Scaling-up Disentanglement for Image Translation

Aviv Gabbay   Yedid Hoshen

School of Computer Science and Engineering

The Hebrew University of Jerusalem, Israel

Project webpage: http://www.vision.huji.ac.il/overlord

Abstract

*Image translation methods typically aim to manipulate a set of labeled attributes (given as supervision at training time e.g. domain label) while leaving the unlabeled attributes intact. Current methods achieve either: (i) disentanglement, which exhibits low visual fidelity and can only be satisfied where the attributes are perfectly uncorrelated. (ii) visually-plausible translations, which are clearly not disentangled. In this work, we propose OverLORD, a single framework for disentangling labeled and unlabeled attributes as well as synthesizing high-fidelity images, which is composed of two stages; (i) **Disentanglement**: Learning disentangled representations with latent optimization. Differently from previous approaches, we do not rely on adversarial training or any architectural biases. (ii) **Synthesis**: Training feed-forward encoders for inferring the learned attributes and tuning the generator in an adversarial manner to increase the perceptual quality. When the labeled and unlabeled attributes are correlated, we model an additional representation that accounts for the correlated attributes and improves disentanglement. We highlight that our flexible framework covers multiple settings as disentangling labeled attributes, pose and appearance, localized concepts, and shape and texture. We present significantly better disentanglement with higher translation quality and greater output diversity than state-of-the-art methods.*

1. Introduction

Learning disentangled representations for different factors of variation in a set of observations is a fundamental problem in machine learning. Such representations can facilitate generalization to downstream generative and discriminative tasks such as novel image synthesis [39] and person re-identification [9], as well as improving interpretability [14], reasoning [34] and fairness [6]. A popular task which benefits from disentanglement is image translation, in which the goal is to translate a given input image from a source domain (e.g. cats) to an analogous image in a target domain (e.g. dogs). Although this task is generally poorly specified, it is often satisfied under the assumption that images in different domains share common attributes (e.g. head pose) which can remain unchanged during translation. In this paper, we divide the set of all attributes that define a target image precisely into two subsets; (i) *labeled* attributes: the attributes that are supervised at training time e.g. whether the image belongs to the "cats" or "dogs" domain; (ii) *unlabeled* attributes: all the remaining attributes that we do not have supervision for e.g. breed of the animal, head pose, background etc. While several methods (e.g. LORD [10], StarGAN [4], Fader Networks [20]) have been proposed for disentangling labeled and unlabeled attributes, we explain and show why they can not deal with cases where the labeled and unlabeled attributes are correlated. For example, when translating cats to dogs, the unlabeled breed (which specifies fur texture or facial shape) is highly correlated with the domain label (cat or dog), and can not be translated between species. This dependency motivates the specification of more fine-grained attributes that we wish the translated image to have. Several methods as MUNIT [16], FUNIT [23] and StarGAN-v2 [5] attempt to learn a domain-dependent style representation which ideally should account for the correlated attributes. Unfortunately, we show that despite their visually pleasing results, the translated images still retain many domain-specific attributes of the source image. As demonstrated in Fig. 2, when translating dogs to wild animals, current methods transfer facial shapes which are unique to dogs and should not be transferred precisely to wild animals, while our model transfers the semantic head pose more reliably.

In this work, we analyze the different settings for disentanglement of labeled and unlabeled attributes. In the case where they are perfectly uncorrelated, we improve over LORD [10] and scale to higher perceptual quality by a novel synthesis stage. In cases where the labeled and unlabeled attributes are correlated, we rely on the existence of spatial transformations which can retain the correlated attributes
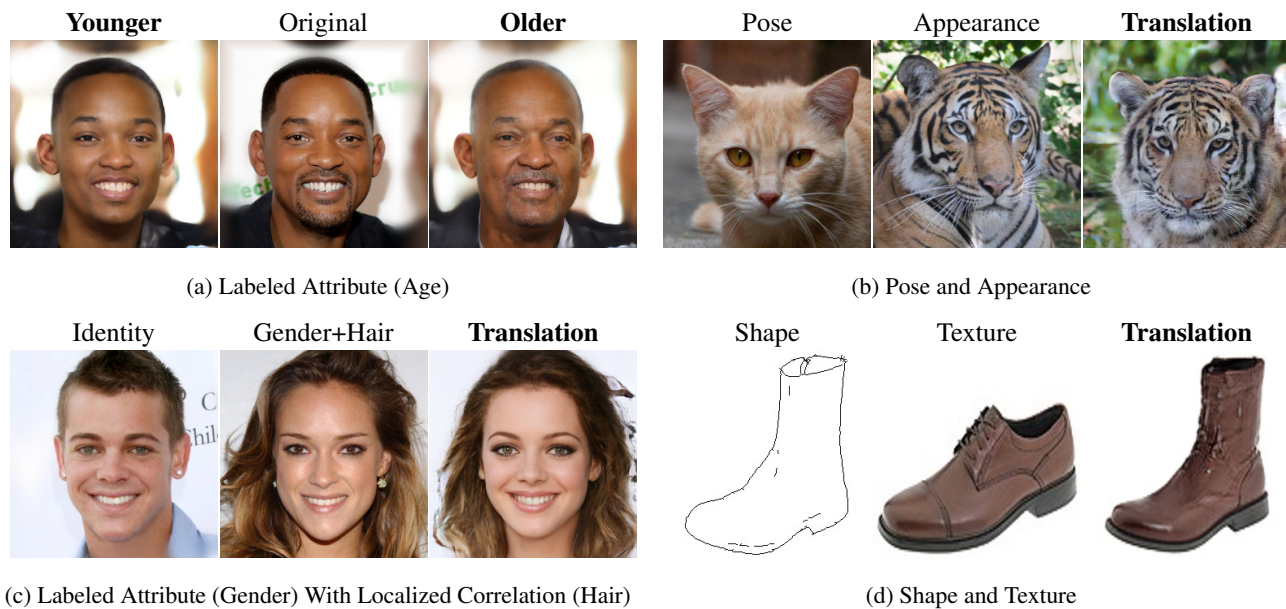
Younger	Original	Older

(a) Labeled Attribute (Age)

Pose	Appearance	**Translation**

(b) Pose and Appearance

Identity	Gender+Hair	**Translation**

(c) Labeled Attribute (Gender) With Localized Correlation (Hair)

Shape	Texture	**Translation**

(d) Shape and Texture

Figure 1: A summary of the different attribute forms covered by our disentanglement framework.

while exhibiting different or no uncorrelated attributes. We suggest simple forms of transformations for learning *pose-independent* or *localized* correlated attributes, by which we achieve better disentanglement both quantitatively and qualitatively than state-of-the-art methods (e.g. FUNIT [23] and StarGAN-v2 [5]). Our approach suggests that adversarial optimization, which is typically used for domain translation, is not necessary for disentanglement, and its main utility lies in generating perceptually pleasing images. Fig. 1 summarizes the settings covered by our framework.

Our contributions are as follows: (i) Introducing a non-adversarial disentanglement method that carefully extends to cases where the attributes are correlated. (ii) Scaling disentanglement methods to high perceptual quality with a final synthesis stage while learning disentangled representations. (iii) State-of-the-art results in multiple image-translation settings within a unified framework.

2. Related Work

Image Translation Translating the content of images across different domains has attracted much attention. In the unsupervised setting, CycleGAN [38] introduces a cycle consistency loss to encourage the translated images to preserve the domain-invariant attributes (e.g. pose) of the source image. MUNIT [16] recognizes that a given content image could be translated to many different styles (e.g. colors and textures) in a target domain and extends UNIT [15] to learn multi-modal mappings by learning style representations. DRIT [22] tackles the same setting using an adversarial constraint at the representation level. MSGAN [26] adds

a regularization term to prevent mode collapse. StarGAN-v2 [5] and DMIT [36] extend previous frameworks to translation across more than two domains. FUNIT [23] further allows translation to novel domains. COCO-FUNIT [31] aims to preserve the structure of the image by conditioning the style of a reference image on the actual content image.

Class-Supervised Disentanglement In this parallel line of work, the goal is to anchor the semantics of each class into a single representation while modeling all the remaining class-independent attributes by a residual representation. Several methods encourage disentanglement by adversarial constraints [8, 32, 27] while other rely on cycle consistency [12] or group accumulation [2]. LORD [10] takes a non-adversarial approach and trains a generative model while directly optimizing over the latent codes. Methods in this area assume that the class label and the unlabeled attributes are perfectly uncorrelated. We show that in cases where this assumption does not hold, these methods fail to disentangle the attributes and perform poorly. Moreover, they work on synthetic and low-resolution datasets and cannot compete in recent image translation benchmarks.

Disentanglement of Shape and Texture Several methods aim to disentangle the shape of an object from its texture. Lorenz *et al.* [25] exploit equivariance and invariance constraints between synthetically transformed images to learn shape and appearance representations. Park *et al.* [29] learn the texture by co-occurrence patch statistics across different parts of the image. In this setting, no supervision is used, limiting current methods to low-level manipulations i.e. no significant semantic changes are performed.

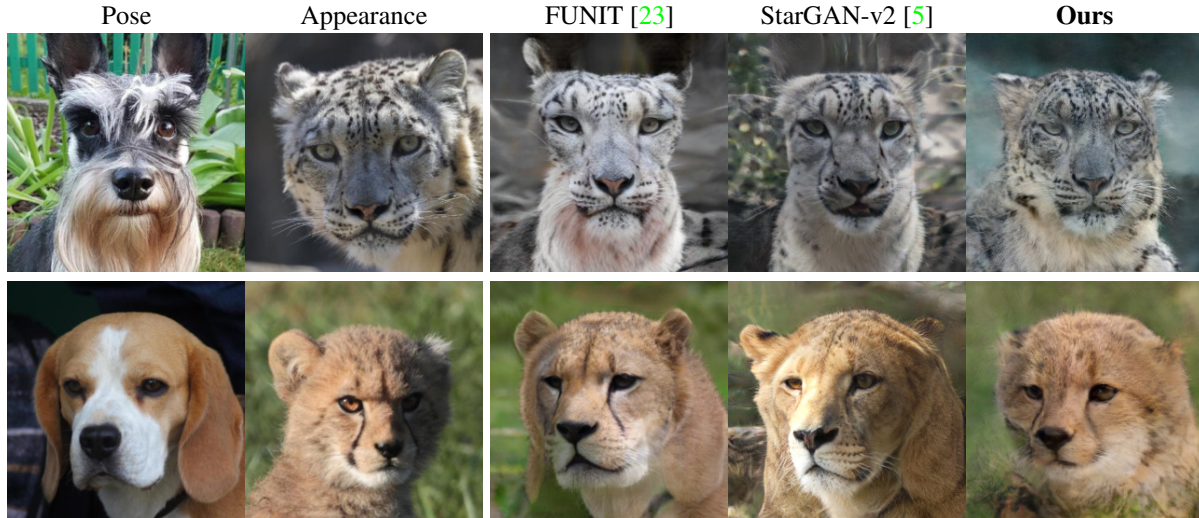| Pose | Appearance | FUNIT [23] | StarGAN-v2 [5] | **Ours** |

Figure 2: Entanglement in image translation models on AFHQ. Domain label (cat, dog or wild) and its correlated attributes (e.g. breed) are guided by the appearance image, while the uncorrelated attributes (e.g. head pose) of the source image should be preserved. Current approaches and their architectural biases tightly preserve the original structure and generate unreliable facial shapes which are unique to the source domain. We disentangle the pose and capture the target breed faithfully.

3. Disentanglement in Image Translation

Assume a set of images $x_1, x_2, ..., x_n \in \mathcal{X}$ in which every image x_i is specified precisely by a set of attributes. Some of the attributes are labeled by y_i, while the rest are unlabeled and denoted as u_i.

Case 1: If the labeled and unlabeled attributes are *approximately uncorrelated*, we seek representations u^{uncorr} for the unlabeled attributes. An unknown function G maps y and u^{uncorr} to x:

$$x_i = G(y_i, u_i^{uncorr}) \qquad (1)$$

Case 2: If the labeled and unlabeled attributes are correlated, y and u^{uncorr} do not uniquely specify x. For example, in animal face images, y defines the species (e.g. cat or dog) and u^{uncorr} specifies uncorrelated attributes such as head pose. The attributes which are correlated with species (e.g. breed) are denoted as u^{corr}, and should be learned as well. G now maps y, u^{corr} and u^{uncorr} to x:

$$x_i = G(y_i, u_i^{corr}, u_i^{uncorr}) \qquad (2)$$

The goal of image translation is to replace the labeled attributes y_i along with their correlated attributes u_i^{corr} of source image x_i by those of target image x_j. Let us briefly review the main ideas proposed in previous work and discuss their drawbacks.

Learning the uncorrelated attributes u^{uncorr} Adversarial methods (e.g. StarGAN-v2, FUNIT, MUNIT) train a domain confusion discriminator on the translated images to retain only the uncorrelated attributes. Unfortunately, we empirically show in Tab. 2 and Fig. 2 that these methods do not learn disentangled representations and the translated images do not retain only the uncorrelated attributes. However, some of the correlated attributes leak into the representations and the translated images. We hypothesize that this failure lies in the challenging adversarial optimization.

Learning the correlated attributes u^{corr} Current methods rely on locality-preserving architectures (e.g. AdaIN [5, 23, 16]) that bias the uncorrelated attributes to represent the spatial structure of the image and the correlated attributes to control its appearance. Unfortunately, architectural biases of this sort inevitably restrict the correlated attributes from altering spatial features when needed. As shown in Fig. 2, the uncorrelated pose is indeed spatial, but some of the correlated attributes (as the facial shape of different breeds) are of spatial nature as well.

4. Proposed Framework: OverLORD

Motivated by the above analysis, we present OverLORD, which is composed of: *disentanglement* and *synthesis*.

4.1. Learning a Disentanglement Model

In order to learn disentangled representations for the labeled and unlabeled attributes, we draw inspiration from
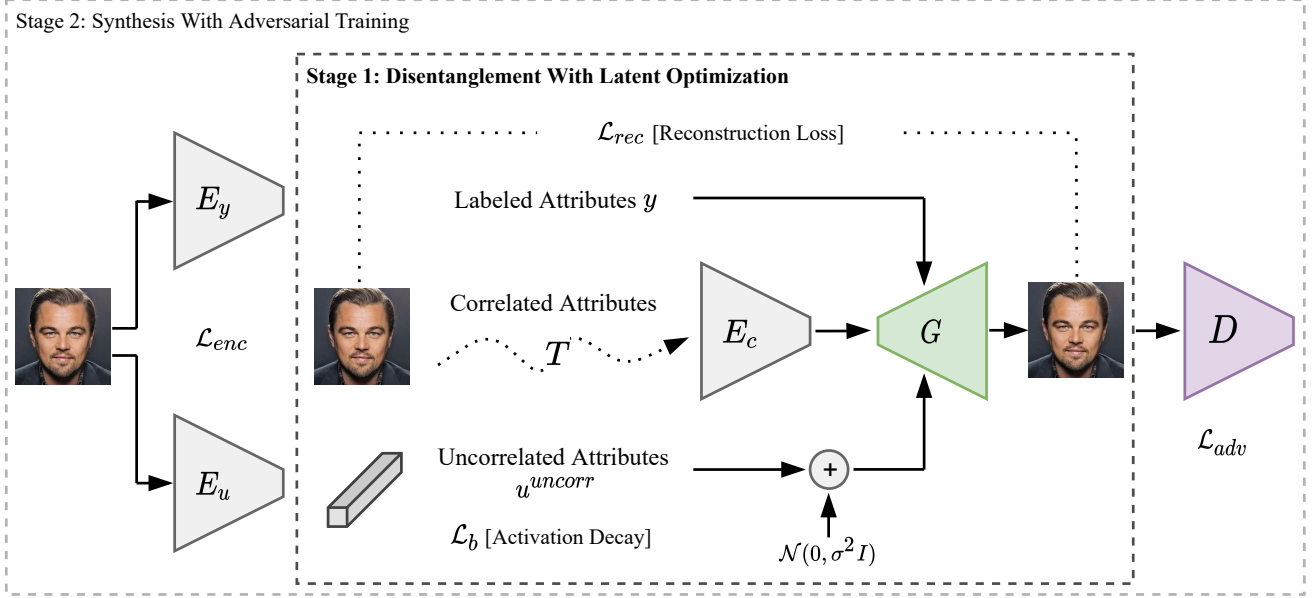
Figure 3: A sketch of our method. In the disentanglement stage, T outputs a transformed version of the image, retaining only attributes that correlate with y. u^{uncorr} is regularized and optimized to recover the minimal residual information required for reconstructing the input image, resulting in the remaining uncorrelated attributes, as they are not represented by y or the output of T. In the synthesis stage, we tune all modules in an amortized fashion using the learned embeddings as targets of the new encoders, E_y and E_u. During this stage, an adversarial discriminator is trained to increase the visual fidelity.

LORD [10] and propose a non-adversarial method based on latent optimization. Differently from LORD, we do not strictly assume that the labeled and unlabeled attributes are uncorrelated. We *relax* this assumption and assume the existence of correlated attributes, as described next.

Unlabeled correlated attributes We design a simple yet effective method for learning some of the correlated attributes independently from the uncorrelated ones. We first form a function T that outputs an image $x^{corr} = T(x)$, which retains the correlated attributes but exhibits different uncorrelated attributes. The exact realization of T depends on the nature of the uncorrelated and correlated attributes. In this paper, we propose two different forms of T:

(i) *Random spatial transformation*: If the uncorrelated attributes should capture the *pose* of an object, wet set T as a sequence of random flips, rotations or crops, making x^{corr} retain attributes that are *pose-independent*:

$$T(x) = (f_{crop} \circ f_{rotate} \circ f_{flip})(x) \qquad (3)$$

(ii) *Masking*: If the correlated attributes are *localized*, we set T to mask-out the uncorrelated attributes, retaining only correlated regions included in the segmentation mask m:

$$T(x; m) = x \odot m \qquad (4)$$

E.g. when translating males to females, T might mask the hair region which is highly correlated with the gender (i.e. masking out all the other regions). Masks can be obtained using external supervision or unsupervised methods [17, 33] and are used only for training. The correlated attributes are modeled by: $u^{corr} = E_c(x^{corr}) = E_c(T(x))$.

Unlabeled uncorrelated attributes As T accounts for the dependency of the labeled and unlabeled attributes, we can assume that u^{uncorr} is *approximately uncorrelated* with y and u^{corr}. To obtain an independent representation, we aim u^{uncorr} to recover the *minimal residual* information, not represented by y and u^{uncorr}, that is required to reconstruct the image. We therefore parameterize a noisy channel consisting of a vector u_i' and an additive Gaussian noise $z \sim \mathcal{N}(0, I)$, $u_i^{uncorr} = u_i' + z$, similarly to the non-adversarial *bottleneck* proposed in [10]. We therefore define the bottleneck loss as: $\mathcal{L}_b = \sum_i \|u_i'\|^2$.

Reconstruction: An image x_i should be fully specified by the representations $y_i, u_i^{corr}, u_i^{uncorr}$:

$$\mathcal{L}_{rec} = \sum_i \ell\big(G(y_i, u_i^{corr}, u_i^{uncorr}), x_i\big) \qquad (5)$$

Our complete objective can be summarized as follows:

$$\min_{u_i', E_c, G} \mathcal{L}_{disent} = \mathcal{L}_{rec} + \lambda_b \mathcal{L}_b \qquad (6)$$

For reconstruction we use ℓ as the VGG-based perceptual loss. Note that we optimize over u'_i directly, as they are *not* parameterized by a feed-forward encoder. Inspired by [10], we leverage latent optimization as it improves disentanglement significantly over encoder-based methods. To get an intuition, we should consider their initialization: each u'_i in latent optimization is initialized i.i.d, and therefore is totally uncorrelated with the labeled attributes. However, a feed-forward encoder starts with near perfect correlation (the labeled attributes can be predicted even from the output of a randomly initialized encoder). We further elaborate on latent optimization and its inductive bias in Appendix A.4.

4.2. Generalization and Perceptual Quality

After the disentanglement stage, we possess representations of the uncorrelated attributes u_i^{uncorr} for every image in the training set. In order to generalize to unseen images and novel labeled attributes (e.g. new face identities), we train feed-forward encoders $E_y : \mathcal{X} \to \mathcal{Y}$ and $E_u : \mathcal{X} \to \mathcal{U}$ to infer the labeled and uncorrelated attributes (E_c is already trained in the previous stage), respectively:

$$\mathcal{L}_{enc} = \sum_i \|E_y(x_i) - y_i\|^2 + \|E_u(x_i) - u_i^{uncorr}\|^2 \quad (7)$$

The reconstruction term is changed accordingly:

$$\mathcal{L}_{gen} = \sum_i \ell\Big(G\big(E_y(x_i), E_c(x_i), E_u(x_i)\big), x_i\Big) \quad (8)$$

These two objectives ensure that the training set can be reconstructed in an amortized fashion ($\mathcal{L}_{gen}$), without violating the disentanglement criterion established in the previous stage ($\mathcal{L}_{enc}$). Note that the targets for $\mathcal{L}_{enc}$ are the ones learned by our own model in the previous stage.

Although these constraints are sufficient for disentanglement and generalization, the generated images exhibit relatively low perceptual quality, as can be seen in Fig. 4. While we argue that achieving disentanglement by adversarial methods is notoriously difficult, as observed in our experiments and in [10], we find that tuning the model with an adversarial term after disentangled representations have been already learned significantly improves the perceptual fidelity. Therefore, we jointly train an unconditional discriminator D and employ an adversarial loss to distinguish between real and reconstructed images:

$$\mathcal{L}_{adv} = \sum_i \log D(x_i) + \log\big(1 - D(\bar{x}_i)\big) \quad (9)$$

The entire objective of the synthesis stage is as follows:

$$\min_{E_y, E_u, G} \max_D \quad \mathcal{L}_{gen} + \lambda_{enc}\mathcal{L}_{enc} + \lambda_{adv}\mathcal{L}_{adv} \quad (10)$$

A sketch of our entire framework is shown in Fig. 3.

5. Experiments

Our framework is evaluated in the two cases considered along this paper: when the labeled and unlabeled attributes are uncorrelated, and when they are correlated.

5.1. Evaluation Protocol

We measure disentanglement by training two auxiliary classifiers to predict the original labeled attributes from: (i) the learned representations of the uncorrelated attributes. (ii) the translated images in which the labeled attributes have been changed. In both metrics, lower accuracy indicates better disentanglement. In cases where annotations for evaluation exist (e.g. CelebA) we measure task-specific metrics as identity similarity (Id), expression (Exp) and head pose (Pose) errors given a target image. Where no annotations exist (e.g. AFHQ, CelebA-HQ) we measure how the translated images fit the target domain (FID) and their diversity (LPIPS). More implementation and evaluation details are provided in Appendix A.1.

5.2. Uncorrelated Attributes

Attribute Manipulation In this setting there is a single labeled attribute which is assumed to be *approximately uncorrelated* with all the unlabeled attributes. We perform an experiment on CelebA [24] and define the facial identity label as the only labeled attribute. Tab. 1 and Fig. 5 show results of disentangling face identity from the unlabeled attributes including head pose, expression and illumination. We compare to LORD and FUNIT as they can handle fine-grained labels (i.e. 10K face identities) and generalize to unseen identities. It can be seen that our method preserves the head pose, expression and illumination well while translating the identity more faithfully than the baselines. Another task which has attracted much attention recently is facial age editing [1, 28]. We define the age (categorized to 8 ranges: 0-9, ..., 70-79) as the only labeled attribute in FFHQ and compare to state-of-the-art aging methods: lifespan [28] and SAM [1]. We demonstrate in Fig. 6 that while the baselines rely on a pretrained face identity loss [7] we are able to preserve the identity better. Finally, we explore the task of translating Males to Females. We define the gender in CelebA-HQ [18] as the labeled attribute and compare to FaderNetworks [20] and mGANprior [11] which operates in the latent space of StyleGAN. As shown in Tab. 3 and Fig. 7, our model achieves near perfect score in fooling a target classifier and generates visually pleasing results. We later show that the assumption that the gender is uncorrelated with the other attributes can be relaxed by masking regions (e.g. hair) that are highly correlated with the gender to improve translation control and reliability.

Shape-Texture In this task the goal is to disentangle the shape of an object from its texture. We define the shape

Table 1: Disentanglement performance on CelebA. Lower classification accuracy of identity from the learned representations of the uncorrelated attributes ($y \leftarrow u^{uncorr}$) and higher error of landmarks regression from identity representations ($y \rightarrow lnd$) indicate better disentanglement. Id = FaceNet cosine similarity, Exp = Expression by RMSE on facial landmarks, Pose = average of yaw, pitch and roll angle errors. *Indicates detail loss.

	Representation		Image		
	$y \leftarrow u^{uncorr}$	$y \rightarrow lnd$	Id	Exp	Pose
FUNIT	0.16	2.6	0.24	3.8	4.7
LORD	10^{-4}*	**3.6**	0.48	3.2	3.5
Ours	10^{-3}	**3.6**	**0.63**	**2.7**	**2.5**
Optimal	10^{-3}	-	1	0	0

as the labeled attribute (represented by an edge map) and demonstrate texture transfer between different shoes on Edges2Shoes [35] in Fig. 1 and Appendix A.3. Note that in order to guide the labeled attribute by an image (and not by a categorical label), we train E_y with the rest of the models in the disentanglement stage instead of the synthesis stage.

5.3. Correlated Attributes

Pose-Appearance When the uncorrelated attributes should encode the pose of an object, we set T to be a series of random transformation of horizontal flips, rotations and crops to let x^{corr} retain all the *pose-independent* attributes. We perform an experiment on AFHQ and define the domain label (cat, dog, wildlife) as the labeled attribute. Tab. 2 and Fig. 2 show that our method outperforms all baselines achieving near perfect disentanglement with better visual quality (FID) and higher translation diversity (LPIPS). Note that while StarGAN-v2 and FUNIT support multiple domains, MUNIT, DRIT and MSGAN are trained multiple times for every possible pair of domains. Moreover, as standard LORD does not distinct between correlated and uncorrelated attributes i.e. can not utilize a reference image, we make an effort to extend it to the correlated case by clustering the images into 512 sub-categories before training (denoted as LORD clusters).

Localized correlation When the correlated attributes are *localized*, T masks their corresponding regions within the image. We repeat the Male-to-Female experiment with masks provided in CelebAMask-HQ [21]. In this experiment, x^{corr} contains only the hair region. As shown in Fig. 7, our method translates the gender with the general target hair style while preserving the uncorrelated attributes including identity, age and illumination better than StarGAN-v2. More results are in Appendix A.3.

Table 2: Disentanglement performance on AFHQ. Classification accuracy of domain label from the learned representations of the uncorrelated attributes ($y \leftarrow u^{uncorr}$), classification accuracy of source domain label from the translated image ($y_i \leftarrow x_{ij}$), translation fidelity (FID) and translation diversity (LPIPS).

	Rep.	Image		
	$y \leftarrow u^{uncorr}$	$y_i \leftarrow x_{ij}$	FID	LPIPS
MUNIT [16]	1.0	1.0	223.9	0.20
DRIT [22]	1.0	1.0	114.8	0.16
MSGAN [26]	1.0	1.0	69.8	0.38
LORD [10]	0.74	0.47	97.1	0
LORD clusters	0.53	0.43	37.1	0.36
StarGAN-v2	0.89	0.75	19.8	0.43
FUNIT [23]	0.94	0.85	18.8	0.44
Ours w/o x^{corr}	0.80	0.79	55.9	0
Ours w/o adv.	**0.33**	**0.38**	29.1	0.45
Ours	**0.33**	0.42	**16.5**	**0.51**
Optimal	0.33	0.33	12.9	-

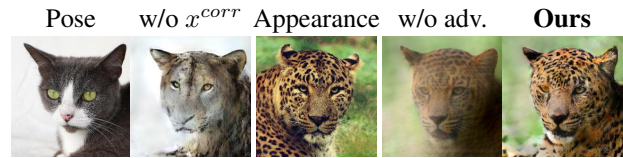Pose   w/o x^{corr} Appearance   w/o adv.   **Ours**

Figure 4: Ablations: w/o x^{corr}: Ignoring the correlated attributes leads to unreliable translations. w/o adv.: the attributes are disentangled but images exhibit low quality. Adversarial synthesis greatly improves perceptual quality.

5.4. Ablation Study

Ignoring correlation with the labeled attribute We train our model on AFHQ without x^{corr}, assuming all the unlabeled attributes are uncorrelated with the labeled attribute. However, as breed is highly correlated with species, the animal faces are translated in an unreliable and entangled manner, as shown in Fig. 4 and in Appendix. A.5. Tab. 2 includes results from the ablation analysis which suggest that this strategy does not reach disentanglement neither at the representation nor at the image level.

Adversarial loss for perceptual quality We train our method without the adversarial loss in the synthesis stage. Tab. 2 suggests that disentanglement is achieved by our non-adversarial framework, while the additional adversarial loss contributes to increasing the output fidelity (FID). Qualitative evidence for the claims in the ablation analysis are presented in Fig. 4 and in Appendix A.5.
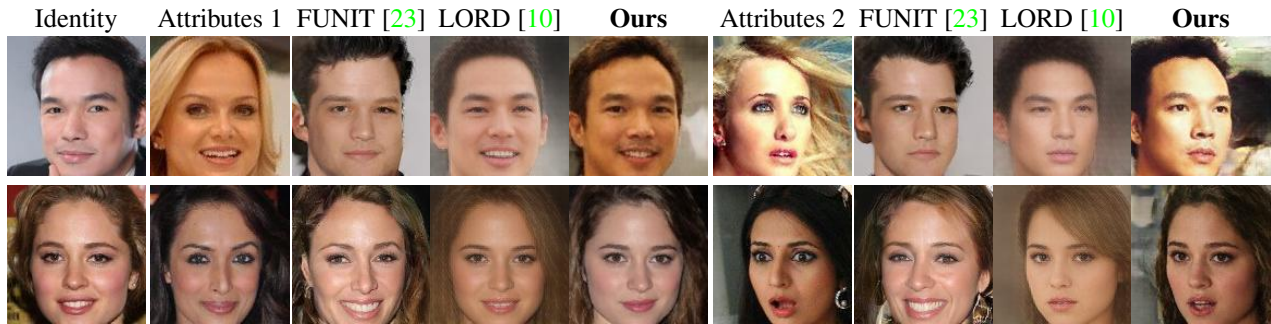
Figure 5: Disentanglement of 10K facial identities from unlabeled attributes (e.g. head pose, expression, illumination). FUNIT preserves pose but fails to model expression and exact identity. LORD captures the identity but generates low-fidelity images. We preserve head pose and facial expression while transferring the identity and generating appealing images.

Figure 6: Disentanglement of age and unlabeled attributes. Lifespan introduces artifacts and does not change hair color, while StyleGAN-based SAM struggles to preserve identity. Note that we **do not use any supervised identity loss**.

Limitations There should be noted two limitations of the proposed framework; (i) We design T for learning *pose-independent* or *localized* correlated attributes, which cov-ers common image translation settings. Nonetheless, other settings may require different forms of T. (ii) As our framework relies on latent optimization and avoids locality-

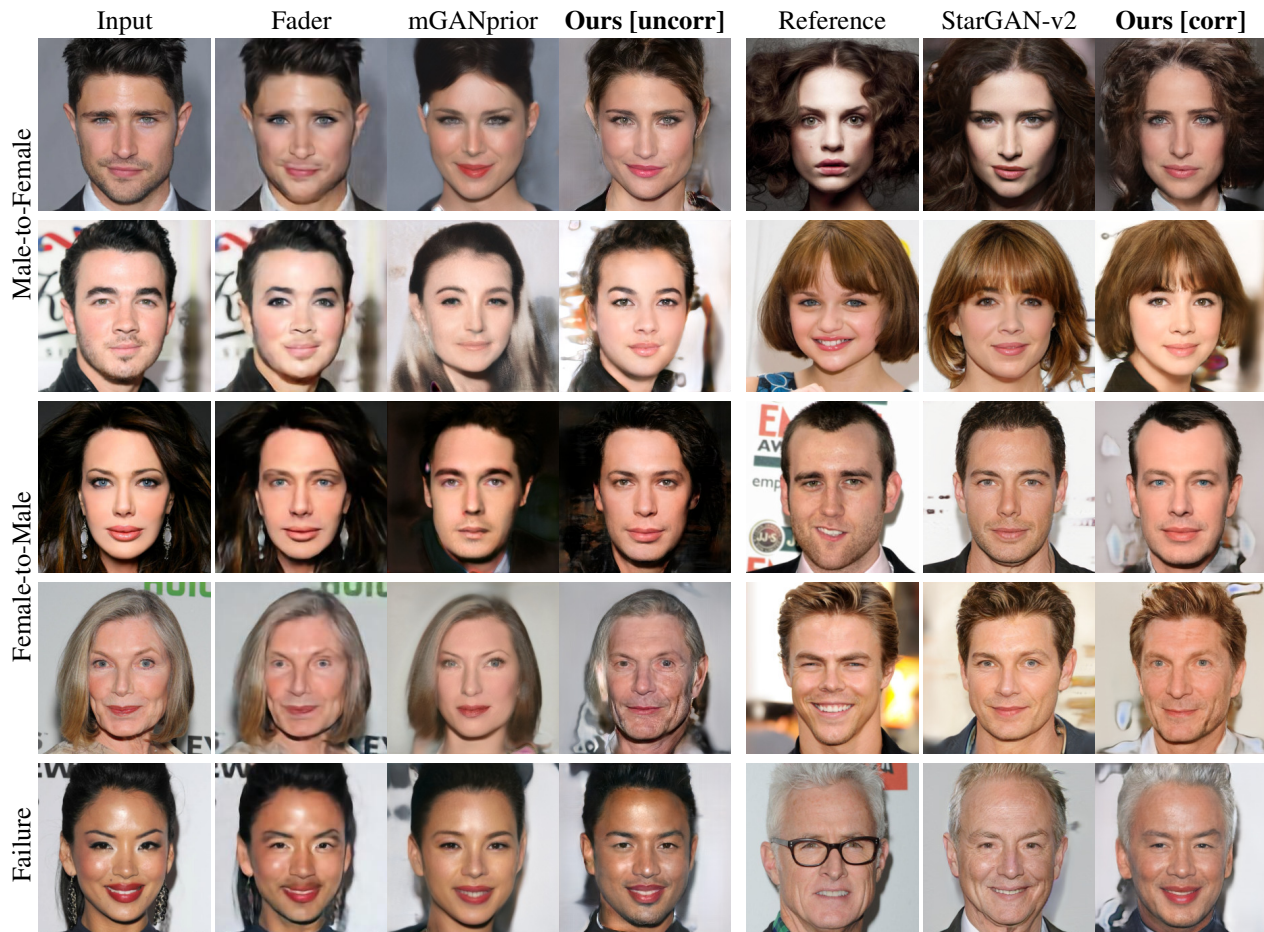| Input | Fader | mGANprior | **Ours [uncorr]** | Reference | StarGAN-v2 | **Ours [corr]** |

Figure 7: Male-to-Female translation in two settings: (i) When the attributes are assumed to be *uncorrelated*, our method makes greater changes than Fader while preserving the identity better than mGANprior. (ii) As translating hairstyle between genders is poorly specified, we model it as the *correlated* attribute and utilize a reference image specifying its target. Our method preserves the uncorrelated attributes including identity, age and illumination better than StarGAN-v2, while transferring hairstyle more faithfully. *Failure case*: as the lipstick is not modeled as a correlated attribute, it transfers over unreliably.

biased architectures for improving disentanglement, it is not well optimized for cases in which the object is not the major part of the image e.g. where the background contains other objects or large variation. We hypothesize that this can be better solved with unsupervised segmentation techniques.

6. Conclusion

We presented OverLORD, a framework for representation disentanglement in the case where supervision is provided for some attributes while not for others. Our model extends latent-optimization-based methods to cases where the attributes are correlated, and scales-up for high-fidelity image synthesis. We further showed how adversarial optimization can be decoupled from representation disentanglement and be applied only to increase the perceptual quality of the generated images. With our unified framework, we

Table 3: Male-to-Female results on CelebA-HQ. Accuracy of fooling a gender classifier and translation fidelity (FID).

	Target Classification		FID	
	F2M	M2F	F2M	M2F
Fader [20]	0.82	0.80	119.7	81.7
mGANprior [11]	0.59	0.76	78.2	45.3
Ours [uncorr]	**0.98**	**0.97**	**54.0**	**42.9**
StarGAN-v2 [5]	**0.98**	**0.99**	**27.9**	20.1
Ours [corr]	**0.98**	**0.99**	28.1	**16.4**
Optimal (Real)	0.99	0.99	15.6	14.0

achieve state-of-the-art performance compared to both general and task-specific methods on various image translation tasks with different forms of labeled attributes.

References

[1] Yuval Alaluf, Or Patashnik, and Daniel Cohen-Or. Only a matter of style: Age transformation using a style-based regression model. *arXiv preprint arXiv:2102.02754*, 2021. 5, 7, 11, 13, 14

[2] Diane Bouchacourt, Ryota Tomioka, and Sebastian Nowozin. Multi-level variational autoencoder: Learning disentangled representations from grouped observations. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018. 2

[3] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 67–74. IEEE, 2018. 11

[4] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8789–8797, 2018. 1

[5] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2020. 1, 2, 3, 8, 11, 19

[6] Elliot Creager, David Madras, Jörn-Henrik Jacobsen, Marissa A Weis, Kevin Swersky, Toniann Pitassi, and Richard Zemel. Flexibly fair representation learning by disentanglement. In *International Conference on Machine Learning*, 2019. 1

[7] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019. 5

[8] Emily L Denton et al. Unsupervised learning of disentangled representations from video. In *Advances in neural information processing systems*, pages 4414–4423, 2017. 2

[9] Chanho Eom and Bumsub Ham. Learning disentangled representation for robust person re-identification. In *Advances in Neural Information Processing Systems*, pages 5297–5308, 2019. 1

[10] Aviv Gabbay and Yedid Hoshen. Demystifying inter-class disentanglement. In *ICLR*, 2020. 1, 2, 4, 5, 6, 7, 11, 12, 25

[11] Jinjin Gu, Yujun Shen, and Bolei Zhou. Image processing using multi-code gan prior. In *CVPR*, 2020. 5, 8, 19

[12] Ananya Harsh Jha, Saket Anand, Maneesh Singh, and VSR Veeravasarapu. Disentangling factors of variation with cycle-consistent variational auto-encoders. In *ECCV*, 2018. 2

[13] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in neural information processing systems*, pages 6626–6637, 2017. 11

[14] Wei-Ning Hsu, Yu Zhang, and James Glass. Unsupervised learning of disentangled and interpretable representations from sequential data. In *Advances in neural information processing systems*, pages 1878–1889, 2017. 1

[15] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1501–1510, 2017. 2

[16] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 172–189, 2018. 1, 2, 3, 6

[17] Xu Ji, João F Henriques, and Andrea Vedaldi. Invariant information clustering for unsupervised image classification and segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9865–9874, 2019. 4

[18] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018. 5, 11

[19] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019. 11

[20] Guillaume Lample, Neil Zeghidour, Nicolas Usunier, Antoine Bordes, Ludovic Denoyer, and Marc'Aurelio Ranzato. Fader networks: Manipulating images by sliding attributes. In *Advances in neural information processing systems*, pages 5967–5976, 2017. 1, 5, 8, 19

[21] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 6, 12

[22] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Kumar Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *European Conference on Computer Vision*, 2018. 2, 6

[23] Ming-Yu Liu, Xun Huang, Arun Mallya, Tero Karras, Timo Aila, Jaakko Lehtinen, and Jan Kautz. Few-shot unsupervised image-to-image translation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 10551–10560, 2019. 1, 2, 3, 6, 7

[24] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. http://mmlab.ie.cuhk.edu.hk/projects/celeba.html. In *ICCV*, 2015. 5, 11

[25] Dominik Lorenz, Leonard Bereska, Timo Milbich, and Bjorn Ommer. Unsupervised part-based disentangling of object shape and appearance. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10955–10964, 2019. 2

[26] Qi Mao, Hsin-Ying Lee, Hung-Yu Tseng, Siwei Ma, and Ming-Hsuan Yang. Mode seeking generative adversarial networks for diverse image synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1429–1437, 2019. 2, 6

[27] Michael F Mathieu, Junbo Jake Zhao, Junbo Zhao, Aditya Ramesh, Pablo Sprechmann, and Yann LeCun. Disentan-

gling factors of variation in deep representation using adversarial training. In *NIPS*, 2016. 2

[28] Roy Or-El, Soumyadip Sengupta, Ohad Fried, Eli Shechtman, and Ira Kemelmacher-Shlizerman. Lifespan age transformation synthesis. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 5, 7, 11, 13, 14

[29] Taesung Park, Jun-Yan Zhu, Oliver Wang, Jingwan Lu, Eli Shechtman, Alexei A Efros, and Richard Zhang. Swapping autoencoder for deep image manipulation. *arXiv preprint arXiv:2007.00653*, 2020. 2

[30] Nataniel Ruiz, Eunji Chong, and James M. Rehg. Fine-grained head pose estimation without keypoints. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2018. 11

[31] Kuniaki Saito, Kate Saenko, and Ming-Yu Liu. Coco-funit: Few-shot unsupervised image translation with a content conditioned style encoder. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 2

[32] Attila Szabó, Qiyang Hu, Tiziano Portenier, Matthias Zwicker, and Paolo Favaro. Challenges in disentangling independent factors of variation. *ICLRW*, 2018. 2

[33] Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, and Luc Van Gool. Unsupervised semantic segmentation by contrasting object mask proposals. *arXiv preprint arXiv:2102.06191*, 2021. 4

[34] Sjoerd van Steenkiste, Francesco Locatello, Jürgen Schmidhuber, and Olivier Bachem. Are disentangled representations helpful for abstract visual reasoning? In *Advances in Neural Information Processing Systems*, pages 14245–14258, 2019. 1

[35] Aron Yu and Kristen Grauman. Fine-grained visual comparisons with local learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 192–199, 2014. 6, 12

[36] Xiaoming Yu, Yuanqi Chen, Shan Liu, Thomas Li, and Ge Li. Multi-mapping image-to-image translation via learning disentanglement. In *Advances in Neural Information Processing Systems*, pages 2994–3004, 2019. 2

[37] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018. 11

[38] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. 2

[39] Jun-Yan Zhu, Zhoutong Zhang, Chengkai Zhang, Jiajun Wu, Antonio Torralba, Josh Tenenbaum, and Bill Freeman. Visual object networks: Image generation with disentangled 3d representations. In *Advances in neural information processing systems*, pages 118–129, 2018. 1