Explainable Noisy Label Flipping for Multi-Label Fashion Image Classification

Beatriz Quintino Ferreira, João P. Costeira, João P. Gomes
ISR-IST, Universidade de Lisboa
{beatrizquintino, jpc, jpg}@isr.tecnico.ulisboa.pt *

Abstract

In online shopping applications, the daily insertion of new products requires an overwhelming annotation effort. Usually done by humans, it comes at a huge cost and yet generates high rates of noisy/missing labels that seriously hinder the effectiveness of CNNs in multi-label classification. We propose SELF-ML, a classification framework that exploits the relation between visual attributes and appearance together with the "low-rank" nature of the feature space. It learns a sparse reconstruction of image features as a convex combination of very few images - a basis - that are correctly annotated. Building on this representation, SELF-ML has a module that relabels noisy annotations from the derived combination of the clean data. Due to such structured reconstruction, SELF-ML gives an explanation of its label-flipping decisions. Experiments on a real-world shopping dataset show that SELF-ML significantly increases the number of correct labels even with few clean annotations.

1. Introduction

Deep Convolutional Neural Networks (DCNNs) are the premier technique for supervised visual learning, especially for multi-class/multi-label classification. However, as application scenarios scale up, obtaining accurate supervisory data for inflated label spaces is simply not feasible in practice. For example, public multi-label fashion datasets [2, 8] have label spaces on the order of 10^3. Since most data is provided through human labeling, errors mount and some labels are highly mislabeled. This is recurrent in fashion e-commerce platforms where relevant attributes are not annotated since they are seldom queried. For *e.g.*, it is infrequent to search for jackets using the label *Long Sleeved*, consequently, this attribute is consistently mislabeled. This issue has a severe negative impact both in retrieval (product variability and discoverability) and in recommendations [16].

Against this backdrop, we propose SELF-ML, a DCNN classification framework that handles highly mislabeled training. By exploring the relation between visual attributes and appearance, SELF-ML infers new labels from the visual reconstruction of similarly-looking well-annotated images.

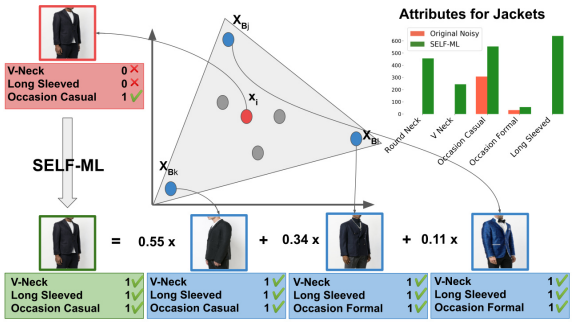


Figure 1: Example of a sparse and explainable reconstruction of a noisy labeled image (with the red frame) obtained by our SELF-ML from a few clean labeled images (in blue).

SELF-ML selects a small set of images forming a "basis" that reconstructs the whole dataset through convex combinations. As illustrated in Fig. 1, SELF-ML learns the "best" reconstruction of a noisy labeled image (in red) from a selection of correctly annotated basis images (in blue), propagating via a flipping model the correct labels (in green). As shown in the example, SELF-ML finds true positives of labels suffering from systematic mislabeling, as *Long Sleeved* and *V-Neck* for jackets. The underlying assumption (empirically verified) is that the feature space learned by DCNNs is "low rank" and approximated by linear subspaces [17], enabling a sparse reconstruction by few points that create the convex hull of meaningful subsets of image features.

The existence of a set of images from which SELF-ML can select a small set to be cleanly annotated is common in real applications [6, 15]. This setting arises when images are collected from the web or in e-commerce platforms, where only a small subset of product images is assigned to be labeled by experts (possibly limited by a labeling budget). The sizes of the two subsets comprising the total dataset T are usually significantly different, with $|N| \gg |C|$, where N and C denote the noisy and the clean set, respectively. Training a robust classifier in this setting is challenging as, on the one hand, the large set N has noisy/missing labels and, on the other hand, the clean set C is too small to train DCNNs. A common practice is to use

*Work partially funded by FCT via grant [PD/BD/114430/2016] and project [UIDB/50009/2020], and by iFetch project co-financed by ANI and FCT under CMU Portugal. We also thank Ricardo Sousa from Farfetch.

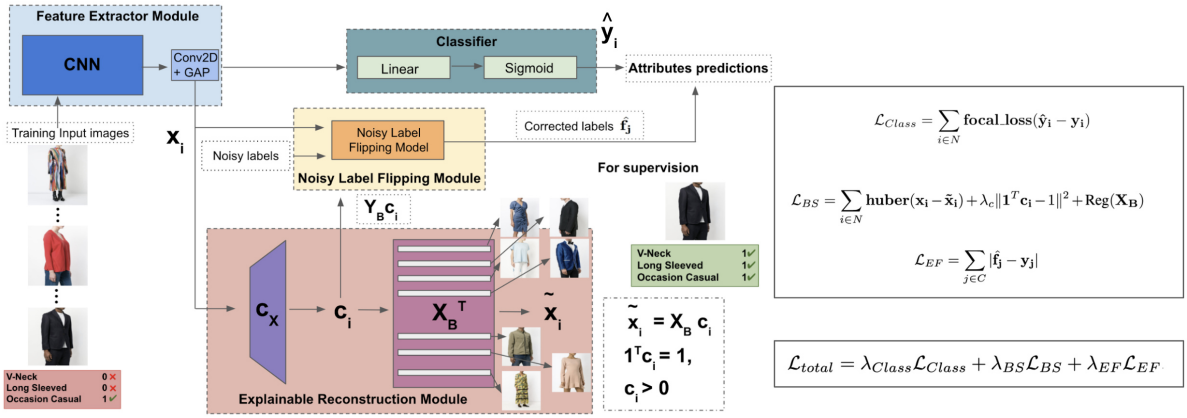


Figure 2: Overview of the proposed SELF-ML framework. SELF-ML has a feature extractor module trained for classification, an explainable reconstruction module that infers a sparse convex combination $\mathbf{c}_i$ of features of visually similar images from a basis of cleanly labeled images $\mathbf{X_B}$, and a label flipping module that uses the combination $\mathbf{c}_i$ and the verified labels from the clean set $\mathbf{Y_B}$ to learn to correct the noisy labels. Finally, these corrected labels are used as supervision to the classifier.

transfer learning by (pre-)training the model on set N and then fine-tuning it on C to improve performance. However, this approach is prone to overfitting the small set [4, 14, 15].

Our contributions are 3-fold: 1) We propose the SELF-ML framework that integrates 4 modules: *feature extraction*, *classifier*, *self-explainable reconstruction*, and *noisy label flipping*. The explainable reconstruction module learns to reconstruct image features, previously learned by the feature extractor and classifier, from a sparse "basis". If the attributes of the images of this basis are "cleaned", the noisy label flipping module learns to correct the noisy labels during model training, leading to remarkable classification gains. SELF-ML is also much less prone to overfitting than fine-tuning. 2) By training SELF-ML to obtain a reconstruction of the input image features based on the set of basis images, our flipping mechanism will be interpretable and explainable. 3) We show SELF-ML effectiveness on a multi-label fashion dataset with label imbalance and noise.

2. Proposed Approach: Self Explainable Label Flipping for Multi-Label Classification

The overview of SELF-ML is shown in Fig. 2. For each image i, we seek to represent the visual embeddings as $\mathbf{x_i} = \mathbf{X_B c_i}$ where $\mathbf{X_B} \in \mathbb{R}^{d \times n_B}$ is a subset of the data and $\mathbf{c_i} \in \mathbb{R}^{n_B}$ is a sparse vector. This representation is derived in the *explainable reconstruction model* by enforcing $\tilde{\mathbf{x}}_\mathbf{i} \approx \mathbf{x_i}$. Classical and deep subspace clustering approaches [1, 5, 9, 17] robustly encode images in a similar way however, they must solve large global optimization problems that require the entire dataset. Consequently, these approaches are not scalable and cannot represent unseen examples.

Admitting that visual similarity implies similar attributes, the labels of a given image $\mathbf{x_i}$ could be straightforwardly regressed from the convex combination of the corresponding clean labels of basis $\mathbf{X_B}$, *i.e.* $\mathbf{Y_B c_i}$. We observed the adequacy of this assumption. However, provided there is enough training, results can be significantly improved by a learned "flipping" model that infers the new labels from the explicit image representation and the noisy labels.

In summary, for a dataset of n_T images T = $\{(\mathbf{I_1}, \mathbf{y_1}), ..., (\mathbf{I_{n_T}}, \mathbf{y_{n_T}})\}$, where $\mathbf{I_i}$ is the i-th image and $\mathbf{y_i} \in \{0, 1\}^{n_{attr}}$ is a vector with the attribute labels (1 indicates attribute presence and 0 its absence) associated to $\mathbf{I_i}$ (if $i \in C$ then labels in $\mathbf{y_i}$ are correct, but if $i \in N$ they were not verified and may be noisy), we train a classification pipeline (comprising the *feature extractor module* and the *classifier*) so that meaningful features ($\mathbf{x_i} \in \mathbb{R}^d$) are learned for each image i. This is trained only on the noisy labeled set N. Features $\mathbf{x_i}$ are the input to the *explainable reconstruction module* that learns to reconstruct features $\mathbf{x_i}$ of each training image based on a sparse decomposition (given by the learned $\mathbf{c_i}$) of the image features that belong to the basis $\mathbf{X_B}$. As this model deals only with visual information it is trained solely on noisy data (set N).

The *label flipping module* builds on what is learned by the reconstruction module since it learns to flip the noisy labels conditioned on the sparse decomposition coefficients. To predict the correct labels for each image this module takes as input the original noisy labels, the cleaned labels of the features learned in $\mathbf{X_B}$ weighted by the $\mathbf{c_i}$ vector, and the image features $\mathbf{x_i}$. This is the only component of SELF-ML that trains on images with cleaned labels (set C).

We remark that, as a result of its design, SELF-ML does not make any assumption on the noisy label distribution.

Selecting a Basis of clean labeled images We implement the explainable reconstruction module using two linear layers: $\mathbf{C_X}$ that regresses coefficients $\mathbf{c}_i$, and $\mathbf{X_B}$, whose learned weights will form a basis of image features. To ensure the learned explainable reconstruction $\tilde{\mathbf{x}}_\mathbf{i} = \mathbf{X_B c_i}$ contains meaningful information about the input image fea-

Method (# clean samples)	**Precision@k***	**Recall@k***	**F1@k***	**AP_{all}**	**MAP**	**Jaccard Similarity**	**0-1 exact match**
Baseline 0 - *only noisy* (0)	68.16	57.74	56.30	75.02	65.69	29.62	1.60
Baseline 1 - *fine-tuned with clean* (500)	83.83	83.89	83.76	87.28	70.91	74.86	41.73
SELF-ML - *jointly trained* (500)	**86.30**	**86.75**	**86.33**	**89.54**	**74.99**	**78.77**	**50.00**

*metrics at top-k, where k is the number of ground-truth labels of each image

Table 1: Multi-label classification performance (in %) for compared methods. SELF-ML outperforms the other methods.

tures, we include the reconstruction error between the extracted feature $\mathbf{x_i}$ and the reconstructed feature $\tilde{\mathbf{x}}_\mathbf{i}$, using the Huber function [3] (that further induces sparsity). We also apply a basis regularizer $\text{Reg}(\mathbf{X_B}) = \|\mathbf{X_B} - \mathbf{X_{sample}}\|_2$ so that the learned basis is close to a set of image features $\mathbf{X_{sample}}$. We let the dimension of $\mathbf{X_B}$ be large (n_B=1000) and since coefficients $\mathbf{c_i}$ will only select very few of these features ($\mathbf{c_i}$ is very sparse, only ≈ 40 out of 1000 elements are nonzero) this regularization is only providing an initialization (the model chooses which images go to the basis).

Labeling the Basis We can follow one of two strategies to obtain the correct labels for the clean basis: 1) we can have an oracle that correctly annotates the images learned to be on the basis $\mathbf{X_B}$ that are associated with nonzero elements of $\mathbf{c_i}$, or 2) we find, for each of the selected images in $\mathbf{X_B}$, the nearest neighbor in the learned feature space from the images in the clean set, in case we have a previously defined clean set C, and hence use the clean labels already gathered.

If we follow 1) we may not collect enough annotations to train the label flipping model, so we can apply a direct rule that flips the labels of the images in the noisy set N based on the coefficients $\mathbf{c_i}$ and correct annotations for the selected basis images. Such an explicit rule can be applied because the obtained $\mathbf{c_i}$ coefficients are intelligible, demonstrating the explainable nature of SELF-ML. Contrarily, if a subset C is available we can train and use the label flipping model, which provides improved label flipping performance.

Explaining Label Flips The coefficients $\mathbf{c}_i$, learned by the reconstruction module, together with the correct labels $\mathbf{Y_B} \in \{0,1\}^{n_{attr} \times n_B}$ (collected by one of the two previous annotation strategies) for the images in basis $\mathbf{X_B}$ help the label flipping module to correct each noisy labeled training image. This model learns a shared space that combines the original noisy labels, the features $\mathbf{x_i}$, and $\mathbf{Y_B c_i} \in \mathbb{R}^{n_{attr}}$ (which provides a label score for each image i based on the correct labels from its reconstruction). The flipping model is trained on clean set C, for which we also have the original noisy labels. This model loss is the absolute distance between flipped labels $\hat{\mathbf{f}}_\mathbf{i}$ and clean ground-truth labels $\mathbf{y_i}$. Note that $\hat{\mathbf{f}}_\mathbf{i}$ will serve as new corrected supervision for the classifier, replacing the original noisy labels $\mathbf{y_i}$ from N.

Model Training According to Fig. 2, SELF-ML is trained by minimizing $\mathcal{L}_{total}$ which combines the classification loss $\mathcal{L}_{Class}$ (we use the Focal Loss [7]), the Basis Selection loss $\mathcal{L}_{BS}$, and the Explainable Flipping loss $\mathcal{L}_{EF}$. λ_{Class}, λ_{BS} and λ_{EF} denote hyperparameters to control the contribution of each loss. Batches are sampled so that 9/10 images are from N and 1/10 from C (to train the label flipping module).

3. Experiments

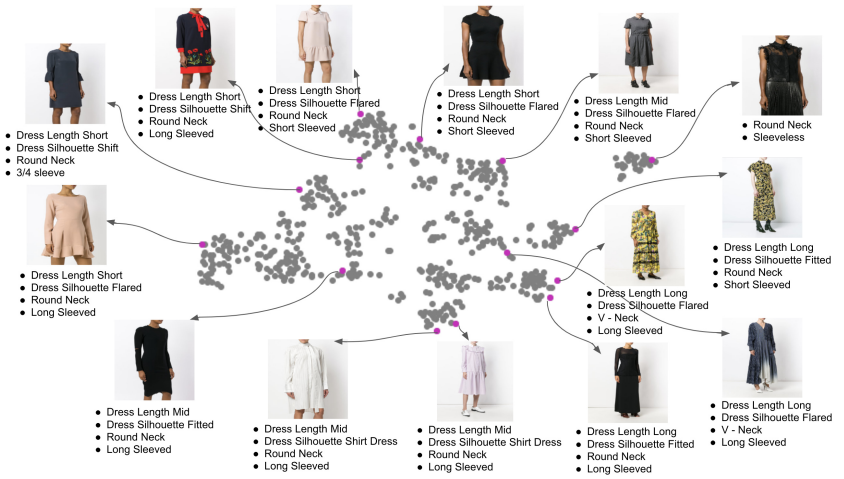
Dataset It is easy to spot mislabeling in widely used datasets as the Deepfashion [8], which undermines its use for benchmarks. Thus, similar to the approach followed in other works [4, 14], we test SELF-ML on an e-commerce fashion dataset, previously used in [10, 11], for which a sampled subset was verified and curated by experts. In near future work, we plan to experiment on public datasets.

This real-world dataset has approx. 60200 images belonging to the 4 most frequent clothing categories (*dresses*, *jackets*, *knitwear* and *tops*), and each image has 17 possible attributes associated to it. Examples of these attributes are: *Round Neck*, *Short Sleeved*, *Dress Length Long*, *Dress Silhouette Flared* or *Occasion Formal*. The average number of attribute annotations per image in the noisy set N is 1.28, and 3.40 in C, revealing the noise present. The clean set has $|C| = 3000$ images, which is $\approx 5\%$ of the total dataset size, a fraction similar to the verification labels used in [6, 15].

Training Details and Experimental Setup All models are implemented with Tensorflow/Keras, with VGG-16 [12] as CNN backbone, and run under the same conditions: stopped training after 2×10^4 batches of size 64 (when we pre-train the CNN we train for 10^4 batches and when we fine-tune we train for additional 10^4), optimized by Adam (with *lr* 10^{-4} and decay 10^{-5}). Images are resized to 224×224 and random rotations, translations and horizontal flips are applied to $1/3$ of the images. We use the focal loss standard parameters ($\gamma = 2$, $\alpha = 0.6$) and train SELF-ML with the following hyperparameters: $\lambda_c = 10$, $\lambda_{Class} = 1$, $\lambda_{BS} = 1$, $\lambda_{EF} = 10$, $d = 512$, and $n_B = 1000$.

We applied a 75%|25% train|test split ratio for both the noisy and the clean sets, named $N_{train}|N_{test}$ and $C_{train}|C_{test}$, respectively. Therefore, when we train a model only on noisy labels we train it on N_{train}, and when we use clean samples for training we sample them from C_{train}. Evaluation is performed on the clean test set C_{test}.

Compared Methods *Baseline 0:* train the feature extractor and classifier on N_{train} and on original noisy labels of C_{train}. This is the lower-bound of the following methods.
Baseline 1: pre-train the feature extractor and the classifier on N_{train} and fine-tune the last layers (Conv2D + GAP and the classifier) with clean labels (sampled from C_{train}).
SELF-ML- jointly trained: pre-train the feature extractor, the classifier, and then the explainable reconstruction on N_{train}, and pre-train the label flipping model on C_{train}. Then jointly train the explainable reconstruction, the classifier, and the noisy label flipping modules, initializing the flipping model with the pre-training weights.



(a) Left: t-SNE of $\mathbf{c_i}$ vectors for a random sample of *dresses* images. The images B_i, for which $c_{i_{B_i}} \approx 1$ (in magenta) are selected to be on basis $\mathbf{X_B}$, approximately form the convex hull of the data points and cover all possible attributes for *dresses*.

(b) Above: Images selected for the basis $\mathbf{X_B}$ are mainly reconstructed by themselves and by few other images (with much lower c_{i_k}).

Figure 3: Examples of what is learned by $\mathbf{X_B}$ and $\mathbf{c_i}$

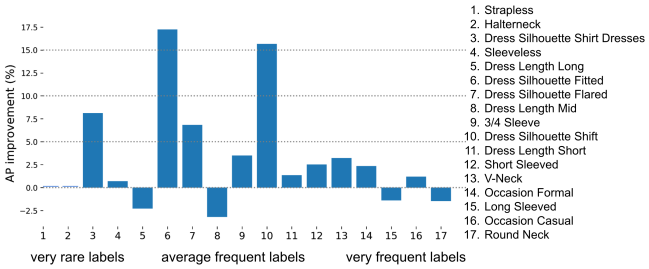


Figure 4: Average Precision gain per label of SELF-ML over Baseline 1. SELF-ML can improve more rare and difficult labels that lead the fine-tuning approach to overfitting.

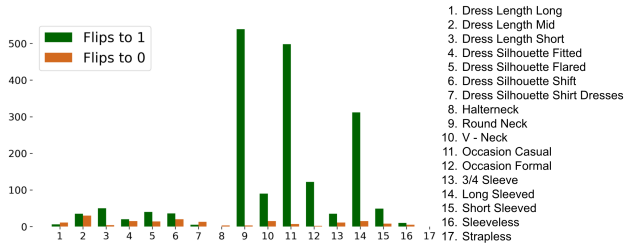


Figure 5: Labels flipped to 1 and to 0 by SELF-ML.

Classification Results Table 1 reports the multi-label classification performance for the compared methods. As anticipated, training only on noisy labels has very poor performance, and Baseline 0 is indeed a lower bound of the remaining methods. In Baseline 1, we observe a clear boost in all metrics. However, fine-tuning the last layers on clean samples is prone to overfitting more rare labels (due to the limited samples available for these labels in the clean set), as pointed out by [4, 14]. This behavior is indicated by a small increase of the MAP. Contrarily, SELF-ML shows special improvement over the fine-tuning approach for the less frequent labels. This is shown in Fig. 4, where we plot the improvement in AP for each label sorted by label frequency. Particularly, the most significant improvements of AP ($\approx$ 5% to 20%) occurred for labels that are more difficult to capture (as they are visually more ambiguous) like *dress silhouettes*. Using the same 500 clean samples, SELF-ML beats all baselines for all metrics by a considerable margin, demonstrating that it is more effective than fine-tuning.

Label Flipping Fig. 5 shows the relabeling (to 1 or to 0) obtained by our flipping model for each attribute, on C_{test}. We observe that it is able to flip a massive number of false negatives for the noisiest attributes (for example, none of the jackets images had positive annotations for *Round Neck* or for *Long Sleeved*), but at the same time, also relabels false positives that are distributed among all attributes.

What is Learned by $\mathbf{X_B}$ and $\mathbf{c_i}$? Considering only the subset of *dresses* images, for clarity, Fig.3a shows the t-SNE [13] of the $\mathbf{c_i}$ vectors learned for this subset. We observe that the $\mathbf{c_i}$ for the few images B_i selected for the basis $\mathbf{X_B}$ (with $c_{i_{B_i}} \approx 1$) indeed approximate well the convex hull of the $\mathbf{c_i}$ vectors for images in this class. We also notice smoothness in the space as nearby images share most of the labels. Additionally, we saw that obtained $\mathbf{c_i}$ vectors are very sparse, and thus only $\approx$ 40 image features from the columns of basis $\mathbf{X_B}$ are used to reconstruct the entire noisy training set (considerably less than the basis size n_B=1000). Although very few, we found that the selected basis images belong to the 4 possible classes and, if correctly annotated, cover all n_{attr}=17 possible labels. Finally, we verified that the images learned for basis $\mathbf{X_B}$ are mainly reconstructed by themselves and by few other similar images with a much lower c_{i_k} coefficient (examples are presented in Fig. 3b).

4. Conclusions

Creating complete and consistent multi-label fashion datasets requires tremendous effort and is extremely costly. We introduce SELF-ML, a DCNN classification framework that selects a set of images that should be cleanly annotated allowing to correct all noisy labels. Besides providing a flipping explanation, it outperforms compared approaches.

References

[1] E. Elhamifar and R. Vidal. Sparse Subspace Clustering : Algorithm, Theory, and Applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(11), 2013.

[2] X. Han, Z. Wu, P. X. Huang, X. Zhang, M. Zhu, Y. Li, Y. Zhao, and L. S. Davis. Automatic Spatially-Aware Fashion Concept Discovery. In *IEEE International Conference on Computer Vision (ICCV)*, 2017.

[3] P. Huber. *Robust Statistics*. John Wiley & Sons, 1981.

[4] N. Inoue, E. Simo-Serra, T. Yamasaki, and H. Ishikawa. Multi-label fashion image classification with minimal human supervision. In *IEEE International Conference on Computer Vision Workshops (ICCVW)*, 2017.

[5] P. Ji, T. Zhang, H. Li, M. Salzmann, and I. Reid. Deep Subspace Clustering Networks. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2017.

[6] K. H. Lee, X. He, L. Zhang, and L. Yang. CleanNet: Transfer Learning for Scalable Image Classifier Training with Label Noise. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[7] T. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. In *IEEE International Conference on Computer Vision (ICCV)*, 2017.

[8] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang. DeepFashion: Powering Robust Clothes Recognition and Retrieval with Rich Annotations. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. Available at: http://mmlab.ie.cuhk.edu.hk/projects/DeepFashion.html, (accessed January 2021).

[9] X. Peng, J. Feng, J. T. Zhou, Y. Lei, and S. Yan. Deep Subspace Clustering. *IEEE Transactions on Neural Networks and Learning Systems*, 2020.

[10] B. Quintino Ferreira, L. Baía, J. Faria, and R. G. Sousa. A Unified Model with Structured Output for Fashion Images Classification. In *KDD'18 Workshop on AI for Fashion*, 2018.

[11] B. Quintino Ferreira, J.P. Costeira, R. G. Sousa, L-Y Gui, and J. P. Gomes. Pose Guided Attention for Multi-label Fashion Image Classification. In *IEEE International Conference on Computer Vision Workshops (ICCVW)*, 2019.

[12] K. Simonyan and A. Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *International Conference on Learning Representations (ICLR)*, 2015.

[13] L. van der Maaten and G. Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9, 2008.

[14] A. Veit, N. Alldrin, G. Chechik, I. Krasin, A. Gupta, and S. Belongie. Learning from noisy large-scale datasets with minimal supervision. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[15] T. Xiao, T. Xia, Y. Yang, C. Huang, and X. Wang. Learning from massive noisy labeled data for image classification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[16] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations (ICLR)*, 2017.

[17] J. Zhang, C. G. Li, C. You, X. Qi, H. Zhang, J. Guo, and Z. Lin. Self-Supervised Convolutional Subspace Clustering Network. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.