Region-Based Emotion Recognition via Superpixel Feature Pooling

Zhihang Ren[1], Yifan Wang[1], Tsung-Wei Ke[2], Yunhui Guo[3], Stella X. Yu[1,4], David Whitney[1]
[1]University of California, Berkeley, [2]Carnegie Mellon University,
[3]University of Texas at Dallas, [4]University of Michigan, Ann Arbor
[1]{peter.zhren,wyf020803,zhimin,dwhitney}@berkeley.edu,
[2]tsungwek@andrew.cmu.edu, [3]yunhui.guo@utdallas.edu,[4]stellayu@umich.edu

Abstract

Perceiving other people's emotional states is fundamentally important for successful social interactions and robotics. Traditional emotion recognition algorithms exclusively focus on facial expressions, ignoring the critical role of background context, which is now known to be necessary to accurately represent and understand the emotions of others. More recent studies have utilized different fusing techniques to combine facial and contextual information in visual scenes, but these approaches are limited to detection-based methods. In this study, we propose a new region-based emotion recognition method via superpixel feature pooling that does not rely on detection. Our proposed method consists of three types of blocks, including an initial over-segmentation block, the superpixel pooling block, and the emotion recognition block. On EMOTIC and VEATIC datasets, our proposed method improves state-of-the-art performance by 68.57% *and* 11.79% *respectively. We also achieve competitive performance on the CAER-S dataset.*

1. Introduction

Recognizing human emotions is routine and necessary to successfully navigate social interactions on a daily basis. Nowadays, as robotic techniques grow fast, it is natural to make future intelligent machines socially aware in the human-populated world. Moreover, emotion recognition may help the autonomous driving system to anticipate pedestrians' or drivers' intentions and react properly. Therefore, understanding emotion perception mechanisms and designing automatic emotion recognition methods are essential for future robotics and autonomous driving developments.

Over the past several years, the interest in utilizing deep learning models to automatically recognize emotional states has grown rapidly. Following a long tradition of research on emotion recognition in the fields of psychology, neuroscience, and vision science, previous computer vision research focused almost exclusively on facial expressions as the key information for emotion recognition. This is unsurprising, as facial expressions seem to be the most direct and inherent way for humans to understand the emotions of others. Consequently, many early datasets annotated only the emotional states of character faces or lab-controlled human interactions, treating them as if they are independent of the context [5, 11, 14, 24–26, 39, 40, 44, 46, 53, 56, 57, 66]. With this massive annotated data involving facial expressions, researchers primarily focused on the analysis of facial expression to predict emotions [6, 62–64]. Later studies also found extra information, such as shoulder location and body pose, could be utilized to infer emotional states [43, 54]. Overall, though, early datasets and recognition models focus strictly on character-specific information to infer emotional states.

Although the character itself—including the facial expression—contains a great deal of information about its emotional state, many studies in Psychology have proven that context information is critically important for accurate emotion perception [3, 4, 10]. In many scenarios, scene context influences the perception of human emotion even though the facial expression is unchanged or very similar [1, 37, 52]. And, scene context can explain as much of the variance in human emotion perception as facial expression [9]. For these reasons, annotations of isolated faces may not accurately reflect true human emotion, and context-based emotion datasets are necessary.

In light of the importance of context in emotion recognition, several datasets that include contextual information have emerged recently [23, 28, 29, 51]. Those datasets not only contain characters but also large areas of the surrounding context. In turn, recent algorithms [17, 21, 30, 31, 41, 45, 65, 68] then focus on different feature extraction methods and fusing techniques for various types of information and visual features.

To locate different types of visual information, such as the face, body, and background scene, previous methods

rely heavily on object detection methods [49]. They often utilize rectangular bounding boxes to select or mask out certain image blocks (Figure 1(b)). Then, visual feature encoding and fusing modules are utilized to represent the visual information of the whole scene.

When humans try to perceive the emotions of other people, they rely on bottom-up and top-down visual processes, where fine-scale visual features and coarse-scale object and scene knowledge mutually facilitate each other [7, 38]. In this process, there are no bounding box structures; instead, there are fine or coarse scales of object and scene regions (superpixels). With this insight, researchers have started to utilize superpixels in a variety of computer vision models [2, 15, 16, 18, 22, 42, 48, 50, 55, 61, 69].

So far, superpixel-based models have been successful in understanding and grouping semantically similar image regions, achieving good performance on part parsing, saliency detection, and image segmentation tasks. As emotion recognition requires the understanding of characters' facial expressions, as well as their interactions with different regions of objects and background scenes, the superpixel approach could be useful for emotion recognition. But, to the best of our knowledge, no emotion recognition method has utilized superpixel-based methods.

In this paper, we propose a new emotion recognition method that utilizes superpixel-level visual features. Our proposed method consists of three types of blocks: 1) an over-segmentation block to initialize the fine-grained segments and generate initial superpixel features; 2) the superpixel pooling block to learn the grouping policy and aggregate the current level finer-scale features to the next level coarser-scale features; and 3) the emotion recognition block for either emotion regression or classification tasks based on the final aggregated feature. We test our proposed method on three public context-aware emotion recognition datasets, EMOTIC [28], CAER [29], and VEATIC [51]. We achieve state-of-the-art performance on EMOTIC [28] and VEATIC [51], with 68.57% improvement on EMOTIC and 11.79% improvement on VEATIC. We also achieve competitive results on CAER-S [29]. Moreover, we show that by using superpixels as feature extraction anchors, we can naturally obtain semantically similar superpixels for free with the learned grouping policy.

In summary, our contribution of this work lies in three aspects:

1. We propose the first region-based emotion recognition method via superpixel feature pooling.
2. We achieve state-of-the-art emotion recognition performance on VEATIC and EMOTIC datasets with 68.57% improvement on EMOTIC and 11.79% improvement on VEATIC.
3. We show that the method can also provide us with clusters that contain semantically similar superpixels via the learned grouping policy.

2. Related Work

2.1. Context-Aware Emotion Recognition

When inferring emotion states, the context-aware emotion recognition methods do not only rely solely on the face or body information but also consider the emotion cues from scene context and background information. Traditional context-aware emotion recognition methods invariably extract multiple representations from various visual information sources and then apply feature fusion to make the final prediction [17, 21, 30, 31, 41, 45, 65, 68]. Object detection methods are widely utilized to identify the information sources, marking them with rectangular bounding boxes. For example, the model released along with the EMOTIC dataset [28] fused the body region feature and the whole image as the context feature via a Convolutional Neural Network (CNN). In this study, we utilize superpixel as the feature anchor for subsequent feature fusing, which does not rely on object detection or bounding boxes.

2.2. Vision Transformers

Vision Transformers (ViT) [13] have achieved amazing performance in image recognition. They treat images as sets of rectangular patch tokens and employ an attention mechanism in learning [60]. ViTs can be computationally expensive. To improve their efficiency, hierarchical transformers aim to reduce the number of tokens by spatial pooling [12, 20, 32, 34]. Other approaches directly prune tokens away according to their significance scores [8, 19, 35, 47, 67]. Our grouping procedure looks like the latter approach. However, we focus on grouping different visual regions for emotion recognition while those methods aim for efficiency. Additionally, we use superpixels as input units instead of square patches.

2.3. Superpixels

Superpixels are sets of locally connected pixels that encapsulate coherent structures, such as colors [48]. Intuitively, superpixels have been utilized in various computer vision tasks that involve dense labeling, including part parsing [42], saliency detection [50], and image segmentation [15, 16, 18, 50, 55, 61]. Recent studies have replaced patches with superpixel tokens in ViT architectures to achieve semantic segmentation [22, 69]. In this study, we adopt superpixels as the visual feature extraction anchors for feature fusing in emotion recognition tasks.

3. Method

Inspired by the human emotion recognition process, we propose the first region-based emotion recognition method via

Figure 1. **Comparison of detection-based methods and our proposed region-based method.** Detection-based methods rely on the bounding box to encode the character and context visual information separately, while our proposed region-based method directly utilizes initial pixel-level features and gradually aggregates similar superpixel features to represent the visual information.

superpixel feature pooling. Our idea revolves around utilizing superpixels to enhance our understanding of characters' facial expressions, along with their interactions within various regions of objects and background scenes. Figure 2 illustrates an overview of the method. The image/frame is at first over-segmented to obtain the fine-grained segments. At each level, the finer-scale superpixels are grouped into coarser-scale superpixels via the superpixel pooling block. The corresponding superpixel features are aggregated according to the learned pooling policy, processed by the visual transformer block to learn better features, and then sent to the next level pooling and aggregation. At last, the emotion recognition block can take the final aggregated feature to complete the emotion recognition task. Now, we introduce each block respectively.

3.1. Over-segmentation Block

Each time, we start with the finest-level pixel grouping, denoted as G_0, i.e., the initial image region grouping. These groupings are based on low-level visual cues and designed to align with image contours. In this paper, we utilize SEEDS [59] to obtain the locally connected and color-wise coherent regions, i.e., the superpixels. Then, we progressively group these superpixels into coarser regions and fuse the corresponding superpixel features to get the final aggregated feature for emotion recognition.

The initial pixel features of the input image, X_{cnn}, are obtained via a convolutional neural network (CNN). These pixel features are then aggregated within each superpixel in G_0 to create the initial superpixel features, referred to as X_G. The aggregation is achieved by averaging each pixel feature within a specific superpixel. After this, we append a class token X_{class}, and positional encodings E_{pos} into the initial features X_G. We set E_{pos} to align with the resolution of the CNN features X_{cnn} and then average it within each superpixel. The resulting input segment features are defined as $Z_0 = [X_{class}; X_G] + [\mathbf{0}; E_{pos}]$

3.2. Superpixel Pooling Block

To form better features for the aggregated superpixel features at each grouping level, such that semantically similar superpixel features at each grouping level would be more similar in the feature space and vice versa, we apply two visual transformer (ViT) blocks before the superpixel pooling. Then, we pool the similar fine-scale regions into a coarser scale and move to the next level. Various pooling strategies can be applied here. In this paper, we adopt a graph pooling strategy [22]. The similarities between different superpixel features in neighbor levels are computed and utilized to quantify the soft assignment probability P_l from a finer level $l-1$ to a coarser level l. Then, the next level of coarser groupings G_l can be determined by the finer level grouping G_{l-1} and the soft assignment probability P_l.

$$G_l = G_{l-1} \times P_l = G_0 \prod_{i=1}^{l} P_i \tag{1}$$

3.3. Emotion Recognition Block

At the final stage, we will have an aggregated feature that contains the combined visual information from separate visual regions. Then, we utilize a multilayer perceptron (MLP) to achieve the emotion recognition task. At last, we utilize either categorical emotion states or continuous emotion ratings to guide the training. We emphasize that no segmentation maps are utilized in the training. The grouping of superpixels is only trained to make good emotion recognition results, though the method naturally learned how to group superpixels efficiently.

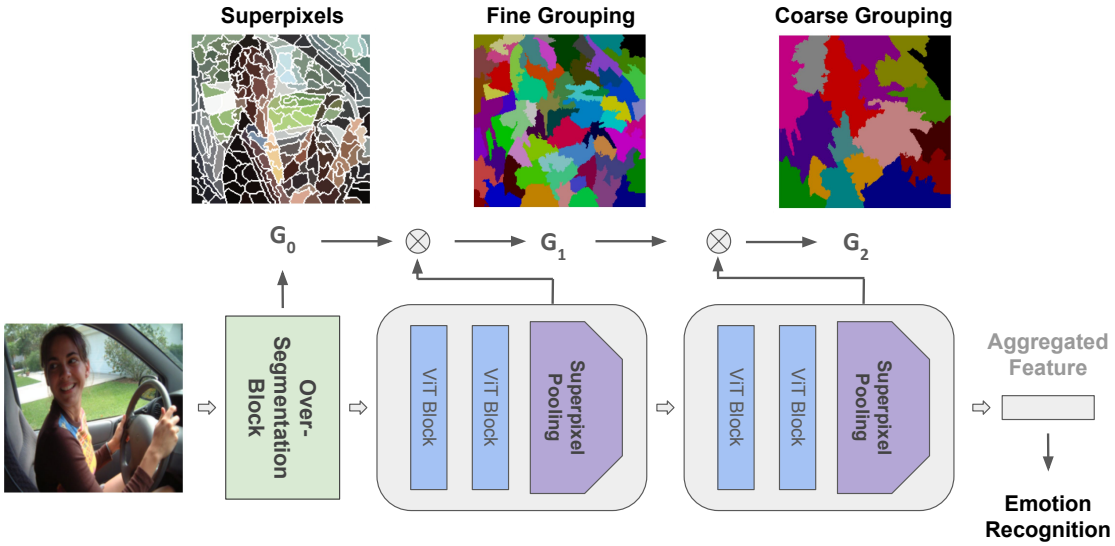

Figure 2. **Overview of the method:** In our proposed method, We start with over-segmented regions, i.e., the superpixels, and then gradually group similar superpixels, and aggregate features of corresponding superpixels. We utilize the final aggregated feature for emotion recognition. Along with the training, we also obtain clusters that contain semantically similar pixels, e.g., the red face region in G_2.

4. Experiment

4.1. Datasets and Evaluation Metrics

Datasets: We conduct our experiments on three standard datasets for the context-aware emotion recognition task, namely EMOTIC [28], CAER-S [29], and VEATIC [51]. EMOTIC contains 23,571 images of 34,320 annotated subjects in uncontrolled environments. The annotation of these images contains the bounding boxes of the target subjects' body regions and 26 discrete emotion categories. CAER-S includes 70k static images extracted from video clips of 79 TV shows to predict emotional states. These images are annotated with 7 emotion categories: Anger, Disgust, Fear, Happy, Sad, Surprise, and Neutral. VEATIC has 124 video clips from Hollywood movies, documentaries, and home videos with continuous valence and arousal ratings of each frame via real-time annotation.

Evaluation Metrics: Following [27, 41, 65], we utilize the standard classification accuracy to evaluate performance on CAER-S. For VEATIC, the root mean square error (RMSE) is used. At last, we utilize the mean Average Precision (mAP) to evaluate the classification results on the EMOTIC.

4.2. State-of-the-art Methods

Given the fact that our model is tested on three datasets, we select several models with different structures tested on each of the corresponding datasets for comparison. For the EMOTIC dataset, we select seven distinct models for comparison. EMOT-Net [27] is a two-branch Convolutional Neural Network model, whose unique branches capture the body features and context features separately. GCN-CNN [68] is a Graph Convolutional Network trying to infer emotion relationships utilizing the affective graph constructed by context elements. CAER-Net [29] is a double-stream Convolutional Neural Network model with an adaptive fusion module focusing on inferring emotion by integrating context information with facial information. RRLA [30] proposed the Body-Object Attention module and Body Part Attention module to estimate the importance of body parts and background information. VRD [21] utilizes both the spatial and semantic features by attention mechanism to learn the impact of each part on emotion recognition. EmotiCon [41] takes advantage of visual attention and depth maps to obtain multi-modal information. And CCIM [65] utilizes causal inference for model training. For CAER-S Dataset, we have added two more models for comparison in addition to the ones mentioned above. SIB-Net [31] is inspired by the study of context-containing order, interaction, and bias relationships. GRERN [17] proposes a framework based on a Graph Convolutional Network to do emotion classification utilizing the region-wise semantic relationships. For VEATIC Dataset, We compare our model to VEATIC-NET [51], which is a two-stream Video Transformer using the attention mechanism to learn

Category	EMOT-NET	GCN-CNN	CAER-NET	RRLA	VRD	EmotiCon	CCIM	Ours
Affection	26.47	47.52	22.36	37.93	44.48	38.55	40.77	**64.22**
Anger	11.24	11.27	12.88	13.73	30.71	14.69	15.48	**65.62**
Annoyance	15.26	12.33	14.42	20.87	26.47	24.68	24.47	**65.81**
Anticipation	57.31	63.2	52.85	61.08	59.89	60.73	95.15	**70.81**
Aversion	7.44	6.81	3.26	9.61	12.43	11.33	19.38	**71.67**
Confidence	**80.33**	74.83	72.68	80.08	79.24	68.12	75.81	60.79
Disapproval	16.14	12.64	15.37	21.54	24.54	18.55	23.65	**65.50**
Disconnection	20.64	23.17	22.01	28.32	34.24	28.73	31.93	**70.84**
Disquietment	19.57	17.66	10.84	22.57	24.23	22.14	26.84	**66.76**
Doubt/Confusion	31.88	19.67	26.07	33.5	25.42	38.43	34.28	**59.45**
Embarrassment	3.05	1.58	1.88	4.16	4.26	10.31	16.73	**60.59**
Engagement	86.69	87.31	73.71	88.12	88.71	86.23	**97.41**	61.17
Esteem	17.86	12.05	15.38	20.5	17.99	25.75	27.44	**63.51**
Excitement	78.05	72.68	70.42	80.11	74.21	80.75	**81.59**	70.22
Fatigue	8.87	12.93	6.29	17.51	22.62	19.35	15.53	**78.61**
Fear	15.7	6.15	7.47	15.56	13.92	16.99	15.37	**66.46**
Happiness	58.92	72.9	53.73	76.01	83.02	80.45	**83.55**	67.35
Pain	9.46	8.22	8.16	14.56	16.68	14.68	17.76	**68.99**
Peace	22.35	30.68	19.55	26.76	28.91	35.72	38.94	**71.30**
Pleasure	46.72	48.37	34.12	55.64	55.47	**67.31**	64.57	63.93
Sadness	18.69	23.9	17.75	30.8	42.87	40.26	45.63	**57.78**
Sensitivity	9.05	4.74	6.94	9.59	15.89	13.94	17.04	**66.48**
Suffering	17.67	23.71	14.85	30.7	46.23	48.05	21.52	**65.32**
Surprise	22.38	8.44	17.46	17.92	16.27	19.6	26.81	**59.14**
Sympathy	15.23	19.45	14.89	15.26	15.37	16.74	47.6	**65.01**
Yearning	9.22	9.86	4.84	10.11	10.04	15.08	12.25	**67.76**
mAP	27.93	28.16	23.85	32.41	35.16	35.28	39.13	**65.96**

Table 1. Average precision (%) of seven recent methods, and our proposed method for each emotion category on the EMOTIC dataset [28]. Overall, our proposed method improves state-of-the-art performance by 68.57%.

Methods	CAER-NET	EMOT-NET	SIB-Net	GCN-CNN	GRERN	RRLA	EmotiCon	VRD	Ours
Accuracy(%)	73.47	74.51	74.56	77.21	81.31	84.82	88.65	90.49	76.54

Table 2. Emotion classification accuracy (%) of eight recent methods, and our proposed method on the CAER-S dataset [29]. Our proposed method performs competitively with recent methods.

Method	RMSE↓		
	Valence	Arousal	Overall
VEATIC-NET	0.3084	0.2410	0.2747
Ours	**0.2577**	**0.2268**	**0.2423**

Table 3. Comparison of our proposed method with the baseline model proposed in VEATIC [51]. Our method outperforms 11.79% compared to the baseline method.

the contextual relationships between frames. We reproduce the results on the corresponding datasets based on the details given by the models above.

4.3. Implementation Details

We conducted supervised training following the setup of DeiT [58]. The model is trained on 4 NVIDIA GeForce RTX 2080 Ti GPUs. For the hyperparameters, we have 4 levels in total for the superpixel pooling. There are 64, 32, 16, and 8 clusters respectively at each grouping level and the batch size of the data is 64. For the superpixel inputs, we utilize 196 pixels as default. We resize the input images into 224×224 and apply normalization to the images. Our model is trained using the AdamW optimizer [33]. For the learning rate schedule, we use a linear warmup of 5 epochs to reach a peak learning rate of 5.0×10^{-4} from 1.0×10^{-6}, followed by a cosine decay of 30 epochs to decay the final learning rate to minimum learning rate of 1.0×10^{-5}.

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

4.4.1 Results on the EMOTIC Dataset.

In Table 1, we see that our proposed method significantly improves the recognition precision of most emotion categories. In particular, compared to EMOT-NET [27], GCN-CNN [68], CAER-NET [29], RRLA [30], VRD [21], EmotiCon [41], and CCIM [65], our proposed method improve the mAP scores by 136.16%, 134.23%, 176.56%, 103.52%, 87.60%, 86.96%, and 68.57% respectively. For

Figure 3. **Visualization of Grouping:** Column (a) raw images; (b) over-segments; (c) finer groupings; (d) coarser groupings; (e) overlaying coarser groupings onto the raw images. Surprisingly, without the supervision of segmentation maps, the proposed method learned the grouping policy for superpixels guided by visual emotion recognition training.

certain emotion categories, the emotion recognition average precision is even improved drastically compared to SOTA performances, such as Yearning (+349.34%), Sensitivity (+290.14%), Pain (+288.46%), Anger (+269.81%), Embarrassment (+262.16%), and Fatigue (+247.52%).

4.4.2 Results on the CAER-S Dataset.

Table 2 shows our proposed method performs competitively with recently released emotion recognition methods. It is worth noting that the CAER dataset utilized very few annotators (six) and has little control over the annotation quality

compared to EMOTIC and VEATIC datasets. As annotation uncertainty and bias may result from the insufficiency of annotators [36, 51], this may influence the interpretation of any model's performance to some extent.

4.4.3 Results on the VEATIC Dataset.

We also test our proposed method on a recently released dataset, VEATIC [51]. Our proposed method improves the RMSE of the overall rating by 11.79%. In terms of valence and arousal, our method improves 16.44% and 5.89% respectively.

4.5. Grouping Visualization

Although our supervised training process does not utilize segmentation maps as guidance, through the learned grouping policy, we show that the semantically similar finer-scale superpixels are pooled to form coarser-scale regions at the next level. As in Figure 3(e), we can find groupings of facial regions and the object/scene regions which the character is interacting with.

Compared to traditional detection-based methods, where the emotion recognition module passively encodes the visual information selected by bounding boxes, our proposed method proactively learns which superpixels to group and aggregate. It is clear to see which regions contribute similarly to the final emotion prediction. Thus, by utilizing superpixels to enhance our understanding of characters' facial expressions, we can achieve more accurate emotion recognition.

5. Conclusion

In this paper, we proposed the first region-based emotion recognition method via superpixel feature pooling. It achieves state-of-the-art emotion recognition performance on VEATIC and EMOTIC datasets. It also achieves competitive results on CAER-S dataset. Moreover, the proposed method can also provide us with clusters that contain semantically similar superpixels via the learned grouping policy.

References

[1] Hillel Aviezer, Ran R Hassin, Jennifer Ryan, Cheryl Grady, Josh Susskind, Adam Anderson, Morris Moscovitch, and Shlomo Bentin. Angry, disgusted, or afraid? studies on the malleability of emotion perception. *Psychological science*, 19(7):724–732, 2008. 1

[2] Moshe Bar. Visual objects in context. *Nature Reviews Neuroscience*, 5(8):617–629, 2004. 2

[3] Lisa Feldman Barrett and Elizabeth A Kensinger. Context is routinely encoded during emotion perception. *Psychological science*, 21(4):595–599, 2010. 1

[4] Lisa Feldman Barrett, Batja Mesquita, and Maria Gendron. Context in emotion perception. *Current directions in psychological science*, 20(5):286–290, 2011. 1

[5] Pablo Barros, Nikhil Churamani, Egor Lakomkin, Henrique Siqueira, Alexander Sutherland, and Stefan Wermter. The omg-emotion behavior dataset. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE, 2018. 1

[6] Uttaran Bhattacharya, Trisha Mittal, Rohan Chandra, Tanmay Randhavane, Aniket Bera, and Dinesh Manocha. Step: Spatial temporal graph convolutional networks for emotion perception from gaits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1342–1350, 2020. 1

[7] Irving Biederman. Recognition-by-components: a theory of human image understanding. *Psychological review*, 94(2): 115, 1987. 2

[8] Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your vit but faster. In *The Eleventh International Conference on Learning Representations*, 2022. 2

[9] Zhimin Chen and David Whitney. Tracking the affective state of unseen persons. *Proceedings of the National Academy of Sciences*, 116(15):7559–7564, 2019. 1

[10] Zhimin Chen and David Whitney. Inferential emotion tracking (iet) reveals the critical role of context in emotion recognition. *Emotion*, 22(6):1185, 2022. 1

[11] Abhinav Dhall, Roland Goecke, Simon Lucey, Tom Gedeon, et al. Collecting large, richly annotated facial-expression databases from movies. *IEEE multimedia*, 19(3):34, 2012. 1

[12] Xiaoyi Dong, Jianmin Bao, Dongdong Chen, Weiming Zhang, Nenghai Yu, Lu Yuan, Dong Chen, and Baining Guo. Cswin transformer: A general vision transformer backbone with cross-shaped windows. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12124–12134, 2022. 2

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

[14] Ellen Douglas-Cowie, Roddy Cowie, Cate Cox, Noam Amir, and Dirk Heylen. The sensitive artificial listener: an induction technique for generating emotionally coloured conversation. In *LREC workshop on corpora for research on emotion and affect*, pages 1–4. ELRA Paris, 2008. 1

[15] Brian Fulkerson, Andrea Vedaldi, and Stefano Soatto. Class segmentation and object localization with superpixel neighborhoods. In *2009 IEEE 12th international conference on computer vision*, pages 670–677. IEEE, 2009. 2

[16] Raghudeep Gadde, Varun Jampani, Martin Kiefel, Daniel Kappler, and Peter V Gehler. Superpixel convolutional networks using bilateral inceptions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 597–613. Springer, 2016. 2

[17] Qinquan Gao, Hanxin Zeng, Gen Li, and Tong Tong. Graph reasoning-based emotion recognition network. *IEEE Access*, 9:6488–6497, 2021. 1, 2, 4

[18] Stephen Gould, Jim Rodgers, David Cohen, Gal Elidan, and Daphne Koller. Multi-class segmentation with relative location prior. *International journal of computer vision*, 80: 300–316, 2008. 2

[19] Saurabh Goyal, Anamitra Roy Choudhury, Saurabh Raje, Venkatesan Chakaravarthy, Yogish Sabharwal, and Ashish Verma. Power-bert: Accelerating bert inference via progressive word-vector elimination. In *International Conference on Machine Learning*, pages 3690–3699. PMLR, 2020. 2

[20] Byeongho Heo, Sangdoo Yun, Dongyoon Han, Sanghyuk Chun, Junsuk Choe, and Seong Joon Oh. Rethinking spatial dimensions of vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11936–11945, 2021. 2

[21] Manh-Hung Hoang, Soo-Hyung Kim, Hyung-Jeong Yang, and Guee-Sang Lee. Context-aware emotion recognition based on visual relationship detection. *IEEE Access*, 9: 90465–90474, 2021. 1, 2, 4, 5

[22] Tsung-Wei Ke, Sangwoo Mo, and X Yu Stella. Learning hierarchical image segmentation for recognition and by recognition. In *The Twelfth International Conference on Learning Representations*, 2024. 2, 3

[23] Sander Koelstra, Christian Muhl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. Deap: A database for emotion analysis; using physiological signals. *IEEE transactions on affective computing*, 3(1):18–31, 2011. 1

[24] Dimitrios Kollias and Stefanos Zafeiriou. Aff-wild2: Extending the aff-wild database for affect recognition. *arXiv preprint arXiv:1811.07770*, 2018. 1

[25] Dimitrios Kollias and Stefanos Zafeiriou. Expression, affect, action unit recognition: Aff-wild2, multi-task learning and arcface. *arXiv preprint arXiv:1910.04855*, 2019.

[26] Jean Kossaifi, Georgios Tzimiropoulos, Sinisa Todorovic, and Maja Pantic. Afew-va database for valence and arousal estimation in-the-wild. *Image and Vision Computing*, 65:23–36, 2017. 1

[27] Ronak Kosti, Jose M Alvarez, Adria Recasens, and Agata Lapedriza. Emotion recognition in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1667–1675, 2017. 4, 5

[28] Ronak Kosti, Jose M Alvarez, Adria Recasens, and Agata Lapedriza. Context based emotion recognition using emotic dataset. *IEEE transactions on pattern analysis and machine intelligence*, 42(11):2755–2766, 2019. 1, 2, 4, 5

[29] Jiyoung Lee, Seungryong Kim, Sunok Kim, Jungin Park, and Kwanghoon Sohn. Context-aware emotion recognition networks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10143–10152, 2019. 1, 2, 4, 5

[30] Weixin Li, Xuan Dong, and Yunhong Wang. Human emotion recognition with relational region-level analysis. *IEEE Transactions on Affective Computing*, 2021. 1, 2, 4, 5

[31] Xinpeng Li, Xiaojiang Peng, and Changxing Ding. Sequential interactive biased network for context-aware emotion recognition. In *2021 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–6. IEEE, 2021. 1, 2, 4

[32] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 2

[33] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5

[34] Xu Ma, Yuqian Zhou, Huan Wang, Can Qin, Bin Sun, Chang Liu, and Yun Fu. Image as set of points. In *The Eleventh International Conference on Learning Representations*, 2022. 2

[35] Dmitrii Marin, Jen-Hao Rick Chang, Anurag Ranjan, Anish Prabhu, Mohammad Rastegari, and Oncel Tuzel. Token pooling in vision transformers. *arXiv preprint arXiv:2110.03860*, 2021. 2

[36] Carlos A Martínez-Miwa and Mario Castelán. On reliability of annotations in contextual emotion imagery. *Scientific Data*, 10(1):538, 2023. 7

[37] Takahiko Masuda, Phoebe C Ellsworth, Batja Mesquita, Janxin Leu, Shigehito Tanida, and Ellen Van de Veerdonk. Placing the face in context: cultural differences in the perception of facial emotion. *Journal of personality and social psychology*, 94(3):365, 2008. 1

[38] Daphne Maurer, Richard Le Grand, and Catherine J Mondloch. The many faces of configural processing. *Trends in cognitive sciences*, 6(6):255–260, 2002. 2

[39] Daniel McDuff, Rana Kaliouby, Thibaud Senechal, May Amr, Jeffrey Cohn, and Rosalind Picard. Affectiva-mit facial expression dataset (am-fed): Naturalistic and spontaneous facial expressions collected. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 881–888, 2013. 1

[40] Gary McKeown, Michel Valstar, Roddy Cowie, Maja Pantic, and Marc Schroder. The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE transactions on affective computing*, 3(1):5–17, 2011. 1

[41] Trisha Mittal, Pooja Guhan, Uttaran Bhattacharya, Rohan Chandra, Aniket Bera, and Dinesh Manocha. Emoticon: Context-aware multimodal emotion recognition using frege's principle. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14234–14243, 2020. 1, 2, 4, 5

[42] Greg Mori, Xiaofeng Ren, Alexei A Efros, and Jitendra Malik. Recovering human body configurations: Combining segmentation and recognition. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, pages II–II. IEEE, 2004. 2

[43] Mihalis A Nicolaou, Hatice Gunes, and Maja Pantic. Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space. *IEEE Transactions on Affective Computing*, 2(2):92–105, 2011. 1

[44] Desmond C Ong, Zhengxuan Wu, Zhi-Xuan Tan, Marianne Reddan, Isabella Kahhale, Alison Mattek, and Jamil Zaki.

Modeling emotion in complex stories: the stanford emotional narratives dataset. *IEEE Transactions on Affective Computing*, 12(3):579–594, 2019. 1

[45] Ioannis Pikoulis, Panagiotis P Filntisis, and Petros Maragos. Leveraging semantic scene characteristics and multistream convolutional architectures in a contextual approach for video-based visual emotion recognition in the wild. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pages 01–08. IEEE, 2021. 1, 2

[46] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. Meld: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv preprint arXiv:1810.02508*, 2018. 1

[47] Yongming Rao, Wenliang Zhao, Benlin Liu, Jiwen Lu, Jie Zhou, and Cho-Jui Hsieh. Dynamicvit: Efficient vision transformers with dynamic token sparsification. *Advances in neural information processing systems*, 34:13937–13949, 2021. 2

[48] Ren and Malik. Learning a classification model for segmentation. In *Proceedings ninth IEEE international conference on computer vision*, pages 10–17. IEEE, 2003. 2

[49] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 2

[50] Zhixiang Ren, Shenghua Gao, Liang-Tien Chia, and Ivor Wai-Hung Tsang. Region-based saliency detection and its application in object recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 24(5):769–779, 2013. 2

[51] Zhihang Ren, Jefferson Ortega, Yifan Wang, Zhimin Chen, Yunhui Guo, Stella X Yu, and David Whitney. Veatic: Video-based emotion and affect tracking in context dataset. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4467–4477, 2024. 1, 2, 4, 5, 7

[52] Ruthger Righart and Beatrice De Gelder. Rapid influence of emotional scenes on encoding of facial expressions: an erp study. *Social cognitive and affective neuroscience*, 3(3): 270–278, 2008. 1

[53] Fabien Ringeval, Andreas Sonderegger, Juergen Sauer, and Denis Lalanne. Introducing the recola multimodal corpus of remote collaborative and affective interactions. In *2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG)*, pages 1–8. IEEE, 2013. 1

[54] Konrad Schindler, Luc Van Gool, and Beatrice De Gelder. Recognizing emotions expressed by body pose: A biologically inspired neural model. *Neural networks*, 21(9):1238–1246, 2008. 1

[55] Abhishek Sharma, Oncel Tuzel, and Ming-Yu Liu. Recursive context propagation network for semantic scene labeling. *Advances in Neural Information Processing Systems*, 27, 2014. 2

[56] Ian Sneddon, Margaret McRorie, Gary McKeown, and Jennifer Hanratty. The belfast induced natural emotion database. *IEEE Transactions on Affective Computing*, 3(1):32–41, 2011. 1

[57] Mohammad Soleymani, Jeroen Lichtenauer, Thierry Pun, and Maja Pantic. A multimodal database for affect recognition and implicit tagging. *IEEE transactions on affective computing*, 3(1):42–55, 2011. 1

[58] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR, 2021. 5

[59] Michael Van den Bergh, Xavier Boix, Gemma Roig, Benjamin De Capitani, and Luc Van Gool. Seeds: Superpixels extracted via energy-driven sampling. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part VII 12*, pages 13–26. Springer, 2012. 3

[60] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2

[61] Xing Wei, Qingxiong Yang, Yihong Gong, Narendra Ahuja, and Ming-Hsuan Yang. Superpixel hierarchy. *IEEE Transactions on Image Processing*, 27(10):4838–4849, 2018. 2

[62] Fanglei Xue, Zichang Tan, Yu Zhu, Zhongsong Ma, and Guodong Guo. Coarse-to-fine cascaded networks with smooth predicting for video facial expression recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2412–2418, 2022. 1

[63] Dingkang Yang, Shuai Huang, Haopeng Kuang, Yangtao Du, and Lihua Zhang. Disentangled representation learning for multimodal emotion recognition. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 1642–1651, 2022.

[64] Dingkang Yang, Haopeng Kuang, Shuai Huang, and Lihua Zhang. Learning modality-specific and-agnostic representations for asynchronous multimodal language sequences. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 1708–1717, 2022. 1

[65] Dingkang Yang, Zhaoyu Chen, Yuzheng Wang, Shunli Wang, Mingcheng Li, Siao Liu, Xiao Zhao, Shuai Huang, Zhiyan Dong, Peng Zhai, et al. Context de-confounded emotion recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19005–19015, 2023. 1, 2, 4, 5

[66] Stefanos Zafeiriou, Dimitrios Kollias, Mihalis A Nicolaou, Athanasios Papaioannou, Guoying Zhao, and Irene Kotsia. Aff-wild: valence and arousal'in-the-wild'challenge. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 34–41, 2017. 1

[67] Wang Zeng, Sheng Jin, Wentao Liu, Chen Qian, Ping Luo, Wanli Ouyang, and Xiaogang Wang. Not all tokens are equal: Human-centric visual analysis via token clustering transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11101–11111, 2022. 2

[68] Minghui Zhang, Yumeng Liang, and Huadong Ma. Context-aware affective graph reasoning for emotion recognition. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*, pages 151–156. IEEE, 2019. 1, 2, 4, 5

[69] Yifan Zhang, Bo Pang, and Cewu Lu. Semantic segmentation by early region proxy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1258–1268, 2022. 2