IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 19 September 2023, accepted 2 October 2023, date of publication 13 October 2023, date of current version 24 October 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3324545

RESEARCH ARTICLE

Domain-Adaptive Vision Transformers for Generalizing Across Visual Domains

YUNSUNG CHO[1], JUNGMIN YUN[2], JUNEHYOUNG KWON[1], AND YOUNGBIN KIM[1], (Member, IEEE)

[1]Graduate School of Advanced Imaging Science, Multimedia and Film, Chung-Ang University, Seoul 06974, South Korea
[2]Department of Artificial Intelligence, Chung-Ang University, Seoul 06974, South Korea

Corresponding author: Youngbin Kim (ybkim85@cau.ac.kr)

This work was supported in part by the Chung-Ang University Graduate Research Scholarship in 2022, in part by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education under Grant NRF-2022R1C1C1008534, and in part by the Institute for Information and Communications Technology Planning and Evaluation (IITP) through the Korea Government (MSIT) (Artificial Intelligence Graduate School Program, Chung-Ang University) under Grant 2021-0-01341.

ABSTRACT Deep-learning models often struggle to generalize well to unseen domains because of the distribution shift between the training and real-world data. Domain generalization aims to train models that can acquire general features from data across different domains, thereby improving the performance on unseen domains. Inspired by the glance-and-gaze approach, which mimics the way humans perceive the real world, we introduce the domain-adaptive vision transformer (DA-ViT) model, which adopts a human cognitive perspective for domain generalization. We merge glance and gaze blocks to initially capture general information from each block and subsequently acquire more detailed and focused information. Unlike previous methods that predominantly employ convolutional neural networks, we adapted the ViT model to learn features that are robust across different visual domains. DA-ViT is pretrained on the ImageNet 1K dataset and designed to adaptively learn features that are generalizable across various visual domains. We evaluated our adapted model for domain generalization and demonstrated that it outperforms the ResNet50 model based on non-ensemble algorithms by 0.7%p on the VLCS benchmark dataset. Our proposed model introduces a new approach for domain generalization that leverages the capabilities of vision transformers to adapt effectively to diverse visual domains.

INDEX TERMS Domain generalization, ViT, masked ViT, cross-attention-based ViT, glance and gaze, human cognitive approach.

I. INTRODUCTION

Deep learning models continue to advance rapidly and are being applied to various real-world tasks. However, in practical usage, many of these models fail to meet research expectations because of disparities between training and real-world data. Despite the development of advanced models, such as convolutional neural network (CNN) architectures and vision transformer (ViT)-based models, real-world applications are plagued with challenges. To address domain generalization, we propose a deep-learning model, **D**omain-**A**daptive **Vi**sion **T**ransformers(DA-ViT), inspired by human cognitive processes.

Our approach is designed to address the limitations associated with existing domain generalization studies that either utilize only CNN-based models or focus solely on learning strategies. The proposed DA-ViT model introduces a cognitive perspective for domain generalization. This approach draws inspiration from the rich history of deep-learning studies that have successfully mimicked human biological and cognitive processes. By incorporating these cognitive insights, we aimed to enhance the performance and robustness of deep-learning models in real-world applications.

The associate editor coordinating the review of this manuscript and approving it for publication was Kathiravan Srinivasan.

A. DOMAIN GENERALIZATION METHODS AND THEIR LIMITATIONS

Recent advances in deep-learning models have demonstrated remarkable performance across diverse domains, such as image object classification [1], [2], segmentation [3], [4], and translation [5]. However, when deep-learning models are deployed in real-world applications, they often suffer from performance degradation owing to the difference in distribution between the training and real-world data. To address this issue, several domain generalization studies have been conducted that feature various strategies, such as data manipulation [6], [7], [8], data representation learning [9], [10] and learning strategies [11], [12], [13]. Data manipulation techniques involve modifying input data to improve generalization by extracting more generalized representations. These techniques include methods such as data augmentation, achieved through randomization [14] and transformations [15], as well as data generation [16], [17]. Data representation learning involves adversarial training [18] to learn domain-invariant representations, domain-unbiased representation learning [9], [19] to perform explicit feature alignment between domains, and domain sharing [20] or specific partial separation [21] for improved generalization. Another approach, learning strategy, aims to enhance generalization by utilizing ensemble [22] and meta-learning [11], [23] techniques to obtain general expressions.

However, these methods have limitations. For instance, data manipulation lacks a theoretical guarantee, and although adversarial training excels in domain adaptation, it lacks meaningful results for generalization. The design of an optimization strategy for a learning strategy is complex [24]. To address these issues and improve deep learning for real-world tasks, we propose DA-ViT, which is inspired by the human vision perspective and exhibits enhanced domain generalization.

B. TRANSFORMERS AND THEIR APPLICATIONS

Transformer-based models such as BERT [25] and the GPT series [26], [27], [28] have significantly enhanced performance across various domains, including translation [29], text classification [30], and question answering [31]. Notably, the extension of transformers to images using models such as ViT [32] represents a breakthrough, diverging from the prevalent use of CNNs, which acquire local image representations. ViT enables learning of global information, thereby revolutionizing image processing. Subsequently, various transformer-based models were developed. For instance, the detection transformer model (DETR) [33], which consists of an encoder and decoder structure, not only simplifies the detection pipeline but also removes many hand-designed components and consequently exhibits good performance in object detection. A pretrained image transformer (IPT) [34] improves image enhancement by solving basic image-processing problems such as denoising and de-raining. The texture transformer network for image super-resolution (TTSR) model [35] approaches the super-resolution problem by texture learning. For representation learning, masked autoencoders (MAEs) [36] have been applied to reconstruction tasks that mask the ViT patches. Multiscale ViT [37] has demonstrated remarkable performance on video tasks without pretraining. However, few studies have explored the application of ViT in representation learning for domain generalization. Therefore, the aim of our study is to enhance vision representation and extend the utility of the model to domain generalization by integrating masked ViT and cross-attention-based ViT.

C. ADVANCES IN DEEP LEARNING TECHNOLOGIES THAT MIMIC HUMANS

Initially, the field of deep learning encompassed emulating human neural activity and learning. Subsequently, researchers have actively explored deep-learning methods that replicate human functions from both biological and cognitive perspectives. The journey to mimic human deep neural networks began with the development of shallow networks such as ResNet and AlexNet, which then progressed into deeper architectures such as ResNet50 and ResNet101 [38]. Additionally, techniques inspired by human visual perception, which utilize both slow and fast methods for pose estimation through video flow analysis, have emerged, further advancing this field [39]. Furthermore, a growing body of research is aimed at approaching object detection from a cognitive perspective [40]. These studies introduced innovative approaches and made significant advancements in the research within their respective domains.

In this paper, we introduce DA-ViT with the ‘‘glance-and-gaze’’ cognitive approach, which is a longstanding concept used in human vision and object recognition. Our implementation of the glance-and-gaze method is intended to improve domain generalization. By combining deep-learning models, which can perform tasks similar to the way humans do, with cognitive methods such as glance and gaze, we anticipate promising results. Our approach mimics how humans extract limited information through a quick glance and improve upon it by focusing their attention. Through glancing, humans perceive only a limited amount of information, such as key features or rough information about an object. A quick initial glance improves the accuracy of information absorption while focusing on and gazing at an object after acquiring preliminary knowledge. To replicate this cognitive perspective, we utilized the masked ViT and cross-attention-based ViT to emulate the functions of glance and gaze, respectively. Within the glance block, we use the masked ViT to acquire approximate information. Within the gaze block, we utilize a cross-attention-based ViT to learn more comprehensive and refined information.

Our contributions are as follows:

- Unlike the conventional CNN-based model for domain generalization, our design utilizes a ViT-based model with a transformer architecture. This design choice is intended to preserve the ability of the model to retain the

strength of learning common expressions across various domains.

- To incorporate a human-like cognitive perspective into domain generalization, we introduce DA-ViT, based on a glance-and-gaze approach. This approach utilizes human-inspired learning techniques to innovate our model structure in domain generalization. We aim to address the current situation in which domain generalization primarily revolves around model ensembles and complex learning strategies, which can be difficult to implement in real-world scenarios. Through this approach, we strive to develop models that emulate human learning processes, ultimately enhancing the feasibility and practicality of domain generalization.
- We evaluated the performance of our proposed DA-ViT model, which integrates human cognitive insights, in different domains. Overall, DA-ViT exhibited a performance improvement of 0.7%p on the VLCS dataset.

II. RELATED WORK

A. DOMAIN GENERALIZATION

Domain generalization aims to learn common representations from the source domain to ensure that they are generalized well on unused datasets or out-of-distribution domain datasets during model learning. Domain generalization is typically categorized into three approaches. Data manipulation, initially proposed as a method for randomly generating learning environments that resemble the real world, aims to generate diverse training data [41]. These datasets help in learning general representations of domains that are not included in the training process. The creation of real-world-like data [42], which mitigates the reality gap through domain randomization [14], contributes to improving model generalization. In addition, techniques such as self-supervised contrast regularization [8] enhance the ability of the model to be generalized through self-supervised learning. Representation learning focuses on identifying invariant representations across diverse domains. It commonly employs techniques such as maximum mean discrepancy [43], [44] and Wasserstein distance [45] to align features across different domains. Learning strategies focus on utilizing general learning techniques to promote generalization through various approaches such as ensemble learning [22], [46], meta-learning [11], [12], [23], and gradient operations [13], [47].

However, these methods have practical limitations and are often considered inadequate [24]. They are customized for specific learning algorithms and model selection [48]. Consequently, this limitation restricts model design, reduces the scalability of domain generalization, and results in performance improvements that are applicable only to limited models. Presently, domain generalization relies primarily on CNN-based models, emphasizing the need to develop new models. In this paper, we introduce a new model for domain generalization designed to address these challenges and satisfy the aforementioned requirements.

B. ViT

The transformer model [49], initially employed in the field of natural-language processing (NLP), has shown remarkable performance in various applications such as language translation [29], question answering [31], and text generation [50]. Recently, there has been a significant research shift from CNN-based approaches to ViT [2], [4], [32], [51], [52]. Transformers are increasingly applied to image-related tasks. ViT-based studies have spurred advancements in computer vision, leading to their use in various tasks such as object detection [51], [52], image classification [2], [32], and semantic segmentation [4].

Furthermore, there has been a growing interest in studying masked ViT models, specifically for improving image attributes [36], [53], [54]. One such approach [36] involves randomly masking patches within input images and reconstructing the masked regions. In this process, the encoder operates on unmasked tokens, whereas the decoder utilizes masked tokens and latent representations to reconstruct the original image. The present study demonstrates the potential of expanding the scope of self-supervised learning in the field of computer vision by utilizing an encoder–decoder architecture. The self-supervisor transformer [53] employs a masked-patch approach to consider both high-level and local features. Specifically, this masking strategy enhances the comprehension of local contextual semantics without compromising the overall image structure. It addresses issues related to the insufficient extraction of local information and the loss of spatial information. Furthermore, studies using masked ViT models have revealed that an effective model design can enhance the learning of image characteristics, even when certain portions of an image are obscured or only specific parts are learned [36], [53].

Within the domain of ViT-based research, cross-attention is employed to learn features simultaneously within images [55], [56] and between images and text [57], [58]. In addition, the dual-branch ViT [55] model combines information across branches of varying scales or patch-embedding sizes to effectively capture multiscale image features. This approach combines two self-attention mechanisms, operating within and between patches [56], with the aim of effectively integrating local features and global information while reducing the computational overhead associated with capturing additional information. In studies examining the relationship between images and word features [57], stacked cross-attention was used to determine the similarity between images and sentences. Meanwhile, in research applying cross-modal attention to image patches and words [58], the focus was on combining global semantic consistency and alignment between local image regions and words. Drawing inspiration from these studies, we propose an approach for image feature learning using cross-attention

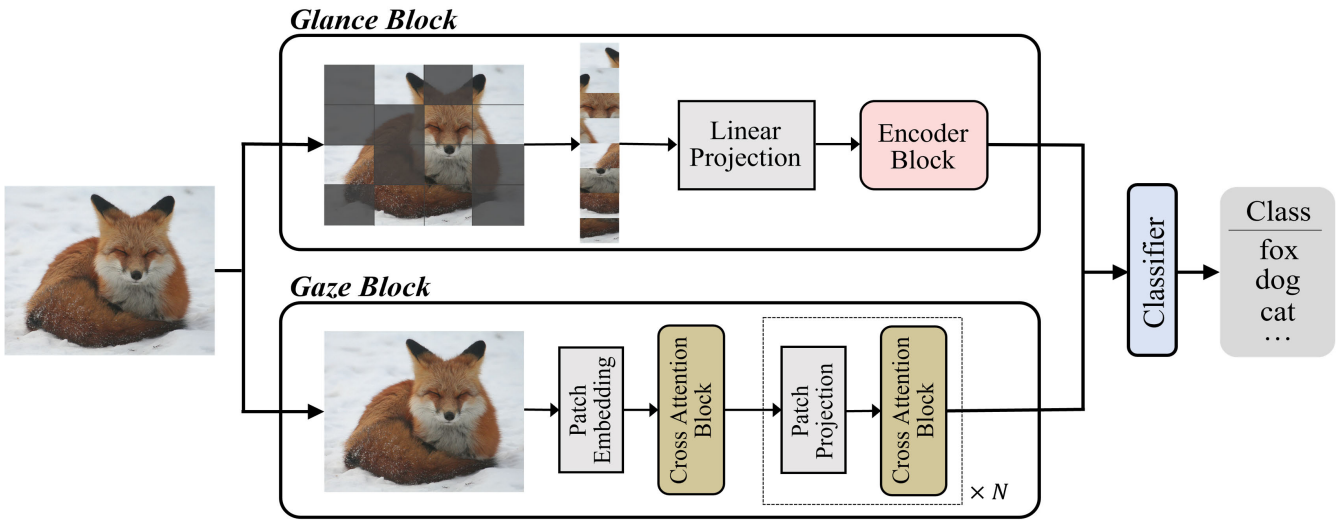

FIGURE 1. Overall architecture of the proposed DA-ViT model. The glance block consists of a masked ViT with specific patches being masked. The gaze block is configured using cross-attention.

blocks that combine masked ViT with pixel self-attention and self-attention between image patches.

C. APPROACHING DEEP LEARNING FROM A HUMAN PERSPECTIVE

In general, the development and advancement of deep-learning research has been driven by the emulation of human biological or cognitive processes [38], [59], [60], [61]. ResNet [38], [61] and Inception [60] are prominent examples of deep-learning architecture that draw inspiration from deep neural networks that resemble the structure of the human brain. In addition, SlowFast [39] presents a video recognition approach inspired by the function of retinal ganglion cells in the human visual system. Specifically, the model processes images through two streams, similar to how human photoreceptors receive visual inputs, and applies them to human pose estimation.

The glance-and-gaze mechanism, which embodies the human cognitive perspective, influences various forms across multiple fields. For example, in applications related to human–object interaction, the glancing transformer [40] swiftly determines whether feature map pixels correspond to interaction points. The model subsequently facilitates adaptive reasoning, enabling judgment-based determination of the nature of human–object interaction. Furthermore, the glance-and-gaze network (GaGNet) [62] expands the concept of glance and gaze, incorporating the cognitive aspect of human perception to improve speech processing and enhancement. In the field of speech processing, GaGNet adopts a dual approach: ''glance,'' which offers an initial estimation, and ''gaze,'' which compensates for the loss of spectral information. Furthermore, GAGNet [63] provides a solution to the structural limitations of the current CNN models in breath sound classification. This network utilizes global features through ''glance'' and localized features through ''gaze,'' leading to significant performance improvements using publicly accessible breath-sound data. The glance-and-gaze transformer (GG-Transformer) [64] is also inspired by the glancing and gazing behavior of human beings when recognizing objects in natural scenes. Unlike our study, which is aimed at domain generalization, GG-Transformer introduces an efficient transformer that reduces the computational and memory costs of self-attention. It combines a self-attention branch with a simple depth-wise convolution layer branch to model both long-range dependencies and local context.

The development of deep learning has been profoundly influenced by the emulation of human biological and cognitive perspectives. Furthermore, studies that began with the construction of models have evolved significantly. When dealing with tasks that exhibit human-like characteristics, models effectively mimic human-like capabilities and actively advance towards acquiring the abilities to learn and reason. In this context, the objective of this study is to develop a model designed for domain generalization, taking inspiration from ''glance and gaze.'' The ultimate goal is to overcome the challenges and limitations that arise when using models in real-world problem-solving situations.

III. METHOD

In the field of visual recognition, humans quickly perceive the presence of an object and its general characteristics with a quick glance. This study introduced DA-ViT, which is an approach that emulates human cognitive processes by integrating masked ViT and cross-attention-based ViT using a glance-and-gaze approach. The overall architecture of the proposed model is illustrated in Fig. 1. In the glance block of the proposed design, the masked ViT is used to infer the approximate information. In the gaze block, cross-attention

is employed to combine pixel attention and interpatch attention, thereby enhancing the richness of the image characteristics. The proposed DA-ViT model integrates both general initial information and specific intricate details to suit each characteristic.

A. GLANCE BLOCK

A pivotal focus of this study lies on developing the ''glance'' aspect from a cognitive perspective. In previous studies investigating human–object interactions [40] or enhanced speech recognition [62], humans perceived the presence or absence of objects and gained a basic understanding of their overall characteristics through glancing [40], [63]. However, this understanding is often incomplete and imprecise. When humans cast a glance, they may inaccurately perceive the existence of objects, resulting in incomplete gathering of information. Although valuable, gaze supplementation only partially compensates for this inadequacy, resulting in a limited amount of information being obtained from the initial glance. To address this limitation, we designed a glance block using masked ViT to better align with human cognitive characteristics. The overall structure of the glance block is illustrated in Fig. 2.

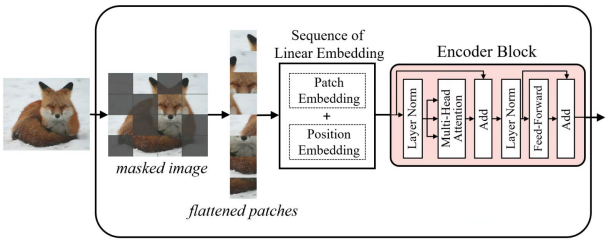

FIGURE 2. Structure of the glance block. Three-quarters of the image patches are masked.

The glance block closely resembles ViT but incorporates mask-based learning as an additional component. More precisely, we transform the original image $x \in \mathbb{R}^{H\times W\times C}$ into $x_{glance} \in \mathbb{R}^{N\times(P^2\times C)}$, where $(H \times W)$ is the size of the original image, C is the number of channels, and $(P \times P)$ is the size of the patch. In the glance block, the patch size is set to 7×7. Subsequently, we incorporate learnable embeddings to capture positional information. Unlike ViT, the proposed model does not include a $[CLS]$ token. Following the partitioning of the input image into patches, we determine the set of patches to be used while masking the remaining patches that will not be utilized. This process is similar to the exclusion of patches from consideration. To be precise, we utilize a random uniform distribution to randomly select a set of patches to be retained. This strategic random sampling significantly reduces redundancy. By employing a stack of encoders in a layered fashion, we obtain approximate information regarding the images to be captured. The following operations are performed within the glance block:

$$z_o = Patch.Emb(x_{glance}) + Pos.Emb \quad (1)$$

$$hatz_{temp1} = MSA(LN(\hat{z}_{n-1})) + \hat{z}_{n-1} \quad (2)$$

$$\hat{z}_{temp2} = MLP(LN(\hat{z}_{temp1})) + \hat{z}_n \quad (3)$$

B. GAZE BLOCK

As previously mentioned, gaze should exhibit the characteristics of focused concentration. In several studies [65], [66], the process of feature learning for images involved establishing relationships between individual pixels. In particular, we developed a gaze block inspired by the cross-attention mechanism found in ViT [56]. This gaze block leverages pixel self-attention within individual patches, divides the image into patch units, and applies cross-attention to establish connections between these patches. We employed a cross-attention-based transformer model that utilizes both pixel self-attention (PiSA) and interpatch self-attention (PaSA). This design enables the system to comprehensively learn pixel-wise information within an image and the relationships between various patches. The structure of the gaze block is shown in Fig. 3.

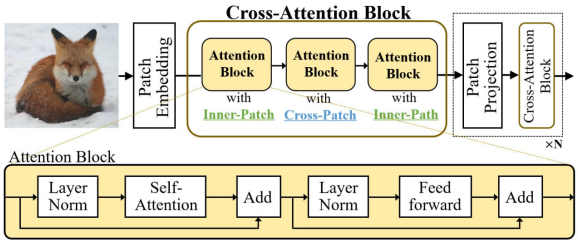

FIGURE 3. Structure of the gaze block. The gaze block consists of cross-attention stacking of the PiSA and PaSA, each with layer normalization [67].

Similar to that observed for word tokens in NLP [25], [28], using all pixels in an image or feature map as tokens results in a significant increase in computational requirements. In computer vision, understanding the relationships among pixels is crucial for object recognition. However, considering every pixel in all the images is computationally challenging. Therefore, we exclusively applied self-attention to the pixels within the divided patches. More specifically, when addressing the computational complexities posed by considering every pixel in all the images, we limited our self-attention to the pixels corresponding to the divided patch. This approach enables the system to capture pixel relationships within individual patches without the need for complete pixel computation across the entire image. In this study, we applied pixel self-attention within each patch using inner patch self-attention [56], as shown in Fig. 4.

However, limiting pixel self-attention solely to a divided patch confines our focus to the pixel correlations within that specific patch. In an image, it is crucial not only to understand the relationships between pixels but also to gain a comprehensive understanding of its content and meaning. Thus, drawing

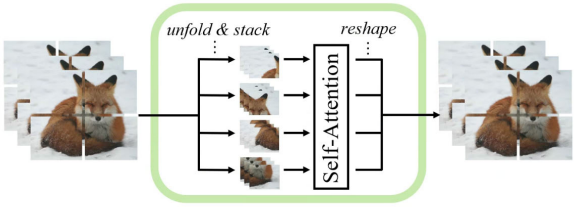

FIGURE 4. Pixel self-attention (PiSA) structure, which includes an attention mechanism with an inner patch. PiSA operates by unfolding and stacking the inputs.

inspiration from cross-patch self-attention [56], we utilize interpatch self-attention to simultaneously capture the overall image context and identify the relationships between patches, as shown in Fig. 5. In the proposed approach, the feature map of each channel is separated and partitioned into patches of size $H/N \times W/N$. Subsequently, we utilize self-attention to facilitate interpatch self-attention, enabling the capture of global information from the entire feature map. Here, N denotes the patch size and H and W denote the height and width of the feature map, respectively. As shown in Fig. 3, we combined pixel and inter-patch self-attention blocks to extract and integrate both the features of a pixel within the patch and those across the entire image and feature map. Specifically, pixel self-attention utilizes relative position encoding [68], [69], [70], whereas interpatch attention adopts absolute position encoding. The reason for this differentiation lies in the nature of interpatch attention, which operates on a complete single-channel feature map.

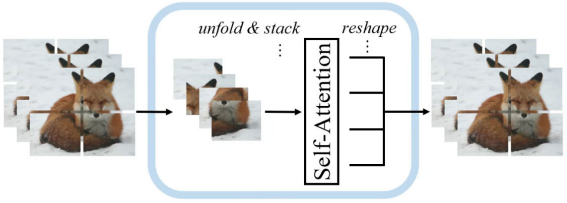

FIGURE 5. Structure of interpatch self-attention (PaSA), which includes an attention mechanism with a cross patch. Interpatch self-attention operates by unfolding and stacking the inputs.

The cross-attention block consists of two pixel self-attention (PiSA) blocks and an interpatch self-attention (PaSA) block. The gaze block is composed of multiple cross-attention blocks, and each stage of the network features a different number of layers and patch-embedding layers. The operations performed within the cross-attention block are as follows.

$$y_o = Patch.Emb(x_{gaze}) + Pos.Emb \tag{4}$$
$$\hat{y}_{temp1} = PiSA(LN(\hat{y}_{n-1})) + \hat{y}_{n-1} \tag{5}$$
$$\hat{y}_{temp2} = MLP(LN(\hat{y}_{temp1})) + \hat{y}_{temp1} \tag{6}$$
$$\hat{y}_{temp3} = PaSA(LN(\hat{y}_{temp2})) + \hat{y}_{temp2} \tag{7}$$
$$\hat{y}_{temp4} = MLP(LN(\hat{y}_{temp3})) + \hat{y}_{temp3} \tag{8}$$
$$\hat{y}_{temp5} = PiSA(LN(\hat{y}_{temp4})) + \hat{y}_{temp4} \tag{9}$$
$$\hat{y}_n = MLP(LN(\hat{y}_{temp5})) + \hat{y}_{temp5} \tag{10}$$

The gaze block component follows the pipeline shown in Fig. 6. In a related study involving cross-attention [56], a patch size of 7×7 was used and multi-head attention was not employed for interpatch self-attention. However, in our gaze block, the patch size is enlarged to 14×14. By increasing the patch size in the gaze block, distinct characteristics beyond those learned from the relationship between patches within the glance block component can be captured. This larger patch size also allows the implementation of more effective pixel self-attention, which leads to improved performance. Furthermore, we incorporated multi-head attention into the gaze block, which enables the extraction of more comprehensive features during training.

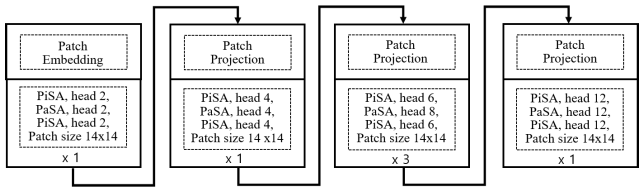

FIGURE 6. Gaze block pipeline.

IV. EXPERIMENTS

A. EXPERIMENTAL SETTINGS

A single V100 GPU was used to train the model. To train the glance block, we used a ViT encoder with three layers. For gaze block training, we employed four layers, each comprising cross-attention blocks. These cross-attention blocks combine pixel and interpatch self-attention. The four layers consist of cross-attention blocks. During the training on the ImageNet 1K dataset [71], we utilized the AdamW optimizer [72] with an initial learning rate of 5e-5. Subsequently, the learning rate was linearly adjusted using a warm-up scheduler [73].

In this study, we followed the training and validation protocols outlined in DomainBed [48] for a fair comparison. To elaborate, we designated data from a specific domain as the target domain; the remaining data served as the source domain. Furthermore, we allocated 20% of the source domain for validation. By utilizing different random seeds, we split the data into training and validation sets and repeated the training and validation procedures three times. The resulting distributions were then averaged across all domains.

The performance of each test was recorded for each dataset. In this experiment, we used an input image of size 224×224. We used the pretrained DA-ViT on ImageNet 1K as the backbone. For optimization, we employed stochastic gradient descent (SGD) with a learning rate of 3e-3, momentum of 0.9, weight decay of 1e-4, and batch size of 64.

B. DATASET

We utilized training and test data to evaluate the effectiveness of DA-ViT. Specifically, our model was trained on ImageNet 1K, a dataset comprising 14,197,122 annotated images

categorized according to the WordNet hierarchy. Since 2010, ImageNet has played a crucial role in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC), which serves as a benchmark for image classification and object detection. Prior to the evaluation of domain generalization, we pre-trained our proposed model on ImageNet 1K. Subsequently, we evaluated the domain generalization performance on the VLCS benchmark dataset [74], which encompasses four domains: VOC2007, LabelMe, Caltech101, and SUN09, containing a total of 10,729 images. Each domain consists of five categories: birds, cars, chairs, dogs, and people.

Furthermore, to evaluate the performance of our proposed model, as well as each glance-and-gaze methodology, we expanded our evaluation to include the PACS [75] and Office-Home datasets [76]. The PACS dataset, which is commonly used for domain adaptation and generalization tasks, consists of four distinct domains: photos, art, cartoons, and sketches, with a total of 9,991 images. The Office-Home dataset, designed for domain adaptation and generalization assessment, encompasses four domains: art, clip art, product, and real. The art domain includes images such as sketches, paintings, and decorations. Clip art consists of clipped art images. Product contains images of objects without backgrounds. Real images encompass images captured using standard cameras. A total of 15,500 images were obtained.

FIGURE 7. (1) Dog label in the VLCS dataset, (2) dog label in the PACS dataset, and (3) alarm clock label in the Office-Home dataset.

C. EXPERIMENTAL RESULTS

By leveraging the proposed DA-ViT model, we assessed its domain generalization performance using the VLCS benchmark dataset. Because the VLCS dataset consists of real-world images, it aligns closely with our objective of replicating human cognitive perspectives in real-world contexts. We compared the performance of DA-ViT with that of the ResNet50 model using a non-ensemble algorithm; the results are presented in Table 1.

TABLE 1. Comparison of DA-ViT with ResNet50 models using non-ensemble algorithms.

	MLDG [77]	MMD [78]	ERM [48]	Fish [79]	*(ours)* ERM
Accuracy	77.1%	76.7%	77.4%	77.8%	**78.1%**

Using the empirical risk-minimization (ERM) method, DA-ViT achieves a performance of 78.1% on the VLCS benchmark dataset. This demonstrates an improvement ranging from 0.3%p to 1.4%p compared with the performance of the ResNet50 model based on different non-ensemble algorithms. When comparing the results obtained using the same ERM algorithm, a 0.7%p improvement is observed. The proposed DA-ViT model outperforms ResNet50 on the VLCS dataset, which comprises real photos. ResNet50 and DA-ViT have 25.6 M and 39.6 M parameters, respectively.

TABLE 2. Experimental comparison of the performance of DA-ViT, Glance, and Gaze.

Model	PACS	VLCS	Office-Home	Avg.
only Glance	34.5%	54.6%	15.1%	34.7%
only Gaze	72.0%	75.2%	51.5%	66.2%
DA-ViT	**78.1%**	**78.1%**	**58.3%**	**71.5%**

We separated our proposed model, DA-ViT, which combines glance and gaze blocks, into Glance and Gaze models. Subsequently, we evaluated the domain generalization performance of the three models, DA-ViT, Glance, and Gaze, using the benchmark dataset. The results of the evaluations are presented in Table 2. DA-ViT consistently exhibits superior performance compared to the individual Glance and Gaze methodologies across the PACS, VLCS, and Office-Home datasets. These results underscore the effectiveness of the proposed model, which captures human cognitive perspectives by integrating glance and gaze.

TABLE 3. Experiments specific to different domains within the VLCS, PACS, and Office-Home datasets.

		VOC2007	LabelMe	Caltech101	SUN09	Avg.
VLCS	*only* Glance	47.8%	53.88%	73.94%	50.88%	56.62%
	only Gaze	69.27%	64.47%	93.2%	**73.99%**	75.23%
	DA-ViT	**75.82%**	**68%**	**95.14%**	73.57%	**78.13%**
		Photo	Art	Cartoon	Sketch	Avg.
PACS	*only* Glance	49.1%	29.6%	35%	24.5%	34.5%
	only Gaze	90.1%	70.2%	68.8%	59%	72.1%
	DA-ViT	**94.5%**	**78.5%**	**75.3%**	**64%**	**78.1%**
		Art	ClipArt	Product	Real	Avg.
Office-Home	*only* Glance	10.1%	11.2%	19.2%	19.9%	15.1%
	only Gaze	41.1%	39.2%	62.1%	63.5%	51.5%
	DA-ViT	**48.6%**	**45%**	**68.3%**	**71.5%**	**58.3%**

Table 3 lists the experimental results of selecting a target domain and utilizing the remaining domains as source domains for training. DA-ViT shows better performance than the models using only Glance and Gaze individually in all domain-specific experiments, except for SUN09 in VLCS. These results demonstrate the effectiveness of the proposed model. Performance is relatively low when using

only glance; however, combining glance and gaze provides better performance than using gaze alone.

However, the performance improvement is relatively small for image types that are significantly different from photographs, such as the sketch domain in the PACS dataset and the clip art domain in the Office-Home dataset. According to the findings presented in Table 3, our analysis highlights the exceptional performance of the proposed model, particularly in handling image categories such as photorealistic images within datasets that encompass a wide range of domains, including paintings, infographics, real photographs, and sketches.

V. CONCLUSION

In this study, we explored domain generalization, a research area that can expand practical deep-learning applications in the real world. Inspired by human cognitive processes, we introduced the DA-ViT model as a novel approach for domain generalization. Analogous to human perception, this approach enables our model to capture both coarse information through glances and more detailed information through a focused gaze. In contrast to prior approaches, we adopted a transformer-based architecture, which has demonstrated its effectiveness in image processing and resulted in significant performance improvements. To implement the glance-and-gaze mechanism, we utilized the combination of a glance block, with a masked ViT, and a gaze block, with a cross-attention-based ViT. Because our goal was to mimic a human-like perception of the real world, we evaluated the domain generalization capability of our proposed model on the VLCS dataset, which consists of real-world images. Our experimental results demonstrate that the synergistic utilization of glance and gaze leads to competitive accuracy, outperforming ResNet50 based on non-ensemble algorithms by 0.7%p.

Our proposed model demonstrates outstanding performance when processing realistic photographic images but exhibits limitations in domain generalization when confronted with non-real-world images such as sketches or cartoons. However, our model is designed to emulate human perception of the real world. We emphasize that our work fundamentally focuses on domain generalization for effective application of deep-learning models in real-world scenarios. Furthermore, the performance of DA-ViT is enhanced when both glance and gaze mechanisms are employed simultaneously. However, this also reveals a limitation in that the performance falls short when relying solely on the glance mechanism. Thus, in the future, we aim to include learning the masking strategy in the glance block rather than a simple random selection. Our study is significant because it represents a pioneering effort to apply a human cognitive perspective to domain generalization utilizing a glance-and-gaze approach. In the future, we intend to incorporate datasets from a broader range of domains and explore a wider array of baselines to enhance our research. This study may enable the development of a more elaborate cognitive model that imitates humans.

REFERENCES

[1] Z. Dai, H. Liu, Q. V. Le, and M. Tan, "CoAtNet: Marrying convolution and attention for all data sizes," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 3965–3977.

[2] M. Ding, B. Xiao, N. Codella, P. Luo, J. Wang, and L. Yuan, "DaViT: Dual attention vision transformers," 2022, *arXiv:2204.03645*.

[3] Z. Liu, H. Hu, Y. Lin, Z. Yao, Z. Xie, Y. Wei, J. Ning, Y. Cao, Z. Zhang, L. Dong, F. Wei, and B. Guo, "Swin transformer v2: Scaling up capacity and resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 11999–12009.

[4] J. Jain, A. Singh, N. Orlov, Z. Huang, J. Li, S. Walton, and H. Shi, "SeMask: Semantically masked transformers for semantic segmentation," 2021, *arXiv:2112.12782*.

[5] H. Xu, B. Van Durme, and K. Murray, "BERT, mBERT, or BiBERT? A study on contextualized embeddings for neural machine translation," 2021, *arXiv:2109.04588*.

[6] R. Khirodkar, D. Yoo, and K. Kitani, "Domain randomization for scene-specific car detection and pose estimation," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2019, pp. 1932–1940.

[7] A. Prakash, S. Boochoon, M. Brophy, D. Acuna, E. Cameracci, G. State, O. Shapira, and S. Birchfield, "Structured domain randomization: Bridging the reality gap by context-aware synthetic data," in *Proc. Int. Conf. Robot. Autom. (ICRA)*, May 2019, pp. 7249–7255.

[8] D. Kim, Y. Yoo, S. Park, J. Kim, and J. Lee, "SelfReg: Self-supervised contrastive regularization for domain generalization," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9599–9608.

[9] C. Shui, B. Wang, and C. Gagné, "On the benefits of representation regularization in invariance based domain generalization," *Mach. Learn.*, vol. 111, no. 3, pp. 895–915, Mar. 2022.

[10] C. Shui, Z. Li, J. Li, C. Gagné, C. X. Ling, and B. Wang, "Aggregating from multiple target-shifted sources," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 9638–9648.

[11] K. Chen, D. Zhuang, and J. M. Chang, "Discriminative adversarial domain generalization with meta-learning based cross-domain validation," *Neurocomputing*, vol. 467, pp. 418–426, Jan. 2022.

[12] Y. Zhao, Z. Zhong, F. Yang, Z. Luo, Y. Lin, S. Li, and N. Sebe, "Learning to generalize unseen domains via memory-based multi-source meta-learning for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 6273–6282.

[13] C. X. Tian, H. Li, X. Xie, Y. Liu, and S. Wang, "Neuron coverage-guided domain generalization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 1302–1311, Jan. 2023.

[14] J. Huang, D. Guan, A. Xiao, and S. Lu, "FSDR: Frequency space domain randomization for domain generalization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 6887–6898.

[15] N. H. Nazari and A. Kovashka, "Domain generalization using shape representation," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 666–670.

[16] P. Li, D. Li, W. Li, S. Gong, Y. Fu, and T. M. Hospedales, "A simple feature augmentation for domain generalization," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 8866–8875.

[17] N. Somavarapu, C.-Y. Ma, and Z. Kira, "Frustratingly simple domain generalization via image stylization," 2020, *arXiv:2006.11207*.

[18] H. Ye, C. Xie, T. Cai, R. Li, Z. Li, and L. Wang, "Towards a theoretical framework of out-of-distribution generalization," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 23519–23531.

[19] M. Planamente, C. Plizzari, E. Alberti, and B. Caputo, "Domain generalization through audio-visual relative norm alignment in first person action recognition," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2022, pp. 163–174.

[20] X. Jin, C. Lan, W. Zeng, and Z. Chen, "Style normalization and restitution for domain generalization and adaptation," *IEEE Trans. Multimedia*, vol. 24, pp. 3636–3651, 2022.

[21] S. Motiian, M. Piccirilli, D. A. Adjeroh, and G. Doretto, "Unified deep supervised domain adaptation and generalization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5716–5726.

[22] M. Segu, A. Tonioni, and F. Tombari, "Batch normalization embeddings for deep domain generalization," *Pattern Recognit.*, vol. 135, Mar. 2023, Art. no. 109115.

[23] Q. Chen, C. Shui, and M. Marchand, "Generalization bounds for meta-learning: An information-theoretic analysis," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 25878–25890.
[24] J. Wang, C. Lan, C. Liu, Y. Ouyang, T. Qin, W. Lu, Y. Chen, W. Zeng, and P. Yu, "Generalizing to unseen domains: A survey on domain generalization," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 8, pp. 8052–8072, May 2022.
[25] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.
[26] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," OpenAI, San Francisco, CA, USA, Tech. Rep., 2018.
[27] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI Blog*, vol. 1, no. 8, p. 9, 2019.
[28] T. B. Brown et al., "Language models are few-shot learners," in *Proc. Adv. Neur. Inf. Process. Sys.*, vol. 33, 2020, pp. 1877–1901.
[29] J. Zhu, Y. Xia, L. Wu, D. He, T. Qin, W. Zhou, H. Li, and T.-Y. Liu, "Incorporating BERT into neural machine translation," 2020, *arXiv:2002.06823*.
[30] S. Garg and G. Ramakrishnan, "BAE: BERT-based adversarial examples for text classification," 2020, *arXiv:2004.01970*.
[31] W. Yang, Y. Xie, A. Lin, X. Li, L. Tan, K. Xiong, M. Li, and J. Lin, "End-to-end open-domain question answering with BERTserini," 2019, *arXiv:1902.01718*.
[32] 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 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.
[33] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 213–229.
[34] H. Chen, Y. Wang, T. Guo, C. Xu, Y. Deng, Z. Liu, S. Ma, C. Xu, C. Xu, and W. Gao, "Pre-trained image processing transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 12294–12305.
[35] F. Yang, H. Yang, J. Fu, H. Lu, and B. Guo, "Learning texture transformer network for image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5790–5799.
[36] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 15979–15988.
[37] Q. Zhang, Y. Xu, J. Zhang, and D. Tao, "ViTAEv2: Vision transformer advanced by exploring inductive bias for image recognition and beyond," *Int. J. Comput. Vis.*, vol. 2023, pp. 1–22, Jan. 2023.
[38] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[39] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "SlowFast networks for video recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6201–6210.
[40] X. Zhong, X. Qu, C. Ding, and D. Tao, "Glance and gaze: Inferring action-aware points for one-stage human-object interaction detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 13229–13238.
[41] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. Abbeel, "Domain randomization for transferring deep neural networks from simulation to the real world," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Sep. 2017, pp. 23–30.
[42] X. B. Peng, M. Andrychowicz, W. Zaremba, and P. Abbeel, "Sim-to-real transfer of robotic control with dynamics randomization," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2018, pp. 3803–3810.
[43] J. Wang, Y. Chen, W. Feng, H. Yu, M. Huang, and Q. Yang, "Transfer learning with dynamic distribution adaptation," *ACM Trans. Intell. Syst. Technol.*, vol. 11, no. 1, pp. 1–25, Feb. 2020.
[44] Y. Zhu, F. Zhuang, J. Wang, G. Ke, J. Chen, J. Bian, H. Xiong, and Q. He, "Deep subdomain adaptation network for image classification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 4, pp. 1713–1722, Apr. 2021.
[45] S. Zhao, M. Gong, T. Liu, H. Fu, and D. Tao, "Domain generalization via entropy regularization," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 16096–16107.
[46] A. D'Innocente and B. Caputo, "Domain generalization with domain-specific aggregation modules," in *Proc. German Conf. Pattern Recognit.* Cham, Switzerland: Springer, 2019, pp. 187–198.
[47] Y. Wang, H. Li, L.-P. Chau, and A. C. Kot, "Embracing the dark knowledge: Domain generalization using regularized knowledge distillation," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 2595–2604.
[48] I. Gulrajani and D. Lopez-Paz, "In search of lost domain generalization," 2020, *arXiv:2007.01434*.
[49] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–12.
[50] L. Gong, J. Crego, and J. Senellart, "Enhanced transformer model for data-to-text generation," in *Proc. 3rd Workshop Neural Gener. Transl.*, 2019, pp. 148–156.
[51] W. Wang, H. Bao, L. Dong, J. Bjorck, Z. Peng, Q. Liu, K. Aggarwal, O. K. Mohammed, S. Singhal, S. Som, and F. Wei, "Image as a foreign language: BEiT pretraining for all vision and vision-language tasks," 2022, *arXiv:2208.10442*.
[52] H. Zhang, F. Li, S. Liu, L. Zhang, H. Su, J. Zhu, L. M. Ni, and H.-Y. Shum, "DINO: DETR with improved denoising anchor boxes for end-to-end object detection," 2022, *arXiv:2203.03605*.
[53] Z. Li, Z. Chen, F. Yang, W. Li, Y. Zhu, C. Zhao, R. Deng, L. Wu, R. Zhao, and M. Tang, "MST: Masked self-supervised transformer for visual representation," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 13165–13176.
[54] K. Han, Y. Wang, H. Chen, X. Chen, J. Guo, Z. Liu, Y. Tang, A. Xiao, C. Xu, Y. Xu, Z. Yang, Y. Zhang, and D. Tao, "A survey on vision transformer," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 87–110, Jan. 2023.
[55] C. R. Chen, Q. Fan, and R. Panda, "CrossViT: Cross-attention multi-scale vision transformer for image classification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 347–356.
[56] H. Lin, X. Cheng, X. Wu, F. Yang, D. Shen, Z. Wang, Q. Song, and W. Yuan, "CAT: Cross attention in vision transformer," 2021, *arXiv:2106.05786*.
[57] K.-H. Lee, X. Chen, G. Hua, H. Hu, and X. He, "Stacked cross attention for image-text matching," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 201–216.
[58] X. Xu, T. Wang, Y. Yang, L. Zuo, F. Shen, and H. T. Shen, "Cross-modal attention with semantic consistence for image–text matching," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 12, pp. 5412–5425, Dec. 2020.
[59] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017.
[60] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.
[61] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 630–645.
[62] A. Li, C. Zheng, L. Zhang, and X. Li, "Glance and gaze: A collaborative learning framework for single-channel speech enhancement," *Appl. Acoust.*, vol. 187, Feb. 2022, Art. no. 108499.
[63] S. Yu, Y. Ding, K. Qian, B. Hu, W. Li, and B. W. Schuller, "A glance-and-gaze network for respiratory sound classification," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 9007–9011.
[64] Q. Yu, Y. Xia, Y. Bai, Y. Lu, A. L. Yuille, and W. Shen, "Glance-and-gaze vision transformer," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 12992–13003.
[65] Z. Tian, C. Shen, and H. Chen, "Conditional convolutions for instance segmentation," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 282–298.
[66] K. Han, A. Xiao, E. Wu, J. Guo, C. Xu, and Y. Wang, "Transformer in transformer," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 15908–15919.
[67] J. Lei Ba, J. Ryan Kiros, and G. E. Hinton, "Layer normalization," 2016, *arXiv:1607.06450*.
[68] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9992–10002.

[69] X. Zhu, H. Hu, S. Lin, and J. Dai, "Deformable ConvNets v2: More deformable, better results," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9300–9308.
[70] H. Bao, L. Dong, F. Wei, W. Wang, N. Yang, X. Liu, Y. Wang, J. Gao, S. Piao, M. Zhou, and H.-W. Hon, "UniLMv2: Pseudo-masked language models for unified language model pre-training," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 642–652.
[71] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.
[72] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. Int. Conf. Learn. Represent.*, 2017, pp. 1–11.
[73] P. Goyal, P. Dollár, R. Girshick, P. Noordhuis, L. Wesolowski, A. Kyrola, A. Tulloch, Y. Jia, and K. He, "Accurate, large minibatch SGD: Training ImageNet in 1 hour," 2017, *arXiv:1706.02677*.
[74] D. Li, Y. Yang, Y.-Z. Song, and T. M. Hospedales, "Deeper, broader and artier domain generalization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5543–5551.
[75] A. Torralba and A. A. Efros, "Unbiased look at dataset bias," in *Proc. CVPR*, Jun. 2011, pp. 1521–1528.
[76] H. Venkateswara, J. Eusebio, S. Chakraborty, and S. Panchanathan, "Deep hashing network for unsupervised domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5385–5394.
[77] D. Li, Y. Yang, Y.-Z. Song, and T. Hospedales, "Learning to generalize: Meta-learning for domain generalization," in *Proc. AAAI Conf. Artif. Intell.*, 2018, vol. 32, no. 1, pp. 1–22.
[78] Z. Li, Z. Cui, S. Wang, Y. Qi, X. Ouyang, Q. Chen, Y. Yang, Z. Xue, D. Shen, and J.-Z. Cheng, "Domain generalization for mammography detection via multi-style and multi-view contrastive learning," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2021, pp. 98–108.
[79] Y. Shi, J. Seely, P. H. S. Torr, N. Siddharth, A. Hannun, N. Usunier, and G. Synnaeve, "Gradient matching for domain generalization," 2021, *arXiv:2104.09937*.

YUNSUNG CHO received the B.S. degree in digital imaging engineering from Chung-Ang University, Seoul, South Korea, in 2020, and the M.S. degree in imaging engineering from the Graduate School of Advanced Image Science, Multimedia and Film, Chung-Ang University. His current research interests include deep learning and computer vision.

JUNGMIN YUN received the B.S. degree in library and information science from Chung-Ang University, Seoul, South Korea, in 2022, where she is currently pursuing the M.S. degree with the Graduate School of Artificial Intelligence. Her current research interests include deep learning and natural language processing.

JUNEHYOUNG KWON received the B.S. degree in philosophy from Kyunghee University, Seoul, South Korea, in 2020, and the M.S. degree in digital imaging from Chung-Ang University, in 2022, where he is currently pursuing the Ph.D. degree with the Graduate School of Artificial Intelligence. His research interests include deep learning and computer vision.

YOUNGBIN KIM (Member, IEEE) received the B.S. and M.S. degrees in computer science and the Ph.D. degree in visual information processing from Korea University, in 2010, 2012, and 2017, respectively. From August 2017 to February 2018, he was a Principal Research Engineer with Linewalks. He is currently an Assistant Professor with the Graduate School of Advanced Imaging Science, Multimedia and Film, Chung-Ang University. His current research interests include data science and deep learning.