mathematics

MDPI

Article

A Masked Self-Supervised Pretraining Method for Face Parsing

Zhuang Li [1,2], Leilei Cao [2], Hongbin Wang [2] and Lihong Xu [1,*]

[1] Department of Control Science and Engineering, Tongji University, Shanghai 201804, China; liz@tongji.edu.cn
[2] Ant Group, Hangzhou 310013, China; caoleilei.cll@antgroup.com (L.C.); hongbin.whb@antgroup.com (H.W.)
* Correspondence: xulhk@163.com

Abstract: Face Parsing aims to partition the face into different semantic parts, which can be applied into many downstream tasks, e.g., face mask up, face swapping, and face animation. With the popularity of cameras, it is easier to acquire facial images. However, pixel-wise manually labeling is time-consuming and labor-intensive, which motivates us to explore the unlabeled data. In this paper, we present a self-supervised learning method attempting to make full use of the unlabeled facial images for face parsing. In particular, we randomly mask some patches in the central area of facial images, and the model is required to reconstruct the masked patches. This self-supervised pretraining is capable of making the model capture facial feature representations through these unlabeled data. After self-supervised pretraining, the model is fine-tuned on a few labeled data for the face parsing task. Experimental results show that the model achieves better performance for face parsing assisted by the self-supervised pretraining, which greatly decreases the labeling cost. Our approach achieves 74.41 mIoU on the LaPa test set fine-tuned on only 0.2% of the labeled data of the whole training data, surpassing the model that is directly trained by a large margin of +5.02 mIoU. In addition, our approach achieves a new state-of-the-art on the LaPa and CelebAMask-HQ test set.

Keywords: face parsing; semantic segmentation; self-supervised learning

MSC: 68U10

check for updates

Citation: Li, Z.; Cao, L.; Wang, H.; Xu, L. A Masked Self-Supervised Pretraining Method for Face Parsing. *Mathematics* **2022**, *10*, 2002. https://doi.org/10.3390/math10122002

Academic Editor: Radu Tudor Ionescu

Received: 15 May 2022
Accepted: 7 June 2022
Published: 10 June 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Face parsing, aiming to partition the face into different semantic parts, belongs to a fine-grained semantic segmentation task. It has been widely used in various downstream tasks, e.g., face recognition [1,2], face mask up [3], face swapping [4,5] and face animation [6–8]. With the rapid development of cameras, face parsing has drawn increasing attention.

In earlier years, some traditional machine learning methods were used to capture the facial structure. Conditional Random Field (CRF) combined with Restricted Boltzmann Machine (RBM) was applied to build the local and global feature representation [9]. Recently, models based on the Fully Convolutional Networks (FCNs) [10] have made great progress in this area. These methods can be roughly divided into two categories based on whether they predict the facial bounding box first. With the localization of facial bounding box, some approaches [11–14] can directly parse on facial region in an image. Instead of predicting a facial bounding box, the other approaches [15–19] parse the whole facial image directly, which solves face parsing by regarding it as a specific semantic segmentation task. Recently, some approaches capture the graph representation for face parsing. Edge Aware Graph Reasoning Network (EAGRNet) [18] and Adaptive Graph Representation Network (AGRNet) [20] are the region-wise (facial appearance, pose, expression, etc.) and component-wise (eyes, mouth, nose, etc.) relations by learning graph representations.

However, most of the literature focuses on building model architectures for face parsing [11,17,21]. In fact, images and annotations are also significant for face parsing [17]. With the popularity of cameras in electronic commerce, it is quite easy to acquire facial images. Conversely, pixel-wise labeling on these facial images is time-consuming and

labor-intensive. A natural question arises: how can we make full use of the unlabeled data to improve the model's accuracy?

Self-supervised representation learning aims to learn semantically meaningful features from data without the requirement of large-scale labels. It is a form of unsupervised learning [22,23]. As a promising alternative, it has drawn massive attention for its data efficiency and generalization ability. Many approaches have been proposed based on this paradigm, which can be divided into three categories: context-based methods, temporal-based methods and contrast-based methods. Masked self-supervised learning is a kind of context-based method. Vincent et al. [24] proposed to generate masks as noise to assist the model to learn useful representations. Doersch et al. [25] divide images into patches, and the algorithms receive random two patches to guess the position of one patch relative to the other. Zhang et al. [26] trained a Convolutional Neural Network (CNN) to map from a grayscale input to a distribution over quantized color value outputs. All of this context-based self-supervised learning provides a good way to train the encoders, and they can be well used in image classification with transfer learning. However, decoders in the Encoder–Decoder models can not be pretrained just like the encoder. They also cannot be applied well to the fine-grained task of face parsing where the class distribution is extremely unbalanced.

In this paper, we propose a new framework for face parsing, which attempts to make full use of unlabeled facial images. The framework consists of two stages: pretraining and fine-tuning. In the pretraining stage, images are masked randomly in the center area and then fed into the model to reconstruct the masked images. No labels are needed for this stage of pretraining; therefore, all images can be used. The pretrained model is expected to capture the facial feature representation. In the next stage, the pretrained model can be fine-tuned on the labeled data for the face parsing task. Compared with the directly supervised learning method without the self-supervised pretraining, the proposed method achieves much better performance. In addition, experimental results show that our method achieves the new state-of-the-art performance on the LaPa and CelebAMask-HQ test set. The major contributions of this paper are summarized as follows: (1) We design a novel framework for face parsing, which consists of pretraining and fine-tuning stages. In the pretraining, the model is trained on the unlabeled data and then fine-tuned on the labeled data for the face parsing task. (2) We propose a masked self-supervised learning method to pretrain the model, which is expected to reconstruct the masked images to acquire the facial feature representation. (3) Extensive experiments are conducted on two challenging benchmarks to show the significant performance improvement of the proposed approach compared with the state-of-the-art methods.

This paper is organized as follows. Section 2 discusses the concept of face parsing and self-supervised learning. Section 3 describes the proposed framework in detail, especially how the masked self-supervised learning works and the network architecture used in our method. In Section 4, we give the detailed experiment settings and compared them with the state-of-the-art algorithms. The experiment results are evaluated and discussed. Finally, Section 5 summarizes the conclusions of this paper.

2. Related Work

2.1. Face Parsing

In the earlier years, face parsing has been actively investigated. Some traditional machine learning approaches for face parsing were proposed. Different from building the local and global feature representation [9] by CRF and RBM, Gaussian Radial Basis Function (RBF) [27] combined with hand-crafted features were used to model facial regions. With the development of deep Convolutional Neural Networks, FCNs were proposed in [10] for semantic segmentation. As for the face parsing, Liu [11] decomposed a unified network into a two-stage model which performed well on small facial components. STN-iCNN [13] extended the Interlinked Convolutional Neural Networks (iCNN) [14] by adding the Spatial Transformer Networks (STN) [28] between two isolated stages, making the end-to-end

joint training be possible. Recently, most of methods focus on face parsing by parsing image directly [15–19]. Luo et al. [19] proposed an effective and efficient hierarchical aggregation network called EHANet, which included a stage contextual attention mechanism and a semantic gap compensation block to build higher-level contextual information and hierarchical information. EHANet and the Boundary-Attention Semantic Segmentation (BASS) method [17] also fully exploited the boundary information. In addition to the CNN models, EAGRNet [18] explored the region-wise relations by learning graph representations, where the edge cues were also used to project significant pixels onto graph vertices on a higher semantic level. AGRNet [20] designed an adaptive and differentiable graph abstraction method to represent facial components; with the refined vertex features, accurate face parsing is required.

2.2. Self-Supervised Learning

Self-supervised learning acquires features from the data itself without any manual labels. As a promising alternative, it has drawn massive attention for its data efficiency and generalization ability. Generally, it can be divided into three categories: context-based methods [24,25,29–31], temporal based methods [32–35] and contrast based methods [36–40]. Context-based methods learn from the sample contextual information itself. Temporal-based methods build feature representations by taking advantage of time constraints, which can be applied to videos. Contrast-based methods are carried out based on contrast constraints, which build representations by learning to encode the similarity or dissimilarity of two things. Based on the contextual information of the data itself, many tasks can be constructed. For example, given an image with a missing region, Context Encoder [29] was trained to predict the missing pixel values. With the input of gray images, a CNN model [26] is required to output the distribution over quantized color values. Masked self-supervised learning is also a kind of context-based method which has attracted many researchers. Noroozi et al. [41] divide images into patches just like [25] and proposed to learn image representations by solving jigsaw puzzles. Vincent et al. [24] proposed to generate masks as noise to assist the model in learning valuable representations. Vision Transformer (ViT) [31] also explored masked patch prediction for self-supervised learning.

3. Approach

In this section, we will describe our approach toward masked self-supervised pretraining for face parsing.

3.1. Overall Framework

The overall framework of our approach is shown in Figure 1. In Step 1, the neural network is pretrained on the masked images, which is expected to reconstruct the input images without the requirement of parsing labels. In the second step, the neural network is fine-tuned on a certain number of images with parsing labels.

3.1.1. Self-Supervised Pretraining

We explore a new masked self-supervised learning method to learn meaningful semantic features from the unlabeled images for pretraining the neural network. In particular, we mask some regions of the images and use a CNN network to reconstruct the masked images. For images with the size of 512×512, the size of the masked patch ranged from 32 pixels to 64 pixels. A certain number of patches (128 in this paper) are masked. For facial images in datasets, the target face is usually located in the central area of the image. To focus on acquiring the facial feature representation, only patches from the central area are masked. The reconstruction loss is computed on the central area. In this paper, we define two-thirds of the whole image as the central area. We use the UNet++ [42] architecture as the encoder and a simple R Decoder with only one convolution layer to reconstruct the masked image. The encoder extracts $n \times H \times W$ features from the masked input with the

size of $3 \times H \times W$ and feeds them into R Decoder, which reconstructs a $3 \times H \times W$ image. The overall framework is shown in Figure 1a.

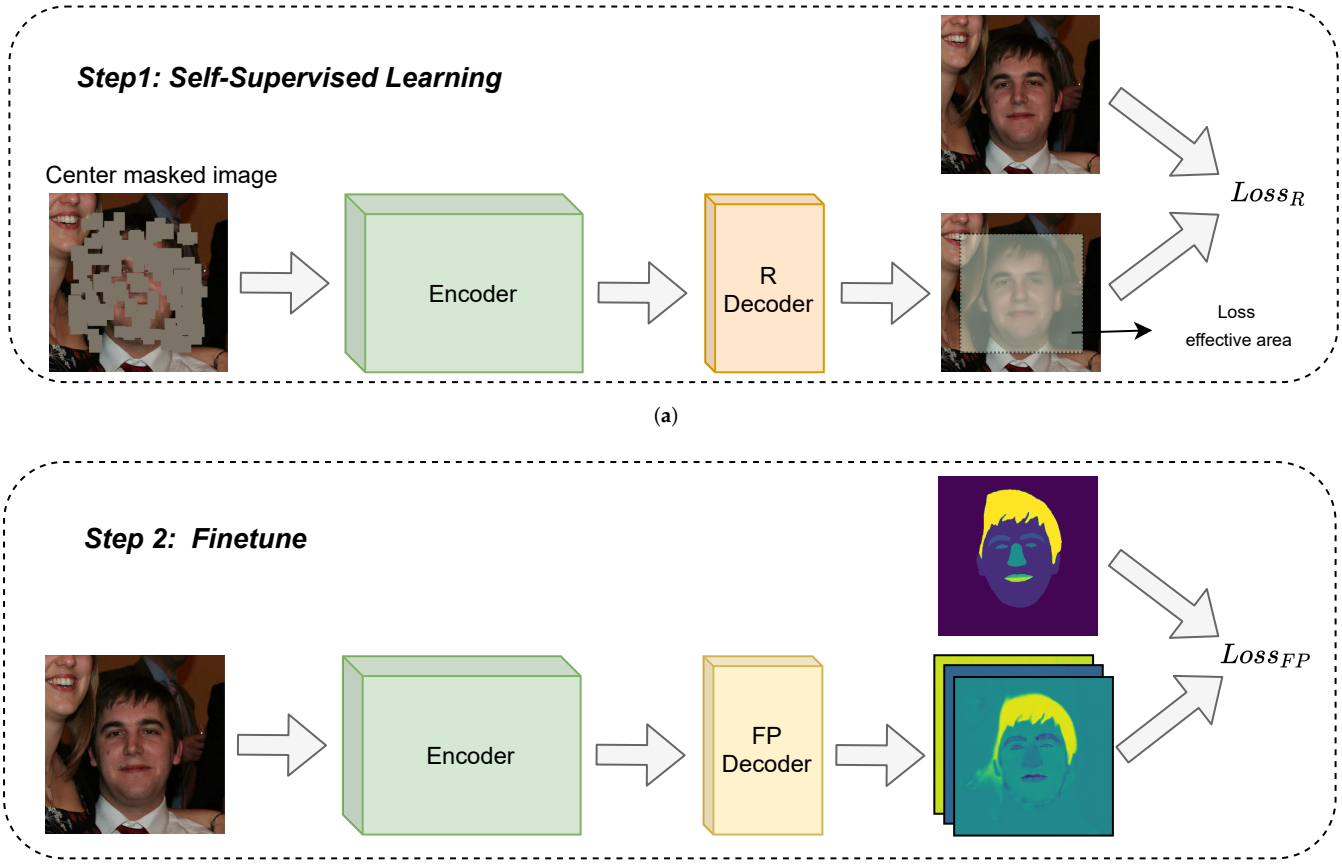


Figure 1. The overall framework of our approach. We build our encoder followed the UNet++ and share the same in two stages. (**a**) Step 1: Masked self-supervised pretraining. (**b**) Step 2: Fine-tune models on the target dataset.

3.1.2. Fine-Tune

After the self-supervised pretraining, the encoder is capable of acquiring facial feature representation. In the fine-tuning stage, we share the same encoder, which is pretrained in the self-supervised learning stage and has the power to learn the facial semantic features. Appended to the encoder, we design a decoder named FP Decoder for face parsing, as shown in Figure 1b.

3.2. CNN Architectures

In the proposed method, we apply different decoders but the same encoder in two stages. In the first stage, we conduct a masked self-supervised learning method to help the encoder acquire facial feature representations. On the other hand, we hope the decoder should be simple to be trained well. Inspired by this, a powerful UNet++ architecture with ResNet50 as the backbone is used as our encoder. As shown in Figure 2, an image with the size of $3 \times H \times W$ is fed into a UNet++ architecture, which acquires $n \times H \times W$ facial feature representations. As for R Decoder and FP Decoder, we only use one convolution layer. Therefore, a powerful encoder can be well pretrained and reused in the fine-tuning stage.

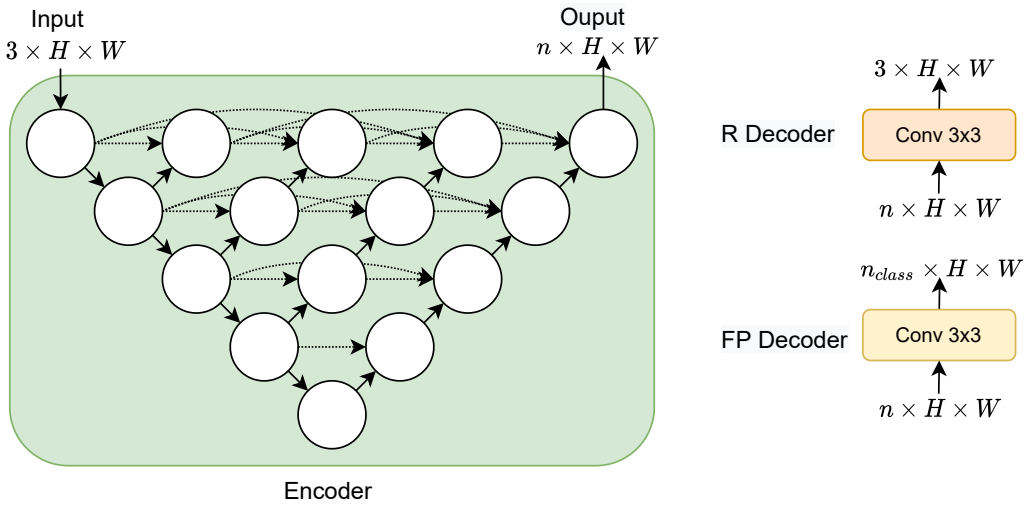


Figure 2. The architecture of Encoder, R Decoder, and FP Decoder in the proposed framework. n_{class} denotes the number of facial parts labeled in the dataset.

3.3. Training

We leverage the masked self-supervised learning to pretrain our encoder and fine-tune the encoder associated with FP Decoder for the face parsing task.

3.3.1. Pretraining

The pretraining stage aims to reconstruct the missing patches from the input masked image. We compute the following loss:

$$Loss_R = \sum_{i=W_s}^{W_e} \sum_{j=H_s}^{H_e} L(\hat{y}_{ij}, y_{ij}) \tag{1}$$

where $L(\hat{y}_{ij}, y_{ij})$ is the pixel loss between the original image and the reconstructed image. In our experiment, the L_1 loss is used in this task. Here, $0 < W_s < W_e < W$, $0 < H_s < H_e < H$, $W_s = \frac{1}{6} * W$, $W_e = \frac{5}{6} * W$, $H_s = \frac{1}{6} * H$, $H_e = \frac{5}{6} * H$.

3.3.2. Fine-Tuning

Face parsing belongs to a semantic segmentation task that assigns each pixel to semantic labels of facial parts. To obtain finer decision boundaries and accurate data distribution, we make use of the combined loss as follows:

$$Loss_{FP} = \lambda_{dice} L_{dice} + \lambda_{ce} L_{ce} \tag{2}$$

where λ_{dice} and λ_{ce} are hyper-parameters. L_{dice} and L_{ce} denote the Dice loss and the Cross-Entropy loss, respectively. In our experiments, $\lambda_{dice} = \lambda_{ce} = 0.5$ gives good performance with respect to the F1 score and mIoU.

4. Experiments

4.1. Datasets and Metrics

4.1.1. Datasets

We evaluate our proposed method on two benchmarks: LaPa [17] and CelebAMask-HQ [12]. LaPa is a new large-scale landmark guided face parsing dataset containing 22,176 facial images with elaborated pixel-wise annotations of 10 semantic face-part categories. The images are divided into 18,176 images for training, 2000 images for validation, and 2000 images for testing. The CelebAMask-HQ dataset is composed of 24,183 training images, 2993 validation images, and 2824 test images with 18 semantic face-part categories.

4.1.2. Metrics

The standard Mean Intersection-over-Union (mIOU) criterion, Pixel Accuracy (pix Acc), mean Pixel Accuracy (Mean Acc), and Mean F1 score (Mean F1) are adopted for evaluation on face parsing results, both for LaPa and CelebAMask-HQ. To maintain consistency with the other approaches, we also report the F1 score performance.

4.1.3. Implementation Details

During the self-supervised learning and fine-tuning steps, random rotation and scale augmentation are employed with all our experiments. To be specific, the rotation angle is randomly selected from $(-30°, 30°)$ and $(-15°, 15°)$ for different steps. The scale factor is randomly selected from (0.75, 1.25) and (0.8, 1.2) for two steps. We choose hyper-parameters in Equation (2) as $\lambda_{dice} = \lambda_{ce} = 0.5$.

All the experiments are implemented with two NVIDIA V100 GPUs. Optimizations for self-supervised learning and fine-tuning steps are Stochastic Gradient Descent (SGD). The input size for both steps is 480×480, which is the nearest $32\times$ size to the 473×473 size used in compared methods due to the network limitation and fair comparisons. The batch size is set to 16 and the learning rate starts at 0.0001 for the first step and 0.00001 for the second step.

We first pretrain (self-supervised learning) the encoder with ResNet50 as a backbone and R Decoder for 500 epochs. The backbone is initialized with a pretrained model on ImageNet. Then, we use the self-supervised learning model to initialize the whole encoder in Figure 2, and we fine-tune the encoder and FP Decoder on the LaPa and CelebAMask-HQ datasets, respectively.

In the fine-tuning stage, there are two settings for validating the performance of our method. We first randomly sample 0.2%, 0.5%, 1%, 5%, and 10% from the training data with labels to fine-tune, respectively. For these experiments, we create five different data folds, and the final performance is the average of all five folds. Then, the entire training data with labels are also used for fine-tuning.

4.2. Quantitative Evaluation

To verify the effectiveness of masked self-supervised learning, we compare our method with our baseline, which is directly trained on the labeled data without masked self-supervised learning. Table 1 summarizes the comparing results on different protocols, which shows that our approach improves the baseline significantly on two benchmarks. Specifically, our approach outperforms the baseline by 5.02 mIoU, 2.91 mIoU, 2.14 mIoU, 1.07 mIoU, and 1.14 mIoU when there are 0.2%, 0.5%, 1%, 5% and 10% labeled data from the LaPa dataset, respectively. As for the comparing results at the entire training data on the LaPa dataset, our approach outperforms the baseline by 0.22 mIoU. The comparing results on the CelebAMask-HQ dataset also show our approach outperforms the baseline by a large margin. For the entire training data setting, we compare our method with other state-of-the-art methods in terms of F1 score. Tables 2 and 3 show the performance comparison between our approach and the state-of-the-art methods trained on the entire data of the LaPa dataset and the CelebAMask-HQ dataset, respectively. Our approach achieves a new state-of-the-art performance with the mean F1 score of 92.5 on LaPa and 86.7 on CelebAMask-HQ, respectively. Our approach also achieves the best performance on some facial semantic parts.

Table 1. Performance comparison between the baseline (without the masked self-supervised learning) and our approach in terms of pixel accuracy, mean accuracy, mIoU and mean F1 score with state-of-the-art methods on the LaPa and CelebAMask-HQ test set.

Dataset		Method	pixAcc.	Mean Acc.	Mean IoU	Mean F1
LaPa	0.2%	Baseline	91.71 ± 0.35	83.09 ± 0.57	69.39 ± 3.21	78.69 ± 0.78
		Ours	93.16 ± 0.33	86.88 ± 0.51	74.41 ± 2.89	82.92 ± 0.67
	0.5%	Baseline	94.58 ± 0.31	86.69 ± 0.41	76.33 ± 2.32	85.21 ± 0.53
		Ours	95.39 ± 0.31	89.58 ± 0.39	79.24 ± 2.13	87.52 ± 0.41
	1%	Baseline	95.50 ± 0.22	88.82 ± 0.32	78.20 ± 1.89	86.73 ± 0.37
		Ours	95.88 ± 0.19	90.03 ± 0.29	80.34 ± 1.75	88.19 ± 0.30
	5%	Baseline	97.34 ± 0.13	91.21 ± 0.23	83.49 ± 0.95	89.57 ± 0.33
		Ours	97.49 ± 0.12	91.74 ± 0.21	84.56 ± 0.76	90.43 ± 0.27
	10%	Baseline	97.68 ± 0.07	91.70 ± 0.21	84.40 ± 0.33	90.41 ± 0.23
		Ours	97.77 ± 0.07	92.43 ± 0.19	85.54 ± 0.25	91.03 ± 0.3
	100%	Baseline	98.11	93.21	87.15	92.4
		Ours	98.39	93.57	87.37	92.5
CelebAMask-HQ	0.2%	Baseline	90.86 ± 0.23	71.72 ± 1.98	63.30 ± 4.12	73.23 ± 1.63
		Ours	91.21 ± 0.22	74.99 ± 1.76	66.48 ± 3.76	75.23 ± 1.22
	0.5%	Baseline	91.66 ± 0.13	75.65 ± 1.14	67.13 ± 3.68	76.15 ± 1.17
		Ours	91.76 ± 0.13	76.43 ± 1.07	68.34 ± 3.32	77.01 ± 1.03
	1%	Baseline	92.51 ± 0.06	78.51 ± 0.83	70.47 ± 3.11	79.61 ± 0.91
		Ours	92.69 ± 0.06	79.05 ± 0.76	71.24 ± 2.73	79.88 ± 0.98
	5%	Baseline	93.67 ± 0.06	82.12 ± 0.33	74.06 ± 1.57	81.37 ± 0.93
		Ours	93.76 ± 0.06	82.37 ± 0.31	74.33 ± 1.21	81.90 ± 0.89
	10%	Baseline	94.22 ± 0.05	82.59 ± 0.21	75.17 ± 0.77	83.04 ± 0.65
		Ours	94.91 ± 0.04	83.02 ± 0.20	75.88 ± 0.63	83.72 ± 0.57
	100%	Baseline	95.16	84.87	77.77	84.64
		Ours	95.16	87.29	79.03	86.73

Table 2. Performance comparison between our approach and state-of-the-art methods on the LaPa test set in terms of F1 score. Bold font indicates the best results in different categories.

Methods	Skin	Hair	L-Eye	R-Eye	U-Lip	I-Mouth	L-Lip	Nose	L-Brow	R-Brow	Mean
Zhao [43]	93.5	94.1	86.3	86.0	83.6	86.9	84.7	94.8	86.8	86.9	88.4
Liu [17]	97.2	96.3	88.1	88.0	84.4	87.6	85.7	95.5	87.7	87.6	89.8
Te [18]	97.3	96.2	89.5	90.0	88.1	90.0	89.0	97.1	86.5	87.0	91.1
Luo [19]	95.8	94.3	87.0	89.1	85.3	85.6	88.8	94.3	85.9	86.1	89.2
Wei [44]	96.1	95.1	88.9	87.5	83.1	89.2	83.8	96.1	86.0	87.8	89.4
Te [20]	**97.7**	**96.5**	91.6	**91.1**	88.5	**90.7**	**90.1**	**97.3**	89.9	90.0	92.3
Baseline	97.6	96.3	91.9	91.4	**89.1**	90.2	89.3	97.0	90.7	90.7	92.4
Ours	97.6	96.4	**92.5**	**92.1**	88.2	89.8	89.3	97.2	**91.5**	**90.8**	**92.5**

Table 3. Performance comparison between our approach and state-of-the-art methods on the CelebAMask-HQ test set in terms of F1 score. Bold font indicates the best results in different categories.

Methods	Face	Nose	Glasses	L-Eye	R-Eye	L-Brow	R-Brow	L-Ear	R-Ear	I-Mouth	U-Lip	L-Lip	Hair	Hat	Earring	Necklace	Neck	Cloth	Mean
Zhao [43]	94.8	90.3	75.8	79.9	80.1	77.3	78.0	75.6	73.1	89.8	87.1	88.8	90.4	58.2	65.7	19.4	82.7	64.2	76.2
Lee [6]	95.5	85.6	**92.9**	84.3	85.2	81.4	81.2	84.9	83.1	63.4	88.9	90.1	86.6	**91.3**	63.2	26.1	**92.8**	68.3	80.3
Luo [19]	96.0	93.7	90.6	86.2	86.5	83.2	83.1	86.5	84.1	93.8	88.6	90.3	93.9	85.9	67.8	30.1	88.8	83.5	84.0
Wei [44]	96.4	91.9	89.5	87.1	85.0	80.8	82.5	84.1	83.3	90.6	87.9	91.0	91.1	83.9	65.4	17.8	88.1	80.6	82.1
Te [18]	96.2	94	92.3	88.6	88.7	85.7	85.2	88	85.7	**95.0**	88.9	91.2	94.9	87.6	68.3	27.6	89.4	85.3	85.1
Te [20]	96.5	93.9	91.8	88.7	89.1	85.5	85.6	88.1	**88.7**	92.0	89.1	91.1	95.2	87.2	69.6	32.8	89.9	84.9	85.5
Baseline	96.5	**94.1**	92.8	90.2	90.3	86.4	86.4	**88.9**	88.6	92.7	**89.8**	**91.4**	**95.5**	87.1	72.5	2.2	91.2	**87.0**	84.6
Ours	**96.6**	**94.1**	**92.9**	**90.3**	**90.4**	**86.6**	**86.6**	88.7	88.6	92.8	**89.8**	**91.4**	**95.5**	87.8	**72.8**	**37.1**	91.1	86.7	**86.7**

4.3. Effects of Masking Patches from Central Area

To verify the effectiveness of masking patches from the central area of the image, we compare with the setting that masks patches from the whole image. Figure 3 shows the reconstructing results from two settings on two benchmarks with masked self-supervised learning, which illustrates that both of these two settings can reconstruct the masked images effectively. However, the fine-tuning results on the face parsing task show the performance gap between these two settings. As shown in Table 4, masking patches from the central area can make the model better capture facial feature representations, which is more suitable for the face parsing task.

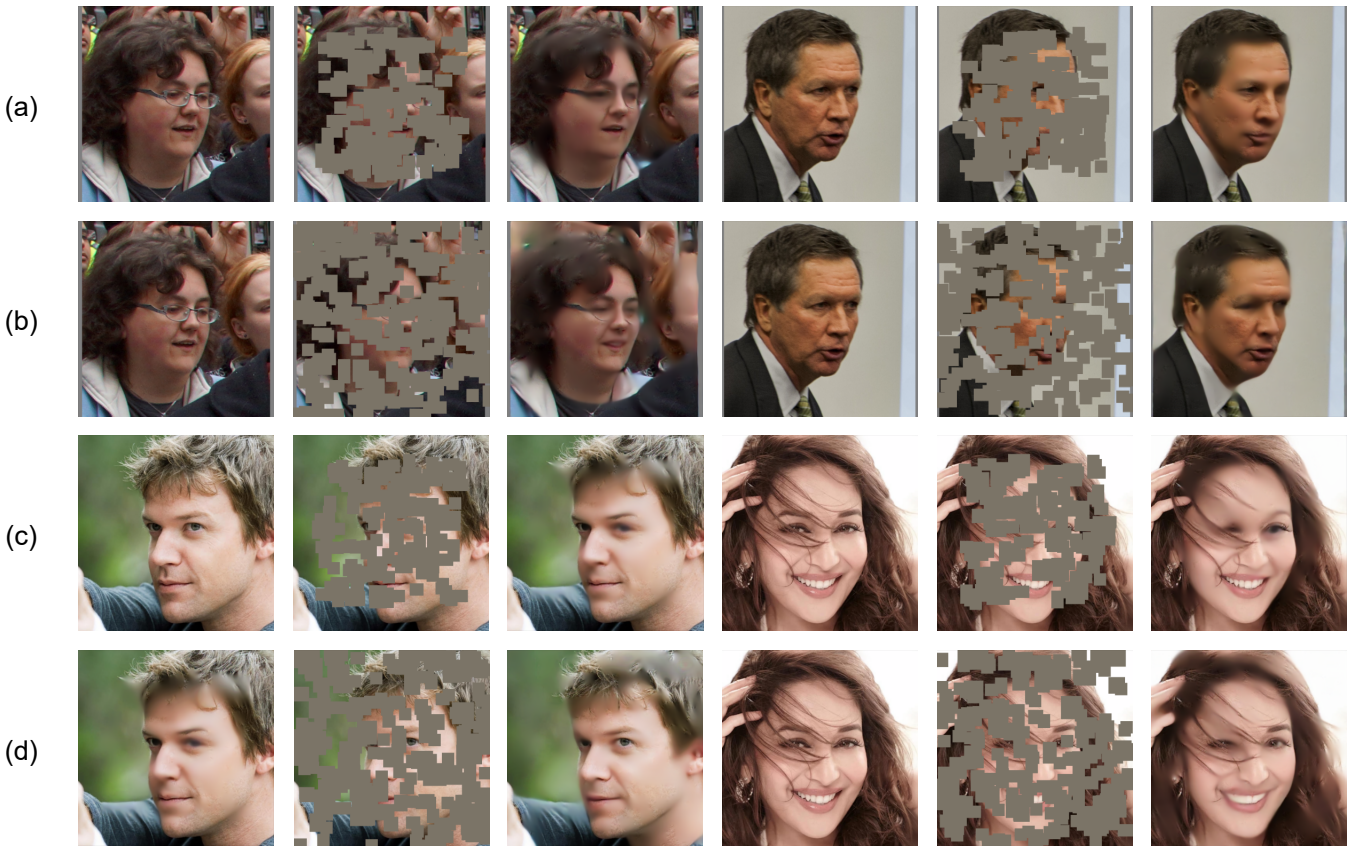


Figure 3. Example reconstruction results on Lapa (**a**,**b**) and CelebAMask-HQ (**c**,**d**) images. The first column shows the original images. The second column is the masked images where patches from the center area and the whole area are masked, respectively. The third column shows the corresponding reconstructed images.

Table 4. Performance comparison between masking patches from the central area and masking patches from the whole area of the image during self-supervised pretraining in terms of mIoU on the LaPa test set. Δ denotes the mIoU gain based on the baseline. Integral-Masked denotes masking patches from the whole area of the image.

Protocol	Method	Mean IoU	Δ
0.2%	Baseline	69.39 ± 3.21	-
	Integral-Masked	68.12 ± 3.11	−1.27
	Ours	74.41 ± 2.89	5.02
0.5%	Baseline	76.33 ± 2.32	-
	Integral-Masked	76.56 ± 2.37	0.23
	Ours	79.24 ± 2.13	2.91
100%	Baseline	87.15	-
	Integral-Masked	87.17	0.02
	Ours	87.37	0.22

4.4. Fine-Tuning Process

In our baseline training, the encoder is pretrained on the ImageNet dataset. Figure 4 shows that without the self-supervised learning, the loss has a faster convergence speed, and mIoU gets better rapidly. However, the proposed method guides the model to achieve better performance in the last. It shows that the original model is easily trapped in local optimization. In other words, the masked self-supervised learning helps the model overstep the local optimum.

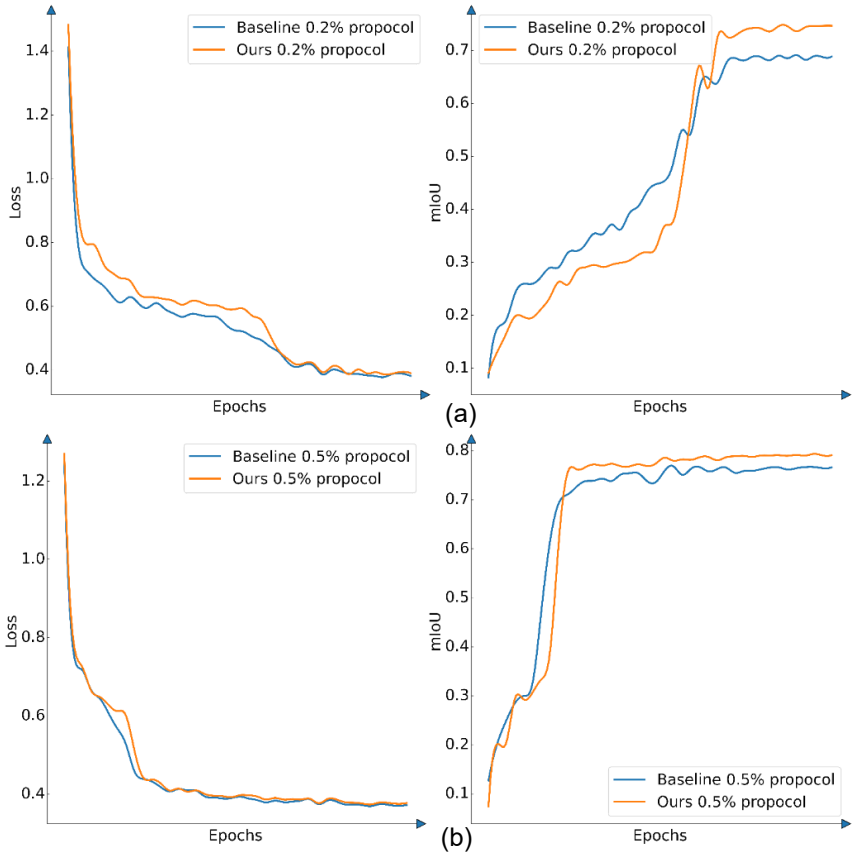


Figure 4. (**a**,**b**) show the difference between the baseline and our method in training loss and test set mIoU with 0.2 and 0.5% protocol training data. As for the mIoU performance, our method has worse performance but ends with better mIoU consistently.

4.5. Discussions

We conduct a deep exploration on how masked self-supervised pretraining is effective. Deep Feature Aggregation Network (DFANet) [45], Depth-wise Asymmetric Bottleneck Network (DABNet) [46], and UNet [47] from Luo [19] obtained poor performance on the necklace part (0.00, 0.01, 0.00 mIoU, respectively) in CelebAMask-HQ. Our baseline also performs badly on this part, as shown in Table 5. Analyzing the statistics of each facial part in CelebAMask-HQ, we found that the pixel that belongs to the necklace only takes up 0.017% of all pixels. Without bells and whistles, models place more optimization on other categories training with semantic masks merely and eventually fall into a local optimum. As a result, categories with very few pixels, such as necklace, can not be trained effectively. Figure 5 shows all feature activation maps with the size of $16 \times 512 \times 512$ from the Encoder output. The activation maps from the baseline have no activation on the necklace part seen from Figure 5e. Nevertheless, in our proposed self-supervised pretraining method, the model is required to reconstruct the image, and the whole training procedure is categorized independently. Accordingly, it makes the model focus on each category fairly. With the proposed masked self-supervised pretraining, our model has the power to obtain the necklace feature representation and acquire the feature activation on the corresponding locations, as shown in Figure 5f.

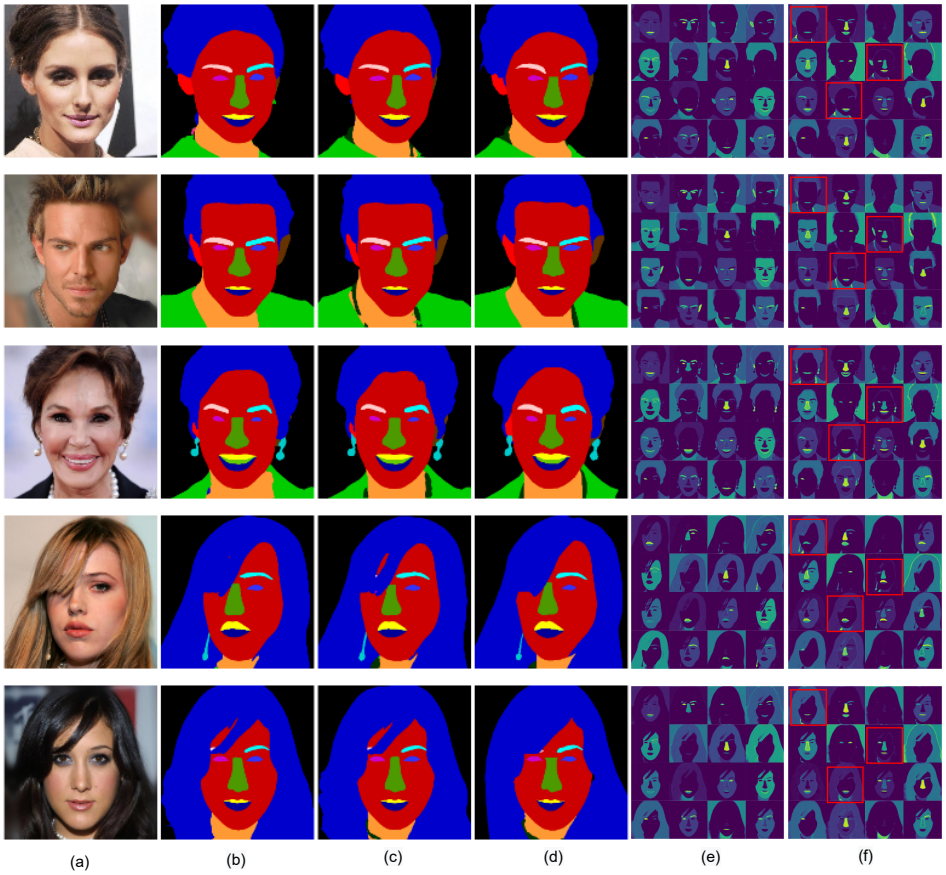


Figure 5. General image parsing results and feature activations on CelebAMask-HQ test set. (**a**): Original images. (**b**): Unet++ with Imagenet pretrained (baseline) parsing results. (**c**): Unet++ pretrained with the proposed self-supervised method. (**d**): Ground truth. (**e**): Encoder feature activation from baseline. (**f**): Encoder feature activation from ours. The baseline model obtains poor performance with necklace, and there is no necklace feature activation. The model pretrained with our methods shows better results. The first two rows even have better results than the ground truth. From the red box in the last columns, our pretraining method helps the model build necklace feature representation correctly.

Table 5. Pixel ratios of face-part categories on the CelebAMask-HQ train set in terms of F1 score.

Face part	Face	Nose	Glasses	L-Eye	R-Eye	L-Brow
Pixel Ratios (%)	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
Pixel Ratios (%)	0.31	0.90	0.24	0.017	4.10	3.35

5. Conclusions

In this paper, we propose a novel self-supervised pretraining method to alleviate the burden of manual labeling on dense facial parts annotations. In our proposed method, the Unet++ is first pretrained on the masked images where the patches from the central area of images are masked with the objective of reconstructing the masked images. After pretraining, our model is fine-tuned on the target face parsing dataset. The experimental results demonstrate that our proposed self-supervised pretraining method can consistently improve the strong baseline (directly trained on the labeled data with ImageNet pretraining) fine-tuned on different proportions of labeled data. Moreover, our method achieves the new state-of-the-art performance on the LaPa dataset and CelebAMask-HQ dataset. Furthermore, the new proposed method helps the model build comprehensive feature representation. The fine-tuned model obtains feature activations accurately on every category through the feature visualization, including ones with very small ratios. Experiment shows that a much better parsing performance is achieved, especially on categories with very small ratios (such as the necklace in CelebAMask-HQ). We think this masked self-supervised pretraining method can also be used for other face-related tasks, e.g., face landmark detection, face generation, and face attribute learning. This will be our future work to analyze the potential effectiveness of the self-supervised pretraining for limited labeled data.

Author Contributions: Conceptualization, Z.L. and H.W.; Funding acquisition, H.W.; Investigation, L.C.; Methodology, Z.L.; Project administration, Z.L. and L.X.; Software, Z.L.; Supervision, H.W.; Writing—original draft, Z.L. and L.C.; Writing—review & editing, L.C. and L.X. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China under Grant 61973337 and the U.S. National Science Foundation's BEACON Center for the Study of Evolution in Action, under Cooperative Agreement DBI-0939454.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Masi, I.; Wu, Y.; Hassner, T.; Natarajan, P. Deep Face Recognition: A Survey. In Proceedings of the 2018 31st SIBGRAPI Conference on 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 Proceedings of 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 the IEEE/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 Faces using 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. *IEEE Signal 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. In Proceedings 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 IEEE Conference 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 Vision Conference, 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 IEEE Conference 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 Neural Networks*; 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 of the 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 IEEE International 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 *European Conference 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 on Computer 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. In Proceedings 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 Active Learning. *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 useful representations 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 the IEEE 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 International Conference 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 Proceedings of 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 Vision and 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 $\times$ 16 Words: Transformers for Image Recognition at Scale. In Proceedings of the International 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-supervised learning 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 International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 2794–2802.

34. Misra, I.; Zitnick, C.L.; Hebert, M. Shuffle and learn: Unsupervised learning using temporal order verification. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2016; Springer: Berlin/Heidelberg, Germany, 2016; pp. 527–544.
35. Wu, J.; Wang, X.; Wang, W.Y. Self-Supervised Dialogue Learning. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; pp. 3857–3867.
36. Oord, A.; Li, Y.; Vinyals, O. Representation learning with contrastive predictive coding. *arXiv* **2018**. arXiv:1807.03748
37. Tian, Y.; Krishnan, D.; Isola, P. Contrastive multiview coding. In Proceedings of the Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 776–794.
38. Chen, T.; Kornblith, S.; Norouzi, M.; Hinton, G. A simple framework for contrastive learning of visual representations. In Proceedings of the International Conference on Machine Learning, Virtual, 13–18 July 2020; PMLR: New York, NY, USA, 2020; pp. 1597–1607.
39. Wu, Z.; Xiong, Y.; Yu, S.X.; Lin, D. Unsupervised feature learning via non-parametric instance discrimination. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 3733–3742.
40. He, K.; Fan, H.; Wu, Y.; Xie, S.; Girshick, R. Momentum contrast for unsupervised visual representation learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 9729–9738.
41. Noroozi, M.; Favaro, P. Unsupervised learning of visual representations by solving jigsaw puzzles. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; Springer: Berlin/Heidelberg, Germany, 2016; pp. 69–84.
42. Zhou, Z.; Siddiquee, M.M.R.; Tajbakhsh, N.; Liang, J. Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. *IEEE Trans. Med. Imaging* **2019**, *39*, 1856–1867. [CrossRef] [PubMed]
43. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid scene parsing network. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2881–2890.
44. Wei, Z.; Liu, S.; Sun, Y.; Ling, H. Accurate facial image parsing at real-time speed. *IEEE Trans. Image Process.* **2019**, *28*, 4659–4670. [CrossRef] [PubMed]
45. Li, H.; Xiong, P.; Fan, H.; Sun, J. Dfanet: Deep feature aggregation for real-time semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 9522–9531.
46. Li, G.; Yun, I.; Kim, J.; Kim, J. Dabnet: Depth-wise asymmetric bottleneck for real-time semantic segmentation. *arXiv* **2019**, arXiv:1907.11357.
47. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015; Springer: Berlin/Heidelberg, Germany, 2015; pp. 234–241.