A Hierarchical Theme Recognition Model for Sandplay Therapy

Xiaokun Feng[1,2], Shiyu Hu[1,2], Xiaotang Chen[1,2,3], and Kaiqi Huang[1,2,3](✉)

[1] School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
[2] Institute of Automation, Chinese Academy of Sciences, Beijing, China
[3] Center for Excellence in Brain Science and Intelligence Technology, Chinese Academy of Sciences, Beijing, China
{fengxiaokun2022,hushiyu2019}@ia.ac.cn, {xtchen,kaiqi.huang}@nlpr.ia.ac.cn

Abstract. Sandplay therapy functions as a pivotal tool for psychological projection, where testers construct a scene to mirror their inner world while psychoanalysts scrutinize the testers' psychological state. In this process, recognizing the theme (*i.e.*, identifying the content and emotional tone) of a sandplay image is a vital step in facilitating higher-level analysis. Unlike traditional visual recognition that focuses solely on the basic information (*e.g.*, category, location, shape, *etc.*), sandplay theme recognition needs to consider the overall content of the image, then relies on a hierarchical knowledge structure to complete the reasoning process. Nevertheless, the research of sandplay theme recognition is hindered by following challenges: (1) Gathering high-quality and enough sandplay images paired with expert analyses to form a scientific dataset is challenging, due to this task relies on a specialized sandplay environment. (2) Theme is a comprehensive and high-level information, making it difficult to adopt existing works directly in this task. In summary, we have tackled the above challenges from the following aspects: (1) Based on carefully analysis of the challenges (*e.g.*, small-scale dataset and complex information) , we present the **HIST** (**HI**erarchical **S**andplay **T**heme recognition) model that incorporates external knowledge to emulate the psychoanalysts' reasoning process. (2) Taking the split theme (a representative and evenly distributed theme) as an example, we proposed a high-quality dataset called **SP**^{2} (**S**and**P**lay **SP**lit) to evaluate our proposed method. Experimental results demonstrate the superior performance of our algorithm compared to other baselines, and ablation experiments confirm the importance of incorporating external knowledge. We anticipate this work will contribute to the research in sandplay theme recognition. The relevant datasets and codes will be released continuously.

Keywords: Sandplay therapy· Sandplay theme recognition· Visual recognition.

1 Introduction

Sandplay therapy functions as a pivotal tool for psychological projection, where individuals construct a scene to mirror their inner world while psychoanalysts scrutinize the individual's psychological state[1, 2]. In this process, recognizing the *theme* (*i.e.*, identifying the content and emotional tone) of a sandplay image is an important step in facilitating higher-level analysis. As shown in Fig. 1, generating a sandplay image and recognizing its theme always include several steps – a client should first construct a sandplay work based on inner thoughts, and the psychoanalyst then analyzes the theme by synthesizing basic semantic information and high-level semantic information (*e.g.*, psychological knowledge).

Although identifying the theme of a sandplay image is an important and valuable issue, the current visual recognition algorithms are unable to adapt well to this task.

Early visual recognition research studies mainly focus on identifying basic semantic information in images, such as category[3], location[4], shape[5], *etc.*Recently, several researchers have shifted their focus to recognizing high-level visual information like emotions[6, 7]. The emotion recognition task enables computer systems to process and comprehend emotional information conveyed by humans, and facilitates natural human-computer interactions[8]. However, unlike existing visual recognition tasks that focuses solely on basic semantic information or emotion, the sandplay theme recognition task is proposed for a more challenging and specialized application scenario. It aims to identify the content and emotional tone expressed in a sandplay image, which can support higher-level tasks such as understanding the psychological state of the creator like psychologists. Therefore, a well-designed sandplay theme recognition method should execute the above analyzing process like experts. It should first consider the overall content of the image, then rely on a hierarchical structure to utilize various knowledge (*e.g.*, basic semantic information, emotion information, and even external knowledge from psychologists), and finally complete the reasoning process.

Nevertheless, two challenges hinder the research of designing a suitable model for the sandplay theme recognition task:

(1) **Obtaining enough high-quality data samples is difficult.** ❶ Unlike general scene visual recognition tasks that can easily collect numerous data samples from the Internet[6, 9], the sandplay theme recognition task relies on a specialized sandplay environment. ❷ Besides, existing research commonly annotates data samples through crowd-sourcing platforms like Amazon Mechanical Turk, while the annotation of sandplay image should be generated by psychology experts. Thus, gathering high-quality and enough sandplay images paired with expert analyses to form a scientific dataset is challenging.

(2) **A sandplay image's theme information is complex, intensifying the difficulty of recognition.** Existing works have focused on recognizing basic semantics or emotions, but they lack a hierarchical framework for comprehensive understanding, making them difficult to adopt directly in the sandplay theme recognition task.

In this paper, we address the difficulties mentioned above and carry out our work, aiming to accomplish the challenging sandplay theme recognition task:

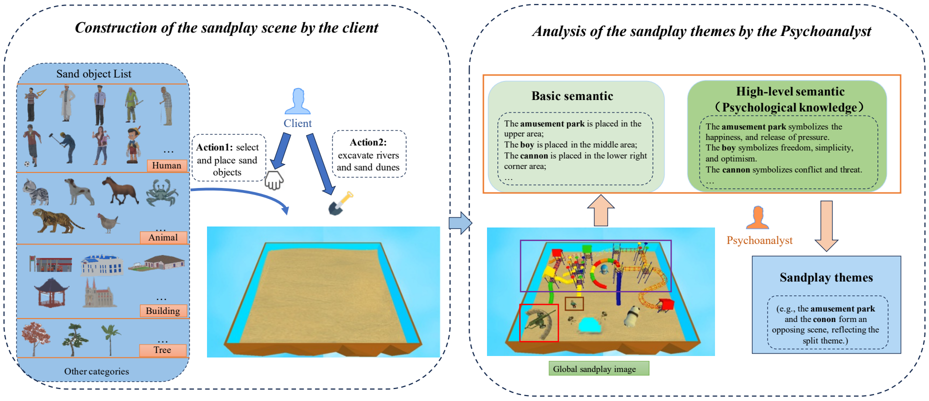

Construction of the sandplay scene by the client
Sand object List
Human
Animal
Building
Tree
Other categories
Client
Action1: select and place sand objects
Action2: excavate rivers and sand dunes
Analysis of the sandplay themes by the Psychoanalyst
Basic semantic
The amusement park is placed in the upper area;
The boy is placed in the middle area;
The cannon is placed in the lower right corner area;
…
High-level semantic (Psychological knowledge)
The amusement park symbolizes the happiness, and release of pressure.
The boy symbolizes freedom, simplicity, and optimism.
The cannon symbolizes conflict and threat.
…
Psychoanalyst
Sandplay themes
(e.g., the amusement park and the conon form an opposing scene, reflecting the split theme.)
Global sandplay image

Fig. 1. Framework of sandplay themes recognition process on 3D electronic sandplay platform. Firstly, the client constructs a sandplay scene based on inner thoughts. Then, the psychoanalyst analyzes the sandplay themes by synthesizing basic semantic information and high-level semantic information (*i.e.*, psychological knowledge).

(1) **Design a hierarchical theme recognition model.** Based on carefully analyzing the challenges of the sandplay theme recognition task (*e.g.*, small-scale datasets and complex information), we present a recognition model named **HIST** (**HI**erarchical **S**andplay **T**heme recognition) that incorporates external knowledge. In light of the analysis process of psychologists, our proposed model comprises two fundamental steps. Firstly, we focus on perceiving the basic semantics of the image, which specifically refers to extracting the categorical information of the objects present in the image. Secondly, based on the perceived categorical information, we incorporate the external knowledge by indexing the corresponding high-level attribute information. Then, our model recognizes the sandplay theme by leveraging the above information. Specifically, to evaluate the effects of small-scale datasets caused by the characteristics of sandplay scenario, we also employ several training strategies to enhance the learning capacity of the model.

(2) **Propose a high-quality dataset to train and evaluate the proposed HIST model.** ❶ We take psychological sandplay as our experimental environment, and collect data samples from a 3D electronic platform (Fig. 1). Specifically, we invite a substantial number of testers to participate in sandplay test and collect data samples following the sandplay analysis process, which ensure the professionalism and scientificity of the data samples. According to statistics, each sandplay sample contains an average of 15 sand objects (selected from 494 sand categories), which reflects the diversity of sandplay samples. After screening and sorting, we collect 5,000 samples. ❷ We engage professional psychoanalysts to annotate the theme of the sandplay sample. Without affecting the integrity of the sandplay analysis process, we select one of the representative themes, namely the split theme, as the object of our psychological recognition (Fig. 1). The definition of split theme refers to a state of isolation and separa-

tion between the various parts of the whole sandplay scene (the split samples are shown in Fig. 3). It reflects the inner integration of the tester and is related to many emotional and personality issues. Finally, we construct a dataset with split theme annotations, denoted as $\mathbf{SP}^2$ (**S**and**P**lay **SP**lit). Statistical analysis reveals that the acquisition of each sandplay sample demands an average of 10 minutes, which reflects the expense and time consumption of obtaining it.

In general, our contributions are as follows:

- We propose a hierarchical sandplay theme recognition model(*i.e.*, HIST) that incorporates external knowledge. By modeling the process of perceiving the basic semantics of images and incorporating corresponding high-level attribute knowledge, our model emulates the analytical ability of psychologists (Section 3).
- Based on the 3D electronic sandplay platform, we construct a dataset named SP^2. All the sandplay samples are carefully collected and annotated, aiming to propose a high-quality dataset for recognizing theme in the sandplay environment (Section 4). Besides, experimental results indicate the excellent performance of our model, and the ablation experiment demonstrates the importance of introducing external knowledge (Section 5).

We anticipate this work will contribute to the research in sandplay theme recognition. The relevant datasets and codes results will be released continuously.

2 Related work

The recognition of sandplay themes can be regarded as an image recognition task, wherein the sandplay image serves as the input data, and the sandplay theme represents the output. In this section, we firstly introduce existing image recognition tasks to highlight the characteristics of sandplay theme recognition task (Section 2.1). Then we introduce the relevant background knowledge in the field of psychological sandplay (Section 2.2).

2.1 Image recognition tasks

Recognizing semantic information from image data is a fundamental research problem in the field of computer vision. Various research tasks aim to model different aspects of human abilities by focusing on different semantic information. For instance, basic image classification[3] tasks aim to emulate human visual perception abilities, while image emotion recognition tasks[9, 10] aim to capture human emotional cognitive abilities. Although these tasks center around different forms of semantic information, their research[9–11] typically utilizes data collected from the Internet (*e.g.*, TV shows, social networks,*etc.*) and relies on crowd-sourcing platforms (like Amazon Mechanical Turk) for semantic label annotations. Consequently, the modeling object of these tasks is primarily centered around ordinary people.

In the context of sandplay theme recognition, the data samples are sampled from the specialized sandplay environment, and the annotation of sandplay themes requires the expertise of psychologists who employ hierarchical cognitive analysis. Through the sandplay theme recognition task, we can explore modeling the hierarchical analytical capabilities of psychoanalysts.

2.2 Projection test and sandplay therapy

Projective test[12] is a well-known and widely used psychoanalysis test in which client offer responses to ambiguous stimuli (*e.g.*, words, images, *etc.*), so as to reveal the hidden conflicts or emotions that client project onto the stimuli. The deeply held feelings and motivations are often not verbalized or even be aware by client, while these subjective psychological semantics can be detected and analyzed through projecting onto the stimuli.

The visual stimulus is the main carrier of projection test, such as Rorschach inkblot image[13], house-tree-person painting and sandplay image[1]. The image carrier of inkblot test is only 10 inkblot pictures, and the painting image in house-tree-person test can only contain 3 elements (*i.e.*, house, tree, person). On the other hand, sandplay therapy usually contains hundreds of miniature figures, and they can be combined and placed in any way. The high degree of interactivity ensures a diverse range of sandplay samples, thus making it an ideal research scene for our work.

However, despite the potential of sandplay to generate numerous data samples, a large-scale sandplay dataset is currently unavailable. For the purpose of using existing data-driven deep neural network models, we invite a substantial number of testers to collect samples. As a result, we build a sandplay dataset consisting of 5,000 diverse samples.

Furthermore, according to the sandplay theme theory developed by Mitchell and Friedman[2], around 20 themes have been identified to encode psychological states (see A.1). This work takes the split theme as the research label, mainly considering the following two reasons: (1)the positive and negative sample size of the split is relatively balanced (see A.3), making it advantageous for model training; (2) split theme exhibits connections with various psychological symptoms, including some common emotional issues (*e.g.*, depression, negative study experiences) and personality problems (*e.g.*, compulsion, paranoia, marginalization, aggression), which reflects the practical application value of the sandplay. For annotation and research on other themes, we leave them for future work.

3 Method

Based on the preceding analysis, sandplay theme recognition task encounters two distinct challenges: small-scale datasets and the intricate processing of complex information (*i.e.*, sandplay themes). In this case, we propose the HIST model, and the framework is shown in Fig. 2. Corresponding to the hierarchical analysis process of psychologists, we model three hierarchical processes: the perception

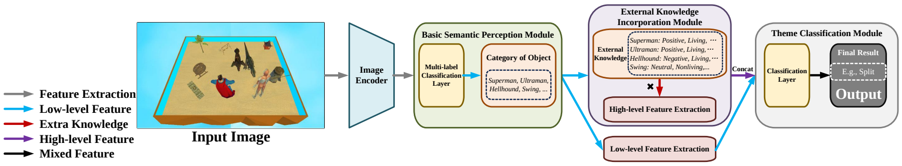

Feature Extraction
Low-level Feature
Extra Knowledge
High-level Feature
Mixed Feature
Input Image
Image Encoder
Basic Semantic Perception Module
Multi-label Classification Layer
Category of Object
Superman, Ultraman, Hellhound, Swing, ...
External Knowledge Incorporation Module
External Knowledge
Superman: Positive, Living, ⋯
Ultraman: Positive, Living, ⋯
Hellhound: Negative, Living, ⋯
Swing: Neutral, Nonliving,...
High-level Feature Extraction
Low-level Feature Extraction
Concat
Theme Classification Module
Classification Layer
Final Result
E.g., Split
Output

Fig. 2. Overview of the HIST model. Correspondingly to the hierarchical analysis process of psychoanalysts, this model consists of basic semantic perception module, external knowledge incorporation module and theme classification module.

of basic semantics (Section 3.1), the incorporation of external knowledge (Section 3.2), and the final theme classification (Section 3.3).

3.1 Basic semantic perception module

Perceiving basic semantic information (such as sand object's category, bounding box, *etc.*) from images is a crucial step for recognizing the theme information. In this process, our model focuses on the category information of the sand objects within the image, and we employ existing multi-label classification[14] techniques to handle this task.

Firstly, we adopt the common ResNet[15] backbone as the image encoder. Given the image $I \in \mathbf{R}^{3\times960\times540}$, we feed it into the image encoder and obtain the corresponding feature vector $F_i \in \mathbf{R}^{1\times L}$, where the length of the feature vector $L = 1,000$. To alleviate the limitations imposed by the small-scale sandplay dataset, we utilize image augmentation techniques to expand the dataset's size. Additionally, we leverage the pre-trained model to extract the initial image features (see Section 5.1).

Then, F_i is fed into the fully connected layers (accompanied by the *Relu* activation function) to obtain the multi-label classification feature vector $F_{cl} \in \mathbf{R}^{1\times N_{cl}}$. Here, N_{cl} represents the summary of sand objects' categories, and $N_{cl} = 494$ in our sandplay environment.

Finally, we perform multi-label classification based on F_{cl}. We apply the *Sigmoid* activation function on F_{cl} to obtain F_{clp}. Each element f^i_{clp} in F_{clp} represents the probability of the model classifying the i_{th} sand object. We construct the classification error L_{cl} based on the true labels $Y_c = \{y^1, y^2, ..., y^{N_{cl}}\}$.

$$L_{cl} = -\frac{1}{N_{cl}}\sum_{i=1}^{N_{cl}}[p^i y^i log f^i_{clp} + (1-y^i)log(1-f^i_{clp})], p^i = \frac{neg_i}{pos_i} \tag{1}$$

where p^i means the weight of the i_{th} object. In order to reduce the long tail distribution problem[16] between different objects, we define p^i as the ratio between i_{th} object's negative sample number neg_i and positive sample number pos_i.

3.2 External knowledge incorporation module

For the sandplay theme recognition task, we consider the psychological attributes of each sand object as external knowledge . According to sandplay therapy, the number of these psychological attributes (denoted as L_h) is fixed, which means that the psychological attributes of each object can be represented by a one-dimensional vector $k_h \in \mathbf{R}^{1\times L_h}$. In scenarios with N_{cl} objects, we can utilize a feature vector $K_h \in \mathbf{R}^{N_{cl}\times L_h}$ to encode the external knowledge. We set $L_h = 7$ (see Section 4.2) in subsequent experiments.

We employ object category indexing to incorporate the external knowledge K_h into our model. By leveraging F_{clp}, we utilize the probability of each object's existence f^i_{clp}, to weight the corresponding high-level semantic vector k^i_h. This process results in the construction of the vector $K_w \in \mathbf{R}^{N_{cl}\times L_h}$.

$$k^i_w = f^i_{clp} \times k^i_h \tag{2}$$

Then, in order to comprehensively analyze the weighted high-level semantic information, we employ self-attention operations to achieve high-level feature extraction. Specifically, we use a transformer encoder[17] module to process and obtain $F_{we} \in \mathbf{R}^{1\times N_{cl}}$, consider it as the integrated external knowledge feature information.

3.3 Theme classification module

Building upon F_{we} and F_i, we can integrate the basic semantic information and external knowledge information of objects. Firstly, we employ the fully connected layers (accompanied by the *Relu* activation function) to further process F_i, obtaining F'_i. Next, we concatenate the F'_i with F_{we}, resulting in the final feature vector, denoted as F_u.

$$F_u = concat\{F_{we} + Relu(Fc\,(F_I))\} \tag{3}$$

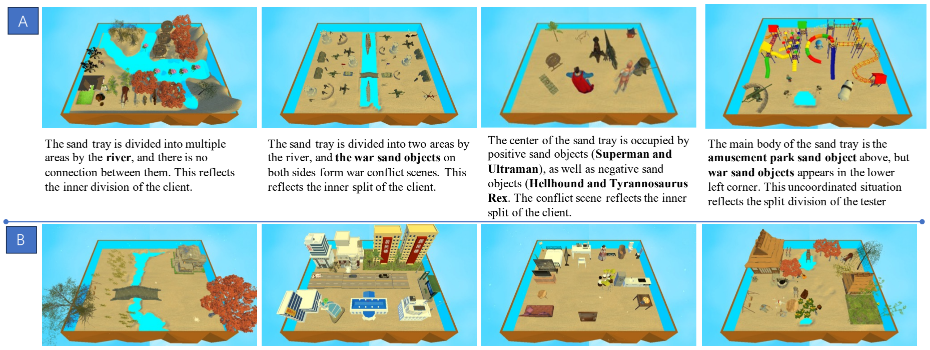
Finally, we feed F_u into the fully connected layers (accompanied by the *Relu* activation function) to obtain the theme category vector $F_t \in \mathbf{R}^{1\times 2}$ (*i.e.*, recognition for a specific sandplay theme can be regarded as a binary classification task). Based on groundtruth labels Y_t, we construct the classification error L_t using the CrossEntropy loss function.

Based on L_{cl} and L_t. we compute the final error L_a by the relative weight coefficient w_{cl}. L_a is used for backpropagation:

$$L_a = L_t + w_{cl} \times L_{cl} \tag{4}$$

4 SP^2 dataset

Considering the challenge of gathering high-quality sandplay images paired with theme annotation from psychoanalysts, we take the representative split theme as the example, and construct the SP^2 dataset. In this section, we will firstly present the process of dataset construction (Section 4.1). Then, we present the psychological attribute information for each sand object (Section 4.2), which can be considered as external knowledge employed by psychoanalysts to achieve hierarchical analysis.



The sand tray is divided into multiple areas by the **river**, and there is no connection between them. This reflects the inner division of the client.

The sand tray is divided into two areas by the river, and **the war sand objects** on both sides form war conflict scenes. This reflects the inner split of the client.

The center of the sand tray is occupied by positive sand objects (**Superman and Ultraman**), as well as negative sand objects (**Hellhound and Tyrannosaurus Rex**. The conflict scene reflects the inner split of the client.

The main body of the sand tray is the **amusement park sand object** above, but **war sand objects** appears in the lower left corner. This uncoordinated situation reflects the split division of the tester

Fig. 3. Examples from the SP^2 dataset. Panel A represents sandplay samples with the split theme, accompanied by explanatory descriptions for the judgment (bold font indicates the concerned sand objects during judgment); Panel B represents sandplay samples without the split theme.

4.1 Dataset construction

Sampling of sandplay samples. We invited a substantial number of testers to participate in the sandplay test and obtained corresponding sandplay samples. In order to ensure the authenticity, we only collected a sandplay sample from each tester. In the end, we sorted out 5,000 sandplay samples.

Labeling of sandplay samples. For each sandplay sample, we provide the basic semantic label (*i.e.*, the name and bounding box of each sand object) and the high-level sandplay theme label (*i.e.*, split label). For the former, with the help of the 3D electronic sandplay platform, we can directly obtain it from the terminal; For the latter, we engaged psychoanalysts to discern split theme through binary classification in order to ensure the objectivity of the split label. Initially, we selected 200 samples and each sample was labeled by five psychoanalysts. Through discussions and deliberations, we established consistent criteria for the recognition of split theme. Then, in order to improve annotation efficiency, each sample in the remaining data is labeled by one psychoanalyst.

Format of sandplay samples. A sandplay sample consists of a global image $I \in \mathbf{R}^{3\times960\times540}$, a set of basic semantic labels, and a split label.

In addition, we conducted statistics on the number of positive and negative samples of split label (positive samples mean the sandplay has the split theme). Among SP^2, there are 2,303 positive samples and 2,697 negative samples, which indicates that the number of positive and negative samples is relatively balanced.

4.2 Psychological attributes of sand objects

During the analysis of sandplay images (as depicted in Fig 1), psychoanalysts integrate the psychological attributes of each sand object to make their final assessments. For instance, the boy represents a positive object, whereas cannon represents a negative object. These psychological attributes associated with the sand objects can be considered as the external knowledge. Drawing from the principles of sandplay analysis, we have organized 7 key psychological attributes for each sand object, including polarity, life attribute, spiritual/material attribute, static/dynamic attribute, *etc.*For more details, please refer to A.2.

5 Experiments

Table 1. Comparison with the state-of-the-art models on SP^2 dataset. The best and second-best results are marked in **bold** and underline.

Model	Acc	F1
AlexNet	0.717	0.676
VGG	0.727	0.707
ResNet	0.742	0.727
ViT	0.731	0.712
MldrNet	0.723	0.710
zhang 2020	0.729	0.707
BiGRU	0.734	0.712
PadNet	0.745	0.722
HIST (Ours)	**0.790**	**0.765**

Table 2. Results of ablation experiments. The best and second-best results are marked in **bold** and underline.

Variants	Acc	F1
ViT-based	0.761	0.745
Without-L_{cl}	0.764	0.750
Without-K_h	0.747	0.723
HIST (Ours)	**0.790**	**0.765**

In this section, we conduct experiments on the SP^2 dataset for evaluation.

5.1 Experimental setup

Implementation details. We employ ResNet-50[15] as the visual encoder, and we initialize it using the model weights pretrained on ImageNet[11]. We firstly reshape the input sandplay image $I \in \mathbf{R}^{3\times960\times540}$ into $I' \in \mathbf{R}^{3\times224\times224}$ to match the input of the model. We use basic image rotation and flipping operations for data augmentation. The batch size is set to 64, and the learning rate is set to $1e-3$. We use SGD optimizer, and the relative weight coefficient w_{cl} is set to 2.

Datasets and metrics. The SP^2 dataset is divided into training, testing and validation sets in the 8:1:1 ratio (see A.3 for detailed information). We evaluate the performance of different models by Accuracy and F1 metric.

5.2 Comparison with the state-of-the-art

Following the two recent authoritative works[6, 18], we compare our model with representative image emotion recognition state-of-the-art methods which are publicly available and adaptable to sandplay images, including BiGRU[19], MldrNet[20], Zhang et al.[18] and PadNet[21]. Additionally, we compare with some classic image classification models, including AlexNet[22], VGG[23], ResNet[15], and ViT[17].

During training our proposed model, two types of supervised information(sand object category and split label) are used. So for the sake of fairness, we provide these two types of supervised information in a multi-task format when training these baseline models. Specifically, a sand object classification head is addded in the final output layer of the model, and the model is trained using both object classification loss and split classification loss.

Based on the results presented in Table 1, it is evident that the performance of our proposed model surpasses that of other models. The essential distinction between our model and other models lies in the incorporation of external knowledge, highlighting the necessity of external knowledge for sandplay theme recognition.

5.3 Ablation study

Effect of visual backbone. The visual backbone in our model is used to encode image. In addition to ResNet[15], ViT[17] serves as another popular choice for a backbone network. Hence, we conduct experiments to evaluate the performance of our model using the ViT backbone network (ViT-based).

Experimental results shown in Table 2 indicate that the ResNet based network outperforms ViT. (which is consistent with the experimental result in Table 1). This discrepancy may be attributed to the dataset size. Because prior studies[24] have shown that ViT has the advantageous performance in large scale datasets, and the dataset size of SP^2 is relatively small.

Effect of semantic information. Our proposed models aim to emulate the reasoning process of psychoanalysts by considering both basic semantic information and high-level semantic information. To evaluate the effects of different semantic inputs, we conduct experiments under two settings. Firstly, we evaluate the model's performance when the sand object category information is not provided, which means setting l_{cl} to 0 (Without-L_{cl}). Secondly, we evaluate the model's performance when external knowledge is not incorporated, which achieve that by masking K_h to 1 (Without-K_h).

Experimental results shown in Table 2 indicate that the lack of category information or external knowledge can reduce the performance of the model. Moreover, the lack of external knowledge results in more severe performance degradation. This result once again reflects the importance of incorporation external knowledge.

6 Conclusion and feature work

Based on the sandplay therapy, sandplay theme recognition task relies on sandplay images and the corresponding theme annotations provided by psychoanalysts. This task offers an opportunity to explore the modeling of psychoanalysts' hierarchical analysis capabilities. Inspired by the analysis process of psychoanalysts, we propose HIST model. To facilitate the sandplay theme recognition task, we construct SP^2 dataset, focusing on split theme. Our proposed model outperforms existing baseline models on the SP^2 dataset, and ablation experiments demonstrates the significance of incorporating external knowledge.

In this work, we leverage a sandplay-based research environment to highlight the significance of external knowledge in the assessment of sandplay theme which represents high-level psychological semantics. Moving forward, we intend to explore incorporating external knowledge into more general scenarios, such as emotion recognition tasks for common images. By incorporating external knowledge into these scenarios, we aim to enhance the machine's recognition capabilities and facilitate more natural human-computer interactions.

References

1. Christian Roesler. Sandplay therapy: An overview of theory, applications and evidence base. *The arts in Psychotherapy*, 64:84–94, 2019.
2. Rie Rogers Mitchell and Harriet S Friedman. *Sandplay: Past, present, and future.* Psychology Press, 1994.
3. Dengsheng Lu and Qihao Weng. A survey of image classification methods and techniques for improving classification performance. *International journal of Remote sensing*, 28(5):823–870, 2007.
4. Zhengxia Zou, Keyan Chen, Zhenwei Shi, Yuhong Guo, and Jieping Ye. Object detection in 20 years: A survey. *Proceedings of the IEEE*, 2023.
5. Yanming Guo, Yu Liu, Theodoros Georgiou, and Michael S Lew. A review of semantic segmentation using deep neural networks. *International journal of multimedia information retrieval*, 7:87–93, 2018.
6. Sicheng Zhao, Xingxu Yao, Jufeng Yang, Guoli Jia, Guiguang Ding, Tat-Seng Chua, Bjoern W Schuller, and Kurt Keutzer. Affective image content analysis: Two decades review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6729–6751, 2021.
7. Jianhua Tao and Tieniu Tan. Affective computing: A review. In *International Conference on Affective computing and intelligent interaction*, pages 981–995. Springer, 2005.

8. Rosalind W Picard. Building hal: Computers that sense, recognize, and respond to human emotion. In *Human Vision and Electronic Imaging VI*, volume 4299, pages 518–523. SPIE, 2001.
9. Quanzeng You, Jiebo Luo, Hailin Jin, and Jianchao Yang. Robust image sentiment analysis using progressively trained and domain transferred deep networks. In *Proceedings of the AAAI conference on Artificial Intelligence*, volume 29, 2015.
10. Brendan Jou, Tao Chen, Nikolaos Pappas, Miriam Redi, Mercan Topkara, and Shih-Fu Chang. Visual affect around the world: A large-scale multilingual visual sentiment ontology. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 159–168, 2015.
11. Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.
12. Hideo Otsuna and Kei Ito. Systematic analysis of the visual projection neurons of drosophila melanogaster. i. lobula-specific pathways. *Journal of Comparative Neurology*, 497(6):928–958, 2006.
13. Kenneth R Gamble. The holtzman inkblot technique. *Psychological Bulletin*, 77(3):172, 1972.
14. Grigorios Tsoumakas and Ioannis Katakis. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining (IJDWM)*, 3(3):1–13, 2007.
15. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
16. Yifan Zhang, Bingyi Kang, Bryan Hooi, Shuicheng Yan, and Jiashi Feng. Deep long-tailed learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.
17. Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.
18. Wei Zhang, Xuanyu He, and Weizhi Lu. Exploring discriminative representations for image emotion recognition with cnns. *IEEE Transactions on Multimedia*, 22(2):515–523, 2019.
19. Xinge Zhu, Liang Li, Weigang Zhang, Tianrong Rao, Min Xu, Qingming Huang, and Dong Xu. Dependency exploitation: A unified cnn-rnn approach for visual emotion recognition. In *IJCAI*, pages 3595–3601, 2017.
20. Tianrong Rao, Xiaoxu Li, and Min Xu. Learning multi-level deep representations for image emotion classification. *Neural processing letters*, 51:2043–2061, 2020.
21. Sicheng Zhao, Zizhou Jia, Hui Chen, Leida Li, Guiguang Ding, and Kurt Keutzer. Pdanet: Polarity-consistent deep attention network for fine-grained visual emotion regression. In *Proceedings of the 27th ACM international conference on multimedia*, pages 192–201, 2019.
22. Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.
23. Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
24. Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? *Advances in Neural Information Processing Systems*, 34:12116–12128, 2021.