# Supplementary Materials

## 1. Additional Experiment Detail

**1.1. Experiment Parameter Settings.** The proposed MODC is a general framework that can adopt any multi-task deep network. In our experiments, we employ the network from three baselines (ASEN, ASEN$_{v2}$, and ASEN++) respectively. All three baselines use a ResNet50 [4] pretrained on ImageNet [2] as the shared backbone network. ASEN++ contains an additional ResNet34 backbone network for the local branch. The training processes follow the design in Section 4 in the main paper. During training, for ASEN and ASEN$_{v2}$, we apply the Adam optimizer with a learning rate of $2 \times 10^{-4}$, and the decay rate is 0.985 per epoch. For ASEN++, we follow the setting of [3]. For image augmentation, We employ the same image augmentation methods as [1]. Particularly, the scale of random image cropping and resizing is in [0.8-1.0] to prevent attribute feature loss, and the color distortion is removed for color-related attributes.

**1.2. Baseline Performance** To get the baseline performance, for ASEN++, we utilize the pretrained weights. On DARN, the performance is slightly different because we only successfully download 203,990 images due to broken URLs. For ASEN and ASEN$_{v2}$, because the pretrained weights are not provided, we train the networks using the original source code, following the original setting. However, in our reported results, the baseline may perform slightly differently from their original report. For example, the work [5] reports the overall MAP@all of ASEN on Fashion is 61.02 and ASEN and ASEN$_{v2}$ have similar performance, while we observe a slightly lower and higher performance on ASEN and ASEN$_{v2}$, respectively.

**1.3. MODC Training** The training of MODC($Net$) can be separated into three stages: warm-up stage, supervised-learning stage, and semi-supervised learning

| Model | MAP@100 for each attribute | | | | | | | | Overall |
|---|---|---|---|---|---|---|---|---|---|
| | skirt length | sleeve length | coat length | path length | collar length | lapel length | neckline length | neck | |
| *ASEN* | 72.20 | 56.74 | 58.87 | 70.48 | 75.43 | 68.57 | 61.82 | 63.73 | 64.70 |
| *MODC(ASEN)$_{top1}$* | **81.09** | **70.58** | **74.28** | **80.68** | **85.33** | **78.94** | **76.32** | **75.61** | *77.10* |
| *ASEN$_{v2}$* | 72.57 | 61.33 | 60.14 | 71.67 | 75.70 | 71.82 | 69.42 | 66.48 | 67.85 |
| *MODC(ASEN$_{v2}$)$_{top1}$* | **80.97** | **74.67** | **72.02** | **80.10** | **86.20** | **82.88** | **81.86** | **81.60** | **79.29** |
| *ASEN++* | 73.65 | 63.90 | 63.90 | 74 | 76.81 | 71.29 | 73.82 | 72.53 | 70.62 |
| *MODC(ASEN++)$_{top1}$* | **81.53** | **73.88** | **76.23** | **82.10** | **84.52** | **80.86** | **84.18** | **82.39** | **80.29** |

Table 1: Performance comparison on MAP@100 of each attribute on FashionAI.

| Model | MAP@100 for each attribute | | | | | | | | | Overall |
|---|---|---|---|---|---|---|---|---|---|---|
| | clothes category | clothes button | clothes color | clothes length | clothes pattern | clothes shape | collar shape | sleeve length | sleeve shape | |
| *ASEN* | 42.57 | 53.58 | 58.47 | 62.62 | 61.91 | 65.04 | 42.43 | 80.71 | 62.25 | 58.72 |
| *MODC(ASEN)$_{top1}$* | **57.02** | **70.60** | **69.42** | **70.18** | **71.00** | **73.04** | **56.08** | **90.09** | **68.56** | **69.45** |
| *ASEN$_{v2}$* | 43.16 | 55.87 | 56.97 | 65.23 | 62.33 | 64.78 | 44.03 | 83.08 | 62.86 | 59.66 |
| *MODC(ASEN$_{v2}$)$_{top1}$* | **55.51** | **69.99** | **68.23** | **72.13** | **72.81** | **71.16** | **56.45** | **87.99** | **70.36** | **69.25** |
| *ASEN++* | 45.12 | 56.40 | 57.71 | 65.22 | 65.08 | 65.81 | 45.47 | 85.45 | 64.93 | 61.09 |
| *MODC(ASEN++)$_{top1}$* | **59.20** | **72.08** | **67.96** | **74.56** | **74.43** | **74.59** | **61.66** | **91.14** | **75.07** | **72.16** |

Table 2: Performance comparison on MAP@100 of each attribute on DARN.

| Model | 1% labeled | | | 5% labeled | | | 50% labeled | | |
|---|---|---|---|---|---|---|---|---|---|
| | MAP@100 | MAP@all | Recall@100 | MAP@100 | MAP@all | Recall@100 | MAP@100 | MAP@all | Recall@100 |
| *ASEN* | 29 | 22.23 | 8.92 | 40.52 | 32.30 | 13.15 | 61.24 | 53.37 | 21.36 |
| *MODC(ASEN)*$_{top1}$ | **39.60** | **30.83** | **13.84** | **58.64** | **48.84** | **21.47** | **75.20** | **69.11** | **28.12** |
| *MODC(ASEN)*$_{top2}$ | 35.45 | 29.05 | 11.81 | 51.36 | 45.48 | 17.99 | 68.25 | 64.12 | 24.81 |
| *ASEN*$_{v2}$ | 30.32 | 22.86 | 9.19 | 42.85 | 34.35 | 13.92 | 64.88 | 57.63 | 22.84 |
| *MODC(ASEN*$_{v2}$*)*$_{top1}$ | **40.66** | **31.83** | **14.28** | **57.32** | **47.97** | **20.25** | **75.93** | **70.26** | **28.38** |
| *MODC(ASEN*$_{v2}$*)*$_{top2}$ | 36.52 | 30.68 | 12.44 | 50.01 | 44.50 | 17.68 | 69.41 | 65.40 | 25.33 |
| *ASEN++* | 29.33 | 22.45 | 8.93 | 39.66 | 31.44 | 12.72 | 62.86 | 55.75 | 22.21 |
| *MODC(ASEN++)*$_{top1}$ | **37.49** | **28.41** | **12.77** | **52.37** | **42.44** | **18.87** | **75.25** | **68.39** | **28.77** |
| *MODC(ASEN++)*$_{top2}$ | 33.67 | 27.35 | 11.05 | 45.27 | 39.10 | 15.70 | 67.40 | 62.88 | 24.57 |

Table 3: Overall semi-supervised learning performance comparison on all attributes of FashionAI.

stage. As introduced in Section 4, the warm-up stage only includes the instance-level loss, while the supervised-learning stage adds the cluster-level loss. The semi-supervised learning stage can be further separated into 4 steps: (1) adding self-supervision for labeled data; (2) adding self-supervision for unlabeled data to introduce unlabeled data; (3) adding instance-level loss with pseudo labels for unlabeled data; (4) adding cluster-level loss with pseudo labeled for unlabeled data. Each step is trained to converge before starting the next.

| Model | Query label | 100% labeled | | | | 10% labeled | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | MAP@100 | MAP@all | Recall@100 | Acc | MAP@100 | MAP@all | Recall@100 | Acc |
| *ASEN* | | 64.70 | 57.37 | 22.77 | - | 49.68 | 41.35 | 16.81 | - |
| *ASEN*$_{v2}$ | | 67.85 | 61.13 | 24.14 | - | **50.20** | **41.89** | **17.06** | - |
| *ASEN++* | | **70.62** | **64.27** | **25.30** | - | 45.12 | 36.66 | 14.90 | - |
| *MODC(ASEN)*$_{top1}$ | unknown | 67.00 | 60.54 | 23.62 | 71.99 | **54.64** | **46.77** | **18.78** | 61.56 |
| *MODC(ASEN*$_{v2}$*)*$_{top1}$ | unknown | 69.94 | 63.57 | 24.75 | 73.59 | 54.56 | 46.59 | 18.66 | **62.30** |
| *MODC(ASEN++)*$_{top1}$ | unknown | **71.83** | **65.73** | **25.76** | **74.58** | 50.80 | 42.49 | 17.20 | 58.07 |
| *MODC(ASEN)*$_{top1}$ | known | 77.10 | 70.02 | 28.89 | 71.99 | **65.29** | **56.64** | **24.32** | 61.56 |
| *MODC(ASEN*$_{v2}$*)*$_{top1}$ | known | 79.29 | 72.51 | 29.78 | 73.59 | 64.78 | 56.36 | 24.13 | **62.30** |
| *MODC(ASEN++)*$_{top1}$ | known | **80.29** | **74.32** | **30.26** | **74.58** | 61.61 | 52.00 | 22.73 | 58.07 |

Table 4: Overall MAP@100, MAP@all, Recall@100, and Acc (pseudo label generation accuracy) on all attributes of FashionAI. Baselines cannot generate pseudo-labels.

| Model | Query label | 100% labeled | | | | 10% labeled | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | MAP@100 | MAP@all | Recall@100 | Acc | MAP@100 | MAP@all | Recall@100 | Acc |
| *ASEN* | | 58.72 | 52.75 | 20.26 | - | 51.35 | 45.10 | 16.03 | - |
| *ASEN*$_{v2}$ | | 59.66 | 54.29 | 20.88 | - | **55.34** | **50.02** | **18.00** | - |
| *ASEN++* | | **61.09** | **55.78** | **21.51** | - | 54.83 | 49.85 | 17.63 | - |
| *MODC(ASEN)*$_{top1}$ | unknown | 59.45 | 54.57 | 20.95 | 41.74 | 55.24 | 49.65 | 18.15 | 34.64 |
| *MODC(ASEN*$_{v2}$*)*$_{top1}$ | unknown | 59.53 | 54.99 | 20.95 | 46.94 | **56.68** | **52.32** | **19.13** | **48.76** |
| *MODC(ASEN++)*$_{top1}$ | unknown | **60.88** | **56.23** | **21.94** | **50.02** | 56.23 | 51.54 | 18.70 | 47.42 |
| *MODC(ASEN)*$_{top1}$ | known | 69.45 | 59.61 | 25.36 | 41.74 | 61.67 | 53.21 | 21.53 | 34.64 |
| *MODC(ASEN*$_{v2}$*)*$_{top1}$ | known | 69.25 | 60.43 | 25.65 | 46.94 | 64.94 | **57.60** | **23.86** | **48.76** |
| *MODC(ASEN++)*$_{top1}$ | known | **72.16** | **62.56** | **26.76** | **50.02** | **65.35** | 57.07 | 23.05 | 47.42 |

Table 5: Overall MAP@100, MAP@all, Recall@100, and Acc (pseudo label generation accuracy) on all attributes of DARN.

The initiative training of ASEN and ASEN$_{v2}$ uses Adam optimizer with $2 \times 10^{-4}$ lr and 0.985 decay rate each epoch. The training on further stages and steps inherits the optimizer. The initial training of ASEN++ global branch uses Adam optimizer with $2 \times 10^{-4}$ lr and 0.9 decay rate every 3 epochs. The local branch applies a new Adam optimizer with $1 \times 10^{-5}$ lr and 0.9 decay rate every 1 epoch. The fine-tuning of ASEN++ with MODC applies a new optimizer with $1 \times 10^{-5}$ lr and 0.985 decay rate every 1 epoch.

## 2. Additional Experiment Results

**2.1. MAP@100 of Each Attribute** Table 1 and Table 2 present the MAP@100 on each attribute on FashionAI and DARN. According to the tables, MODC improves the top 100 retrieval performance on each attribute, for each network respectively. The best performers on each network are in **bold**. The global best performers are underlined.

**2.2. MODC Performance on Semi-supervised Learning** On FashionAI, we conduct various ratios of "labeled"/"unlabeled" subset split, including 1%/99%, 5%/95%, 10%/90%, 50%/50%. Except the performance comparison of 10%/90% that is shown in the main paper, the rest is shown in Table 3. The improvement is consistent with the observations in Section 5 in the main paper. Particularly, we observe MODC(ASEN++) performs better when employing more labeled data, which offers insights on why it performs the best in supervised learning but not the best in semi-supervised learning.

**2.3. MODC Performance with Unknown Query Image Labels** MODC allows us to leverage the query image labels and enable the prioritized retrieval strategy for fashion retrievals. When query image labels are known, the performance is significantly improved with MODC, as analyzed in Section 5 in the main paper. When query image labels are unknown, MODC can generate pseudo-labels. Table 4 and Table 5 demonstrate the performance of leveraging generated pseudo-labels for query images for fashion retrieval. We observe that based on generated pseudo-labels, MODC is still able to introduce improvement when applying various networks, compared with the baseline methods. We also observe that when the query image labels are unknown, the better quality of the generated pseudo-labels leads to larger improvement on the performance of fashion retrieval.

**2.4. Limitations** One of the key factors of our performance boost can be derived from the prioritized retrieval in class-specific embedding spaces. It relies on the quality of the query image labels/pseudo-labels for accurate prioritization. For fashion retrieval, query image labels are usually available. However, in some cases, the query image labels may be missing, requiring MODC to generate pseudo labels. In these cases, the quality of the generated pseudo labels affects

the fashion retrieval performance. According to Table 4 and Table 5, the performance of fashion retrieval when query image labels are known is significantly better than unknown cases, and higher accuracy of pseudo label prediction tends to yield larger performance improvement on fashion retrieval. Another way to address the missing query image label situation is to extract query image labels from the text descriptions, which requires further exploration.

## 3. Fine-grained Representation Distribution

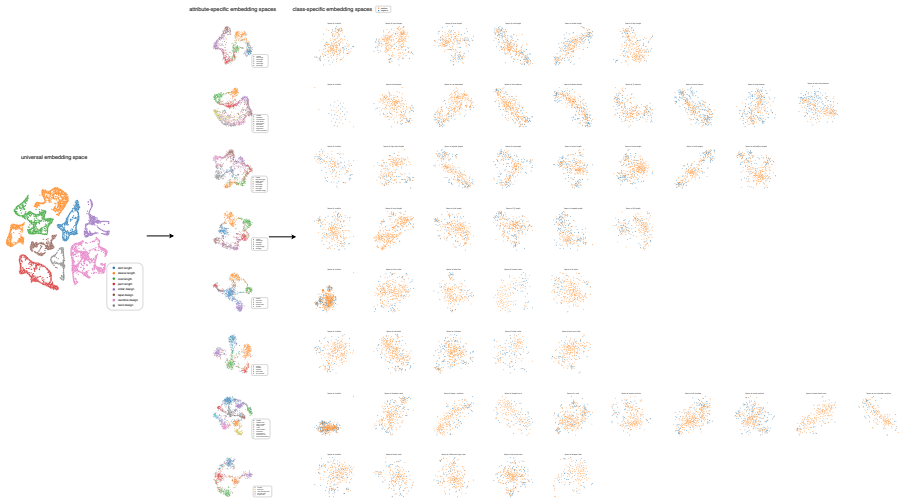


Fig. 1: The fine-grained representation distribution on universal embedding space, attribute-specific embedding spaces, and class-specific embedding spaces. The representations are conducted by MODC(ASEN$_{v2}$) on FashionAI test split. Best viewed in color and zoomed in.

The t-SNE visualization of the fine-grained representation on FashionAI test split is shown in Figure 1. It includes three-scale spaces and demonstrates that the learned fine-grained representations are well-distributed on all three-scale spaces.

## 4. More Examples of Attribute-Specific Fashion Retrieval with MODC


Fig. 2: Examples of attribute-specific fashion retrieval. The retrieval is conducted by MODC(ASEN++) on FashionAI test split. Green shows positive retrieves, while red shows negatives. Best viewed in color and zoomed in.

## References


1. Chen, X., He, K.: Exploring simple siamese representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15750–15758 (2021) 1
2. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009) 1
3. Dong, J., Ma, Z., Mao, X., Yang, X., He, Y., Hong, R., Ji, S.: Fine-grained fashion similarity prediction by attribute-specific embedding learning. arXiv preprint arXiv:2104.02429 (2021) 1
4. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016) 1
5. Ma, Z., Dong, J., Long, Z., Zhang, Y., He, Y., Xue, H., Ji, S.: Fine-grained fashion similarity learning by attribute-specific embedding network. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 11741–11748 (2020) 1