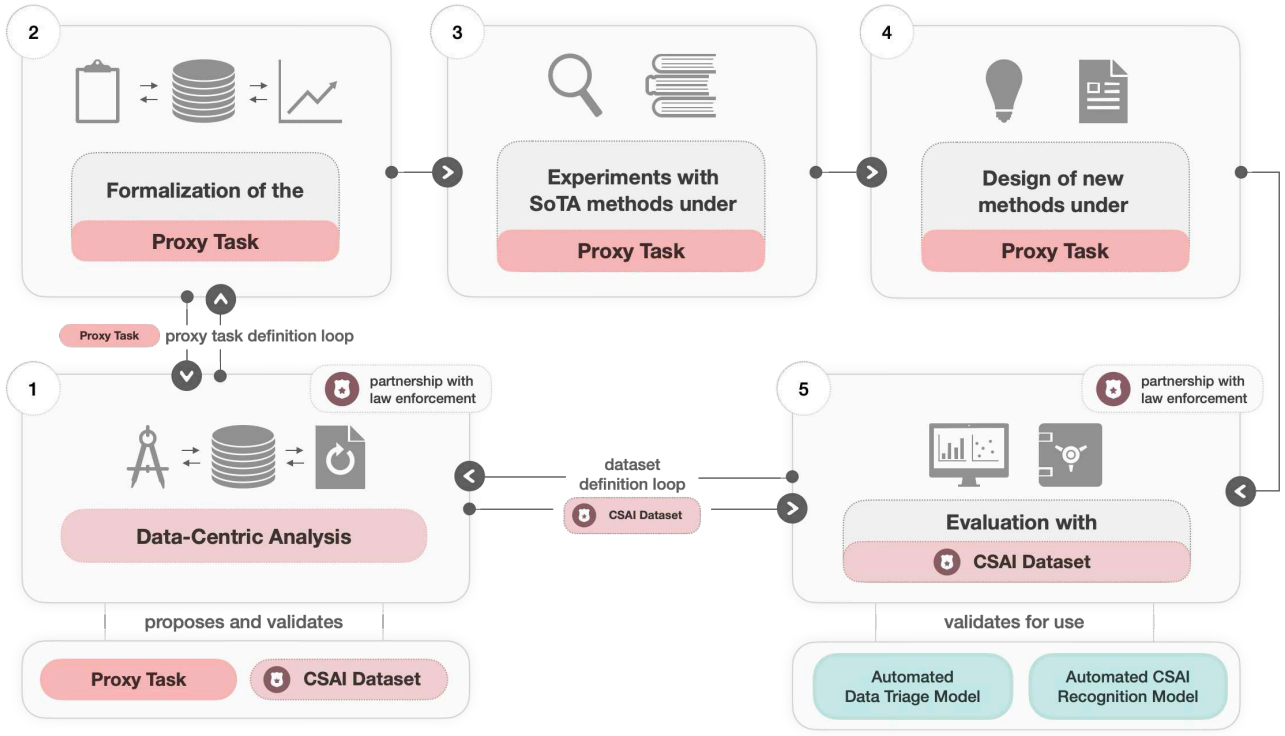



**Figure 2: Diagram depicting the cycle of steps on our protocol. We believe that Proxy Tasks and CSAI evaluation datasets should be designed with a Data-Centric mindset; Proxy tasks are then used in lieu of real CSAI data to minimize model-CSAI interactions; real CSAI is used only through our partnerships with LEAs to evaluate final models.**

large datasets and we turn our attention instead to methods from the general few-shot learning literature.

*Few-shot learning (FSL)* is a machine learning method that learns from a few labeled examples, unlike traditional methods that require large annotated datasets. It is particularly useful for real-world scenarios where obtaining extensive labeled data is unfeasible. Most FSL methods learn how to transfer knowledge from a large dataset related to the target task to a small labeled dataset with samples of interest [21].

Many studies in the field adopt a meta-learning paradigm, where the model learns to learn [16, 49, 53, 58]. For that, a model learns from the large dataset (meta-training), mimicking the few-shot task; then, the learned model is applied to the target task dataset (meta-testing). The learned model can be used as model initializer [16], optimizer [41], or metric learner [49].

In this case study, we focus on metric learning approaches, which aim to learn how to compare samples. If a model knows how to measure the similarity between samples, it can also classify unseen samples by comparing them [26]. More specifically, we employ embedding learning [49, 53], a technique under the metric learning umbrella. The idea is to learn how to generate the best embedding function for the samples so that same-class sample embeddings are closer to embeddings of samples from different classes. The core advantage of this approach is that once the embedding function is learned the weights are not updated on the target task and classification is performed only through comparing the learned representations; for our use case that means that no weights are trained or learned from sensitive data, minimizing model-data interaction.

## 4.2 Few-shot Learning Problem Definition

In FSL, the goal is to classify unseen samples using only a few labeled data. Such a task is referred to as $N$-way $K$-shot classification, where $N$ is the number of classes considered and $K$ is the number of labeled samples per class. It is common for FSL methods to consider a transfer learning approach, where a model is first pre-trained on a base set $D_{train}$ and then the FSL evaluation considered under a separate unseen set $D_{test}$, where the classes from each set are disjointed: $C_{train} \cap C_{test} = \emptyset$.

We opted for methods that follow this approach specifically so that knowledge accumulated from public data of scene imagery can later be employed within the domain of CSAI. In further detail, this pre-training stage follows the *episodic meta-learning* concept introduced by Vinyals et al. [58]; for each training step, an episode is built by simulating one $N$-way $K$-shot task. The episode $E$ is composed of a support set $S_{sup} = \{S^{nk}_{sup} = \{x^{nk}_{sup}, y^{nk}_{sup}\} \mid n = 1, ..., N; k = 1, ..., K\}$ and a query set $S_{qry} = \{S^{n}_{qry} = \{x^{n}_{qry}, y^{n}_{qry}\} \mid n = 1, ..., N\}$, where the former represents a small training set and the latter a test set in the simulated task. *Episodic meta-learning* then is to learn — often through end-to-end gradient descent — to improve performance on these episodes. A matching protocol, *episodic meta-testing*, is

used for evaluating methods; the difference is that within meta-training episodes are sampled from the base set $D_{train}$ and within meta-testing from the unseen set $D_{test}$; because of our specific choice of using embedding learning methods, no gradient steps are taken during *episodic meta-testing* and learning is done through comparing sample representations (i.e., as in a nearest neighbors classifier).

## 4.3 Comparative Analysis

*4.3.1 Methodology.* We chose eleven FSL methods from the literature for comparison. Most of these comply with the Embedding Learning paradigm for FSL; differences can be found in model architecture and specific loss designs. Specific hyperparameters were replicated from the original studies.

*Pre-training Datasets.* Methods are pre-trained on *mini*ImageNet, with P>M>F [21] and CPEA [19] being the exceptions presenting models trained with ImageNet-1K [13].

*Fine-tuning Dataset.* This is the first study to address FSL for scene classification. Testing is to be done with the Places8 dataset, proposed by Valois et al. [56] to represent indoor scenes common in CSAI. It is also sensible to have a Base set within the domain of scenes. Following the design of *mini*ImageNet, we present *Places600*, a subset of Places365 [67] designed to have 600 samples per class, excluding classes with insufficient samples, and not to have class overlap with Places8. We ended up with a dataset that contains 268 classes. *Places600* is used to fine-tune methods for the domain of scenes. The design of *Places600* and the choice of Places8 as a testing ground are part of what we defined as the (2) Formalization of the Proxy Task step within our protocol.

*Training and Evaluation Details.* Unless specified otherwise by the original authors, models were trained for 100 epochs composed of 2,000 episodes each. For evaluation, 10,000 episodes are randomly sampled from Places8 with the 8-way 5-shot setup, allowing all 8 classes to participate in all tests. Each evaluation episode contains 15 test queries per class. A single NVIDIA Tesla T4 was used in all experiments, and the training time for each experiment was an average of less than two days. All code and data (*Places600)* are available at https://github.com/thamycoelho.

*4.3.2 Results.* The comparison of state-of-the-art FSL methods and selection of a best performing model on the selected indoor scenes benchmark comprise steps (3) and (4) of our protocol. We report the comparison result in Table 1. For a fair comparison, these results all used *mini*ImageNet for pre-training. We can see that the model based on PMF, fine-tuned on *Places600,* was the best performing. The core contribution of the authors of PMF [21] was showing that an updated architecture combined with larger-scale pre-training surpassed other approaches in FSL. With their proposed model performing the best, we follow their experimental design and compare PMF and second-best method CPEA using different backbone networks and the full ImageNet-1K dataset for pre-training.

The result of this comparison can be found in Table 2. PMF continues to perform better. More importantly, these results surpass the ones that used *mini*ImageNet by a large margin, corroborating

**Table 1: Results of few-shot methods on Places8 validation set. Pre-trained and fine-tuned top-1 accuracy results are reported. Best result presented in bold.**

| Method | Backbone | Pre-trained | Fine-tuned |
|---|---|---|---|
| ProtoNet [49] | ResNet-12 | 37.48 ± 0.10 | 43.13 ± 0.10 |
| MAML [16] | Conv-64 | 34.42 ± 0.09 | 36.42 ± 0.10 |
| RelationNet [53] | ResNet-12 | 30.49 ± 0.09 | 38.69 ± 0.10 |
| Baseline++ [11] | ResNet-18 | – | 38.36 ± 0.09 |
| LEO [42] | ResNet-18 | 31.09 ± 0.09 | 32.66 ± 0.91 |
| FEAT [63] | ResNet-18 | 45.43 ± 0.09 | 45.34 ± 0.10 |
| CTransf. [14] | ResNet-34 | 46.67 ± 0.10 | 45.07 ± 0.22 |
| PMF [21] | ViT Small | 45.49 ± 0.10 | **52.86** ± 0.10 |
| HCTransf.[20] | ViT Small | 44.39 ± 0.10 | 44.03 ± 0.10 |
| SSFormers [10] | ResNet-12 | 41.20 ± 0.11 | 46.27 ± 0.12 |
| CPEA [19] | ViT Small | 49.25 ± 0.30 | 51.65 ± 0.30 |

**Table 2: Results of best-performing few-shot methods, pre-trained on ImageNet-1K to show the impact of large-scale data for pre-training. Top-1 accuracy on the Places8 validation set shown.**

| Method | Backbone | Pre-trained | Fine-tuned |
|---|---|---|---|
| PMF [21] | ViT Small | 69.76 ± 0.09 | **71.86** ± 0.10 |
| | ResNet-50 | 59.63 ± 0.10 | 61.29 ± 0.10 |
| CPEA [19] | ViT Small | 64.74 ± 0.30 | 67.79 ± 0.30 |

Hu et al. [21]'s findings. We also show how PMF performs with a ResNet-50 backbone and confirm that the newer Transformer-based ViT is the better choice. Our final model is then a PMF model, pre-trained on ImageNet-1K with the ViT Small backbone; with this model at hand, we have further analyzed its hyperparameters and found that through using stepped learning rate decay every 10 epochs and a temperature of 0.07, an accuracy of 72.50% is achieved.

With our best model at hand, we can move our analysis to the test set of Places8. The model achieved an average accuracy of 73.50 ± 0.09%. Through the confusion matrix (Fig. 3), we can observe that *bedroom*, *living room* and *child's room* were often misclassified between them; there is also confusion between *child's room* against *bedroom* and *classroom*. To better observe these missteps, we plot the embeddings from our model using t-SNE [57]; Fig. 4 shows feature vectors for these classes have their clusters but that they indeed overlap with each other.

The standard FSL evaluation protocol makes use of test samples to compose episodes. To push the boundaries of our model beyond FSL, we design a custom *Comparable Protocol*. To avoid what would be characterized as a "leak" in traditional classification, our protocol samples the support set (5-shot, 5 samples per class) from the validation set of Places8 and uses the entire test set (40,534 samples) for evaluation for each episode. When we follow the standard FSL evaluation protocol, the query set in each epoch is balanced, but we can not assume that for the *comparable protocol*. Importantly, this protocol does not relinquish the claim to being a FSL protocol;


**Figure 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 FSL evaluation protocol. Ground truth labels are on the vertical axis; predicted labels on the horizontal axis.**

we after all use only a few samples (5) in the support set. The difference is that the support and the query sets can have samples from different domains and, more importantly, not using test samples to compose support sets means the results obtained with this protocol are *comparable* against other traditional machine learning uses of the benchmark.

Under the *comparable protocol*, we achieve an accuracy of 69.25 ± 0.05%. This is a competent result; Valois et al. reported 71.6% with their self-supervised pipeline while actually fine-tuning weights on the entirety of the large Places8 training set; our model, on the other hand, used only 5 samples per Places8 class and did not employ fine-tuning, only comparisons among sample representations.

## 4.4 Evaluation on Child Sexual Abuse Imagery

We conducted the final tests in collaboration with a Brazilian Federal Police agent, per step (5) on our protocol: Evaluation with CSAI Dataset. The agent annotated 374 randomly chosen samples with the indoor classes to evaluate the model on indoor scenes. This set was sampled from an internal collection of images (4,592) categorized into: *CSAI*, *suspected CSAI*, *drawing*, *people*, *porn*, and *other*. As the samples were randomly selected, they only contained 6 out of 8 classes from Places8: *bathroom*, *bedroom*, *child's room*, *living room*, *studio*, and *swimming pool*.

We followed both the traditional FSL evaluation and our *comparable protocol*. In both, we evaluated the model for 10,000 episodes. For the few-shot evaluation protocol, instead of selecting 15 samples per class to compose the query set, due to the limited number of annotated samples, all remaining samples not in the support set were used to compose the query set. Fig. 5 shows the confusion matrix for both experiments.

For the few-shot protocol, the average accuracy from the episodes with 95% confidence is 63.38 ± 0.09%. From the confusion matrix, the major confusion occurs among *bedroom*, *child's room*, and *living room*, a similar behavior observed in the evaluation on Places8. The model was good at classifying *bathroom* and *studio*. Given those results, we believe the few-shot proposed model is generalized for the CSAI environment from only a few samples.

For the *comparable protocol*, the model could classify well *swimming pool* and *bathroom*; there was also a considerable confusion between *bedroom* and *child's room*, which was expected given the results on Places8. The model could not classify *living room*, classifying most of the samples as *bedroom* and *child's room*. For the samples from *studio*, the model is also confused with *swimming pool* and *child's room*. A similar behavior was observed by Valois et al. [56], who obtained a balanced accuracy of 36.7%. Our model achieved a balanced accuracy of 43.43 ± 0.09% with 95% confidence.

Although the results on the CSAI indoor dataset were lower than the results obtained on Places8, indicating a domain shift, the model still performed better than the self-supervised model proposed by Valois et al. [56]. This suggests that the proposed model is less specialized for Places8 and more capable of generalization. As for the difference between this result with the *comparable protocol* and the FSL protocol presented before, we highlight that on the latter each episode's support set is composed of images from the CSAI indoor collection while the former uses public images from Places8 as support; this distinction is further evidence of both the domain shift between Places8 and CSAI and of our model's capability for generalization with few samples.

To understand the importance of the features of indoor scenes to the CSAI recognition itself, since the model is trained to compare samples, we evaluated it on the 4,592 internal collection of images; none of the classes from this collection are present in Places8. Therefore, we only performed the FSL evaluation protocol. Fig. 6 shows the confusion matrix of the experiment.

The model achieved an average accuracy through all the episodes for the CSAI dataset with 95% confidence of 50.82 ± 0.11%. From the confusion matrix, we can observe confusion among *CSAI*, *suspected CSAI*, and *porn*. It makes sense, given that those samples are probably visually similar, making it difficult even for police agents to differentiate those classes. However, the model effectively classified *drawing*, possibly due to their distinct domain from the training data. There was a slight confusion between *other* and *drawing*, probably because the class *other* contains images of documents, banknotes, and handwritten documents. Hence, the model probably understood some drawings as documents and vice versa. The model struggled to classify *people*, often confusing them with *CSAI*, *suspected CSAI*, and *porn*. We believe that this confusion is potentially due to the presence of nudity or seminudity without sexual connotations. Considering only *CSAI* and *suspected CSAI* classes and mapping all the others to *not CSAI*, the balanced accuracy increases to 53.44 ± 0.16%.

We also performed evaluation on the region-based annotated child pornography dataset (RCPD) [29], a CSAI benchmark composed of over 2,000 samples among regular and CSAI images, produced by Brazilian Federal Police. We performed only the FSL evaluation protocol, since the dataset does not contain indoor annotation. In each of the 10,000 episodes, five samples per class (10 samples

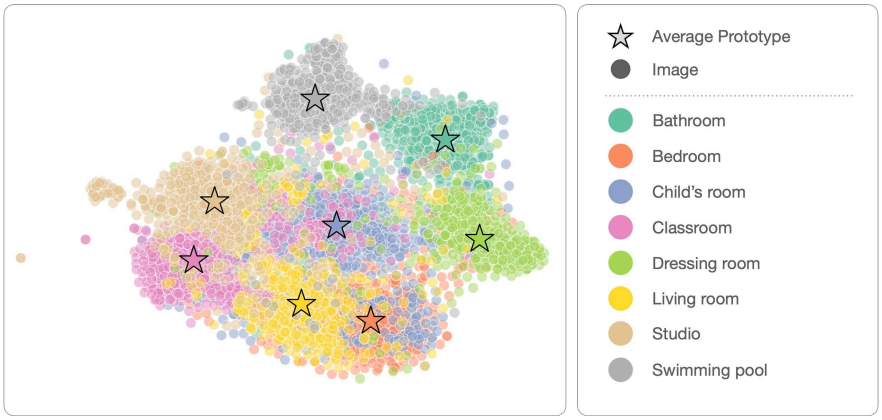
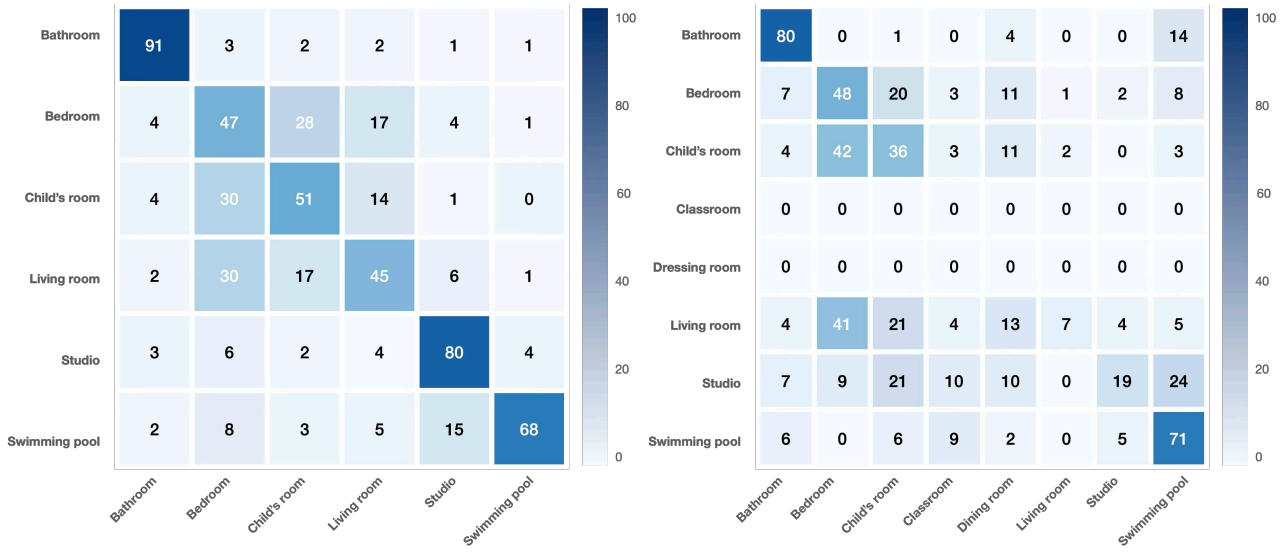



**Figure 4: t-SNE on two dimensions for the Places8 test set. Many children's room samples (blue dots) are closer to the bedroom prototype (orange star) than to the child's room (blue star). We observe the same phenomenon between classroom samples (pink dots) and child's room samples (blue dots).**




**(a) Few-shot protocol** **(b) Comparable protocol**

**Figure 5: Confusion matrix with the model's results on the CSAI samples classified into indoor. Values reported are the accuracy (%) for prediction per class. Ground truth labels are on the vertical axis; predicted labels on the horizontal axis.**

total) were randomly selected for the support set and 15 samples per class (30 samples total) for the query set.

The model achieved an average accuracy with 95% confidence of 65.40 ± 0.16%. Fig. 7 shows that the model was good at classifying CSAI samples, while it struggled with non-CSAI samples. This behavior is probably due to images containing nudity or seminudity that were not CSAI. We can see nonetheless that such a model would be useful to immediately identify images that are suspect of being CSAI.

The results from both datasets indicate that the model can classify samples reasonably well even with little data in the support set. This suggests the potential of using indoor scene classification to help in CSAI investigations. Combined with other complementary features, it could build a robust CSAI classifier.

## 4.5 Case Study Main Takeaways and Future Work

With this case study we have shown the potential not only of our protocol for investigating future Proxy tasks for CSAI recognition but we have also shown how few-shot learning and indoor scene classification can benefit CSAI investigation. Our results demonstrate the benefits of leveraging pre-trained models and adapting them with limited indoor scene data relevant to CSAI investigations.

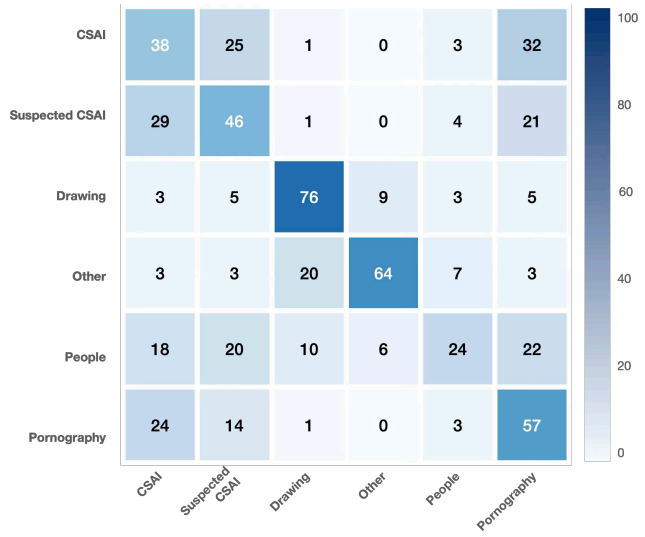



**Figure 6: Confusion matrix with the model's results on the samples with CSAI annotation. Values reported are the accuracy (%) for prediction per class. Ground truth labels are on the vertical axis; predicted labels on the horizontal axis.**

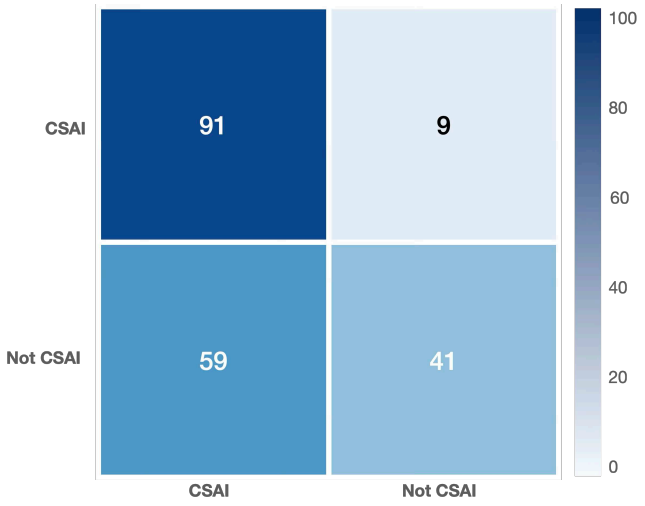



**Figure 7: Confusion matrix with the model's results on RCPD [29]. Values reported are the accuracy (%) for prediction per class. Ground truth labels are on the vertical axis; predicted labels on the horizontal axis.**

We highlight how the few-shot protocols allow for models to be used on a different domain than what they were tested on and particularly embedding based FSL allows for this use without training model weights on sensitive data. This allows for such models to be shared across law enforcement agencies without fear that sensitive data will leak through model weights.

While five samples per class is a standard benchmark for FSL, future work could explore varying sample sizes in the support set could optimize performance for CSAI recognition tasks. This investigation could determine the ideal number of samples to balance accuracy and efficiency in real-world applications.

Furthermore, investigating diverse classifiers beyond cosine similarity-based ones, such as Euclidean distance, linear classifiers, or appended neural networks, could bring more accurate results [10, 14, 32, 33, 44, 53]. Further addressing the inherent domain shift between standard datasets and CSAI is also paramount.

Finally, we would like to raise awareness to a limitation present in our case study and also common to most models trained on Proxy Tasks and used on CSAI. When making use of public benchmarks and public datasets from related tasks we are risking "contaminating" the CSAI recognition task with a non-neutral representation of the related sub-task [25] carried forward from the original dataset's intent and biases; this distancing from original purpose to final evaluation further muddles the implications of these data choices. For example, in preliminary talks with LEAs regarding the Places365 dataset they have already pointed out how the depiction of "bedroom" and "child's bedroom" in Places365 often presents a skewed "perfect" or "magazine-like" idea of these classes. This bias is certain to have an impact when models are used to identify bedrooms from distinct backgrounds and under more difficult lighting conditions.

## 5 Conclusion

Building a method that will be applied to a dataset that is never seen is a considerable challenge; not only that, but creating a classifier that will be used for data that sometimes does not have a precise classification is also difficult. With this manuscript we wanted to create a path to make more research on CSAI possible. Through a formal definition of Proxy Tasks, authors can have a better chance of comparing their results on public benchmarks for these related sub-tasks of CSAI Recognition.

We furthermore encourage that these Proxy Tasks are well thought out, that law-enforcement is never left out of the conversation and that we do not forget that both the data and the law-enforcement agencies that store that data will have constraints on how much labeled data there is, how it can be used and how models should or should not interact with such data.

We introduced a common protocol for investigating these Proxy Tasks and showed a case study on few-shot learning for indoor scene classification; through our case study we were able to observe both the benefits and the shortcomings of this Proxy Task. Our expectation is that systems used by LEAs in the future are actually constituted of multiple Proxy Task-based models, each highlighting one aspect of a full investigation of these images, from scene classification to nudity detection and age estimation.

## 6 Ethical Considerations Statement

All CSAI data was only ever handled by LEAs authorized to do so and CSAI data was never taken out of the premises of partner law enforcement agencies authorized to handle it; we have provided LEAs with our own code for evaluation on their authorized hardware. Furthermore, the goal of this manuscript is to introduce a way to study the problem of child sexual abuse imagery (CSAI) detection while minimizing the interactions between models and CSAI and avoiding entirely interactions between CSAI and practitioners

that are not LEAs. We are furthermore interested in galvanizing research on CSAI detection so that LEAs are less likely to have to check CSAI manually, a task that negatively impacts them daily.

While working with partnership LEAs, we prioritized minimizing their interaction with CSAI caused by our research; of note is that our case study experiments with CSAI were only executed once, with the model found to be best suited for the Proxy Task. We hope this practice becomes a pattern in the field, again to avoid increasing the burden on LEAs.

## 7 Adverse Impacts Statement

We list below our considerations of two potential adverse impacts of our work.

*Evaluation only with Proxy Tasks.* While we discourage this practice, it is possible that some practitioners that do not have an established partnership with LEAs make use of Proxy Tasks and their attached benchmarks isolated from evaluation on CSAI or conversations with LEAs. While achieving better metrics on Proxy Tasks is encouraged, we highlight the importance of keeping the stakeholders of the final task in mind and always seek their guidance and help with executing final evaluations on real CSAI.

*Proxy Tasks and their models could be collected with malicious intent.* We have highlighted that one of the advantages of experimenting first with Proxy Tasks is that, because they were not trained and do not carry any information about CSAI within their weights, these models can be freely shared and reproduced, improving potential for collaboration. A malicious actor could collect these models and what aspects of CSAI each model represents to either check their own illegal CSAI against them or use them to search for CSAI in large public datasets.

## Acknowledgments

This work is partially funded by FAPESP 2023/12086-9, PIND/FAEPEX UNICAMP 2597/23, and the Serrapilheira Institute R-2011-37776. T. Coelho is also funded by Becas Santander and Alumni Grant (Instituto de Computação). 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] Mhd Wesam Al-Nabki, Eduardo Fidalgo, Roberto A Vasco-Carofilis, Francisco Janez-Martino, and Javier Velasco-Mata. 2020. Evaluating performance of an adult pornography classifier for child sexual abuse detection. *arXiv preprint arXiv:2005.08766* (2020).

[2] Felix Anda, Nhien-An Le-Khac, and Mark Scanlon. 2020. DeepUAge: improving underage age estimation accuracy to aid CSEM investigation. *Forensic Science International: Digital Investigation* 32 (2020), 300921.

[3] Apple. 2021. *CSAM Detection - Technical Summary.* Technical Report. Apple.

[4] Sandra Avila, Nicolas Thome, Matthieu Cord, Eduardo Valle, and Arnaldo de A. AraúJo. 2013. Pooling in Image Representation: The Visual Codeword Point of View. *Computer Vision and Image Understanding* 117, 5 (2013), 453–465.

[5] Rubel Biswas, Victor González-Castro, E Fidalgo, and Deisy Chaves. 2019. Boosting child abuse victim identification in Forensic Tools with hashing techniques. *V Jornadas Nacionales de Investigación en Ciberseguridad* 1 (2019), 344–345.

[6] Elie Bursztein, Einat Clarke, Michelle DeLaune, David M. Elifff, Nick Hsu, Lindsey Olson, John Shehan, Madhukar Thakur, Kurt Thomas, and Travis Bright. 2019. Rethinking the Detection of Child Sexual Abuse Imagery on the Internet. In *The World Wide Web Conference* (San Francisco, CA, USA) *(WWW '19)*. Association for Computing Machinery, New York, NY, USA, 2601–2607. https://doi.org/10.1145/3308558.3313482

[7] Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. 2021. Extracting Training Data from Large Language Models. In *30th USENIX Security Symposium (USENIX Security 21)*. USENIX Association, 2633–2650. https://www.usenix.org/conference/usenixsecurity21/presentation/carlini-extracting

[8] Modesto Castrillón-Santana, Javier Lorenzo-Navarro, Carlos M Travieso-González, David Freire-Obregón, and Jesús B Alonso-Hernández. 2018. Evaluation of local descriptors and CNNs for non-adult detection in visual content. *Pattern Recognition Letters* 113 (2018), 10–18.

[9] Deisy Chaves, Eduardo Fidalgo, Enrique Alegre, Francisco Jánez-Martino, and Rubel Biswas. 2020. Improving Age Estimation in Minors and Young Adults with Occluded Faces to Fight Against Child Sexual Exploitation.. In *VISIGRAPP (5: VISAPP)*. 721–729.

[10] Haoxing Chen, Huaxiong Li, Yaohui Li, and Chunlin Chen. 2023. Sparse Spatial Transformers for Few-Shot Learning. *Science China Information Sciences* 66, 11 (Nov. 2023), 210102. https://doi.org/10.1007/s11432-022-3700-8

[11] Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Frank Wang, and Jia-Bin Huang. 2019. A Closer Look at Few-shot Classification. In *International Conference on Learning Representations*. https://openreview.net/forum?id=HkxLXnAcFQ

[12] Mateus de Castro Polastro and Pedro Monteiro da Silva Eleuterio. 2010. Nudetective: A forensic tool to help combat child pornography through automatic nudity detection. In *Workshops on Database and Expert Systems Applications*. 349–353.

[13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *Conference on Computer Vision and Pattern Recognition*. 248–255.

[14] Carl Doersch, Ankush Gupta, and Andrew Zisserman. 2020. CrossTransformers: spatially-aware few-shot transfer. In *Advances in Neural Information Processing Systems*, Vol. 33. Curran Associates, Inc., 21981–21993.

[15] Pedro Eleuterio, Mateus Polastro, and Brazilian Federal Police. 2012. An Adaptive Sampling Strategy for Automatic Detection of Child Pornographic Videos. In *International Conference on Forensic Computer Science*.

[16] Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*. 1126–1135.

[17] Abhishek Gangwar, Víctor González-Castro, Enrique Alegre, and Eduardo Fidalgo. 2021. AttM-CNN: Attention and metric learning based CNN for pornography, age and Child Sexual Abuse (CSA) Detection in images. *Neurocomputing* 445 (2021), 81–104.

[18] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. MIT press.

[19] Fusheng Hao, Fengxiang He, Liu Liu, Fuxiang Wu, Dacheng Tao, Jun Cheng, et al. 2023. Class-Aware Patch Embedding Adaptation for Few-Shot Image Classification. In *Internation Conference on Computer Vision*. 18905–18915.

[20] Yangji He, Weihan Liang, Dongyang Zhao, Hong-Yu Zhou, Weifeng Ge, Yizhou Yu, et al. 2022. Attribute surrogates learning and spectral tokens pooling in transformers for few-shot learning. In *Conference on Computer Vision and Pattern Recognition*. 9119–9129.

[21] Shell Xu Hu, Da Li, Jan Stühmer, Minyoung Kim, and Timothy M. Hospedales. 2022. Pushing the Limits of Simple Pipelines for Few-Shot Learning: External Data and Fine-Tuning Make a Difference. In *Conference on Computer Vision and Pattern Recognition*.

[22] Microsoft Inc. 2020. PhotoDNA Cloud Services. https://www.microsoft.com/en-us/PhotoDNA.

[23] Juliane A Kloess, Jessica Woodhams, and Catherine E Hamilton-Giachritsis. 2021. The Challenges of Identifying and Classifying Child Sexual Exploitation Material: Moving towards a More Ecologically Valid Pilot Study with Digital Forensics Analysts. *Child Abuse & Neglect* 118 (2021), 105166.

[24] Juliane A Kloess, Jessica Woodhams, Helen Whittle, Tim Grant, and Catherine E Hamilton-Giachritsis. 2019. The Challenges of Identifying and Classifying Child Sexual Abuse Material. *Sexual Abuse* 31, 2 (2019), 173–196.

[25] Bernard Koch, Emily Denton, Alex Hanna, and Jacob Foster. 2021. Reduced, Reused and Recycled: The Life of a Dataset in Machine Learning Research. In *35th Conference on Neural Information Processing Systems (NeurIPS 2021)*.

[26] Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov. 2015. Siamese neural networks for one-shot image recognition. In *ICML Deep Learning Workshop*.

[27] Camila Laranjeira da Silva, João Macedo, Sandra Avila, and Jefersson dos Santos. 2022. Seeing without looking: Analysis pipeline for child sexual abuse datasets. In *ACM Conference on Fairness, Accountability, and Transparency*. 2189–2205.

[28] Hee-Eun Lee, Tatiana Ermakova, Vasilis Ververis, and Benjamin Fabian. 2020. Detecting child sexual abuse material: A comprehensive survey. *Forensic Science International: Digital Investigation* 34 (2020), 301022.

[29] Joao Macedo, Filipe Costa, and Jefersson A. dos Santos. 2018. A Benchmark Methodology for Child Pornography Detection. In *Conference on Graphics, Patterns and Images (SIBGRAPI)*. 455–462.

[30] Jay Mahadeokar and Gerry Pesavento. 2016. Open Sourcing a Deep Learning Solution for Detecting NSFW Images. *Retrieved August* 24 (2016), 2018.

[31] Mark Mazumder, Colby Banbury, Xiaozhe Yao, Bojan Karlaš, William Gaviria Rojas, Sudnya Diamos, Greg Diamos, Lynn He, Douwe Kiela, David Jurado, David Kanter, et al. 2022. DataPerf: Benchmarks for Data-Centric AI Development. *arXiv preprint arXiv:2207.10062* (2022).

[32] Nikhil Mishra, Mostafa Rohaninejad, Xi Chen, and Pieter Abbeel. 2018. A Simple Neural Attentive Meta-Learner. In *International Conference on Learning Representations*.

[33] Boris Oreshkin, Pau Rodríguez López, and Alexandre Lacoste. 2018. TADAM: Task dependent adaptive metric for improved few-shot learning. In *Advances in Neural Information Processing Systems*, Vol. 31.

[34] Tanvi A Patel, Vipul K Dabhi, and Harshadkumar B Prajapati. 2020. Survey on Scene Classification techniques. In *International Conference on Advanced Computing and Communication Systems*. 452–458.

[35] Claudia Peersman, Christian Schulze, Awais Rashid, Margaret Brennan, and Carl Fischer. 2016. iCOP: Live forensics to reveal previously unknown criminal media on P2P networks. *Digital Investigation* 18 (2016), 50–64.

[36] Mateus Polastro and Pedro Eleuterio. 2010. Nudetective: A Forensic Tool to Help Combat Child Pornography through Automatic Nudity Detection. In *Workshops on Database and Expert Systems Applications*. 349–353.

[37] Mateus Polastro and Pedro Eleuterio. 2012. A Statistical Approach for Identifying Videos of Child Pornography at Crime Scenes. In *International Conference on Availability, Reliability and Security*. 604–612.

[38] Neoklis Polyzotis and Matei Zaharia. 2021. What Can Data-Centric AI Learn from Data and ML Engineering? *arXiv preprint arXiv:2112.06439* (2021).

[39] Jiayan Qiu, Yiding Yang, Xinchao Wang, and Dacheng Tao. 2021. Scene Essence. In *Conference on Computer Vision and Pattern Recognition*. 8322–8333.

[40] Ariadna Quattoni and Antonio Torralba. 2009. Recognizing indoor scenes. In *Conference on Computer Vision and Pattern Recognition*. 413–420.

[41] Sachin Ravi and Hugo Larochelle. 2016. Optimization as a model for few-shot learning. In *International conference on learning representations*.

[42] Andrei A. Rusu, Dushyant Rao, Jakub Sygnowski, Oriol Vinyals, Razvan Pascanu, Simon Osindero, and Raia Hadsell. 2019. Meta-Learning with Latent Embedding Optimization. In *International Conference on Learning Representations*.

[43] Napa Sae-Bae, Xiaoxi Sun, Husrev T Sencar, and Nasir D Memon. 2014. Towards Automatic Detection of Child Pornography. In *IEEE International Conference on Image Processing*. 5332–5336.

[44] Victor Garcia Satorras and Joan Bruna Estrach. 2018. Few-Shot Learning with Graph Neural Networks. In *International Conference on Learning Representations*.

[45] Christoph Schröer, Felix Kruse, and Jorge Marx Gómez. 2021. A Systematic Literature Review on Applying CRISP-DM Process Model. *Procedia Computer Science* 181 (2021), 526–534.

[46] Christian Schulze, Dominik Henter, Damian Borth, and Andreas Dengel. 2014. Automatic Detection of CSA Media by Multi-Modal Feature Fusion for Law Enforcement Support. In *International Conference on Multimedia Retrieval*.

[47] Hongje Seong, Junhyuk Hyun, and Euntai Kim. 2020. Fosnet: An end-to-end trainable deep neural network for scene recognition. *IEEE Access* 8 (2020), 82066–82077.

[48] Josef Sivic and Andrew Zisserman. 2003. Video Google: A text retrieval approach to object matching in videos. In *International Conference on Computer Vision*. 1470–1477.

[49] Jake Snell, Kevin Swersky, and Richard Zemel. 2017. Prototypical Networks for Few-shot Learning. In *Advances in Neural Information Processing Systems*, Vol. 30.

[50] Siddharth Srivastava and Gaurav Sharma. 2024. OmniVec2 - A Novel Transformer Based Network for Large Scale Multimodal and Multitask Learning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 27402–27414. https://doi.org/10.1109/CVPR52733.2024.02588

[51] Clare Strickland, Juliane A Kloess, and Michael Larkin. 2023. An exploration of the personal experiences of digital forensics analysts who work with child sexual abuse material on a daily basis:"you cannot unsee the darker side of life". *Frontiers in Psychology* 14 (2023), 1142106.

[52] Lukas Struppek, Dominik Hintersdorf, Daniel Neider, and Kristian Kersting. 2022. Learning to Break Deep Perceptual Hashing: The Use Case NeuralHash. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 58–69. https://doi.org/10.1145/3531146.3533073

[53] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. 2018. Learning to compare: Relation network for few-shot learning. In *Conference on Computer Vision and Pattern Recognition*. 1199–1208.

[54] André Tabone, Kenneth Camilleri, Alexandra Bonnici, Stefania Cristina, Reuben Farrugia, and Mark Borg. 2021. Pornographic content classification using deep-learning. In *ACM Symposium on Document Engineering*. 1–10.

[55] U. K. Her Majesty's Government. 2021. Tackling Child Sexual Abuse Strategy.

[56] Pedro H. V. Valois, João Macedo, Leo S. F. Ribeiro, Jefersson A. dos Santos, and Sandra Avila. 2025. Leveraging Self-Supervised Learning for Scene Classification in Child Sexual Abuse Imagery. *Forensic Science International: Digital Investigation* (2025).

[57] Laurens vd Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research* 9 (2008), 2579–2605.

[58] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. 2016. Matching networks for one shot learning. *Advances in Neural Information Processing Systems* 29 (2016), 3630–3638.

[59] Paulo Vitorino, Sandra Avila, Mauricio Perez, and Anderson Rocha. 2018. Leveraging Deep Neural Networks to Fight Child Pornography in the Age of Social Media. *Journal of Visual Communication and Image Representation* 50 (2018), 303–313.

[60] Paulo Vitorino, Sandra Avila, and Anderson Rocha. 2016. A Two-tier Image Representation Approach to Detecting Child Pornography. In *XII Workshop de Visão Computational*. 129–134.

[61] Bryce Westlake, Martin Bouchard, and Richard Frank. 2012. Comparing Methods for Detecting Child Exploitation Content Online. In *2012 European Intelligence and Security Informatics Conference*. 156–163. https://doi.org/10.1109/EISIC.2012.25

[62] Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. 2010. SUN database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 3485–3492. https://doi.org/10.1109/CVPR.2010.5539970

[63] Han-Jia Ye, Hexiang Hu, De-Chuan Zhan, and Fei Sha. 2020. Few-shot learning via embedding adaptation with set-to-set functions. In *Conference on Computer Vision and Pattern Recognition*. 8808–8817.

[64] Andrew Young, Stuart Campo, and Stefaan Verhulst. 2019. Responsible Data for Children: Synthesis Report. (2019).

[65] Daochen Zha, Zaid Pervaiz Bhat, Kwei-Herng Lai, Fan Yang, Zhimeng Jiang, Shaochen Zhong, and Xia Hu. 2023. Data-Centric Artificial Intelligence: A Survey. *arXiv preprint arXiv:2303.10158* (2023).

[66] Wanrong Zhang, Olga Ohrimenko, and Rachel Cummings. 2021. Attribute Privacy: Framework and Mechanisms. *arXiv preprint arXiv:2009.04013* (2021).

[67] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2017. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40, 6 (2017), 1452–1464.