CRAFT: Complementary Recommendation by Adversarial Feature Transform

Cong Phuoc Huynh[1⋆], Arridhana Ciptadi[1⋆], Ambrish Tyagi[1], and Amit Agrawal[1]

Amazon Lab126, Sunnyvale, California, USA
{conghuyn,ambrisht,aaagrawa}@amazon.com

Abstract. We propose a framework that harnesses visual cues in an unsupervised manner to learn the co-occurrence distribution of items in real-world images for complementary recommendation. Our model learns a non-linear transformation between the two manifolds of source and target item categories (e.g., tops and bottoms in outfits). Given a large dataset of images containing instances of co-occurring items, we train a generative transformer network directly on the feature representation by casting it as an adversarial optimization problem. Such a conditional generative model can produce multiple novel samples of complementary items (in the feature space) for a given query item. We demonstrate our framework for the task of recommending complementary top apparel for a given bottom clothing item. The recommendations made by our system are diverse, and are favored by human experts over the baseline approaches.

Keywords: recommender systems · complementary recommendation · generative adversarial network · unsupervised learning · adversarial learning

1 Introduction

Recommendation algorithms are central to many commercial applications, particularly for online shopping. In domains such as fashion, customers are looking for clothing recommendations that visually complement their current outfits, styles, and wardrobe. Traditional content-based and collaborative recommendation algorithms [1, 18] do not make use of the visual cues to suggest complementary items. Among these, collaborative filtering [16, 25] is a commonly used approach, which primarily relies on behavioral and historical data such as co-purchases, co-views, and past purchases to suggest new items to customers. In contrast to these approaches, we address the problem of recommending complementary items for a given query item based on visual cues. Our proposed approach is general (modeling visual co-occurrences) and can be applied to different domains such as fashion, home design, etc.

⋆ Equal contributors

GAN Input	Generative (w/ random seed)	Output	Example
N/A	Yes	Image	Image Generation [8]
Image	No	Image	Image-to-Image Translation [35]
Image + Attribute	No	Image	Image Manipulation [17]
Synthetic Image	No	Image	Adding Realism [27]
Synthetic Image	Yes	Image	Adding Realism [2]
Image	No	Features	Domain Adaptation [29, 31]
Features	Yes	Features	Ours

Table 1. Similarities and differences between our approach and those that use adversarial loss for training.

We develop an unsupervised learning approach for Complementary Recommendation using Adversarial Feature Transform (CRAFT), by learning the co-occurrence of item pairs in real images. Here we assume that the co-occurrence frequency of item pairs is a strong indicator of the likelihood of their complementary relationship. We define an adversarial process to train a conditional generative adversarial network that can learn the joint distribution of item pairs by observing samples from the real distribution, i.e., image features of co-occurring items. Instead of direct image synthesis, the generative transformer in CRAFT is trained in the feature space and is able to generate diverse features thanks to the input random vector. The transformed feature vectors are used to recommend images corresponding to the nearest neighbors in the feature space.

The proposed feature transformation approach is novel and unique, with several advantages over existing image and feature generation methods using Generative Adversarial Network (GAN) [8]. While the quality of visual image generation using GANs has improved significantly (especially for faces [13]), it still lacks the realism required for fashion/apparel recommendation. In contrast to image transformation approaches using GAN approach [2, 12, 27, 35], which operate in the image space, CRAFT directly generates features of the recommended items. Therefore, it bypasses the need for generating synthetic images and enables a simpler and more efficient network architecture. This improves the stability of CRAFT during training and avoids common pitfalls such as mode collapse [3]. Another advantage is that our generative model can provide multiple complementary items by learning the joint distribution in the feature space, rather than a fixed mapping provided by image translation approaches.

2 Related Work

Generative Adversarial Networks (GAN): The original GAN [8] and variants [7] have been presented as a powerful framework for learning generative models of complex data distributions for various tasks including image generation [13, 33], image-to-image translation [12, 35], domain adaptation [2, 27, 29, 31], etc. A recent work by Zhu et al. [36] used the GAN framework to generate new clothing on a wearer. Our approach differs from these methods since we

do not aim to generate an image of the complementary item. Instead, we use the adversarial training framework to learn the joint distribution between the source and target *features* in an *unsupervised* manner. The GAN paradigm has also found applications in the areas of image manipulation and image transformation [2, 12, 27, 35]. While such an approach can be applied to transform a given image into that of a complementary item, it only provides a fixed mapping. In contrast, our method adopts a generative model that can provide multiple complementary items by learning the joint distribution in the feature space. Further, contrary to methods such as CycleGAN [35] and Zhu et al. [36] that perform image-to-image translation using raw pixels, our approach works directly in the feature space. Feature-based domain adaptation approaches attempt to directly learn a visual encoder for the target domain [29] or a domain-invariant encoder [31] through optimizing an adversarial loss defined on the source, target and augmented features. In contrast, our method learns a generative transformer network that operates in the feature space. Table 2 shows similarities and differences between our approach and those that use adversarial loss for training.

Unsupervised Learning: Recent applications of unsupervised learning for visual tasks include object discovery in videos [4]. In addition, there have been demonstrations of self-supervised learning [6] for the tasks of image colorization, image in-painting, hole filling, jigsaw puzzle solving from image patches, future frame prediction using video data [32], etc. In the fashion domain, annotated data are typically used for predicting fashion-related attributes and matching street-to-catalog images [14, 21]. These approaches involve visual search to find *similar* looking items, whereas our approach is focused on finding complementary items. Furthermore, our approach is unsupervised: we only take as input a set of images to learn the feature transformation between complementary objects.

Recommendation: There is a rich body of literature on using behavioral customer data such as browsing and purchasing history to develop recommender systems [16]. Specific to the fashion domain, McAuley et al. [24] employed convolution neural network (CNN) features and non-visual data to build a personalized model of user's preference. In [10, 30], the authors proposed to learn visual compatibility via a common embedding across categories. In [9], the authors proposed to learn a bi-directional Long Short Term Memory (LSTM) model in a supervised manner, to suggest items that complement each other in an entire outfit.

The aforementioned recommendation approaches use customer's behavioral data as training labels. Behavioral signals do not necessarily reflect that items viewed or purchased together are visually complementary. In contrast, our unsupervised approach learns item co-occurrences from only visual data. In multiple methods [9, 24, 30], the recommendation model is non-generative in the sense that it can only evaluate the compatibility between two given items. In [10], the diversity of recommendation is limited by the (fixed) number of embeddings employed. In contrast, our *generative* model is not subject to such a constraint, thanks to its ability to sample an infinite amount of noise vectors.

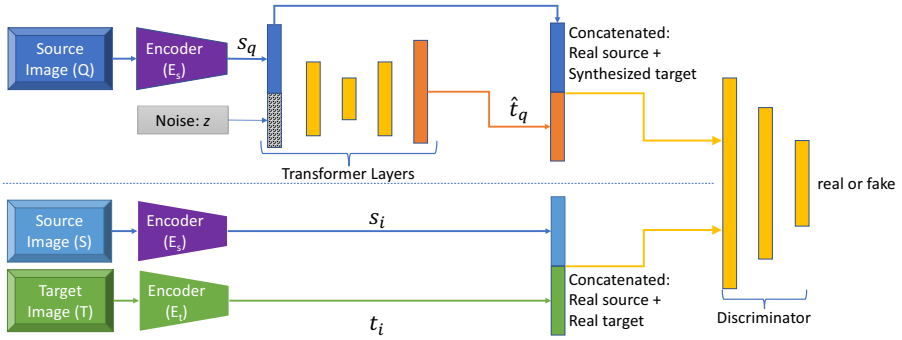


Fig. 1. Architecture for the CRAFT framework. The transformer component is trained to generate the target features conditioned on the source features and a sampled noise vector.

3 Generative Feature Transformer

This section describes our generative recommendation model based on the co-occurrence of item pairs in real-world images. We hypothesize that learning the joint distribution of such pairs can be useful for recommending new items that complement a given query. We adopt an adversarial learning paradigm, where our transformer network learns to generate features of the complementary items conditioned on the query item.

3.1 Network Architecture

Figure 1 depicts the overall architecture of the proposed CRAFT network. The source and the target feature encoders, E_s and E_t, respectively, are fixed and are used to generate feature vectors for training and inference. Typically, it is advisable to use application-specific feature representations, e.g., apparel feature embeddings for clothing recommendations. However, a general representation such as one pre-trained on ImageNet [5] or MS-COCO [20] offer robust alternatives.

Our architecture resembles traditional GAN designs with two main components: a conditional feature transformer and a discriminator. The role of the feature transformer is to transform the source feature s_q into a complementary target feature $\hat{t}_q$. The input to the transformer also consists of a random noise vector z sampled uniformly from a unit sphere in a d_z-dimensional space. By design, the transformer is generative since it is able to sample various features in the target domain.

As discussed, since our approach works in the feature space, we can adopt a simple architecture for the feature transformer and discriminator. The transformer consists of several fully-connected layers, each followed by batch normalization [11] and leaky ReLU [22] activation layers. The discriminator is commensurate to the transformer in capacity, consisting of the same number of layers.

This helps balance the power between the transformer and the discriminator in the two-player game, leading to stable training and convergence.

3.2 Training & Generating Recommendations

Our training data consists of N co-occurring feature pairs $\mathcal{C} = \{(s_i, t_i), i = 1, \ldots, N\}$, where $s_i \in \mathbb{R}^{d_s}$ and $t_i \in \mathbb{R}^{d_t}$ denote the features corresponding to the source and the target images, respectively. Given a sample s_q from the source space, the complementary recommendation task is to generate target features $\{\hat{t}_q\}$ that maximizes the likelihood that the pair $(s_q, \hat{t}_q)$ belongs to the joint distribution $p_{\mathcal{C}}$ represented by the training data. Note that the source features fed into the transformer (s_q) and discriminator (s_i) are generally different from each other. To this end, we model the composition of layers in the feature transformer and the discriminator as two functions $T_\phi(s, z) : (s, z) \mapsto \hat{t}$ and $D_\theta(s, t) : (s, t) \mapsto [0, 1]$, respectively. Here, ϕ and θ are the learnable parameters of the two players, transformer and discriminator, respectively, and (s, t) is a pair of source and target feature vectors, and z is a random noise vector.

The training process emulates an adversarial game between the feature transformer and the discriminator, where the discriminator aims to classify feature pairs as real (co-occurring) or synthetic. On the other hand, the feature transformer synthesizes target features $\{\hat{t}_q\}$ conditioned on a given source feature s_q. Its objective is to fool the discriminator into the belief that $\hat{t}_q$ co-occurs with s_q. The feedback from the discriminator encourages the transformer to produce a target feature $\hat{t}_q$ so as to maximize the co-occurrence probability of the synthetic pair.

The adversarial game can be formulated as a mini-max optimization problem. The optimization approach can be implemented by alternating the training of the discriminator and the feature transformer. The overall objective function of the adversarial training process is formulated in Equation 1.

$$\min_{\phi} \max_{\theta} \mathcal{L} \triangleq \mathbb{E}_{(s_i,t_i)\sim p_{\mathcal{C}}} \log D_\theta(s_i, t_i) + \mathbb{E}_{z\sim p_z, s_q\sim p_s} \log(1 - D_\theta(s_q, T_\phi(s_q, z))), \tag{1}$$

where p_z and p_s are the probability distribution function (pdf) of the random noise and the source feature.

In the discriminator step (D-step), the discriminator's goal is to assign a binary label, i.e., 0 to the synthesized feature pair $(s_q, \hat{t}_q)$, where $\hat{t}_q = T_\phi(s_q, z)$, and 1 to an actual pair (s_i, t_i). The discriminator's goal is to maximizes the cross entropy loss in Equation 1. Meanwhile, the feature transformer maximizes the likelihood that the discriminator recognizes synthetic pairs as belonging to the data-generating (joint) distribution $\mathcal{C}$, i.e., assigning a label 1 to such pairs. Therefore, the transformer step (T-step) aims to minimize the second term on the right-hand side of Equation 1.

3.3 Generating Recommendations

The recommendation workflow is depicted in Figure 3.3. Here, we retain only the transformer's layers shown in Figure 1 for recommendation. From a given

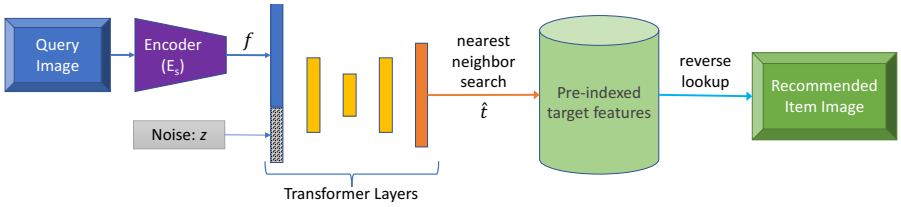


Fig. 2. Generating recommendations using the proposed CRAFT network.

query image, we first extract its features via a pre-trained source encoder, E_s. The query feature, s_q, along with a sampled noise vector, z, is fed to the CRAFT network to generate the feature vector $\hat{t}_q$ for a complementary item. This allows us to generate a diverse set of complementary recommendations by sampling the underlying conditional probability distribution function. We then perform a nearest neighbor search on a pre-indexed candidate subspace with $\hat{t}_q$ as the query vector. Actual recommendation images are retrieved by a reverse lookup that maps the selected features to the original target images.

4 Experiments

In this section, we describe the experimental setup and results of applying CRAFT to the problem of complementary apparel recommendation. Specifically, we train the generative transformer network to synthesize features for top clothing items that are visually compatible to a given query bottom item.

4.1 Datasets

We trained the proposed CRAFT network from scratch on unlabeled images, without the need for any human annotation to determine the complementary relationship. To achieve this, we collected 474,184 full-length outfit images of fashion enthusiasts, each containing a top and a bottom item. From each outfit, we extract the regions of interest (ROIs) of co-occurring pairs of tops and bottoms for training. We trained a semantic segmentation network [34] on the "Human Parsing in the Wild" dataset [19] to parse the collected images into top and bottom segments. We consolidate the original labels in the dataset into 15 labels, where top clothing items correspond to the label "upper-clothes" and bottom ones correspond to "pants" and "skirt". Subsequently, we obtained tight bounding boxes/regions of interest (ROIs) around the resulting top and bottom segments. In this manner, the training pairs are obtained automatically without the need for manual annotations of the complementary relationship.

In our experiments, we extract the global averaging pooling feature of the pre-trained Inception-v4 model [28] after performing a forward pass on the top and bottom ROIs. Rather than working in the original 1536-dimensional feature space, we opt for the top 128 PCA components to stabilize the training and reduce the computational load.

4.2 Training and Network Parameters

We use the Adam optimizer [15] with starting learning rate of 0.0002 for both the discriminator and the transformer networks. To improve training stability, we use one-sided label noise [26]. Each minibatch for training the discriminator consists of an equal proportion of synthetic and real feature pairs. The transformer is composed of 3 fully connected layers with 256 channels in the first two layers and 128 channels in the third. The discriminator is composed of 3 fully connected layers with 256, 256, and 1 channel(s), respectively. The noise vector z is uniformly sampled from the unit sphere in $\mathbb{R}^{128}$. We use leaky ReLU ($\alpha = 0.2$) and batch normalization for the first two layers of both the transformer and the discriminator.

4.3 Baseline Algorithms

We compare the CRAFT algorithm with the following baseline methods.

Random recommendations: A trivial baseline generates random recommendations from a given set of candidate options, referred to as *Random*. A random selection can offer a diverse set of target items, but they may not necessarily be complementary to the query item.

Nearest neighbors of source items: In addition, we consider a relevant and good baseline method, which operates by finding nearest neighbors of the query/source feature, and recommend the *corresponding* target items, i.e. the one that co-occurs with the neighboring source items in the training data. We refer to this method as *NN-Source*.

Incompatible recommendations: Lastly, we illustrate that CRAFT not only learns to recommend complementary items, but also the concept of visual incompatibility. The *Incompatible* recommendation method suggests tops that are assigned low discriminator scores by the generative transformer of CRAFT.

4.4 Visualization of The Discriminator Output

In this section, we visualize how the learned transformer network *dynamically* reacts to given queries in terms of assigning compatibility scores for candidate tops. To visualize the space of candidate top items, we projected them to a two-dimensional (2D) subspace using t-SNE [23]. The discriminator output can be seen as a proxy for the compatibility score between any top and a given query item. The left-hand side of each subplot in Figure 3 shows 2D embedding of all the tops in the dataset, color coded by the discriminator/compatibility score for each top with the given bottom item (shown on the right-hand side). Note that the compatibility scores for candidate tops change with the query bottom. The yellow colors in the t-SNE plot denote low compatibility, while shades of orange to red denote high compatibility (see color bar). It is interesting to note how universal items such as blue jeans or gray pants are compatible with a large set of candidate tops, while rare bottoms like the richly textured pattern skirt shown on the bottom row are compatible with only a handful of tops. This illustrates that our network is able to model the distribution of real item pairs.

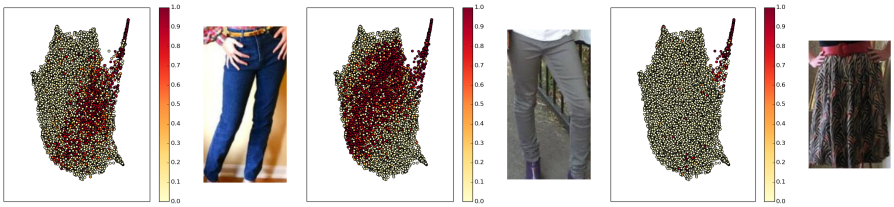


Fig. 3. Each subplot shows a 2D t-SNE embedding of all the candidate tops (left) with the corresponding query image (right). The colors represent the discriminator score for tops conditioned on the query (red: high score, yellow: low score). Note that the discriminator is able to learn that common bottoms such as blue jeans and gray pants are compatible with a wide range of tops as compared to rarer query items such as the patterned skirt shown in the last subplot.

4.5 Qualitative Results

Figure 4 shows qualitative results of the different recommendation methods for two query items. For this experiment, we generated 8 top recommendations from each algorithm and asked a fashion specialist to identify the top items that complement the given bottom query. While all of the approaches produce visually diverse recommendations, not all of them are compatible with the query. For a common bottom outfit such as dark jeans (Figure 4(a)), NN-Source perform as well as our algorithm (CRAFT), while for a less common bottom such as bright pink skirt (Figure 4(b)) they perform worse (see Section 4.7 for a more thorough analysis). This is aligned with our intuition that the quality of NN-Source recommendation highly depends on the proximity of the neighbors of the query. Interestingly, the *Incompatible* algorithm demonstrates its ability to learn the concept of visual *incompatibility*: it often produces unusual outfit recommendation (e.g., the fur top as the third item in Figure 4(a)).

4.6 User Study Design

When recommendations are provided from an open ended set, they are difficult to evaluate in absolute terms. For subjective domains such as fashion, it is preferable to obtain input from domain experts who are familiar with nuances involved in making style-appropriate recommendations. We adopt A/B testing as the main methodology to compare our proposed approach to the baselines. Here, we evaluate the relevance of recommendations generated by each algorithm by measuring their *acceptance* by domain experts.

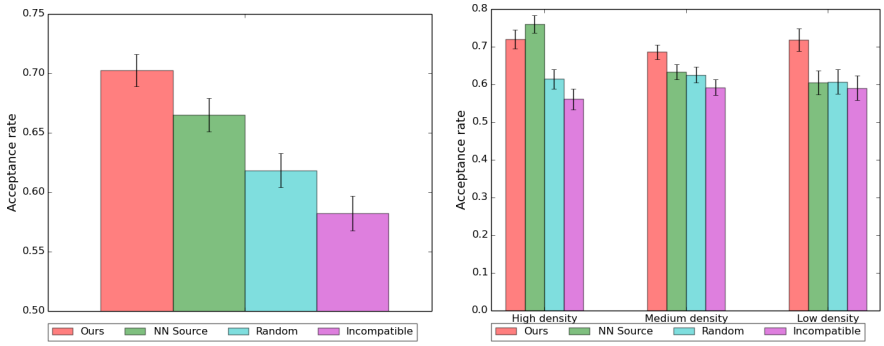
We approached a panel of four fashion specialists (FS) to provide feedback on recommendations generated by various algorithms. Each FS was presented with 17 recommendations for a given query (bottom) item, for each of the four algorithms. Among these recommendations, the FS were asked to select those that they judge to be complementary to the query. We used a total of 64 different query bottoms in this study, ranging for popular bottoms such as blue jeans to

(a) Recommendations for dark jeans

(b) Recommendations for a pink skirt

Fig. 4. Top to bottom: recommendations for two queries by CRAFT, NN-Source and the Incompatible algorithms. Highlighted in green are the items accepted marked by a fashion specialist as complementary to the query input, whereas rejected items are in red. Best viewed in color. Our approach generates better and diverse recommendations.

less common bottoms such as richly patterned skirts. The images were presented to FS in a random order to eliminate any bias for the algorithm or query items. Since some FS are in general more selective than others, we need to normalize for their individual bias. To achieve this, we add the *actual* top worn by the user in the query outfit to the set of 17 recommendations at a random location. We normalize the FS acceptance scores by their likelihood of selecting the actual top as an acceptable recommendation. Note that we only perform analysis on the newly recommended tops, and exclude the original top from our results.



(a) Overall acceptance rates (b) Binned acceptance rates

Fig. 5. Mean acceptance rate of recommendations by fashion specialists (error bars indicate 95% confidence intervals). (a) Overall acceptance rating for each algorithm. (b) Acceptance ratings binned according to the density (high, medium, low) of query items.

4.7 Quantitative Analysis

Figure 5(a) shows the average rate of acceptance of generated recommendations for all FS for the four algorithms. As discussed, acceptance rates were normalized by the probability of each FS accepting the actual top for the given query bottom. The error bar denotes the 95% confidence interval for each of the results. Non-overlapping error bars indicate that the differences between the two results are statistically significant. The NN-Source algorithm has the overall acceptance score of 66.5 ± 1.4 and outperforms the *Random* and *Incompatible* baseline algorithms as expected. The CRAFT approach generates recommendations with the highest FS acceptance score (70.3 ± 1.4).

Stratification by Feature Space Density It is even more interesting to break down the analysis of the results in terms of the density of the query items in the feature space. To this end, we approximate the density of each query point by taking the average distance to $K = 25$ nearest neighbors and bin the queries into low, medium, and high density regions, respectively. Figure 5(b) shows the average recommendation acceptance rate provided by FS for each algorithm in each density region. Again, the error bars denote the 95% confidence interval for each result. For queries that fall in the high density regions, the difference between CRAFT and the NN-Source algorithm is statistically insignificant (error bars overlap). This is expected since nearest neighbor search is a good estimator of the joint top-bottom density for high density regions, where a large number of samples are available. This is expected since nearest neighbor search is a good estimator of the conditional distribution of tops given a bottom for high density regions, where a large number of bottoms are available. However, the NN-Source

algorithm starts to degrade at the medium density level, and eventually degenerates to similar performance as the *Random* and the *Incompatible* recommendation algorithms for low density regions. In contrast, the performance of CRAFT is consistent across all regions and is better than baseline algorithms for mid and low density regime. Thus, the proposed conditional transformer is able to generalize well irrespective of the density of the neighborhood surrounding the query item.

5 Conclusion and Future Work

We presented CRAFT, an approach to visual complementary recommendation by learning the *joint* distribution of co-occurring visual objects in an unsupervised manner. Our approach does not require annotations or labels to indicate complementary relationships. The feature transformer in CRAFT samples a *conditional* distribution to generate diverse and relevant item recommendations for a given query. The recommendations generated by CRAFT are preferred by the domain experts over those produced by competing approaches.

By modeling the feature level distributions, our framework can potentially enable a host of applications, ranging from domain adaptation to one- or few-shot learning. The current work could be extended to incorporate the end-to-end learning of domain-related encoders as part of the generative framework.

References

1. Adomavicius, G., Tuzhilin, A.: Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. IEEE Trans. Knowledge and Data Engineering **17**(6), 734–749 (Jun 2005)
2. Bousmalis, K., Silberman, N., Dohan, D., Erhan, D., Krishnan, D.: Unsupervised pixel-level domain adaptation with generative adversarial networks. In: CVPR. pp. 95–104 (2017)
3. Che, T., Li, Y., Jacob, A.P., Bengio, Y., Li, W.: Mode regularized generative adversarial networks. In: ICLR (2017)
4. Croitoru, I., Bogolin, S.V., Leordeanu, M.: Unsupervised learning from video to detect foreground objects in single images. In: ICCV. pp. 4335–4343 (2017)
5. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A Large-Scale Hierarchical Image Database. In: CVPR (2009)
6. Doersch, C., Zisserman, A.: Multi-task self-supervised visual learning. In: ICCV (2017)
7. Donahue, J., Krähenbühl, P., Darrell, T.: Adversarial feature learning. In: International Conference on Learning Representations (2017)
8. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S.: Generative adversarial nets. In: NIPS. pp. 2672–2680 (2014)
9. Han, X., Wu, Z., Jiang, Y.G., Davis, L.S.: Learning fashion compatibility with bidirectional lstms. In: ACM on Multimedia Conference. pp. 1078–1086 (2017). https://doi.org/10.1145/3123266.3123394
10. He, R., Packer, C., McAuley, J.: Learning compatibility across categories for heterogeneous item recommendation. In: IEEE 16th International Conference on Data Mining, ICDM. pp. 937–942 (2016)

11. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: ICML. pp. 448–456 (2015)
12. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial nets. In: CVPR (2017)
13. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive Growing of GANs for Improved Quality, Stability, and Variation. In: ICLR (2018)
14. Kiapour, M.H., Han, X., Lazebnik, S., Berg, A.C., Berg, T.L.: Where to buy it: Matching street clothing photos in online shops. In: ICCV (2015)
15. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. ICLR (2015), http://arxiv.org/abs/1412.6980
16. Koren, Y., Bell, R.: Advances in collaborative filtering. In: Ricci, F., Rokach, L., Shapira, B., Kantor, P.B. (eds.) Recommender Systems Handbook, pp. 145–186 (2011)
17. Lample, G., Zeghidour, N., Usunier, N., Bordes, A., Denoyer, L., Ranzato, M.: Fader networks: Manipulating images by sliding attributes. In: Advances in Neural Information Processing Systems. pp. 5963–5972 (2017)
18. Lew, M.S., Sebe, N., Djeraba, C., Jain, R.: Content-based multimedia information retrieval: State of the art and challenges. ACM Trans. Multimedia Comput. Commun. Appl. **2**(1), 1–19 (2006)
19. Liang, X., Liu, S., Shen, X., Yang, J., Liu, L., Dong, J., Lin, L., Yan, S.: Deep human parsing with active template regression. IEEE Trans. Pattern Anal. Mach. Intell. **37**(12), 2402–2414 (Dec 2015). https://doi.org/10.1109/TPAMI.2015.2408360, http://dx.doi.org/10.1109/TPAMI.2015.2408360
20. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L., Pajdla, T., Schiele, B., Tuytelaars, T.: Microsoft COCO: Common Objects in Context. Springer International Publishing (2014)
21. Liu, S., Song, Z., Wang, M., Xu, C., Lu, H., Yan, S.: Street-to-shop: Cross-scenario clothing retrieval via parts alignment and auxiliary set. ACM Multimedia pp. 1335–1336 (2012)
22. Maas, Andrew L and Hannun, Awni Y and Ng, Andrew Y.: Rectifier nonlinearities improve neural network acoustic models. In: ICML. vol. 30 (2013)
23. van der Maaten, L., Hinton, G.: Visualizing high-dimensional data using t-SNE. Journal of Machine Learning Research **9**, 2579–2605 (2008)
24. McAuley, J.J., Targett, C., Shi, Q., van den Hengel, A.: Image-based recommendations on styles and substitutes. In: SIGIR (2015)
25. Melville, P., Mooney, R.J., Nagarajan, R.: Content-boosted collaborative filtering for improved recommendations. In: Eighteenth National Conference on Artificial Intelligence. pp. 187–192 (2002)
26. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training GANs. In: NIPS. pp. 2234–2242 (2016)
27. Shrivastava, A., Pfister, T., Tuzel, O., Susskind, J., Wang, W., Webb, R.: Learning from simulated and unsupervised images through adversarial training. In: CVPR (2017)
28. Szegedy, C., Ioffe, S., Vanhoucke, V.: Inception-v4, inception-resnet and the impact of residual connections on learning. AAAI **abs/1602.07261** (2017)
29. Tzeng, E., Hoffman, J., Darrell, T., Saenko, K.: Adversarial discriminative domain adaptation. In: CVPR (2017)
30. Veit, A., Kovacs, B., Bell, S., McAuley, J., Bala, K., Belongie, S.: Learning visual clothing style with heterogeneous dyadic co-occurrences. In: International Conference on Computer Vision (ICCV) (2015)

31. Volpi, R., Morerio, P., Savarese, S., Murino, V.: Adversarial feature augmentation for unsupervised domain adaptation. In: Computer Vision and Pattern Recognition (2018)
32. Vondrick, C., Torralba, A.: Generating the future with adversarial transformers. In: CVPR (2017)
33. Wang, T.C., Liu, M.Y., Zhu, J.Y., Tao, A., Kautz, J., Catanzaro, B.: High-resolution image synthesis and semantic manipulation with conditional gans. In: Computer Vision and Pattern Recognition (2018)
34. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: CVPR (2017)
35. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: ICCV (2017)
36. Zhu, S., Fidler, S., Urtasun, R.: Be your own prada: Fashion synthesis with structural coherence. In: ICCV (2017)