Feature-level Frankenstein: Eliminating Variations for Discriminative Recognition

Xiaofeng Liu[1,2,4†], Site Li[1†], Lingsheng Kong[2], Wanqing Xie[3,4*], Ping Jia[2], Jane You[5], B.V.K. Kumar[1]
1. Carnegie Mellon University, Pittsburgh, PA, USA
2. CIOMP, Chinese Academy of Sciences, Changchun, China
3. Harbin Engineering University, Harbin, China
4. Harvard University, Cambridge, MA, USA
5. Dept. of Computing, The Hong Kong Polytechnic University, Hong Kong, China
†contribute equally {liuxiaofeng, site} cmu@gmail.com, ⋆corresponding author

Abstract

Recent successes of deep learning-based recognition rely on maintaining the content related to the main-task label. However, how to explicitly dispel the noisy signals for better generalization remains an open issue. We systematically summarize the detrimental factors as task-relevant/irrelevant semantic variations and unspecified latent variation. In this paper, we cast these problems as an adversarial minimax game in the latent space. Specifically, we propose equipping an end-to-end conditional adversarial network with the ability to decompose an input sample into three complementary parts. The discriminative representation inherits the desired invariance property guided by prior knowledge of the task, which is marginally independent to the task-relevant/irrelevant semantic and latent variations. Our proposed framework achieves top performance on a serials of tasks, including digits recognition, lighting, makeup, disguise-tolerant face recognition, and facial attributes recognition.

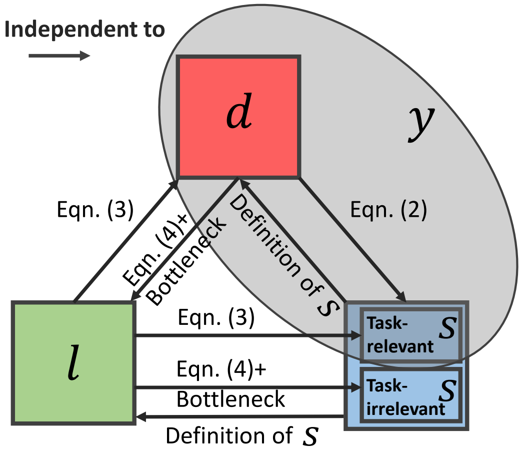


Figure 1. Illustration of the expected separations of the observation x, which associated with the discriminative representation d (red), latent variation l (green) and semantic variations s (blue). Our framework explicitly enforces them marginally independent to each other. The d and task-dependent s are related to the main-recognition task label y.

1. Introduction

Extracting a discriminative representation for the task at hand is an important research goal of recognition. We targeting for the problem of explicitly eliminating the detrimental variations following the prior knowledge of our task to achieve better generalization. It is challenging since the training set contains images annotated with multiple semantic variations of interest, but there is no example of the transformation (*e.g.*, gender) as the unsupervised image translation [33, 6], and the latent variation is totally unspecified.

Following the terminology used in previous multi-class dataset (including a main-task label and several side-labels) [20, 46, 44], we propose to define three complementary parts as in Fig 1.

The factors relate to the side-labels is named as the $semantic\ variations$ (s), which can be either $task-relevant/irrelevant$ depending on whether they are marginally independent to the main recognition task or not [59]. The $latent\ variation$ (l) summarizes the remaining properties unspecified by main and semantic labels. How the DNN can systematically learn a $discriminative\ representation$ (d) to be informative for the main recognition task, while marginally independent to both multiple s and unspecified l remains challenging.

Several efforts have been made to enforce the main task representation invariant to a single task-*irrelevant* (independent) semantic factor, such as pose, expression or illumination-invariant face recognition via neural preprocessing [19, 56] or metric learning [35]. These methods bear the same drawback that the cost used to regularize the representation is pairwise, which does not scale well as the number of values that the attribute can take could be large. Since the invariance we care about can vary greatly across tasks, these approaches require us to design a new architecture each time when a new invariance is required.

Moreover, a basic assumption in their theoretical analysis is that the attribute is $irrelevant$ to the prediction, which limits its capabilities in analyzing the task-$relevant$ (dependent) semantic labels [5, 59]. These labels are usually used to achieve the attribute-enhanced recognition via the feature aggregation in multi-task learning [16, 22, 49, 32] (*e.g.*, gender, age and ethnicity can shrink the search space for face identification).

However, the invariance *w.r.t.* those attributes are also desired in some specific tasks. For example, the makeup face recognition system should be invariant to age, hair color *etc*. Similarly, the gender and ethnicity are sensitive factors in fairness/bias-free classification when predicting the credit and health condition of a person. These semantic labels and the main task label are related due to the inherent bias within the data. A possible solution is setting this attribute as a random variable of a probabilistic model and reasoning about the invariance explicitly [39, 9, 61]. Since the divergence between a pair of distributions is used as the criteria to induce the invariance, the number of pairs to be processed grows quadratically with the number of attributes, which can be computationally expensive for the multiple variations in practice.

Another challenge is how to achieve better generalization by dispelling those latent variations without the label. For instance, we may expect our face recognition system not only be invariant to the expression following the side label, but also applicable to different race, which do not has side label. We note this problem also share some similarity with feature disentanglements in image generation area [44, 14], while our goal is to improve content classification performance instead of synthesizing high-quality images.

Motivated by the aforementioned difficulties, we propose to enable a system which can dispel *a group of* undesired *task-irrelevant/relevant* and $latent$ variations in an unsupervised manner: we do not need paired semantic transformation example [33, 6] and latent labels.

Specifically, we resort to an end-to-end conditional adversarial training framework. Our approach relies on an encoder-decoder architecture where, given an input image x with its main-task label y and to-be dispelled semantic variation label s, the encoders maps x to a discriminative representation d and a latent variation l, and the decoder is trained to reconstruct x given (d,s,l). We configure a semantic discriminator condition to s only, and two classifiers with inverse objectives which condition to d and l, respectively to constrain the latent space for manipulating multiple variations for better scalability.

The main contributions of this paper are summarized as:
• It is able to explicitly learn a task-specific discriminative representation with desired invariance property by systematically incorporating prior domain knowledge of the task. The to-be dispelled $multiple$ semantic variations could be either task-$dependent$ or task-$independent$ semantic variations, and the unspecified $latent$ variation can also been eliminated in an $unsupervised$ manner.
• Semantic discriminator and two inverse classifiers are introduced to constrain the latent space and result in a simpler training pipeline and better scalability.
• The semantic and latent variation representations are jointly disentangled and preserved as the complementary parts. The flexible swapping of those factors makes different image transform with a unified model.

Experimental results on source-independent digits classification, lighting-tolerant (Extrended YaleB) face recognition, makeup-invariant face recognition and disguised face recognition benchmarks show that the proposed model outperforms existing discriminative approaches. We further show that our framework is generic enough to accommodate hand-written style transforms by switching the disentangled latent or semantic codes.

2. Related work

Multi-task learning is a typical method to utilize multi-class label. It has been observed in many prior works that jointly learning of main-task and $relevant$ side tasks can help improve the performance in an aggregation manner [16, 22, 49, 32], while we are targeting for dispelling.

Generative Adversarial Networks (GANs) has aroused increasing attraction. Conventionally, under the two-player (*i.e.*, generator and discriminator) formulation, the vanilla GANs [11] are good at generating realistic images, but their potential for recognition remains to be developed. The typical method use GANs as a preprocessing step of image, which is similar to the “denoise”, and then use these processed images for normal training and testing [56, 19, 47, 54, 60, 36, 41]. We deploy the trained network for predictions directly as a feature extractor.

Comparing with the pixel-level GANs [56, 19, 47, 54, 60, 36, 41], our feature-level competition results in much simpler training schemes and nicely scales to multiple attributes. Moreover, they usually cannot dispel task-relevant s, $e.g.$, dispel gender from identity cannot get verisimilar face image for subsequent network training.

Besides, they usually focus on a single variation for a specific task. Actually, the most of GANs and adversarial domain adaptation [57, 4, 31] use binary adversarial objective and applied for no more than two distributions.

It is worth noting that some works of GANs, e.g., Semi-Supervised GAN [24] and DR-GAN [56] have claimed that they consider multiple side labels. Indeed, they have added a new branch for the multi-categorical classification, but their competing adversarial loss only confuses the discriminator by using two distributions (real or generated) and no adversarial strategies are adopted between different categories in the auxiliary multi-categorical classifier branch.

We are different from them in two aspects: **1**) the input of semantic discriminator is feature, instead of real/synthesized image; **2**) the goal of encoder needs to match or align the feature distribution between any two different attributes, instead of only real/fake distribution, and there is no "real" class in semantic discriminator.

Fairness/bias-free classification also targets a representation that is invariant to certain task-relevant(dependent) factor (*i.e.*, bias) hence makes the predictions fair [7]. As data-driven models trained using historical data easily inherit the bias exhibited in the data, the Fair VAEs [39] tackled the problem using a Variational Autoencoder structure [25] approached with maximum mean discrepancy (MMD) regularization [30]. [62] proposed to regularize the l_1 distance between representation distributions of data with different nuisance variables to enforce fairness. These methods have the same drawback that the cost used to regularize the representation is pairwise, which does not scale well for multiple semantic variations [59, 5]. [59] propose to combine this concept with adversarial training, which has the similar framework as the Fader Networks [27] for image generation.

Latent variation disentangled representation is closely related to our work. It trying to separate the input into two complementary codes according to their correlation with the task for image transform in single label dataset setting [3]. Early attempts [52] separate text from fonts using bilinear models. Manifold learning and VAEs were used in [8, 24] to separate the digit from the style. What-where encoders [64] combined the reconstruction criteria with discrimination to separate the factors that are relevant to the labels. Unfortunately, their approaches cannot be generalized to unseen identities. [44, 58] added the GANs objective into the VAEs objective to relax this restriction using an intricate triplet training pipeline. [18, 37, 14, 2, 21, 21] further reduced the complexity. Inspired by them, we make our framework implicitly invariant to unspecified l for better generality in a simple yet efficient way. [44].

3. Methodology

3.1. The problem definition

We formalize the task of Feature-level Frankenstein (FLF) framework as follows: Given a training set $\mathcal{D} = \{x^1, s^1, y^1\}, \cdots, \{x^M, s^M, y^M\}$, of M samples $\{image, semantic\ variations, class\}$, we are interested in the task of disentangling the feature representation of x to be three complementary parts, *i.e.*, discriminative representation d, semantic variation s and latent variation l. These three codes are expected to be marginally independent with each other, as illustrated schematically in Fig. 1. In the case of face, typical semantic variations including gender, expressions $etc.$ All the remaining variability unspecified by y and s fall into the latent part l. We note that there are two possible dependency scenarios of s and y as discussed in Sec. 1. This will not affect the definition of l, and the information related to y should incorporate d and some of the task-dependent s.

3.2. The structure of representations

For the latent variation encoding, we choose the l to be a vector of real value rather than a one-hot or a class ordinal vector to enable the network to be generalized to identities that are not presented in the training dataset as in [44, 2]. However, as the semantic variations are human-named for a specific domain, this concern is removed. In theory, s can be any type of data (*e.g.*, continuous value scalar/vector, or a sub-structure of a natural language sentence) as long as it represents a semantic attribute of x under our framework. For simplicity, we consider here the case where s is a N-dimensional binary variable for N to-be controlled semantic variations. Regarding the multi-categorical labels, they are factorized to multiple binary choices. The domain adaptation could be a special case of our model when the semantic variation is the Bernoulli variable which takes the one-dimensional binary value (*i.e.*, $s = \{0, 1\}$), representing the domains.

3.3. Framework architecture

The model described in Fig. 2 is proposed to achieve our objective based on an encoder-decoder architecture with conditional adversarial training.

At inference time, a test image is encoded to the d and l in the latent space, and the d can be used for recognition task with desired invariant property $w.r.t.$ the s. Besides, the user can choose the combination of (d,s,l) that are fed to the decoder for different image transforms.

Informative to main-recognition task. The discriminative encoder E_d with parameter θ_{E_d} maps an input image to its discriminative representation $d = E_d(x)$ which is informative for the main recognition task and invariant to some semantic attributes. By invariance, we mean that given two samples x^1, x^2 from a subject class ($y^1 = y^2$) but with different semantic attribute labels ($s^1 \neq s^2$), their d^1 and d^2 are expected to be the same. Given the obtained d, we expect to predict its corresponding label y with the classifier C_d to model the distribution $p_{C_d}(y|x)$. The task of C_d

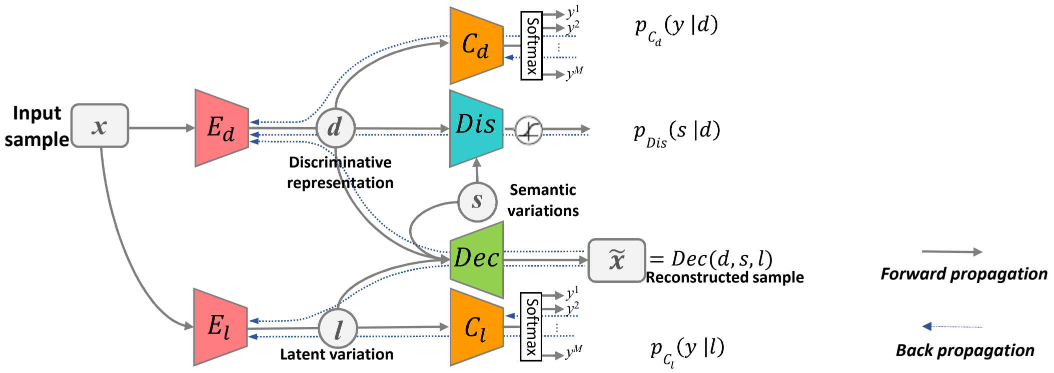


Figure 2. The proposed Feature-level Frankenstein framework, where the x is encoded into 3 parts (i.e., d, l, s) by two encoders, and a combination of (d, l, s) can be reconstructed to $\tilde{x}$ via decoder. The adversarial trained dis and classifiers are used to constrain the latent feature space.

and the first objective of the E_d is to ensure the accuracy of the main recognition task. Therefore, we update them to minimize :

$$\min_{E_d,C_d} \mathcal{L}_{C_d} = \mathbb{E}_{x,y\sim q(x,s,y)} - \mathtt{log} p_{C_d}(y|E_d(x)) \quad (1)$$

where we use the categorical cross-entropy loss for the classifier. The $q(x, s, y)$ is the true underlying distribution that the empirical observations are drawn from.

Eliminating semantic variations. The discriminator *Dis* output probabilities of an attribute vector $p_{Dis}(s|d)$. In practical implementation, this is made by concatenating d and binary attributes code s for input and outputs the [0,1] values using the sigmoid unit. Its loss depends on the current state of semantic encoders and is written as:

$$\min_{Dis} \max_{E_d} \mathcal{L}_{Dis} = \mathbb{E}_{x,s\sim q(x,s,y)} - \mathtt{log} p_{Dis}(s|E_d(x)) \quad (2)$$

Concretely, the *Dis* and E_d form an adversarial game, in which the *Dis* is trained to detect an attribute of data by maximizing the likelihood $p_{Dis}(s|d)$, while the E_d fights to conceal it by minimizing the same likelihood. Eq.(2) guarantees that d is marginally independent to s. Supposing that a semantic variation follows the Bernoulli distribution, the loss is formulated as $-\{s\mathtt{log}Dis(d) + (1-s)\mathtt{log}(1-Dis(d))\}$. The proposed framework is readily amenable to control multiple attributes by extending the dimension of semantic variation vector. With N to-be dispelled semantic variations, we have $\mathtt{log}p_{Dis}(s|d) = \sum_{i=1}^{N}\{\mathtt{log}p_{Dis}(s_i|d)\}$. Note that even with binary attribute values at the training stage, each attribute can be considered as a continuous variable during inference to choose how much a specific attribute is perceivable in the generated images.

As discussed in Sec. 2, our semantic discriminator is essentially different from conventional GANs. The feature-level competition also similar to adversarial auto-encoder [43], which match the intermediate feature with a prior distribution (Gaussian). However, we are conditioned to another vector s, and require the encoder align the distribution between any two s, instead of only real/fake.

Algorithm 1 Training the FLF framework

```
$\theta \leftarrow$initialize network parameters
repeat
    $\{x, s, y\} \leftarrow$ random mini-batch from dataset
    $d \leftarrow E_d(x)$ $l \leftarrow E_l(x)$ $\tilde{x} \leftarrow Dec(d, s, l)$
    $\mathcal{L}_{C_d} \leftarrow \mathbb{E}_{x,y\sim q(x,s,y)} - \mathtt{log}p_{C_d}(y|E_d(x))$
    $\mathcal{L}_{Dis} \leftarrow \mathbb{E}_{x,s\sim q(x,s,y)} - \mathtt{log}p_{Dis}(s|E_d(x))$
    $\mathcal{L}_{C_l} \leftarrow \mathbb{E}_{x,y\sim q(x,s,y)} - \mathtt{log}p_{C_l}(y|E_l(x))$
    $\mathcal{L}_{rec} \leftarrow \mathbb{E}_{x,s,y\sim q(x,s,y)} \|Dec(d, l, s) - x\|_2^2$
    //Update parameters according to gradients
    $\theta_{E_d} \leftarrow \nabla_{\theta_{E_d}}(\mathcal{L}_{C_d} - \alpha\mathcal{L}_{Dis} + \beta\mathcal{L}_{rec})$;
    $\theta_{E_l} \leftarrow \nabla_{\theta_{E_l}}(\lambda\mathcal{L}_{rec} - \mathcal{L}_{C_l})$; $\theta_{rec} \leftarrow \nabla_{\theta_{rec}}\mathcal{L}_{rec}$;
    $\theta_{C_d} \leftarrow \nabla_{\theta_{C_d}}\mathcal{L}_{C_d}$;$\theta_{C_l} \leftarrow \nabla_{\theta_{C_l}}\mathcal{L}_{C_l}$;$\theta_{dis} \leftarrow \nabla_{\theta_{dis}}\mathcal{L}_{dis}$
until dead line
```

Eliminating latent variation. To train the latent variation encoder E_l, we propose a novel variant of adversarial networks, in which the E_l plays a minimax game with a classifier C_l instead of a discriminator. The C_l inspects the background latent variation l and learns to predict class label correctly, while the E_l is trying to eliminate task-specific factors d by fooling C_l to make false predictions.

$$\min_{C_l} \max_{E_l} \mathcal{L}_{C_l} = \mathbb{E}_{x,y\sim q(x,s,y)} - \mathtt{log}p_{C_l}(y|E_l(x)) \quad (3)$$

Since the ground truth of d is unobservable, we use the y in here, which incorporate d and main-task relevant s. We also use softmax ouput unit and cross-entropy loss in our implementations. In contrast to using three parallel VAEs [44], the adversarial classifiers are expected to alleviate the costly training pipeline and facilitate the convergence.

Complementary constraint. The decoder *Dec* is a deconvolution network to produce a new version of the input image given the concatenated codes (d, s, l). These three parts should contain enough information to allow the reconstruction of the input x. Herein, we measure the similarity of the reconstruction with the self-regularized mean squared error (MSE) for simply:

$$\min_{E_d, E_l, Dec} \mathcal{L}_{rec} = \mathbb{E}_{x,s,y \sim q(x,s,y)} \|Dec(d, s, l) - x\|_2^2 \quad (4)$$

This design contributes to variation separation in an implicit way, and makes the encoded features more inclusive of the image content.

Independent analysis

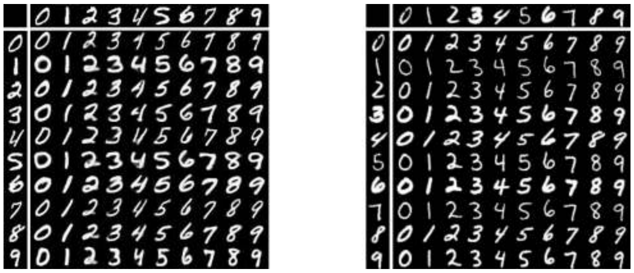
The three complementary parts are expected to uncorrelated to each other. The s is marginally independent to the d and s, since its short code cannot incorporate the other information. We learn the d to be discriminative to the main recognition task and marginally independent to s by maximizing the certainty of making main task predictions (Eqn. (1)) and uncertainty of inferring the semantic variations given the d (Eqn. (2)). Given the l, minimizing the certainty of making main task (y) predictions in Eqn. (3) can makes l marginally independent to the d and some of the task-dependent s.

Considering the complexity of our framework, we do not strictly require our learned l to be marginally independent to task-irrelevant s. The ground truth label of l also does not exist in the datasets to supervise the d to be marginally independent to latent variation l. Instead, we limit the output dimension of E_d and E_l as an information bottleneck to implicitly require d and l incorporate little unexpected information [55]. Additionally, a reconstruction loss is utilized as the complementary constraint, which avoids the d and l containing nothing.

4. Experiments

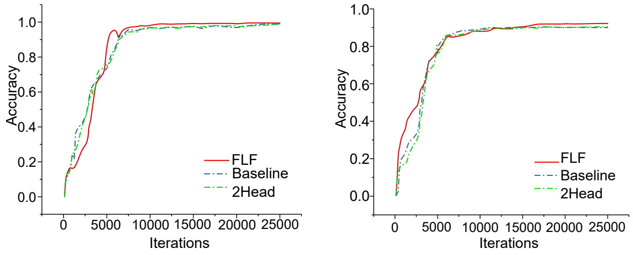
To illustrate the behavior of the **F**eature-**l**evel **F**rankenstein (FLF) framework, we quantitatively evaluate the discriminative representation with desired invariance property on three different recognition tasks and also offer qualitative evaluations by visually examining the perceptual quality of conditional face generation. As the frequent metrics (*e.g.*, log-likelihood of a set of validation samples) are not meaningful for perceptual generative models [53], we measure the information associated with the semantic variations s or main-task label y that is contained in each representation part to evaluate the degree of disentanglement as in [44, 39].

Several trade-off parameters constrained between 0 and 1 are used to balance the judiciously selected loss functions.



(a) (b)

Figure 3. A visualization grid of MNIST image swapping. We fix the semantic variation to index the MNIST dataset, while swap the discriminative representation and latent variations. The images are generated using l (writing style) from the leftmost digit and d (number) from the digit at the top of the colum using (a) our method is comparable to (b) [44] with triplet-training, using more than 3 times fewer training time.



(a) (b)

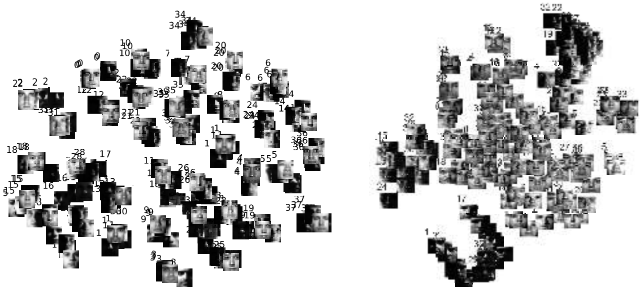
Figure 4. The digital number recognition accuracies of the proposed and baseline CNNs that are trained using MNIST+SVHN and tested on MNIST (a) and SFEW (b).

The E_l is trained to minimize the $(-\mathcal{L}_{C_l} + \lambda \mathcal{L}_{rec})$, where the λ is used to weight the relevance of the latent representation with the class label, and the quality of reconstruction.

In all our experiments, we utilize the Adam optimization method [23] with a learning rate of 0.001 and beta of 0.9 for the training of the encoders-decoder network, discriminator and classifiers. We use a variable weight for the discriminator loss coefficient α. We initially set α to 0 and the model is trained as a normal auto-encoder. Then, α is linearly increased to 0.5 over the first 500,000 iterations to slowly encourage the model to produce invariant representations. This scheduling turned out to be critical in our experiments. Without it, we observed that the E_d was too affected by the loss coming from the discriminator, even for low values of α. All the models were implemented using TensorFlow.

4.1. Source-independent digits classification

We construct a combined dataset DIGITS with the MNIST [28] and SVHN [45] to verify the ability of FLF for digit classification with the prior information about background variation. The MNIST is a digit dataset, in which each sample is a 28 × 28 black and white digit image with the class label from 0 to 9. Street View House Numbers



(a) FLF (b) Original x[59]

Figure 5. t-SNE [42] visualization of images in Extended YaleB. The original images (b) are clustered according to their lighting environments, while the discriminative representation learned by our framework (a) is more likely to cluster with only identities.

(SVHN) is a collection of house numbers collected from Google street view images, and each is a colored image with a size of 32 × 32, containing printing type digits in the natural environment. We resize SVHN images to 28 × 28 and replicate the MNIST samples to three channels to combine these two training sets. As we know, the background between these two datasets has large contrast, and should be disentangled for our digits classification task. Herein, we construct a binary variable s which takes the value of 0 if the sample is from MNIST with the clean background and a value of 1, otherwise.

We show the average accuracy of digits recognition when trained on DIGITS dataset and tested on MNIST and SVHN in Fig. 4. We adopt the encoder and classifier structure in [44] for the E_d and C_d in FLF and our baseline CNN model which does not aim for an invariant feature is used for comparison. As desired, our source-invariant representation is better at classifying digits in both sub-tasks than the baseline. The two heads multi-task network that predicts digits number and source sharing the parameters also achieves the similar result as the baseline. It is not efficient to utilize the task-independent semantic labels. Moreover, the 10-class & 10-class two-heads network require semantic labels in testing stage to choose which head should be used for a specific input. The precise semantic labels are hard to acquire in real-world applications. These networks also suffer from the same drawback that the network design will be very complicated for multiple to-be disentangle semantic variations.

In order to quantitatively measure the disentanglement, we try to measure the amount of information related to digit-classes y and the semantic variations s from the extracted representation d and l following the [44, 39]. Herein, we report the testing results in MNIST dataset for comparison in Table 1. As we can see, the latent variation representations l are almost agnostic to the class label y and semantic variation label s, while the discriminative representations d archive high recognition accuracy to predict y and incorporate few semantic variations s. We note that the FLF is trained in DIGITS, while [44] cannot support semantic variation disentanglement and is trained only in MNIST. The higher accuracy of $(y \mid d)$ than [44] is expected, since the extra information transferred from SVHN with the prior knowledge of background variance.

By fixing $s = 0$ (*i.e.,* MNIST dataset) and swapping the other two parts (*i.e.,* d and l), we get the same function as in [44], which corresponds to solving image analogies. Herein, the l component represents the hand-writing style and d focuses on the class of digits. In Fig. 3. we present our results of swapping and compare it with existing state-of-art methods. The style l and content d look well separated and the visual properties in SVHN do not appear in our generated samples. To the best of our judgment, the three disentangled parts are almost independent to each other. Despite we do not have a loss function to disentangle l from d, and the labels for the latent variation are usually unavailable in realistic scenarios, the limited output dimension of E_d achieves the separation in an efficient way. Without the triplet-training protocol, our training is much faster than [44]. On our NVIDIA K40 GPU, the loss typically converges within 20 mins for DIGITS, while the triplet-training [44] takes more than one hour for only MNIST to get the results with comparable visual quality. This gap will be more appealing when applied to a larger dataset. Note that SVHN makes little contribution to our MNIST digit generation considering the difference between hand writing style and the unified print fonts of house numbers.

4.2. Lighting-tolerant face recognition

For our lighting-tolerant classification task, we use the Extended Yale B dataset [10]. It comprises face images from 38 subjects under 5 different lighting conditions, *i.e.*, front, upper left, upper right, lower left, or lower right. We aim to predict the subject identity y using d. The semantic variable s to be purged here is the lighting condition, while the latent variation l does not have practical meaning in this dataset setting. We follow the two-layer E_d structure and train/test split of [39, 30]. 190 samples are utilized for training and all remaining 1,096 images are used for testing.

The numerical results of recognition using E_d and C_d are shown in Table 1. We compare it with the state-of-the-art methods that use MMD regularizations $etc.$, to remove the affects of lighting conditions [39, 30]. The advantage of our framework about factoring out lighting conditions is shown by the improved accuracy 90.1%, while the best baseline achieves an accuracy of 86.6%. Although the lighting conditions can be modeled very well with a Lambertian model, we choose to use a generic neural network to learn invariant features, so that the proposed method can be readily applied to other applications.

Att.Id	Attr.Def	Att.Id	Attr.Def	Att.Id	Att.Def	Att.Id	Att.Def
1	5'O Shadow	11	***Gray Hair****	21	Male	31	***Sideburns***
2	Arched Eyebr	12	Big Lips	22	***Mouth Open***	32	***Smiling***
3	Bushy Eyebr	13	Big Nose	23	***Mustache***	33	***Straight Hair***
4	***Attractive***	14	Blurry	24	Narrow Eyes	34	***Wavy Hair***
5	***Eyes Bags***	15	Chubby	25	***No Beard***	35	***Earrings***
6	***Bald****	16	Double Chin	26	Oval Face	36	***Hat***
7	***Bangs***	17	***Eyeglasses***	27	***Pale Skin***	37	***Lipstick***
8	***Black Hair****	18	***Goatee***	28	Pointy Nose	38	Necklace
9	***Blond Hair****	19	***Makeup***	29	Hairline	39	Necktie
10	***Brown Hair****	20	Cheekbones	30	***Rosy Cheeks***	40	***Young****

Table 1. Summary of the 40 face attributes provided with the CelebA and LFWA dataset. We expect the network learns to be invariant to the bolded and italicized attributes for our makeup face recognition task. *We noticed the degrades of recognition accuracy in CelebA dataset when dispelling these attributes.

Method	Accuracy on MNIST(y)/DIGITS(s)			
	$(y \mid d)$	$(s \mid d)$	$(y \mid l)$	$(s \mid l)$
Mathieu 2016	99.2%	-	13.0%	-
Proposed	**99.5%**	**52.6%***	**58.2%**	**8.1%***
Method	Accuracy on Extended YaleB			
	$(y \mid d)$	$(s \mid d)$	$(y \mid l)$	$(s \mid l)$
Original x as d	78.0%*	96.1%*	-	-
Li [30]	82%	-	-	-
Louizos [39]	84.6%	56.5%*	-	-
Daniel [5]	86.6%	47.9%*		-
Proposed	**90.1%**	**26.2%***	**8.7%**	**30.5%***

Table 2. Classification accuracy comparisons. We expect the accuracy of classifying y or s from l to be a lower value. A better discriminative representation d has a higher accuracy of classifying y and a lower accuracy in predicting s. *Following the setting in [39], we utilize the Logistics Regression classifier for the accuracy of predicting the s and using original x to predict y. The to be dispelled s represents source dataset (i.e., domain) on DIGITS, and represents lighting condition on Extened YaleB, both are main-task $irrelevant$ semantic variations.

In terms of removing s, our framework can filter the lighting conditions since the accuracy of classifying s from d drops from 56.5% to 26.2% (halved), as shown in Table 1. We note that 20% is a chance performance for 5 class illumination, when the s is totally dispelled. This can also be seen in the visualization of two-dimensional embeddings of the original x Fig. 5 (b) and the discriminative representations d extracted by FLF (a). We see that the original images are clustered based on the lighting conditions. The clustering based on CNN features are almost well according to the identity, but still affected by the lighting and results in a 'black center'. As soon as we remove the lighting variations via FLF, images are distributed almost only according to the identity of each subject.

4.3. Attributes-invariant face recognition

Makeup face recognition

We evaluate the desired makeup-invariance property of our learned discriminative representation on three makeup benchmarks. To be detailed, we train our framework using CelebA dataset [38] which is a face dataset with 202,599 face images from more than 10K subjects, with 40 different attribute labels where each label is a binary value. We adapt our E_d and C_d from VGG-16 [50], and the extracted d in testing stage are directly utilized for the open-set recognitions [34], without fine-tuning on the makeup datasets as our VGG baseline method.

PR 2017 Dataset [51] collected 406 makeup and non-makeup images from the Internet of 203 females. TCSVT 2014 dataset [12] incorporate 1002 face images. FAM dataset [17] involves 222 males and 297 females, with 1038 images belonging to 519 subjects in total. It is worth noticing that all these images are acquired under uncontrolled condition. We follow the protocol provided in [29], and the rank-1 average accuracy of FLF and state-of-the-art methods are reported in Table 3 as quantitative evaluation. The performance of [29], VGG-baseline and FLF are benefited from the large scale training dataset in CelebA. We note that the CelebA used in FLF and baseline, and even larger MS-Celeb-1M databases [13] used in [29] have incorporated several makeup variations.

With the prior information about the makeup recognition datasets, we systematically enforce our network to be invariant to the *makeup and skin*(attr.id 5,19,27,30,37), *attractive*(attr.id 4),*age*(attr.id 40), *hair*(attr.id 6-11,33,34), *beard*(attr.id 18,23,25,31), *expressions*(attr.id 22,32), *accessories*(attr.id 17,35,36), which incorporate both the id-relevant attributes (e.g., hair color) and id-irrelevant attributes (e.g., smiling/not). Dispelling these id-relevant attributes usually degrades the recognition accuracy in original CelebA dataset, but achieve better generalization ability on makeup face recognition datasets.

Since these attributes are very likely to be changed for the subjects in makeup face recognition datasets, our FLF can extracts more discriminative feature for better general-

Dataset	PR2017			TCSVT2014			FAM		
Performance	Methods	Acc	TPR	Methods	Acc	TPR	Methods	Acc	TPR
	Sun et al.[51]	68.0%	-	Sun et al.[51]	82.4%	-	Hu et al.[17]	62.4%	-
	Li et al.[29]	92.3%	38.9%	Li et al.[29]	94.8%	65.9%	Zheng et al.[65]	82.6%	-
	VGG	82.7%	34.7%	VGG	84.5%	59.5%	VGG	80.8%	48.3%
	Proposed	**94.6%**	**45.9%**	Proposed	**96.2%**	**71.4%**	Proposed	**91.4%**	**58.6%**

Table 3. Comparisons of the rank-1 accuracy and TPR@FPR=0.1% on three makeup datasets.

Methods	Rank-1 accuracy
VGG	85.4%
19-head (1ID+18attr)	81.1%
FLF	**92.7%** (↑22.7%)

Table 4. Face recognition accuracy on CelebA dataset

Methods	backbone	CelebA	LFWA
[38]	AlexNet	87.30%	83.85%
[40]	VGG-16	91.20%	-
[37]	InceptionResNet	87.82%	83.16%
[15]	ResNet50	91.81%	85.28%
FLF	VGG-16	**93.26%**	**87.82%**

Table 5. Face attribute recognition accuracy on CelebA and LFWA dataset. Two datasets are trained and tested separately.

Methods	Genuine Acceptance Rate	
	@1%FAR	@0.1%FAR
VGG [26]	33.76%	17.73%
FLF	51.78% (↑18.02%)	38.64% (↑20.91%)
MIRA [63]	89.04%	75.08%
FLF+MIRA	**91.30%** (↑2.26%)	**78.55%** (↑3.47%)

Table 6. Face recognition on DFW dataset

ization ability.

By utilizing the valuable side labels (both main-task and attributes) in CelebA in a controllable way, we achieve more than 10% improvement over the baseline, and outperforms STOA by $\geq$5.5% $w.r.t$ TPR@FPR=0.1% in all datasets.

Face and face attributes recognition

We also take the open-set identification experiments in CelebA with an ID-independent 5-fold protocol. In Table 1, we have shown which 18 attributes can increase the generalization in CelebA, while 6 attributes will degrade the accuracy in CelebA while improving the performance in Makeup face recognition. The accuracy of FLF on CelebA after dispelled 18 kinds of attributes is significantly better than its baselines. The VGG does not utilize the attribute label, the 19-head is a typical multi-task learning framework which can be distracted by task-$irrelevant\ s$. We note that CelebA is essentially an attributes datasets, there are few works use it as the testing dataset $w.r.t.$ identity.

Inversely, we can flexibly change our main-task as attribute recognition and dispel the identity information. As shown in Table 5, FLF outperforms the previous methods with a relatively simple backbone following the standard evaluation protocol of CelebA and LFWA [38] benchmarks.

The 5 hours training takes on a K40 GPU is 3× faster than pixel-level IcGAN [48, 1], without the subsequent training using the generated image for recognition and the inference time in the testing phase is the same as VGG.

Disguised face recognition

The disgust face in the wild (DFW) dataset [26] is a recently leased benchmark, which has 11157 images from 1000 subjects. The mainstream methods usually choose CelebA as pre-training dataset, despite it has a slightly larger domain gap with CelebA than these makeup datasets. In Table 6, we show the FLF can largely improve the VGG baseline by 18% and 20.9% $w.r.t$ GAR@1%FAR and GAR@0.1%FAR respectively. It can also be used as a pre-training scheme (FLF+MIRA) to complementary with the state-of-the-art methods for better performance.

5. Conclusions

This paper presents a solution to extract discriminative representation inheriting the desired invariance property for both the latent and semantic variations. Besides, the variations are factorized and maintained in their codes. Such separation can facilitate both the recognition and generation. As a result, we show that the invariant representation is learned, and the three parts are complementary to each other. In the future, we plan to facilitate several interesting works in privacy data, generation and domain adaptation, *etc.*.

6. Acknowledgement

This work was supported in part by the National Natural Science Foundation 61627819 and 61727818, Hong Kong Government General Research Fund GRF 152202/14E, PolyU Central Research Grant G-YBJW, Youth Innovation Promotion Association, CAS (2017264), Innovative Foundation of CIOMP, CAS (Y586320150).

References

[1] Gans comparison without cherry-picking. https://github.com/khanrc/tf.gans-comparison. 8

[2] J. Bao, D. Chen, F. Wen, H. Li, and G. Hua. Towards open-set identity preserving face synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6713–6722, 2018. 3

[3] Y. Bengio et al. Learning deep architectures for ai. *Foundations and trends® in Machine Learning*, 2(1):1–127, 2009. 3

[4] J. Cao, O. Katzir, P. Jiang, D. Lischinski, D. Cohen-Or, C. Tu, and Y. Li. Dida: Disentangled synthesis for domain adaptation. *arXiv preprint arXiv:1805.08019*, 2018. 2

[5] M. Daniel, G. Shuyang, B. Rob, V. S. Greg, and G. Aram. Invariant representations without adversarial training. In *NIPS*, 2018. 2, 3, 7

[6] H. Dong, P. Neekhara, C. Wu, and Y. Guo. Unsupervised image-to-image translation with generative adversarial networks. *arXiv preprint arXiv:1701.02676*, 2017. 1, 2

[7] H. Edwards and A. Storkey. Censoring representations with an adversary. *arXiv preprint arXiv:1511.05897*, 2015. 3

[8] A. Elgammal and C.-S. Lee. Separating style and content on a nonlinear manifold. In *CVPR*, volume 1, 2004. 3

[9] Y. Fu, T. M. Hospedales, T. Xiang, and S. Gong. Learning multimodal latent attributes. *IEEE TPAMI*, 36(2):303–316, 2014. 2

[10] A. S. Georghiades, P. N. Belhumeur, and D. J. Kriegman. From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE TPAMI*, 23(6):643–660, 2001. 6

[11] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *NIPS*, pages 2672–2680, 2014. 2

[12] G. Guo, L. Wen, and S. Yan. Face authentication with makeup changes. *IEEE Transactions on Circuits and Systems for Video Technology*, 24(5):814–825, 2014. 7

[13] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *ECCV*, pages 87–102. Springer, 2016. 7

[14] N. Hadad, L. Wolf, and M. Shahar. A two-step disentanglement method. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 772–780, 2018. 2, 3

[15] K. He, Y. Fu, W. Zhang, C. Wang, Y.-G. Jiang, F. Huang, and X. Xue. Harnessing synthesized abstraction images to improve facial attribute recognition. In *IJCAI*, pages 733–740, 2018. 8

[16] G. Hu, Y. Hua, Y. Yuan, Z. Zhang, Z. Lu, S. S. Mukherjee, T. M. Hospedales, N. M. Robertson, and Y. Yang. Attribute-enhanced face recognition with neural tensor fusion networks. In *ICCV*, 2017. 2

[17] J. Hu, Y. Ge, J. Lu, and X. Feng. Makeup-robust face verification. In *ICASSP*, pages 2342–2346. IEEE, 2013. 7, 8

[18] Q. Hu, A. Szabó, T. Portenier, M. Zwicker, and P. Favaro. Disentangling factors of variation by mixing them. *arXiv preprint arXiv:1711.07410*, 2017. 3

[19] R. Huang, S. Zhang, T. Li, R. He, et al. Beyond face rotation: Global and local perception gan for photorealistic and identity preserving frontal view synthesis. *arXiv preprint arXiv:1704.04086*, 2017. 2

[20] D. Jayaraman, F. Sha, and K. Grauman. Decorrelating semantic visual attributes by resisting the urge to share. In *CVPR*, pages 1629–1636, 2014. 1

[21] A. H. Jha, S. Anand, M. Singh, and V. Veeravasarapu. Disentangling factors of variation with cycle-consistent variational auto-encoders. *arXiv preprint arXiv:1804.10469*, 2018. 3

[22] H. Jiang, R. Wang, S. Shan, Y. Yang, and X. Chen. Learning discriminative latent attributes for zero-shot classification. In *CVPR*, pages 4223–4232, 2017. 2

[23] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[24] D. P. Kingma, S. Mohamed, D. J. Rezende, and M. Welling. Semi-supervised learning with deep generative models. In *NIPS*, pages 3581–3589, 2014. 3

[25] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 3

[26] V. Kushwaha, M. Singh, R. Singh, M. Vatsa, N. Ratha, and R. Chellappa. Disguised faces in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, volume 8, 2018. 8

[27] G. Lample, N. Zeghidour, N. Usunier, A. Bordes, L. Denoyer, et al. Fader networks: Manipulating images by sliding attributes. In *NIPS*, pages 5969–5978, 2017. 3

[28] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 5

[29] Y. Li, L. Song, X. Wu, R. He, and T. Tan. Anti-makeup: Learning a bi-level adversarial network for makeup-invariant face verification. *AAAI*, 2018. 7, 8

[30] Y. Li, K. Swersky, and R. Zemel. Learning unbiased features. *arXiv preprint arXiv:1412.5244*, 2014. 3, 6, 7

[31] Y. Li, X. Tian, M. Gong, Y. Liu, T. Liu, K. Zhang, and D. Tao. Deep domain generalization via conditional invariant adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 624–639, 2018. 2

[32] Y. Li, R. Wang, H. Liu, H. Jiang, S. Shan, and X. Chen. Two birds, one stone: Jointly learning binary code for large-scale face image retrieval and attributes prediction. In *ICCV*, pages 3819–3827, 2015. 2

[33] M.-Y. Liu, T. Breuel, and J. Kautz. Unsupervised image-to-image translation networks. In *Advances in Neural Information Processing Systems*, pages 700–708, 2017. 1, 2

[34] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song. Sphereface: Deep hypersphere embedding for face recognition. In *CVPR*, volume 1, 2017. 7

[35] X. Liu, B. Kumar, J. You, and P. Jia. Adaptive deep metric learning for identity-aware facial expression recognition. In *CVPRW*, pages 522–531, 2017. 2

[36] Y. Liu, Z. Wang, H. Jin, and I. Wassell. Multi-task adversarial network for disentangled feature learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3743–3751, 2018. 2

[37] Y. Liu, F. Wei, J. Shao, L. Sheng, J. Yan, and X. Wang. Exploring disentangled feature representation beyond face identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2080–2089, 2018. 3, 8

[38] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *ICCV*, pages 3730–3738, 2015. 7, 8

[39] C. Louizos, K. Swersky, Y. Li, M. Welling, and R. Zemel. The variational fair autoencoder. *arXiv preprint arXiv:1511.00830*, 2015. 2, 3, 5, 6, 7

[40] Y. Lu, A. Kumar, S. Zhai, Y. Cheng, T. Javidi, and R. S. Feris. Fully-adaptive feature sharing in multi-task networks with applications in person attribute classification. In *CVPR*, volume 1, page 6, 2017. 8

[41] L. Ma, Q. Sun, S. Georgoulis, L. Van Gool, B. Schiele, and M. Fritz. Disentangled person image generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 99–108, 2018. 2

[42] L. v. d. Maaten and G. Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9:2579–2605, 2008. 6

[43] A. Makhzani, J. Shlens, N. Jaitly, I. Goodfellow, and B. Frey. Adversarial autoencoders. *arXiv preprint arXiv:1511.05644*, 2015. 4

[44] M. F. Mathieu, J. J. Zhao, J. Zhao, A. Ramesh, P. Sprechmann, and Y. LeCun. Disentangling factors of variation in deep representation using adversarial training. In *NIPS*, pages 5040–5048, 2016. 1, 2, 3, 4, 5, 6

[45] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Y. Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep learning and unsupervised feature learning*, volume 2011, page 5, 2011. 5

[46] P. Peng, Y. Tian, T. Xiang, Y. Wang, M. Pontil, and T. Huang. Joint semantic and latent attribute modelling for cross-class transfer learning. *IEEE transactions on pattern analysis and machine intelligence*, 2017. 1

[47] X. Peng, X. Yu, K. Sohn, D. N. Metaxas, and M. Chandraker. Reconstruction-based disentanglement for pose-invariant face recognition. *intervals*, 20:12, 2017. 2

[48] G. Perarnau, J. van de Weijer, B. Raducanu, and J. M. Álvarez. Invertible conditional gans for image editing. *arXiv preprint arXiv:1611.06355*, 2016. 8

[49] V. Sharmanska, N. Quadrianto, and C. H. Lampert. Augmented attribute representations. In *ECCV*, pages 242–255. Springer, 2012. 2

[50] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 7

[51] Y. Sun, L. Ren, Z. Wei, B. Liu, Y. Zhai, and S. Liu. A weakly supervised method for makeup-invariant face verification. *Pattern Recognition*, 66:153–159, 2017. 7, 8

[52] J. B. Tenenbaum and W. T. Freeman. Separating style and content. In *NIPS*, pages 662–668, 1997. 3

[53] L. Theis, A. v. d. Oord, and M. Bethge. A note on the evaluation of generative models. *arXiv preprint arXiv:1511.01844*, 2015. 5

[54] Y. Tian, X. Peng, L. Zhao, S. Zhang, and D. N. Metaxas. Cr-gan: learning complete representations for multi-view generation. *arXiv preprint arXiv:1806.11191*, 2018. 2

[55] N. Tishby and N. Zaslavsky. Deep learning and the information bottleneck principle. In *Information Theory Workshop (ITW), 2015 IEEE*, pages 1–5. IEEE, 2015. 5

[56] L. Tran, X. Yin, and X. Liu. Disentangled representation learning gan for pose-invariant face recognition. In *CVPR*, volume 3, page 7, 2017. 2, 3

[57] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell. Adversarial discriminative domain adaptation. In *CVPR*, volume 1, page 4, 2017. 2

[58] T. Xiao, J. Hong, and J. Ma. Dna-gan: Learning disentangled representations from multi-attribute images. *arXiv preprint arXiv:1711.05415*, 2017. 3

[59] Q. Xie, Z. Dai, Y. Du, E. Hovy, and G. Neubig. Adversarial invariant feature learning. In *NIPS*, pages 585–596, 2017. 1, 2, 3, 6

[60] X. Yin, X. Yu, K. Sohn, X. Liu, and M. Chandraker. Towards large-pose face frontalization in the wild. In *Proc. ICCV*, pages 1–10, 2017. 2

[61] W. Zellinger, T. Grubinger, E. Lughofer, T. Natschläger, and S. Saminger-Platz. Central moment discrepancy (cmd) for domain-invariant representation learning. *arXiv preprint arXiv:1702.08811*, 2017. 2

[62] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork. Learning fair representations. In *International Conference on Machine Learning*, pages 325–333, 2013. 3

[63] K. Zhang, Y.-L. Chang, and W. Hsu. Deep disguised faces recognition. In *CVPR Workshop on Disguised Faces in the Wild*, volume 4, page 5, 2018. 8

[64] J. Zhao, M. Mathieu, R. Goroshin, and Y. Lecun. Stacked what-where auto-encoders. *arXiv preprint arXiv:1506.02351*, 2017. 3

[65] Z. Zheng and C. Kambhamettu. Multi-level feature learning for face recognition under makeup changes. In *Automatic Face & Gesture Recognition (FG 2017), 2017*, pages 918–923. IEEE, 2017. 8