Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Issue 6, No.3, June : 2024

A MASKED SELF-SUPERVISED FACE PARSING PRETRAINING TECHNIQUE

Uroosa Fatima Zaidi Research Scholar, Department of CSE, Rabindranath tagore university, Bhopal, INDIA uroosa.zaidi@gmail.com
Dr. Rakesh Kumar Professor, Department of CSE, Rabindranath tagore university, Bhopal, INDIA
Dr Md Ilyas khan Professor, Department of CSE, Prestige Institute of Management and Research, Bhopal, INDIA

ABSTRACT:
Face masking, face switching, and face animation are just a few of the downstream jobs that can benefit from Face Parsing's goal of dividing the face into distinct semantic components. It is now simpler to obtain face photographs due to the widespread use of cameras. But pixel-by-pixel hand labeling islabor-and time-intensive, which encourages us to investigate the unlabeled data. In this research, we describe a self-supervised learning approach that aims to maximize the face parsing potential of the unlabeled facial photos. Specifically, we mask certain patches in the middle of face photos at random, and the model has to rebuild the patches that are masked. Through this unlabeled data, the model is able to collect representations of face features because to its self-supervised pretraining. The model is optimized for the face parsing problem on a small set of labeled data following self-supervised pretraining. According to experimental data, the model performs better for face parsing with the help of self-supervised pretraining, which significantly lowers the labeling cost. Our method succeeds in With only 0.2% of the training data labeled, the LaPa test set yielded 74.41 mIoU, outperforming the manually trained model by a significant margin of +5.02 mIoU. Furthermore, our method sets a new state-of-the-art on the test sets for LaPa and CelebAMask-HQ.

KEYWORDS: f
ace parsing; semantic segmentation; self-supervised learning

INTRODUCTION

Face parsing is a type of fine-grained semantic segmentation job that seeks to divide the face into distinct semantic components. Numerous downstream tasks, such as face recognition, face mask up face switching and face animation , have made extensive use of it. As cameras have developed so quickly, face parsing has gained more and more attention. The face structure was captured using several conventional machine learning techniques in the past. The local and global feature representation was constructed by combining the Conditional Random Field (CRF) with the Restricted Boltzmann Machine (RBM) Lately, models that utilize Fully Convolutional Networks (FCNs) have demonstrated remarkable advancements in this field. Generally speaking, these techniques fall into one of two groups depending on whether they anticipate the face bounding box first. Certain methods allow for direct parsing of the face region in an image by using the localization of the facial bounding box. Rather than The other methods predict a facial bounding box, but they parse the entire facial picture directly, which addresses face parsing by treating it as a distinct semantic segmentation job. Some methods for face parsing have recently been able to capture the graph representation. EAGRNet, or the Edge Aware Graph Reasoning Network, and Modular Graph Illustration Network (AGRNet) [20] are the component-wise (eyes, mouth, nose, etc.) and region-wise (facial appearance, position, emotion, etc.) relations via learning graph representations.But the majority of the researchfocuses on creating face model structures.
parsing As a matter of fact, annotations and photos are important for face parsingSince cameras are widely used in electronic commerce, obtaining face photos is not difficult. On the other hand, categorizing these face photos pixel-by-pixel takes a lot of effort and time. Naturally, the question of how to fully utilize the unlabeled data to raise the accuracy of the model emerges. The goal of self-supervised representation learning is to extract semantically significant features from data without the

need for extensive labeling. That type of unsupervised gaining knowledge It has garnered a lot of interest as a viable substitute because to its capacity for generalization and data efficiency. Based on this paradigm, several strategies have been put forth and fall into three categories: temporally-based and context-based techniques and tactics based on contrast. One type of context-based approach is masked self-supervised learning. It was suggested by Vincent et al. [24] to create masks as noise to help the model acquire helpful representations. In order to estimate the location of one patch in relation to the other, Doersch et al. [25] split pictures into patches, which are then randomly assigned to two different methods. A Convolutional Neural Network (CNN) was trained by Zhang et al. [26] to map from a grayscale input to a distribution of quantized color value outputs. Everything that Encoders can be trained effectively using context-based self-supervised learning, and they can be effectively applied to image classification using transfer learning. Nevertheless, unlike encoders, decoders in encoder-decoder models are not pretrained. They arelikewise unable to be effectively used for the granular job of face parsing in situations with very uneven class distribution. In this study, we provide a novel face parsing system that aims to fully use unlabeled facial photos. There are two phases to the framework: pretraining and fine-tuning. pictures are randomly masked in the central region during the pretraining phase, and the masked pictures are subsequently supplied into the model for reconstruction. This doesn't require any labeling. phase of pretraining; as a result, any picture may be utilized. It is anticipated that the pretrained model would accurately depict the face features. Using the labeled data for the face parsing task, the pretrained model may be improved in the next phase. In contrast to the direct The suggested strategy performs significantly better than the self-supervised pretraining supervised learning method. Furthermore, on the LaPa and CelebAMask-HQ test set, testing findings demonstrate that our technique achieves the new state-of-the-artperformance.

The following is a summary of this paper's main contributions: (1) We create a brand-new face parsing system including pretraining and fine-tuning phases. Within the Pretraining: For the face parsing job, the model is trained on the unlabeled data and then refined on the labeled data. (2) To pretrain the model, we provide a masked self-supervised learning technique. The model is anticipated

Torebuild the masked pictures in order to get the depiction of face features. (3) In-depth tests are carried out on two difficult benchmarks to demonstrate the noteworthy increase in performance of the suggested strategy in comparison to the most advanced techniques. The structure of this document is as follows. The ideas of face parsing and self-supervised learning are covered in Section 2. Part 3 provides a detailed description of the suggested structure, particularly how our method's network design and masked self-supervised learning function. We provide the full experiment parameters and compare them with the most recent methods in Section 4. The experiment's findings are assessed and spoken about. Lastly, Section 5 provides a summary of the paper's findings.

Related Work

FaceParsing

Face parsing has been the subject of active research in the past few years. For face parsing, a few conventional machine learning techniques were suggested. As opposed to constructing the Facial regions were modeled using Gaussian Radial Basis Function (RBF) [27], local and global feature representation [9] by CRF and RBM, and manually created features. Deep Convolutional Neural Networks have led to the proposal of FCNs for semantic segmentation in [10]. Regarding face parsing, Liu [11] split a unified network into a two-stage model that did a good job on discrete facial elements. By including the Spatial Transformer Networks (STN) [28] in between two isolated stages, STN-iCNN [13] expanded the Interlinked Convolutional Neural Networks (iCNN) [14], creating an end-to-end be able to train together. Nowadays, the majority of techniques concentrate on face parsing by direct picture parsing [15–19]. An efficient and successful hierarchical aggregation network named EHANet was suggested by Luo et al. [19] and featured a stage contextual attention mechanism and a Semantic gap correction block for constructing hierarchical and contextual information at a higher level. The border information was also completely utilized by EHANet and the border-Attention Semantic

Segmentation (BASS) technique [17]. EAGRNet [18] investigated the region-wise relations in addition to the CNN models by learning graph representations, where the edge cues were additionally utilized to project important pixels onto graph vertices on a higher level of semantics. To represent facial components, AGRNet [20] developed an adaptable and differentiable graph abstraction technique; correct face parsing is needed with
Theenhanced vertex characteristics.

Self-Supervised Learning

Without the need for human labeling, self-supervised learning extracts characteristics directly from the data itself. It has garnered a lot of interest as a viable substitute because to its capacity for generalization and data efficiency. Generally speaking, it falls into three groups: techniques based on context techniques temporal techniques and contrast techniques Context-based techniques pick up knowledge from the contextual data samples themselves. Time restrictions are used by temporal-based approaches to construct feature representations that may be applied to movies. Contrast-based techniques work with contrast restrictions, learning to encode the similarity or dissimilarity to create representations.
comprises two items. Numerous tasks can be created based on the contextual information contained in the data itself. For instance, Context Encoder [29] was trained to anticipate the missing pixel values given an image that had a missing region. Using grayscale picture input, a CNN To produce the distribution across quantized color values, a model is needed. Researchers have been drawn to masked self-supervised learning as another type of context-based methodology. Similar to ,Noroozi et al. split pictures into patches and suggest using jigsaw puzzles to help learners acquire visual representations. It was suggested by Vincent et al. [24] to produce masks as noise in order to help the model acquire useful representations. Masked patch prediction was also investigated by Vision Transformer (ViT) for self-supervised learning.

.

Approach

We will outline our method for masked self-supervised face parsing pretraining in this part. Overarching Structure Figure depicts our approach's general structure. The neural network is pretrained on the masked pictures in Step 1, which should allow it to rebuild the input images without the need for label parsing. The neural network is refined on a specific number of pictures with parsing labels in the second stage. Pretraining Under Self-Supervision
To acquire relevant semantic characteristics from the unlabeled images for pretraining the neural network, we investigate a novel masked self-supervised learning technique. Specifically, we employ a CNN network to rebuild the masked areas of the pictures after masking parts of them. pictures. The masked patch's size varied from 32 to 64 pixels for photos that were 512 by 512 pixels in size. A certain amount of patches in this paperare hidden. The target face in facial pictures found in datasets is typically found in the image's center. Only patches from the core region are hidden in order to concentrate on obtaining the face feature representation. The core area is used to calculate the reconstruction loss. In this work, the center region is defined as two-thirds of the entire picture. We recreate the data using a basic R decoder with a single convolution layer and the UNet++ [42] architecture as the encoder. veiled picture. Using a masked input of size 3 x H x W, the encoder extracts n x H x W features, which it then feeds into the R Decoder to recreate a 3 x H x W picture.

Figure 1a displays the overall framework.

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Issue 6, No.3, June : 2024

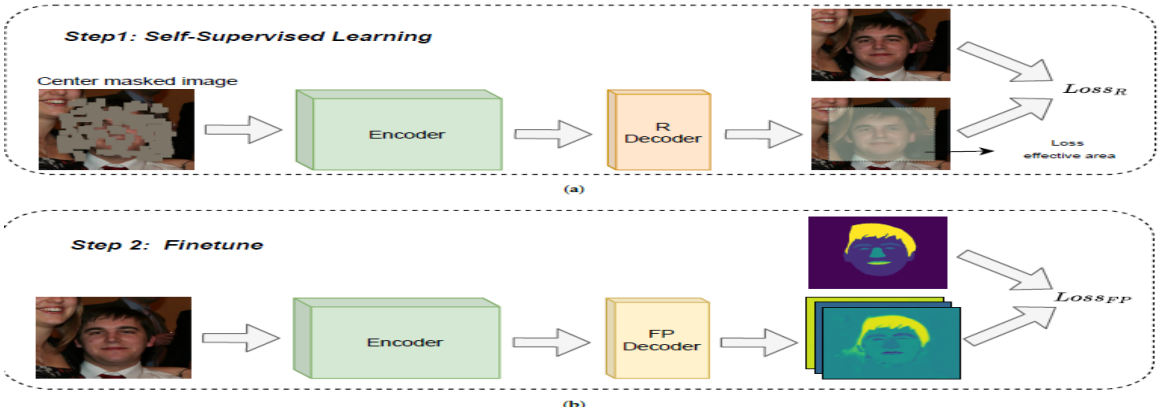

Figure 1. The overall framework of our approach. We build our encoder followed the UNet++ and share the same in two stages. (**a**) Step 1: Masked self-supervised pretraining. (**b**) Step 2: Fine-tune models on the target dataset.

Fine-Tune

Following the self-supervised pretraining, the encoder may learn to represent face features. We use the same encoder in the fine-tuning step, which was pretrained at the stage of self-supervised learning and is capable of picking up semantic aspects of faces. As seen in Figure 1b, we create a decoder for face parsing called FP Decoder and append it to the encoder.

.

CNN Architectures

The suggested approach uses the same encoder in two stages but separate decoders. We use a masked self-supervised learning technique in the first step to assist the encoder in learning representations of face features. However, we're hoping that the decoder need should be easy to learn properly. This served as inspiration for our encoder, which is a robust UNet++ architecture with ResNet50 serving as the core. A UNet++ architecture receives an image with a size of 3_ H _W, as seen in Figure 2, and uses it to extract n_ H _W face feature representations. Our R and FP decoders have a single convolution layer. Consequently, during the fine-tuning phase, a potent encoder may be effectively pretrained and repurposed.

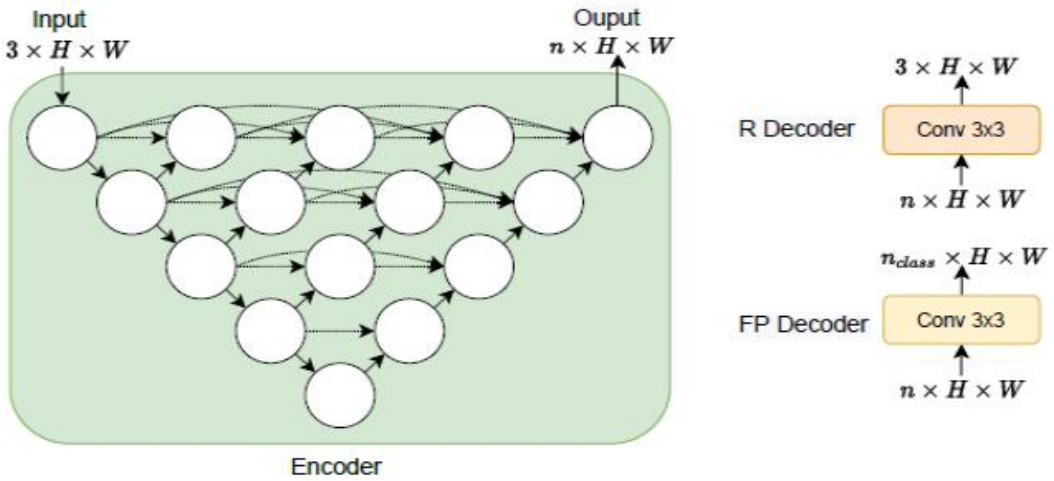

Figure 2. The architecture of Encoder, R Decoder, and FP Decoder in the proposed framework. Nclass denotes the number of facial parts labeled in the dataset

Fine-Tuning Process

The encoder is pretrained on the ImageNet dataset for our baseline training. Figure 4 illustrates how mIoU improves quickly and the loss has a faster convergence speed in the absence of self-supervised learning. Nonetheless, the suggested approach directs the model to accomplish improved performance

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Issue 6, No.3, June : 2024

in the previous one. It demonstrates how easy the original model becomes mired in local optimization. Stated differently, the model is assisted in surpassing the
local optimum by means of masked self-supervised learning.

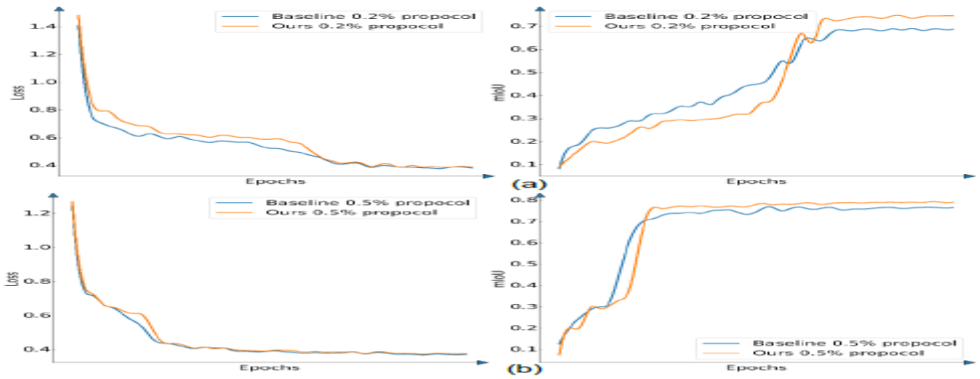

Figure 3 .(**a**,**b**) show the difference between the baseline and our method in training loss and test set mIoU with 0.2 and 0.5% protocol training data. As for the mIoU performance, our method has worse performance but ends with better mIoU consistently.

Discussions

We investigate the effectiveness of masked self-supervised pretraining in detail. Luo's UNet, Depth-wise Asymmetric Bottleneck Network (DABNet), and Deep Feature Aggregation Network (DFANet) all performed poorly on the necklace portion of CelebAMask-HQ (0.00, 0.01, 0.00 mIoU, respectively). Table illustrates how poorly our baseline scores on this section as well. After analyzing the data of every face region in CelebAMask-HQ, we discovered that the necklace-related pixel barely makes up 0.017% of all the pixels. In the absence of more features, models prioritize optimization.only use semantic masks for training on additional categories, ultimately reaching a local optimum. As a result, it is unable to adequately learn categories like jewelry that have extremely few pixels. All feature activation maps with a dimension of 16 X 512 X 512 are displayed in Figure . the output of the encoder. The necklace portion depicted in Figure 5e is not activated in the baseline activation maps. However, the model must recreate the picture in our suggested self-supervised pretraining technique, and the whole training process is classified separately. As a result, it forces the model to fairly concentrate on each category. By using the suggested masked self-supervised pretraining, our model may acquire the representation of the necklace feature.

Consequently, as indicated in Figure, obtain the feature activation on the appropriate spots.

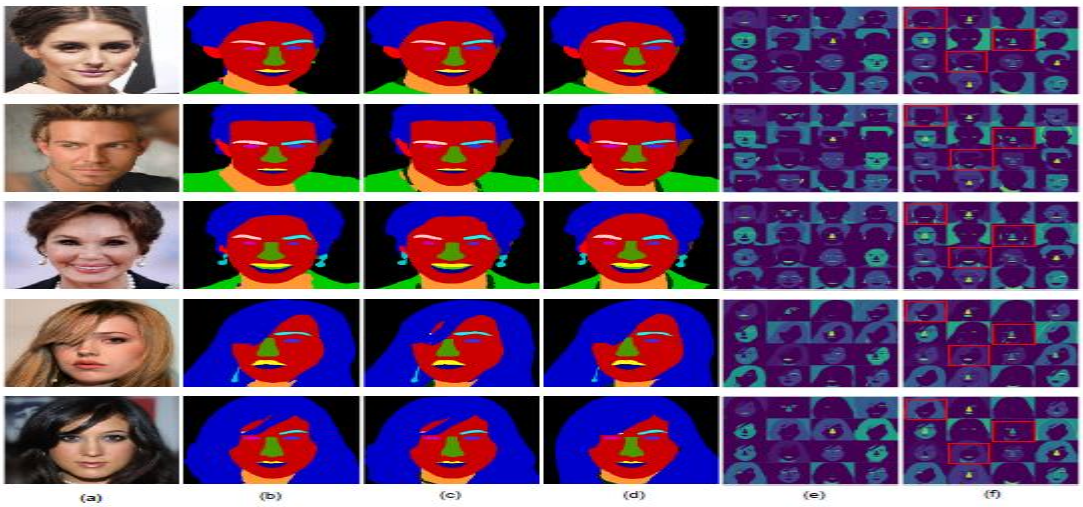

Figure 4 CelebAMask-HQ test set general image parsing findings and feature activations. Original pictures (a). (b): Unet++ parsing results using Imagenet pretrained (baseline). (c): Using the

suggested self-supervised approach, Unet++ was pretrained. (d): The real world. (e): Encoding function activation starting at the beginning. (f): Activating our encoder functionality.

With necklace, the basic model performs poorly and does not activate the necklace function. Better outcomes are shown by the pretrained model using our techniques. Even the first two rows outperform the ground truth in terms of performance. Our pretraining technique aids the model in accurately building necklace feature representation, as shown by the red box in the final columns.

Table . Pixel ratios of face-part categories on the CelebAMask-HQ train set in terms of F1 scor

Face part	Face	Nose	Glasses	L-Eye	R-Eye	L-Brow
Pixel Rations (%)	25.34	2.06	0.27	0.22	0.22	0.42
Face part	R-Brow	L-Ear	R-Ear	I-Mouth	U-Lip	L-Lip
Pixel Ratios (%)	0.41	0.46	0.39	0.30	0.41	0.68
Face part	hair	hat	earring	Necklace	Neck	Cloth
Fixel Ratios (%)	0.31	0.90	0.24	0.017	4.10	3.35

Conclusions

In this study, we provide a new approach to reduce the manual labeling work on dense face parts annotations: self-supervised pretraining. Our suggested procedure involves prettrainingUnet++ on the masked pictures, where the patches from the centralThe goal of masking a region of photos is to rebuild the masked images. Our model is refined on the target face parsing dataset following pretraining. The experimental findings show that the strong baseline (directly trained on the labeled data with ImageNet pretraining) fine-tuned on various fractions of labeled data may be consistently improved by our suggested self-supervised pretraining technique. Additionally, our approach reaches the new cutting-edge performance on the CelebAMask-HQ and LaPa datasets. Additionally, the newly suggested approach aids in the model's development of a thorough feature representation. The improved model reliably gets

feature activations on all categories—even those with extremely modest ratios—through the feature visualization. The experiment demonstrates that much improved parsing performance is attained, particularly for categories with extremely tiny ratios (like the necklace in CelebAMask-HQ). We believe that more face-related problems, such as face landmark identification, face creation, and face attribute learning, may be addressed with similar disguised self-supervised pretraining technique. Our next task will be to examine the self-supervised pretraining's possible efficacy for small amounts of labeled data.

References

1. Masi, I.;Wu, Y.; Hassner, T.; Natarajan, P. Deep Face Recognition: A Survey. In Proceedings of the 2018 31st SIBGRAPI Conferenceon Graphics, Patterns and Images (SIBGRAPI), Paraná, Brazil, 29 October 29–1 November 2018; pp. 471–478. [CrossRef]
2. Adjabi, I.; Ouahabi, A.; Benzaoui, A.; Taleb-Ahmed, A. Past, Present, and Future of Face Recognition: A Review. Electronics **2020**,9, 1188. [CrossRef]
3. Ou, X.; Liu, S.; Cao, X.; Ling, H. Beauty emakeup: A deep makeup transfer system. In Proceedings of the ACM Multimedia,Amsterdam, The Netherlands, 15–19 October 2016; pp. 701–702.
4. Kemelmacher-Shlizerman, I. Transfiguring portraits. ACM Trans. Graph. **2016**, 35, 1–8. [CrossRef]
5. Nirkin, Y.; Masi, I.; Tuan, A.T.; Hassner, T.; Medioni, G. On face segmentation, face swapping, and face perception. In Proceedingsof the 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), Xi'an, China, 15–19 May
2018; pp. 98–105.
6. Lee, C.H.; Liu, Z.;Wu, L.; Luo, P. Maskgan: Towards diverse and interactive facial image manipulation. In Proceedings of theIEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 5549–5558.

7. Zhang, H.; Riggan, B.S.; Hu, S.; Short, N.J.; Patel, V.M. Synthesis of High-Quality Visible Faces from Polarimetric Thermal Facessing Generative Adversarial Networks. Int. J. Comput. Vis. **2018**, 127, 845–862. [CrossRef]
8. Zhang, K.; Zhang, Z.; Li, Z.; Qiao, Y. Joint face detection and alignment using multitask cascaded convolutional networks. IEEESignal Process. Lett. **2016**, 23, 1499–1503[CrossRef]
9. Kae, A.; Sohn, K.; Lee, H.; Learned-Miller, E. Augmenting CRFs with Boltzmann machine shape priors for image labeling. InProceedings of the IEEE International Conference on Computer Vision, Sydney, Australia, 8 April 2013; pp. 2019–2026.
10. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEEConference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
11. Liu, S.; Shi, J.; Liang, J.; Yang, M.H. Face parsing via recurrent propagation. In Proceedings of the 28th British Machine VisionConference, BMVC 2017, London, UK, 4–7 September 2017; pp. 1–10.
12. Lin, J.; Yang, H.; Chen, D.; Zeng, M.; Wen, F.; Yuan, L. Face Parsing with RoI Tanh-Warping. In Proceedings of the IEEEConference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 5654–5663.
13. Yin, Z.; Yiu, V.; Hu, X.; Tang, L. End-to-End Face Parsing via Interlinked Convolutional Neural Networks. arXiv**2020**,arXiv:2002.04831.
14. Zhou, Y.; Hu, X.; Zhang, B. Interlinked convolutional neural networks for face parsing. In International Symposium on NeuralNetworks; Springer: Berlin/Heidelberg, Germany, 2015; pp. 222–231.
15. Wei, Z.; Sun, Y.;Wang, J.; Lai, H.; Liu, S. Learning adaptive receptive fields for deep image parsing network. In Proceedings ofthe IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2434–2442.
16. Liu, S.; Yang, J.; Huang, C.; Yang, M.H. Multi-objective convolutional learning for face labeling. In Proceedings of the IEEEInternational Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 3451–3459.
17. Liu, Y.; Shi, H.; Shen, H.; Si, Y.;Wang, X.; Mei, T. A New Dataset and Boundary-Attention Semantic Segmentation for Face Parsing.In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; pp. 11637–11644.
18. Te, G.; Liu, Y.; Hu, W.; Shi, H.; Mei, T. Edge-aware Graph Representation Learning and Reasoning for Face Parsing. In EuropeanConference on Computer Vision; Springer: Cham, Switzerland, 2020; pp.258–274.
19. Luo, L.; Xue, D.; Feng, X. EHANet: An Effective Hierarchical Aggregation Network for Face Parsing. Appl. Sci. **2020**, 10, 3135.[CrossRef]
20. Te, G.; Hu,W.; Liu, Y.; Shi, H.; Mei, T. Agrnet: Adaptive graph representation learning and reasoning for face parsing. IEEE Trans.Image Process. **2021**, 30, 8236–8250. [CrossRef] [PubMed]
21. Luo, P.;Wang, X.; Tang, X. Hierarchical face parsing via deep learning. In Proceedings of the IEEE International Conference onComputer Vision, Providence, RL, USA, 16–21 June 2012; pp. 2480–2487.
22. Dike, H.U.; Zhou, Y.; Deveerasetty, K.K.; Wu, Q. Unsupervised Learning Based On Artificial Neural Network: A Review. InProceedings of the 2018 IEEE International Conference on Cyborg and Bionic Systems (CBS), Shenzhen, China, 25–27 October 2018;
pp. 322–327. [CrossRef]
23. Khaldi, Y.; Benzaoui, A.; Ouahabi, A.; Jacques, S.; Taleb-Ahmed, A. Ear Recognition Based on Deep Unsupervised ActiveLearning. IEEE Sensors J. **2021**, 21, 20704–20713. [CrossRef]
24. Vincent, P.; Larochelle, H.; Lajoie, I.; Bengio, Y.; Manzagol, P.A.; Bottou, L. Stacked denoising autoencoders: Learning usefulrepresentations in a deep network with a local denoising criterion. J. Mach. Learn. Res. **2010**, 11, 3371–3408

25. Doersch, C.; Gupta, A.; Efros, A.A. Unsupervised visual representation learning by context prediction. In Proceedings of theIEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1422–1430.
26. Zhang, R.; Isola, P.; Efros, A.A. Colorful image colorization. In European Conference on Computer Vision; Springer:Berlin/Heidelberg, Germany, 2016; pp. 649–666.
27. Smith, B.M.; Zhang, L.; Brandt, J.; Lin, Z.; Yang, J. Exemplar-based face parsing. In Proceedings of the IEEE InternationalConference on Computer Vision, Sydney, Australia, 8 April 2013; pp. 3484–3491.
28. Jaderberg, M.; Simonyan, K.; Zisserman, A.; Kavukcuoglu, K. Spatial transformer networks. Adv. Neural Inf. Process. Syst. **2015**,28, 2017–2025.
29. Pathak, D.; Krahenbuhl, P.; Donahue, J.; Darrell, T.; Efros, A.A. Context encoders: Feature learning by inpainting. In Proceedingsof the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2536–2544.
30. Wang, X.; Girshick, R.; Gupta, A.; He, K. Non-local neural networks. In Proceedings of the IEEE Conference on Computer Visionand Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7794–7803.
31. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold,G.; Gelly, S.; et al. An Image is Worth 16 _ 16 Words: Transformers for Image Recognition at Scale. In Proceedings of theInternational Conference on Learning Representations, Virtual Event, Austria, 3–7 May 2021.
32. Sermanet, P.; Lynch, C.; Chebotar, Y.; Hsu, J.; Jang, E.; Schaal, S.; Levine, S.; Brain, G. Time-contrastive networks: Self-supervisedlearning from video. In Proceedings of the 2018 IEEE International Conference on Robotics and Automation (ICRA), Brisbane,
Australia, 21–25 May 2018; pp. 1134–1141.
33. Wang, X.; Gupta, A. Unsupervised learning of visual representations using videos. In Proceedings of the IEEE InternationalConference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 2794–2802.