VISION TRANSFORMER EQUIPPED WITH NEURAL RESIZER ON FACIAL EXPRESSION RECOGNITION TASK

Hyeonbin Hwang[1*]   *Soyeon Kim*[1*]

Wei-Jin Park[2]   Jiho Seo[2]   Kyungtae Ko[2]   Hyeon Yeo[1†]

[1]KAIST, Korea   [2]ACRYL, Korea

ABSTRACT

When it comes to wild conditions, Facial Expression Recognition is often challenged with low-quality data and imbalanced, ambiguous labels. This field has much benefited from CNN based approaches; however, CNN models have structural limitations to see the facial regions in distance. As a remedy, Transformer has been introduced to vision fields with a global receptive field but requires adjusting input spatial size to the pretrained models to enjoy its strong inductive bias at hands. We herein raise the question of whether using the deterministic interpolation method is enough to feed low-resolution data to Transformer. In this work, we propose a novel training framework, Neural Resizer, to support Transformer by compensating information and downscaling in a data-driven manner trained with loss function balancing the noisiness and imbalance. Experiments show our Neural Resizer with F-PDLS loss function improves the performance with Transformer variants in general and nearly achieves the state-of-the-art performance.

Index Terms— Facial Expression Recognition, Vision Transformer, Learnable Resizer, Deep Learning

1. INTRODUCTION

Facial expressions recognition (FER) has recently drawn much attention and has accomplished tremendous achievements thanks to the advancement of deep learning techniques. However, when it comes to in-the-wild (ITW) scenarios, several inherent problems are posed that make this task more challenging than classic vision benchmarks [1, 2].

First of all, datasets obtained from crowdsourcing inevitably elicit large image variance as in size, alignment, occlusion, and pose [3]. To fully utilize a convolutional neural network(CNN), training images are often preprocessed along with normalization techniques to suppress other variations. Nevertheless, the inherent ambiguity on individual facial expression become more evident on low-quality data, smoothing out the fine-grained features which contribute significantly to the overall performance.

The second challenge is the inherent problem of the FER task originating from the human subjectivity when annotating [2, 4] – noisy labels and class imbalance. Compared to annotating some discrete objects in an image, facial expression classes can not be consistently annotated by individuals with different backgrounds. Meanwhile, most of the real-world datasets inevitably face imbalanced classes, resulting in a bias against the minor classes [5]. While careful data filtering strategies may help, given that the FER task generally has a limited number of data, one should take a risk of having either fewer training examples for major classes or limited augmented data for minor classes.

Several previous works have focused to resolve the aforementioned challenges with CNN based approach. In respect to the architecture, considering that the FER task heavily depends on fine-grained key features, CNN with attention network [6, 7] was introduced to give more weights for the important regions. Based on such CNN models, [8] devised a weighted loss on each class's relative portion of the total dataset, [9] used loss values as cues to catch the outliers and the noisy data samples, and [10] proposed adaptive label smoothing loss to resolve different confusion distribution of labels inspired by [11, 12].

Orthogonal to these works, recent studies explore the applicability of Vision Transformer owing to its advantages on global reception fields with self-attention [13–15]. However, one of the drawbacks is that images must be all uniformly resized, generally *upscaled from the low-resolution*, to fit into the pretrained Transformer input spatial size. However, such practice may not be optimal because the recommended approach for adoption in the downstream target task is to use a higher resolution than the pretrained dataset [16].

Contribution. Our contribution is summarized to threefold: *(i)* We propose a novel training framework to fully provide Vision Transformer variants with strong generalization power by a trainable resizing module, Neural Resizer. *(ii)* We propose a novel loss function that puts importance on minor as well as ambiguous classes adaptively. *(iii)* We empirically show that one of the SOTA transformer, Swin-Transformer [17], achieves comparable performance to the state-of-the-art CNN based approaches when jointly trained with our proposed framework.

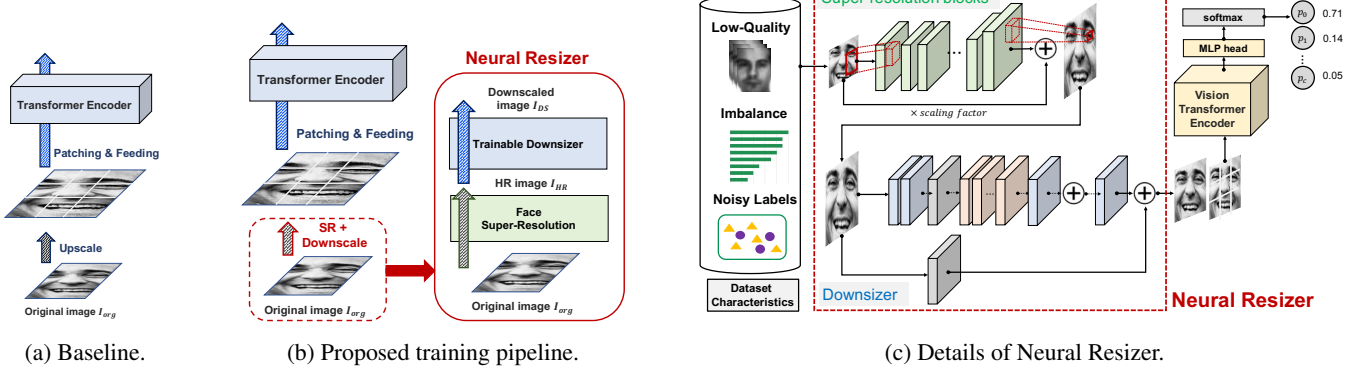* Equal contributions.

† Corresponding author

(a) Baseline.	(b) Proposed training pipeline.	(c) Details of Neural Resizer.

Fig. 1: **(a),(b)** The comparison of the naive use case and the proposed framework, Neural Resizer, to apply vision transformer variants in FER Task. Blue blocks : are only involved in parameter updating. **(c)** Before feeding to Transformer directly, we organize a two-step process, which first maps low-resolution images to a high resolution followed by a trainable downsizer to match the fixed input spatial size for Transformer.

2. PROPOSED ARCHITECTURE

2.1. Neural Resizer for Vision Transformer

Our goal throughout Neural Resizer(NR), f_Θ , is to feed low-quality image in a way that Vision Transformer can well adapt to a domain-specific task. Inspired by previous works on trainable downscaling [18, 19], we thoroughly follow the generalized formulation including the compensation process to enrich the image information. Let I_{org}, I_{hr}, I_{ds} denote original image, high-resolution image and down-scaled image where each of the spatial size is $I_{org} \in R^{H \times W \times C}$, $I_{hr} \in R^{hH \times hW \times hC}$, $I_{ds} \in R^{tH \times tW \times tC}$ respectively and h, t is the scaling factor for the high-resolution and target size. Then, the Neural Resizer is formulated as an approximate mapping function $f_\Theta : R^{H \times W \times C} \mapsto R^{tH \times tW \times tC}$. To train f_Θ, we decompose the function as

$$f_\Theta = g_\theta \circ T(f_{sr}(I_{org})) \circ g_\phi + T(f_{sr}(I_{org})) \quad (1)$$

where f_{sr}, T denote a pretrained super-resolution function and a conventional deterministic interpolation method which do not involve in updating parameter process respectively, and $\Theta \in \{\theta, \phi\}$ denotes trainable parameters. In practice, f_{sr} can be any type of off-the-shelf pretrained super resolution function and g_θ, g_ϕ can be implemented using simple stacked CNN layers.

2.2. Learning with Noisy and Imbalanced Labels

In the real world, not only accuracy of annotation by humans but the balance between the classes is also hardly guaranteed. To minimize human error, *voting* is often employed to annotate each data by multiple annotators to construct a class distribution. [10] predefined voting information and class distribution of datasets as prior knowledge and implemented modified label smoothing (Prior Distribution Label Smoothing,

PDLS) loss function to overcome conventional Cross Entropy which applies to only a single class with the highest vote count.

Nevertheless, PDLS still overlooked the imbalance issue. To handle both of the aforementioned issues, we propose a novel loss function called F-PDLS(Focal PDLS), which applies noise distribution to the classifying task while compensating noise balance to improve accuracy additionally, inspired by [11]. Our loss is designed to utilize prior knowledge of the label's confusion and focus on difficult tasks(classes) to classify correctly so that the model can cover the imbalance of labeled ITW datasets as shown in Eq.2 and Eq.3.

$$L_{F-PDLS} = -\sum_{c \in C}(1 - \sigma(z_c))^\gamma * L_{PDLS}^c \quad (2)$$

$$L_{PDLS}^c = (t_c * \alpha + d_{kc} * (1 - \alpha)) * log(\sigma(z_c)) \quad (3)$$

where z_c is the output of the model and L_{PDLS}^c is PDLS loss for each class $c \in C$ respectively.

All notations in Eq.2 are similar to PDLS loss function in [10] except a modulating factor to the PDLS loss expressed as $(1 - \sigma(z_c))^\gamma$ where γ is a positive-valued focusing parameter.

3. EXPERIMENTS

3.1. Datasets

FERPlus consists of 28,709 training images, 3,589 validation images and 3,589 test images. They are all in grayscale and size of 48 × 48. Each image was manually assigned to one of 8 predefined expression classes - neutral, happiness, surprise, sadness, anger, disgust, fear, contempt - by in total of 10 annotators.

RAF-DB contains 29,672 facial images annotated with basic or compound expressions independently labeled by 40

Table 1: Comparison with various state-of-the-art **small-sized** Transformers on FERPlus, tested with sole backbone architecture and our framework

Models	CE + Vanilla	F-PDLS + Proposed
ViT [16]	88.84	88.87
DeiT [20]	88.00	88.09
ConViT [21]	88.12	88.53
XCiT [22]	88.22	88.81
Swin-S	88.69	**89.28**

Table 2: Ablation study on the effect of each module, when **downscaling** and **upscaling** images, tested with **Swin-S**, on FERPlus, using F-PDLS

Setting	model	STN	Up.	Down.	Acc.
a	Swin-S	-	Bi.	-	88.69
b	Swin-S	-	Bi.	LTR	88.53
c	Swin-S	-	SR	Bi.	89.03
d	Swin-S	-	SR	LTR	89.28
e	**Swin-B**	✓	**SR**	**LTR**	**89.50**

Table 3: Comparison across the effect of the loss function, tested on both Vanilla Swin-T and Proposed architecture, **with Swin-B**

Loss	Vanilla	Proposed
Cross-Entropy	88.72	88.87
PDLS [10]	88.69	88.91
F-PDLS (ours)	**88.78**	**89.50**

annotators. It only consists of seven different expressions, excluding contempt. For our experiment, we use aligned RAF-DB, where all 12,271 training images and 3,068 test images are resized to 100×100.

3.2. Implementation Details

As our backbone architecture, we utilize the pretrained Transformer families released from the official library with the expected image resolution of 224. We used Deep Iterative Collaboration(DIC) model [23] pretrained with facial dataset(e.g, Helen [24]) for super-resolution, while in terms of the resizer, we follow the previous work configuration [19]. For the hyperparameter setting, we trained our model for 100 epochs with a batch size of 64. We set the initial learning rate to 1e-4, while decreasing it one quarter every 10 epochs. We empirically employ constructive data augmentation techniques because the FER task is sensitive to certain types of augmentation such as rotation. To evaluate on FERPlus, we follow the test setting referred in [4], and since its images consist of 48×48 spatial size, we upscale them into high-resolution of 384×384. In contrast, RAF-DB dataset has 100×100 size images, so we first downscale them to 48×48 and after upscale to 384×384 due to internal API issues.

3.3. Evaluation

3.3.1. Main Results

Table.1 shows the overall result of our proposed framework tested with various backbone Transformers. We observe a general trend in the increase of accuracy with our proposed framework, demonstrating its broad applicability with many Transformer-based variants. Also, Table.3 confirms this idea as the performance increases regardless of the loss function choice. Specifically, while ViT [16] has the best accuracy in the vanilla setting, we observe that Swin Transformer [17] equipped with Neural Resizer and F-PDLS achieves the best accuracy. We conjecture that our approach is well compatible with Swin Transformer(Swin-T) architecture which flexibly observes features map at various scales. In other words, its sequential fine-unit patch merging step can distinguish the fine-grained features more attentively by our proposed approach.

Therefore, we use Swin-T in the following experiments and denote Swin-S, Swin-B, and Swin-L for Swin-Small, Swin-Base, and Swin-Large, respectively.

3.3.2. Effectiveness of Each Module

Quantitative Results Table.2 shows the ablation study of replacing each module with deterministic bilinear interpolation. When using Swin-S, our method **(d)** outperforms interpolation based traditional approach **(a)** by **0.59%** in total. It can be concluded that Learning-To-Resize Module(LTR) [19] itself has limitations when applied to interpolated low resolution data, but works better when downsized from the higher resolution with more information. The gist from these results is the importance of data quality before neurally downsizing. Thus, we exploit a simple module, Spatial Transformer Network [25], to align the necessary features before feeding to Super Resolution (SR) modules as an extended study. Even though the depth of Swin-T does not show a strong correlation with the maximum accuracy attained, there is a tendency towards slight improvement for deeper models. Thus, we select the middle level Swin-B when applying STN. Setting **(e)** shows our best result, which outperforms our base case **(a)** by **0.81%**.

Qualitative Results To visually deliver the role of NR, the examples in Figure 2 presents that our framework successfully captures fine-grained features like the line of the wrinkles compared to the deterministic interpolation approaches. That is, image shape is not notably changed, but the edges of the discriminant features are more conspicuously accented which facilitates the classification process. Compared to the previous work [15] which applies typical vision method(e.g, LBP [26]) and attentional selective fusion branch, we consider our trained resizer plays the incorporated

Fig. 2: Example of the trainable resizer output. **First row :** the result of deterministic resizer(e.g, bilinear interpolation), the proposed trained resizer and the absolute difference between (a) and (b) from the left. **Second row:** More examples of the difference.

Table 4: Comparison with other state-of-the-art methods for In-the-wild FER task. ***** denotes accuracy trained with Swin-Large

Type	Method	FERPlus	RAF-DB
CNN	RAN [6]	89.16	86.90
	SCN [7]	89.35	88.14
	PSR [10]	89.75	88.98
Transformer	LBP + CVT [15]	88.81	88.14
	MVT [14]	88.88	87.03
	VIT + SE [13]	-	86.18
	ours	**89.50**	**88.57***

role in one shot. In addition, it can save significant computing and memory footprint while training, as we only need a single branch(NR) to train and a single image to feed to NR before Transformer while the previous work has to feed both original and LBP-applied images and needs to train dual branches.

3.3.3. The Effectiveness of Loss

In addition to compensating for low-quality data, our other goal is to mitigate the inherent problems of FER task - class imbalance and noisy labels. To demonstrate the effectiveness of F-PDLS to the problems, we conduct the experiments on Cross-Entropy and Prior-Distribution Label Smoothing [10] loss function as the comparative standards under the proposed setting **(e)** in Table.2. Interpreting the results in Table.3, while F-PDLS achieves the highest accuracy with a trivial increase of **0.06%**, there is no notable difference between the loss functions when used in vanilla Swin-T. However, the gain on Transformer equipped with NR increases to **0.63%**. To conclude with the empirical results, we hypothesize that our NR plays a role as a magnifier to F-PDLS which puts importance on minor and ambiguous samples. In other words, NR helps the Transformer to look closely at the fine-grained features which were ignored when the only image itself is fed to the Transformer directly.

3.3.4. Comparison With State-Of-The-Art Networks in FER

Table.4 shows our performance in two ITW datasets compared to other state-of-the-art approaches in the FER task including the recent transformer-based networks. To the best of our knowledge, our model outperforms all of the transformer-variants and achieves the second highest performance on FERPlus with **89.50%**, and **88.57%** on RAF-DB.

4. CONCLUSION

In this paper, we propose a novel training framework to leverage Transformer under the realistic FER task condition with the proposed trainable resizer and a loss function called Neural Resizer and F-PDLS respectively. Neural Resizer first restores the image to a higher resolution to leverage enhanced information in low-quality image data. Then, the enriched image is downscaled in a data-driven manner by the downsizer. By jointly optimizing the downsizer with Transformer with F-PDLS, which puts importance on minor as well as ambiguous classes, the proposed framework helps Transformer to adapt to the target task better. The experiments suggest that our framework, as well as loss function, have improved the performance of Transformer variants in general. Furthermore, we empirically show that Swin-Transformer achieves the competitive results compared to the strong baselines.

5. REFERENCES

[1] Shan Li and Weihong Deng, "Deep facial expression recognition: A survey," *IEEE Transactions on Affective Computing*, p. 1–1, 2020.

[2] Shan Li, Weihong Deng, and JunPing Du, "Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2852–2861.

[3] Hwanjun Song, Minseok Kim, Dongmin Park, Yooju Shin, and Jae-Gil Lee, "Learning from noisy labels with deep neural networks: A survey," *arXiv preprint arXiv:2007.08199*, 2020.

[4] Emad Barsoum, Cha Zhang, Cristian Canton Ferrer, and Zhengyou Zhang, "Training deep networks for facial expression recognition with crowd-sourced label distribution," in *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, 2016, pp. 279–283.

[5] Isabelle Hupont and Carles Fernández, "Demogpairs: Quantifying the impact of demographic imbalance in deep face recognition," in *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*. IEEE, 2019, pp. 1–7.

[6] Kai Wang, Xiaojiang Peng, Jianfei Yang, Debin Meng, and Yu Qiao, "Region attention networks for pose and occlusion robust facial expression recognition," 2019.

[7] Kai Wang, Xiaojiang Peng, Jianfei Yang, Shijian Lu, and Yu Qiao, "Suppressing uncertainties for large-scale facial expression recognition," 2020.

[8] Quan T Ngo and Seokhoon Yoon, "Facial expression recognition based on weighted-cluster loss and deep transfer learning using a highly imbalanced dataset," *Sensors*, vol. 20, no. 9, pp. 2639, 2020.

[9] Xiaobo Wang, Shuo Wang, Jun Wang, Hailin Shi, and Tao Mei, "Co-mining: Deep face recognition with noisy labels," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9358–9367.

[10] Thanh-Hung Vo, Guee-Sang Lee, Hyung-Jeong Yang, and Soo-Hyung Kim, "Pyramid with super resolution for in-the-wild facial expression recognition," *IEEE Access*, vol. 8, pp. 131988–132001, 2020.

[11] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[12] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma, "Learning imbalanced datasets with label-distribution-aware margin loss," *arXiv preprint arXiv:1906.07413*, 2019.

[13] Mouath Aouayeb, Wassim Hamidouche, Catherine Soladie, Kidiyo Kpalma, and Renaud Seguier, "Learning vision transformer with squeeze and excitation for facial expression recognition," 2021.

[14] Hanting Li, Mingzhe Sui, Feng Zhao, Zhengjun Zha, and Feng Wu, "Mvt: Mask vision transformer for facial expression recognition in the wild," 2021.

[15] Fuyan Ma, Bin Sun, and Shutao Li, "Robust facial expression recognition with convolutional visual transformers," 2021.

[16] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," 2021.

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

[18] Heewon Kim, Myungsub Choi, Bee Lim, and Kyoung Mu Lee, "Task-aware image downscaling," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 399–414.

[19] Hossein Talebi and Peyman Milanfar, "Learning to resize images for computer vision tasks," 2021.

[20] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou, "Training data-efficient image transformers distillation through attention," 2021.

[21] Stéphane d'Ascoli, Hugo Touvron, Matthew Leavitt, Ari Morcos, Giulio Biroli, and Levent Sagun, "Convit: Improving vision transformers with soft convolutional inductive biases," 2021.

[22] Alaaeldin El-Nouby, Hugo Touvron, Mathilde Caron, Piotr Bojanowski, Matthijs Douze, Armand Joulin, Ivan Laptev, Natalia Neverova, Gabriel Synnaeve, Jakob Verbeek, and Hervé Jegou, "Xcit: Cross-covariance image transformers," 2021.

[23] Cheng Ma, Zhenyu Jiang, Yongming Rao, Jiwen Lu, and Jie Zhou, "Deep face super-resolution with iterative collaboration between attentive recovery and landmark estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 5569–5578.

[24] Vuong Le, Jonathan Brandt, Zhe Lin, Lubomir Bourdev, and Thomas S Huang, "Interactive facial feature localization," in *European conference on computer vision*. Springer, 2012, pp. 679–692.

[25] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu, "Spatial transformer networks," 2016.

[26] Caifeng Shan, Shaogang Gong, and Peter W McOwan, "Robust facial expression recognition using local binary patterns," in *IEEE International Conference on Image Processing 2005*. IEEE, 2005, vol. 2, pp. II–370.