Embracing Context-Aware Emotion Recognition: A Feature Relevance-Based Strategy

Tran Minh Hai, Tran Nguyen Quynh Tram
Department of Information Technology
HCMC University of Foreign Language-
Information Technology,Vietnam
tranminhhai1506@gmail.com, tramtnq@huflit.edu.vn

Nguyen Quoc Huy
Department of Information Technology
Sai Gon University
Vietnam
nqhuy@sgu.edu.vn

Do Nhu Tai
Institute of Intelligent and Interactive Technologies
University of Economics Ho Chi Minh City-UEH
Vietnam
taidn@ueh.edu.vn

Kim Soo-Hyung
Department of Artificial Intelligence Convergence
Chonnam National University
South Korea
shkim@jnu.ac.kr

***Abstract*—In recent times, facial expression recognition has experienced significant progress, primarily attributed to the advancements in deep learning techniques focusing on extracting meaningful facial features. However, despite these remarkable developments, the practical application of these methods still faces challenges. Accurately predicting human emotions necessitates the consideration of multiple factors, including body language and contextual cues, rather than relying solely on facial expressions. In this paper, we propose an innovative approach that leverages the collective power of multiple models to extract comprehensive features from both facial and bodily expressions. Through this method, we achieved exceptional accuracy in predicting 26 different emotions using the EMOTIC dataset. Notably, our approach outperforms recent methods that solely concentrate on the facial and body features of individuals, underscoring the value of incorporating multi-modal information in emotion recognition.**

***Keywords*-Facial Emotions recognition; Emotion in Context; Deep learning; Multimodal**

I. Introduction

Facial expressions are crucial for human social communication, impacting various aspects of daily life [1]. They serve as significant non-verbal communication methods, influencing work success, personal relationships, and psychological well-being. Facial expressions, such as happiness, sadness, anger, surprise, disgust, and agreement, allow people to understand and interact with each other in social contexts [2]. In computer vision, extensive research aims to enable computers to comprehend human emotions, leading to seamless human-computer interactions [3]. However, facial recognition remains challenging due to variations in facial angles. Moreover, in modern society, face concealment or manipulation has become common.

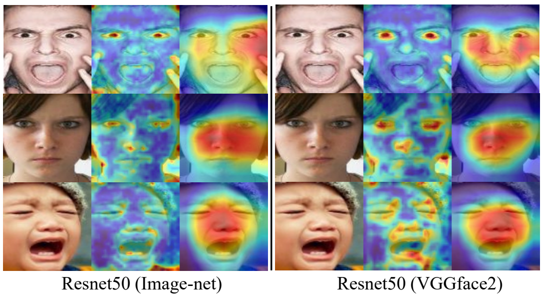

Fig. 1: GradCam visualization on two pre-train weights. When utilizing GradCam, the variance can be observed in the emphasis on facial features between the ResNet50 model trained with two different weights: ImageNet and VGGFace2.

Facial emotion recognition is essential for human-computer interaction. It utilizes diverse methods, including voice, facial expressions, and gestures. Factors like surrounding scenery and human psychology play a role in this recognition process [4]. Applications range from criminal psychology evaluation and driving safety improvement [5] to educational analysis. Automated systems for detecting facial expressions are also valuable for assessing client psychology in commerce [**?**].

However, previous studies on emotion recognition have often overlooked the crucial role of context. Numerous psychological publications [6] have highlighted the significance of considering the contextual factors that influence the experience of emotions. Not only facial expressions but also body language and the environment in which an event takes place can convey emotions. Consequently, relying solely on facial expressions has been proven inadequate for accurately predicting emotions in practical applications. In this study, we improve the effectiveness of contextual emotion recognition by utilizing components related to prognosticative outcomes and employing the proper pre-trained weights. For the primary object face feature extraction, we take the technique of reusing pre-trained weights that have been learned on two datasets, FER2013 [7] and RAF-DB [8] shown in Fig. 1. In order to complement the body part information in the following feature extraction model, we take the segment feature part from the main subject's body. 26 different varieties of real feelings are produced by conventionally combining the output attributes.

The main contributions are as follows: (1) train deep learning models on the FER2013 and RAF-DB datasets using Image-net [9] and VGGface2 [10] as well as two important pre-trained weights; (2) propose a network that combines multiple models, each of which extracts features from an image, such as the face, body, and body segmentation feature images.

II. Related Works

A. Emotion Recognition

The ICML 2013 FER2013 competition [7] serves as a benchmark for evaluating facial emotion recognition models. The CNN model has undergone significant modifications, resulting in accuracy ranging from 65% to 72.7%. Liu et al. [12] combined three CNN models to improve performance, but their best model achieved only 62.44% accuracy. Tang introduced a CNN model combined with a Linear Support Vector Machine (SVM) [13] and achieved 70.2% accuracy, outperforming other models and winning the competition. In recent years, networks with deeper layers, like VGGNet [14] and ResNet [15], have made breakthroughs and achieved results of 72.7% to 73.5% accuracy when trained on the FER2013 dataset.

B. Context Emotion Recognition

Emotion recognition in computer vision, including context, has gained attention recently. EMOTIC [11] is an image dataset trained on deep-learning approaches for recognizing emotions in context. The authors proposed combining various models to extract features from the body and context. However, facial features, which significantly impact emotion prediction, are disregarded in this approach. While features of the face can still be discovered when dividing the body for feature extraction, they may lack detail if the face region is small. As a result, this approach achieves only 28.33 accuracy in mean Average Precision.

The writers of the article [16] have used facial features to complement information for the main subject's body part by extracting the face as an additional extraction. Furthermore, to fully exploit the context's character, the authors provided a visual-based emotional state prediction method. Using spatial features, to discover the relationship between the main object and its neighbors in the context. However, it has not yet made excellent use of pre-trained weights on the face dataset for facial feature extraction. The study in the article [17] also took advantage of the facial features from the main subject's body, however, the authors did not use any pre-trained weights so the results only achieved mean Average Precision of 24.06.

Research on emotion recognition in context has focused on exploiting associated and facial elements. However, other impacts in the image will still affect the emotional impression of the main subject. That was also a driving factor in developing an idea for this contextual emotion recognition.

III. Proposed Method

A. Motivation

In this section, we propose a comprehensive approach for contextual emotion recognition shown in Fig. 2, addressing challenges faced in previous studies, including limited facial information and inadequate pre-trained weights. Our method utilizes essential face-to-body information to focus solely on emotions, avoiding the complexities of handling the entire context, which can affect emotion prediction. Furthermore, we introduce a body segmentation feature from the main subject's body image to ensure accurate extraction of body features. Our approach involves three models, each with suitable pre-trained weights, to process the main subject's

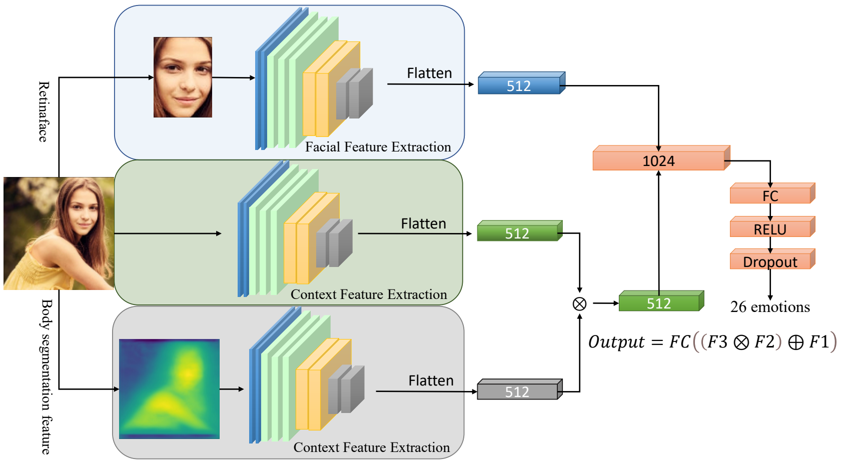
Retinaface
Flatten
512
Facial Feature Extraction
1024
FC
RELU
Dropout
26 emotions
Flatten
512
Context Feature Extraction
512
Body segmentation feature
Flatten
512
Context Feature Extraction
Output = FC((F3 ⊗ F2) ⊕ F1)

Fig. 2: Multiple models for contextual emotion recognition.

face image, body image, and segmented body image for emotion recognition shown in Fig. 2. The model's output for predicting emotions is presented in the formula below.

$$Output = FC\left((F3 \otimes F2) \oplus F1\right) \tag{1}$$

where FC is the Fully Connected layer, F1, F2, and F3 are the features from the models in turn from top to bottom, $\otimes$ is the element-wise multiplication and $\oplus$ is the concatenating operator

B. Facial Feature Extraction

To optimize the facial feature extraction process, we decided to train our model using two datasets: FER2013 [7] and RAF-DB [8]. We also reused the pre-trained weights from these datasets for the facial feature extraction on EMOTIC dataset and conducted a comparative analysis of their effectiveness. The models utilized for training and feature extraction included VGG11, VGG13 [14], ResNet34, and notably, ResNet50 [15] with pre-trained weights from VGGface2 [10].

C. Context Feature Extraction

In our approach, we incorporate two multi-models, to perform feature extraction tasks on the main subject. The model in the middle (Figure 2) focuses on extracting features related to the body of the subject, benefiting from the specialized weights from the Place365 dataset [18]. This choice enhances the model's ability to capture intricate details and contextual information associated with the subject's body.

On the other hand, the final model is specifically designed for segmenting the main subject and extracting segment features. It utilizes pre-trained weights from the ImageNet dataset [9], which provide a comprehensive understanding of various visual concepts, enabling the model to effectively delineate and extract meaningful features for the main subject's segmentation.

D. Data processing

1) Facial from Body: Within the EMOTIC dataset, the annotated bodies exhibit a wide range of diversity, including various objects or ongoing activities involving individuals. Since the face plays a pivotal role in predicting human emotions, we performed facial cropping from the annotated human bodies in the EMOTIC dataset shown in Fig. 3. For this purpose, we utilized RetinaFace [19], an advanced deep learning-based face detection tool for Python equipped with facial landmarks.

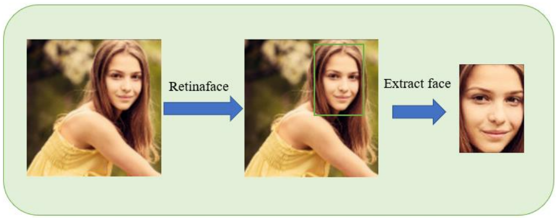

Fig. 3: Extracting facial images from the body.

2) Body segmentation feature from Body: As mentioned earlier, body images often contain surrounding details that can introduce noise into the process of extracting features from the main body object. Therefore, we propose generating images that solely focus on the body features by leveraging the strengths of the Unet network [20] for body segmentation. These segmented features will then be used to enhance the body part and improve overall effectiveness.

IV. Experiments and Results

A. *Datasets*

1) Facial Emotion Recognition Datasets: **The Facial Expression Recognition 2013 (FER2013)** [7] FER2013 is a significant dataset widely used for facial expression recognition. It is divided into three subsets: 28,709 images for training, 3,589 images for validation, and 3,589 images for testing.

The Real-world Affective Faces (RAF-DB) [8] dataset is a large-scale facial reaction database with approximately 30,000 facial images collected from the Internet. For our experiments, we utilized the basic set, which comprises 12,271 images for training and 3,068 images for testing.

2) EMOTIC Dataset: The EMOTIC dataset [11] is constructed by combining images from MSCOCO, Ade20K, and manually downloaded images from Google search. This amalgamation results in a challenging collection of images depicting individuals engaged in various activities, captured in different locations, and exhibiting diverse emotional states. The dataset comprises a total of 18,316 images, annotated with 23,788 individuals. There are 26 discrete emotions and 3 emotions along continuous dimensions. The training set consists of 70% of the images, the validation set contains 10% of the images, and the remaining 20% of the images constitute the test set.

B. *Experimental settings*

In this study, all input images, comprising facial images, body images, and segmented body feature maps, were uniformly resized to a standard dimension of 224x224x3. To enhance the model's generalization and mitigate overfitting, data augmentation techniques were employed during the training process. The augmentation methods included vertical flipping (vertical transformation) to introduce mirrored versions of the original images and random rotation within a range of [-25, 25] degrees to add diversity to the training data.

Loss function: Because this is a multi-class, multiple-label problem, it is considered a regression problem with a weighted Euclidean loss to compensate for the data set's existing class imbalance. The author of [11] also experimented and discovered that using Kullback-Leibler divergence or multi-class multi-classification hinge loss was more efficient. The following is the definition of the loss function:

$$Loss = \frac{1}{N}\sum_{i=1}^{N} w_i\left(\hat{y}_i - y_i\right) \tag{2}$$

where N is the number of emotional classes ($N = 26$ in this case), (y_i) is the estimated output for the $i-th$ class and y_i is the correct class for that emotion. Parameters w is the weight assigned to each class. The weight value is defined as $w_i = \frac{1}{ln(c+p_i)}$. In there, p_i is the probability that i and $c-th$ categories are parameters to control the valid range of values for w_i. All our methods are evaluated using Average Precision (AP).

Training: All models were trained within 50 epochs with RAdam, a learning rate of 0.0001, a weight decay of 0.0001, and a batch size of 48. During the training process, if there were 2 consecutive epochs without changing validation accuracy, the learning rate would be reduced by 10 times.

C. *Results and Discussion*

In our facial emotion recognition experiments, we employed pre-trained weights from ImageNet as the initial weights for the models. The main backbones used in the experiments were VGG11, VGG13, ResNet34, and ResNet50. Furthermore, for the models listed in Table I, we also incorporated additional pre-trained weights from VGG-Face2 specifically for the ResNet50 architecture.

Table I: Experimental results on FER datasets.

Model	Pre-trained	FER2013 (%)	RAF-DB (%)
Resnet34	Image-Net	72.80%	86.70%
Resnet50	Image-Net	73.40%	86.99%
VGG11	Image-Net	70.41%	85.23%
VGG13	Image-Net	70.66%	85.20%
Resnet50	VGGface2	74.30%	88.90%

Table I shows that ResNet50 model with VGGface2 pre-trained weights demonstrates higher accuracy on both the FER2013 and RAF-DB datasets compared to other models. For FER2013, accuracy ranges from 70% to 74%, while on RAF-DB,

Table II: Results on EMOTIC dataset[a].

	Model 1	**Model 2**	**Model 3**	**mAP**
1	Resnet50_I	Resnet34_I	_	27.0
2	Resnet50_V	Resnet34_I	_	27.8
3	Resnet50_I	Resnet34_I	Resnet18_I	26.6
4	Resnet50_V	Resnet34_I	Resnet18_I	27.7
5	Resnet50_I	Resnet18_P	Resnet18_I	25.6
6	Resnet50_V	Resnet18_P	Resnet18_I	26.7
7	VGG11_I	Resnet34_I	Resnet18_I	26.7
8	VGG13_I	Resnet34_I	Resnet18_I	26.8
9	VGG11_I^{**}	Resnet34_I	Resnet18_I	28.4
10	VGG13_I^{*}	Resnet34_I	Resnet18_I	28.8
11	Resnet50_V^{**}	Resnet18_P	_	27.5
12	Resnet50_V^{*}	Resnet18_P	_	29.3
13	Resnet50_I^{*}	Resnet18_P	Resnet34_I	28.5
14	Resnet50_I^{**}	Resnet18_P	Resnet34_I	27.3
15	Resnet50_V^{**}	Resnet18_P	Resnet34_I	29.2
16	Resnet50_V^{*}	Resnet18_P	Resnet34_I	30.4
17	Resnet50_V^{**}	Resnet50_P	Resnet34_I	29.8
18	**Resnet50_V^{*}**	**Resnet50_P**	**Resnet34_I**	**30.7**

[a] I: Image-net, V: VGGface2 and P: Place365. $*$: was trained on RAF-DB and $**$: was trained on FER2013. Model 1: Facial Feature Extraction, Model 2: Context Feature Extraction on Body, Model 3: Context Feature Extraction on Segment Body

Table III: Compare with state-of-the-art methods

Emotions Categories	Kositi [11]	Hoang [16]	Mittal [17]	**Ours**
Affection	27.85	37.07	29.87	30.89
Anger	09.49	18.67	08.52	22.81
Annoyance	14.06	20.74	09.65	22.12
Anticipation	58.64	57.96	46.23	57.94
Aversion	07.48	10.81	06.27	12.17
Confidence	78.35	76.81	51.92	80.00
Disapproval	14.97	19.65	11.81	19.30
Disconnection	21.32	30.16	31.74	27.87
Disquietment	16.89	19.48	07.57	22.39
Doubt/Confusion	29.63	21.76	21.62	20.37
Embarrassment	03.18	02.65	08.43	03.02
Engagement	87.53	87.47	78.68	86.20
Esteem	17.73	15.25	18.32	15.94
Excitement	77.16	72.49	73.19	70.13
Fatigue	09.70	16.38	06.34	12.7
Fear	14.14	05.95	14.29	07.03
Happiness	58.26	79.99	52.52	80.05
Pain	08.94	12.19	05.75	12.14
Peace	21.56	24.68	13.53	23.91
Pleasure	45.46	50.05	58.26	54.02
Sadness	19.66	30.46	19.94	33.88
Sensitivity	09.28	06.87	03.16	08.85
Suffering	18.84	31.18	15.38	36.38
Surprise	18.81	14.11	05.29	12.82
Sympathy	14.71	12.81	22.38	15.36
Yearning	08.34	09.57	04.94	09.26
mAP	27.38	30.20	24.06	**30.70**

it ranges from 85% to 86%. This highlights the model's effectiveness in facial emotion recognition tasks for both datasets. The models above will serve as feature extraction models for facial analysis in the EMOTIC dataset. Similarly, the following experiments will further demonstrate the efficacy of these models in the context of the EMOTIC dataset.

In our EMOTIC experiments, we performed an elimination experiment using a range of models with various pre-trained weights to assess their effectiveness. Table II displays the results of the methods evaluated based on mean Average Precision (mAP). Regarding the model for facial extraction, the utilization of VGGface2 weights in conjunction with the Resnet50 model proves to be more effective in facial feature extraction compared to using ImageNet weights, as evidenced by the results from experiment 1 and experiment 2. Subsequent experiments, 3 to 8, consistently yield mAP scores ranging between 25.5 and 27.5. These experiments highlight the varying degrees of contribution that pre-trained weights make to each of the three models, further emphasizing the significance of the choice of pre-trained weights in achieving optimal performance. **In experiments 9 to 18**, we employed models previously trained on the FER2013 and RAF-DB datasets as models to extract facial features. Notably, the use of the RAF-DB model's weights for feature extraction resulted in significantly improved performance, as observed prominently in experiments 11 and 12, where the model for segmentation's body feature was excluded from the training process.

As for the model for body extraction, utilizing Place 365 weights and ImageNet proved to be more effective. These experimental findings underscore the importance of selecting appropriate pre-trained weights for each model, as they significantly impact the overall performance in context feature extraction tasks. The best results were shown in the 18th experiment with a mAP of 30.7.

Presented in Table III are comprehensive AP calculation results for each method, with a particular emphasis on the precise computation and extraction of facial features and body parts of the main subject. Notably, our method stands out for its exceptional efficiency, boasting the highest mAP score among all approaches. Furthermore, we have achieved remarkable accuracy in successfully identifying a wide spectrum of emotions, encompassing Anger, Annoyance, Aversion, Confidence, Disquietment, Happiness, Sadness, and Suffering. These results underscore the effectiveness and robustness of our approach to emotion recognition, making it a promising and valuable contribution to the field.

V. Conclusion

In this paper, we have successfully leveraged pre-trained facial models for emotion classification and applied them to feature extraction in the EMOTIC dataset. We thoroughly utilized facial emotion information from models trained on the RAF-DB dataset. Additionally, we have achieved a breakthrough by effectively incorporating segment-level features, resulting in the highest performance to date when utilizing image body captions in the EMOTIC dataset. Although some emotion classes still exhibit lower results, overall, significant improvements have been made in the recognition of the remaining emotions. This underscores the effectiveness of my approach in enhancing emotion recognition and represents a notable advancement in the field.

ACKNOWLEDGEMENT

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) under the Artificial Intelligence Convergence Innovation Human Resources Development (IITP-2023-RS-2023-00256629) grant funded by the Korea government(MSIT). This research was funded by University of Economics Ho Chi Minh City, Vietnam.

REFERENCE

[1] L. E. Ishii, J. C. Nellis, K. D. Boahene, P. Byrne, and M. Ishii, “The Importance and Psychology of Facial Expression,” Otolaryngol. Clin. North Am., vol. 51, no. 6, pp. 1011–1017, Dec. 2018.

[2] L. M. Mayo, J. Lindé, H. Olausson, and M. Heilig, “Putting a good face on touch: Facial expression reflects the affective valence of caress-like touch across modalities,” Biol. Psychol., vol. 137, pp. 83–90, 2018.

[3] G. Tavares, A. Mourao, and J. Magalhaes, “Crowdsourcing facial expressions for affective-interaction,” Comput. Vis. Image Underst., vol. 147, pp. 102–113, 2016.

[4] H. Aviezer, Y. Trope, and A. Todorov, “Body cues, not facial expressions, discriminate between intense positive and negative emotions,” Science, vol. 338, no. 6111, pp. 1225–1229, 2012.

[5] J. Yang, T. Qian, F. Zhang, and S. U. Khan, “Real-time facial expression recognition based on edge computing,” IEEE Access, vol. 9, pp. 76178–76190, 2021.

[6] L. F. Barrett, B. Mesquita, and M. Gendron, “Context in emotion perception,” Curr. Dir. Psychol. Sci., vol. 20, no. 5, pp. 286–290, 2011.

[7] I. J. Goodfellow et al., “Challenges in representation learning: A report on three machine learning contests,” in Neural Information Processing: 20th International Conference, ICONIP 2013, Daegu, Korea, November 3-7, 2013. Proceedings, Part III 20, Springer, 2013, pp. 117–124.

[8] S. Li, W. Deng, and J. Du, “Reliable Crowdsourcing and Deep Locality-Preserving Learning for Expression Recognition in the Wild,” in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI: IEEE, Jul. 2017, pp. 2584–2593.

[9] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” Commun. ACM, vol. 60, no. 6, pp. 84–90, 2017.

[10] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, “Vggface2: A dataset for recognising faces across pose and age,” in 2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018), IEEE, 2018, pp. 67–74.

[11] R. Kosti, J. M. Alvarez, A. Recasens, and A. Lapedriza, “Context Based Emotion Recognition using EMOTIC Dataset,” IEEE Trans. Pattern Anal. Mach. Intell., pp. 1–1, 2019.

[12] K. Liu, M. Zhang, and Z. Pan, “Facial expression recognition with CNN ensemble,” in 2016 international conference on cyberworlds (CW), IEEE, 2016, pp. 163–166.

[13] Y. Tang, “Deep learning using linear support vector machines,” ArXiv Prepr. ArXiv13060239, 2013.

[14] Simonyan, K., and A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition,” 3rd International Conference on Learning Representations (ICLR 2015), Computational and Biological Learning Society, 2015, pp. 1–14.

[15] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778. Accessed: Apr. 03, 2023.

[16] M.-H. Hoang, S.-H. Kim, H.-J. Yang, and G.-S. Lee, “Context-aware emotion recognition based on visual relationship detection,” IEEE Access, vol. 9, pp. 90465–90474, 2021.

[17] T. Mittal, P. Guhan, U. Bhattacharya, R. Chandra, A. Bera, and D. Manocha, “EmotiCon: Context-Aware Multimodal Emotion Recognition using Frege’s Principle.” arXiv, Mar. 14, 2020. Accessed: Dec. 13, 2022.

[18] A. López-Cifuentes, M. Escudero-Vinolo, J. Bescós, and Á. García-Martín, “Semantic-aware scene recognition,” Pattern Recognit., vol. 102, p. 107256, 2020.

[19] S. I. Serengil and A. Ozpinar, “Lightface: A hybrid deep face recognition framework,” in 2020 Innovations in Intelligent Systems and Applications Conference (ASYU), IEEE, 2020, pp. 1–5.

[20] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18, Springer, 2015, pp. 234–241.