Mixed Autoencoder for Self-supervised Visual Representation Learning

Kai Chen[1] Zhili Liu[1,2] Lanqing Hong[2] Hang Xu[2] Zhenguo Li[2] Dit-Yan Yeung[1]
[1]Hong Kong University of Science and Technology [2]Huawei Noah's Ark Lab
{kai.chen,zhili.liu}@connect.ust.hk {honglanqing,xu.hang,li.zhenguo}@huawei.com
dyyeung@cse.ust.hk

A. Proof of Eq. (5)

In this section, we prove Eq. (5) with the terminologies maintained consistent with Sec. 3. We start to prove when $r = 0.5$ (*i.e.*, two clean images within a single group). Denote $\boldsymbol{X}_1, \boldsymbol{X}_2$ as two random variables representing two input images, and $\boldsymbol{M}$ as the random mask, which can be considered as a constant here since it is independently generated with $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$. Then, according to Eq. (1), the mixed input can be represented as,

$$\sigma_{mix}(\{\boldsymbol{X}_1, \boldsymbol{X}_2\}, \boldsymbol{M}) = \mathbb{1}(\boldsymbol{M} = 1)\boldsymbol{X}_1 + \mathbb{1}(\boldsymbol{M} = 2)\boldsymbol{X}_2, \tag{A1}$$

while the MAE input can be represented as,

$$\sigma_{MAE}(\boldsymbol{X}_1, \boldsymbol{M}) = \{\boldsymbol{X}_{1,l} | M_l = 1\} = \mathbb{1}(\boldsymbol{M} = 1)\boldsymbol{X}_1 + \mathbb{1}(\boldsymbol{M} = 2)\overrightarrow{\mathbf{0}}. \tag{A2}$$

Therefore, given $\boldsymbol{X}_1$ as the reconstruction target, we can represent the mutual information (MI) between the mixed input and the reconstruction target $\boldsymbol{X}_1$ as,

$$\begin{aligned}
I(\sigma_{mix}(\{\boldsymbol{X}_1, \boldsymbol{X}_2\}, \boldsymbol{M}); \boldsymbol{X}_1) &= I(\mathbb{1}(\boldsymbol{M} = 1)\boldsymbol{X}_1 + \mathbb{1}(\boldsymbol{M} = 2)\boldsymbol{X}_2; \boldsymbol{X}_1) \\
&= H(\boldsymbol{X}_1) - H(\boldsymbol{X}_1 | \mathbb{1}(\boldsymbol{M} = 1)\boldsymbol{X}_1 + \mathbb{1}(\boldsymbol{M} = 2)\boldsymbol{X}_2) \\
&= H(\boldsymbol{X}_1) - H(\boldsymbol{X}_1 | \mathbb{1}(\boldsymbol{M} = 1)\boldsymbol{X}_1 + \mathbb{1}(\boldsymbol{M} = 2)\overrightarrow{\mathbf{0}} + \mathbb{1}(\boldsymbol{M} = 1)\overrightarrow{\mathbf{0}} + \mathbb{1}(\boldsymbol{M} = 2)\boldsymbol{X}_2) \\
&= H(\boldsymbol{X}_1) - H(\boldsymbol{X}_1 | \mathbb{1}(\boldsymbol{M} = 1)\boldsymbol{X}_1 + \mathbb{1}(\boldsymbol{M} = 2)\overrightarrow{\mathbf{0}}, \mathbb{1}(\boldsymbol{M} = 1)\overrightarrow{\mathbf{0}} + \mathbb{1}(\boldsymbol{M} = 2)\boldsymbol{X}_2) \\
&\geq H(\boldsymbol{X}_1) - H(\boldsymbol{X}_1 | \mathbb{1}(\boldsymbol{M} = 1)\boldsymbol{X}_1 + \mathbb{1}(\boldsymbol{M} = 2)\overrightarrow{\mathbf{0}}) \\
&= I(\sigma_{MAE}(\boldsymbol{X}_1, \boldsymbol{M}); \boldsymbol{X}_1),
\end{aligned} \tag{A3}$$

where $H(\cdot)$ is the entropy. The conclusion above also holds when $\boldsymbol{X}_2$ is considered as the reconstruction target symmetrically.

Note that although independent in the data space, $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are not independent in the feature space, because the **global self-attention** would introduce inevitable information leakage from $\boldsymbol{X}_2$ to $\boldsymbol{X}_1$, which would be enhanced for "relevant" $\boldsymbol{X}_2$ (*e.g.*, the green cucumber in Fig. 3), while restrained for "irrelevant" ones (*e.g.*, the blue sky) by assigning different attention weights. As shown in Fig. 6, the $\mathrm{TopK}(\cdot)$ sampling accuracy converges to around 80%, suggesting that the information leakage does exist in practice. Therefore, the independence is affected, and the equality in Eq. (A3) does not hold.

For $r \in (0, 0.5)$, there are more than two images within a single group $\{\boldsymbol{X}_1, \boldsymbol{X}_2, ..., \boldsymbol{X}_{1/r}\}$. Considering $\boldsymbol{X}_1$ as the reconstruction target, we can first mix all images except $\boldsymbol{X}_1$ to generate a pseudo $\hat{\boldsymbol{X}}_2$ as,

$$\hat{\boldsymbol{X}}_2 = \sigma_{mix}(\{\overrightarrow{\mathbf{0}}, \boldsymbol{X}_2, \boldsymbol{X}_3, ..., \boldsymbol{X}_{1/r}\}, \boldsymbol{M}) = \mathbb{1}(\boldsymbol{M} = 1)\overrightarrow{\mathbf{0}} + \sum_{i=2}^{1/r} \mathbb{1}(\boldsymbol{M} = i)\boldsymbol{X}_i, \tag{A4}$$

which is then mixed with $\boldsymbol{X}_1$ following Eq. (A1). Therefore, the conclusion in Eq. (A3) can still be satisfied.

As discussed above, the usage of the **global self-attention** in ViT [8] is another indispensable factor to achieve the MI increasement in Eq. (A3). Therefore, in this paper, we propose the homologous attention to replace the global self-attention together with the homologous contrastive loss as verification for our *MixedAE*.

config	value
optimizer	AdamW [27]
base learning rate	$7.5e^{-5}$
weight decay	0.05
optimizer momentum	$\beta_1, \beta_2 = 0.9, 0.95$
batch size	4096 (B), 2048 (L)
learning rate schedule	cosine decay [26]
warmup epochs	40
augmentation	RandomResizedCrop
reconstruction target	normalized pixels [13]

Table A1. **Pre-training settings.**

config	value
optimizer	AdamW
base learning rate	$5e^{-4}$ (B), $1e^{-3}$ (L)
weight decay	0.05
optimizer momentum	$\beta_1, \beta_2 = 0.9, 0.999$
batch size	1024 (B), 512 (L)
learning rate schedule	cosine decay
warmup epochs	5
training epochs	100 (B), 50 (L)
augmentation	RandAug (9, 0.5)
label smoothing [31]	0.1
mixup [37]	0.8
cutmix [36]	1.0
drop path [16]	0.1

Table A2. **Fully fine-tuning settings.**

B. More Implementation Details

Pre-training. The default settings are provided in Tab. A1. We use xavier_uniform [11] to initialize all Transformer layers, following the official code of ViT [8]. Normalized pixels [13] are utilized as the reconstruction target, and the mask patch size is set to be 32×32, following [22, 35]. In practice, we utilize the sine-cosine encodings [32] for segment embeddings, which is added to the input of each Transformer layer following [25]. If no otherwise specified, the compose mixing mode is by default adopted to generate a single mixed sample for each image group.

ImageNet classification. The default settings are provided in Tab. A2. We mainly adopt the fully fine-tuning transfer setting to fine-tune the parameters of the backbone and the classification head simultaneously. We utilize the layer-wise learning rate decay strategy [6] following [1]. In practice, we sweep the decay ratio in $\{0.65, 0.7, 0.75\}$ following [22, 39].

ADE20K semantic segmentation. We use UperNet [34] as the segmentor following BEiT [1]. The input resolution is 512×512, and the batch size is set to be 16. The learning rate is set to be $3e^{-4}$ with the layer-wise learning rate decay ratio as 0.65 for ViT-Base and 0.75 for ViT-Large. We conduct fine-tuning for 160K iterations, and evaluate the performance without the multi-scale augmentation.

COCO object detection and instance segmentation. We utilize the Cascade Mask R-CNN [3, 14] following iBOT [39]. Multi-scale training is adopted with the shorted side randomly resized between 480 and 800 while the longer side no larger than 1333. The batch size is 16, the initial learning rate is $1e^{-4}$, and the layer-wise learning rate decay ratio is set to be 0.75. We adopt the standard $1\times$ schedule to train for 12 epochs, and decrease the learning rate by $10\times$ at epoch 9 and 11.

Downstream classification. We mainly follow the setups in [23, 24] to evaluate the transfer performance on 11 downstream classification datasets, including both the fine-grained datasets (*e.g.*, Aircraft [28], Cars [17], Flowers [29], Food [2], Pets [30] and SUN397 [33]), and the coarse-grained ones (*e.g.*, Caltech101 [19], CIFAR10 [18], CIFAR100 [18], DTD [5] and VOC2007 [10]). Specifically, we adopt a linear classification head upon the pre-trained ViT backbone and fully fine-tune the whole model simultaneously for 5000 iterations. The SGD optimizer and the cosine learning rate schedule are adopted. We grid search the optimal learning rate among $\{1e^{-3}, 3e^{-3}, 1e^{-2}, 3e^{-2}\}$, and set the weight decay to be 0.

C. More Experiments

Scaling property of *MixedAE*. We further build *MixedAE* with the standard ViT-Large [8] as the backbone architecture, and pre-train for 1600 epochs on ImageNet-1K [7] following the same optimization recipe with ViT-Base as in Appendix B. As demonstrated in Tab. A3, *MixedAE* still outperforms MAE consistently with ViT-Large, especially on downstream dense perception tasks [12, 20, 21, 38] thanks to the object-aware pre-training, revealing the scalability of our proposed *MixedAE*.

Method	Pre-train Epochs	Pre-train GPU-days	ImageNet Top-1 Acc.	ADE20K mIoU	COCO AP^{bb}	COCO AP^{bb}_{50}	COCO AP^{bb}_{75}	COCO AP^{mk}	COCO AP^{mk}_{50}	COCO AP^{mk}_{75}
iBOT [39]	1000*	285.0	85.0	52.2	49.9	69.5	54.1	42.9	66.5	45.9
MAE [13]	1600	151.1	85.9	53.6	54.0	72.6	59.0	46.3	69.6	50.4
MixedAE	1500	159.7	86.0	53.8	54.5	73.4	59.3	46.9	70.4	50.9
MixedAE	1600	170.4	**86.2**$^{+0.3}$	**54.0**$^{+0.4}$	**54.6**$^{+0.6}$	**73.5**$^{+0.9}$	**59.4**$^{+0.4}$	**47.1**$^{+0.8}$	**70.7**$^{+1.1}$	**51.0**$^{+0.6}$

Table A3. **Transfer performance comparison with ViT-Large [8].** 1) Scalability: our *MixedAE* outperforms MAE consistently even with ViT-Large. 2) Efficiency: *MixedAE* achieves a better trade-off between the computational overhead and the transfer performance regardless of the architecture size. *: effective epochs following iBOT [39].

	Masked SA	acc	mIoU
MAE		**82.7**	**46.1**
Naïve		82.4	45.0
Mixing	✓	82.6	45.9

(a) **Main cause of performance degradation** is indeed the information leakage brought by naïve mixing without homologous recognition.

$\mathcal{L}_{recon}$	$\mathcal{L}_{HomoCon}$	acc	mIoU
✓		82.4	45.0
	✓	7.8	8.3
✓	✓	**82.7**	**46.4**

(b) **Functionality of the $\mathcal{L}_{HomoCon}$.** When adopting $\mathcal{L}_{HomoCon}$ alone, *MixedAE* cannot even achieve reasonable transfer performance.

Mixing	SE	acc	mIoU
✓		82.2	44.9
	✓	81.1	42.9
✓	✓	**82.7**	**46.4**

(c) **Necessity of the mixing and segment embeddings.** The best transfer performance is achieved when both are adopted.

Table A4. **More *MixedAE* ablation experiments** with ViT-B/16. Default settings are marked in gray.

Moreover, we further report the performance of the 1500-epoch pre-trained *MixedAE* with ViT-Large to maintain similar computational overhead with MAE in Tab. A3. *MixedAE* still obtains consistent improvements over MAE, while outperforming iBOT with a 1.8× acceleration, suggesting that *MixedAE* can achieve a better trade-off between the computational overhead and the transfer performance regardless of the architecture size.

Ablation settings in Tab. 3. We conduct the ablation of new components based on previous results. Starting from the naïve baseline in Sec. 3.1, we first ablate the mixing ratio r in Tab. 3(a). Accordingly, we fix $r = 0.25$ to explore the position and positives of the homologous contrastive in Tab. 3(b)(c). Similarly, we ablate homologous attention in Tab. 3(d)(e) based on results of Tab. 3(c), and finally summarize all components in Tab. 3(f).

Main cause of performance degradation of naïve mixing. To verify that the information leakage brought by the mutual information increasement, as proved in Appendix A, is indeed the main cause of performance degradation of the naïve mixing baseline introduced in Sec. 3.1 instead of the optimization difficulty, we further build a mixing baseline by: 1) applying the masked self-attention to perform self-attention only within homologous patches to prevent information leakage without homologous recognition (*i.e.*, neither homologous $\mathrm{TopK}(\cdot)$ attention nor homologous contrastive loss); 2) adopting exactly the same optimization recipe with MAE. As demonstrated in Tab. A4a, the model achieves 82.6% accuracy and 45.9 mIoU, comparably with MAE, suggesting that the information leakage is definitely the culprit here.

Functionality of the $\mathcal{L}_{HomoCon}$. To verify that the homologous contrastive loss performs more like a *regularization term* instead of an *individual self-supervision* as in [9, 39], we further pre-train a *MixedAE* with $\mathcal{L}_{HomoCon}$ only following the settings in Sec. 4.4. As demonstrated in Tab. A4b, *MixedAE* performs well when the reconstruction loss $\mathcal{L}_{recon}$ is utilized only or together with the homologous contrastive loss $\mathcal{L}_{HomoCon}$. However, when adopting $\mathcal{L}_{HomoCon}$ only, *MixedAE* cannot even achieve reasonable transfer performance, suggesting that $\mathcal{L}_{HomoCon}$ cannot work alone without $\mathcal{L}_{recon}$.

Necessity of mixing. To verify to necessity of adopting mixing augmentation in our *MixedAE*, we extend MAE with the homologous contrastive $\mathcal{L}_{HomoCon}$ by applying Eq. (7) to patches across different images in groups of 4 for a fair comparison with *MixedAE*, which achieves 81.1% accuracy and 42.9 mIoU as demonstrated in Tab. A4c (2nd & 3rd rows), significantly worse than our default *MixedAE*, revealing the necessity of using mixing augmentation.

Necessity of segment embeddings. As shown in Tab. A4c (1st & 3rd rows), we build a *MixedAE without segment embeddings* and achieve 82.2% accuracy and 44.9 mIoU, significantly worse than our default *MixedAE*, suggesting the importance of adopting segment embeddings to provide necessary information for homologous recognition.

D. More Analysis

Exploration for other augmentations. As discussed in Sec. 3.1, based on mixing, we observe that common augmentation strategies will increase mutual information (MI) between the model input and the reconstruction target, suggesting that data augmentations like random augmentation and color jittering might *not be suitable* or *require specific designs* for MIM, which will be explored in the future work.

Limitations. Although demonstrating significant performance, we notice that the $\mathrm{TopK}(\cdot)$ sampling accuracy in homologous attention still cannot achieve 100% as shown in Fig. 6, for which there exist several potential reasons accounting: 1) The background patches might be included during random cropping inevitably, which are difficult for attention-based methods to recognize. 2) There is still further improvement space for *MixedAE*. For example, more strict verification than homologous contrastive loss is an appealing future work direction.

E. More Visualizations

We provide more visualizations of the attention mapes learnt by MAE and *MixedAE* on images from ImageNet-1K [7], ADE20K [38] and Microsoft COCO [21] datasets in Fig. A1. As demonstrated, our *MixedAE* can generate more reasonable and discriminative attention maps, revealing the effectiveness of *MixedAE* to conduct object-aware pre-training without any specifically designed modules (*e.g.*, K-means [4] and object discovery network [15]).

References

[1] Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021. 2

[2] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *ECCV*, 2014. 2

[3] Zhaowei Cai and Nuno Vasconcelos. Cascade R-CNN: delving into high quality object detection. In *CVPR*, 2018. 2

[4] Kai Chen, Lanqing Hong, Hang Xu, Zhenguo Li, and Dit-Yan Yeung. Multisiam: Self-supervised multi-instance siamese representation learning for autonomous driving. In *ICCV*, 2021. 4

[5] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *CVPR*, 2014. 2

[6] Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. Electra: Pre-training text encoders as discriminators rather than generators. In *ICLR*, 2020. 2

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 2, 4

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1, 2, 3

[9] Alaaeldin El-Nouby, Gautier Izacard, Hugo Touvron, Ivan Laptev, Hervé Jegou, and Edouard Grave. Are large-scale datasets necessary for self-supervised pre-training? *arXiv preprint arXiv:2112.10740*, 2021. 3

[10] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. In *ICCV*, 2010. 2

[11] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *AISTATS*, 2010. 2

[12] Jianhua Han, Xiwen Liang, Hang Xu, Kai Chen, Lanqing Hong, Chaoqiang Ye, Wei Zhang, Zhenguo Li, Xiaodan Liang, and Chunjing Xu. Soda10m: Towards large-scale object detection benchmark for autonomous driving. *arXiv preprint arXiv:2106.11118*, 2021. 2

[13] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. *arXiv preprint arXiv:2111.06377*, 2021. 2, 3

[14] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017. 2

[15] Olivier J Hénaff, Skanda Koppula, Evan Shelhamer, Daniel Zoran, Andrew Jaegle, Andrew Zisserman, João Carreira, and Relja Arandjelović. Object discovery and representation networks. *arXiv preprint arXiv:2203.08777*, 2022. 4

[16] Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q Weinberger. Deep networks with stochastic depth. In *ECCV*, 2016. 2

[17] Jonathan Krause, Jia Deng, Michael Stark, and Li Fei-Fei. Collecting a large-scale dataset of fine-grained cars. In *Workshop on Fine-Grained Visual Categorization*, 2013. 2

[18] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. *Master's thesis, University of Tront*, 2009. 2

[19] Fei-Fei Li, Fergus Rob, and Perona Pietro. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *CVPRW*, 2004. 2

[20] Kaican Li, Kai Chen, Haoyu Wang, Lanqing Hong, Chaoqiang Ye, Jianhua Han, Yukuai Chen, Wei Zhang, Chunjing Xu, Dit-Yan Yeung, et al. Coda: A real-world road corner case dataset for object detection in autonomous driving. *arXiv preprint arXiv:2203.07724*, 2022. 2

[21] 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 *ECCV*, 2014. 2, 4

[22] Jihao Liu, Xin Huang, Yu Liu, and Hongsheng Li. Mixmim: Mixed and masked image modeling for efficient visual representation learning. *arXiv preprint arXiv:2205.13137*, 2022. 2

[23] Zhili Liu, Kai Chen, Jianhua Han, Lanqing Hong, Hang Xu, Zhenguo Li, and James T. Kwok. Task-customized masked autoencoder via mixture of cluster-conditional experts. In *ICLR*, 2023. 2

[24] Zhili Liu, Jianhua Han, Kai Chen, Lanqing Hong, Hang Xu, Chunjing Xu, and Zhenguo Li. Task-customized self-supervised pre-training with scalable dynamic routing. In *AAAI*, 2022. 2

[25] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021. 2

[26] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. In *ICLR*, 2016. 2

[27] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 2

[28] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 2

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

[30] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *CVPR*, 2012. 2

[31] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, 2016. 2

[32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 2

[33] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*, 2010. 2

[34] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *ECCV*, 2018. 2

[35] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. *arXiv preprint arXiv:2111.09886*, 2021. 2

[36] Sangdoo Yun, Dongyoon Han, Sanghyuk Chun, Seong Joon Oh, Junsuk Choe, and Youngjoon Yoo. CutMix: Regularization strategy to train strong classifiers with localizable features. In *ICCV*, 2019. 2

[37] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. MixUp: Beyond empirical risk minimization. In *ICLR*, 2018. 2

[38] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. In *IJCV*, 2019. 2, 4

[39] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. *arXiv preprint arXiv:2111.07832*, 2021. 2, 3

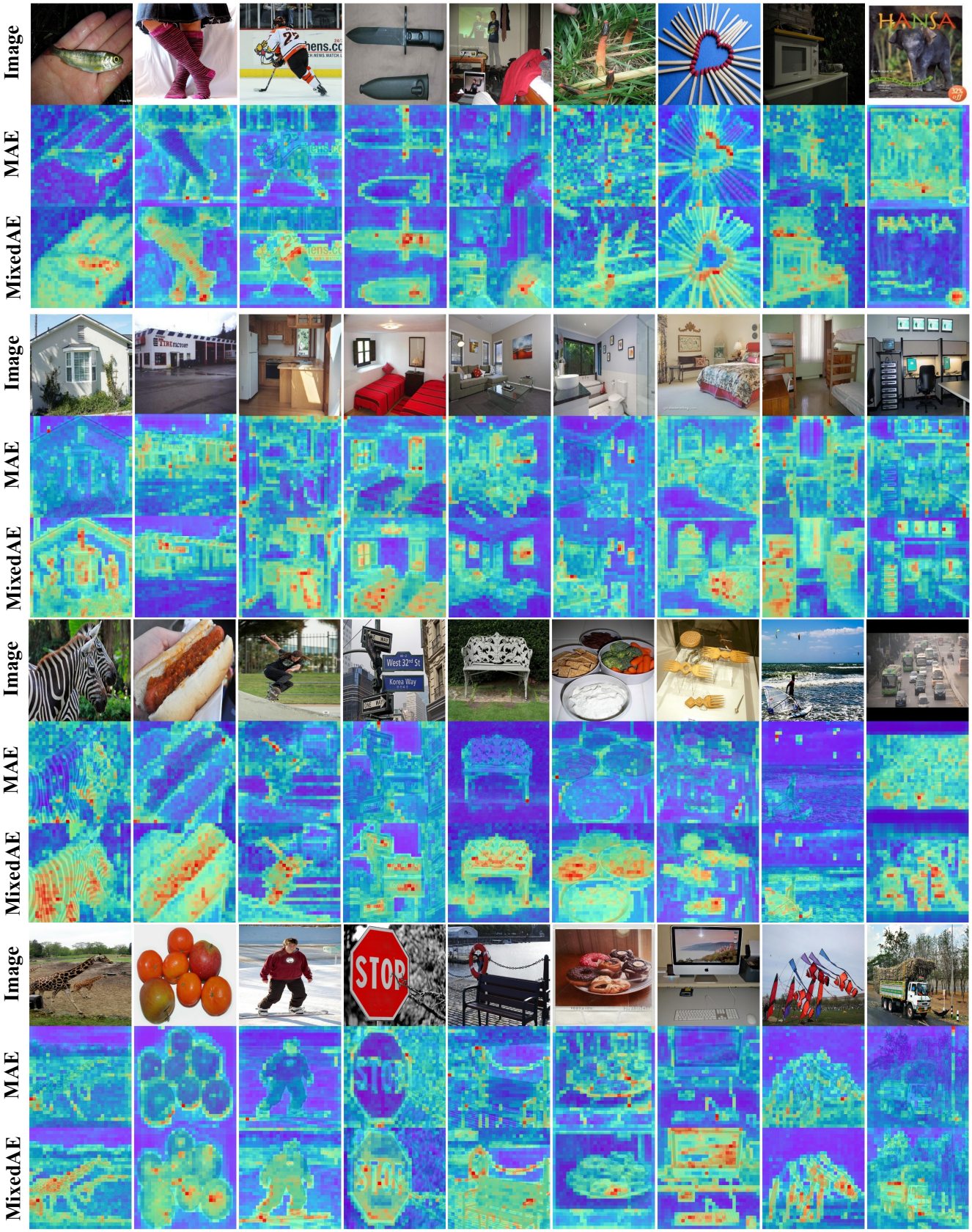

Figure A1. **More visualizations of attention maps** from ImageNet-1K (1-3 rows), ADE20K (4-6 rows) and COCO (7-12 rows).