Spotting the Unseen: Reciprocal Consensus Network Guided by Visual Archetypes

Wenbo Hu[1,2], **Hongjian Zhan**[1], **Xinchen Ma**[1], **Yue Lu**[1*], **Ching Y. Suen**[2]

[1]Shanghai Key Laboratory of Multidimensional Information Processing, Shanghai, China
[2]Centre for Pattern Recognition and Machine Intelligence, Concordia University, Montreal, Canada
{wenbo,xcma}@stu.ecnu.edu.cn, {hjzhan, ylu}@cee.ecnu.edu.cn, suen@cse.concordia.ca

Abstract

Humans often require only a few visual archetypes to spot novel objects. Based on this observation, we present a strategy rooted in "spotting the unseen" by establishing dense correspondences between potential query image regions and a visual archetype, and we propose the Consensus Network (CoNet). Our method leverages relational patterns intra and inter images via Auto-Correlation Representation (ACR) and Mutual-Correlation Representation (MCR). Within each image, the ACR module is capable of encoding both local self-similarity and global context simultaneously. Between the query and support images, the MCR module computes the cross-correlation across two image representations and introduces a reciprocal consistency constraint, which can incorporate to exclude outliers and enhance model robustness. To overcome the challenges of low-resource training data, particularly in one-shot learning scenarios, we incorporate an adaptive margin strategy to better handle diverse instances. The experimental results indicate the effectiveness of the proposed method across diverse domains such as object detection in natural scenes, and text spotting in both historical manuscripts and natural scenes, which demonstrates its sparkling generalization ability. Our code is available at: https://github.com/infinite-hwb/conet.

Introduction

The need to "spot the unseen" is prevalent in real-world scenarios, such as historical document analysis, where researchers might encounter previously unrecorded characters. With the advances of deep learning over the past decades, while state-of-the-art detection or recognition models have showcased impressive performance, they still have a fundamental limitation: under the closed-world assumption, they can only predict classes observed and annotated in their training dataset (Han et al. 2022; Ren et al. 2015; Liu et al. 2015; Zhou 2022). Ideally, we would like the object detection model to be able to detect all objects in the world.

To tackle this challenge, several object detection methods rooted in open-set learning have emerged (Neal et al. 2018; Huang et al. 2022; Jeong, Choi, and Kim 2021; Joseph et al. 2021; Han et al. 2022; Wu et al. 2022). In practical applications, these methods label unseen categories as "unknown", offering the advantage of automatically discovering unseen categories but failing to categorize them further. Another drawback of these approaches is their reliance on vast amounts of training data, a requirement that isn't always feasible in certain application scenarios. Traditional sliding-window-based spotting methods (Wicht, Fischer, and Hennebert 2016; Chan and Riek 2020) though seemingly promising for addressing the open recognition issue, often underperform in practice. This underperformance can likely be attributed to variations in object morphology across different categories and the breakdown of data distribution assumptions when faced with novel categories.

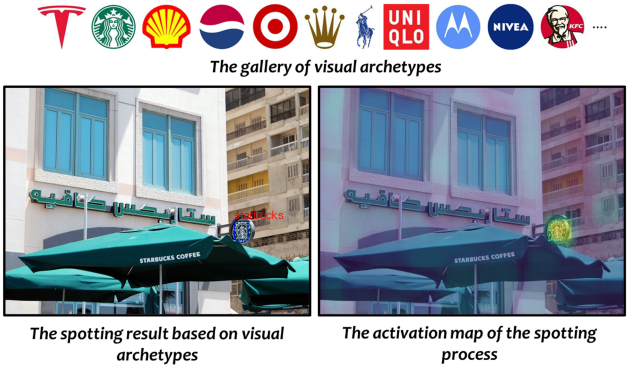

Figure 1: Given a scene image and a set of visual archetypes, our goal is to perform accurate spotting in the scene image based on these visual archetypes.

More recent approaches have leveraged one-shot or few-shot learning for object detection (Fan et al. 2019; Hsieh et al. 2019; Osokin, Sumin, and Lomakin 2020; Zhang et al. 2020; Cheng, Wang, and Long 2021; Jiang et al. 2023). Yet, these methods frequently face challenges when dealing with images of closely related categories. In scenarios like text spotting in historical documents or logo spotting in natural scenes, the novel classes might be mistaken for confusable categories. Moreover, utilizing deep learning in low-data regimes often results in a diminished quality of the feature space. Acknowledging these limitations in existing techniques, we are compelled to seek a novel approach better equipped to manage such intricacies.

*Corresponding author.

In this work, we propose the "***Co**nsensus **Net**work*" (CoNet) to spot the unseen categories that capitalize on the benefits of one-shot learning while alleviating its limitations. Instead of leaning on class-specific knowledge, we harness visual archetypes to discern dense correspondences between potential query regions in images and these archetypes. This approach offers greater flexibility: for spotting new categories, we can simply augment the system with additional visual archetypes. The concept of "consensus" is manifest in how CoNet arrives at an agreement on spotting objects through visual archetypes, as elegantly illustrated in Figure 1. While most prior research has focused on object detection, there is a noticeable lack of studies addressing the detection of unknown texts. Our strategy not only excels in conventional areas such as natural scene object detection but also pioneers advancements in under-explored domains like historical manuscript analysis and natural scene text detection. Our work's contributions can be summarized as follows:

- We present an auto-correlation representation approach that effectively integrates local features and contextual information, significantly boosting precision in one-shot detection tasks and reducing false positives.
- We propose a mutual-correlation representation module, designed to facilitate the learning of robust feature alignments, markedly enhancing our model's ability to detect unseen categories.
- The model's adaptability and excellence are validated across various scenarios. It outperforms the state-of-the-art on multiple benchmarks, with its robust components confirmed by ablation studies.

Related Work

Open-Set Learning

Open-Set Learning (OSL) addresses knowledge gaps during training by aiming to identify samples from seen classes and distinguish those from unseen classes during testing (Bendale and Boult 2016). Recent strides in OSL have deployed pseudo-unknown sample generation (Neal et al. 2018; Huang et al. 2022; Pal et al. 2022) or prototype-based methodologies (Joseph et al. 2021; Han et al. 2022; Wu et al. 2022) to recognize unknown classes. The challenge intensifies for open-set learning in object detection, particularly in low-resource scenarios, where limited training samples can lead to model overfitting and compromise performance. Moreover, recent zero-shot-based methods, viewed as a form of OSL (Wang et al. 2019a; Cao et al. 2020; Zhang, Du, and Dai 2020; Chen, Li, and Xue 2021), achieve open-set recognition by leveraging strokes and structural patterns of Chinese characters. However, these methods are primarily for single-character recognition, not page-level analysis, and extending them to other language recognition tasks has proven to be challenging.

Few-Shot/One-Shot Object Detection

For few-shot/one-shot object detection techniques, which align more closely with our task, excel at identifying unseen classes with merely a handful of labeled support samples. In recent years, unique applications of few-shot learning have emerged, such as logo recognition (Bhunia et al. 2019) and sketch-guided object localization in natural images (Tripathi et al. 2020), underscoring the versatile potential of few-shot techniques. LSTD (Chen et al. 2018), a method grounded in transfer learning, mitigates overfitting through knowledge migration from source to target domain, supplemented by regularization techniques. FSOD (Fan et al. 2019) deploys an attention mechanism to generate a cross-correlation map and, subsequently, class-specific proposals. CoAE (Hsieh et al. 2019), on the other hand, employs non-local blocks for co-attention implementation, achieving one-shot object detection by spotlighting correlated feature channels. Our approach also leverages attention mechanisms to enhance the model's discriminative capabilities. OS2D (Osokin, Sumin, and Lomakin 2020) establishes correspondences via dense correlation matching of features and advocates for a feed-forward geometric transformation model, realized through full convolution, for feature alignment. Mirroring OS2D, our method employs a similar geometric matching-based framework but digs deeper, capturing more profound semantic information from limited samples. We also introduce consistency-constrained strategies for filtering bias and noise, resulting in more accurate detection results. Moreover, we dynamically manage positive and negative samples and apply adaptive margin loss for optimal training.

Proposed Method

The architecture of CoNet is shown in Figure 2, which is composed of three core modules. In this section, we present the technical details of each module and subsequently describe our learning strategy.

Auto-Correlation Representation (ACR)

Given a pair of images, namely a query image, I_q, and a support image, I_s, our shared backbone feature extractor produces base representations, denoted as $A_q \in \mathbb{R}^{C\times H_s\times W_s}$ and $A_s \in \mathbb{R}^{C\times H_s\times W_s}$. The support image can be perceived as a visual archetype, whereas the query image is the target of our detection or search process. Within this context, we propose a novel ACR module. This module encodes the semantic information of local regions into a vector, thereby generating a global context feature map. The ACR module takes the base representation[1] and transforms it to emphasize the more relevant regions in an image.

Local Correlation Computation. Drawing on (Huang et al. 2019), we propose to calculate local correlations. We define a local neighborhood of size (k, k) around each position in the feature map $\hat{A}_{ij}$ where $\hat{A}_{ij} \in C \times 1 \times 1$. This approach allows us to evaluate the similarity between each spatial position and its specific local neighborhood. The defined region is denoted as $l_{ij}^{k\times k} \in \mathbb{R}^{C\times k\times k}$, where k represents the local neighborhood's size. To accommodate border locations in the computation, we apply zero-padding of size $(k-1)/2$ to the feature map A, resulting in the padded

[1]For notational simplicity, we omit subscripts q and s of A_q and A_s in this section.

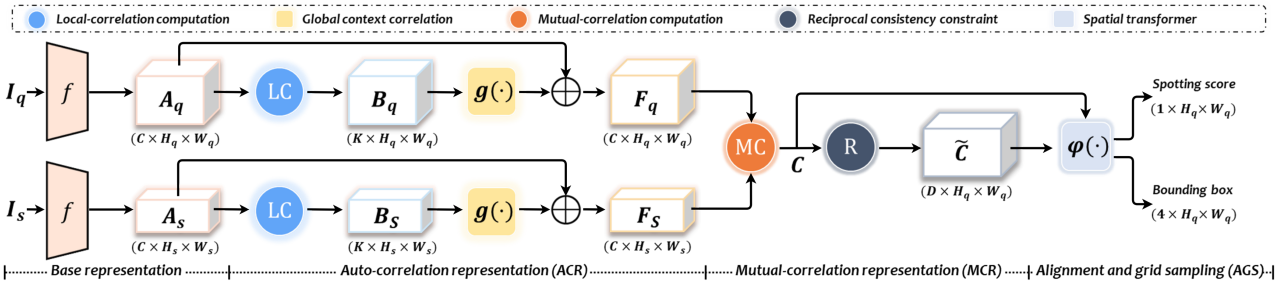

Figure 2: CoNet Architecture Overview. The ACR module processes base representations, transforming them into local self-similarity tensors and refining to auto-correlation representations. The MCR module calculates the mutual-correlation between image pairs and ensures reciprocal consistency. The AGS module then aligns these features, leading to the final results.

map $\hat{A} \in \mathbb{R}^{C\times(H+k-1)\times(W+k-1)}$. Subsequently, we calculate the dot product between $\hat{A}ij$ and each neighbor's position in the local region $L_{ij}^{k\times k}$, giving rise to the local auto-correlation representation $\vec{v}_{ij} \in \mathbb{R}^{k^2\times 1}$. In particular, the local correlation map L is computed as $L = \{\vec{v}_{11}, ..., \vec{v}_{HW}\}$, where $L \in \mathbb{R}^{k^2\times H\times W}$. To finalize the local correlation computation, we combine the local correlation map with the base feature representation, forming the cascaded visual representation $B \in \mathbb{R}^{K\times H\times W}$, where K denotes the dimension of feature B and is given by $K = k^2 + C$.

Global Context Correlation. To distill richer contextual semantics and auto-correlation patterns from feature B, we apply a series of procedures, including a 1×1 convolution and softmax function to calculate attention weights, along with global attention to extract global context features, as shown in Figure 3(a). Subsequently, we employ a bottleneck transform to capture channel-wise dependencies and perform an element-wise addition for feature fusion, yielding the auto-correlation representation $F = A+g(B)$, where $F \in \mathbb{R}^{C\times H\times W}$. This approach augments the base features with contextual correlations, guiding the module on "what to observe" within the image.

Mutual-Correlation Representation (MCR)

The MCR module takes an input pair of query and support forms from ACRs, represented as F_q and F_s, and produces a correlation map $\widetilde{C}$ to enhance semantic alignment.

Mutual-Correlation Computation. We utilize the global context correlation features from the ACRs to establish a dense mutual correlation, thereby enriching the correspondence relations by integrating data from both the query and support images. For the individual features F_q and F_s, we form a 4-dimensional correlation tensor $C \in \mathbb{R}^{H_q\times W_q\times H_s\times W_s}$ that captures the similarities across all spatial location pairs, as described below:

$$C_{qs}(x, y) = \phi(F_q(x), F_s(y)), \tag{1}$$

where x and y denote the positions of the feature representations in input images I_q and I_s respectively, and $\phi(\cdot, \cdot)$ represents the cosine similarity between these features.

Reciprocal Consistency Constraint. The correlation tensor C gauges similarities at various positions, essentially acting as a metric for direct matching probability. Sometimes, due to the minimal inter-class distance among image classes, the query image displays multiple response regions, which can result in inaccurate spotting. In general, similarity exhibits a symmetric behavior during the matching phase. The generated correspondence should satisfy a one-to-one mapping constraint, computing the associations $x \rightarrow y$ and its reverse $y \rightarrow x$. To this end, we apply the consistency constraint to weight all candidate matching points and initially identify the priority matching points. The matching weight of a specific point x_i in F_q to any point y in F_s is $\omega_{q\rightarrow s} = \frac{C_{qs}(x_i,y)}{max[C_{qs}(x_i,y)]}$. Similarly, the matching weight of a particular point y_j in F_s to any point x in F_q is $\omega_{s\rightarrow q} = \frac{C_{qs}(x,y_j)}{max[C_{qs}(x,y_j)]}$. By using this constraint, the correlation map C is refined to facilitate the selection of the most probable matches and assist in inferring pixel-level correspondences. This updated correlation map, denoted by $\widetilde{C}_{qs}$, filters outliers and augments the network's fitting capability:

$$\widetilde{C} = \omega_{q\rightarrow s} \cdot \omega_{s\rightarrow q} \cdot C_{qs}. \tag{2}$$

The correlation map $\widetilde{C}$, containing scores from all pairwise comparisons, offers rich contextual information. For a candidate pair (x_i, y_j), if they obey the reciprocal consistency constraint, then $\omega_{q\rightarrow s} = \omega_{s\rightarrow q} \approx 1$. If not, the product $\omega_{q\rightarrow s} \cdot \omega_{s\rightarrow q}$ becomes zero.

Alignment and Grid Sampling (AGS)

After obtaining the correlation map $\widetilde{C}$, we employ the AGS module to establish the spatial correspondence between the support and the query image. Within the AGS, we integrate the Convolutional Block Attention Module (CBAM) (Woo et al. 2018). It adaptively recalibrates feature maps both in channel-wise and spatial dimensions, emphasizing the most informative features. Following the CBAM, we introduce two successive convolutional layers, building upon the idea presented by (Rocco, Arandjelovic, and Sivic 2018) that further parses the context and correlations among features. These layers, set with a stride of 1 and without padding, incorporate batch normalization and ReLU activation. Lastly, as depicted in Fig. 3(b), we append a final fully connected layer, which regresses to the transformation parameters θ.

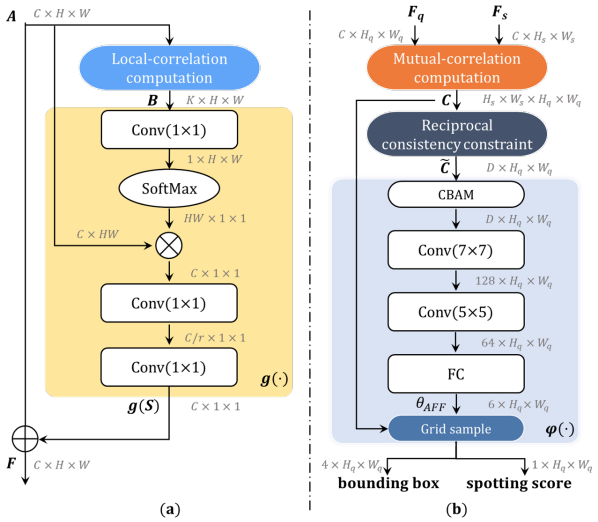

Figure 3: The architectures of ACR and AGS modules. (a) The ACR module encodes information B from local regions into a global context map F; (b) The MCR module refines mutual-correlation representation C, then is processed by the AGS module for final results.

Upon deriving the affine transformation parameters θ, we use a grid sampler to establish a point grid on the query image, ensuring alignment with the support image. This grid, once formed, serves as a tool for the resampling of correlation maps, which elucidates the feature similarities between the query and support images. The values extracted from this resampling process, in essence, represent match scores that shed light on the spatial resemblance between the two images. Following the extraction of these matching scores, a class pooling mask is used to screen and average them to derive the final spotting scores. After spatial alignment, we generate a grid of default bounding boxes across the entire image space, which are then transformed by the transformation parameters θ to produce candidate bounding boxes.

Adaptive Margin Contrastive Learning

Inspired by the studies (Chopra, Hadsell, and LeCun 2005; Wang et al. 2019b; Osokin, Sumin, and Lomakin 2020), we adopt a margin-based contrastive loss function. However, in scenarios with insufficient data, our method faces the challenge of limited training samples, constraining the model's ability to generalize to unseen data. In light of this challenge, we introduce a contrastive loss complemented with an adaptive margin. This adaptive margin distinctively adjusts the positive margin m_{pos} and negative margin m_{neg} for every individual positive s_i^{pos} and negative sample s_i^{neg}:

$$\begin{aligned} m_{pos}^{adapt} &= m_{pos} + [exp(-\mu(1 - s_i^{pos}))]/\mu, \\ m_{neg}^{adapt} &= m_{neg} - [exp(-\mu(s_i^{neg}))]/\mu, \end{aligned} \tag{3}$$

where $s_i \in [-1, 1]$ is the spotting score, trained to be high for positives and low for negatives, with the initial constraints m_{pos} and m_{neg} ($m_{pos} > m_{neg}$) and μ regulating the dynamic margin.

By adopting this adaptive margin approach, we enhance the discriminative power of our model. High spotting scores for positive instances elevate the positive margin, while high spotting scores for negative instances reduce the negative margin. This forces positive scores above and negative scores below their respective margins, cultivating a wide separation. This adaptive margin accommodates samples of varying difficulty levels, empowering the model to fine-tune its approach to handle both complex and straightforward samples. It allows us to define the hinge-embedding loss with adaptive margins for positive and negative instances, expressed as, $l_{pos}^i = max(m_{pos}^{adapt} - s_i^{pos}, 0)$, and $l_{neg}^i = max(s_i^{neg} - m_{neg}^{adapt}, 0)$. The contrastive loss function with the adaptive margin can be stated as:

$$L_c = \sum_i (l_{pos}^i + w_i l_{neg}^i). \tag{4}$$

Similar to the ranked list loss (Wang et al. 2019b), w_i is defined as $w_i = exp(T \cdot (s_i^{neg} - m_{neg}))$, $s_i - m_{neg} > 0$. Our adaptive margin approach lies in its dynamic adaptability to the unique characteristics of each example, efficiently addressing the issue of imbalance and enhancing the model's capability to handle traditionally challenging instances.

Experiments

Experimental Setting

Datasets. We test our model's adaptability on multiple datasets spanning object detection, historical manuscript, and natural scene text recognition. For natural scene object detection, we employ the Grozi-3.2k (George and Floerkemeier 2014), supplemented by the tests on Dairy and Paste datasets (Osokin, Sumin, and Lomakin 2020). Here, 'Paste-f' includes all evaluation data, while 'Paste-v', a 'Paste-f' subset, is curated to exclude misoriented objects. These datasets predominantly feature retail items from supermarkets. For logo localization, the dataset FlickrLogos-32 (Kalantidis et al. 2011) is employed to assess the model's capability in discerning various brand imprints. For historical manuscript analyses, we use the Dongba Hieroglyphics (DBH), VML-HD (Kassis et al. 2017), and Tripitaka Koreana in Han (TKH) (Yang et al. 2018). In the realm of natural scene text recognition, we employ the ICDAR-13 (Karatzas et al. 2013) and KAIST (Jung et al. 2011; Lee et al. 2010). The KAIST dataset's English and Korean subsets are referred to as KAIST-E and KAIST-K respectively. Detailed dataset descriptions can be found in the appendix.

Implementation Details. During training, we randomly crop patches from all query images to a size of 800×800. Our method constructs a 7-level pyramid, with scales spanning from 0.5 to 1.6. We designate the default local neighbor region k as 5. Parameters for the loss function are set as follows: $m_{pos} = 0.6$, $m_{neg} = 0.5$, and μ at 3. For feature extraction, we leverage the third blocks of ResNet-50 (He et al. 2016) pre-trained on ImageNet (Russakovsky et al. 2015). Geometric matching is achieved through a 6-degree-of-freedom affine transformation. Our model, implemented in PyTorch 1.9, utilizes the Adam optimizer (Kingma and

Method	GroZi-3.2k¶			Dairy			Paste-v			Paste-f		
	mAP	Recall	F1	mAP	Recall	F1	mAP	Recall	F1	mAP	Recall	F1
DSW	57.24	72.86	64.11	32.50	58.41	41.76	6.49	7.50	6.96	6.46	5.90	6.17
CoAE	85.12	**97.86**	**91.05**	54.65	87.83	67.38	30.92	49.54	38.07	30.17	53.94	38.69
OS2D	**87.35**	94.23	90.66	72.35	91.88	80.96	56.95	55.52	56.23	52.56	43.76	47.76
CoNet	87.08	95.30	91.00	**73.39**	**92.03**	**81.66**	**64.46**	**72.74**	**68.35**	**58.27**	**55.54**	**56.88**

Table 1: Performance evaluation (%) across datasets GroZi-3.2k, Dairy, Paste-v, and Paste-f. The model is trained on GroZi-3.2k *'Seen'* categories with evaluations on *'Unseen'* categories within GroZi-3.2k indicated by asterisks (¶). Further evaluations on Dairy, Paste-v, and Paste-f employ weights derived from GroZi-3.2k's training process.

Figure 4: Visualization results for objects in natural scene images. Green boxes denote categories encountered during training (*'Seen'*), blue boxes represent categories that are not present during the training process (*'Unseen'*).

Ba 2014), with settings of a learning rate at 1e-4. It's trained on an Nvidia GeForce RTX 3090. In evaluations, we apply standard Pascal VOC metrics (Everingham et al. 2010) such as mAP, recall, and F1 score. We adopt an IoU threshold of 0.5 specifically for the mAP metric.

Comparison Results

For comparison, we consider the following methods: (1) DSW (Wicht, Fischer, and Hennebert 2016), an established unsupervised method grounded in the "sliding window" technique, which we faithfully reproduce. To ensure a fair comparison, DSW applies the same pre-trained network backbone with CoNet, and maintains precise aspect ratios for the image feature maps; (2) CoAE (Hsieh et al. 2019), where we introduce a multi-scale pyramid structure to the original CoAE implementation for a fair comparison; (3) OS2D (Osokin, Sumin, and Lomakin 2020), a one-shot object detection method rooted in dense correlation matching features. We select these techniques as they offer solutions akin to CoNet. All methods, including ours, are trained and tested under consistent conditions for a fair comparison.

Method	Unseen			Seen		
	mAP	Recall	F1	mAP	Recall	F1
DSW	7.84	20.19	11.29	3.06	16.12	5.15
CoAE	8.45	18.56	11.61	**27.06**	**29.07**	**28.03**
OS2D	16.32	25.62	19.93	18.82	25.95	21.82
CoNet	**20.50**	**25.99**	**22.92**	22.25	26.38	24.14

Table 2: Performance evaluation (%) on the FlickrLogos-27.

Method	Unseen			Seen		
	mAP	Recall	F1	mAP	Recall	F1
DSW	51.70	59.64	55.39	39.08	35.76	37.35
CoAE	62.64	60.36	61.48	72.55	93.55	81.72
OS2D	98.08	99.29	98.68	89.22	94.76	91.91
CoNet	**99.42**	**100.0**	**99.71**	**90.71**	**95.34**	**92.97**

Table 3: Performance evaluation (%) on the DBH.

Method	Unseen			Seen		
	mAP	Recall	F1	mAP	Recall	F1
DSW	62.77	74.15	67.99	20.70	46.83	32.36
CoAE	39.07	60.17	47.38	27.30	53.07	36.05
OS2D	94.39	98.31	96.31	55.31	52.45	53.84
CoNet	**99.20**	**100.0**	**99.60**	**72.02**	**73.37**	**72.69**

Table 4: Performance evaluation (%) on the VML-HD.

Object Detection in Natural Scenes. We evaluate comparison methods for object detection in natural scenes. Figure 4 shows the qualitative comparison to the OS2D, distinguishing between *'Unseen'* and *'Seen'* categories. Grozi-3.2k, Dairy, and Paste predominantly feature retail products. Table 1 presents the performance of each method across different datasets, with the top results highlighted in **bold**. To gauge our model's generalization, we evaluate the Dairy and Paste datasets—treating their categories as *'Unseen'* classes—using weights from the GroZi-3.2k training. Although CoNet not achieve the best results on the *'Unseen'* classes of GroZi-3.2k, it outperformed on the other datasets. Compared to the second-best method, CoNet's F1 score showed improvements of 0.7%, 12.12%, and 9.12% for each of the Dairy, Paste-v, and Paste-f datasets, respectively.

Logo Detection in Natural Scenes. Logo detection, a specific form of object detection, faces challenges in real-world scenarios. Logos can exhibit varied appearances due to lighting, occlusions, rotations, and scales. Despite using

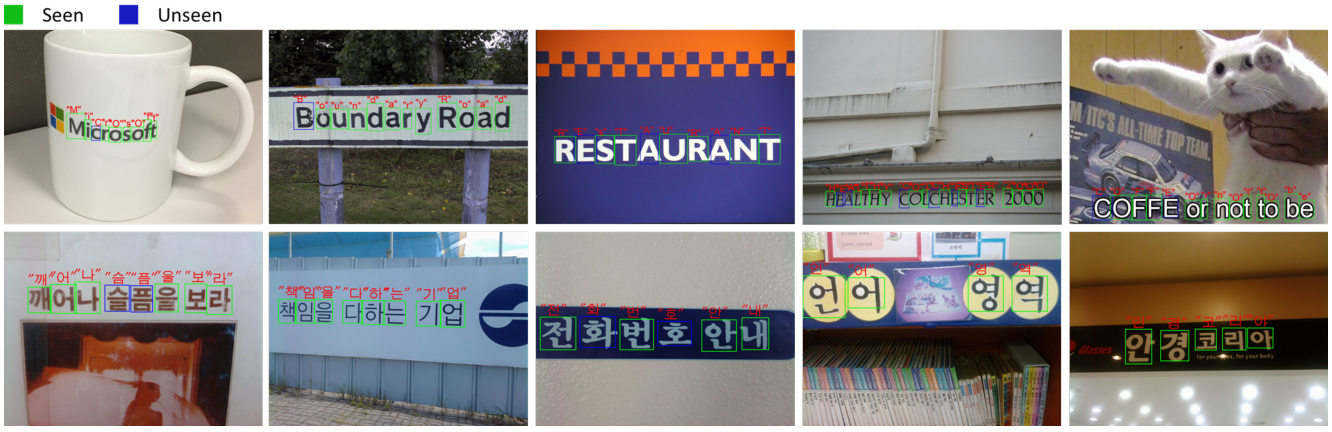

Figure 5: Visualization of CoNet for English and Korean in natural scene images. Green boxes denote categories encountered during training (*'Seen'*), blue boxes represent categories that are not present during the training process (*'Unseen'*).

Method	Set 1			Set 2			Set 3			Set 4		
	mAP	Recall	F1	mAP	Recall	F1	mAP	Recall	F1	mAP	Recall	F1
DSW	39.85	92.03	55.62	45.41	92.00	60.81	37.20	86.87	52.09	38.69	90.48	54.20
CoAE	36.63	52.03	42.99	36.86	42.56	39.51	28.65	43.64	34.59	42.92	47.82	45.24
OS2D	90.45	99.77	94.88	85.65	99.45	92.04	85.02	97.96	91.03	87.38	98.74	92.71
CoNet	**92.05**	**99.75**	**95.75**	**89.96**	**99.63**	**94.55**	**85.95**	**99.42**	**92.20**	**89.90**	**99.58**	**94.49**

Table 5: Performance evaluation (%) on *'Unseen'* categories from the partitioned TKH dataset.

Method	Set 1			Set 2			Set 3			Set 4		
	mAP	Recall	F1	mAP	Recall	F1	mAP	Recall	F1	mAP	Recall	F1
DSW	38.31	81.84	52.19	45.09	85.57	59.06	34.75	81.53	48.73	37.69	82.53	51.75
CoAE	54.12	84.98	66.13	63.50	86.86	73.37	51.22	80.96	62.74	58.11	82.52	68.20
OS2D	90.96	95.04	92.96	87.66	96.41	91.83	82.43	96.23	88.80	88.63	95.69	92.02
CoNet	**92.95**	**95.33**	**94.13**	**89.24**	**96.72**	**92.83**	**83.76**	**97.01**	**89.90**	**91.05**	**96.03**	**93.47**

Table 6: Performance evaluation (%) on *'Seen'* categories from the partitioned TKH dataset.

image pyramids in both our method and the comparative approaches, logo detection presents more difficulties than datasets like Grozi-3.2k. Table 2 displays the performance of the logo detection task on FlickrLogos-27. Experiments find that our method performs better on *'Unseen'* categories, but CoAE performs better on *'Seen'* categories. Nevertheless, the performance obtained by several methods is not high enough, which we guess is due to the low quality of visual archetypes provided by the dataset.

Text Spotting in Historical Manuscript. We evaluate each method's performance on historical manuscript images. Tables 3-4 outline the performance scores for various methods on the DBH and VML datasets. Our proposed CoNet method evidently outperforms the others. Remarkably, CoNet and a few other methods excel in *'Unseen'* categories compared to *'Seen'* categories, potentially due to fewer untrained categories and the relative ease of spotting some *'Unseen'* categories. To assess the model's robustness against diverse data distributions, while conserving computational resources and enhancing computational efficiency, we subdivide the TKH test images into four subsets. Tables 5-6 illustrate each method's performance on the TKH dataset. CoNet consistently outperforms all other methods. Interestingly, we observed that CoAE struggles with text spotting tasks, unsupervised methods like DSW surpass supervised ones like CoAE on the *'Unseen'* class of TKH.

Method	Unseen			Seen		
	mAP	Recall	F1	mAP	Recall	F1
DSW	8.07	46.18	13.74	2.71	27.67	49.94
CoAE	23.48	62.43	34.12	**56.11**	80.19	**66.03**
OS2D	32.34	88.65	47.39	25.00	70.04	37.80
CoNet	**63.04**	**93.35**	**75.26**	46.19	**84.48**	59.72

Table 7: Performance evaluation (%) on the ICDAR-13.

Text Spotting in Natural Scenes. Further testing CoNet's performance, we conduct broader testing using text detection and recognition data from natural scenes. Due to the lack of corresponding datasets for certain lesser-known languages and texts, we employ natural scene data in English and Korean to simulate new classes of problems that could

Method	Unseen			Seen		
	mAP	Recall	F1	mAP	Recall	F1
DSW	42.02	88.31	56.95	12.07	66.51	20.43
CoAE	46.04	71.43	55.99	79.89	**99.07**	88.45
OS2D	68.28	96.10	79.83	65.57	97.91	78.54
CoNet	**87.93**	**97.40**	**92.42**	**81.07**	98.84	**89.08**

Table 8: Performance evaluation (%) on the KAIST-E.

Method	Unseen			Seen		
	mAP	Recall	F1	mAP	Recall	F1
DSW	53.46	91.72	67.55	52.31	95.23	67.52
CoAE	78.04	**98.82**	87.21	82.82	99.43	90.37
OS2D	79.43	**98.82**	88.07	75.26	**99.84**	85.88
CoNet	**85.62**	**98.82**	**91.75**	**83.00**	99.76	**90.61**

Table 9: Performance evaluation (%) on the KAIST-K.

arise from small text samples in open scenes. In contrast to historical manuscript images, natural scene images present their own challenges, including complex backgrounds with numerous interfering factors. Figure 5 presents the visualization results of CoNet. Tables 7-9 outline performance of each method for English and Korean natural scenes. Though CoNet may not lead in mAP or Recall on some *'Seen'* categories, it consistently excels in F1 score across most datasets. On the *'Unseen'* category in ICDAR-13, KAIST-English, and KAIST-Korean, CoNet surpasses the runner-up mAP by 30.7%, 19.65%, and 6.19%, respectively.

Ablation Study

Effectiveness of Auto-correlation Representation Module. Table 10 demonstrates the significance of the ACR module. When we remove ACR, the model forgoes self-correlation learning and directly uses the base representation A instead of its output F. In comparison to the base representation, the auto-correlation representation exhibits superior generalization to *'Unseen'* classes. Based on our results, omitting the ACR leads to notable performance drops across various datasets: decreases of 7.07%, 4.03%, and 5.54% are observed in Paste-f, KAIST-E, and KAIST-K, respectively.

Effectiveness of Reciprocal Consistency Constraint. As illustrated in Table 10, bypassing the *Reciprocal Consistency Constraint* (RCC) and directly inputting the correlation map C to the AGS module, a notable performance degradation occurs. The declines amount to 7.21%, 4.87%, and 9.76% for Paste-f, KAIST-E, and KAIST-K datasets, respectively. The RCC ensures the model incorporates means to exclude outliers, enhancing the model's robustness.

Effectiveness of Adaptive Margin Training Strategy. To demonstrate our proposed adaptive margin contrastive learning strategy, we compare it against fixed margins. As displayed in Table 11, our adaptive margin strategy results in superior scores. This demonstrates the operational utility of the adaptive threshold loss during the training process. Theoretically, this is attributed to the dynamic adaptability of the adaptive margins, which tailors the threshold based on the unique attributes of each data instance.

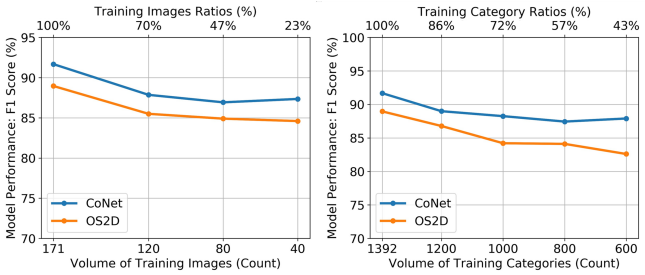

Figure 6: Performance evaluation on TKH *'Unseen'* categories after sub-sampling training images.

ID	ACR	RCC	Paste-f	KAIST-E	KAIST-K
1	✓	✓	**56.88**	**92.42**	**91.75**
2	✗	✓	49.81	88.39	86.21
3	✓	✗	49.67	87.55	81.99

Table 10: Ablation study results from our model. We assess the F1 scores across different datasets for *'Unseen'* classes.

ID	Training strategy	Dairy	VML-HD	KAIST-E
1	Adaptive margin	**81.66**	**99.60**	**92.42**
2	Fixed margin	78.99	98.99	69.46

Table 11: Performance obtained by different training strategies. We assess the F1 scores across different datasets for *'Unseen'* classes.

Sub-Sampling Experiment. We assess CoNet's resilience in resource-constrained settings by varying the number of images or categories during training using the TKH dataset. Figure 6 illustrates a direct correlation between the increase in training images or categories and the enhancement in F1 score performance. Notably, as the training data diminishes, compared with the most competitive method OS2D, CoNet consistently outperforms, reflecting its robustness even with limited data.

Conclusions

In this work, we introduce CoNet, a novel approach that utilizes visual archetypes for the identification of untrained categories. With a strong integration of local features and contextual information, CoNet enhances detection precision while minimizing false positives. Central to our advanced framework is an innovative auto-correlation representation module, in tandem with a mutual-correlation representation module, both working synergistically to bolster detection competency. To optimize the training process, an adaptive margin strategy is deployed, deftly managing sample dynamics. CoNet harnesses structural correlations among visual features for superior generalization to unseen classes, triggering notable enhancements in one-shot learning performance. Abundant experiments across various domains substantiate CoNet's robust performances against existing methods.

Acknowledgments

This work was jointly supported by the National Key Research and Development Program of China under Grant No. 2020AAA0107903, the National Natural Science Foundation of China under Grant No. 62176091.

References

Bendale, A.; and Boult, T. E. 2016. Towards Open Set Deep Networks. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1563–1572.

Bhunia, A. K.; Bhunia, A. K.; Ghose, S.; Das, A.; Roy, P. P.; and Pal, U. 2019. A deep one-shot network for query-based logo retrieval. *Pattern Recognition*, 96: 106965.

Cao, Z.; Lu, J.; Cui, S.; and Zhang, C. 2020. Zero-shot Handwritten Chinese Character Recognition with hierarchical decomposition embedding. *Pattern Recognition*, 107: 107488.

Chan, D. M.; and Riek, L. D. 2020. Unseen Salient Object Discovery for Monocular Robot Vision. *IEEE Robotics and Automation Letters*, 5(2): 1484–1491.

Chen, H.; Wang, Y.; Wang, G.; and Qiao, Y. 2018. LSTD: A Low-Shot Transfer Detector for Object Detection. *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 32(1).

Chen, J.; Li, B.; and Xue, X. 2021. Zero-Shot Chinese Character Recognition with Stroke-Level Decomposition. *ArXiv*, abs/2106.11613.

Cheng, M.; Wang, H.; and Long, Y. 2021. Meta-Learning-Based Incremental Few-Shot Object Detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(4): 2158–2169.

Chopra, S.; Hadsell, R.; and LeCun, Y. 2005. Learning a similarity metric discriminatively, with application to face verification. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1: 539–546.

Everingham, M.; Van Gool, L.; Williams, C. K.; Winn, J.; and Zisserman, A. 2010. The Pascal Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision (IJCV)*, 88: 303–338.

Fan, Q.; Zhuo, W.; Tang, C.-K.; and Tai, Y.-W. 2019. Few-Shot Object Detection With Attention-RPN and Multi-Relation Detector. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4013–4022.

George, M.; and Floerkemeier, C. 2014. Recognizing Products: A Per-exemplar Multi-label Image Classification Approach. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 440–455.

Han, J.; Ren, Y.; Ding, J.; Pan, X.; Yan, K.; and Xia, G.-S. 2022. Expanding Low-Density Latent Regions for Open-Set Object Detection. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9581–9590.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778.

Hsieh, T.-I.; Lo, Y.-C.; Chen, H.-T.; and Liu, T.-L. 2019. One-Shot Object Detection with Co-Attention and Co-Excitation. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32.

Huang, S.; Ma, J.; Han, G.; and Chang, S.-F. 2022. Task-Adaptive Negative Envision for Few-Shot Open-Set Recognition. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 7171–7180.

Huang, S.; Wang, Q.; Zhang, S.; Yan, S.; and He, X. 2019. Dynamic Context Correspondence Network for Semantic Alignment. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2010–2019.

Jeong, M.; Choi, S.; and Kim, C. 2021. Few-shot Open-set Recognition by Transformation Consistency. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 12561–12570.

Jiang, X.; Li, Z.; Tian, M.; Liu, J.; Yi, S.; and Miao, D. 2023. Few-shot Object Detection via Improved Classification Features. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 5386–5395.

Joseph, K.; Khan, S.; Khan, F. S.; and Balasubramanian, V. N. 2021. Towards Open World Object Detection. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5830–5840.

Jung, J.; Lee, S.; Cho, M. S.; and Kim, J. H. 2011. Touch TT: Scene Text Extractor Using Touchscreen Interface. *ETRI Journal*, 33(1): 78–88.

Kalantidis, Y.; Pueyo, L. G.; Trevisiol, M.; van Zwol, R.; and Avrithis, Y. 2011. Scalable triangulation-based logo recognition. In *Proceedings of the 1st ACM International Conference on Multimedia Retrieval*, 1–7.

Karatzas, D.; Shafait, F.; Uchida, S.; Iwamura, M.; i Bigorda, L. G.; Mestre, S. R.; Mas, J.; Mota, D. F.; Almazan, J. A.; and De Las Heras, L. P. 2013. ICDAR 2013 Robust Reading Competition. *International Conference on Document Analysis and Recognition (ICDAR)*, 1484–1493.

Kassis, M.; Abdalhaleem, A.; Droby, A.; Alaasam, R.; and El-Sana, J. 2017. VML-HD: The historical Arabic documents dataset for recognition systems. *International Workshop on Arabic Script Analysis and Recognition (ASAR)*, 11–14.

Kingma, D. P.; and Ba, J. 2014. Adam: A Method for Stochastic Optimization. *arXiv preprint arXiv:1412.6980*.

Lee, S.; Cho, M. S.; Jung, K.; and Kim, J. H. 2010. Scene Text Extraction with Edge Constraint and Text Collinearity. *International Conference on Pattern Recognition (ICPR)*, 3983–3986.

Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; and Berg, A. C. 2015. SSD: Single Shot MultiBox Detector. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Neal, L.; Olson, M.; Fern, X.; Wong, W.-K.; and Li, F. 2018. Open Set Learning with Counterfactual Images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 613–628.

Osokin, A.; Sumin, D.; and Lomakin, V. 2020. OS2D: One-Stage One-Shot Object Detection by Matching Anchor Features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 635–652.

Pal, D.; Bundele, V.; Sharma, R.; Banerjee, B.; and Jeppu, Y. 2022. Few-Shot Open-Set Recognition of Hyperspectral Images with Outlier Calibration Network. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 3801–3810.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 39: 1137–1149.

Rocco, I.; Arandjelovic, R.; and Sivic, J. 2018. Convolutional Neural Network Architecture for Geometric Matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(11): 2553–2567.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115: 211–252.

Tripathi, A.; Dani, R. R.; Mishra, A.; and Chakraborty, A. 2020. Sketch-guided object localization in natural images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 532–547.

Wang, T.; Xie, Z.; Li, Z.; Jin, L.; and Chen, X. 2019a. Radical Aggregation Network for Few-shot Offline Handwritten Chinese Character Recognition. *Pattern Recognition Letters*, 125: 821–827.

Wang, X.; Hua, Y.; Kodirov, E.; Hu, G.; Garnier, R.; and Robertson, N. M. 2019b. Ranked List Loss for Deep Metric Learning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5207–5216.

Wicht, B.; Fischer, A.; and Hennebert, J. 2016. Deep Learning Features for Handwritten Keyword Spotting. *2016 23rd International Conference on Pattern Recognition (ICPR)*, 3434–3439.

Woo, S.; Park, J.; Lee, J.-Y.; and Kweon, I. S. 2018. CBAM: Convolutional Block Attention Module. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Wu, Z.; Lu, Y.; Chen, X.; Wu, Z.; Kang, L.; and Yu, J. 2022. UC-OWOD: Unknown-Classified Open World Object Detection. *Proceedings of the European Conference on Computer Vision (ECCV)*, 193–210.

Yang, H.; Jin, L.; Huang, W.; Yang, Z.; Lai, S.; and Sun, J. 2018. Dense and Tight Detection of Chinese Characters in Historical Documents: Datasets and a Recognition Guided Detector. *IEEE Access*, 6: 30174–30183.

Zhang, J.; Du, J.; and Dai, L. 2020. Radical Analysis Network for Learning Hierarchies of Chinese Characters. *Pattern Recognition*, 103: 107305.

Zhang, S.; Wen, L.; Lei, Z.; and Li, S. Z. 2020. RefineDet++: Single-Shot Refinement Neural Network for Object Detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(2): 674–687.

Zhou, Z.-H. 2022. Open-environment machine learning. *National Science Review*, 9(8): nwac123.