Memory-Efficient Transfer Learning with Fading Side Networks via Masked Dual Path Distillation

Supplementary Material

In the Supplementary Material, more details and experiments are organized as follows, we additionally provide more introduction of **Datasets and Metrics** in Sec. A, the **Implementation Details** in Sec. B, the rigorous proof of **Back-propagation through Large Backbone** in Sec. C, the expanded explanation of **Baselines** in Sec. D, and **More Ablation Studies** on several modules in Sec. E.

A. Datasets and Metrics

Image-Text Retrieval (ITR): We employ Flickr30k [46] and MSCOCO [25] for image-text matching task. The Flickr30k dataset contains 31,783 colloquial images with 158,915 captions, partitioned into 29,783 training images, 1,000 validation images, and 1,000 test images. Its captions emphasize explicit compositional semantics (*e.g.*, object-attribute-spatial relationships), posing fine-grained matching challenges. The MSCOCO dataset comprises 123,287 complex scene images annotated with 616,435 captions, and is divided into 113,287 training images, 5,000 validation images, 5,000 test images, which is the standard Karpathy split. MSCOCO requires deeper relational reasoning due to contextual object interactions and scene dynamics. Both datasets evaluate Image-to-Text (I-T), Text-to-Image (T-I) retrieval using Recall@1, with additionally employing RSum (*i.e.*, sum of all six Recall@K scores where K=1,5,10) as the holistic metric reflecting their compositional complexity.

Video-Text Retrieval (VTR): For video-text retrieval, we employ MSR-VTT [45] and MSVD [4]. The MSR-VTT (*i.e.*, Microsoft Research Video to Text) dataset contains 10,000 web video clips (total duration 41 hours) sourced from YouTube, with each video annotated with 20 English captions. Following the standard split, we employ the 1k-A protocol, where 9,000 videos with all corresponding captions for training, and 1,000 pairs are for testing. Characterized by high diversity across 20 categories (*e.g.*, sports, music, news), its captions describe complex temporal dynamics and object interactions. Moreover, The MSVD (*i.e.*, Microsoft Video Description) dataset comprises 1,970 short video clips with approximately 80,000 multilingual captions, partitioned into 1,200 training videos, 100 validation videos, and 670 testing videos. Noted for fine-grained temporal alignment challenges, MSVD captions emphasize precise action-object localization. Both datasets evaluate Video-to-Text (V-T), Text-to-Video (T-V) retrieval using Recall@1 and RSum for evaluation.

Question Answering (VQA&GQA): For question answering tasks, we evaluate on VQAv2 [14] and GQA [19]. The VQAv2 dataset addresses prior language bias issues by pairing 204,721 COCO images with 1,105,904 questions. It employs the standard split: 82,783 images with 443,757 questions for training, 40,504 images with 214,354 questions for validation, and 81,434 images with 447,793 questions for testing. Questions require diverse reasoning about object attributes, actions, and scene context. The GQA dataset features 113,018 images with 22,669,678 questions generated from scene graphs to ensure compositional rigor. Its balanced split contains 943,000 training questions (∼70%), 132,062 validation questions (∼10%)), and 264,159 testing questions (∼20%), emphasizing structural reasoning over 1,704 object categories. We evaluate performance on both Test-Dev and Test-Std splits via the official EvalAI system.

Visual Grounding (VG): We utilize the RefCOCO, RefCOCO+ [47], and RefCOCOg [28] derived from MSCOCO images for visual grounding. RefCOCO contains 19,994 images with 50,000 bounding boxes annotated by 142,210 expressions, whose standard UNC split comprises 120,624 training expressions, 10,834 validation expressions, and 5,675/5,095 Test A/B expressions. Test A focuses on bounding boxes containing person instances, while Test B involves non-person objects. RefCOCO+ shares the same image set but introduces stricter constraints: 49,856 referred objects with 141,564 expressions that explicitly prohibit location words, divided into 120,191 training, 10,758 validation, and 5,726/4,889 Test A/B expressions. RefCOCOg differs substantially with 26,711 images, 54,822 referred objects, and 104,560 longer, grammatically complex expressions, while they are categorized into train, validation, and test, with 85,474, 7,323, and 9,592 samples. Primary evaluation uses precision@0.5 to measure localization accuracy of predicted bounding boxes against human annotations.

Language-only: For language-only task, we employ the General Language Understanding Evaluation (GLUE) benchmark [40] consolidates eight NLP tasks into four core categories, consisting of *linguistic acceptability* (CoLA [41]), *sentiment analysis* (SST-2 [37]), *similarity and paraphrase* (MRPC [10], QQP, STS-B [3]), and *natural language inference* (MNLI [42], QNLI [35], RTE [2]). Evaluation employs task-specific metrics: classification Accuracy metric for SST-2, MNLI, RTE, and QNLI; F1-score augmented with accuracy for MRPC and QQP; Matthew's Correlation for the class-imbalanced data of CoLA; and

Table A. Detailed Hyper-parameters of MDPD on ITR, VTR, VQA, GQA, and VG tasks. Among them, *AdamW* is adopted as the optimizer uniformly.

Task	Model	Learning Rate	Optimizer (β_1, β_2, Weight Decay)	Batch Size	Total Epochs	Warmup Strategy
ITR	VSE∞	5×10^{-4}	0.9, 0.999, 1×10^{-2}	112	25	linear
VTR	CLIP4Clip	1×10^{-4}	0.9, 0.98, 1×10^{-2}	128	5	cosine
VQA	CLIP-ViL	5×10^{-4}	0.9, 0.999, 1×10^{-2}	256	5	linear
GQA	CLIP-ViL	1×10^{-4}	0.9, 0.999, 1×10^{-2}	256	5	linear
VG	MDETR	5×10^{-4}	0.9, 0.999, 0	8	10	linear

Pearson-Spearman Correlation for similarity scoring of STS-B.

Vision-only: For vision-only task, we employ the Visual Task Adaptation Benchmark (VTAB-1K) [48] systematically evaluates transfer learning capabilities across 19 diverse vision datasets unified under a standardized low-data regime, categorized into three distinct task types: (1) *Natural* tasks (CIFAR-100 [22], Caltech101 [12], DTD [7], Flowers102 [30], Pets [31], SVHN [29], Sun397 [43]) featuring object-centric photographs with moderate complexity; (2) *Specialized* tasks (Patch Camelyon [39], EuroSAT [16], Resisc45 [6], Retinopathy) [15] comprising domain-specific imagery and medical images; (3) *Structured* tasks (Clevr/count [20], Clevr/distance [20], DMLab [1], KITTI-Dist [13], dSprites/location, dSprites/orientation, SmallNORB/azimuth [23], SmallNORB/elevation [23]) emphasizing geometric relationships and spatial reasoning. Each dataset provides 1,000 training images with predefined validation/test splits. Evaluation reports Top-1 Accuracy metric for classification tasks, with final performance aggregated via uniform averaging across all 19 datasets.

B. Implementation Details

For vision-language (VL) tasks, Table A comprehensively details the hyper-parameter configurations. More specifically, for ITR task using VSE∞, we set the batch size to 112, and maintain consistency with pre-trained models for all other tasks, while scale learning rate by a factor of 10, and set the reduction factor and mask rate λ to 2 and 0.5, respectively. For GLUE benchmark evaluations, we set learning rate, reduction factor and batch size to 3×10^{-3}, 8, and 100, respectively. Consistent with the methodology in LST [38], we implement the layer-dropping strategy: for the T5-base architecture, this entails removal of the 0^{th}, 4^{th}, and 8^{th} encoder/decoder layers; for T5-large, we omit all even-indexed encoder and decoder layers.

Beyond task-specific hyper-parameters, we carefully calibrate the weighting coefficients for multi-objective optimization. Specifically, we assign the logits-based and feature-based distillation loss in deep or shallow layers with weights of 1×10^{-4}, 6×10^{-5} and 4×10^{-5}, respectively, while 1 for the primary Supervised Fine-Tuning objective. This balanced scheme ensures commensurate contribution from each optimization component during gradient updates. All experiments are conducted on an NVIDIA GeForce RTX 3090Ti GPU.

C. Back-propagation through Large Backbone

Consider a neural network comprising L sequential layers, where the transformation at the i^{th} layer is defined as $f_i(\mathbf{x}) = \sigma_i(\mathbf{W}_i\mathbf{x} + \mathbf{b}_i)$. This composite function depends on the previous layer's output, parameterized by the weight matrix $\mathbf{W}_i$, bias vector $\mathbf{b}_i$, and nonlinear activation function $\sigma_i(\cdot)$. We denote the pre-activation output as $\mathbf{z}_{i+1}$ and the post-activation output as $\mathbf{a}_{i+1}$, establishing the layer-wise propagation:

$$\mathbf{a}_{i+1} = \sigma_i(\mathbf{z}_{i+1}) = \sigma_i(\mathbf{W}_i\mathbf{a}_i + \mathbf{b}_i). \tag{1}$$

Network parameters are optimized via stochastic gradient descent (SGD) by minimizing a scalar loss function $\mathcal{L}$ applied to the final layer output. The backpropagation algorithm computes gradients for $\mathbf{W}_i$ and $\mathbf{b}_i$ through recursive application of the multivariate chain rule:

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \mathbf{W}_i} &= \frac{\partial \mathcal{L}}{\partial \mathbf{a}_{i+1}} \frac{\partial \mathbf{a}_{i+1}}{\partial \mathbf{z}_{i+1}} \frac{\partial \mathbf{z}_{i+1}}{\partial \mathbf{W}_i} = \frac{\partial \mathcal{L}}{\partial \mathbf{a}_{i+1}} \sigma_i' \mathbf{a}_i^\top, \\ \frac{\partial \mathcal{L}}{\partial \mathbf{b}_i} &= \frac{\partial \mathcal{L}}{\partial \mathbf{a}_{i+1}} \frac{\partial \mathbf{a}_{i+1}}{\partial \mathbf{z}_{i+1}} = \frac{\partial \mathcal{L}}{\partial \mathbf{a}_{i+1}} \sigma_i', \end{aligned} \tag{2}$$

where $\sigma_i' \equiv \frac{d\sigma_i}{d\mathbf{z}_{i+1}}$ denotes the activation gradient, and $\frac{\partial \mathcal{L}}{\partial \mathbf{a}_{i+1}}$ represents the upstream gradient from subsequent layers. This upstream gradient is recursively computed via backward propagation from layer $i+2$:

$$\frac{\partial \mathcal{L}}{\partial \mathbf{a}_{i+1}} = \frac{\partial \mathcal{L}}{\partial \mathbf{a}_{i+2}} \frac{\partial \mathbf{a}_{i+2}}{\partial \mathbf{z}_{i+2}} \frac{\partial \mathbf{z}_{i+2}}{\partial \mathbf{a}_{i+1}} = \frac{\partial \mathcal{L}}{\partial \mathbf{a}_{i+2}} \sigma_{i+1}' \mathbf{W}_{i+1}^\top. \tag{3}$$

As formalized in Equations (2) and (3), the backpropagation algorithm incurs substantial computational overhead

due to floating-point operations (FLOPs) required for two critical gradient components: 1) The activation gradients $\{\mathbf{a}\}$ corresponding to updated parameters $\{\mathbf{W}\}$, and 2) The activation derivatives $\{\sigma'\}$ that must be cached throughout the computational graph, where $\{\cdot\}$ denotes sets of activations, parameters, or gradients. Existing Parameter-Efficient Transfer Learning (PETL) techniques, including Adapter [17], Prompt-Tuning [24], and LoRA [18], mitigate memory footprint by reducing the parameter update set $|\{\mathbf{W}\}|$ through learning only sparse parameter subsets, where $|\{\cdot\}|$ means the size of set $\{\cdot\}$. Consequently, the memory allocated for activation storage $|\{\mathbf{a}\}|$ proportionally decreases. However, the dominant computational burden during backpropagation stems from computing gradient terms involving $\{\sigma'\}$ - the derivatives of activation functions. Crucially, $|\{\sigma'\}|$ remains undiminished in these methods since:

$$|\{\sigma'\}| = \sum_{i=1}^{L} \dim(\mathbf{z}_i). \tag{4}$$

This persistence occurs because PETL methods typically introduce trainable parameters into network inputs or intermediate structures while keeping the backbone frozen. Nevertheless, they still require full computation of σ' across all backbone operations, necessitating: 1) Complete evaluation of activation gradients through the entire computational graph, 2) Storage of intermediate derivatives at each layer, and 3) Backpropagation through all nonlinear transformations.

Since activation dimensions generally satisfy $|\{\mathbf{a}\}| = |\{\sigma'\}|$ (barring dimensionality-altering activations), the theoretical memory reduction ceiling becomes:

$$\text{Memory}_{\text{BP}} = \underbrace{|\{\mathbf{a}\}|}_{\text{reduced}} + \underbrace{|\{\sigma'\}|}_{\text{unchanged}} \leq 50\% \text{ reduction}. \tag{5}$$

Therefore, while PETL methods reduce parameter update costs, they still incur substantial FLOPs and memory requirements proportional to backbone complexity, as full error backpropagation through frozen layers remains mandatory.

Based on the foregoing computational analysis, **side network** is proposed as a memory-efficient alternative. This lightweight network maintains same structure to the backbone network while scaling all weight matrices and hidden state dimensions by a reduction factor $r \geq 2$. Thus, the original backpropagation memory footprint $|\{\mathbf{a}\}| + |\{\sigma'\}|$ is fundamentally transformed in this paradigm. Crucially, the side network *decouples* from the backbone's computational graph during backpropagation, requiring gradient computation only through its own structure. Consequently, its memory consumption reduces to:

$$\text{Memory}_{\text{BP}}^{\text{side}} = \frac{|\{\mathbf{a}\}| + |\{\sigma'\}|}{r} \tag{6}$$

This yields a critical comparative advantage: when $r > 2$, the side network achieves strictly lower memory consumption than the theoretical minimum of Parameter-Efficient Transfer Learning (PETL) methods, which remain bounded by:

$$\text{Memory}_{\text{BP}}^{\text{PETL}} \geq \frac{|\{\mathbf{a}\}| + |\{\sigma'\}|}{2} \tag{7}$$

Thus, side networks establish a new efficiency frontier for Memory-Efficient Transfer Learning (METL), with memory savings growing linearly with r while maintaining functional capacity.

D. Baselines

We select various transfer paradigms for comprehensive and challenging validation:

-*VSE*∞ [5] with BERT-base [9] model and ResNeXt-101(32×8d) [44] backbone pre-trained on Instagram (WSL) on Flickr30K [46], MSCOCO1K and MSCOCO5K [25] for the **ITR** task;

-*CLIP4Clip* [27] with the pre-trained CLIP [33] using Text Transformer [32] and ViT-B/32 [11] on MSR-VTT [45] and MSVD [4] for the **VTR** task;

-*CLIP-ViL* [36] that applies the CLIP image backbone [33] and encodes the text into word embeddings, followed by a cross-modal Transformer on VQAv2 [14] and GQA [19] for the **QA** task;

-*MDETR* [21] that integrates a pre-trained ResNet-101, RoBERTa-base [26] with an encoder-decoder Transformer on RefCOCO, RefCOCO+ [47] and RefCOCOg [28] for the **VG** task;

-*T5-series* [34] that imports text encoder and auto-regressive decoder, while following [38], we drop 6, 24 layers of side network (3, 12 layers each in encoder and decoder) for *T5-base* and *T5-large* on GLUE benchmark [40] for the **NLP** task;

ViT-base [11] witch consists of 86 million parameters, while pre-trained on ImageNet-21K [8] is the most commonly used backbone across prior works (*e.g.*, image classification, video classification, *etc.*), and is adopted on VTAB-1K [48] for the **CV** task.

E. More Ablation Studies

We extensively conduct more ablation studies to verify the effectiveness of our proposed method and the selected hyper-parameters.

Influence of the mask ratio λ. To explore the impact of the mask ratio and verify the effectiveness of adopting $\lambda = 0.5$, we conduct an ablation study of λ on ITR task. As demonstrated in Figure A, the results reveal that the performance of our method is sensitive to λ. Specifically,

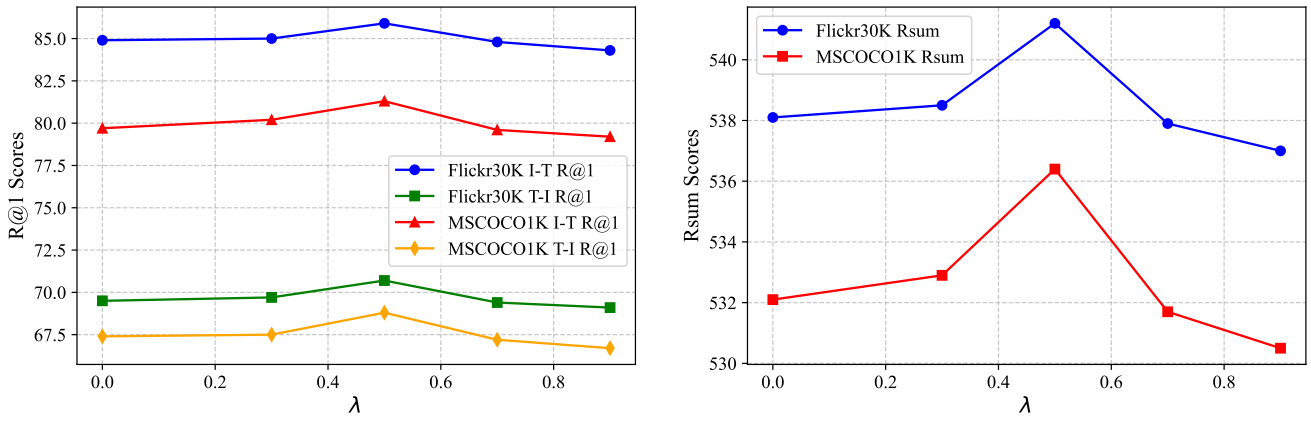

Figure A. Ablation study on the effect of the mask rate λ in our method. (**Left**) The Rsum (%) performance on Flickr30K and MSCOCO1K datasets. (**Right**) The R@1 (%) performance for sentence retrieval ("I-T") and image retrieval ("T-I") on Flickr30K and MSCOCO1K datasets.

Table B. Ablation results (%) on the generation blocks for feature transformation in the student network on distinct datasets for ITR task. The best results are highlighted in **bold**.

Generation Blocks	Params. (M) ↓	Memory (G) ↓	Flickr30K			MSCOCO1K			MSCOCO5K		
			I-T ↑	T-I ↑	Rsum ↑	I-T ↑	T-I ↑	Rsum ↑	I-T ↑	T-I ↑	Rsum ↑
Self-Attention	12.6	16.0	84.7	69.5	537.4	79.8	67.2	531.5	60.4	46.4	444.9
Convolution	**12.4**	**15.9**	**85.9**	**70.7**	**541.2**	**81.3**	**68.8**	**536.4**	**61.9**	**46.8**	**448.1**

Table C. Ablation results (%) on adopting Hierarchical Feature-based Distillation strategy across shallow **(Top)**, deep **(Middle)**, and all **(Bottom)** layers on distinct datasets for ITR task. The best results are highlighted in **bold**. and the second best results are underlined.

	Layers	Params. (M) ↓	Memory (G) ↓	Flickr30K			MSCOCO1K			MSCOCO5K		
				I-T ↑	T-I ↑	Rsum ↑	I-T ↑	T-I ↑	Rsum ↑	I-T ↑	T-I ↑	Rsum ↑
Imitation	1	**12.4**	**15.8**	83.4	68.5	536.8	78.6	65.8	529.1	59.1	44.7	441.3
	1,2	**12.4**	**15.8**	83.7	68.8	536.7	78.9	66.1	529.2	59.7	44.3	441.6
	5,6	**12.4**	**15.8**	84.6	69.2	537.4	79.6	67.1	531.2	60.5	45.6	443.8
	$1 \sim 6$	**12.4**	15.9	84.5	69.5	537.6	79.4	67.3	530.9	60.8	45.5	444.0
	All	12.4	15.9	84.9	69.7	538.3	80.0	67.8	532.6	61.4	45.8	445.7
Generation	7,8	**12.4**	**15.8**	84.6	69.2	537.6	79.3	66.5	530.2	60.8	44.5	442.9
	11,12	**12.4**	**15.8**	85.0	69.5	538.4	79.8	67.1	531.3	61.4	44.8	444.3
	12	**12.4**	**15.8**	85.2	69.8	538.6	80.2	67.4	532.6	61.2	45.4	445.2
	$7 \sim 12$	**12.4**	15.9	84.6	69.3	538.1	79.9	67.2	531.6	60.7	45.2	444.1
	All	12.4	16.1	85.2	70.1	539.5	80.6	67.9	533.4	61.1	45.9	446.0
	All	**12.4**	15.9	**85.9**	**70.7**	**541.2**	**81.3**	**68.8**	**536.4**	**61.9**	**46.8**	**448.1**

the model achieves the highest accuracy when $\lambda = 0.5$ on both Flickr30K and MSCOCO1K datasets, indicating that $\lambda = 0.5$ optimally balances the preservation of critical information and the introduction of diversity during feature distillation. As λ increases beyond 0.5, the performance of our method gradually declines. This degradation can be attributed to the excessive masking, which leads to the loss of valuable information, thereby diminishing the effectiveness of feature distillation and impairing the student network's ability to learn useful information from the teacher network.

Necessity of generation blocks. We adopt a convolutional projector as the generation module for feature transformation in the student network. In order to validate its effectiveness, we conduct an ablation study comparing it with the self-attention mechanism. As shown in Table B, the convolutional projector achieves superior performance on ITR task, demonstrating its effectiveness in mitigating the feature discrepancies between the teacher and student networks. Beyond its advantage in accuracy, the convolutional projector also incurs remarkable advantages in computational efficiency, as it requires fewer parameters and consumes less training memory. These benefits stem from their fundamental differences in computational design. The self-attention mechanism computes attention scores for all input element pairs, leading to quadratic complexity with respect to the input size. This design significantly increases memory consumption and computational overhead during training. In contrast, the convolutional module operates on local receptive fields, capturing spatially localized features with a linear computational complexity with respect to the input size. This efficiency makes the convolutional projector computationally lightweight, thus being more suitable for applications on resource-constrained scenarios for large-scale tasks.

Necessity of distillation across all layers. To verify the necessity of performing feature distillation across all layers, we empirically investigate the performance of our method with different layer combinations when performing distillation. As shown in Table C, the results indicate that performing Hierarchical Feature-based Distillation strategy across all layers yields the highest accuracy on ITR task. Furthermore, when feature distillation is applied solely to the shallow layers (*i.e.*, layer 1$\sim$6), the performance is notably lower compared to distillation applied to the deep layers (*i.e.*, layer 7$\sim$12). The performance degradation is attributed to the fact that shallow layers contain less semantic information, which limits their contribution to the overall task. In contrast, deeper layers encode semantically rich features, leading to superior results. To further demonstrate the effectiveness of HFD independently, we also conduct ablation experiments with *imitation-only* and *generation-only* (*i.e.,* feature-based distillation is performed across all layers utilizing either imitation or generation methods), the results are indicated in gray. These results underscore the importance of leveraging features from all layers of the network during the distillation process, providing evidence that our approach effectively facilitates the knowledge transfer from the backbone to the side network.

References

[1] Charles Beattie, Joel Z Leibo, Denis Teplyashin, Tom Ward, Marcus Wainwright, Heinrich Küttler, Andrew Lefrancq, Simon Green, Víctor Valdés, Amir Sadik, et al. Deepmind lab. *arXiv preprint arXiv:1612.03801*, 2016. 2

[2] Luisa Bentivogli, Peter Clark, Ido Dagan, and Danilo Giampiccolo. The fifth pascal recognizing textual entailment challenge. *TAC*, 7(8):1, 2009. 1

[3] Daniel Cer, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio, and Lucia Specia. Semeval-2017 task 1: Semantic textual similarity-multilingual and cross-lingual focused evaluation. *arXiv preprint arXiv:1708.00055*, 2017. 1

[4] David Chen and William B Dolan. Collecting highly parallel data for paraphrase evaluation. In *Annual Meeting of the Association for Computational Linguistics*, pages 190–200, 2011. 1, 3

[5] Jiacheng Chen, Hexiang Hu, Hao Wu, Yuning Jiang, and Changhu Wang. Learning the best pooling strategy for visual semantic embedding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 15789–15798, 2021. 3

[6] Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10):1865–1883, 2017. 2

[7] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3606–3613, 2014. 2

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 248–255. Ieee, 2009. 3

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*, pages 4171–4186, 2019. 3

[10] Bill Dolan and Chris Brockett. Automatically constructing a corpus of sentential paraphrases. In *Third international workshop on paraphrasing (IWP2005)*, 2005. 1

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the International Conference on Learning Representations*, 2021. 3

[12] Li Fei-Fei, Robert Fergus, and Pietro Perona. One-shot learning of object categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(4):594–611, 2006. 2

[13] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013. 2

[14] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 6904–6913, 2017. 1, 3

[15] Ben Graham. Kaggle diabetic retinopathy detection competition report. *University of Warwick*, 22(9), 2015. 2

[16] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019. 2

[17] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *Proceedings of the International Conference on Machine Learning*, pages 2790–2799, 2019. 3

[18] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *Proceedings of the International Conference on Learning Representations*, 2022. 3

[19] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 6700–6709, 2019. 1, 3

[20] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2901–2910, 2017. 2

[21] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1780–1790, 2021. 3

[22] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 2

[23] Yann LeCun, Fu Jie Huang, and Leon Bottou. Learning methods for generic object recognition with invariance to pose and lighting. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages II–104. IEEE, 2004. 2

[24] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, 2021. 3

[25] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision*, pages 740–755, 2014. 1, 3

[26] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:.11692*, 364, 2019. 3

[27] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. Clip4clip: An empirical study of clip for end to end video clip retrieval. *arXiv preprint arXiv:2104.08860*, 2021. 3

[28] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the Computer Vision and Pattern Recognition*, pages 11–20, 2016. 1, 3

[29] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Baolin Wu, Andrew Y Ng, et al. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep learning and unsupervised feature learning*, page 7. Granada, 2011. 2

[30] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pages 722–729. IEEE, 2008. 2

[31] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3498–3505. IEEE, 2012. 2

[32] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 3

[33] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, and Jack Clark. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*, pages 8748–8763, 2021. 3

[34] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. 3

[35] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016. 1

[36] Sheng Shen, Liunian Harold Li, Hao Tan, Mohit Bansal, Anna Rohrbach, Kai-Wei Chang, Zhewei Yao, and Kurt Keutzer. How much can clip benefit vision-and-language tasks? *arXiv preprint arXiv:2107.06383*, 2021. 3

[37] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642, 2013. 1

[38] Yi-Lin Sung, Jaemin Cho, and Mohit Bansal. Lst: Ladder side-tuning for parameter and memory efficient transfer learning. In *Proceedings of the Advances in Neural Information Processing Systems*, pages 12991–13005, 2022. 2, 3

[39] Bastiaan S Veeling, Jasper Linmans, Jim Winkens, Taco Cohen, and Max Welling. Rotation equivariant cnns for digital pathology. In *International Conference on Medical image computing and computer-assisted intervention*, pages 210–218. Springer, 2018. 2

[40] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*, 2018. 1, 3

[41] Alex Warstadt, Amanpreet Singh, and Samuel R Bowman. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641, 2019. 1

[42] Adina Williams, Nikita Nangia, and Samuel R Bowman. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*, 2017. 1

[43] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3485–3492. IEEE, 2010. 2

[44] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the Computer Vision and Pattern Recognition*, pages 1492–1500, 2017. 3

[45] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5288–5296, 2016. 1, 3

[46] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014. 1, 3

[47] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *Proceedings of the European Conference on Computer Vision*, pages 69–85, 2016. 1, 3

[48] Xiaohua Zhai, Joan Puigcerver, Alexander Kolesnikov, Pierre Ruyssen, Carlos Riquelme, Mario Lucic, Josip Djolonga, Andre Susano Pinto, Maxim Neumann, Alexey Dosovitskiy, et al. A large-scale study of representation learning with the visual task adaptation benchmark. *arXiv preprint arXiv:1910.04867*, 2019. 2, 3