VQAMix: Conditional Triplet Mixup for Medical Visual Question Answering

Haifan Gong, *Student member, IEEE*, Guanqi Chen, *Student member, IEEE*, Mingzhi Mao, Zhen Li, Guanbin Li, *Member, IEEE*

***Abstract*— Medical visual question answering (VQA) aims to correctly answer a clinical question related to a given medical image. Nevertheless, owing to the expensive manual annotations of medical data, the lack of labeled data limits the development of medical VQA. In this paper, we propose a simple yet effective data augmentation method, VQAMix, to mitigate the data limitation problem. Specifically, VQAMix generates more labeled training samples by linearly combining a pair of VQA samples, which can be easily embedded into any visual-language model to boost performance. However, mixing two VQA samples would construct new connections between images and questions from different samples, which will cause the answers for those new fabricated image-question pairs to be missing or meaningless. To solve the missing answer problem, we first develop the Learning with Missing Labels (LML) strategy, which roughly excludes the missing answers. To alleviate the meaningless answer issue, we design the Learning with Conditional-mixed Labels (LCL) strategy, which further utilizes language-type prior to forcing the mixed pairs to have reasonable answers that belong to the same category. Experimental results on the VQA-RAD and PathVQA benchmarks show that our proposed method significantly improves the performance of the baseline by about 7% and 5% on the averaging result of two backbones, respectively. More importantly, VQAMix could improve confidence calibration and model interpretability, which is significant for medical VQA models in practical applications. All code and models are available at https://github.com/haifangong/VQAMix.**

***Index Terms*— Visual question answering, VQAMix, Medical image, Vision and language, Medical questions and answers, Data augmentation**

I. Introduction

Medical visual question answering (VQA) is a domain-specific VQA task aiming at predicting the correct answer given a medical image and a corresponding clinical question. This task requires a system to thoroughly understand the content of medical images and the semantics of the given question, even involving common sense understanding. Medical VQA can assist clinicians to obtain a second opinion on diagnosis and enhance their confidence in interpreting complex medical images. Besides, timely feedback from medical VQA techniques can be provided to patients who are interested in their disease status, which is helpful for patients to better understand their health status.

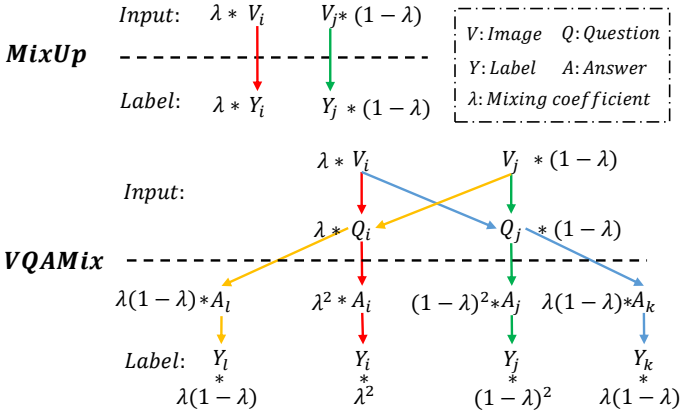

Fig. 1. MixUp vs. VQAMix. In MixUp, two images scaled by random weights (i.e., $\boldsymbol{\lambda}$, $\mathbf{1}-\boldsymbol{\lambda}$) are combined linearly, and their corresponding labels are fused with the same weight. In VQAMix, two image-question pairs $\{\boldsymbol{V_i}, \boldsymbol{Q_i}\}$ and $\{\boldsymbol{V_j}, \boldsymbol{Q_j}\}$ are mixed. When the mixed sample is sent to the VQA model, the linguistic feature extracted from $\boldsymbol{Q_i}$ will interact with the visual feature extracted from $\boldsymbol{V_j}$, which constructs a new connection $\{\boldsymbol{V_j}, \boldsymbol{Q_i}\}$. So is $\{\boldsymbol{V_i}, \boldsymbol{Q_j}\}$. Thus, the label for the mixed image-question pair consists of four answer labels ($\boldsymbol{Y_i}$ for $\{\boldsymbol{V_i}, \boldsymbol{Q_i}\}$, $\boldsymbol{Y_j}$ for $\{\boldsymbol{V_j}, \boldsymbol{Q_j}\}$, $\boldsymbol{Y_k}$ for $\{\boldsymbol{V_i}, \boldsymbol{Q_j}\}$ and $\boldsymbol{Y_l}$ for $\{\boldsymbol{V_j}, \boldsymbol{Q_i}\}$). And the weights of those answer labels are the probabilities of occurrence of those image-question pairs. The answer $\boldsymbol{A}$ is encoded as a one-hot vector $\boldsymbol{Y}$.

However, due to the lack of large-scale well-annotated datasets, the research of medical VQA is still in its infancy. Considering the types of problems in the VQA-MED [1] dataset are too simple to match the complex problem types in the real world, while the SLAKE [2] dataset carries out additional segmentation mask annotation for each image at design time, but such pixel-level annotation in real scenes is often time-consuming. There are only two manually-crafted medical VQA dataset appropriate: (1) VQA-RAD [3], which is composed of 315 radiology images and 3,515 question-answer (QA) pairs; (2) PathVQA [4], which contains 4,998 images with 32,795 QA pairs. In contrast, VQA 2.0 [5], one of the

This work was supported in part by the Guangdong Basic and Applied Basic Research Foundation (No.2020B1515020048), in part by the National Natural Science Foundation of China (No.61976250 and No.U1811463) and in part by the Guangzhou Science and technology project (No.202102020633). (Haifan Gong and Guanqi Chen contributed equally to this work. Corresponding author: Guanbin Li)

Haifan Gong, Guanqi Chen, Guanbin Li, and Mingzhi Mao are with the School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou 510000, China (e-mail: gonghf@mail2.sysu.edu.cn; chengq26@mail2.sysu.edu.cn; liguanbin@mail.sysu.edu.cn; mcsmmz@mail.sysu.edu.cn), Zhen Li is with The Chinese University of Hong Kong, Shenzhen, 518000, China (e-mail:lizhen@cuhk.edu.cn)

most widely used VQA datasets in the general scene, consists of 204,721 images and more than 1 million QA pairs. Hence, compared with general VQA, the annotated data of medical VQA is quite limited.

In the literature on medical VQA, Nguyen *et al.* [6] adopted transfer learning to address the data limitation problem. They constructed a few-shot classification task and a denoising reconstruction task to pre-train two convolutional neural networks separately based on external datasets and then combined the two pre-trained networks to extract visual features on the medical VQA dataset, which enhances the quality of visual representation. However, they did not consider the impact of linguistic representation on the medical VQA task, nor did they address the problem of data limitation on clinical question-answer pairs.

On the other hand, data augmentation is also a good solution to relieve data limitations. For the VQA task, it is challenging to maintain the semantic correctness of the augmented VQA samples [7]. Generally, previous data augmentation strategies for VQA can be roughly divided into three categories: language-only, vision-only, and multi-modal approaches. For language-only augmentation methods [8]–[10], they generate new questions and answers by rules and/or external language models. Vision-only augmentation method [11] attempts to produce new images by using a GAN-based re-synthesis model to remove objects that are irrelevant to the question and its corresponding answer. The multi-modal augmentation method [7] generates new VQA samples from the perspective of adversarial attacks. However, those methods heavily rely on complex models and even external datasets to increase the training data.

In this paper, we propose VQAMix, a data augmentation method, to generate new VQA samples during the training process. Technically, VQAMix combines two training samples with a random coefficient to improve the diversity of the training data without relying on external data. Unlike MixUp [12], which linearly weights the input image and label, VQAMix is related to combining the image-question pairs. Thus, combining (v, q, a) tuples directly will inevitably lead to the missing and meaningless answers problem: the missing answers are attributed to the combination of the two fabricated image-question pairs in Figure 1, while the meaningless answers are due to mixing (v, q, a) tuples without any type constrain on the answer. To address the above-mentioned issues, we propose the Conditional Triplet Mixup (CTM) scheme that contains two learning schemes. Firstly, we introduce a straightforward Learning with Missing Labels (LML) strategy to roughly exclude those missing answers. However, the labels in this scheme are meaningless, which has a negative impact on the learning of models. Thus, we further propose a Learning with Conditional-mixed Labels (LCL) strategy to further purify supervisory signals. Specifically, we leverage the prior knowledge of the question's category to constrain the mixup process. In particular, we only mixup the (v, q, a) tuples with the questions that belong to the same category, to force the candidate's answer to be meaningful. Conclusively, this design pursues the data augmentation on both images and questions, and thus facilitates the learning of visual and linguistic representations.

To demonstrate the effectiveness of our proposed method, we conduct comprehensive evaluations on the VQA-RAD [8] and PathVQA [4] benchmarks. Guided by the CTM strategies, our proposed VQAMix gains significant improvement applying to different medical VQA models. In particular, a strong baseline model [6] trained with VQAMix achieves state-of-the-art performance. Besides, the experimental results also demonstrate that VQAMix can improve the confidence calibration, which means that the predicted scores can better reflect the actual likelihood of correctness, and therefore can be used as a guarantee for the reliability of the model.

The contribution of this work is summarized as follows:

- We present a data augmentation method VQAMix to mitigate data limitation in medical VQA. Technically, VQAMix linearly combines two training samples with a random coefficient to generate a new synthetic sample. To our best knowledge, we are the first to apply MixUp to the vision and language domain.
- We propose the Learning with Missing Label (LML) and Learning with Conditional-mixed Label (LCL) strategy, to alleviate the missing answer and meaningless answer resulting from the combination of (v, q, a) tuples, respectively. LML roughly discards the missing Answers, while LCL further utilizes the language category prior to constraining the mixed answer being meaningful, which makes the mixup process more reasonable.
- We conduct extensive experiments on two medical VQA benchmarks to demonstrate the effectiveness of our proposed method. Experimental results show that our proposed method not only improves the performance but also shows a great confidence calibration and interpretability of medical VQA.

The rest of the paper is organized as follows. Section 2 reviews related works on VQA, Medical VQA, and data augmentation. Section 3 introduces the mechanisms of VQAMix and conditional learning strategies in detail. Section 4 evaluates the VQAMix on VQA-RAD [8] and PathVQA [4] benchmarks, including comparisons with state-of-the-art methods and an ablation study. Section 5 and Section 6 give the discussion and conclusion, respectively.

II. Related Work

A. Visual Question Answering

VQA has been a prevailing research field in recent years. A lot of large-scale VQA datasets and VQA algorithms have been proposed, which boost the development of VQA. Existing VQA algorithms mainly focus on reasoning on multi-modal representation, including attention mechanisms, compositional methods, and bilinear pooling schemes. Attention mechanisms [13]–[17] aim to adaptively focus on the relevant image regions based on the question representation. Compositional methods [18]–[22] attempt to compose several modules with different functions for answer inference. As for bilinear pooling, [23]–[26], [26], [27] proposed to employ the compact bilinear pooling methods to obtain joint representations of visual and linguistic features.

Many existing medical VQA approaches directly apply general VQA models to the medical domain. Typically, [3], [28]–[34] utilize an ImageNet pre-trained CNN like VGG or ResNet to extract feature of medical images and leverage a LSTM or transformer-based model to capture feature of questions. And they borrowed the cross-modal reasoning strategies from general VQA (e.g., SAN [35], MCB [23], BAN [24]) with a simple classifier for answers prediction. However, due to the large variance and scarcity of data, direct adaptation from the natural domain to the medical domain is prone to over-fitting. To deal with this problem, Nguyen *et al.* [6] utilized a meta-learning method MAML [36] and designed an unsupervised denoising reconstruction task to pre-train a visual feature extractor on external medical datasets to capture suitable visual representation for subsequent cross-modal reasoning. However, they only consider the data limitation in the aspect of medical image and neglect the influence of linguistic representation. In this paper, different from [6], we attempt to mitigate the data limitation problem by an ingenious cross-modal MixUp between two pairs of images and questions for data augmentation.

B. Data Augmentation

MixUp and its variants. MixUp was first introduced by Zhang *et al.* [12] for image classification. It works by generating synthetic samples by linearly combining a pair of images as well as their labels using the same mixing coefficients. MixUp has shown its superiorities in reducing the memorization of corrupt labels and improving the robustness of the model. Verma *et al.* [37] extended MixUp to feature level to produce better-generalized models. In addition to image classification, MixUp has also been applied to text classification. Guo *et al.* [38] attempted to adopt MixUp on word embeddings and sentence embeddings for data augmentation, which resulted in significant accuracy improvement on different models. Guo *et al.* [39] further improved MixUp in text classification by considering non-linear mixing operations. In this paper, we design the conditional VQAMix to medical VQA to improve the diversities of training data as a solution to data limitation.

Data augmentation in VQA. Compared to image and text classification, a few works have been done on overcoming data augmentation in VQA. Kafle *et al.* [8] firstly attempted to generate new questions by using preset templates and LSTM. Ray *et al.* [9] proposed to generate a set of logically consistent QA pairs to enhance the consistency of VQA models, relying on external datasets to improve the diversity of the generated questions. Similarly, the work of [10] aimed to solve the problem that VQA models lack consistency. It utilized a visual question generation model to generate a rephrasing question and required the VQA model to predict consistent answers among the rephrasing questions and original questions. From the perspective of image consistency, Agarwal *et al.* [11] generated new samples by removing inconsequential objects to enhance the robustness of VQA models against visual semantic variations. Tang *et al.* [7] proposed to generate adversarial examples for both images and questions to improve the performance of VQA models and resist adversarial attacks. Recently, Gong *et al.* [40] designed a multi-task learning framework by generating the pseudo labels to the unlabeled data according to the modal of medical image. It is worth noting that all of those methods rely heavily on complex models and even external datasets to augment data. This paper introduces a simple yet effective data augmentation method, which can be used to generate new VQA samples during the training process without additional networks or datasets.

III. Methodology

VQAMix is designed for data augmentation, which enhances the generalization of the model by increasing the diversity of training samples while reducing the possibility of rote memorization. Figure 2 shows an overview of our proposed method VQAMix which is composed of three operations: images mixing, questions mixing and labels mixing. In the following, we define the vanilla setting of medical VQA, then elaborate on the proposed VQAMix. After that, we introduce two learning strategies to handle the missing label issues and the meaningless labels issue, respectively.

A. Vanilla Setting for Medical VQA

The mainstream formulation of medical VQA is to select the answer A which best fits the given image V and question Q from the candidate answer set $\mathbb{A}$ according to previous works [6], [29], [31], [40]–[43]. For this purpose, we need to learn a function $F(\cdot)$ which maps each $\{V, Q\}$ pair to a score vector S where S_i denotes the score for the i-th candidate answer. The function $F(\cdot)$ usually consists of four parts: image encoder $F_v(\cdot)$, question encoder $F_q(\cdot)$, cross-modal reasoning $F_{cm}(\cdot)$, and classifier $F_{cls}(\cdot)$. We elaborate on each part below.

Image encoder. To extract the visual features of images for medical VQA, we usually utilize pre-trained convolution neural networks to obtain the intermediate or hierarchical feature maps.

Question encoder. The linguistic features of questions are commonly extracted by the recurrent neural networks (e.g., LSTM [44], GRU [45]) or transformer-based neural networks (e.g., BERT [46], [47]). Given padded word tokens, we first encode each word to word embedding, then we obtain the linguistic features with the above-mentioned neural networks.

Cross-modal reasoning. Cross-modal reasoning, also called cross-modal feature fusion, aims to encode the relationship between the visual features and linguistic features. Conventionally, attention mechanisms and bilinear pooling schemes are applied to model the above-mentioned relationship. The representative cross-modal reasoning modal includes SAN [35], BAN [24], MCB [23], etc.

Classifier. Since medical VQA is generally defined as a classification problem, distinct answers are considered as different categories. A common choice of the classifier is the multilayer perceptron (MLP).

With the above mentioned four components, we get the score vector S by the following formulation:

$$S = sigmoid(F_{cls}(F_{cm}(V_{feat}, Q_{feat}))), \tag{1}$$

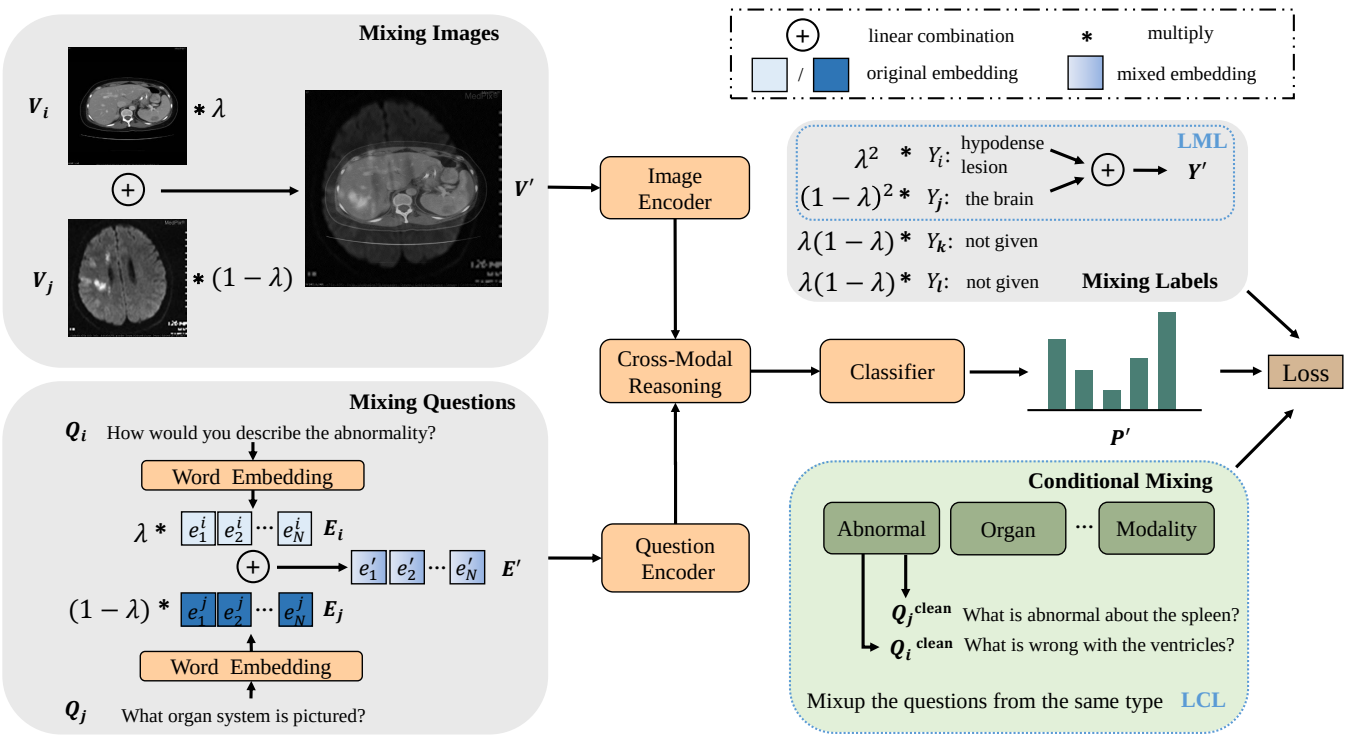

Fig. 2. An overview of our proposed VQAMix enhanced by Learning with Missing Labels (LML) and Learning with Conditional-mixed Labels (LCL) strategies. Two VQA samples are combined linearly in the training phase. To ensure that the mixed label can be used to supervise the learning of VQA models, both LML and LCL scheme discards those two unspecified labels to solve the missing labels issue. Moreover, to avoid meaningless answers, the LCL scheme further utilizes the category of the question to avoid the model suffering from meaningless mixed labels.

where V_{feat} and Q_{feat} denotes the visual feature and question feature extracted by $F_v(\cdot)$ and $F_q(\cdot)$, respectively.

After that, the predicted answer is the one with the highest score.

B. VQAMix

Let $\{V, Q, A\}$ denote a sample in the VQA dataset, where V, Q, A refer to an input image, a given question, and the corresponding answer respectively. The answer A is encoded as a one-hot vector Y. Given two VQA training samples $\{\{V_i, Q_i, A_i\}, \{V_j, Q_j, A_j\}\}$, VQAMix is used to generate a new training sample $\{V', Q', A'\}$. Firstly, VQAMix obtains a mixing coefficient λ from a Beta distribution $Beta(\alpha, \alpha)$, where α is a hyperparameter. Then, a linear combination is applied to two input images V_i, V_j with the mixing coefficient λ to obtain the mixed image V' as follows:

$$V' = \lambda V_i + (1-\lambda)V_j. \tag{2}$$

Different from image mixing, questions mixing is not applied at the input space, since the input space of the question is not continuous. Thus, VQAMix adopts the linear combination of two input questions at the embedding space:

$$E' = \lambda E_i + (1-\lambda)E_j. \tag{3}$$

where E' is the mixed question's embedding representation, E_i and E_j represent the embedding representation of the question Q_i and Q_j. More specifically, E_i and E_j are extracted by the following process. Two input questions are firstly trimmed to a maximum of N words, and they are zero-padding when their lengths are less than N. Thus, we get $Q_i = \{w_1^i, w_2^i, ..., w_N^i\}$ and $Q_j = \{w_1^j, w_2^j, ..., w_N^j\}$. Then, we utilize word embedding to map each word to a D-dim vector and obtain the representation embedding of the two questions as $E_i = \{e_1^i, e_2^i, ..., e_N^i\}$ and $E_j = \{e_1^j, e_2^j, ..., e_N^j\}$.

Given the mixed image and the mixed question, there exists 4 image-question pairs in this situation, including $\{V_i, Q_i\}$, $\{V_j, Q_j\}$, $\{V_i, Q_j\}$, and $\{V_j, Q_i\}$. The answers of $\{V_i, Q_i\}$ and $\{V_j, Q_j\}$ are A_i and A_j respectively, while the answers of $\{V_i, Q_j\}$ and $\{V_j, Q_i\}$ are not given. We suppose that the answers of $\{V_i, Q_j\}$ and $\{V_j, Q_i\}$ are A_k and A_l. And the answer labels of $\{V_i, Q_i\}$, $\{V_j, Q_j\}$, $\{V_i, Q_j\}$, and $\{V_j, Q_i\}$ are Y_i, Y_j, Y_k, and Y_l respectively. Then the labels mixing process can be constructed as:

$$Y' = \lambda^2 Y_i + (1-\lambda)^2 Y_j + \lambda(1-\lambda)Y_k + \lambda(1-\lambda)Y_l, \tag{4}$$

The coefficient of each label represents the corresponding image-question pair's probability distribution (also refer to Figure 1).

Due to that the answers A_k and A_l are missing, directly using Y' in the Eq. 4 to supervise the learning of VQA models could be non-optimal. Moreover, the mixed labels might be unmeaningful as there is no constrain of the mixed type of (V', Q'). Based on these consideration, we propose the conditional triplet mixup scheme to handle the above mentioned issues.

C. Conditional Triplet Mixup Scheme

Learning with missing labels. To handle the missing label issues in this work, we propose a simple and straightforward strategy named "Learning with Missing Labels" (LML) that directly discards those labels, which is expressed as:

$$Y' = \lambda^2 Y_i + (1-\lambda)^2 Y_j. \tag{5}$$

With this strategy, we calculate the binary cross-entropy loss between the predicted score S' (after sigmoid function) and the noisy label Y' to train VQA models:

$$L(S', Y') = \frac{1}{C}\sum_{c=1}^{C}[Y'_c \log(S'_c) + (1-Y'_c)\log(1-S'_c)], \tag{6}$$

where C is the number of answers in the candidate answer set $\mathbb{A}$.

Learning with conditional-mixed labels. In the strategy of LML, there exist noise components in label Y', which may negatively affect the performance of deep neural networks. To deal with this problem, we propose another strategy to make the mixed labels being meaningful, termed as "Learning with Conditional-mixed Labels (LCL)".

Considering the missing labels are intrinsically caused by mixing the answers that are not in the same domain, we propose conditional mixing to make the model learn with the conditional-mixed label. Specifically, there are three ways to implement the conditional mixing: (1) only mixup the (v, q, a) tuples with the same imaging model; (2) only mixup the (v, q, a) tuples with the same question category; (3) mixup the (v, q, a) tuples with the same image model and question category. However, here comes the question that which strategy should we use for meaningful data augmentation? In this work, we propose to mixup the (v, q, a) tuples with questions in the same category based on the following concerns: (1) the question and the answer are closer in the latent space compared with the question and the image, and the type of the question could directly reflect the type of answer, thus makes the mixed label to be meaningful; (2) the images of different modals are easy to distinguish, and the images are much more limited in medical VQA task compared with question pair, so that mixup images from the different modal could improve the diversity of image; (3) as some questions are related to the model & organ of the image, constraining the images from the same model & organ will reduce the uncertainty during the training process, thus making the model overfitted on these samples. For example, assuming there are two mixed pairs with (q1: what is the modal of the imaging; v1: a CT imaging; a1: CT), (q2: What modality might be the figure belong to; v2: an MRI imaging; a2: MRI). If we mix these pairs, the model can better learn the feature representation from the images with different modalities. Thus, we proposed the conditional question constrain, the category-specific question set $\mathbb{Q}_c$ which could be formulated as:

$$\mathbb{Q}_c = \{q_i \in Q | \text{ category of q equals to each other}\}, \tag{7}$$

where the category of the questions are obtained by the "question type" in the corresponding dataset. Based on mixing the (v, q, a) tuples with the question q in the same $\mathbb{Q}_c$, the mixed category of answers could be meaningful and we define the label of meaningful answers as Y''. With the Y'', we can calculate the loss to ignore the answers with unknown presence to reduce the impact of noise:

$$L(S', Y'') = \frac{1}{C}\sum_{c=1}^{C}[Y''_c \log(S'_c) + (1-Y''_c)\log(1-S'_c)], \tag{8}$$

where C is the number of answers in the candidate answer set $\mathbb{A}$. Thus, let B be the batchsize, the final loss of a training batch is:

$$L_B = \sum_{b=1}^{B} L_b. \tag{9}$$

TABLE I
DETAILS OF MEDICAL VISUAL QUESION ANSWERING DATASET.

Dataset	Data category	Training set	Validation set	Test set
VQA-RAD [3]	Images	315	-	315
	QA pairs	3,064	-	451
PathVQA [4]	Images	3,021	987	990
	QA pairs	19,755	6,279	6,761

IV. EXPERIMENTS

In this section, we comprehensively evaluate the performance of VQAMix on two commonly used medical VQA benchmarks: the VQA-RAD [3] and the PathVQA [4]. We compare it with several medical VQA models and other augmentation-based methods. After that, we carry out experiments to deeply analyze the proposed method, including the influence of VQAMix in different layers and the performance of different conditional mixing methods.

A. Dataset

The details of the dataset used in this work are shown in Table I. The VQA-RAD dataset [3] contains 315 radiology images and 3515 question-answer (QA) pairs generated by clinicians. And the whole dataset is split into a training set of 3064 QA pairs and a test set of 451 QA pairs. There are 11 categories of clinical questions: abnormality, attribute, color, count, modality, organ, plane, positional reasoning, object/condition presence, size, and others. According to different answer forms, questions can be divided into two parts: closed-ended questions with answers "yes/no" and other limited choices, and open-ended questions whose answers do not have a limited structure. This dataset contains a total of 458 kinds of answers in the training set, and the questions are different from each other.

The PathVQA dataset [4] contains 4,998 pathology images with 32,795 question-answer pairs, which are collected from PEIR digital library and the pathology book. There are eight types of questions, which includes closed-form, 'what', 'where', 'when', 'how', 'why', 'whose', and 'how much'. The minimum and the maximum number of questions with respect to one image are 1 and 14, respectively.

B. Metrics

To quantitatively measure the model's performance, we adopt accuracy as the evaluation criterion. Let P_i and Y_i denote the prediction and the label of sample i in the test set, and $\mathbb{T}$ represents the set of samples in the test set. The model accuracy is calculated as:

$$accuracy = \frac{1}{|\mathbb{T}|}\sum_{i\in\mathbb{T}}\mathbf{1}(P_i = Y_i), \tag{10}$$

To measure the confidence calibration of the network, we follow [48] to use two widely-used evaluation criteria: Expected Calibration Error (ECE) and Maximum Calibration Error (MCE). Expected Calibration Error reflects the difference in expectation between confidence and accuracy. Before calculation, we divide all the predictions on the test set into $M = 15$ interval bins of equal size. Then, it is computed by:

$$ECE = \sum_{m=1}^{M}\frac{|\mathbb{B}_m|}{|\mathbb{T}|}\left|acc\left(\mathbb{B}_m\right) - conf\left(\mathbb{B}_m\right)\right|, \tag{11}$$

where $\mathbb{B}_m$ denotes the set of samples whose predicted top-1 scores fall into the m-th bin. $acc(\mathbb{B}_m)$ and $conf(\mathbb{B}_m)$ are the accuracy and confidence of $\mathbb{B}_m$, which are defined as:

$$acc(\mathbb{B}_m) = \frac{1}{|\mathbb{B}_m|}\sum_{i\in\mathbb{B}_m}\mathbf{1}(P_i = Y_i), \tag{12}$$

$$conf(\mathbb{B}_m) = \frac{1}{|\mathbb{B}_m|}\sum_{i\in\mathbb{B}_m}\hat{S}_i, \tag{13}$$

where $\hat{S}_i$ is the predicted top-1 score of sample i, which represents the confidence of the prediction.

Maximum Calibration Error reflects the worst-case deviation between accuracy and confidence, which is obtained by:

$$MCE = \max_{m\in\{1,\ldots,M\}}\left|\,acc\left(\mathbb{B}_m\right) - conf\left(\mathbb{B}_m\right)\right|. \tag{14}$$

C. Experiments Setup

Baseline methods. To illustrate the generality of our proposed method, we use two medical VQA models as baselines including SAN-MEVF [6] and BAN-MEVF [6]. SAN-MEVF utilizes the proposed MEVF module as an image encoder, which is composed of two pre-trained branches. One branch that contains four 3×3 convolutional layers and a mean-pooling layer is pre-trained by the MAML algorithm [36] to address the few-shot classification task, while the other branch that consists of three 3×3 convolutional layers and two max-pooling layers is a convolutional denoising auto-encoder. SAN-MEVF leverages a GloVe [50] word embedding and a LSTM to capture question features. Then it uses a SAN module [35] to carry out multi-step reasoning on image and question features. Finally, a 2-layer MLP is adopted to predict the answer. As for BAN-MEVF, its composition is basically the same as SAN-MEVF, except for the cross-modal reasoning strategy. BAN-MEVF utilizes a BAN module [24] to obtain the joint representations of the image and question features and then feeds the multi-modal representations to the classifier.

Implementation details. Following the setting of the baseline methods mentioned above, the length of question N is set to 12. We apply the pre-trained MEVF module [6] to extract the 128-d visual feature. For linguistic features, we applied GloVe [50] to encode a word into a 600-d embedding. Then a 1024-d LSTM is applied to extract the semantic information from the question. All codes are implemented in PyTorch-1.10 and the models are trained on a single NVIDIA Tesla V100 GPU with 32GB memory. We use the Adam optimizer with a learning rate of 0.02 and with a weight decay of 0.0001 to train all the models. The batch size is set to 32. All models are trained for 80 epochs. For evaluation, we only use the manually collected (v, q, a) tuples (i.e., free-form) in the test set of VQA-RAD [3], and the test set of PathVQA [4]. All the reported results are obtained by averaging the results of 5 different seeds.

D. Comparison with State-of-the-art Methods

As shown in Table II, our proposed method is compared with 9 existing medical VQA models including SAN [3], MCB [3], HQS [34], BAN [6], SAN-MEVF [6], BAN-MEVF [6], BAN-MEVF+CR [42], BAN-MEVF+MMQ [41], CMSA-MTPT [40]. We re-implement two baselines with our training strategy for a fair comparison. Furthermore, we also re-implement BAN-MEVF+CR [42] by following its available code[1], since its reported result is based on both free-form and rephrased questions instead of just on free-form questions. Besides, we re-implement two advanced VQA data-augment methods, DAVQA [8] and SEADA [7], and apply them to the strong baseline model BAN-MEVF. We use the publicly available implementation of BAN-MEVF+SEADA [7][2]. Except for those re-implemented models, the results of other models are cited from their corresponding papers. Before the comparison, we first briefly introduce several existing state-of-the-art methods.

- SAN [35] modeled multi-step reasoning progress with a stacked attention model which gradually discovers the image regions related to the given question.
- MCB [23] proposed a multi-modal compact bilinear pooling mechanism that reduces the computational cost of feature fusion by mapping the outer product to a lower-dimensional space.
- HQS [34] is a hierarchical deep the multi-modal network that segregates questions with a query-specific approach to give the answer.
- BAN [24] stemmed from the bilinear multi-modal fusion in MCB [23]. For the sake of lower computational cost, it utilizes the low-rank bilinear pooling to reduce the rank of the weight matrix.
- SAN-MEVF [6] applied the MEVF framework on SAN [35] to better extract the feature of medical images. The details of the MEVF framework are illustrated in Section 4.1.

[1]https://github.com/Awenbocc/med-vqa
[2]https://github.com/zaynmi/seada-vqa

TABLE II

COMPARISON RESULTS WITH STATE-OF-THE-ART METHODS ON THE TEST SET OF VQA-RAD AND THE TEST SET OF PATHVQA. * INDICATES THE RE-IMPLEMENTED RESULT RUNNING ON OUR DEVICE WITH 5 DIFFERENT SEEDS. THE BEST RESULT IS SHOWN IN **BOLD**.

Setting	Models	VQA-RAD Dataset			PathVQA Dataset		
		Closed	Open	All	Closed	Open	All
Baselines	MCB [23]	60.6%	25.4%	46.5%	-	-	-
	HQS [34]	63.4%	12.9%	41.1%	-	-	-
	SAN [35]	57.2%	24.2%	44.0%	59.4%	1.6%	30.5%
	BAN [24]	66.5%	27.6%	51.0%	68.2%	2.9%	35.6%
	SAN-MAML [49]	69.7%	38.2%	57.1%	75.3%	5.4%	40.5%
	BAN-MAML [49]	72.4%	40.1%	60.7%	79.5%	5.9%	42.9%
	SAN-MEVF [6]	74.1%	40.7%	60.8%	81.0%	6.0%	43.6%
	BAN-MEVF [6]	75.1%	43.9%	62.6%	81.4%	8.1%	44.8%
	BAN-MEVF+CR* [42]	$79.3_{\pm 1.1}\%$	$52.4_{\pm 0.9}\%$	$68.5_{\pm 1.0}\%$	-	-	-
	CMSA-MTPT* [40]	$77.8_{\pm 0.4}\%$	$52.8_{\pm 1.8}\%$	$67.9_{\pm 0.8}\%$	-	-	-
	SAN-MEVF+MMQ* [41]	$73.0_{\pm 1.4}\%$	$46.3_{\pm 1.8}\%$	$62.3_{\pm 1.1}\%$	$83.7_{\pm 0.4}\%$	$9.6_{\pm 0.5}\%$	$46.8_{\pm 0.3}\%$
	BAN-MEVF+MMQ* [41]	$72.4_{\pm 0.9}\%$	$52.0_{\pm 1.1}\%$	$64.3_{\pm 0.7}\%$	$82.1_{\pm 0.5}\%$	$11.8_{\pm 0.6}\%$	$47.1_{\pm 0.4}\%$
Augmentation	BAN-MEVF+DAVQA* [8]	$76.2_{\pm 1.4}\%$	$51.2_{\pm 1.4}\%$	$66.2_{\pm 1.3}\%$	$83.4_{\pm 0.2}\%$	$8.5_{\pm 0.4}\%$	$46.1_{\pm 0.2}\%$
	BAN-MEVF+SEADA* [7]	$72.4_{\pm 1.5}\%$	$49.6_{\pm 2.0}\%$	$63.3_{\pm 1.2}\%$	$81.3_{\pm 0.3}\%$	$9.1_{\pm 0.5}\%$	$45.3_{\pm 0.4}\%$
	SAN-MEVF+VQAMix-C	$74.0_{\pm 2.4}\%$	$53.8_{\pm 1.9}\%$	$65.9_{\pm 1.9}\%$	$\mathbf{84.4}_{\pm 0.2}\%$	$12.1_{\pm 0.5}\%$	$48.4_{\pm 0.2}\%$
	BAN-MEVF+VQAMix-C	$\mathbf{79.6}_{\pm 1.5}\%$	$\mathbf{56.6}_{\pm 1.3}\%$	$\mathbf{70.4}_{\pm 1.1}\%$	$83.5_{\pm 0.2}\%$	$\mathbf{13.4}_{\pm 0.6}\%$	$\mathbf{48.6}_{\pm 0.3}\%$

- BAN-MEVF [6] used the bilinear attention network [24] to fuse the visual features extracted by the MEVF framework and the semantic textual feature.
- BAN-MEVF+CR [42] established a conditional reasoning-based framework that decouples the confounder of open-ended questions and close-ended questions.
- BAN-MEVF+MMQ [41] is a multiple meta-model quantifying method to learn meta-annotation for medical visual question answering.
- BAN-MEVF+DAVQA [8] proposed to expand the training data in visual question answering by leveraging a language model to generate new questions.
- BAN-MEVF+SEADA [7] designed a framework that expands the training data with the question translation and adversary training.
- CSMA-MTPT [40] proposed a multi-task pre-training method to enforce the image encoder and the feature fusion module to learn both linguistic compatibility feature and the visual concept.

Quantitative comparison with the state of the art methods. It can be seen from the left part of Table II that a strong baseline, BAN-MEVF [6], is equipped with our proposed VQAMix and LCL strategy achieves the best performance and significantly outperforms all the existing methods. In particular, the proposed method considerably surpasses the state-of-the-art approach CMSA-MTPT [40] by 1.6% and 2.7% w.r.t. accuracy on open-ended and closed-ended questions respectively. It is worth noting that our proposed method does not increase the parameters of networks.

Besides, compared with the two baseline methods, our proposed VQAMix with two learning strategies brings obvious performance improvement. Specifically, SAN-MEVF [35] gains 4.3% and 11.7% accuracy improvement on closed-ended and open-ended questions with the proposed VQAMix for training. VQAMix with LCL scheme effectively increases the performance of BAN-MEVF [24] by 7.9% accuracy overall.

Furthermore, we compare the proposed method VQAMix with two other data augmentation methods DAVQA [8] and SEADA [7].We apply those methods to the framework of BAN-MEVF [6]. As shown in Table II, the proposed method significantly outperforms DAVQA [8] by 5.3% overall. And our algorithm exceeds SEADA [7] by 7.2% accuracy on overall questions. It is worth noting that DAVQA [8] and SEADA [7] rely on additional model (e.g., NMT [51]) to generate new questions, which is time-consuming.

The right part of Table. II presents the quantitative results on the PathVQA [4] dataset. It can be seen that the proposed BAN-MEVF+VQAMix-C outperformed the previous BAN-MEVF+MMQ [41] by 2.5% on overall answers, which is due to the fact that the MMQ [41] merely consider the overfitting in the medical VQA domain. Although the MEVF [6] framework uses the auto-encoder method to alleviate data limitation by enhancing the visual feature representation, the VQAMix dramatically improves the performance by 3.8% on account of enhancing data diversity both linguistically and visually.

Trade-off analysis on efficiency and accuracy. We study the efficiency-accuracy trade-off of 5 state-of-the-art models on the VQA-RAD dataset, which is shown in Figure 3. The larger the circle is, the larger the number of parameters is in the model. The X and Y axes represent inference speed and accuracy, respectively. As we can see, the proposed BAN-VQAMix model significantly outperforms other baselines with similar parameters by a large margin, and it is about 8 times faster than the BAN-CR [42] with better performance, which demonstrates that the proposed VQAMix is quite effective and efficient.

E. Ablation Study

Ablation study of conditional triplet mixup strategies. To demonstrate the superiority of the proposed conditional triplet mixup strategies, we evaluate the learning methods based on the BAN-MEVF [6] on the VQA-RAD dataset.

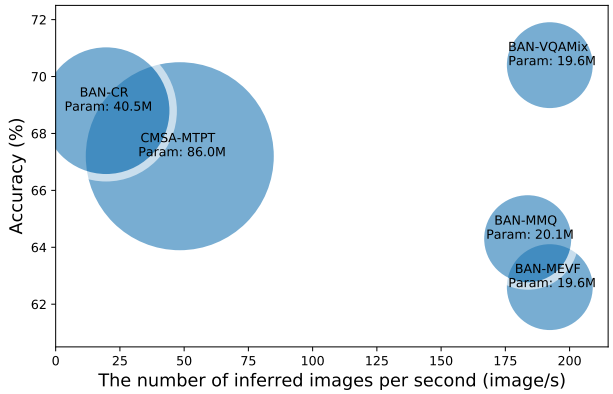

Fig. 3. Complexity (parameters), inference speed, and accuracy of the state-of-the-art methods on the VQA-RAD dataset. The model closer to the upper right corner performed better.

TABLE III

ABLATION STUDY ON THE DIFFERENT LEARNING METHODS OF VQAMIX BASED ON BAN-MEVF BACKBONE. THE STUDENT'S T-TEST IS COMPARED BETWEEN THE BEST PERFORMING METHOD (I.E., VQAMIX-COND-Q) AND OTHER METHODS WITH RESPECT TO THE VALUES OBTAINED BY 5 RUNS UNDER 5 SEEDS.

Methods	V	Q	Accuracy	P-value
VQAMix-LNL			$67.7_{\pm 0.8}\%$	<0.001 (0.0008)
VQAMix-LML			$68.4_{\pm 1.1}\%$	<0.005 (0.0014)
VQAMix-Cond-V	✓		$68.5_{\pm 0.9}\%$	<0.05 (0.0317)
VQAMix-Cond-V-FB	✓		$50.4_{\pm 0.7}\%$	<0.001 (1.06E-5)
VQAMix-Cond-Q (LCL)		✓	$70.4_{\pm 1.1}\%$	-
VQAMix-Cond-VQ-IN	✓	✓	$69.0_{\pm 0.9}\%$	<0.05 (0.0147)
VQAMix-Cond-VQ-UNION	✓	✓	$68.2_{\pm 0.8}\%$	<0.005 (0.0032)

The first row named "VQAMix-LNL" denotes the learning with the noisy label that directly takes all the terms in Eq.4 and uses Y_i and Y_j to replace Y_k and Y_l directly. The second row named "VQAMix-LML" denotes the learning with missing label strategy in Eq.5. "VQAMix-Cond-V" denotes only mixing the (v, q, a) tuples with the image in the same category and organ. "VQAMix-Cond-V-FB" denotes mixing the (v, q, a) tuples with the image in the same category/organ using the "Fixed batch" sampling strategy, which means the samples in a mini-batch are selected from the same organ and modal. "VQAMix-Cond-Q" mixing the (v, q, a) tuples with the question in the same category. "VQAMix-Cond-VQ-IN" denotes mixup (v, q, a) tuples with the question and image in the same category and organ. 'VQAMix-Cond-VQ-UNION" denotes mixing the (v, q, a) tuples with the questions or the images in the same category. In this work, we use the student's t-test to demonstrate the effectiveness of the proposed methods by comparing the overall accuracy of the best-performed "VQAMix-Cond-Q" and other methods under 5 different seeds. The requirements of the student's t-test are: (1) the compared sets of data contain the same number of elements; (2) comparisons of two units/groups; (3) the data are normally distributed. Since the first two requirements are satisfied (i.e., all samples in the group are the overall accuracy that comes from five different seeds), we focus on confirming the (3) assumption. As there are 7 groups of data with 5 in each group (i.e., the limited number of the data is less than 50), we adopt the Shapiro-Wilk statistic to prove whether the sampled data are normally distributed, and the result shows that the data in each group are normally distributed.

TABLE IV

EXPERIMENT RESULTS OF THE PROPOSED METHODS ON THE TESTSET OF VQA-RAD BASED ON FIVE-FOLD CROSS-VALIDATION. "BAN-MEVF" IS REPRESENTED BY "BAN-M" FOR SHORT. THE STUDENT'S T-TEST IS COMPARED BETWEEN THE BEST PERFORMING METHOD (I.E., BAN-M+VQAMIX-LCL) AND OTHER METHODS ON THE METRIC OF ACCURACY AND MACRO F1, RESPECTIVELY.

Models	Accuracy	P-value	Macro F1	P-value
BAN-M	$58.7_{\pm 1.3}\%$	0.0006	$29.3_{\pm 2.3}\%$	0.0002
BAN-M+VQAMix-LML	$62.4_{\pm 0.9}\%$	0.007	$41.2_{\pm 2.4}\%$	0.04
BAN-M+VQAMix-LCL	$64.7_{\pm 1.2}\%$	-	$44.3_{\pm 2.0}\%$	-

As shown in Table. III, the "VQAMix-LNL" underperformance as it intrinsically contains the non-existing answers which are regarded as noise. By neglecting the terms of non-existing answers, "VQAMix-LML" boosts the performance by 0.9% on average. The "VQAMix-Cond-V" merely improves the performance compared with the "VQAMix-LML", as it restricts the diversity of the mixed pairs. Moreover, all the conditional mixup methods outperform the "VQAMix-LML", while the "VQAMix-Cond-Q" (i.e., "VQAMix-Cond-LCL") strategy outperforms the "VQAMix-LML" by around 2%, which confirms our assumption in Section 3-C. It is also worth noting that although the "Fixed Batch" strategy used in "VQAMix-Cond-V-FB" could improve the possibility of mixing in the mini-batch, it is harmful to the optimization algorithm (e.g., Adam used in this work) as it will make the model over-fit on the mini-batch's data that come from the same organ & modal.

To alleviate the uncertainty caused by data partitioning, we further conduct the 5-fold cross-validation on the VQA-RAD dataset based on the BAN-MEVF backbone in Table IV with the metric of accuracy and Macro F1 score. As we can see, our VQA-Mix method is quite significant as it can not only improve the accuracy, but also significantly boosts the Macro F1 score (i.e., exceeds the baseline by 15%), which is quite significant as the class imbalance is quite common in the medical visual question answering domain.

Influence of VQAMix in different layers. To verify the effectiveness of each component in our proposed framework shown in Figure 2, we conduct experiments on several designed schemes, including mixing images only, mixing questions only, and applying VQAMix on the latent representations (i.e., feature maps). As shown in Table V, VQAMix with LCL strategy is applied to different layers of BAN-MEVF even partly applied. M_2 is our proposed scheme which mixes images before the image encoder and mixes questions before the question encoder. It achieves the highest accuracy on the overall questions, which confirms the superiority of our proposed mixing strategy shown in Figure 2. As for M_0 and M_1, both of them only apply the mixing operation in single-modal. Compared to M_2, their performance drops sharply, which suggests that it is of significance to mix images and

TABLE V

EXPERIMENTAL RESULTS OF APPLYING VQAMIX-LCL IN DIFFERENT LAYERS. REPORTED RESULTS ARE THE ACCURACY AND ECE SCORE OF OVERALL QUESTIONS ON THE VQA-RAD TEST SET. VMIX: MIXING IMAGES. QMIX: MIXING QUESTIONS. BE: BEFORE ENCODER. AE: AFTER ENCODER. THE BEST RESULT IS SHOWN IN **BOLD**.

Methods	VMix		QMix		Accuracy↑	ECE↓
	BE	AE	BE	AE		
M_0	✓	×	×	×	$64.0_{\pm 1.1}\%$	$15.1_{\pm 1.6}\%$
M_1	×	×	✓	×	$64.8_{\pm 0.9}\%$	$21.0_{\pm 2.4}\%$
M_2	✓	×	✓	×	$\mathbf{70.4}_{\pm 1.1}\%$	$\mathbf{11.4}_{\pm 1.2}\%$
M_3	×	✓	✓	×	$69.8_{\pm 1.6}\%$	$11.8_{\pm 1.3}\%$
M_4	✓	×	×	✓	$64.3_{\pm 0.9}\%$	$14.2_{\pm 1.6}\%$
M_5	×	✓	×	✓	$63.7_{\pm 1.4}\%$	$13.4_{\pm 1.8}\%$

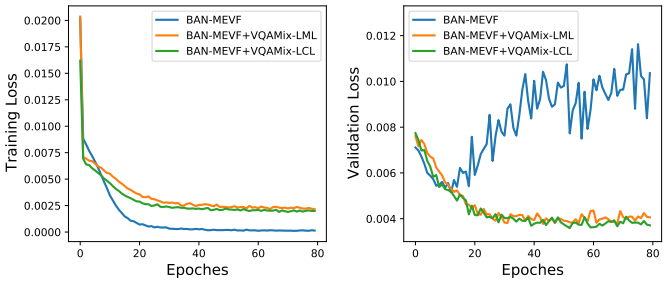

Fig. 4. Visualization results of the training loss and validation loss between the BAN-MEVF, the proposed VQAMix-LML, and VQAMix-LCL cross epochs on VQA-RAD dataset.

questions before encoder. As for M_5, its result reflects that mixing both images and questions on the features extracted by encoders harms the performance of the model, which is similar to the result reported in [12].

To uncover the rationale underlying better calibration, we further evaluate the calibration performance with ECE score by applying mixup to different layers of the VQAMix-LCL, which is shown in the right row of TableV. Compared with applying the mixup to the questions, mixing the images could lead to a much more calibrated result. Moreover, mixup before the encoder could lead to better calibration than mixup after the encoder.

V. DISCUSSION

In this section, we first analyze the results of the training and validation loss curve. Then we provide insights on the hyperparameter of Beta distribution and the exponent of the mixing coefficient, and analysis the model calibration and interpretability by visualizing some medical VQA samples. Finally, we provide some insights on why VQAMix works well on the medical visual question answering task.

A. Analysis on the Loss Curve

We visualize the training and validation losses in Figure 4. The loss curves of the vanilla BAN-MEVF, VQAMix-LML, and the VQAMix-LCL are colored in blue, orange, and green, respectively. Both VQAMix-LML and VQAMix-LCL could effectively alleviate the over-fitting on the training set by empirical risk minimization [12] and data-adaptive regularization [52]. More importantly, the losses of VQAMix-LCL

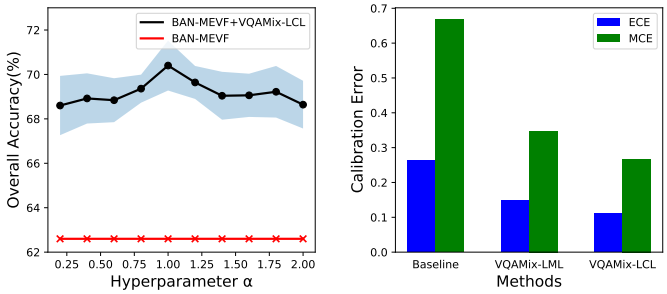

Fig. 5. Evaluation result of hyperparameters sensitivity and confidence calibration. In the left part, the red line denotes the performance of the BAN-MEVF while the black line represents the proposed BAN-MEVF+VQAMIX-LCL. The light blue area represents the error interval obtained by the seeds of 5 different runs. The right part shows the calibration result on the model BAN-MEVF (i.e., "Baseline" in the figure), BAN-MEVF equipped with VQAMix-LML, and BAN-MEVF equipped with VQAMix-LCL.

are lower than VQAMix-LML in the training phrases while achieving a lower loss in the validation phrases. The reason for this phenomenon is that VQAMix-LCL could generate more reasonable (v, q, a) tuples with the question category constrain, thus avoiding the model being affected by the noisy (v, q, a) tuples with the un-meaningful label.

B. Analysis on Beta Distribution and Exponent of Mixing Coefficient

Influence of hyperparameters on Beta distribution. The hyperparameter α decides a Beta distribution that is used to generate a mixing coefficient. Under the premise of no special conditions or different sampling rules, two samples of a pair should be equivalent, so the sampling needs to be between 0-1 and symmetric at about 0.5, and the Beta distribution exactly conforms to this characteristic. Given the specific α, the relationship between the distribution and the possible mixed images is shown in Figure 6. When alpha is lower than 1, the distribution would be a u-shape, and one image may dominate the mixture of images (e.g., Figure 6-(a)). When alpha equals 1, the Beta distribution is equivalent to uniform distribution that random samples images and questions without preference. In this situation, all kinds of different combinations are going to happen with equal probability (e.g., Figure 6-(b)). While alpha is greater than 1, the Beta distribution is bell-shaped, and the mixed images tend to evenly reflect the two mixed images (e.g., Figure 6-(c)).

To investigate its impact on performance, we evaluate VQAMix with different mixing rates. We sample the mixing rates from the beta distribution (i.e., Beta (α, α)) with the hyperparameter alpha varies from 0.2 to 2 with step 0.2. The results are shown in Figure 5 (left plot). For all α values considered, VQAMix significantly improves upon the baseline (62.6%). And the best performance is achieved when $\alpha = 1.0$. It is worth noting that we use the same hyper-parameters (i.e., $\alpha = 1$) to train the models on the PathVQA dataset.

Analysis on the exponent of the mixup coefficient. The rationale underlying this formulation is that the λ represents the possibility of mixup (v, q, a) pair. In that situation, when we only mix the images or questions, the power of λ should

TABLE VI
ANALYSIS ON THE EXPONENT OF MIXING COEFFICIENT BASED ON BAN-MEVF BACKBONE THAT ENHANCED BY VQAMIX-LCL. THE RESULTS ARE OBTAINED BY AVERAGING THE ACCURACY UNDER 5 DIFFERENT SEEDS.

Exponent value	1	2	3
Accuracy	$46.6_{\pm 1.0}\%$	$70.4_{\pm 1.1}\%$	$68.9_{\pm 1.3}\%$

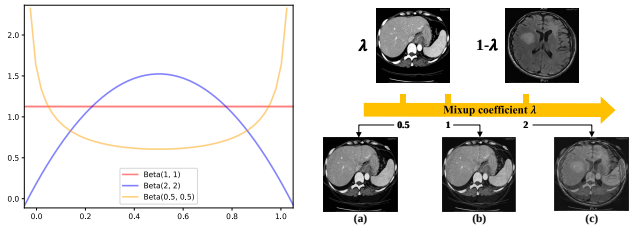

Fig. 6. Visualization result of the shape of Beta distribution under different alpha value, and the mixed images based on different mixing coefficient.

be 1. When we mix the (v, q, a) pair, the power of λ should be 2. We further conduct the experiments on the influence of the power value, which is shown in Table VI. The power value of 1 is not working as it disobeys the probability of mixing, which further confirms our design is reasonable as it could reflect the possibility distribution.

C. Analysis on Model Calibration and Interpretability

Effect of confidence calibration. The calibration of the model represents the matching degree between prediction confidence and accuracy, which is essential in the medical domain. In this part, we evaluate the confidence calibration of three medical VQA models including BAN-MEVF, BAN-MEVF+VQAMix-LML, and BAN-MEVF+VQAMix-LCL, with the ECE and MCE illustrated in Section 4.2. The results are shown in Figure 5 (right plot). Both VQAMix-LML and VQAMix-LCL can improve confidence calibration. This implies that neural networks tend to be more calibrated on the test set with the newly generated samples during the training progress, as the mixed (v, q, a) tuple with the meaningful label learning strategies mitigates the discrepancy between the average confidence and average accuracy. This implies that neural networks trained by our proposed method tend to be better calibrated on the test set since our proposed method can mitigate the discrepancy between the average confidence and average accuracy.

Interpretability analysis based on class-map activation. As shown in Figure 7, we conduct a qualitative comparison between the baseline BAN-MEVF [6] and our proposed VQAMix-LCL which is applied to BAN-MEVF with the help of grad class activation map (Grad-CAM) [53]. The visualized examples include four modalities (i.e., CT, MRI, X-Ray, and FFPE) of radiology images on different organs or tissues. The first column is the original images, the second column shows the visualization results of the baseline BAN-MEVF, and the last column shows the visualization results of our model BAN-MEVF+VQAMix-LCL.

The first row to the left shows an abdomen CT image. Our model focuses on the boundary area of the image while the baseline focuses on the center, which is not the region of the skeletal joint. The second row to the left shows a brain MRI image. The baseline locates non-organic areas, which leads to an incorrect answer. The first row to the right introduces an example of the chest X-Ray image. As the question focuses on whether the position of the image is tilted, medical VQA models should pay attention to as many regions as possible. In comparison, our model correctly answers the question by focusing on a larger area of the image. The second row to the right involves the example of the pathology FFPE images. The question asked the model to be aware of the cells in the imaging rather than the upper part of the imaging. Thus, the proposed method gives the right answer by focusing on the cells in the imaging. These examples have demonstrated that our model can focus on the reasonable regions of the image to answer the question, by taking advantage of the VQAMix strategy and the Meaningful label handling schedule.

From this figure, our model can give a reasonable answer even facing the complex Med-VQA task, by taking advantage of the MixUp strategy and the Meaningful label learning schedule. Still, there is a long way to go to achieve better interpretability.

D. Discussion on Why VQAMix Works

Mixup [12] has been a prevailing data augmentation-based technology to boost the performance of the model, and there are many successful attempts [37]–[39], [54], [55]. Still, the rationale underlying Mixup is unclear until a recent work [52] that theoretically proves Mixup corresponds to approximately minimizing an upper bound of the adversarial loss while serving as a data-adaptive regularization that reduces overfitting. Thus, we proposed the VQA-Mix that mainly focuses on the handle the overfitting of the limited training data. Still, we can't guarantee that all the mixed (v, q, a) pairs are with meaningful labels, even with the conditional mixed strategies proposed in this work. Nevertheless, by embedding the VQAMix into the current VQA models, we witness great success not only in the boost of accuracy and the Macro F1 score but also leads to better model calibration.

Moreover, according to [52] that mixup is served as the adversarial loss, we can explain the constraining the question type in another way. As the mixed labels are closer in the latent space in the learning with conditional-mixed strategy, the adversarial loss could be much more effective. Thus, the VQAMix process could achieve better performance and interpretability. In VQAMix, we generated the VQA pair by linearly combining the original VQA pairs. This process will be unavoidable to generate meaningless pairs. Still, the meaningless generated pairs with the soft labels can avoid the model over-fitting the limited samples in the training set with the hard label. Thus, we reasonably guess that VQAMix may contribute to the model's interpretability by taking the soft label as the supervision during the training process according

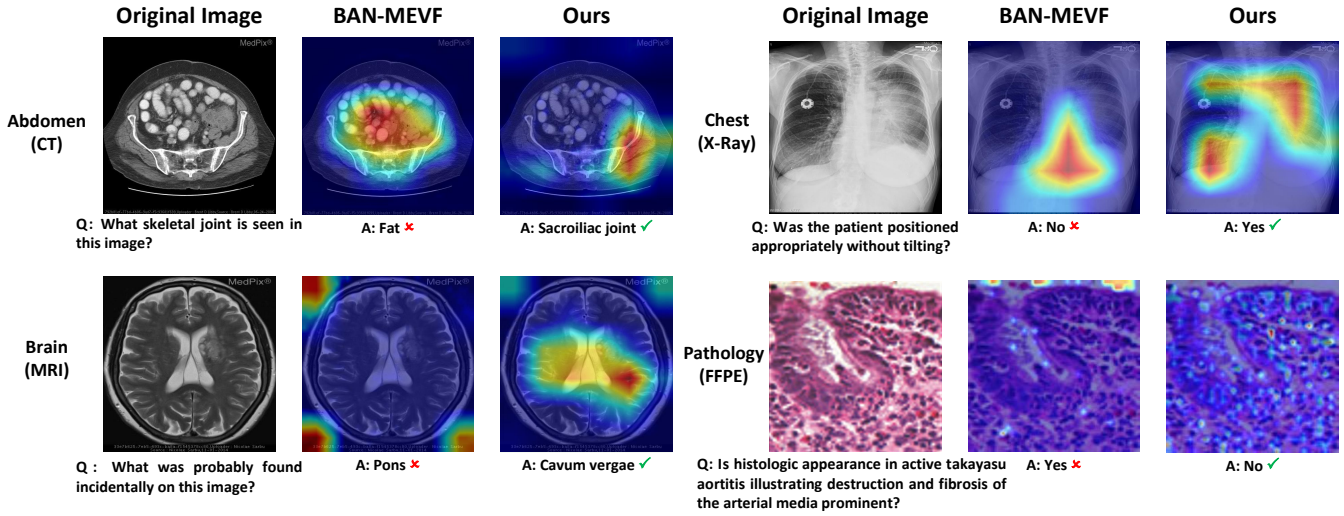

Fig. 7. Several visualized results between the baseline BAN-MEVF and our proposed method that embeds the VQAMix-LCL into the BAN-MEVF backbone. These examples are selected from the CT, MRI, X-Ray, and Formalin-Fixed Paraffin-Embedded (FFPE) modality with its corresponding organ (tissue).

to [56]. And we also believe that it is valuable to further discuss the impact of meaningless mixed images on the model in the future work to increase the interpretability of the model.

VI. Conclusion

In this paper, we present a new data augmentation method VQAMix to relieve data limitations in medical VQA. Technically, VQAMix combines two training samples with a random coefficient to improve the diversity of the training data without relying on external data. To alleviate the inherent missing answer issue and meaningless answer problem resulting from the combination of (v, q, a) tuples, we first develop the learning with the missing label strategy, which roughly discards the missing answers. After that, we further established the learning with conditional-mixed labels by adding the conditional mixup constrain with the prior knowledge of language category, which makes labels meaningful. Extensive experimental results on the VQA-RAD and PathVQA benchmarks show that our proposed method brings significant gains to different models. Furthermore, VQAMix could improve confidence calibration to make the predicted score better reflect the accuracy, and provides more reasonable class activation maps, which is meaningful for medical VQA models in practical applications.

For future works, our group may dig deeper into the reasoning process of medical VQA, which is a crucial issue in the current VQA models. Also, it would be interesting to broaden the VQAMix to other VQA methods which contain the open answer set. We leave these works to future efforts.

References

[1] A. B. Abacha, S. A. Hasan, V. V. Datla, J. Liu, D. Demner-Fushman, and H. Müller, "Vqa-med: Overview of the medical visual question answering task at imageclef 2019.," in *CLEF (Working Notes)*, 2019.

[2] B. Liu, L.-M. Zhan, L. Xu, L. Ma, Y. Yang, and X.-M. Wu, "Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering," in *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pp. 1650–1654, IEEE, 2021.

[3] J. J. Lau, S. Gayen, A. B. Abacha, and D. Demner-Fushman, "A dataset of clinically generated visual questions and answers about radiology images," *Scientific data*, vol. 5, no. 1, pp. 1–10, 2018.

[4] X. He, Z. Cai, W. Wei, Y. Zhang, L. Mou, E. P. Xing, and P. Xie, "Towards visual question answering on pathology images," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Virtual Event, August 1-6, 2021*, pp. 708–718, 2021.

[5] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh, "Making the v in vqa matter: Elevating the role of image understanding in visual question answering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6904–6913, 2017.

[6] B. D. Nguyen, T.-T. Do, B. X. Nguyen, T. Do, E. Tjiputra, and Q. D. Tran, "Overcoming data limitation in medical visual question answering," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 522–530, Springer, 2019.

[7] R. Tang, C. Ma, W. E. Zhang, Q. Wu, and X. Yang, "Semantic equivalent adversarial data augmentation for visual question answering," in *European Conference on Computer Vision*, pp. 437–453, Springer, 2020.

[8] K. Kafle, M. Yousefhussien, and C. Kanan, "Data augmentation for visual question answering," in *Proceedings of the 10th International Conference on Natural Language Generation*, pp. 198–202, 2017.

[9] A. Ray, K. Sikka, A. Divakaran, S. Lee, and G. Burachas, "Sunny and dark outside?! improving answer consistency in VQA through entailed question generation," in *Proceedings of the 2019 Conference on Empirical Methods in NaturalLanguage Processing and the 9th International Joint Conference onNatural Language Processing, Hong Kong, China, November 3-7, 2019*, pp. 5859–5864, 2019.

[10] M. Shah, X. Chen, M. Rohrbach, and D. Parikh, "Cycle-consistency for robust visual question answering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6649–6658, 2019.

[11] V. Agarwal, R. Shetty, and M. Fritz, "Towards causal vqa: Revealing and reducing spurious correlations by invariant and covariant semantic editing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9690–9698, 2020.

[12] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *International Conference on Learning Representations*, 2018.

[13] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6077–6086, 2018.

[14] H. Nam, J.-W. Ha, and J. Kim, "Dual attention networks for multimodal reasoning and matching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 299–307, 2017.

[15] P. Gao, Z. Jiang, H. You, P. Lu, S. C. Hoi, X. Wang, and H. Li, "Dynamic fusion with intra-and inter-modality attention flow for visual question answering," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6639–6648, 2019.

[16] L. Peng, Y. Yang, Z. Wang, X. Wu, and Z. Huang, "Cra-net: Composed relation attention network for visual question answering," in *Proceedings of the 27th ACM International Conference on Multimedia*, pp. 1202–1210, 2019.

[17] J. Yu, W. Zhang, Y. Lu, Z. Qin, Y. Hu, J. Tan, and Q. Wu, "Reasoning on the relation: Enhancing visual representation for visual question answering and cross-modal retrieval," *IEEE Transactions on Multimedia*, vol. 22, no. 12, pp. 3196–3209, 2020.

[18] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein, "Learning to compose neural networks for question answering," in *Proceedings of NAACL-HLT*, pp. 1545–1554, 2016.

[19] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein, "Neural module networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 39–48, 2016.

[20] R. Hu, J. Andreas, M. Rohrbach, T. Darrell, and K. Saenko, "Learning to reason: End-to-end module networks for visual question answering," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 804–813, 2017.

[21] J. Johnson, B. Hariharan, L. Van Der Maaten, J. Hoffman, L. Fei-Fei, C. Lawrence Zitnick, and R. Girshick, "Inferring and executing programs for visual reasoning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2989–2998, 2017.

[22] R. Hu, J. Andreas, T. Darrell, and K. Saenko, "Explainable neural computation via stack neural module networks," in *Proceedings of the European Conference on Computer Vision*, pp. 53–69, 2018.

[23] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach, "Multimodal compact bilinear pooling for visual question answering and visual grounding," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 457–468, 2016.

[24] J.-H. Kim, J. Jun, and B.-T. Zhang, "Bilinear attention networks," in *Advances in Neural Information Processing Systems*, pp. 1564–1574, 2018.

[25] Z. Yu, J. Yu, C. Xiang, J. Fan, and D. Tao, "Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering," *IEEE transactions on neural networks and learning systems*, vol. 29, no. 12, pp. 5947–5959, 2018.

[26] J. Kim, K. W. On, W. Lim, J. Kim, J. Ha, and B. Zhang, "Hadamard product for low-rank bilinear pooling," in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, OpenReview.net, 2017.

[27] H. Ben-younes, R. Cadène, M. Cord, and N. Thome, "MUTAN: multimodal tucker fusion for visual question answering," in *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pp. 2631–2639, IEEE Computer Society, 2017.

[28] X. Yan, L. Li, C. Xie, J. Xiao, and L. Gu, "Zhejiang university at imageclef 2019 visual question answering in the medical domain.," in *CLEF (Working Notes)*, 2019.

[29] M. Vu, R. Sznitman, T. Nyholm, and T. Löfstedt, "Ensemble of streamlined bilinear visual question answering models for the imageclef 2019 challenge in the medical domain," in *CLEF 2019*, vol. 2380, 2019.

[30] B. Jung, L. Gu, and T. Harada, "bumjun_jung at vqa-med 2020: VQA model based on feature extraction and multi-modal feature fusion," in *Working Notes of CLEF 2020 - Conference and Labs of the Evaluationv Forum, Thessaloniki, Greece, September 22-25, 2020*, vol. 2696 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2020.

[31] G. Chen, H. Gong, and G. Li, "HCP-MIC at VQA-Med 2020: Effective visual representation for medical visual question answering," in *Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, September 22-25, 2020*, vol. 2696 of *CEUR Workshop Proceedings*, 2020.

[32] H. Gong, R. Huang, G. Chen, and G. Li, "SYSU-HCP at VQA-Med 2021: A data-centric model with efficient training methodology for medical visual question answering," in *CLEF 2021 – Conference and Labs of the Evaluation Forum, September 21–24, 2021, Bucharest, Romania*, CEUR Workshop Proceedings, 2021.

[33] F. Ren and Y. Zhou, "Cgmvqa: A new classification and generative model for medical visual question answering," *IEEE Access*, vol. 8, pp. 50626–50636, 2020.

[34] D. Gupta, S. Suman, and A. Ekbal, "Hierarchical deep multi-modal network for medical visual question answering," *Expert Systems with Applications*, vol. 164, p. 113993, 2021.

[35] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola, "Stacked attention networks for image question answering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 21–29, 2016.

[36] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, vol. 70, pp. 1126–1135, 2017.

[37] V. Verma, A. Lamb, C. Beckham, A. Najafi, I. Mitliagkas, D. Lopez-Paz, and Y. Bengio, "Manifold mixup: Better representations by interpolating hidden states," in *International Conference on Machine Learning*, pp. 6438–6447, 2019.

[38] H. Guo, Y. Mao, and R. Zhang, "Augmenting data with mixup for sentence classification: An empirical study," *arXiv preprint arXiv:1905.08941*, 2019.

[39] H. Guo, "Nonlinear mixup: Out-of-manifold data augmentation for text classification.," in *AAAI*, pp. 4044–4051, 2020.

[40] H. Gong, G. Chen, S. Liu, Y. Yu, and G. Li, "Cross-modal self-attention with multi-task pre-training for medical visual question answering," in *ACM International Conference on Multimedia Retrieval(ICMR)*, 2021.

[41] T. Do, B. X. Nguyen, E. Tjiputra, M. Tran, Q. D. Tran, and A. Nguyen, "Multiple meta-model quantifying for medical visual question answering," in *MICCAI*, 2021.

[42] L.-M. Zhan, B. Liu, L. Fan, J. Chen, and X.-M. Wu, "Medical visual question answering via conditional reasoning," in *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 2345–2354, 2020.

[43] M. H. Vu, T. Löfstedt, T. Nyholm, and R. Sznitman, "A question-centric model for visual question answering in medical imaging," *IEEE Transactions on Medical Imaging*, vol. 39, pp. 2856–2868, 2020.

[44] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[45] K. Cho, B. van Merriënboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder–decoder approaches," in *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pp. 103–111, 2014.

[46] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*, pp. 4171–4186, 2019.

[47] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "Biobert: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020.

[48] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *International Conference on Machine Learning*, pp. 1321–1330, 2017.

[49] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *ICML*, 2017.

[50] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing*, pp. 1532–1543, 2014.

[51] G. Klein, Y. Kim, Y. Deng, J. Senellart, and A. M. Rush, "Opennmt: Open-source toolkit for neural machine translation," in *Proceedings of ACL 2017, System Demonstrations*, pp. 67–72, 2017.

[52] L. Zhang, Z. Deng, K. Kawaguchi, A. Ghorbani, and J. Zou, "How does mixup help with robustness and generalization?," in *International Conference on Learning Representations*, 2021.

[53] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.

[54] Y. Wang, W. Wang, Y. Liang, Y. Cai, and B. Hooi, "Mixup for node and graph classification," in *Proceedings of the Web Conference 2021*, pp. 3663–3674, 2021.

[55] J.-H. Kim, W. Choo, and H. O. Song, "Puzzle mix: Exploiting saliency and local statistics for optimal mixup," in *International Conference on Machine Learning*, pp. 5275–5285, PMLR, 2020.

[56] S. Thulasidasan, G. Chennupati, J. A. Bilmes, T. Bhattacharya, and S. E. Michalak, "On mixup training: Improved calibration and predictive uncertainty for deep neural networks," in *NeurIPS*, 2019.