Contrastive and Consistent Learning for Unsupervised Human Parsing

Xiaomei Zhang[1], Feng Pan[3], Ke Xiang[3], Xiangyu Zhu[1(✉)], Chang Yu[1,2], Zidu Wang[1,2], and Zhen Lei[1,2]

[1] CBSR&NLPR, CASIA, Beijing, China
{xiaomei.zhang,xiangyu.zhu,chang.yu,zlei}@nlpr.ia.ac.cn, wangzidu2022@ia.ac.cn
[2] School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
[3] Zhejiang Sunny Optical Intelligence Technology Co., Ltd., Yuyao, China
{fpan,xiangke}@sunnyoptical.com

Abstract. How to learn pixel-level representations of human parts without supervision is a challenging task. However, despite its significance, a few works explore this challenge. In this work, we propose a contrastive and consistent learning network (C^2L) for unsupervised human parsing. C^2L mainly consists of a part contrastive module and a pixel consistent module. We design a part contrastive module to distinguish the same semantic human parts from other ones by contrastive learning, which pulls the same semantic parts closer and pushes different semantic ones away. A pixel consistent module is proposed to obtain spatial correspondence in each view of images, which can select semantic-relevant image pixels and suppress semantic-irrelevant ones. To improve the pattern analysis ability, we perform a sparse operation on the feed-forward networks of the pixel consistent module. Extensive experiments on the popular human parsing benchmark show that our method achieves competitive performance.

Keywords: Unsupervised human parsing · Part contrastive module · Pixel consistent module

1 Introduction

Human parsing aims to assign a class label to each pixel of the human body in an image. Various applications make use of it, including human behavior analysis, clothing style recognition and retrieval, clothing category classification and so on. However, most works focus on supervised methods. A major drawback of supervised methods is that they need pixel-wise semantic labels for every image in a dataset. These datasets are a labor-intensive process that spends significant amounts of time and money. To remedy this situation, weakly-supervised methods employ weaker forms of supervision, *e.g.*, image-level labels [1], bounding boxes [2] and scribbles [3], and semi-supervised methods use partially labeled

W. Deng et al. (Eds.): CCBR 2022, LNCS 13628, pp. 226–236, 2022.
https://doi.org/10.1007/978-3-031-20233-9_23

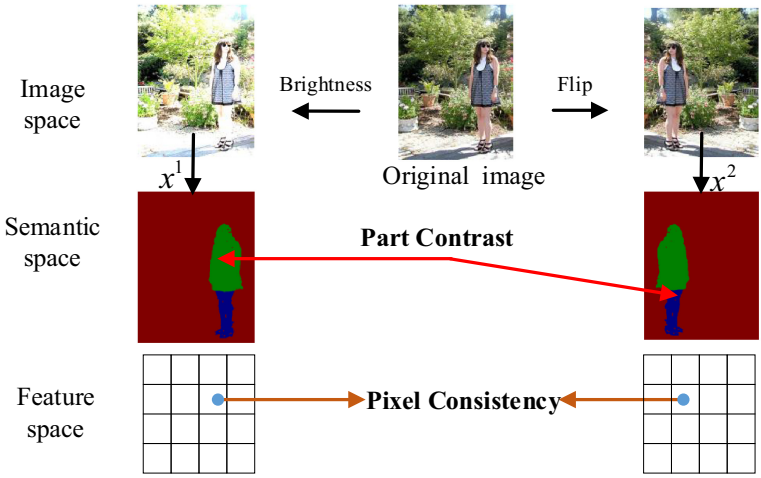

Fig. 1. An illustration of the proposed contrastive and consistent learning method for unsupervised human parsing. In this method, two views are randomly augmented from an image, and the parts with different semantics of two views are encouraged to be contrastive, and the pixels with the same semantics from the corresponding features of the two views are encouraged to be consistent.

examples to train the module. Although above methods can reduce labor consumption, training networks still rely on some form of supervision.

In this paper, we deal with this problem by introducing a novel unsupervised human parsing approach, which does not need annotated training data. More concretely, we aim to learn pixel-level representations for unsupervised human parsing by contrastive and consistent learning that consists of a part contrastive module and a pixel consistent module.

The major challenge of unsupervised human parsing is to identify part semantics. The insight of our part contrastive module is to leverage part-level representations to learn part semantics. Recently, self-supervised representation learning methods [4–6] show how to obtain the classification of images with unlabeled training data. They compute features to capture the category of a whole image, thus they cannot meet the need for the classification of parts in human parsing. Therefore, we use multiple high-level features, capturing the semantic characteristic of each human part. According to the characteristics, each pixel is assigned to its corresponding category. Figure 1 shows our motivation that the same categories (*e.g.*, upper-body) is distinguished from the other categories (*e.g.*, lower-body) by contrastive loss. In this way, the same semantic parts are pulled closer and different semantic ones are pushed away.

The part-level representations cannot be effective for dense pixel classification, because they ignore the spatial correspondence. To amend this problem, we design a pixel consistent module. As shown in Fig. 1 that the pixels with the same semantics from the corresponding features of the two views are encouraged to be consistent. The module extracts the spatial correlation in each view of images,

aiming to select semantic-relevant pixels and suppress semantic-irrelevant ones. Specifically, we first reshape input features into patches and obtain the semantic relevance of pixels. Then, we select semantic-relevant pixels and suppress semantic-irrelevant ones by a sparse operation, which makes the module pay attention to the foreground and improves the accuracy of prediction. In summary, our contributions are threefold:

1. We propose a novel contrastive and consistent learning (C^2L) network to solve the challenging unsupervised human parsing problem, which has attracted less attention in human understanding community.
2. A part contrastive module is designed to distinguish the same semantic human parts from other ones by contrastive learning, and a pixel consistent module is presented to obtain spatial correspondence of two views of input images.
3. Extensive experiments on a popular human parsing benchmark show that our method achieves competitive performance.

2 Related Work

Unsupervised Human Parsing. There have only been a few attempts in the literature to tackle human parsing under a fully unsupervised setting. Hung *et al.* [7] learned part features that are semantically consistent across images and achieved good results in their paper. Lorenz *et al.* [8] presented a method to disentangle object shape and appearance to obtain a part modeling result. Liu *et al.* [9] followed the above methods [8] to disentangle object shape and appearance and proposed a self-supervised part classification loss. Different from the above methods, our C^2L does not require predefined constraints, *e.g.*, saliency map [7], elliptical assumption of the shape of human parts [8] and background cut [9]. We are more interested in learning a model that can predict part-level semantic information without supervision.

Contrastive Learning. Contrastive learning [4,10,11] has been developing rapidly, which learns representations to discriminate positive image pairs (constructed from different augmentations of the same images) from dissimilar, negative image pairs. Varied strategies are proposed to choose appropriate negative pairs. In MoCo [4,5], a memory buffer and a momentum encoder were designed to provide negative samples. In SimCLR [10,11], the negative samples were the large training mini-batches. Some papers [12,13] designed methods to alleviate the bias issue caused by incorrect (false) negative images by modifying the contrastive loss function. PC^{2}Seg [14] sampled the negative examples strategically rather than changing the loss function. Compared with these attempts, we choose semantic inconsistent features as negative pairs.

3 Proposed Method

3.1 Overall Framework

In this paper, we propose a new network called Contrastive and Consistent Learning (C^2L) that aims to assign every pixel a label with unlabeled training

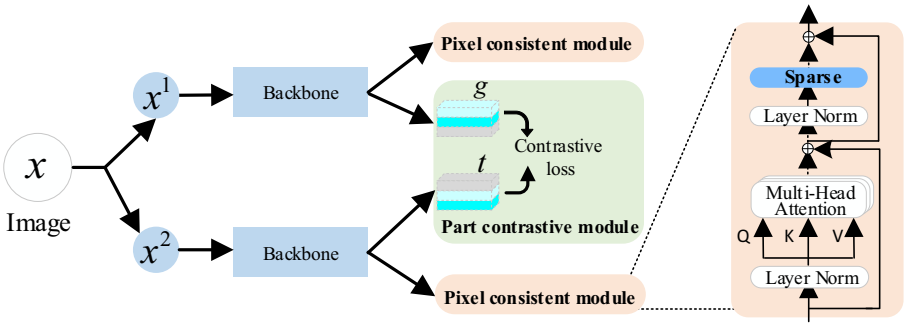

Fig. 2. Architecture of the proposed contrastive and consistent learning for unsupervised human parsing.

data. Specifically, as shown in Fig. 2, given an input image, the random photometric transforms generate two views. These views are sent into the backbone network, *e.g.*, ResNet [15] or any other convolutional neural network to obtain original features, and then we send these features to the part contrastive module and the pixel consistent module, respectively. The part contrastive module can distinguish the same semantic human parts from other ones, at the level of the global feature. The pixel consistent module can obtain spatial correspondence between pixels to adaptively aggregate semantically consistent pixels, improving the accuracy of the prediction.

3.2 Part Contrastive Module

The part contrastive module uses a 1×1 *conv* to reduce the channel number of original features and a group convolutional layer to decrease the computation, and then reshapes them to $g = \{g_0, \ldots, g_k\}$ and $t = \{t_0, \ldots, t_k\}$. We define a set of encoded keys $t_0, \ldots, t_k$ for each encoded query g_i. The encoded keys and queries are generated from different views of the input image, respectively. However, here each key and query no longer represents the whole view, and encodes a human part of a view. The positive key t_+ encodes the same part of the two views, which is one of the N feature vectors from another view of the same image. Note that N usually corresponds to the number of labels in a dataset. Hungarian-matching [16] is the sampling strategy to ensure that the positive key t_+ encodes the same semantic part with encoded query g_i. While the negative keys t_- encode the other parts of the different view. We use a contrastive loss function InfoNCE [17], it can pull g_i close to the positive key t_+ while pushing it away from other negative keys t_-:

$$\mathcal{L}_r = \sum_{i=0}^{k} -log \frac{\exp(g_i \cdot t_+/\tau)}{\exp(g_i \cdot t_+/\tau) + \sum_{t_-} \exp(g_i \cdot t_-/\tau)}, \tag{1}$$

where τ denotes a temperature hyper-parameter as in [18].

3.3 Pixel Consistent Module

The part-level representations generated by the part contrastive module cannot meet the demand of dense pixel classification, because they ignore the spatial correspondence. To implement this, we utilize the transformer encoder architecture [19]. Specifically, we use the multi-head self-attention mechanism to extract the spatial correlation of each spatial element. The transformer models the relation by refining the feature embeddings of each element with consideration to all the other elements. Formally, we reshape original features to query $Q \in R^{HW \times C}$, key $K \in R^{HW \times C}$, and value $V \in R^{HW \times C}$ which denote the input triplets of the self-attention module, where H, W and C denote height, width and channel number of the original features m, respectively. We do not use the fixed positional embeddings in the network. Then, the spatial correspondence $F \in R^{HW \times C}$ is obtained through the standard multi-head self-attention layer, with the whole process defined as $F = multi(m)$. Through end-to-end training on a human parsing dataset, the spatial correlation is obtained.

However, some works [20,21] have suggested that performing selection of spatial correlation is critical for pattern analysis. Therefore, we employ a sparse operation on the feed-forward networks (FFNs) in the transformer encoder. We begin with a brief review of sparse code algorithms. Sparse code [22] aims to learn a useful sparse representation of any given data. The mathematical representation of the general objective function for this problem can help:

$$\min_{\alpha \in R^k} \frac{1}{2} \|x - D\alpha\|_2^2 + \lambda \|\alpha\|_1, \tag{2}$$

where $x \in R^{HW \times C}$ is the given data, $D \in R^{HW \times k}$ is the decoder matrix, λ is a regularization parameter. In general, we have $k < C$, and the loss function should be small if D is "good" at representing the signal x. It is well known that L_1 loss penalty yields a sparse solution α and $\alpha \geq 0$. To prevent D from being arbitrarily large (which would lead to arbitrarily small values of α), it is common to constrain its columns $(d_i)_{j=1}^k$ to have an L_2 norm less than or equal to one. We call c the convex set of matrices verifying this constraint:

$$c = \{D \in R^{HW \times k} s.t. \forall j = 1, \ldots, k, d_j^T d_j \leq 1\}. \tag{3}$$

We perform pixel consistent operations on the feed-forward networks (FFNs) by extending the original sparse code to a spatial correlation. The spatial correlation (the outputs of the multi-head self-attention layer F) is the input of the sparse code. Thus the pixel consistent loss is defined as:

$$\mathcal{L}_c = \min_{\alpha \in R^k} \frac{1}{2} \|F - D\alpha\|_2^2 + \lambda \|\alpha\|_1, \tag{4}$$

Then, the $D\alpha$ is the output of sparse code, and it connects with the outputs of the multi-head self-attention layer by short connections as the output of our pixel consistent module.

Algorithm 1. C^2L pseudocode

P_i^1, P_i^2 : random photometric augmented version
G_i : random geometric augmented version
f_θ : the backbone
S_σ, C_τ: the pixel consistent module and the part contrastive module, respectively
for $(x_i) \sim \mathcal{D}$ do
    $y^1_{i,:}, g^1_{i,:} \leftarrow S_\sigma / C_\tau(G_i(f_\theta(P_i^1(x_i))))$
    $y^2_{i,:}, g^2_{i,:} \leftarrow S_\sigma / C_\tau(f_\theta(G_i(P_i^2(x_i))))$
end for
$\mu^1, z^1 \leftarrow BatchKMeans(y^1_{ip} : i \in [N], p \in [HW])$
$\mu^2, z^2 \leftarrow BatchKMeans(y^2_{ip} : i \in [N], p \in [HW])$
for $(x_i) \sim \mathcal{D}$ do
    $y^1_{i,:}, g^1_{i,:} \leftarrow S_\sigma / C_\tau(G_i(f_\theta(P_i^1(x_i))))$
    $y^2_{i,:}, g^2_{i,:} \leftarrow S_\sigma / C_\tau(f_\theta(G_i(P_i^2(x_i))))$
    $\mathcal{L}_r \leftarrow \mathcal{L}_{contrastive}(g^1_{i,:}, g^2_{i,:})$
    $\mathcal{L}_{view} \leftarrow \mathcal{L}_{clust}(y^1_{i,:}, \mu^1, z^1) + \mathcal{L}_{clust}(y^2_{i,:}, \mu^2, z^2)$
    $\mathcal{L}_c \leftarrow \mathcal{L}_c$
    $\mathcal{L}_{total} \leftarrow \mathcal{L}_{view} + \mathcal{L}_r + \mathcal{L}_c$
    $f_\theta, S_\sigma, C_\tau \leftarrow backward(\mathcal{L}_{total})$
end for

3.4 Pseudo Code of C^2L

Algorithm 1 provides the pseudo code of C^2L for this unsupervised method. We follow PiCIE [23], for each image x_i in the dataset. We randomly sample two photometric transformations, P_i^1 and P_i^2. And then, P_i^1 and P_i^2 are sent into the backbone f_θ and the geometric transformations G_i in different order to improve the robustness of the network. Finally, features are sent into pixel consistent module S_σ and part contrastive module C_τ, respectively. This yields two features for each pixel p in each image x_i:

$$y^1_{i,:}, g^1_{i,:} \leftarrow S_\sigma / C_\tau(G_i(f_\theta(P_i^1(x_i)))), y^2_{i,:}, g^2_{i,:} \leftarrow S_\sigma / C_\tau(f_\theta(G_i(P_i^2(x_i)))), \quad (5)$$

We employ clustering separately in the two views to get two sets of pseudo-labels and centroids:

$$\mu^1, z^1 = \arg\min_{z,\mu} \sum_{i,p} \left\| y^1_{ip} - \mu_{yip} \right\|^2, \mu^2, z^2 = \arg\min_{z,\mu} \sum_{i,p} \left\| y^2_{ip} - \mu_{yip} \right\|^2, \quad (6)$$

Given these two sets of centroid and pseudo-labels, the features are adhered to the clustering labels in a cluster loss [23]. Now that we have two views, we want this to be true in each view:

$$\mathcal{L}_{view} \leftarrow \mathcal{L}_{clust}(y^1_{i,:}, \mu^1, z^1) + \mathcal{L}_{clust}(y^2_{i,:}, \mu^2, z^2), \quad (7)$$

Overall, the total loss for our C^2L can be formulated as:

$$\mathcal{L}_{total} \leftarrow \mathcal{L}_{view} + \alpha\mathcal{L}_r + \beta\mathcal{L}_c, \quad (8)$$

where α and β are the weight to balance the two terms. α is set to 0.5 and β is set to 0.3, which is validated by experiments.

Table 1. Ablation study for every module. The baseline is PiCIE. Pacm denotes our part contrastive module. Picmnosparse denotes our pixel consistent module without the sparse code operation. Picm denotes our pixel consistent module.

#	Baseline	Pacm (ours)	Picmnosparse (ours)	Picm (ours)	mIoU (%)
1	✓				9.62
2	✓	✓			12.33
3	✓	✓	✓		14.61
4	✓	✓		✓	16.27

4 Experiments

4.1 Datasets and Evaluation Metrics

ATR dataset [24] contains 7700 multi-person images with challenging poses and viewpoints (6000 for training, 700 for validation and 1000 for testing). In this paper, we merge the ground truth to the upper-body, lower-body and background, respectively, evaluating performance. Evaluation metrics for ATR, following supervised human parsing [25,26], the performance is evaluated in terms of mean pixel Intersection-over-Union (mIoU).

4.2 Implementation Details and Baseline

For all experiments, we use the Feature Pyramid Network [27] with ResNet-18 [15] backbone pre-trained on ImageNet [28]. The fusion dimension of the feature pyramid is 128 instead of 256. Following PiCIE [23], the cluster centroids are computed with mini-batch approximation with GPUs using the FAISS library [29]. For the baseline, we do not use image gradients as an additional input when we use ImageNet-pretrained weight. For optimization, we adopt Adam. As for the crop size of the dataset, we resize images to 320×320 as the input size. The mini-batch size for k-means is 192, and the batch size for training and testing is 96.

The baseline is PiCIE [23] which is an explicit clustering method. PiCIE clusters the feature vectors of given images and uses the cluster assignment as labels to train the network. Since the size of images explodes the number of feature vectors to cluster, PiCIE applies mini-batch k-means to first compute the cluster centroids, assign labels, and then train the network.

4.3 Ablation Study

Ablation of Each Module. We conduct ablation studies with Resnet-18 as our backbone and report all the performance on the ATR validation set. For starters, we evaluate the performance of the baseline (PiCIE), as the result in Tabel 1 (#1). To verify the effect of the part contrastive module, we remove the pixel consistent module in Fig. 2. The experiment result is shown in Table 1 (#2). This modification improves the performance to 12.33%(2.71%↑) with negligible

Table 2. Ablation study for weight α and β. $\alpha = 0.5$ and $\beta = 0.3$ achieve the best prediction.

α	mIoU (%)	Acc. (%)	$\beta(\alpha = 0.5)$	mIoU (%)	Acc. (%)
0.0	9.62	34.58	0.0	15.17	68.44
0.3	12.01	58.60	**0.3**	**16.27**	**71.32**
0.5	**12.33**	**69.30**	0.5	16.06	69.02
0.7	12.06	65.21	0.7	15.33	68.51
1.0	11.63	52.13	1.0	15.32	68.05

additional parameters. We further evaluate the role of the pixel consistent module. As for this module, we replace its sparse code with normal FFNs. The result is shown in Table 1 (#3), obtaining the performance of 14.61%. We add sparse code to the pixel consistent module, the accuracy has been further improved to achieve 16.27%. Compared with the baseline, C^2L achieves a great improvement.

Ablation of Hyper-Parameters. Table 2 examines the sensitivity to hyper-parameters of C^2L. The hyper-parameter α, β in Eq. (8) serve as the weight to balance the contrastive loss and sparse code loss. We report the results of different α, β in the left and right of Table 2, respectively. We first conduct experiments to obtain the best α. In the left of Table 2, it shows a trend that the segmentation performance improves when we increase the α. when $\alpha = 0.5$, the performance achieves the best result. As shown in the right of Table 2, when $\beta = 0$, our C^2L without sparse code operation. By increasing β, $\beta = 0.3$ achieves the best prediction.

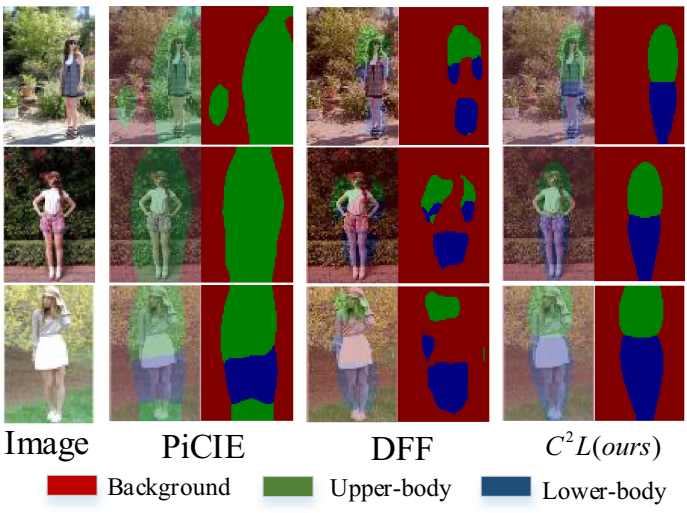

Fig. 3. Qualitative comparison results on ATR for unsupervised human parsing.

Table 3. The quantitative comparison of unsupervised human parsing on ATR.

Method	mIoU (%)	Acc. (%)
PiCIE [23]	9.62	34.58
DFF [30]	12.63	56.15
$C^2L(Ours)$	16.27	71.32

4.4 Comparison on Unsupervised Human Parsing

The unsupervised human parsing from unlabeled images is a challenge that has not been well explored. DFF [30] proposes to use non-negative matrix factorization upon the CNN features to obtain the semantic concepts, which need to optimize on the whole datasets during inference to keep semantic consistency.

To visualize the part segmentation result, we show some resulting images from ATR in Fig. 3. We can find that our method can correctly segment most parts. What's more, the foreground can be extracted from the complex background. This is because our part contrastive module can distinguish the same human parts from other ones by contrastive learning and our pixel consistent module obtains spatial correspondence in each view of images to improve foreground extracting. Results in Table 3 validate the effectiveness of our method.

5 Conclusion

In this paper, we propose contrastive and consistent learning (C^2L), a novel unsupervised human parsing method. It encourages human parts with different semantics of two views to be contrastive and the pixels from the corresponding features of the two views to be consistent. C^2L mainly consists of two modules, including a part contrastive module and a pixel consistent module. Both the quantitative and qualitative results demonstrate the superiority of C^2L.

Acknowledgement. This work was supported in part by the National Key Research & Development Program (No. 2020YFC2003901), Chinese National Natural Science Foundation Projects (No. 62206280, 62176256, 61876178, 61976229 and 62106264), the Youth Innovation Promotion Association CAS (No. Y2021131).

References

1. Pathak, D., Krahenbuhl, P., Darrell, T.: Constrained convolutional neural networks for weakly supervised segmentation. In: ICCV. (2015)
2. Dai, J., He, K., Sun, J.: Boxsup: exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In: ICCV (2015)
3. Lin, D., Dai, J., Jia, J., He, K., Sun, J.: Scribblesup: scribble-supervised convolutional networks for semantic segmentation. In: CVPR (2016)

4. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: CVPR (2020)
5. Chen, X., Fan, H., Girshick, R., He, K.: Improved baselines with momentum contrastive learning. arXiv preprint arXiv:2003.04297 (2020)
6. Wang, X., Zhang, R., Shen, C., Kong, T., Li, L.: Dense contrastive learning for self-supervised visual pre-training. In: CVPR (2021)
7. Hung, W.C., Jampani, V., Liu, S., Molchanov, P., Yang, M.H., Kautz, J.: Scops: self-supervised co-part segmentation. In: CVPR (2019)
8. Lorenz, D., Bereska, L., Milbich, T., Ommer, B.: Unsupervised part-based disentangling of object shape and appearance. In: CVPR (2019)
9. Liu, S., Zhang, L., Yang, X., Su, H., Zhu, J.: Unsupervised part segmentation through disentangling appearance and shape. In: CVPR (2021)
10. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: ICML (2020)
11. Chen, T., Kornblith, S., Swersky, K., Norouzi, M., Hinton, G.: Big self-supervised models are strong semi-supervised learners. arXiv preprint arXiv:2006.10029 (2020)
12. Chuang, C.Y., Robinson, J., Yen-Chen, L., Torralba, A., Jegelka, S.: Debiased contrastive learning. arXiv preprint arXiv:2007.00224 (2020)
13. Huynh, T., Kornblith, S., Walter, M.R., Maire, M., Khademi, M.: Boosting contrastive self-supervised learning with false negative cancellation. arXiv preprint arXiv:2011.11765 (2020)
14. Zhong, Y., Yuan, B., Wu, H., Yuan, Z., Peng, J., Wang, Y.X.: Pixel contrastive-consistent semi-supervised semantic segmentation. In: ICCV (2021)
15. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
16. Kuhn, H.W.: The Hungarian method for the assignment problem. NRL **2**(1–2), 83–97 (1955)
17. Oord, A.V.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
18. Wu, Z., Xiong, Y., Yu, S.X., Lin, D.: Unsupervised feature learning via non-parametric instance discrimination. In: CVPR (2018)
19. Vaswani, A., et al.: Attention is all you need. In: NeurIPS (2017)
20. Donoho, D.L.: Compressed sensing. TIT **52**(4), 1289–1306 (2006)
21. Wright, J., Yang, A.Y., Ganesh, A., Sastry, S.S., Ma, Y.: Robust face recognition via sparse representation. TPAMI **31**(2), 210–227 (2008)
22. Mairal, J., Bach, F., Ponce, J., Sapiro, G.: Online dictionary learning for sparse coding. In: ICML (2009)
23. Cho, J.H., Mall, U., Bala, K., Hariharan, B.: Picie: unsupervised semantic segmentation using invariance and equivariance in clustering. In: CVPR (2021)
24. Liang, X., et al.: Deep human parsing with active template regression. TPAMI **37**(12), 2402–2414 (2015)
25. Li, T., Liang, Z., Zhao, S., Gong, J., Shen, J.: Self-learning with rectification strategy for human parsing. In: CVPR (2020)
26. Yuan, Y., Chen, X., Wang, J.: Object-contextual representations for semantic segmentation. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12351, pp. 173–190. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58539-6_11
27. Lin, T.Y., Dollar, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: CVPR (2017)

28. Jia, D., Wei, D., Socher, R., Li, L.J., Kai, L., Li, F.F.: Imagenet: a large-scale hierarchical image database. In: CVPR (2009)
29. Johnson, J., Douze, M., Jégou, H.: Billion-scale similarity search with gpus (2017)
30. Collins, E., Achanta, R., Süsstrunk, S.: Deep feature factorization for concept discovery. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) Computer Vision – ECCV 2018. LNCS, vol. 11218, pp. 352–368. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01264-9_21