An End-to-End Class-Aware and Attention-Guided Model for Object State Classification

Filippos Gouidis[1,2], Konstantinos Papoutsakis[1], Theodore Patkos[1], Antonis Argyros[1,2], Dimitris Plexousakis[1,2]
[1]Foundation For Research and Technology - Hellas, Heraklion, Greece
[2]University of Crete, Heraklion, Greece
Email: {gouidis,papoutsa,patkos,argyros,dp}@ics.forth.gr

***Abstract*—Object State Classification (OSC) is a critical task in computer vision, enabling systems to understand the functional state of objects. This work proposes a novel end-to-end architecture for OSC that leverages the inherent relationship between object classification and state recognition. Our approach first classifies the object and then uses object-specific attention mechanisms to focus on relevant features for state classification. This two-stage design allows the model to effectively capture object-state dependencies while maintaining modularity and flexibility. We conduct an extensive ablation study to analyze the impact of key parameters, such as attention mechanisms and loss weighting, and evaluate our method against three baselines across four benchmark datasets. Experimental results demonstrate that our approach outperforms competing methods by a significant margin, achieving state-of-the-art performance.**

***Index Terms*—Object State Classification, Object Recognition, Attention Mechanisms, Multi-task Learning**

I. Introduction

An object's usability is inherently linked to its current state. For instance, the actions that an object can afford — such as opening, closing, turning on, or turning off—are dictated by its current state, i.e. whether it is "open" or "closed". Accurately assessing an object's state is therefore crucial for autonomous systems and intelligent agents, allowing them to interpret the object's status and affordances, anticipate possible interactions, and plan appropriate system actions effectively. Thus, the recognition of object states and their transitions is essential for understanding an object's condition, the interactions performed on it, and its affordances [1]. This underscores the importance of the **Object State Classification (OSC)** task in computer vision, which enhances the functionality of AI systems in applications such as learning object affordances [2], recognizing human-object interactions [3], reasoning about state transitions [4], and assessing task completion or failure [5].

Despite its significance, research on OSC remains relatively sparse compared to the extensive work on object classification. Recent years, however, have seen growing interest in this area, with several studies addressing the problem [6], [7], [8], [9]. OSC presents unique challenges: subtle visual details often distinguish different states. States exhibit significant intra-class variability, as visually distinct objects can share the same state. Moreover, states are inherently object category-dependent, meaning they are only applicable to specific object types, further complicating the classification process. Finally, defining precise boundaries between state classes can be highly challenging.

In this work, we introduce a novel method for OSC, grounded on the observation that although object classification and state recognition are both classification tasks, they require distinct approaches. Specifically, state classification (SC) depends heavily on object classification (OC), suggesting that SC should be performed after OC, using the object class as an informative cue. Building on this, we propose an end-to-end architecture where object classification serves as an auxiliary component, enhancing the accuracy and robustness of state recognition through object-specific attention.

The main contributions of this work are as follows:

- A novel end-to-end architecture for state classification that uses object class information to improve state classification.
- An extensive ablation study to analyze the impact of key parameters on state classification.
- An evaluation against 3 baseline methods on 4 benchmark datasets, showing that the proposed approach leads to significant performance improvements.

II. Related Work

Object State Classification: Object states are typically treated as a sub-category of visual attributes — machine-detectable and human-understandable concepts that describe objects [10]. Traditional approaches to attribute classification rely on convolutional neural networks (CNNs) trained with discriminative classifiers on annotated datasets [11]. However, existing datasets for attribute and state classification are limited in scale, diversity, and category coverage [12], [6], [13], [14], [15], and research specifically focused on state classification remains relatively sparse [7], [8], [16], [17], [18], [19], [20], as opposed to the more generic tasks of attribute classification and attribute-based object recognition [21], [22], [23], [24], [25], [26], [27].

Attention-based Models: Recent advancements in attention mechanisms and transformer architectures have shown promise in fine-grained recognition tasks, including object-state classification. Multi-head attention, in particular, has been effective in capturing diverse aspects of feature representations [28], [29], [30], [31]. However, the application of

these techniques to object-state classification remains under-explored. Our work bridges this gap by introducing a novel architecture that leverages object-specific attention for state classification, addressing the limitations of existing methods.

Modular Architectures for Multi-Task Learning: Modular architectures have gained traction in multi-task learning, where separate modules are designed to handle distinct tasks while sharing common features [32], [33], [34], [35]. In the context of object-state classification, modular designs allow for the decoupling of object and state recognition, enabling flexible and scalable systems. Our work builds on these ideas by introducing a modular architecture that combines object classification with object-specific attention mechanisms for OSC.

III. Methodology

Let O denote a set of object classes, S a set of object states and I the set of images, respectively. We assume that each image $i \in I$ contains an object $o \in O$ that is situated in a state $s \in S$. Given the image i, the objective of OSC is to identify the object state $s \in S$. Importantly, although the OSC task does not deal explicitly with the classification of the object classes, the characteristics of the set O, i.e., size, variability, etc., affect the difficulty of the problem significantly.

We propose a novel end-to-end network architecture for Object State Classification (OSC), comprising two key components: an Object Classifier and a State Classifier. The Object Classifier extracts object-specific features and predicts the object class, while the State Classifier leverages attention mechanisms to determine the object's state. A distinguishing aspect of our approach is the dependency of the attention mechanisms on the predicted object class, facilitating object-specific feature refinement for improved state classification. Furthermore, the modular design of our framework allows flexibility in selecting the object classifier, accommodating both off-the-shelf and custom models. The training and inference pipeline follows a four-stage process:

1) Extraction of deep feature representations using the object classifier.
2) Prediction of the object class using the object classifier head.
3) Application of object-specific attention to the extracted features based on the predicted object class.
4) Prediction of the state class using the refined features.

The basic network components of the model are:

Backbone Feature Extractor: The backbone extracts feature representations from the input image i. The final layer is removed, and the penultimate layer's output (a D-dimensional feature vector) is used as input for subsequent classifiers. For instance, if ResNet-101 is used as the backbone, $D = 2048$.

Object Classifier Head: The object classifier head maps the extracted features to N_{obj} object classes using a single linear layer:

$$y_{\text{obj}} = \text{Softmax}(W_o \mathbf{x} + b_o), \tag{1}$$

where $W_o \in \mathbb{R}^{N_{\text{obj}} \times D}$ and $b_o \in \mathbb{R}^{N_{\text{obj}}}$ are learnable parameters.

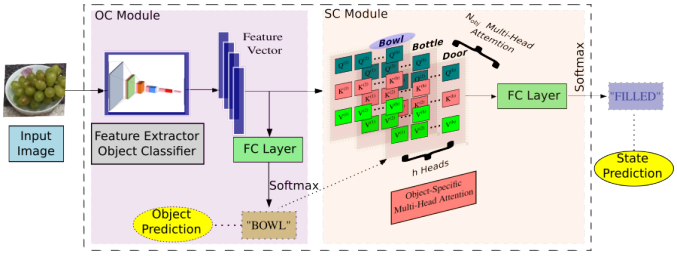

Fig. 1: Illustration of the proposed architecture which consists of two components: an object and a state classifier. The object classifier processes an input image, extracts features, and predicts the object class. These features are passed to the state classifier, which employs a multi-head attention mechanism. The attention weights are determined by the predicted object class, enabling object-specific state classification. Best viewed with zoom and color.

Multi-Head Attention State Classifier: The state classifier employs multi-head attention to focus on diverse aspects of the extracted feature representations. Let $\mathbf{X} \in \mathbb{R}^{B \times D}$ denote the input feature matrix, where B is the batch size and D is the feature dimensionality.

Base attention formulation. For each of the H attention heads, the model learns query, key, and value projections:

$$\mathbf{Q}_h = \mathbf{X}\mathbf{W}_h^Q, \quad \mathbf{K}_h = \mathbf{X}\mathbf{W}_h^K, \quad \mathbf{V}_h = \mathbf{X}\mathbf{W}_h^V, \tag{2}$$

where $\mathbf{W}_h^Q, \mathbf{W}_h^K, \mathbf{W}_h^V \in \mathbb{R}^{D \times d_k}$, and $d_k = D/H$. The attention weights and context vectors are computed as:

$$\mathbf{A}_h = \text{softmax}\left(\frac{\mathbf{Q}_h \mathbf{K}_h^\top}{\sqrt{d_k}}\right), \quad \mathbf{C}_h = \mathbf{A}_h \mathbf{V}_h. \tag{3}$$

The outputs of all heads are concatenated and projected:

$$\mathbf{C} = \text{Concat}(\mathbf{C}_1, \ldots, \mathbf{C}_H), \quad \mathbf{Z} = \mathbf{C}\mathbf{W}_o, \tag{4}$$

where $\mathbf{W}_o \in \mathbb{R}^{D \times D}$. Finally, a residual connection and layer normalization yield:

$$\mathbf{Y} = \text{LayerNorm}(\mathbf{Z} + \mathbf{X}). \tag{5}$$

Object-dependent attention. To incorporate object-specific priors, the attention parameters are conditioned on the predicted object class y_{obj}. Each object class $c \in \{1, \ldots, N_{\text{obj}}\}$ has a distinct set of projection matrices:

$$\mathbf{W}_{h,c}^Q, \ \mathbf{W}_{h,c}^K, \ \mathbf{W}_{h,c}^V \in \mathbb{R}^{D \times d_k}, \quad \mathbf{W}_{o,c} \in \mathbb{R}^{D \times D}. \tag{6}$$

For a predicted object class $y_{\text{obj}} = c$, the corresponding parameters are selected:

$$\mathbf{Q}_{h,c} = \mathbf{X}\mathbf{W}_{h,c}^Q, \quad \mathbf{K}_{h,c} = \mathbf{X}\mathbf{W}_{h,c}^K, \quad \mathbf{V}_{h,c} = \mathbf{X}\mathbf{W}_{h,c}^V, \tag{7}$$

$$\mathbf{A}_{h,c} = \text{softmax}\left(\frac{\mathbf{Q}_{h,c} \mathbf{K}_{h,c}^\top}{\sqrt{d_k}}\right), \quad \mathbf{C}_{h,c} = \mathbf{A}_{h,c} \mathbf{V}_{h,c}. \tag{8}$$

The object class-specific attention output is then obtained as:

$$\begin{aligned} \mathbf{C}_c = \text{Concat}(\mathbf{C}_{1,c}, \ldots, \mathbf{C}_{H,c}), \mathbf{Z}_c = \mathbf{C}_c \mathbf{W}_{o,c}, \\ \mathbf{Y} = \text{LayerNorm}(\mathbf{Z}_c \mathbf{W}_{o,c} + \mathbf{X}). \end{aligned} \tag{9}$$

TABLE I: Ablation study for the parameter γ. First/Second value in each cell corresponds to Weighted Accuracy/Average Accuracy. The reported values have been averaged over the 5 different values of the number of attention heads (see Table II).

γ	OSDD	CGQA	MIT	VAW
0.0	13.3/12.5	12.2/12.4	10.4/9.9	11.4/12.8
0.1	**63.7**/**60.5**	**36.4**/33.4	**49.9**/**41.5**	**34.6**/35.9
0.2	62.6/60.2	33.0/35.4	46.5/37.4	32.7/35.3
0.3	61.2/60.1	35.8/34.2	46.5/37.9	33.6/35.5
0.4	61.8/59.1	32.7/**36.2**	45.8/37.6	33.4/35.3
0.5	60.4/59.1	**36.4**/34.5	47.5/39.8	**34.3**/35.8
0.6	60.3/58.4	34.5/31.7	46.8/38.3	34.3/34.4
0.7	60.0/58.5	32.4/30.7	43.1/35.4	32.8/33.8
0.8	58.3/57.2	33.5/32.8	44.4/37.6	33.0/32.8
0.9	55.7/53.2	32.1/30.7	38.1/31.1	29.8/28.5
1.0	56.5/54.4	24.7/29.6	40.2/35.6	26.6/27.7

The resulting object-conditioned representation $\mathbf{Y}$ is passed to the state classifier, which produces the final state probabilities:

$$y_{\text{st}} = \text{softmax}(\mathbf{Y}\mathbf{W}_s + \mathbf{b}_s), \tag{10}$$

where $\mathbf{W}_s \in \mathbb{R}^{D \times N_{\text{st}}}$ and $\mathbf{b}_s \in \mathbb{R}^{N_{\text{st}}}$ are learnable parameters.
Training Objective: The network is optimized using a weighted loss function that balances object and state classification:

$$\mathcal{L} = (1-\gamma)\mathcal{L}_{\text{obj}} + \gamma\mathcal{L}_{\text{st}}, \tag{11}$$

where $\mathcal{L}_{\text{obj}}$ and $\mathcal{L}_{\text{state}}$ are the losses for object and state classification, respectively, and $\gamma \in [0, 1]$ controls the relative importance of object and state classification.

The object and state classification losses $\mathcal{L}_{\text{obj}}$ and $\mathcal{L}_{\text{state}}$ are computed using the cross-entropy loss function as follows:

$$\mathcal{L}_{\text{obj}} = -\frac{1}{B}\sum_{i=1}^{B}\sum_{j=1}^{N_{\text{obj}}} y_{\text{obj},ij}\log(\hat{y}_{\text{obj},ij}), \tag{12}$$

$$\mathcal{L}_{\text{st}} = -\frac{1}{B}\sum_{i=1}^{B}\sum_{k=1}^{N_{\text{st}}} y_{\text{st},ik}\log(\hat{y}_{\text{st},ik}), \tag{13}$$

where B is the batch size, N_{obj} is the number of object classes, $y_{\text{obj},ij}$ is the ground truth label for object class j in sample i, $\hat{y}_{\text{obj},ij}$ is the predicted probability for object class j in sample i. N_{st} is the number of state classes, $y_{\text{st},ik}$ is the ground truth label for state class k in sample i, and $\hat{y}_{\text{st},ik}$ is the predicted probability for state class k in sample i.

IV. Experimental Evaluation

Implementation Details: The visual backbone employed for object classification is the CNN-based ResNet-101 model [36]. Training was conducted for $30 \times N$ iterations, where N denotes the number of target state classes. The model was optimized using stochastic gradient descent (SGD) with a learning rate of 0.0001 and a momentum of 0.9. The evaluation was performed on four benchmark Object State Datasets: OSDD [7], CGQA [15], MIT [6], and VAW [37].

Evaluation Scenarios: We consider 2 evaluation scenarios:

- *Intra-dataset evaluation*, where training and testing are conducted on the same dataset.

TABLE II: Ablation study for the number of attention heads. First/Second value in each cell corresponds to Weighted Accuracy/Average Accuracy. The reported values have been averaged over the 11 different values of the γ parameter (see Table I).

Heads	OSDD	CGQA	MIT	VAW
2	60.0/58.2	**34.9**/32.1	44.6/37.5	**33.4**/**33.9**
4	**60.4**/**58.4**	34.8/**34.3**	**46.5**/**38.6**	33.3/33.4
8	59.8/57.6	33.0/33.4	44.4/36.7	32.4/33.6
16	60.1/57.7	32.4/32.5	44.2/36.3	32.1/33.4
32	60.0/**58.4**	30.7/32.2	44.6/37.1	31.5/33.2

- *Inter-dataset evaluation*, where training is performed on one dataset and testing on other datasets.

Metrics: Model assessment employs two evaluation metrics:

- *Weighted Accuracy (WA)* – WA aggregates class-wise accuracies weighted by the relative frequency of each class, followed by computing the weighted average.
- *Average Accuracy (AA)* – AA assigns equal weights to all classes, computing their average accuracy.

Ablation Study: The ablation study investigates the impact of the γ parameter and the number of attention heads on model performance. Experiments were conducted using 11 different γ values and five distinct attention head configurations. Due to space constraints, we present only the results for the model trained on OSDD (see Table I and Table II). Regarding the γ parameter, the best performance was observed at $\gamma = 0.1$, followed by $\gamma = 0.2$. Conversely, the worst performance was recorded for $\gamma = 0$. Overall, performance deteriorated as γ increased, highlighting the role of object classification in the whole procedure.For the number of attention heads, the optimal results were obtained with 4 heads, followed by 2 heads. The lowest performance was observed with eight heads. Unlike the γ parameter, no clear trend was evident in the impact of the number of attention heads. These findings suggest that a small number of attention heads is sufficient for optimal performance, as increasing the number of heads does not yield further improvements.
Comparison with baseline methods: To evaluate the effectiveness of our proposed model, we compare it against three baseline methods:

- *Object-Agnostic State Classifier (OA-SC)*: A classifier trained to recognize object states without explicitly leveraging object class information.
- *Two-Stage State Classifier (TS-SC)*: A sequential approach where the object class is first predicted, followed by an object-specific state classifier to determine the object state.
- *Object-State Pair Classifier (OSPC)*: A model trained to jointly classify both the object class and its state.

To ensure a fair comparison, the three baseline methods employ ResNet-101 as visual backbone which is the same classifier that our model uses as object classifier. Although all 3 baseline methods share some characteristics with our

TABLE III: Experimental results. 1st/2nd/3rd/4th value in each cell corresponds to the performance of our method/OA-SC/TS-SC/OSPC, respectively. WA: Weighted Accuracy. AA: Average Accuracy. Bold/Underline: Best/Second best performance.

Metric	Train \ Test	OSDD	CGQA	MIT	VAW
WA	**OSDD**	**64.4** / 63.2 / 64.3 / 43.3	**39.1** / 37.2 / 16.2 / 25.9	**50.0** / 29.1 / 14.7 / 30.0	**39.5** / 37.2 / 36.4 / 27.1
AA		60.5 / **61.0** / 57.4 / 46.1	**34.8** / 32.7 / 15.5 / 23.1	**42.4** / 30.5 / 5.6 / 20.5	**37.9** / 34.0 / 26.4 / 20.3
WA	**CGQA**	**36.2** / 32.1 / 23.2 / 25.7	**69.8** / 65.7/68.7 / 48.5	**44.6** / 20.9 / 44.0 / 37.1	**52.2** / 48.2 / 15.7 / 32.9
AA		**39.0** / 18.4 / 28.6 / 29.4	47.7 / 40.1 / **53.2** / 34.5	**37.9** / 25.2 / 28.3 / 24.3	**41.4** / 19.4 / 19.8 / 21.6
WA	**MIT**	35.1 / 35.6 / **37.1** / 25.1	35.5 / 20.0 / **53.8** / 20.7	76.8 / **81.0** / 66.0 / 61.6	**35.9** / 33.7 / 10.7 / 25.0
AA		**46.8** / 20.8 / 22.3 / 20.7	**35.6** / 34.0 / 16.7 / 21.8	75.5 / **76.4** / 55.0 / 54.9	**44.8** / 22.4 / 6.8 / 25.2
WA	**VAW**	**35.4** / 29.6 / 30.9 / 24.0	**66.5** / 65.6 / 11.2 / 51.5	**43.2** / 40.0 / 12.5 / 35.3	63.6 / 62.1 / **65.6** / 51.4
AA		**34.1** / 29.5 / 30.9 / 26.2	**52.1** / 50.1 / 9.8 / 38.4	**41.1** / 37.0 / 5.4 / 37.7	54.4 / 50.2 / **64.0** / 35.6

TABLE IV: Method ranking according to the two metrics. 1st / 2nd / 3rd / 4th value in each cell corresponds to our method / OA-SC / TS-SC / OSPC. WA: Weighted Accuracy. AA: Average Accuracy.

Metric	Ranking	OSDD	CGQA	MIT	VAW
WA	1	3/0/1/0	2/0/2/0	3/1/0/0	3/0/1/0
WA	2	0/2/2/0	2/1/1/0	1/1/1/1	1/3/0/0
AA	1	3/0/1/0	3/0/1/0	3/1/0/0	3/0/1/0
AA	2	1/0/2/1	1/3/0/0	1/1/1/1	1/1/0/2
WA/AA	1	6/0/2/0	5/1/2/0	6/2/0/0	6/0/2/0
WA/AA	2	1/2/4/1	3/4/1/0	2/2/2/2	2/4/0/2

TABLE V: Aggregated ranking of the 4 evaluated methods.

Ranking	Ours	OA-SC	TS-SC	OSPC
1st	23	3	6	0
2nd	8	12	9	5
1st/2nd	31	15	15	5
1st/2nd (%)	96.9%	46.9%	46.9%	15.6%

approach, they also differ in key aspects. OA-SC is object-agnostic and, therefore, cannot leverage object class information to guide state classification. This limitation further hinders OA-SC because object classes are typically more visually salient than state classes, making it challenging for an object-agnostic method to disentangle the relevant state features from potentially misleading object features.

TS-SC is the method that most closely resembles our approach, as it also follows a two-stage procedure. However, there are crucial differences: TS-SC utilizes object prediction to select an appropriate state classifier, effectively reducing the number of possible state classes by limiting them to the states associated with a given object class. While this approach has the advantage of reducing the classification search space, it does not refine the extracted object features to produce a more informed state prediction. Instead, TS-SC applies the same feature extraction process for both object classification and state classification, treating them as independent tasks. This approach is sub-optimal because object features inherently contain valuable information for SC. Furthermore, a significant limitation of TS-SC is that it requires training $N_{\text{obj}}+1$ models, where N_{obj} is the number of object classes, making real-time inference impractical.

OSPC is also object-agnostic, despite learning to predict joint object-state labels. However, it does so in a semantically unaware manner. This means that for OSPC, labels such as "open book," "closed book," "open door," and "folded shirt" are treated as entirely separate classes. The model does not recognize that "open book" and "closed book" share the same object class, nor that "open book" and "open door" share the same state class. As a result, OSPC struggles to differentiate state-related features from object-class features. Moreover, a major drawback of this method is its large search space, i.e., $N_{\text{sts}} \times N_{\text{objs}}$. Based on these observations, we hypothesize the following behavior among the three baseline methods. First, we expect TS-SC to achieve the highest performance, while OSPC is likely to perform the worst. Second, we anticipate that the performance advantage of our model will be more pronounced in inter-dataset evaluation scenarios.

The results of the experimental evaluation are summarized in Table III, Table IV, and Table V. Across a total of 16 different evaluation scenarios (four intra-dataset and 12 inter-dataset), our method achieves the best performance in 11 cases using the AA metric and 12 cases using the WA metric. Overall, our approach outperforms competing methods in 23 out of 32 instances. Furthermore, in 31 out of 32 scenarios, our method ranks either first or second. If we examine solely the most representative metric for each experimental scenario, we can see that in the case of intra-dataset evaluation (WA metric), our method ranks first in 2 out of 4 cases, while in the case of inter-dataset evaluation (AA metric), it achieves the best performance in 11 out of 12 cases. Among the competing methods, the TS-SC model achieves the highest performance, whereas the OSPC model performs the worst. However, if we consider the significant drawbacks of TS-SC, i.e., increased computational cost and lack of real-time inference, and focus on comparing our model with the other 2 baselines we can see that our method outperforms these two baselines in 28 out of 32 scenarios. The previous results corroborate our hypotheses.

V. Conclusion

We propose a novel end-to-end method for Object State Classification (OSC) using a two-stage architecture that first classifies the object and then applies object-specific attention for state prediction. Our approach leverages the insight that object classification aids state recognition. Through ablation studies and evaluation on four benchmark datasets, we demonstrate superior performance. In future work, we plan to extend our method by refining further the employed attention mechanisms.

REFERENCES

[1] Lorenzo Jamone, Emre Ugur, Angelo Cangelosi, Luciano Fadiga, Alexandre Bernardino, Justus Piater, and José Santos-Victor, "Affordances in psychology, neuroscience, and robotics: A survey," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 10, no. 1, pp. 4–25, 2016.

[2] Ching-Yao Chuang, Jiaman Li, Antonio Torralba, and Sanja Fidler, "Learning to act properly: Predicting and explaining affordances from images," in *IEEE CVPR*, 2018, pp. 975–983.

[3] Xiaolong Wang, Ali Farhadi, and Abhinav Gupta, "Actions $\sim$ Transformations," in *IEEE CVPR*. jun 2016, vol. 2016-Decem, pp. 2658–2667, IEEE.

[4] Ali Farhadi, Ian Endres, Derek Hoiem, and David Forsyth, "Describing objects by their attributes," *2009 IEEE IEEE CVPR, CVPR 2009*, pp. 1778–1785, 2009.

[5] Tim J Schoonbeek, Tim Houben, Hans Onvlee, Fons van der Sommen, et al., "Industreal: A dataset for procedure step recognition handling execution errors in egocentric videos in an industrial-like setting," in *IEEE WACV*, 2024, pp. 4365–4374.

[6] Phillip Isola, Joseph J. Lim, and Edward H. Adelson, "Discovering states and transformations in image collections," *IEEE CVPR*, vol. 07-12-June, pp. 1383–1391, 2015.

[7] F. Gouidis, T. Patkos, A. Argyros, and D. Plexousakis, "Detecting object states vs detecting objects: A new dataset and a quantitative experimental study," in *VISAPP*, 2022, vol. 5, pp. 590–600.

[8] Tomáš Souček, Jean-Baptiste Alayrac, Antoine Miech, Ivan Laptev, and Josef Sivic, "Multi-Task Learning of Object State Changes from Uncurated Videos," 2022.

[9] N. Saini, H. Wang, A. Swaminathan, V. Jayasundara, B. He, K. Gupta, and A. Shrivastava, "Chop amp; learn: Recognizing and generating object-state compositions," in *IEEE ICCV*, Los Alamitos, CA, USA, oct 2023, pp. 20190–20201, IEEE Computer Society.

[10] Kun Duan, Devi Parikh, David Crandall, and Kristen Grauman, "Discovering localized attributes for fine-grained recognition," in *2012 IEEE IEEE CVPR*. IEEE, 2012, pp. 3474–3481.

[11] Krishna Kumar Singh and Yong Jae Lee, "End-to-end localization and ranking for relative attributes," in *ECCV*. Springer, 2016, pp. 753–769.

[12] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling, "Learning to detect unseen object classes by between-class attribute transfer," in *IEEE CVPR*. IEEE, 2009, pp. 951–958.

[13] Genevieve Patterson and James Hays, "Coco attributes: Attributes for people, animals, and objects," in *ECCV*. Springer, 2016, pp. 85–100.

[14] Aron Yu and Kristen Grauman, "Semantic jitter: Dense supervision for visual comparisons via synthetic images," in *IEEE ICCV*, 2017, pp. 5570–5579.

[15] Massimiliano Mancini, Muhammad Ferjad Naeem, Yongqin Xian, and Zeynep Akata, "Learning Graph Embeddings for Open World Compositional Zero-Shot Learning," *IEEE Trans. on PAMI*, 2022.

[16] Nirat Saini, Khoi Pham, and Abhinav Shrivastava, "Disentangling visual embeddings for attributes and objects," in *IEEE CVPR*, June 2022, pp. 13658–13667.

[17] Filippos Gouidis, Katerina Papantoniou, Konstantinos Papoutsakis, Theodore Patkos, Antonis Argyros, and Dimitris Plexousakis, "Fusing domain-specific content from large language models into knowledge graphs for enhanced zero shot object state classification," in *Proceedings of the AAAI Symposium Series*, 2024, vol. 3, pp. 115–124.

[18] Filippos Gouidis, Katerina Papantoniou, Konstantinos Papoutsakis, Theodore Patkos, Antonis Argyros, and Dimitris Plexousakis, "Llm-aided knowledge graph construction for zero-shot visual object state classification," in *ICPRS*. IEEE, 2024, pp. 1–7.

[19] Filippos Gouidis, Konstantinos Papoutsakis, Theodore Patkos, Antonis Argyros, and Dimitris Plexousakis, "Exploring the impact of knowledge graphs on zero-shot visual object state classification," in *VISAPP*, 2024, pp. 738–749.

[20] Filippos Gouidis, Konstantinos Papoutsakis, Theodore Patkos, Antonis Argyros, and Dimitris Plexousakis, "Recognizing unseen states of unknown objects by leveraging knowledge graphs," in *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2025, pp. 8648–8659.

[21] Zeynep Akata, Florent Perronnin, Zaid Harchaoui, and Cordelia Schmid, "Label-embedding for attribute-based classification," in *CVPR*, 2013.

[22] An Yan, Yu Wang, Yiwu Zhong, Chengyu Dong, Zexue He, Yujie Lu, William Yang Wang, Jingbo Shang, and Julian McAuley, "Learning concise and descriptive attributes for visual recognition," in *ICCV*, 2023, pp. 3090–3100.

[23] Farhad Pourpanah, Moloud Abdar, Yuxuan Luo, Xinlei Zhou, Ran Wang, Chee Peng Lim, Xi-Zhao Wang, and QM Jonathan Wu, "A review of generalized zero-shot learning methods," *IEEE transactions on pattern analysis and machine intelligence*, vol. 45, no. 4, pp. 4051–4070, 2022.

[24] Wenjia Xu, Yongqin Xian, Jiuniu Wang, Bernt Schiele, and Zeynep Akata, "Attribute prototype network for any-shot learning," *International Journal of Computer Vision*, vol. 130, no. 7, pp. 1735–1753, 2022.

[25] Muhammad Ferjad Naeem, Yongqin Xian, Federico Tombari, and Zeynep Akata, "Learning graph embeddings for compositional zero-shot learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 953–962.

[26] Liunian Li, Zi-Yi Dou, Nanyun Peng, and Kai-Wei Chang, "Desco: Learning object recognition with rich language descriptions," *Advances in Neural Information Processing Systems*, vol. 36, pp. 37511–37526, 2023.

[27] Shiming Chen, Ziming Hong, Yang Liu, Guo-Sen Xie, Baigui Sun, Hao Li, Qinmu Peng, Ke Lu, and Xinge You, "Transzero: Attribute-guided transformer for zero-shot learning," in *Proceedings of the AAAI conference on artificial intelligence*, 2022, vol. 36, pp. 330–338.

[28] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin, "Attention is all you need," *NIPS*, vol. 30, 2017.

[29] Jie Hu, Li Shen, and Gang Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[30] Hang Zhang, Chongruo Wu, Zhongyue Zhang, Yi Zhu, Haibin Lin, Zhi Zhang, Yue Sun, Tong He, Jonas Mueller, R Manmatha, et al., "Resnest: Split-attention networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 2736–2746.

[31] Wei Xu and Yi Wan, "Ela: Efficient local attention for deep convolutional neural networks," *arXiv preprint arXiv:2403.01123*, 2024.

[32] Ishan Misra, Abhinav Shrivastava, Abhinav Gupta, and Martial Hebert, "Cross-stitch networks for multi-task learning," in *IEEE CVPR*, 2016, pp. 3994–4003.

[33] Alex Kendall, Yarin Gal, and Roberto Cipolla, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," in *IEEE CVPR*, 2018.

[34] Xiangyun Zhao, Haoxiang Li, Xiaohui Shen, Xiaodan Liang, and Ying Wu, "A modulation module for multi-task learning with applications in image retrieval," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 401–416.

[35] Shikun Liu, Edward Johns, and Andrew J Davison, "End-to-end multi-task learning with attention," in *CVPR*, 2019, pp. 1871–1880.

[36] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *IEEE CVPR*, 2016, pp. 770–778.

[37] Khoi Pham, Kushal Kafle, Zhe Lin, Zhihong Ding, Scott Cohen, Quan Tran, and Abhinav Shrivastava, "Learning to predict visual attributes in the wild," in *Proceedings of the IEEE/CVF CVPR*, June 2021, pp. 13018–13028.