Learning Shared Visual Representations for Manga Title and Target Classification: A Multi-Task Closed and Open-Set Evaluation

Rosario Licciardello[0009−0007−4586−6632], Georgia Fargetta[0000−0002−6444−1564], Alessandro Ortis[0000−0003−3461−4679], and Sebastiano Battiato[0000−0001−6127−2470]

University of Catania, Department of Mathematics and Computer Science, Viale Andrea Doria 6, 95125 Catania, Italy
1000069335@studium.unict.it,{georgia.fargetta, alessandro.ortis, sebastiano.battiato}@unict.it

Abstract. Manga pages carry rich visual semantics at multiple levels of granularity, from the broad demographic target of a series to its specific visual identity, yet learning representations that jointly capture both dimensions remains unexplored. Multi-Task Learning (MTL) is studied as a principled approach to this problem, jointly optimising demographic and title-level supervision on Manga109 across fourteen model configurations and evaluating on both closed-set classification and retrieval. Multi-task models consistently improve closed-set demographic classification over single-task baselines (+3.7 pp), with negligible loss on title discrimination, and produce embedding spaces that preserve coherent structure along both semantic axes simultaneously, a property no single-task model achieves. In open-set retrieval on unseen commercial titles, multi-task embeddings match title-specialised models on R@1, yet both classification and retrieval metrics degrade substantially across all configurations, underscoring that zero-shot generalization to unseen visual identities remains an open challenge. An evaluation framework covering fourteen model configurations across four metrics is released, providing a structured baseline for future work on scalable manga indexing and recommendation.

Keywords: Manga Classification · Multi-Task Learning · Open-Set Recognition · Visual Embeddings · Manga109

1 Introduction

Deep learning has significantly advanced image classification across various domains. Manga panel classification presents unique challenges due to the stylized nature of illustrations, the presence of text, and the varying artistic styles across different works. Graphic images such as manga are sparse and high-contrast, unlike natural images which are rich in gradients [9, 3]. Among deep learning techniques, CNNs have become the dominant framework for comic classification [16].

The Manga109 dataset [1], introduced by Matsui et al. [13], provides a diverse collection of panels from 109 titles and represents the primary benchmark in this field. Prior research has predominantly focused on single-task classification with a limited number of classes, often based on artist styles [22]. In contrast, our study adopts a Multi-Task Learning (MTL) framework to concurrently categorize 109 manga titles and their demographic categories, exploring whether dual supervision provides a stronger inductive bias than single-task training. Accurate manga recognition serves multiple practical goals: it enables age-based content classification into groups such as *shonen*, *shojo*, *seinen*, and *josei*; it supports intellectual property protection by authenticating original art styles [16]; and it provides a foundation for visual embedding-based recommendation systems. While open-set retrieval remains challenging, our results demonstrate that joint supervision offers a more robust starting point than single-task approaches. Our study contributes to the field in three key ways. First, it investigates the impact of joint demographic and title supervision on closed-set classification and embedding-space retrieval under a unified MTL framework. Second, it systematically benchmarks fourteen ResNet-50 configurations to identify the optimal balance for multi-task manga representation. Third, it evaluates generalization through open-set retrieval on unseen commercial titles, showing that multi-task embeddings match title-specialised models while substantially outperforming demographic-only baselines. The remainder of this paper is structured as follows: Section 2 reviews related work; Section 3 describes the datasets; Section 4 presents the methodology; Section 5 reports results and analysis; Section 6 concludes.

2 Related Works

Scientific research on comic data remains limited compared to natural image processing, due to the complexity of the medium and the scarcity of large-scale open datasets caused by copyright restrictions. Early classification efforts on Manga109 were restricted to small subsets, such as artist identification across eight categories [22]. Subsequent work shifted toward supporting recommender systems: Vie et al. [19] extracted tag information from posters to address cold-start, Soni et al. [17] generated latent embeddings for anime titles via autoencoders, and Kim et al. [11] combined visual style extraction with synopsis analysis. Cross-domain transfer between anime and manga has also been shown to reduce recommendation error [2]. The Manga109 dataset [1, 13] has driven progress toward holistic manga understanding. Feng et al. [7] showed that panel layouts alone serve as a stylistic fingerprint sufficient for high-accuracy title identification, confirming that structural organisation encodes work-specific visual identity. Vaze et al. [18] established that closed-set accuracy is highly correlated with open-set generalisation, motivating robust title classification as a prerequisite for practical deployment. In this context, the adoption of adversarially resilient architectures, such as ShieldNet [12], demonstrates how dynamic weighting and gradient regularization can further enhance model stability across

diverse and potentially adversarial input distributions. Zuccarà et al. [23] further addressed open-set recognition by introducing a synthetic unknown class via adversarial learning to improve feature space cohesion under unseen inputs. In the MTL literature, [14] showed that jointly learning panel detection, character detection [5], and balloon segmentation reduces processing time with no loss in performance. Nguyen et al. [15] extended this to manga character analysis, demonstrating that an auxiliary album classification task acts as a regulariser that improves character feature learning over all single-modal baselines. Beyond comics, CNN-based classification under severe class imbalance and high intra-class visual similarity has been studied in other specialised visual domains [6, 4], where richer supervision and enlarged training signals have been shown to compensate for label scarcity in minority classes, a finding that motivates the auxiliary title supervision adopted in this work. Despite these advances, no prior work has jointly predicted manga titles and demographic targets at scale. To assess generalisation, we evaluate these embeddings through open-set retrieval on 20 modern commercial titles external to the Manga109 dataset, including five works per demographic category. While this investigation identifies significant domain-specific limitations, it establishes an empirical backbone for building more scalable and robust visual discovery systems in the future.

3 Dataset

Closed-Set Dataset. Manga109 is a benchmark dataset comprising 109 manga titles. For this experiment, the corpus consists of 12,655 images at a resolution of 256×256 pixels, predominantly in grayscale. Each title is annotated with one of four demographic target labels: Shonen (young boys), Shojo (young girls), Seinen (young men), and Josei (young women). These labels are not part of the official Manga109 annotations and were sourced from a public Kaggle competition [20]. The dataset exhibits a class imbalance: Shonen is the dominant category (43 titles, 5,116 images), followed by Seinen (26 titles, 2,991 images), Shojo (25 titles, 2,840 images), and Josei (15 titles, 1,708 images). The dataset was partitioned into training, validation, and test sets following an 80/10/10 ratio, stratified at the image level by title, resulting in 10,123 training, 1,266 validation, and 1,266 test images. The split is deterministic and cached via an MD5 hash of its defining parameters, guaranteeing identical partitions across all experiments. Images are converted to RGB and center-cropped to 224×224 pixels. Training images undergo on-the-fly augmentation: random resized cropping (scale $\in [0.8, 1.0]$), horizontal flipping, random rotation (± 15), and color jitter (brightness ± 0.2, contrast ± 0.2, saturation ± 0.1). Validation and test images are resized to 256 pixels and center-cropped to 224×224. All images are normalised using ImageNet statistics ($\mu = [0.485, 0.456, 0.406]$, $\sigma = [0.229, 0.224, 0.225]$), consistent with the ResNet-50 pre-training protocol.

Open-Set Dataset. To evaluate the zero-shot generalization capabilities of the models beyond the Manga109 distribution, a dedicated open-set dataset was constructed. This corpus comprises 20 unseen, highly successful commercial manga

Table 1: Composition of the open-set evaluation dataset. The corpus consists of 20 modern commercial titles perfectly balanced across the four demographic targets. Exactly the first 300 pages of each title were extracted (6,000 pages total).

Title	**Target**	**Decade**	**Publisher**
Ao Haru Ride	Shojo	2010s	Shueisha
Berserk	Seinen	1990s	Hakusensha
Cardcaptor Sakura	Shojo	1990s	Kodansha
Chainsaw Man	Shonen	2010s	Shueisha
Chihayafuru	Josei	2000s	Kodansha
Dandadan	Shonen	2020s	Shueisha
Dr. STONE	Shonen	2010s	Shueisha
Fruits Basket	Shojo	1990s	Hakusensha
Honey and Clover	Josei	2000s	Shueisha
Jujutsu Kaisen	Shonen	2010s	Shueisha
Kaguya-sama: Love is War	Seinen	2010s	Shueisha
Nana	Josei	2000s	Shueisha
One Piece	Shonen	1990s	Shueisha
Ouran High School Host Club	Shojo	2000s	Hakusensha
Oyasumi Punpun	Seinen	2000s	Shogakukan
Paradise Kiss	Josei	2000s	Shodensha
Princess Jellyfish	Josei	2000s	Kodansha
Tokyo Ghoul	Seinen	2010s	Shueisha
Vinland Saga	Seinen	2000s	Kodansha
Yona of the Dawn	Shojo	2010s	Hakusensha

titles serialised in major Japanese magazines. To maintain strict consistency in the evaluation phase, exactly the first 300 pages from the initial volumes of each title were extracted, yielding a uniform and strictly disjoint evaluation corpus of 6,000 pages. Unlike the Manga109 dataset, where a single image typically corresponds to a two-page spread, each image in this open-set dataset strictly represents a single page. To ensure a rigorous evaluation, the dataset is perfectly balanced across the four target demographic classes, featuring exactly five titles each for Shonen, Shojo, Seinen, and Josei. To prevent confounding factors and spurious correlations, the selection deliberately enforces variance across secondary editorial attributes. As detailed in Table 1, the corpus encompasses diverse major publishing houses (Shueisha, Kodansha, Hakusensha, Shogakukan) and distinct modern publication eras spanning from the 1990s to the 2020s. This compositional balance ensures that performance metrics on open-set retrieval, clustering, and classification tasks reflect genuine structural and demographic generalisation, rather than superficial overfitting to specific artistic trends of a single decade or specific editorial guidelines.

Ethical and Copyright Considerations. All commercial manga volumes included in this open-set evaluation corpus were legally purchased by the authors. In compliance with copyright regulations regarding data analysis and machine

learning applications, the use of these works is strictly confined to computational feature extraction and statistical evaluation. To respect intellectual property rights and prevent any potential redistribution or reconstruction of the original works, the raw images will not be made publicly available. However, full reproducibility of the experiments is ensured: the exact titles are detailed in this manuscript, and the extraction protocol systematically targets the first 300 pages of each respective series, allowing independent researchers to seamlessly reconstruct the dataset by acquiring the corresponding commercial volumes.

4 Methodology

All model variants share a ResNet-50 [8] backbone pretrained on ImageNet. The selection of this specific architecture is guided by preliminary internal evaluations suggesting its suitability for manga visual feature extraction. The focus of this study is not to maximize absolute classification metrics with the latest state-of-the-art vision architectures, but rather to employ a well-established baseline. This choice helps ensure that the observed behaviors, particularly the interplay between title and demographic objectives and the resulting zero-shot generalisation, can be attributed to the multi-task learning framework itself, limiting potential confounding factors tied to more complex architectural choices. Following the removal of its original classification head, The 2048-dimensional feature vector produced by the global average pooling layer is projected into a 512-dimensional embedding space via a fully connected layer. Depending on the variant, a BatchNorm1d layer may follow the projection before the activation function ϕ, which is either ReLU or Tanh. The resulting embedding $\mathbf{z} \in R^{512}$ is then passed to one or two task-specific linear classifiers. All models are trained with the Adam optimizer at a fixed learning rate of 10^{-4} and a batch size of 128, retaining the checkpoint with the highest validation accuracy. A total of 14 configurations are investigated, obtained by combining task type, activation function, presence of BatchNorm1d, and loss weighting strategy. Eight are single-task baselines trained independently for title or target classification. The title variants attach a linear classifier $\mathbf{W}_t \in R^{109\times512}$ to $\mathbf{z}$ and minimise $\mathcal{L}_{\text{title}}$ over 109 classes; the target variants attach $\mathbf{W}_g \in R^{4\times512}$ and minimise $\mathcal{L}_{\text{target}}$ over four demographic categories. The remaining six are multitask models attaching both classifiers simultaneously, investigating two loss formulations. The equal-weight variant (4 configurations: ReLU/Tanh $\times$ with/without BatchNorm1d) uses fixed unit weights:

$$\mathcal{L}_{\text{sum}} = \mathcal{L}_{\text{title}} + \mathcal{L}_{\text{target}}. \tag{1}$$

The uncertainty-weighted variant (2 configurations: ReLU and Tanh, both with BatchNorm1d) treats task weights as learnable parameters following the homoscedastic uncertainty framework of Kendall et al. [10]. A scalar log-variance s_i is initialised to zero for each task i and optimised jointly with the model parameters:

$$\mathcal{L}_{\text{uw}} = e^{-s_1}\,\mathcal{L}_{\text{title}} + s_1 + e^{-s_2}\,\mathcal{L}_{\text{target}} + s_2. \tag{2}$$

The factor e^{-s_i} down-weights tasks with higher estimated uncertainty, while s_i prevents the weights from collapsing to zero. The reparameterised variant in which $s_i = \log \sigma_i^2$ is optimised directly yields a regularisation term equivalent to $2 \log \sigma_i$, providing a slightly stronger penalty against uncertainty collapse compared to the original formulation. All 14 configurations are trained across 5 random seeds $\{42, 123, 7, 99, 2024\}$, and results are reported as mean $\pm$ standard deviation. Closed-Set Evaluation is done on the Manga109 test split, each model is evaluated by per-class precision, recall, F1-score (macro and weighted), and overall accuracy for both tasks. Open-set retrieval and target classification are used to assess generalization to unseen titles. The backbone is used as a fixed feature extractor on the 20 commercial titles described in Section 3. Given a query page, all remaining pages are ranked by cosine similarity. Reported metrics are classification accuracy (Acc), mean Average Precision (mAP), and Recall@1 (R@1). Two semantic diagnostics complement retrieval: *TargetConsistency@10* (TC_D) measures the fraction of top-10 results sharing the query's demographic label. Separability is quantified by the Silhouette coefficient, evaluated separately for title and demographic labels.

5 Experiments

This section reports the implementation details and training setup common to all 14 configurations. All models are trained for a maximum of 200 epochs. For single-task models, the checkpoint with the highest validation accuracy is retained; for multi-task models, the checkpoint maximising the average validation accuracy across both tasks is used. All experiments are repeated across five random seeds $\{7, 42, 99, 123, 2024\}$, with random states fixed for Python, NumPy, and PyTorch and deterministic CUDA kernels. All experiments are run on a single Quadro P5000 GPU with 16 GB of VRAM.

5.1 Closed-Set Classification

Table 2 reports test accuracy and macro-averaged F1 for demographic and title classification on Manga109. The best demographic accuracy is achieved by the fixed-sum multi-task configuration with ReLU activation and batch normalisation (MT-sum, ReLU, Norm; $\mathrm{Acc}_D = 90.4\%$), closely followed by the uncertainty-weighted ReLU variant (MT-unc, ReLU, Norm; $\mathrm{Acc}_D = 90.3\%$) and the unnormalised fixed-sum Tanh model (MT-sum, Tanh; $\mathrm{Acc}_D = 90.1\%$). All multi-task configurations outperform the strongest single-task demographic baseline (ST, Tanh, Norm; $\mathrm{Acc}_D = 86.7\%$) by up to 3.7 percentage points, indicating that joint supervision consistently strengthens the demographic signal. The gain is most pronounced for the Josei class, where multi-task models improve per-class accuracy by up to 9.4 points over the corresponding single-task network. For title classification, the highest accuracy is obtained by the single-task model with ReLU and batch normalisation (ST, ReLU, Norm; $\mathrm{Acc}_T = 82.1\%$), while the best multi-task configuration reaches 81.9% (MT-sum, ReLU, Norm),

Table 2: Closed-set classification results on Manga109. Metrics are reported as percentages (%). Standard deviations over 5 seeds are shown in subscripts. Acc_D and F1_D: demographic head (4 classes); Acc_T and F1_T: title head (109 classes). F1 scores are macro-averaged. Models are abbreviated (u=unc, s=sum, R=ReLU, T=Tanh, N=Norm). Single-task (ST) models are combined into single rows. Bold: best per column.

	Demographic Target		**Title**	
Model	Acc_D	F1_D	Acc_T	F1_T
Multi-task (MT)				
MT-u-R-N	$90.3_{\pm 1.3}$	$89.5_{\pm 1.5}$	$81.4_{\pm 0.4}$	$81.8_{\pm 0.3}$
MT-u-T-N	$89.8_{\pm 0.7}$	$88.8_{\pm 0.7}$	$81.2_{\pm 0.6}$	$81.4_{\pm 0.8}$
MT-s-R	$89.9_{\pm 1.1}$	$89.0_{\pm 1.0}$	$79.8_{\pm 0.6}$	$79.9_{\pm 0.8}$
MT-s-R-N	$\mathbf{90.4}_{\pm 0.9}$	$\mathbf{89.6}_{\pm 0.9}$	$81.9_{\pm 0.5}$	$82.1_{\pm 0.6}$
MT-s-T	$90.1_{\pm 0.6}$	$89.2_{\pm 0.8}$	$81.2_{\pm 1.0}$	$81.4_{\pm 1.2}$
MT-s-T-N	$89.5_{\pm 0.6}$	$88.5_{\pm 0.6}$	$81.3_{\pm 0.9}$	$81.5_{\pm 0.8}$
Single-task (ST)				
ST-R	$86.4_{\pm 0.8}$	$85.0_{\pm 1.0}$	$80.9_{\pm 0.4}$	$81.1_{\pm 0.6}$
ST-R-N	$86.5_{\pm 0.6}$	$85.5_{\pm 0.9}$	$\mathbf{82.1}_{\pm 1.6}$	$\mathbf{82.4}_{\pm 1.6}$
ST-T	$85.6_{\pm 0.7}$	$84.2_{\pm 0.8}$	$81.8_{\pm 0.6}$	$81.8_{\pm 0.9}$
ST-T-N	$86.7_{\pm 0.8}$	$85.4_{\pm 1.2}$	$81.5_{\pm 0.9}$	$81.7_{\pm 0.8}$

a gap of 0.2 percentage points. This shows that the demographic objective does not substantially harm title discrimination. Among multi-task variants, configurations with ReLU and batch normalisation yield the highest title accuracy, whereas Tanh without normalisation favours the demographic head, confirming the task-specific sensitivity introduced by activation and normalisation choices. The uncertainty-weighted configurations (MT-unc) achieve performance intermediate between the fixed-sum and single-task baselines on both heads, without a consistent advantage over the simpler fixed-sum strategy in the closed-set setting. The F1 scores follow the same trends and corroborate the conclusions drawn from accuracy. Training times are essentially identical across all configurations (≈311–313 min), confirming that architectural differences introduce negligible computational overhead.

5.2 Closed-Set Retrieval and Embedding Analysis

Retrieval quality is assessed directly in the embedding space produced by each model, without any re-ranking or post-processing. For each test image, the full test set minus the query itself serves as the gallery, and cosine similarity is used as the distance measure throughout. The retrieval ground truth is defined at the *title* level: a gallery page is considered relevant if and only if it belongs to the same title as the query. This choice reflects the primary objective of the embedding space, namely to group pages of the same manga title into com-

Table 3: Embedding-space retrieval and cluster quality on Manga109, evaluated on the closed-set test split. Retrieval metrics (mAP_T, R@1_T) are reported as percentages (%); silhouette scores (Sil_D, Sil_T) are in the $[-1, 1]$ range; TC_D denotes demographic target consistency at k=10. Standard deviations over 5 seeds are shown in subscripts. Models are abbreviated (u=unc, s=sum, R=ReLU, T=Tanh, N=Norm). Bold: best per column.

Model	mAP_T	R@1_T	Sil_D	Sil_T	TC_D
Multi-task (MT)					
MT-u-R-N	$67.2_{\pm 1.0}$	$80.3_{\pm 0.6}$	$0.123_{\pm 0.007}$	$0.272_{\pm 0.007}$	$0.872_{\pm 0.013}$
MT-u-T-N	$66.2_{\pm 0.9}$	$80.7_{\pm 1.3}$	$0.174_{\pm 0.006}$	$0.264_{\pm 0.006}$	$\mathbf{0.878}_{\pm 0.004}$
MT-s-R	$63.0_{\pm 0.7}$	$77.3_{\pm 0.9}$	$0.179_{\pm 0.008}$	$0.268_{\pm 0.008}$	$0.864_{\pm 0.011}$
MT-s-R-N	$68.6_{\pm 0.7}$	$80.9_{\pm 1.5}$	$0.124_{\pm 0.005}$	$0.287_{\pm 0.003}$	$0.874_{\pm 0.009}$
MT-s-T	$70.0_{\pm 1.6}$	$\mathbf{81.4}_{\pm 1.6}$	$0.154_{\pm 0.005}$	$0.338_{\pm 0.015}$	$0.876_{\pm 0.011}$
MT-s-T-N	$66.7_{\pm 0.7}$	$80.8_{\pm 1.3}$	$0.162_{\pm 0.005}$	$0.267_{\pm 0.009}$	$0.873_{\pm 0.010}$
Single-task – demographic (ST-D)					
ST-D-R	$10.8_{\pm 0.6}$	$21.7_{\pm 1.6}$	$\mathbf{0.515}_{\pm 0.014}$	$-0.400_{\pm 0.021}$	$0.834_{\pm 0.008}$
ST-D-R-N	$22.9_{\pm 0.7}$	$38.9_{\pm 1.7}$	$0.268_{\pm 0.017}$	$-0.108_{\pm 0.011}$	$0.839_{\pm 0.008}$
ST-D-T	$8.3_{\pm 0.3}$	$16.2_{\pm 1.1}$	$0.506_{\pm 0.015}$	$-0.485_{\pm 0.016}$	$0.822_{\pm 0.012}$
ST-D-T-N	$14.8_{\pm 0.7}$	$29.1_{\pm 1.4}$	$0.436_{\pm 0.012}$	$-0.278_{\pm 0.022}$	$0.842_{\pm 0.008}$
Single-task – title (ST-T)					
ST-T-R	$61.7_{\pm 1.6}$	$77.3_{\pm 0.9}$	$0.024_{\pm 0.004}$	$0.258_{\pm 0.013}$	$0.781_{\pm 0.012}$
ST-T-R-N	$67.2_{\pm 1.6}$	$80.4_{\pm 1.6}$	$0.021_{\pm 0.002}$	$0.268_{\pm 0.014}$	$0.804_{\pm 0.012}$
ST-T-T	$\mathbf{71.0}_{\pm 0.8}$	$\mathbf{81.4}_{\pm 1.0}$	$0.030_{\pm 0.003}$	$\mathbf{0.346}_{\pm 0.012}$	$0.836_{\pm 0.010}$
ST-T-T-N	$67.7_{\pm 1.4}$	$81.2_{\pm 1.0}$	$0.021_{\pm 0.002}$	$0.267_{\pm 0.013}$	$0.815_{\pm 0.015}$

pact, well-separated regions. As a consequence, single-task demographic models (ST-D) are evaluated under a criterion orthogonal to their training objective, which structurally limits their retrieval performance while producing high demographic silhouette scores (Sil_D). This asymmetry is intentional and informative: it quantifies the cost of discarding title supervision on the geometry of the learned representation. The silhouette scores Sil_D and Sil_T are computed on the full test embedding set using cosine distance, and measure the degree to which demographic and title clusters are internally cohesive and mutually separated, respectively.

Table 3 presents retrieval performance and silhouette scores evaluated on the closed-set test split. The highest mAP and R@1 are achieved by the single-task title model with Tanh activation (ST-T, Tanh; $\text{mAP}_T = 71.0\%$, $\text{R@1}_T = 81.4\%$). The best multi-task configuration for retrieval is MT-sum, Tanh ($\text{mAP}_T = 70.0\%$, $\text{R@1}_T = 81.4\%$), matching the single-task title model on R@1 and falling within 1.0 percentage point on mAP, confirming that joint training does not substantially degrade title-level discriminability in the embedding space. Single-task demographic models exhibit markedly lower retrieval performance,

with R@1 ranging from 16.2% (ST-D, Tanh) to 38.9% (ST-D, ReLU, Norm), reflecting the structural mismatch between a demographic training objective and a title-level retrieval criterion. The silhouette scores reveal a fundamental trade-off between the two semantic axes. Single-task title models yield the highest title silhouette (ST-T, Tanh; $\mathrm{Sil}_T = 0.346$) but near-zero demographic silhouette, indicating that title-level training produces an embedding space with no recoverable demographic structure. Single-task demographic models exhibit the opposite behaviour: they attain high demographic silhouette values (ST-D, ReLU; $\mathrm{Sil}_D = 0.515$) but strongly negative title silhouette scores (down to -0.485 for ST-D, Tanh), indicating that title-level clusters are effectively collapsed. Multi-task configurations occupy a consistent intermediate position on both axes: MT-sum, Tanh achieves $\mathrm{Sil}_T = 0.338$ and $\mathrm{Sil}_D = 0.154$, preserving meaningful structure along both semantic dimensions simultaneously. Regarding demographic target consistency (TC_D), multi-task models attain the highest values overall (0.864–0.878), with MT-unc, Tanh, Norm achieving the best score ($\mathrm{TC}_D = 0.878$), consistent with their joint demographic supervision. ST-D models score slightly lower (0.822–0.842) despite being trained exclusively on demographics, while ST-T models show the lowest consistency (0.781–0.836), as expected given the complete absence of a demographic training signal. With isolated exceptions, batch normalisation tends to reduce demographic silhouette in both ST-D and MT models, suggesting a regularisation effect that compresses class-specific variance in the embedding space.

5.3 Open-Set Retrieval and Demographic Classification

The open-set evaluation measures generalization to titles entirely absent from training, effectively framing a zero-shot retrieval and classification scenario. Queries and gallery pages belong to the previously described external dataset of commercial manga titles, ensuring complete separation from the Manga109 training domain. Retrieval ground truth is defined at the title level, using cosine similarity as the ranking function. Page-level demographic accuracy ($\mathrm{Acc}^{\mathrm{page}}_{D,os}$) is estimated without open-set supervision by applying the trained demographic classification head directly to each page embedding and computing the fraction of correctly classified pages; this metric is unavailable for ST-T (title) models.

Table 4 reports open-set retrieval performance, page-level demographic accuracy, and silhouette scores on the held-out titles. The best $\mathrm{mAP}_{T,os}$ is achieved by MT-unc, Tanh, Norm (26.2%), marginally ahead of ST-T, ReLU (26.1%) and ST-T, ReLU, Norm (25.9%). For instance-level retrieval, ST-T, ReLU, Norm achieves the highest $\mathrm{R@1}_{T,os}$ (67.9%). Batch normalisation consistently improves open-set retrieval within the ST-T family: $\mathrm{R@1}_{T,os}$ increases from 64.0% to 67.9% for ReLU and from 60.3% to 67.2% for Tanh, confirming that normalising the embedding distribution improves generalization to unseen titles. A similar but more pronounced effect is observed for ST-D models, where batch normalisation raises $\mathrm{R@1}_{T,os}$ from 36.2% to 54.0% (ReLU) and from 33.7% to 46.9% (Tanh), suggesting that the regularisation benefit is larger when the training

Table 4: Open-set retrieval and cluster quality on unseen commercial manga titles. Retrieval metrics ($\text{mAP}_{T,os}$, $\text{R@1}_{T,os}$) and demographic page accuracy ($\text{Acc}^{\text{page}}_{D,os}$) are reported as percentages (%); silhouette scores ($\text{Sil}_{D,os}$, $\text{Sil}_{T,os}$) are in the $[-1, 1]$ range; $\text{TC}_{D,os}$ denotes open-set demographic target consistency at k=10. Standard deviations over 5 seeds are shown in subscripts. Models are abbreviated (u=unc, s=sum, R=ReLU, T=Tanh, N=Norm). ST-T models have no demographic head (–). Bold: best per column.

Model	$\text{mAP}_{T,os}$	$\text{R@1}_{T,os}$	$\text{Sil}_{D,os}$	$\text{Sil}_{T,os}$	$\text{Acc}^{\text{page}}_{D,os}$	$\text{TC}_{D,os}$
Multi-task (MT)						
MT-u-R-N	$25.8_{\pm 1.0}$	$66.4_{\pm 0.4}$	$0.035_{\pm 0.007}$	$0.013_{\pm 0.008}$	$48.4_{\pm 1.1}$	$0.740_{\pm 0.010}$
MT-u-T-N	$\mathbf{26.2}_{\pm 0.3}$	$67.2_{\pm 1.1}$	$0.038_{\pm 0.003}$	$0.013_{\pm 0.004}$	$49.5_{\pm 1.2}$	$0.751_{\pm 0.005}$
MT-s-R	$24.9_{\pm 0.7}$	$61.0_{\pm 1.3}$	$0.048_{\pm 0.006}$	$0.008_{\pm 0.007}$	$50.0_{\pm 0.8}$	$0.717_{\pm 0.009}$
MT-s-R-N	$25.3_{\pm 0.6}$	$65.6_{\pm 1.1}$	$0.037_{\pm 0.005}$	$0.016_{\pm 0.005}$	$48.6_{\pm 0.8}$	$0.737_{\pm 0.009}$
MT-s-T	$23.6_{\pm 0.8}$	$59.0_{\pm 1.1}$	$0.043_{\pm 0.005}$	$-0.005_{\pm 0.005}$	$\mathbf{50.6}_{\pm 1.5}$	$0.710_{\pm 0.009}$
MT-s-T-N	$25.1_{\pm 0.7}$	$64.9_{\pm 1.3}$	$0.039_{\pm 0.006}$	$0.009_{\pm 0.010}$	$49.2_{\pm 2.0}$	$0.746_{\pm 0.003}$
Single-task – demographic (ST-D)						
ST-D-R	$12.4_{\pm 0.3}$	$36.2_{\pm 1.9}$	$0.010_{\pm 0.014}$	$-0.229_{\pm 0.012}$	$48.7_{\pm 1.2}$	$0.571_{\pm 0.011}$
ST-D-R-N	$19.1_{\pm 0.5}$	$54.0_{\pm 1.1}$	$0.040_{\pm 0.008}$	$-0.058_{\pm 0.011}$	$50.0_{\pm 3.1}$	$0.677_{\pm 0.003}$
ST-D-T	$11.8_{\pm 0.5}$	$33.7_{\pm 0.8}$	$0.004_{\pm 0.012}$	$-0.289_{\pm 0.014}$	$49.6_{\pm 1.6}$	$0.547_{\pm 0.005}$
ST-D-T-N	$14.9_{\pm 0.4}$	$46.9_{\pm 0.2}$	$0.025_{\pm 0.016}$	$-0.168_{\pm 0.017}$	$49.4_{\pm 3.1}$	$0.640_{\pm 0.010}$
Single-task – title (ST-T)						
ST-T-R	$26.1_{\pm 0.7}$	$64.0_{\pm 0.9}$	$\mathbf{0.049}_{\pm 0.007}$	$0.022_{\pm 0.008}$	–	$0.741_{\pm 0.006}$
ST-T-R-N	$25.9_{\pm 0.3}$	$\mathbf{67.9}_{\pm 0.8}$	$0.031_{\pm 0.004}$	$\mathbf{0.027}_{\pm 0.005}$	–	$0.747_{\pm 0.006}$
ST-T-T	$24.0_{\pm 0.5}$	$60.3_{\pm 0.7}$	$0.033_{\pm 0.002}$	$0.007_{\pm 0.007}$	–	$0.706_{\pm 0.005}$
ST-T-T-N	$25.9_{\pm 0.5}$	$67.2_{\pm 0.6}$	$0.035_{\pm 0.004}$	$0.018_{\pm 0.008}$	–	$\mathbf{0.758}_{\pm 0.008}$

objective is misaligned with the retrieval criterion. Notably, the closed-set ranking is partially reversed in the open-set scenario: ST-T, Tanh, which achieved the best closed-set mAP (71.0%), drops to 24.0% in the open-set evaluation, while its normalised counterpart generalises more effectively (25.9%). For page-level demographic classification, the highest accuracy is obtained by MT-sum, Tanh ($\text{Acc}^{\text{page}}_{D,os} = 50.6\%$), followed by ST-D, ReLU, Norm and MT-sum, ReLU (50.0%). The differences among the top configurations are within one standard deviation, indicating that demographic classification on an entirely unseen domain is a highly challenging zero-shot task where architectural choices have limited impact. All models with a demographic head cluster in a narrow range (48.4%–50.6%), suggesting that the learned proxy centroids provide a consistent but modest signal for zero-shot generalisation. The open-set silhouette scores are substantially lower than their closed-set counterparts across all configurations. Title silhouettes for ST-T models remain slightly positive (0.007–0.027), while ST-D models yield negative title silhouettes (down to −0.289 for ST-D, Tanh), replicating the closed-set pattern at reduced magnitude. Nearly all multi-task

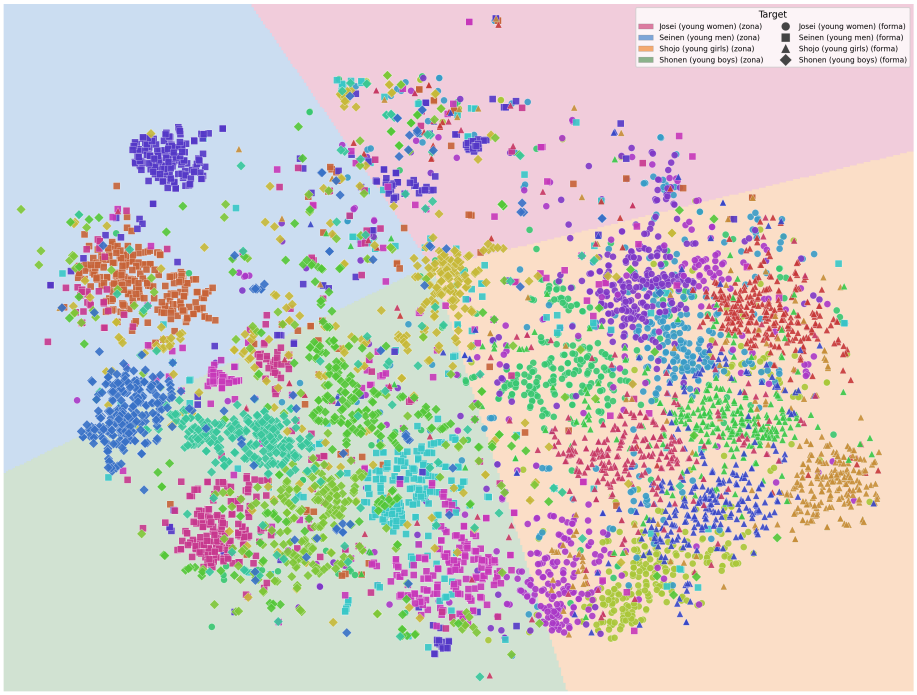
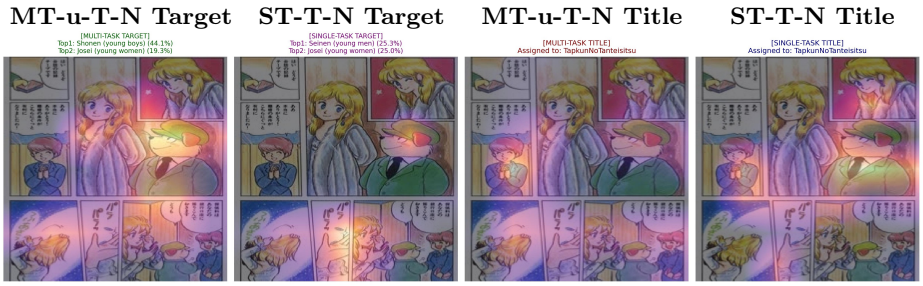


Fig. 1: t-SNE projection of the open-set embedding space for the MT-unc model. Points represent individual unseen titles (colored by title, shaped by ground-truth demographic); background Voronoi regions indicate demographic decision boundaries.

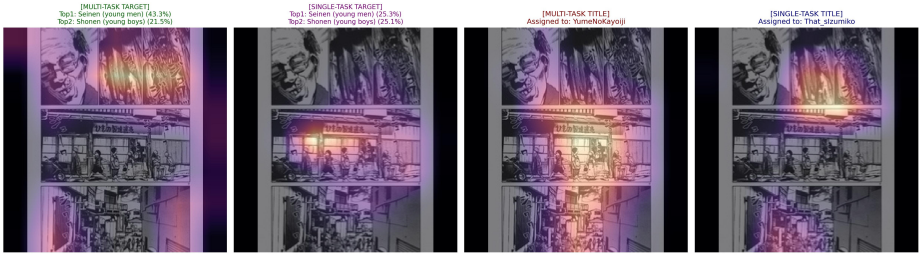
models maintain both silhouettes in the positive range, though the values are small, indicating that cluster cohesion attenuates when generalising to unseen titles. Open-set demographic target consistency ($TC_{D,os}$) follows a pattern that partially diverges from the closed-set results: ST-T, Tanh, Norm achieves the highest score overall ($TC_{D,os} = 0.758$), marginally ahead of MT-unc, Tanh, Norm (0.751), while ST-D models drop substantially (0.547–0.677). The relatively strong performance of ST-T models on this metric suggests that title-supervised embeddings generalise better demographically to unseen domains, likely because title and demographic labels are correlated in the training distribution, allowing title-level representations to implicitly capture demographic structure that transfers to new titles. Overall, these results underscore the inherent complexity of zero-shot manga representation: while MTL preserves robustness better than demographic-only models, the performance gap between closed-set and open-set scenarios highlights the critical need for further research in domain-generalised manga recommendation.

5.4 Qualitative Analysis



(a) Representative Case (Low Entropy) - Closed-set: Shonen title *TapkunNoTanteisitsu*.

(b) Ambiguous Case (High Entropy) - Closed-set: Josei title *Raphael*.



(c) Generalization Case - Open-set: Seinen title *Oyasumi Punpun*.

Fig. 2: Score-CAM visualization comparing **MT-u-T-N** (Multi-task, uncertainty + tanh + norm) and **ST-T-N** (Single-task, tanh + norm) models across demographic (Target) and Title heads. MT-u-T-N models demonstrate a more focused and semantically consistent attention compared to ST-T-N counterparts, especially in ambiguous and open-set scenarios.

In this section, visual evidence is used to analyze the mechanisms driving the observed performance trends. The impact of batch normalisation and joint supervision is most evident in open-set scenarios. As shown in the t-SNE projection of unseen titles (Figure 1), the multi-task model organises the embedding space into four broadly separable demographic regions even without prior exposure to these titles, suggesting that joint supervision induces a degree of semantic

organisation that generalises beyond the training domain. However, title-level cluster cohesion remains limited, consistent with the near-zero open-set silhouette scores reported in Table 4. Figure 2 compares Score-CAM[21] activations for MT and ST models across three scenarios. In the representative closed-set case (Figure 2a), the MT-Target head concentrates on the protagonist while ST-Target disperses activation across background textures. In the ambiguous case (Figure 2b, *Raphael*), ST-Title fragments attention over the bottom landscape, whereas MT-Title localises on the character's profile; the MT-Target head similarly suppresses background regions that mislead the ST-Target model. In the open-set case (Figure 2c, *Oyasumi Punpun*), MT-Target correctly predicts Seinen (43.3%) against a near-uniform ST-Target distribution (25.3% Seinen, 25.1% Shonen), with activation concentrated on the characters' faces. Neither model recovers the correct title identity, as expected given that both titles are unseen during training; MT-Title nonetheless attends to character regions while ST-Title activates on the urban background.

5.5 Limitations and Future Work

Limitations of this study include the reliance on the ResNet-50 architecture and the structural imbalance of the Manga109 dataset, which poses significant challenges for minority demographic classes. Consequently, the reported performance metrics represent an empirical baseline rather than a fully optimized bound. These constraints reflect broader difficulties in the manga domain, where data scarcity and copyright restrictions often hinder the collection of large-scale standardized datasets. Future research should investigate more modern backbones, such as Vision Transformers, to capture stylistic complexities that may elude traditional convolutional networks. The use of non-Euclidean latent spaces, particularly Poincaré embeddings, could better model the hierarchical relationships between specific titles and broader demographic targets. Additionally, specialized objectives such as contrastive or triplet losses may further improve cluster cohesion and embedding quality. Extending the multi-task framework to include additional attributes like genre or publication era, while exploring alternative loss weighting strategies, can further advance scalable manga visual analysis.

6 Conclusions

A systematic evaluation of fourteen ResNet-50 configurations demonstrates that joint optimisation of demographic and title objectives consistently strengthens closed-set demographic classification, with the best multi-task model reaching $\mathrm{Acc}_D = 90.4\%$ and $\mathrm{Acc}_T = 81.9\%$ simultaneously, a gain of 3.7 pp over the strongest single-task demographic baseline. In embedding-space retrieval on seen titles, the best multi-task configuration achieves $\mathrm{mAP}_T = 70.0\%$ and $\mathrm{R@1}_T = 81.4\%$, within 1.0 pp of the best title-specialised model on mAP_T and matching it on $\mathrm{R@1}_T$, while producing latent spaces with positive silhouette scores along both the demographic and title axes simultaneously, a property no

single-task model achieves. In open-set retrieval, multi-task embeddings reach $R@1_{T,os} = 67.2\%$, within one standard deviation of the best title-specialised baseline (67.9%), and outperform the best demographic-only model by 13.2 pp, suggesting that dual-objective supervision encourages more transferable visual features. Open-set demographic page accuracy clusters in a narrow range across all configurations (48.4%–50.6%, differences within one standard deviation), indicating that architectural choices have limited impact on zero-shot demographic generalization at this scale. The closed-set to open-set gap remains substantial: mAP_T drops from 71.0% (best closed-set model, ST-T-T) to 26.2% (best open-set model, MT-u-T-N), and title-level silhouette scores collapse toward zero under domain shift, confirming that cluster structure largely dissolves on unseen titles. Future work should explore transformer-based backbones, contrastive pretraining, and open-set metric learning to reduce this gap and support scalable manga indexing and recommendation.

References

1. Aizawa, K., Fujimoto, A., Otsubo, A., Ogawa, T., Matsui, Y., Tsubota, K., Ikuta, H.: Building a manga dataset "Manga109" with annotations for multimedia applications. IEEE MultiMedia **27**(2), 8–18 (2020). https://doi.org/10.1109/MMUL.2020.2987895
2. Bahaweres, R.B., Almujaddidi, A.R.: Improving recommender systems performance with cross-domain scenario: Anime and manga domain studies. In: Proceedings of the 9th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI). pp. 361–365. IEEE (2022). https://doi.org/10.23919/EECSI56542.2022.9946560
3. Battiato, S., Gallo, G., Impoco, G., Stanco, F.: An efficient re-indexing algorithm for color-mapped images. IEEE Transactions on Image Processing **13**(11), 1419–1423 (2004)
4. Fargetta, G., Anile, S., Priano, L., Spata, M.O., Ortis, A., Battiato, S.: Bridging ai and wildlife conservation: Classifying wildcats from genetically validated images. Ecological Informatics **92**, 103467 (2025). https://doi.org/https://doi.org/10.1016/j.ecoinf.2025.103467
5. Fargetta, G., Fiorito, E.R., Licciardello, R., Vivera, P., Ortis, A., Battiato, S.: You only look at manga faces: A YOLOv12-based approach on manga datasets and benchmark. PeerJ Computer Science (2026)
6. Fargetta, G., Ortis, A., Anile, S., Battiato, S.: Evaluation of cnns for wildcats classification in real world scenario. In: Garcia, F.P., Jamil, A., Hameed, A.A., Ortis, A., Ramirez, I.S. (eds.) Recent Trends and Advances in Artificial Intelligence. pp. 15–25. Springer Nature Switzerland, Cham (2024)
7. Feng, S., Yoshinaga, T., Hayashi, K., Washio, K., Kamigaito, H.: How panel layouts define manga: Insights from visual ablation experiments. In: Proceedings of the Annual Meeting of the Cognitive Science Society. vol. 47 (2025)
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778 (2016)
9. Huang, J., Mumford, D.: Statistics of natural images and models. In: Proceedings. 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No PR00149). vol. 1, pp. 541–547. IEEE (1999)

10. Kendall, A., Gal, Y., Cipolla, R.: Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7482–7491 (2018)
11. Kim, B.H., Lee, C.s., Lim, S.K.: A study on a webtoon recommendation system with autoencoder and domain-specific finetuning BERT with synthetic data considering art style and synopsis. International Journal on Advanced Science, Engineering and Information Technology **14**(6), 1944–1950 (2024). https://doi.org/10.18517/ijaseit.14.6.20443
12. Manzoor, A., Fargetta, G., Ortis, A., Battiato, S.: Shieldnet: A novel adversarially resilient convolutional neural network for robust image classification. Applied Sciences **16**(3) (2026). https://doi.org/10.3390/app16031254
13. Matsui, Y., Ito, K., Aramaki, Y., Fujimoto, A., Ogawa, T., Yamasaki, T., Aizawa, K.: Sketch-based manga retrieval using manga109 dataset. Multimedia tools and applications **76**, 21811–21838 (2017)
14. Nguyen, N.V., Rigaud, C., Burie, J.C.: Comic MTL: optimized multi-task learning for comic book image analysis. International Journal on Document Analysis and Recognition (IJDAR) **22**(3), 265–284 (2019). https://doi.org/10.1007/s10032-019-00330-3
15. Nguyen, N.V., Rigaud, C., Revel, A., Burie, J.C.: Manga-MMTL: Multimodal multitask transfer learning for manga character analysis. In: Document Analysis and Recognition – ICDAR 2021. Lecture Notes in Computer Science, vol. 12822, pp. 410–425. Springer (2021). https://doi.org/10.1007/978-3-030-86331-9_27
16. Sharma, R., Kukreja, V.: Image segmentation, classification and recognition methods for comics: A decade systematic literature review. Engineering Applications of Artificial Intelligence **131**, 107715 (2024)
17. Soni, B., Thakuria, D., Nath, N., Das, N., Boro, B.: RikoNet: A novel anime recommendation engine. Multimedia Tools and Applications **82**, 32329–32348 (2023). https://doi.org/10.1007/s11042-023-14710-9
18. Vaze, S., Han, K., Vedaldi, A., Zisserman, A.: Open-set recognition: A good closed-set classifier is all you need. In: Proceedings of the International Conference on Learning Representations (ICLR) (2022)
19. Vie, J.J., Yger, F., Lahfa, R., Clement, B., Cocchi, K., Chalumeau, T., Kashima, H.: Using posters to recommend anime and mangas in a cold-start scenario. In: Proceedings of the 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). vol. 3, pp. 21–26. IEEE (2017). https://doi.org/10.1109/ICDAR.2017.287
20. Vu, T.: [ee331] manga classification. https://kaggle.com/competitions/manga (2019), kaggle
21. Wang, H., Wang, Z., Du, M., Yang, F., Zhang, Z., Ding, S., Mardziel, P., Hu, X.: Score-cam: Score-weighted visual explanations for convolutional neural networks. In: Proceedings - 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, CVPRW 2020. pp. 111–119. IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, IEEE Computer Society, United States (Jun 2020). https://doi.org/10.1109/CVPRW50498.2020.00020
22. Young-Min, K.: Feature visualization in comic artist classification using deep neural networks. Journal of Big Data **6**(1), 56 (2019)
23. Zuccarà, R., Fargetta, G., Ortis, A., Battiato, S.: Exploiting adversarial learning and topology augmentation for open-set visual recognition. In: 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 3416–3424 (2025). https://doi.org/10.1109/CVPRW67362.2025.00326