

GameTileNet: A Semantic Dataset for Low-Resolution Game Art in Procedural Content Generation

Yi-Chun Chen¹, Arnav Jhala²

¹Yale University

²North Carolina State University

ychen74@alumni.ncsu.edu, yi-chun.chen@yale.edu, ahjhala@ncsu.edu

Abstract

GameTileNet is a dataset designed to provide semantic labels for low-resolution digital game art, advancing procedural content generation (PCG) and related AI research as a vision-language alignment task. Large Language Models (LLMs) and image-generative AI models have enabled indie developers to create visual assets, such as sprites, for game interactions. However, generating visuals that align with game narratives remains challenging due to inconsistent AI outputs, requiring manual adjustments by human artists. The diversity of visual representations in automatically generated game content is also limited because of the imbalance in distributions across styles for training data. GameTileNet addresses this by collecting artist-created game tiles from OpenGameArt.org under Creative Commons licenses and providing semantic annotations to support narrative-driven content generation. The dataset introduces a pipeline for object detection in low-resolution tile-based game art (e.g., 32x32 pixels) and annotates semantics, connectivity, and object classifications. GameTileNet is a valuable resource for improving PCG methods, supporting narrative-rich game content, and establishing a baseline for object detection in low-resolution, non-photorealistic images.

Introduction

While Large Language Models (LLMs) hold great potential for generating diverse game narratives by enriching objects and scenes, translating these narrative elements into visual representations remains a challenge. Combining LLMs with visual models like diffusion models has produced appealing results, but aligning narrative and visual content in games often requires significant manual modifications and artistic judgment. (Domsch 2013; Shapiro, Peña-Herborn, and Hancock 2012; Ang 2006; Koenitz 2024; Williams 2024)

Effective visual representations require integrating game materials with the narrative context to create immersive experiences for players. Therefore, visual elements like sprites and environments are essential for engagement and interaction. The game world requires a variety of visual assets to align with narrative and player choices. (Bulatovic Trygg, Skult, and Smed 2024; Tyndale and Ramsomair 2016)

One potential solution is the semantic mapping between narrative elements and game materials. This approach en-

Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

ures that visual assets align with narrative themes, allowing for the dynamic generation of visuals that reflect narrative changes. Semantic mapping helps maintain synergy between the game’s visual and narrative aspects by streamlining the creation of procedurally generated content.

In this paper, we introduce GameTileNet, a dataset of game materials with semantic labels designed to facilitate such semantic mapping. Focused on top-down, tile-based pixel art, GameTileNet sources materials from OpenGameArt.org under Creative Commons licenses, providing a diverse range of assets.

Pixel art tiles, typically low-resolution (e.g., 8x8, 16x16, or 32x32 pixels), use minimal memory and processing power, making them ideal for retro aesthetics. They rely on restricted color palettes to maintain visual consistency and their grid-based structure allows for modular game design, creating varied environments from a small set of assets.

The limited resolution and simplified details of pixel art make it difficult for computer vision models to distinguish subtle variations, classify objects, and accurately segment scenes. Furthermore, the niche nature of pixel art games means there is often a need for labeled data, hindering model training.

This study explores a pipeline for labeling tasks within this low-resolution dataset, focusing on pixel art game tiles. We demonstrate the use of GameTileNet through procedural content generation, using scene descriptions to create levels dynamically. The table 1 showcases one of the results. These examples show how semantic mapping can align visual representations with generated narratives

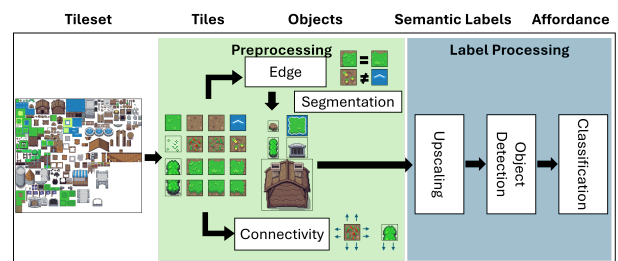


Figure 1: Object processing pipeline for pixel art game tiles. Each step supports accurate detection and labeling of low-resolution game elements.


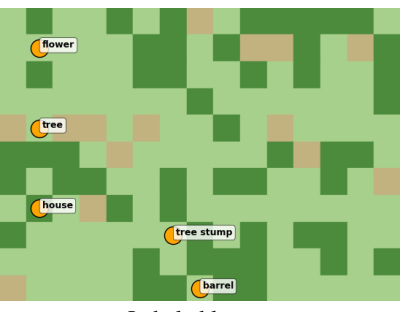
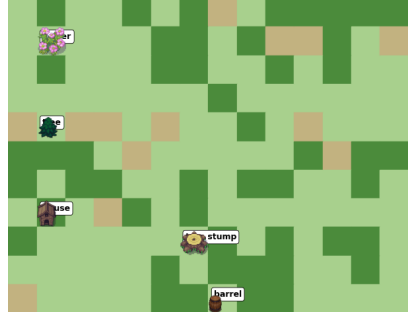
Place: Forest	Found Tiles	Showcase	
<p>House below Tree Tree to the right of Barrel Flower above Tree Tree stump to the left of Tree</p>	 <p>House Tree</p> <p>Barrel Flower</p> <p>Tree stump</p>	 <p>Labeled layout</p>	 <p>Mapped tiles</p>

Table 1: Create a game scene based on scene descriptions by aligning semantic labels in GameTileNet with narrative spatial relations. Left: labeled object layout. Right: mapped game tiles.

Related Work

Tile Based Game Design and Semantics

Tile-based game designs offer modular components that simplify game development by dividing environments into manageable units. This type of design supports efficient collision detection, level design, and resource optimization (van der Spuy 2010; Lafitte and Weiss 2009). Beyond the structural role, tiles carry semantic meanings and affordances essential for guiding player interactions. Bentley and Osborn’s Videogame Affordances Corpus illustrates how visual affordances signal potential actions, underscoring the need for tiles that reflect gameplay intent through semantic mapping (Bentley and Osborn 2019). Recent advancements, such as Jadhav and Guzdial’s tile embeddings, provide affordance-rich representations that simplify embedding interactive elements in procedural generation (Jadhav and Guzdial 2021). At the same time, context-based segmentation systems enhance the alignment of game visuals with narrative cues using AI (Gabriel and Clua 2023). Similar challenges of aligning visual assets with narrative structure have also been explored in comic generation systems using semantic layout and procedural asset control (Chen and Jhala 2024, 2023). Tools like Tile Dreamer further streamline asset management, facilitating modular and dynamic level design (Karouzaki et al. 2007).

Grid-based procedural content generation (PCG) is a key for creating scalable, tile-based game environments. Foundational surveys by Togelius et al. and Hendrikx et al. explore search-based and modular PCG techniques that balance structured and random tile arrangements (Togelius et al. 2010; Hendrikx et al. 2013). As described by Yannakakis and Togelius, experience-driven PCG adapts content to player engagement (Yannakakis and Togelius 2011). PCG via machine learning (PCGML) leverages neural networks to generate context-sensitive levels, with Guzdial et al. and Hu et al. showing how machine learning and large language models facilitate narrative-driven tile placement (Guздial, Snodgrass, and Summerville 2022; Hu, Zhao, and Liu 2024). In addition, structured approaches to linking narrative events to visual layouts have been proposed in other

domains such as comics, providing a conceptual foundation for applying narrative-based graphs to the game tile arrangement (Chen 2025).

The Challenge of Low-resolution Images

Studies in low-resolution image recognition have shown that such low-detail characteristics impact the accuracy and usability of computer vision models. The study “80 Million Tiny Images” demonstrated that while a large dataset of low-resolution images can be beneficial for object recognition, limitations of the image details make distinguishing fine features difficult (Torralba, Fergus, and Freeman 2008). After that, studies such as Wang et al. further highlight the performance drop on deep networks working with very low-resolution images (e.g., 16x16 pixels), where they proposed the use of domain adaptation and super-resolution methods to counterbalance resolution loss (Wang et al. 2016). Semantic coherence and visual structure in low-resolution visual content have been addressed through hierarchical graph representations in visual narrative domains (Chen 2025).

Dataset Overview and Annotation Schema

In this section, we describe our GameTileNet dataset and present descriptive statistics that characterize its expressive range.

Dataset Summary: GameTileNet includes 2,142 labeled game objects collected from top-down pixel art assets on OpenGameArt.org, a community-driven repository of open game materials. All assets are used under Creative Commons licenses, and attribution is preserved according to license terms. The dataset includes tiles from 67 tilesets and individual images created by 15 different artist groups.

As illustrated in Table 2, game tile resources are available in two formats: full tilesets (which require segmentation) and individual object tiles. Each segmented tile is labeled with object names, semantic tags, connectivity, affordance types, and hierarchical metadata. Labels are provided by both annotators and original asset authors, supplemented by algorithmic methods.

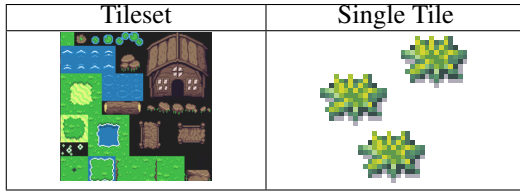


Table 2: Two types of game tile materials in OpenGameArt.org: a tileset and a single image.

Creative Commons License: All images included in GameTileNet are used under one of three Creative Commons licenses: CC0, CC-BY (3.0 or 4.0), or CC-BY-SA (3.0 or 4.0). We ensure full compliance with all license terms, including attribution and share-alike requirements when applicable.

Annotation Schema and Object Categories

Our annotation schema focuses on describing tile content to support narrative mapping and scene composition. Labels originate from asset filenames, author-defined tags, and human annotations. After preprocessing, we normalized 315 author names and 86 annotator-provided tags into a unified set of 361 object names.

Affordance Taxonomy and VGDL Alignment

Our affordance labels are informed by the Video Game Description Language (VGDL) (Schaul 2013), which categorizes objects by their gameplay roles. We adapt this taxonomy through a visual-semantic lens suitable for 2D tile-based scenes. The five final categories are:

- **Terrain:** Walkable or background surfaces such as grass or stone floors.
- **Environmental Object:** Static elements that provide context, including trees, fences, or architectural features.
- **Interactive Object:** Elements that players can trigger or move through, such as doors, levers, or switches.
- **Item/Collectible:** Objects that players can acquire or use, including potions, weapons, or scrolls.
- **Character/Creature:** Playable avatars, non-playable characters (NPCs), and enemies like goblins or shop-keepers.

Annotation Schema: Tile Types and Segment Labels

Each segmented tile is labeled according to both its visual integrity and its semantic role. The annotation types are as follows:

- **Tiles:** 32×32 pixel units extracted from larger tilesets.
- **Complete:** Tiles containing entire objects, visually enclosed with no cropping at the edges.
- **Partial:** Tiles showing truncated or occluded objects that extend beyond the tile boundary.
- **Texture:** Repeating material surfaces, such as wood grain or patterned stone.

- **Complete_texture:** Fully enclosed and tileable texture regions suitable for procedural repetition.
- **Partial_texture:** Texture samples that are visually cropped or unsuitable for seamless repetition.

We define usable tiles as either complete visual entities or texture patches that are structurally appropriate for terrain composition.

Core Annotation Tasks

Building on the defined schema, we structure the annotation process around four core tasks:

- **Adjacency (Edge Similarity):** Annotators judge if adjacent tile edges belong to the same object. Binary labels were collected for 200 randomly sampled tile pairs.
- **Tile Connectivity:** Annotators specify the directions (e.g., top-right, top-left) a tile can connect to others. This supports tile-based map assembly.
- **Object Detection:** Tiles are labeled with semantic object names.
- **Object Categorization:** Names are grouped into functional categories to reduce label noise and improve generalization.

While fixed vocabularies simplify labeling, subjective cases (e.g., partial crops) remain challenging. Interfaces used in these tasks are shown in Table 3.

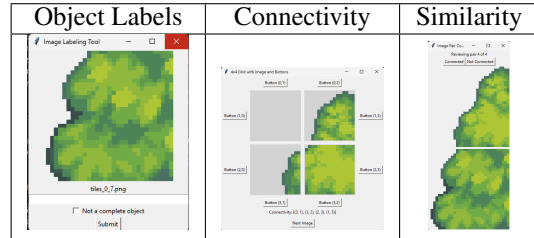


Table 3: Labeling interfaces for different tasks.

Consistency Assessment To evaluate annotation consistency, we performed a relabeling experiment. A subset of 50 tiles was randomly selected from the validation set and re-annotated by the same annotators after a two-week interval.

Author-Labeled Object Tiles

We include 1,898 additional object tiles originally labeled by asset creators. These objects come with free-text names and categories reflecting fantasy game conventions. The distribution includes: **monsters** (1,012), **items** (732), **tiles** (91), **terrain** (33), **player characters** (29), and **beings** (10).

To ensure consistency, we group these into six normalized categories—*creatures*, *items*, *player units*, *tiles*, *terrain*, and *narrative agents*—and standardize their labels.

Hierarchical Semantic Labeling Schema

Each object in the dataset is annotated with a four-level semantic schema, inspired by layered scene annotation practices such as ADE20K (Zhou et al. 2017, 2019):

- **Detailed Name:** The specific identifier for an object, often derived from file names or author labels.
- **Group Label:** A normalized, canonical name representing object variants under a single term.
- **Supercategory:** A higher-level classification that groups related objects, such as terrain or character types.
- **Affordance Label:** The gameplay function of the object, informed by VGDL-style interaction roles.

This schema enables both fine-grained and abstract reasoning across visual game elements and supports affordance-aware detection and symbolic alignment tasks.

Dataset Statistics

Table 4 give the statistics of usable units from 10995 game tiles, defined by tile completeness. Table 5 presents the statistics of annotated data, including the Adjacency (Edge Similarity) and Connectivity of game tiles, the object names of tile content, and the classes to which the objects belong. Additionally, subfigures 2a and 2b illustrate the frequency of objects appearing in GameTileNet.

Affordance Label Distribution Analysis To understand the diversity and overlap of affordance types in our dataset, we analyzed the annotated labels from two perspectives: (1) treating each unique combination of affordances as a single class, and (2) counting each individual label occurrence independently.

Combined Label View. Figure 3a (left) shows the distribution of combined affordance labels. The most frequent class is Characters, followed by Items and Collectibles and multi-label combinations such as Environmental Object | Items and Collectibles. This view highlights how often objects serve multiple gameplay functions, especially among background elements or items.

Separated Label View. Figure 3b (right) presents the distribution when labels are split and counted individually. Here, Characters remains dominant, but we also observe a high frequency of Items and Collectibles, Environmental Object, and Interactive Object. This reveals significant label co-occurrence, except for Characters, which tends to appear alone.

Affordance Overlap Insight. These two views together confirm that many game objects have overlapping functional roles—such as being both interactable and collectible—while Characters remain semantically distinct. This overlapping structure has important implications for downstream modeling.

Affordance Prediction as a Multi-Label Classification Task The affordance annotations in our dataset reflect a multi-label structure—many objects are associated with more than one gameplay function. To preserve this structure, we formulate affordance prediction as a **multi-label classification problem**, where each object can be assigned multiple affordance labels. This approach better reflects real-world

usage and supports flexible downstream applications such as gameplay reasoning and adaptive object interaction.

We adopt a lightweight model using CLIP embeddings and a multi-label MLP classifier. The training pipeline supports multi-hot targets and uses binary cross-entropy loss to optimize prediction across all label dimensions.

Automatic Annotation Pipeline

We design a compact pipeline for object detection in low-resolution pixel art tiles (Figure 1). Each tileset is first split into 32×32 units, which are grouped by visual similarity and summarized into a compact dictionary for efficient matching. Optional background padding is added to enhance context, and tiles are embedded into feature vectors using pre-trained visual encoders (e.g., CLIP). To improve clarity, images are upscaled before being passed to object detection models, whose outputs are used to assign affordance labels.

To support reuse of GameTileNet in PCG tasks, we evaluate the quality and robustness of each pipeline stage: segmentation, connectivity, and affordance prediction, rather than comparing against a universal baseline (which does not exist for low-resolution tile art). This component-wise validation ensures the pipeline produces reliable labels that can generalize to unseen tiles.

Adjacency (Edge Similarity): To determine whether adjacent tiles form a connected pair, we measure structural similarity at their shared boundaries. Given two adjacent tiles, we extract the connecting boundary rows or columns, convert them to grayscale (removing color variation), and compute the Structural Similarity Index Measure (SSIM) (Wang et al. 2004). SSIM evaluates luminance (μ), contrast (σ), and structural correlation ($\sigma_{B_1 B_2}$), offering a perceptually motivated metric that outperforms raw pixel comparison. Its value ranges from -1 (structurally different) to 1 (identical). The similarity between two boundary regions B_1 and B_2 is defined as:

$$SSIM(B_1, B_2) = \frac{(2\mu_{B_1}\mu_{B_2} + C_1)(2\sigma_{B_1 B_2} + C_2)}{(\mu_{B_1}^2 + \mu_{B_2}^2 + C_1)(\sigma_{B_1}^2 + \sigma_{B_2}^2 + C_2)}.$$

Segmentation: For segmenting objects inside tile sets, we employ a recursive region-growing approach that detects connected segments based on adjacency and similarity. The process begins with an unvisited tile, from which the algorithm recursively explores all four neighboring tiles (right, left, down and up), forming a connected segment. A tile is considered part of the same segment if it meets two conditions: (1) it is not primarily transparent, ensuring that only meaningful tiles are included, and (2) it passes a similarity threshold check with its adjacent tile. Each segment is anchored by its first encountered tile, designated as the parent, which groups all connected tiles under it.

Dataset and Training We curated a binary classification dataset with thousands of labeled tile segments. A lightweight ResNet18 model pretrained on ImageNet was fine-tuned on these examples. Tiles were resized to 64×64 , and the data was split into training, validation, and test sets (70%15%15%). Training used cross-entropy loss, the Adam optimizer, and early stopping with a maximum of 15 epochs.

Tileset	Total	Partial	Complete	P_texture	C_texture	Usable_rate
Segment_algo_recursive	1818	1064	737	279	17	56.8%
Segment_model	2612	1754	841	17	171	39.4%

Table 4: Summary of tile segment counts and usable rate from two segmentation methods. The usable rate measures the proportion of extracted segments that result in complete, recognizable objects or textures (e.g., a full chest) rather than partial or cropped fragments.

	Adjacency	Connectivity	# Objects Labeled	Object Groups	# Supercategories	Affordance Labeled
Annotator	200	100	244	86	26	2142
Author	X	X	1898	315	X	X
Algo & Models	90.5%	43%	X	X	X	See Table 8

Table 5: Summary of annotation statistics from three sources: human annotators, original content authors, and automatic models. "Adjacency" and "Connectivity" refer to tile relationships. "Affordance Labeled" indicates the number of tiles annotated with affordance types or prediction accuracy (detailed in Table 8).

Performance We conducted three training runs to evaluate stability. The best model achieved 92.2% test accuracy, with an average validation accuracy of 91.8%. Accuracy typically plateaued by epoch 8–10, demonstrating efficient convergence and minimal overfitting.

Discussion The classifier reliably distinguishes complete from partial tiles using low-level visual features. This supports automated pre-filtering of segments for reuse, reducing manual annotation effort and improving scalability for applying our framework to new game assets.

Connectivity Tile-based games require seamless transitions between adjacent tiles to ensure visually coherent environments. We define tile connectivity in terms of directional adjacency, where each tile is divided into eight directional edge segments (e.g., top-left, right-down). We propose a hybrid rule-based algorithm that combines visual similarity and edge transparency to infer directional connectivity between tiles. For each tile and each edge direction, we first check whether the corresponding edge segment is non-transparent, based on the alpha channel of the image. If the edge is non-transparent, it is marked as a potential connection. If a neighboring tile exists in that direction, we extract the overlapping edge regions from both tiles and compute the Structural Similarity Index Measure (SSIM). If the SSIM score exceeds a predefined threshold (e.g., 0.6), the edge is considered a valid directional connection. This dual-filtering approach helps avoid both false connections from transparent edges and false negatives due to visual variation.

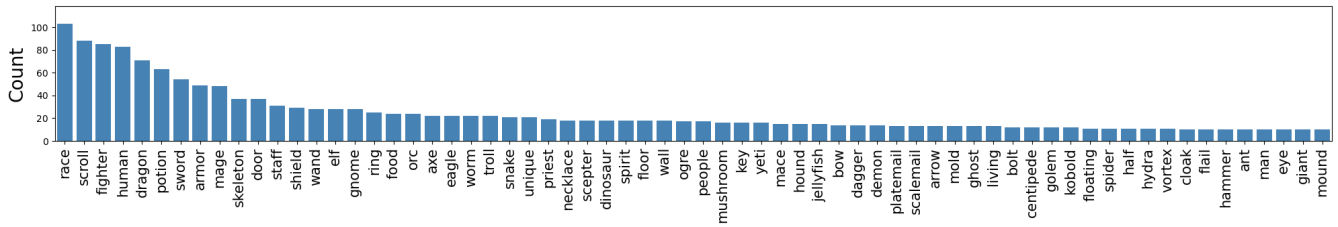
Comparison with Manual Annotations To assess the quality of automatically generated connectivity labels, we compared the predicted edges against manually annotated ground truth using precision, recall, F1 score, and exact match rate. The best-performing variant achieved an average F1 score of 0.63 and an exact match rate of 43%, indicating a moderate level of agreement with human annotations. The balanced precision and recall further support the trustworthiness of the predicted labels for downstream tasks. We adopt this version as our refined connectivity annotation baseline and use it in subsequent experiments.

Semantic Labels The dataset includes two types of semantic labels: (1) object names, and (2) categorical classes. To normalize game-related words into categories such as Characters, Items, Terrain, Non-human, and Attribute, annotators used GPT-4 for suggestions in ambiguous cases (e.g., when file names were unclear), then verified or corrected the results. Although not formally benchmarked, this process improved consistency and reduced label noise.

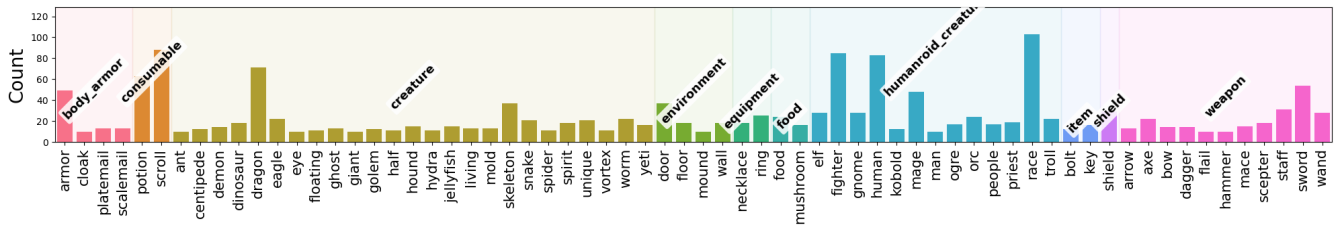
Preprocessing In the preprocessing steps, we utilize the Natural Language Toolkit (NLTK) for initial cleaning (Bird, Klein, and Loper 2009; Loper and Bird 2002). The cleaning process includes three steps: first, checking for and removing duplicated words; second, filtering the words to retain only nouns; and third, identifying and eliminating plurals and transformed forms. *Environmental Objects*: Tree, Rock, Water, Grass, Pathway, Building. *Interactive Objects*: Door, Key, Chest, Switch, Lever. *Characters and NPCs*: Player, Enemy, NPC, Boss. *Items and Collectibles*: Potion, Coin, Gem, Weapon, Armor

Tile Upscaling Methods for Visual Enhancement Low-resolution tiles often suffer from aliasing and loss of detail, which hinders accurate object detection. We evaluate five upscaling methods to enhance 32×32 pixel game tiles prior to recognition:

- **Bicubic Interpolation (PIL)**: A baseline method using traditional image resizing via Python Imaging Library.
- **ESRGAN**: A generative adversarial network trained for perceptual super-resolution on natural images, preserving sharp edges and textures.
- **SwinIR**: A transformer-based method designed for image restoration tasks, which balances fidelity and structure preservation.
- **Stable Diffusion (Fidelity)**: Image super-resolution using diffusion models conditioned on the original image for high-detail fidelity.
- **Stable Diffusion (Img2Img)**: A creative variant of Stable Diffusion that allows semantic enhancement via denoising strength and prompt guidance.



(a) Object counts for each group label, sorted by frequency.



(b) Group labels clustered by supercategories, with color-coded class regions.

Figure 2: Distribution and semantic structure of object categories in GameTileNet. (a) Group label frequencies, limited to labels with more than 10 occurrences. (b) Supercategories over the full set; colored regions show dataset-wide coverage.

Figure 6 illustrates visual differences across methods on the same input tile.

Detection and Semantic Matching We evaluate semantic understanding of game tiles using two approaches: a conventional object detector (YOLOv5s) and a vision-language captioning model (BLIP). Both are applied to original and upscaled images to assess the impact of enhancement on recognition.

YOLOv5s, pretrained on COCO, provides bounding-box detections with fixed real-world labels, which are poorly suited for fantasy-themed, low-resolution tiles. While upscaling improves detection quantity, domain mismatch prevents alignment with our semantic categories. In contrast, BLIP generates free-form captions in an open-vocabulary setting, enabling recognition of stylized or unseen objects. We compare these captions to annotated group labels, supercategories, and affordances to assess semantic alignment.

Overall, YOLO offers spatial grounding but struggles with vocabulary fit, while BLIP supports descriptive and domain-adaptive recognition without localization. BLIP’s outputs show promise for bootstrapping semantic labels, clustering assets, or supporting interactive authoring in creative workflows.

Evaluating Caption–Label Similarity To assess whether vision-language captioning methods such as BLIP can help recognize the semantics of game tiles, we compare the generated captions with curated object labels. Each image is associated with three types of ground-truth labels: (1) **group labels**, (2) **supercategories**, and (3) **affordance labels**. These labels were manually annotated by authors or external annotators.

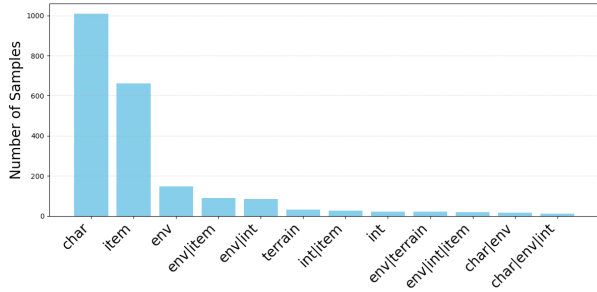
We evaluate the similarity between generated captions and ground-truth labels along three levels: **Direct Match**:

We check whether any exact word in the caption matches a word from the group labels, supercategories, or affordances. **Synonym Match**: Using WordNet-based synonym sets, we expand each label to include common synonyms and check if any appear in the caption. **Semantic Similarity**: We encode both the caption and each label term using a Sentence-BERT encoder and compute cosine similarity scores. A match is considered positive if the similarity score exceeds a threshold (e.g., 0.3).

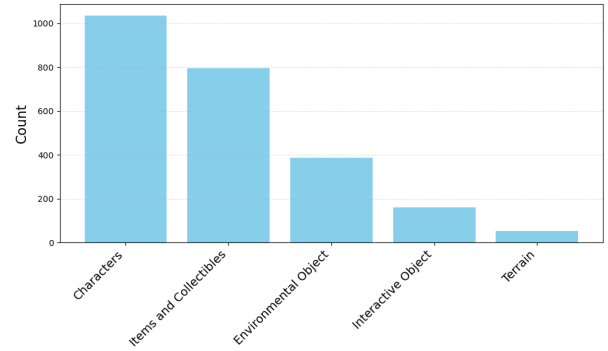
These three matching levels help evaluate whether captions preserve or convey the intended semantic information of the game tile images, and which upscaling method supports this best. The final results are reported as counts of matches across all images per label type and match level.

Evaluating Object Recognition and Semantic Label Alignment To assess how different upscaling methods affect the interpretability of game tiles, we evaluate two forms of automatic annotation: (1) object detection using YOLOv8 and (2) caption–label alignment using BLIP-generated image captions. Table 7 summarizes the detection counts and match statistics, while Figure 5 visualizes the proportion of caption–label matches by match type and label type.

YOLO Detection and Its Limitations. Despite upscaling improving YOLO detection counts—for instance, SwinIR yields 1306 detections and Real-ESRGAN 1138 compared to none at 32×32 resolution—the matches against our annotated labels remain *zero*. This reveals a fundamental limitation: YOLO models are trained on fixed, real-world categories (e.g., *traffic light*, *bicycle*), which rarely overlap with our fantasy-themed or stylized game objects. As a result, while YOLO detects more shapes after upscaling, these labels are largely irrelevant for domain-specific semantic understanding.



(a) Combined affordance label distribution. Each bar is a unique label combination.



(b) Separated label distribution. Labels are split and counted independently.

Figure 3: Affordance label distributions across annotated object tiles. **Left:** Unique label combinations show the diversity of functional roles. **Right:** Individual label counts reveal frequent co-occurrence among affordances, except for Characters, which rarely overlaps.

Original	Bicubic (PIL)	ESRGAN	SwinIR	Stable Diffusion (Fidelity)	Stable Diffusion (Img2Img)

Table 6: Visual comparison of five upscaling techniques on a 32×32 tile (Environmental Object: yellow flower on grass). Stable Diffusion is shown in both fidelity-preserving and generative (img2img) variants.

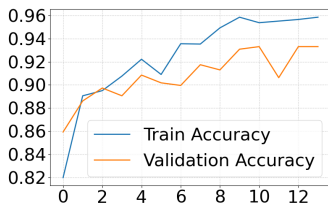


Figure 4: Training and validation accuracy across epochs for the completeness classification task (y-axis: accuracy, x-axis: epochs).

Caption-Based Semantic Matching. In contrast, the BLIP captioning model—designed for open-vocabulary generation—shows stronger alignment with human-provided labels. As shown in Figure 5, upscaling improves semantic similarity across all label types: **group labels**, **supercategories**, and **affordance labels**. SwinIR and Real-ESRGAN achieve the best alignment, with Real-ESRGAN slightly leading in affordance matching and SwinIR in supercategories.

Match Types and Upscaling Effects. We evaluate three levels of alignment: **Direct Matches:** Exact string overlap (e.g., “barrel” in both caption and label). **Synonym**

Matches: Lexical overlap via WordNet-style synonym sets. **Semantic Similarity:** Cosine similarity using sentence embeddings (e.g., “wizard” and “mage”). While direct and synonym matches remain relatively sparse, semantic similarity matches increase notably after upscaling, indicating that image quality helps generate captions that conceptually align with intended labels, even when phrased differently.

Affordance Classification via CLIP and Multi-Label Prediction We frame affordance classification as a multi-label task, allowing tiles to have overlapping roles (e.g., Characters and Items and Collectibles). We use a pre-trained CLIP model (ViT-B/32) to extract 512-dimensional embeddings, followed by a lightweight multi-layer perceptron (MLP) with a sigmoid output layer to predict affordance types. The model is trained with binary cross-entropy loss, and a fixed threshold (e.g., 0.5) is applied at inference to assign multiple labels per tile.

Predicted Results. Table 8 shows per-label precision, recall, and F1-score on a held-out test set. The model performs well overall, with strong results for Characters and Items and Collectibles, likely due to their visual distinctiveness and higher representation. Interactive Object performs lower, suggesting greater visual ambiguity. Micro- and macro-averaged F1 scores are 0.92 and 0.86, indicating robust performance across both common and rare

Image Source	Total Img.	YOLO Det. (#)	Dir. Match (G/S/A)	Syn. Match (G/S/A)	Sem. Match (G/S/A)
Original (32x32)	1898	N/A	104/3/–	143/4/24	207/177/122
Bicubic	1898	1058	166/8/–	197/9/56	335/338/105
Real-ESRGAN	1898	1138	170/2/–	204/22/311	432/383/374
SwinIR	1889	1306	118/15/–	149/20/181	276/300/214
SD (fidelity)	1898	35	51/2/–	68/2/2	96/181/18
SD (img2img)	1898	1171	146/9/–	192/10/45	278/333/102

Table 7: Combined caption–label match statistics and YOLO detection counts across upscaling methods. G = group label, S = supercategory, A = affordance label. Although YOLO detections increased after upscaling (e.g., 1306 detections for SwinIR), none matched the target labels, highlighting a vocabulary mismatch.

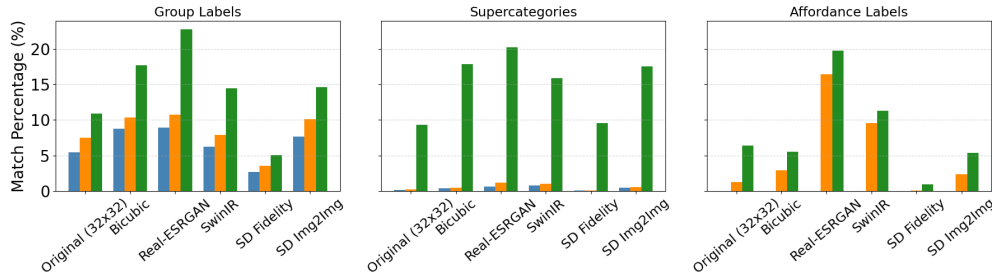


Figure 5: Percentage of caption–label matches across different upscaling methods. Each bar shows the proportion of matches between BLIP-generated captions and author-annotated labels: **Group Labels**, **Supercategories**, and **Affordance Labels**. We measure three types of matches— **Direct Match**, **Synonym Match**, and **Semantic Similarity**— using lexical and embedding-based comparisons. SwinIR and Real-ESRGAN upscaled images show notably better alignment with captions, especially in semantic terms.

labels.

Label	Prec.	Rec.	F1	Supp.
Characters	0.99	0.96	0.98	200
Environmental Obj.	0.89	0.88	0.88	90
Interactive Obj.	0.50	0.65	0.56	31
Items & Collectibles	0.93	0.97	0.95	160
Terrain	1.00	0.86	0.92	7

Table 8: Per-label classification report for affordance prediction.

Narrative Scene Generation

As a proof of concept, we demonstrate how GameTileNet supports narrative-to-scene generation by combining predicate extraction, semantic tile matching, and affordance-based rendering.

Overview and Pipeline: We propose a pipeline that transforms narrative descriptions into semantically coherent game scenes. The process begins with terrain generation using Cellular Automata (CA), followed by semantic object matching, spatial placement guided by knowledge graphs (KGs), and final visual rendering.

Terrain Generation with Cellular Automata We use Cellular Automata (CA) to generate base terrain layers that form connected and navigable regions. These maps serve as

the structural foundation for placing story-relevant content. CA ensures that terrain tiles are walkable and supports dynamic expansion for patch-based refinements.

Semantic Object Matching: To map narrative elements to visual assets, we construct a semantic embedding index of all annotated game tiles. Each object, annotated with names, group labels, supercategories, and affordance types, is encoded using the all-MiniLM-L6-v2 Sentence Transformer. During generation, narrative entities are embedded and matched to game tiles via cosine similarity, ensuring visual-semantic alignment. Affordance categories help disambiguate semantically similar objects (e.g., "guardian" as a character vs. statue).

Narrative Prompting and Frame Decomposition: Inspired by Word2World (Nasir, James, and Togelius 2024), we prompt an LLM to generate short adventure stories and decompose them into three key time frames using structured predicate-style descriptions. An example story and corresponding spatial breakdown are shown in Table 9.

- Prompt_1: "Generate a short adventure story, 100 words."
- Prompt_2: "Retrieve three key time frames from the story, and describe them with [Object] [Relation] [Object] type of scene descriptions."

Relation alias, fine-tuned model with categories, Find closest tiles in the categories.




Time Frame 1	Time Frame 2	Time Frame 3
Elara discovers the ancient map	Elara faces the treacherous paths.	Elara meets the Guardian dragon.
Hollow oak contains ancient map Elara stands near hollow oak Sunlight filters through forest canopy	Elara walks along rocky path Wild creatures hide behind dense bushes Treacherous paths lead to Crystal Cavern	Crystal Cavern entrance glows with shimmering light Guardian dragon sits atop crystal throne Elara stands before Guardian dragon
		

Table 9: Three key time frames extracted from a generated story, represented by spatial predicates and visual renderings.

"In the heart of the Enchanted Forest, young Elara discovered an ancient map hidden within a hollow oak. It led her to the legendary Crystal Cavern, rumored to grant the finder a single wish. Braving treacherous paths and wild creatures, Elara reached the cavern's shimmering entrance. Inside, she faced the Guardian, a majestic dragon. With courage and wit, she solved the Guardian's riddle, earning her the wish. Elara wished for peace in her war-torn village. As she exited the cavern, the skies cleared, and harmony was restored, proving that bravery and hope could transform the world."

Knowledge Graph Construction: For each scene, we construct a knowledge graph (KG) from narrative predicates in the form $\langle \text{subject} \rangle \langle \text{relation} \rangle \langle \text{object} \rangle$. Entities are classified by affordance types (e.g., character, item, terrain) and enriched with metadata like suggested terrain or placement confidence. Predicate relations are mapped to spatial relations (e.g., *on top of*, *below*) and connected through spatial and temporal edges. We build: **Scene-level KGs**, which capture objects, relations, and spatial structure. **Merged KGs**, which integrate multiple scenes and include precedes edges for narrative order.

Scene Synthesis and Rendering: Using the KG and terrain map, each frame is synthesized as a tile-based scene:

- The terrain layer is initialized using CA with connectivity constraints.
- Each object is mapped to a visual tile asset and positioned using rule-based placement based on spatial relations (e.g., "above" \rightarrow vertical adjacency).
- Object tiles are overlaid by sprite rendering, preserving size, centering, and aspect ratio.

Each scene is output as both a numerical matrix and a visual rendering. This representation enables further integration into games or interactive narrative systems. This

layered narrative-to-scene pipeline bridges symbolic storytelling with 2D game synthesis. By combining semantic embeddings, spatial reasoning, and procedural terrain generation, we demonstrate a scalable approach to grounding generative narratives in visually structured environments.

While limited in scope, this module highlights the potential of GameTileNet for narrative-driven PCG. We leave a deeper evaluation of temporal consistency and agent coordination to future work in an extended study.

Discussion and Conclusion

GameTileNet provides a structured semantic dataset tailored for low-resolution pixel art games, with annotations that support affordance prediction, narrative alignment, and symbolic scene generation. Our experiments demonstrate that combining vision-language models with upscaling techniques opens promising directions for interpreting stylized assets, particularly when conventional object detectors fall short.

While the dataset is detailed enough for direct use in PCG pipelines, challenges remain in domain adaptation and object consistency, especially for ambiguous or multi-functional tiles. Our narrative scene generation pipeline shows how symbolic structures and spatial reasoning can bridge story elements and visual output, offering a foundation for future interactive tools or co-creative systems. We see this as a step toward grounding generative AI in playable, narrative-rich game environments. We hope GameTileNet supports researchers and developers working on PCGML, semantic asset reuse, and the integration of storytelling in 2D games.

Availability

The GameTileNet dataset, including annotation schemas and scripts, is publicly available at <https://github.com/RimiChen/2024-GameTileNet>.

References

- Ang, C. S. 2006. Rules, gameplay, and narratives in video games. *Simulation & Gaming*, 37(3): 306–325.
- Bentley, G. R.; and Osborn, J. C. 2019. The videogame affordances corpus. In *2019 Experimental AI in Games Workshop*.
- Bird, S.; Klein, E.; and Loper, E. 2009. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O’Reilly Media, Inc.
- Bulatovic Trygg, N.; Skult, P.; and Smed, J. 2024. Narrative Design. In *Encyclopedia of Computer Graphics and Games*, 1225–1230. Springer.
- Chen, Y.-C. 2025. Structured Graph Representations for Visual Narrative Reasoning: A Hierarchical Framework for Comics. *arXiv preprint arXiv:2506.10008*.
- Chen, Y.-C.; and Jhala, A. 2023. A customizable generator for comic-style visual narrative. *arXiv preprint arXiv:2401.02863*.
- Chen, Y.-C.; and Jhala, A. 2024. Collaborative Comic Generation: Integrating Visual Narrative Theories with AI Models for Enhanced Creativity. In *Proceedings of the 3rd Workshop on Artificial Intelligence and Creativity*, volume 3810.
- Domsch, S. 2013. *Storyplaying: Agency and narrative in video games*. De Gruyter.
- Gabriel, L.; and Clua, E. W. 2023. A Semantic Segmentation System for generating context-based tile-maps. In *Proceedings of the 22nd Brazilian Symposium on Games and Digital Entertainment*, 124–133.
- Guzdial, M.; Snodgrass, S.; and Summerville, A. J. 2022. Grid-Based DNN PCGML. In *Procedural Content Generation via Machine Learning: An Overview*, 135–158. Springer.
- Hendrikx, M.; Meijer, S.; Van Der Velden, J.; and Iosup, A. 2013. Procedural content generation for games: A survey. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 9(1): 1–22.
- Hu, C.; Zhao, Y.; and Liu, J. 2024. Game generation via large language models. In *2024 IEEE Conference on Games (CoG)*, 1–4. IEEE.
- Jadhav, M.; and Guzdial, M. 2021. Tile embedding: a general representation for level generation. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 17, 34–41.
- Karouzaki, E.; Savidis, A.; Katzourakis, A.; and Stephanidis, C. 2007. Tile dreamer: game tiles made easy. In *Universal Access in Human Computer Interaction. Coping with Diversity: 4th International Conference on Universal Access in Human-Computer Interaction, UAHCI 2007, Held as Part of HCI International 2007, Beijing, China, July 22-27, 2007, Proceedings, Part I 4*, 382–391. Springer.
- Koenitz, H. 2024. Narrative in video games. In *Encyclopedia of computer graphics and games*, 1230–1238. Springer.
- Lafitte, G.; and Weiss, M. 2009. An almost totally universal tile set. In *International Conference on Theory and Applications of Models of Computation*, 271–280. Springer.
- Loper, E.; and Bird, S. 2002. NLTK: The Natural Language Toolkit.
- Nasir, M. U.; James, S.; and Togelius, J. 2024. Word2World: Generating Stories and Worlds through Large Language Models. *arXiv preprint arXiv:2405.06686*.
- Schaul, T. 2013. A video game description language for model-based or interactive learning. In *2013 IEEE Conference on Computational Intelligence in Games (CIG)*, 1–8. IEEE.
- Shapiro, M. A.; Peña-Herborn, J.; and Hancock, J. T. 2012. Realism, imagination, and narrative video games. In *Playing video games*, 323–340. Routledge.
- Togelius, J.; Yannakakis, G. N.; Stanley, K. O.; and Browne, C. 2010. Search-based procedural content generation. In *Applications of Evolutionary Computation: EvoApplications 2010: EvoCOMPLEX, EvoGAMES, EvoIASP, EvoINTELLIGENCE, EvoNUM, and EvoSTOC, Istanbul, Turkey, April 7-9, 2010, Proceedings, Part I*, 141–150. Springer.
- Torralba, A.; Fergus, R.; and Freeman, W. T. 2008. 80 million tiny images: A large data set for nonparametric object and scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 30(11): 1958–1970.
- Tyndale, E.; and Ramsomair, F. 2016. Keys to Successful Interactive Storytelling: A Study of the Booming” Choose-Your-Own-Adventure” Video Game Industry. *Journal of Educational Technology*, 13(3): 28–34.
- van der Spuy, R. 2010. Tile-Based Game Design. In *AdvancED Game Design with Flash*, 503–621. Springer.
- Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4): 600–612.
- Wang, Z.; Chang, S.; Yang, Y.; Liu, D.; and Huang, T. S. 2016. Studying very low resolution recognition using deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4792–4800.
- Williams, T. 2024. Tickr AI Research and Engineering Blog. <https://www.tickr.com/blog/author/timwilliams/>. Accessed: 2024-10-22.
- Yannakakis, G. N.; and Togelius, J. 2011. Experience-driven procedural content generation. *IEEE Transactions on Affective Computing*, 2(3): 147–161.
- Zhou, B.; Zhao, H.; Puig, X.; Fidler, S.; Barriuso, A.; and Torralba, A. 2017. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 633–641.
- Zhou, B.; Zhao, H.; Puig, X.; Xiao, T.; Fidler, S.; Barriuso, A.; and Torralba, A. 2019. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127: 302–321.