MSAT: MULTI-SCALE SEMANTIC-ALIGNED TRANSFORMER FOR MULTI-LABEL IMAGE CLASSIFICATION

Zhi-Jie Wang[a, b], *Zhuoyang Chen*[b], *Yongqi Liu*[a], *Naikang Zhong*[b], *Xiao Lin*[b†]

[a]College of Computer Science, Chongqing University
[b]Institute of Artificial Intelligence in Education, Shanghai Normal University

ABSTRACT

Multi-label image classification needs to recognize multiple coexisting and semantically correlated labels within a single image. Existing Transformer-based methods employ label queries for cross-attention but often rely on single-scale features or randomly initialized queries, limiting their ability to model diverse object scales and semantic consistency. In this paper, we propose a new Transformer-based framework, called MSAT (Multi-scale Semantic Aligned Transformer). MSAT incorporates a Scale-Specific Alignment (SSA) module, which leverages independent encoders and parallel decoders to explicitly align shared label queries with multi-scale features. Moreover, a Dual-source Semantic Prior (DSP) module combines pre-trained word embeddings with learnable task-specific embeddings to enhance both stability and adaptability of queries. Extensive experiments based on benchmarking datasets demonstrate that MSAT achieves favorable performance, verifying the effectiveness and competitiveness of our proposed method.

***Index Terms*—** Multi-Label Image Classification, Multi-Scale Feature Fusion, Semantic Alignment, Semantic-Aware Decoder

1. INTRODUCTION

Unlike traditional image classification [1, 2, 3], which assigns a single dominant category to an image, multi-label image classification aims to recognize multiple labels within a single image [4, 5, 6]. This task presents a greater challenge, as real-world images typically contain multiple coexisting and semantically related labels. These labels exhibit strong correlations and predictable co-occurrence patterns (e.g., "sofa" often appears with "living room"), implying that models should capture such complex interdependencies rather than merely identifying each label in isolation. As for the models that fail to simultaneously learn semantic structure and their alignment with visual contexts, they are hardly to recognize all targets in complex scenes [7, 8]. Therefore, a central focus to advance this field lies in effectively aligning visual content with semantic labels.

In recent years, Transformers [9] have introduced novel perspectives for multi-label image classification [10, 11, 12]. Their cross-attention mechanism in Transformer allows each label to directly retrieve relevant regions from image features. This manner is somewhat similar to a "query" operator. Thereby, label query driven [13] Transformer models have been extensively developed [14, 15, 16]. A representative approach, Query2Label [17], treats category labels as query tokens and uses cross-attention in the decoder to extract category-specific features from feature maps, eliminating the need for explicit region proposals. Although this method is effective, it relies on randomly initialized query vectors that lack semantic guidance, forcing the model to learn category-query mappings purely from visual data. Then, TSFormer [18] proposed a dual-branch decoder that incorporates pre-trained word embeddings into a semantic stream and progressively refines them with visual features, improving performance on various benchmarks. However, both TSFormer and Query2Label rely on single-scale features, limiting their ability to handle objects at diverse scales.

We also noted that, recent works have tried to integrating multi-scale features[19, 20]. For example, FL-Tran [21] aligns and fuses multi-scale features at the encoder level and incorporates Transformer-based learning, improving performance on targets with large size variance. Nevertheless, such methods focus primarily on encoder-side fusion, and lack explicit alignment of semantic queries to individual scale representations during decoding. Moreover, excessive cross-scale interaction may lead to overfitting. In general, current approaches do not fully leverage the semantic priors in query design, or underutilize explicit multi-scale decoding, restricting their discriminative power in complex scenarios.

To alleviate the above issues, we propose MSAT, a Multi-scale Semantic Aligned Transformer, for multi-label image classification. MSAT integrates two novel modules into an encoder-decoder Transformer framework: the Scale-Specific Alignment (SSA) module and the Dual-source Semantic Prior (DSP) module. The SSA module employs three independent Transformer encoders to process features at different scales, preserving distinct discriminative information at each level. For decoding, three parallel decoders utilize shared category queries to perform cross-attention with their respective scale-specific features, achieving explicit semantic alignment across different scales. The DSP module generates the category queries; it combines pre-trained word embeddings with learnable embeddings to provide stable semantic priors.

To summarize, our main contributions are as follows:

- We propose MSAT, a novel end-to-end Transformer framework for multi-label image classification.
- We design a Scale-Specific Alignment (SSA) module that utilizes independent encoders and parallel decoders to explicitly model semantic interactions at different scales.
- We develop a Dual-source Semantic Prior (DSP) module that integrates pre-trained word embeddings with learnable task-specific embeddings to enhance both the stability and adaptability of category queries.
- We conduct extensive experiments based on benchmarking datasets (MS-COCO and PASCAL VOC 2007), demonstrating the competitiveness of our proposed model.

† denotes corresponding author: Xiao Lin(lin6008@shnu.edu.cn)

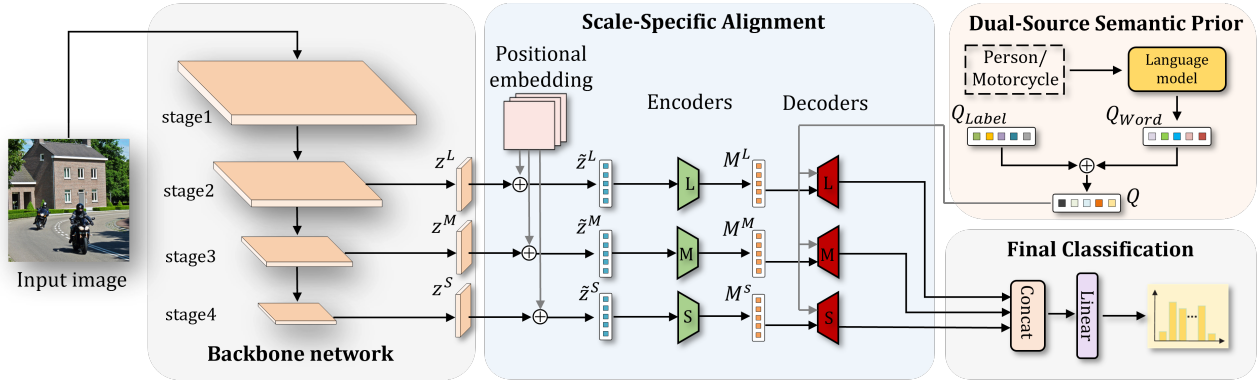

Fig. 1. The overall framework of MSAT.

2. METHOD

2.1. Overview of Our Method

We propose **MSAT**, a multi-scale semantic-aligned Transformer model for multi-label image classification. As depicted in Fig. 1, MSAT first extracts multi-scale features through a hierarchical backbone, and then processes them via two main modules: (1)Scale-Specific Alignment (SSA) module, containing independent multi-scale encoding and parallel multi-scale decoding. The independent multi-scale encoding ensures that each scale preserves its discriminative representation independently; while the parallel multi-scale decoding assigns an exclusive Transformer decoder to each scale, explicitly aligning the unified semantic queries with the visual features at that scale. (2)Dual-source Semantic Prior (DSP) module, which employs pre-trained word embeddings with learnable task-specific embeddings to generate class query embeddings to the SSA decoders, providing rich semantic priors. The outputs from the three decoders are concatenated along the class dimension and fused to produce the final multi-label predictions. Next, we will address the details of our model (Sec. 2.2∼2.5).

2.2. Backbone Network

We employ ConvNeXt-B [22] as the visual backbone, which extracts multi-scale features through four successive downsampling stages. Specifically, for an input image of size $H\times W$, the backbone produces three feature maps at stages 2, 3, and 4: $F^L\in\mathbb{R}^{C_2\times\frac{H}{8}\times\frac{W}{8}}$, $F^M\in\mathbb{R}^{C_3\times\frac{H}{16}\times\frac{W}{16}}$,$F^S\in\mathbb{R}^{C_4\times\frac{H}{32}\times\frac{W}{32}}$, where C_2=256, C_3=512, and C_4=1024, and the superscripts, namely L, M, and S denote large-, medium- and small-scale feature maps, respectively. The further process of each feature maps can be implemented as:

$$Z^X = \text{Proj}(\text{Flatten}(F^X)) \tag{1}$$

where, $Z^X\in\mathbb{R}^{N_X\times D}$, $X\in\{S,M,L\}$, $N_X = H_X \times W_X$, and $Proj(Flatten(\cdot))$ denotes that each feature map F^X is flattened, then projected to a unified embedding space of dimension D. This produces three sequence representations Z^S, Z^M, and Z^L that correspond to fine-grained, medium-scale, and global features, respectively. The alignment of different scale features in the channel dimension is achieved so that the fine-grained local features, structural detail features, and global semantic features extracted in the latter three stages are all represented in the same vector space.

2.3. Scale-Specific Alignment

In Scale-Specific Alignment (SSA) module, Z^X from each scale of the backbone is then fed into an independent Transformer encoder (with identical architecture but unshared parameters). Each encoder performs multi-head self-attention and feed-forward network operations on its input sequences to capture long-range dependencies within the same scale and enhance the feature representation. Before encoding, sinusoidal positional embeddings P^S, P^M, $P^L\in\mathbb{R}^{N_X\times D}$ are added to inject spatial location information:

$$\tilde{Z}^X = Z^X + P^X \tag{2}$$

where, X has the same meaning with that in Equation 1. After L_e stacked layers of multi-head self-attention and feed-forward networks, we obtain scale-specific visual memories:

$$\tilde{Z}^X_\ell = \text{EncLayer}\left(\tilde{Z}^X_{\ell-1}\right),\quad \ell = 1,\ldots,L_e \tag{3}$$

where, $EncLayer(\cdot)$ denotes the operator inside each Encoder layer. Through the L_e stacked layers, we obtain the Memory $\mathbf{M}^{(X)} = \tilde{Z}^X_{L_e}$. It is worth noting that we do not force the direct fusion of different scale features at the encoding end, but use the SSA strategy. This ensures that each scale preserves its discriminative representation independently, preventing premature cross-scale coupling and laying the foundation for semantic alignment.

Then, we employ three parallel Transformer decoders, each dedicated to one visual scale. For scale $X\in\{S,M,L\}$, the decoder takes category queries Q as input and attends to the corresponding visual memory M^X as key-value pairs. And for clarity, the category query matrix Q used by all decoders is constructed by the Dual-source Semantic Prior (DSP) module detailed in Sec. 2.4. Each decoder layer performs self-attention among queries, capturing label co-occurrence patterns, followed by cross-attention between queries and scale-specific features:

$$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{D}}\right)V \tag{4}$$

where, $Q\in\mathbb{R}^{C\times D}$ and $K,V\in\mathbb{R}^{N_X\times D}$. Through cross-attention, the query vector is able to adaptively retrieve the most relevant visual cues for the category from the sequence of features at the corresponding scale. After L_d stacked layers, the decoder outputs: $O^X=Q^{L_d}$, where $O^X\in\mathbb{R}^{C\times D}$ denotes the scale-specific category embeddings. Thus, each category c obtains embeddings o^L_c, o^M_c, o^S_c from different scales, aligned by shared queries. As a result, we

aligned the responses of semantic queries at each scale by category on the decoding side, laying the foundation for subsequent fusion.

2.4. Dual-source Semantic Prior

On the semantic side, we aim to provide each category with a query vector enriched with semantic priors. To this end, we adopt a Dual-source Semantic Prior (DSP) Module, which combines pre-trained word embeddings with label embeddings. For C categories, we firstly choose one from two embedding tables,

$$Q_{\text{Word}} = \begin{cases} Q_{\text{bert}}, & \text{if embed\_type = BERT} \\ Q_{\text{glove}}, & \text{if embed\_type = GloVe} \end{cases} \tag{5}$$

where, $Q_{\text{Word}} \in \mathbb{R}^{C\times D}$ and embed_type is the configuration parameter given by command. Considering the different functions of pre-trained embeddings: for example, BERT [23] provides contextualized representations and GloVe [24] provides global statistical information, we may flexibly select either to address different scenarios.

During training, one embedding table is loaded from pre-trained weights while the other is randomly initialized, depending on the configuration parameterembed_type.

$$Q = Q_{Word} + Q_{Label} \tag{6}$$

where, Q_{Label} are learnable label embeddings for specific task, and Q,$Q_{Label}\in\mathbb{R}^{C\times D}$. Both tables are updated via backpropagation. This hybrid design combines stable semantic priors with task-adaptive flexibility, ensuring robust and generalizable label queries.

Particularly, the fused category query matrix Q remains shared across all scales of the Transformer decoder. It implies that the same category consistently employs a unified semantic identity for cross-scale alignment across different granularity levels (i.e., coarse, medium, and fine). In other words, this mechanism guarantees semantic consistency while enhancing the robustness and generalizability of query representations.

2.5. Final Classification

To integrate outputs from different scales, we concatenate the embeddings of each category c:

$$o_c^{cat} = [o_c^L \parallel o_c^M \parallel o_c^S] \tag{7}$$

where, $o_c^{cat}\in\mathbb{R}^{3D}$ is the embeddings concatenated from three scales to form a unified representation. Then a learnable linear fusion layer maps this vector back to dimension D:

$$\hat{o}_c = W_{fuse}(o_c^{cat}) \tag{8}$$

where, W_{fuse} is a learnable fusion projection matrix and $\hat{o}_c \in \mathbb{R}^D$ is the final fused representation. Finally, we introduce a binary prediction header for the fused representation $\widehat{o_c}$ for each category. Here $\widehat{o_c}$ is mapped to a scalar score using a fully-connected layer, and the probability of the category's existence is obtained through Sigmoid activation:

$$p_c = \sigma(f_c(\hat{o}_c)) \tag{9}$$

The decoding process runs in parallel across the three scales, with complexity proportional to the sequence length N_X of each scale. Compared with simple concatenation-based multi-scale decoding, our design improves both efficiency and modularity while maintaining category-level alignment. This explicit per-category multi-scale fusion ensures strict semantic alignment across scales, avoiding the dominance of large-scale features and improving recognition of small and long-tailed categories.

3. EXPERIMENTS

In this section, we evaluate the proposed multi-label image classification method quantitatively and qualitatively.

3.1. Experimental Settings

To verify the generalization performance of our model, we conducted experiments based on widely used multi-label image classification benchmarks: MS-COCO [25] and PASCAL VOC 2007 [26]. We adopted ConvNeXt-B as the backbone and trained our model on a single NVIDIA V100 GPU. The training schedule is set to 100 epochs with a batch size of 16. We use the AdamW optimizer with an initial learning rate of 1×10^{-5} and a weight decay of 1×10^{-4}. A "ReduceLROnPlateau" scheduler is applied, with linear warmup for the first 2 epochs. The maximum number of training epochs is 100, and early stopping is used with a patience of 4. The loss function is the Asymmetric Loss (ASL) with hyperparameters γ_{neg}=4, γ_{pos}=1, and τ=0.

Following prior works [17, 18, 27], we also use the following metrics: Average Precision (AP), mean AP (mAP), Overall Precision (OP), Overall Recall (OR), Overall F1 (OF1), per-Category Precision (CP), per-Category Recall (CR), and per-Category F1 (CF1). They are computed as: $mAP=\frac{\sum_{c=1}^{C}\text{AP}_c}{C}$; $OP=\frac{\sum_i M_c^i}{\sum_i M_p^i}$; $OR=\frac{\sum_i M_c^i}{\sum_i M_g^i}$; $OF1=\frac{2\times OP\times OR}{OP+OR}$, $CP=\frac{1}{C}\sum_i \frac{M_c^i}{M_p^i}$; $CR=\frac{1}{C}\sum_i \frac{M_c^i}{M_g^i}$; $CF1=\frac{2\times CP\times CR}{CP+CR}$, where c, and C denote current specific class, and number of all classes respectively.

3.2. Comparison with State-of-the-Art Methods

We compare our method against a variety of state-of-the-art approaches, including CNN-based models (e.g., ML-GCN [28]), graph-based models (e.g., SSGRL [7]), and Transformer-based methods (e.g., Q2L [17], TSFormer [18]).

3.2.1. Performance on MS-COCO

Table 1. Comparison with state-of-the-art methods on MS-COCO.

Method	mAP	CP	CR	CF1	OP	OR	OF1
ML-GCN [28]	83.0	85.1	72.0	78.0	85.8	75.4	80.3
SSGRL [7]	83.8	89.9	68.5	76.8	91.3	70.8	79.7
TSGCN [29]	83.5	81.5	72.3	76.7	84.9	75.3	79.8
ML-Decoder (TResNet-L) [30]	**90.0**	–	–	–	–	–	–
ML-VPT [31]	86.4	79.7	80.3	80.0	81.6	82.5	82.0
Q2L (TResNet-L) [17]	89.2	86.3	81.4	83.8	86.5	83.3	84.9
M3TR [27]	87.5	88.4	77.2	82.5	88.3	79.8	83.8
TSFormer [18]	88.9	88.3	79.2	83.5	88.5	81.5	84.9
PatchCT [32]	88.3	83.3	**82.3**	82.6	84.2	**83.7**	83.8
MSAT (ours)	**90.0**	**90.9**	79.6	**84.9**	**91.5**	81.1	**86.0**

Table 1 presents the comparison results when MS-COCO is used. We observe that MSAT achieves 90.0% mAP, matching the CNN-based ML-Decoder (90.0%) and surpassing other Transformer models (e.g., 88.9% of TSFormer). MSAT also demonstrates superior performance on precision/recall metrics, with CP reaching 90.9%, the highest among compared methods, showing balanced improvements across categories. The CF1 score achieves 84.9%, improving by about 1.4 points over TSFormer. Overall precision OP reaches 91.5%, about 3% higher than TSFormer. Although our model does not achieve the highest category-wise or overall recall

Table 2. Comparison with state-of-the-art methods on VOC 2007.

Method	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mAP
ML-GCN [28]	99.5	98.5	98.6	98.1	80.8	94.6	97.2	98.2	82.3	95.7	86.4	98.2	98.4	96.7	99.0	84.7	96.7	84.3	98.9	93.7	94.0
SSGRL [7]	99.5	97.1	97.6	97.8	82.6	94.8	96.7	98.1	78.0	97.0	85.6	97.8	98.3	96.4	98.8	84.9	96.5	79.8	98.4	92.8	93.4
TSGCN [29]	98.9	98.5	96.8	97.3	87.5	94.2	97.4	97.7	84.1	92.6	89.3	98.4	98.0	96.1	98.7	84.9	96.6	87.2	98.4	93.7	94.3
ML-Decoder [30]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	96.6
ML-VPT [31]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	95.6
Q2L [17]	99.9	98.9	99.0	98.4	87.7	**98.6**	**98.8**	99.1	84.5	98.3	89.2	99.2	99.2	**99.2**	99.3	90.2	98.8	88.3	99.5	95.5	96.1
M3TR [27]	99.9	99.3	99.1	99.1	84.0	97.6	98.0	99.0	85.9	99.4	93.9	99.5	99.4	98.5	99.2	90.3	99.7	91.6	99.8	96.0	96.5
TSFormer [18]	**100.0**	99.2	99.2	98.6	86.4	97.2	98.4	98.9	88.9	**99.5**	95.3	**99.7**	99.6	99.1	**99.4**	90.0	99.6	93.7	**99.9**	96.7	97.0
PatchCT [32]	**100.0**	99.4	98.8	**99.3**	87.2	**98.6**	**98.8**	99.2	87.2	99.0	**95.5**	99.4	**99.7**	98.9	99.1	**91.8**	99.5	**94.5**	99.5	96.3	97.1
MSAT (ours)	**100.0**	**99.6**	**99.4**	99.1	**89.1**	98.0	98.5	**99.4**	**91.3**	**99.5**	95.2	99.6	99.4	99.1	99.3	90.8	**100.0**	94.1	99.5	**97.2**	**97.4**

(CR/OR) due to its precision-oriented strategy, it attains consistently high mAP, CF1, and precision scores.

3.2.2. Performance on VOC2007

Table 2 shows the experimental results on the VOC2007. MSAT achieves a new state-of-the-art mAP of 97.4%, slightly outperforming PatchCT (97.1%) and TSFormer (97.0%). It attains 100% AP in certain categories (e.g., "aero") and shows notable gains on challenging classes like "bottle" and "chair", which benefit from fine-grained multi-scale alignment. While advantages are less pronounced on larger or common objects (e.g., "bus", "dog"), MSAT maintains balanced and robust performance across all categories, demonstrating the effectiveness of its semantic-aligned multi-scale design.

3.3. Ablation Study

Table 3. Ablation results on VOC 2007.

SSA Module	Shared Queries	Dual-Emb	mAP
×	✓	✓	97.1
✓	×	✓	96.8
✓	✓	w/o BERT	97.0
✓	✓	w/o Label	97.2
✓	✓	✓	**97.4**

To validate the contribution of each component, we conduct ablation study on the VOC 2007 dataset. Table 3 shows the results. We can see that the full model achieves 97.4% mAP. When we replace the independent scale-specific encoders with a single shared encoder (i.e., removing SSA), the value drops to 97.1%, confirming the importance of preserving scale-discriminative features. In addition, when we remove shared queries from DSP across decoders(using separate query sets per scale) the vaule drops to 96.8%, emphasizing the important role of semantic consistency across scales. In addition, when we remove the dual-source embedding, i.e., using only label embeddings (97.0%) or only word embeddings (97.2%), the performance also reduces, demonstrating the benefit of integrating both semantic sources. As a whole, these results confirm that SSA, shared queries from DSP, and dual embeddings are essential and complementary to the model's performance.

3.4. Visualization Results

We visualize class activation maps (CAM), which reflect class-discriminative cues rather than precise object boundaries. As shown in Fig. 2, the large-scale branch mainly focuses on global semantics (e.g., fork, clock, bicycle, umbrella; see the 1-4 row, respectively),

Fig. 2. Visualization of Cross-Scale Complementarity in SSA: (a) input, (b) large_scale, (c) medium_scale, (d) medium_scale, (e) fused.

while the medium- and small-scale branches highlight fine-grained structures (e.g., cellphone, bicycle, traffic lights, car; see also the 1-4 row, respectively). These details are often overlooked in the large-scale map, but the SSA module ensures that the model can re-attend to such subtle structures. It is important to achieve better performance. The fused CAM, which simultaneously covers both types of features, further verifies our SSA design philosophy, namely, preserving scale-specific evidence and integrating it at the category level.

4. CONCLUSION

Existing Transformer-based methods for multi-label image classification often rely on single-scale features or randomly initialized queries, limiting their ability to model diverse object scales and semantic consistency. In this paper, we have proposed MSAT, a multi-scale semantic-aligned Transformer framework designed to address the challenges of multi-label image classification. By introducing the SSA module and the DSP module, our model effectively captures objects at diverse scales while leveraging rich semantic priors from pre-trained word embeddings. Extensive experiments on MS-COCO and PASCAL VOC 2007 demonstrate that MSAT achieves competitive performance, confirming the importance of explicit semantic alignment and multi-scale modeling in complex visual recognition tasks.

5. REFERENCES

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Commun. ACM*, pp. 84–90, 2017.

[2] D. Yu, X. Liu, and B. Yang, "Zero-shot image classification with logic adapter and rule prompt," *in WWW*, 2024.

[3] T. B. Sifuentes, A.Nazari, I.Razzak, M. R. Bouadjenek, A. R. Kelly, D. Ierodiaconou, and Elizabeth S. Oh, "Marine-tree: A large-scale marine organisms dataset for hierarchical image classification," *in CIKM*, 2022.

[4] E. Gibaja and S. Ventura, "Multi-label learning: A review of the state of the art and ongoing research," *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, pp. 411–444, 2014.

[5] M.A.Shah, K.Viswanathan, C.Lu, A.Fuxman, Z.Li, A.Timofeev, C.Jia, and C.Sun, "Inferring context from pixels for multimodal image classification," *in CIKM*, 2019.

[6] A.Balayn, P.Soilis, C.Lofi, J.Yang, and A.Bozzon, "What do you mean? interpreting image classification with crowdsourced concept extraction and analysis," *in WWW*, 2021.

[7] T. Chen, M. Xu, X. Hui, H. Wu, and L. Lin, "Learning semantic-specific graph representation for multi-label image recognition," in *ICCV*, 2019, pp. 522–531.

[8] Z.Ding, A.Wang, H.Chen, Q.Zhang, P.Liu, Y.Bao, W.P.Yan, and J.Han, "Exploring structured semantic prior for multi label recognition with incomplete labels," *in CVPR*, pp. 3398–3407, 2023.

[9] A. Vaswani, N. M. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *NeurIPS*, 2017.

[10] J. Lanchantin, T. Wang, V. Ordonez, and Y. Qi, "General multi-label image classification with transformers," in *CVPR*, 2021, pp. 16473–16483.

[11] X. Cheng, H. Lin, X. Wu, D. Shen, F. Yang, H. Liu, and N. Shi, "Mltr: Multi-label classification with transformer," in *ICME*, 2022, pp. 1–6.

[12] R.Liu, J.Huang, T.H. Li, and G.Li, "Causality compensated attention for contextual biased visual recognition," in *ICLR*, 2023.

[13] J.Ye, J.He, X.Peng, W.Wu, and Y.Qiao, "Attention-driven dynamic graph convolutional network for multi-label image recognition," in *ECCV*, 2020.

[14] Y. Wu, S. Feng, G. Zhao, and Y. Jin, "Transformer driven matching selection mechanism for multi-label image classification," *IEEE Trans. Circuits Syst. Video Technol.*, pp. 924–937, 2024.

[15] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *ICLR*, 2021.

[16] C. Chen, Q. Fan, and R. Panda, "Crossvit: Cross-attention multi-scale vision transformer for image classification," *in ICCV*, pp. 347–356, 2021.

[17] S. Liu, L. Zhang, X. Yang, H. Su, and J. Zhu, "Query2label: A simple transformer way to multi-label classification," *arXiv*, 2021.

[18] X. Zhu, J. Cao, J. Ge, W. Liu, and B. Liu, "Two-stream transformer for multi-label image classification," in *MM*, 2022, pp. 3598–3607.

[19] X.Wang, L.Feng, D.Wang, and P.Niu, "A robust wavelet domain multi-scale texture descriptor for image classification," *Expert Syst. Appl.*, vol. 265, pp. 126000, 2024.

[20] G.Zhang, Z.Luo, Y.Yu, Z.Tian, J.Zhang, and S.Lu, "Towards efficient use of multi-scale features in transformer-based object detectors," *in CVPR*, pp. 6206–6216, 2022.

[21] W. Zhou, P. Dou, T. Su, H. Hu, and Z. Zheng, "Feature learning network with transformer for multi-label image classification," *Pattern Recognit.*, p. 109203, 2023.

[22] Z. Liu, H. Mao, C. Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," in *CVPR*, 2022, pp. 11966–11976.

[23] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *NAACL-HLT*, 2019, pp. 4171–4186.

[24] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *EMNLP*, 2014, pp. 1532–1543.

[25] T. Y. Lin, M. Maire, S. J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *ECCV*, 2014.

[26] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge: A retrospective," *Int. J. Comput. Vis.*, vol. 111, pp. 98–136, 2014.

[27] J. Zhao, Y. Zhao, and J. Li, "M3tr: Multi-modal multi-label recognition with transformer," in *ACM MM*, 2021.

[28] Z. M. Chen, X. S. Wei, P. Wang, and Y. Guo, "Multi-label image recognition with graph convolutional networks," in *CVPR*, 2019, pp. 5172–5181.

[29] J. Xu, H. Tian, Z. Wang, Y. Wang, W. Kang, and F. Chen, "Joint input and output space learning for multi-label image classification," *IEEE Trans. Multimedia*, vol. 23, pp. 1696–1707, 2021.

[30] T. Ridnik, G. Sharir, A. Ben-Cohen, E. Ben-Baruch, and A. Noy, "Ml-decoder: Scalable and versatile classification head," in *WACV*, 2021, pp. 32–41.

[31] L. Ma, S. Xu, M. K. Xie, L. Wang, D. Sun, and H. Zhao, "Correlative and discriminative label grouping for multi-label visual prompt tuning," in *CVPR*, 2025, pp. 25434–25443.

[32] M. Li, D. Wang, X. Liu, Z. Zeng, R. Lu, B. Chen, and M. Zhou, "Patchct: Aligning patch set and label set with conditional transport for multi-label image classification," in *ICCV*, 2023, pp. 15302–15312.