Metrics for Exposing the Biases of Content-Style Disentanglement

Xiao Liu[*,1] Spyridon Thermos[*,1] Gabriele Valvano[*,1,2] Agisilaos Chartsias[1]
Alison O'Neil[1,3] Sotirios A. Tsaftaris[1,4]

[1] School of Engineering, University of Edinburgh, Edinburgh, UK
[2] IMT School for Advanced Studies Lucca, Lucca, Italy
[3] Canon Medical Research Europe Ltd., Edinburgh, UK
[4] The Alan Turing Institute, London, UK

{Xiao.Liu,SThermos,G.Valvano,Agis.Chartsias,S.Tsaftaris}@ed.ac.uk Alison.ONeil@eu.medical.canon

Abstract

A recent spate of state-of-the-art semi- and unsupervised solutions for challenging computer vision tasks encode image "content" into a spatial tensor and image appearance or "style" into a vector. Most of these solutions use the term disentangled *for their representations and employ different "biases" such as model design, learning objectives, and data, to achieve good performance in spatially equivariant tasks (e.g. image-to-image translation). While considerable effort has been made to measure disentanglement in vector representations, we have lacked metrics for spatial content and vector style representations. In this paper, we propose such metrics to characterize the degree of disentanglement in terms of how (un)correlated and informative the content and style representations are, and we further examine its relation to task performance. In particular, we first identify key design choices and learning constraints on three popular models that employ content-style disentanglement and derive ablated versions. Secondly, we use our metrics to ascertain the role of each bias. Our experiments reveal a "sweet spot" between disentanglement, task performance and latent space interpretability. Our metrics are not task-dependent; thus, they can help guide either the design of new future models or the selection of viable models such that this ideal "sweet spot" is achieved in any task where content-style representations are useful. Code is available at* https://github.com/vios-s/CSDisentanglement_Metrics_Library.

1. Introduction

Recent work in representation learning argues that to achieve explainable and compact representations one should separate out, or *disentangle*, the underlying explanatory factors into different dimensions of the considered latent space [3, 25]. In other words, it is beneficial to obtain representations that can separate latent variables that capture sensitive and useful information for the task at hand, from the ones that are less informative or even distracting [1]. Over the years, disentanglement has been exploited to improve task performance, model generalization, and representation interpretability [13, 17, 21, 33, 42, 44, 48, 57]. However, Locatello et al. [38, 39, 52] indicate that unsupervised disentanglement is ill-posed and hence impossible. Instead, we can achieve disentanglement via restrictions and inductive priors [38, 39] which we can see as different forms of "bias" imposed by model design (design bias), learning objectives (learning bias), and data (data bias).

Inspired by these findings, we set out to reveal such choices of bias in state-of-the-art (SOTA) approaches that employ disentanglement. Our particular focus is "content-style" disentanglement where an imaging input is decomposed into spatial "content" and vector "style" representations. In principle, content variables should contain semantic information required for spatially equivariant tasks such as segmentation and pose estimation, whereas style variables contain information that controls image appearance such as color intensity and texture. These decompositions have been employed to offer semi- or un-supervised solutions for challenging computer vision tasks.

In practice, contrary to extensive research on quantifying the degree of disentanglement between vectors [11, 19, 20, 30, 34, 45, 55], the separation between the content and style latent spaces is not typically assessed. In fact, to the best of our knowledge, there are no metrics to expose the relationship between biases and content-style disentanglement, and by extension the relationship of content-style disentanglement with model performance and latent space interpretability.

Herein, we attempt to bridge these gaps with the following **contributions**:

[*] Authors contributed equally.

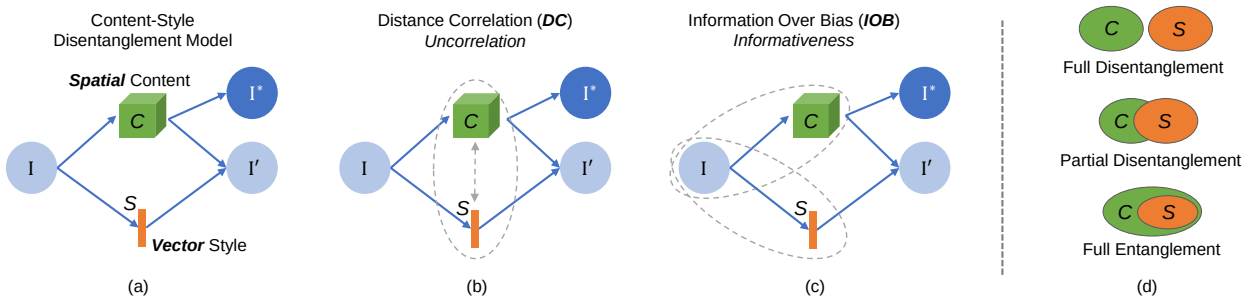

Figure 1: **(a)** A schematic representation of disentanglement between spatial content C and vector style S in the context of a primary and a secondary spatially equivariant task (I', I^*). Proposed metrics that measure the distance correlation **(b)** between C and S (or between a latent variable and input I), and the amount of information encoded in C or S with respect to the I bias **(c)**. **(d)** A visual description of degrees of C-S (dis)entaglement.

- We introduce a set of complementary metrics to quantitatively evaluate content-style disentanglement (concept depicted in Fig. 1) in terms of: a) the amount of information encoded in each latent feature (informativeness); and b) the uncorrelation between the encoded *spatial tensor* content and *vector* style features (a proxy for independence).
- We identify key biases in three SOTA models that employ content-style disentanglement, and expose how the biases affect disentanglement and task performance (utility). We focus on the popular vision tasks of image translation, segmentation, and pose estimation.
- We perform extensive experiments, where we find that lower disentanglement benefits task performance when a specific style-related prior is not violated, and that performance is correlated with latent variable informativeness. We also qualitatively assess the representation semanticness (interpretability).

2. Related Work

2.1. Content-Style Disentanglement

Decoupling the style and content of an image has been extensively explored in Image-to-Image translation (I2I) [29,35,36]. Outside I2I, content-style disentanglement has been used for many applications, such as semantic segmentation [10] and pose estimation [9], where the content has been used as a robust representation for a downstream task. In general, most methods use variants of (variational) auto-encoders to derive latent spaces that capture content and style information.

All of these models achieve content-style disentanglement using different biases, such as specific architectural choices (*e.g.* use of AdaIN [28], content Binarization [10]), learning objectives (*e.g.* KL divergence, latent regression loss, and de-correlation losses in vector representations [8, 49]), or supervisory signals (*e.g.* using the content for a segmentation task [10]). However, the precise effect of each bias on the resulting disentanglement and model performance is not thoroughly explored.

2.2. Disentanglement Evaluation Methods

The ideal approach for evaluating content-style disentanglement should: i) offer the ability to compare latent factors which are tensors of different dimensions (*e.g.* the style is a vector whereas the content is a spatial multi-channel tensor); ii) be quantitative; and iii) not require ground truth information about the factors. Currently, we lack such an approach, but below we review related inspiring work.

A classical approach is *latent traversals*: a visualization that shows how traversing single latent dimensions generates variations in the image reconstruction. Latent traversals do not need ground truth information about the factors, and can be used in mixed tensor spaces (*e.g.* as shown in [10] and [40]) to offer a qualitative evaluation. Alternatively, latent traversals can be combined with pre-trained networks to measure the perceptual distance between the produced embeddings [30].

On the other hand, there is a considerable effort in quantitatively evaluating the representations learned by VAEs and GANs. However, all of them rely on vector representations, and some also peruse ground truth knowledge of the latent factors. In particular, some methods try to associate known factors of variations (*e.g.* rotation and translation) with specific latent dimensions [26, 31]. Others measure the ability to isolate one factor in a single vector latent variable [34], measuring compactness or modularity [11,20,55], linear separability [30], consistency and restrictiveness [47], and explicitness of the representation [45]. On top of that, there are interesting works on measuring the informativeness of a specific factor in a vector latent variable *w.r.t.* the input, as well as measuring the independence among factors

and their interpretability [19, 20].

2.3. Impact

The ability to transfer these concepts from the *vector*-based disentanglement (where they are defined) to the content-style disentanglement, which incorporates both spatial and vector representations,[1] will expand our understanding of the relation between disentanglement and the: a) various biases adopted by each model; b) task performance; c) representation interpretability.

3. Metrics for Content-Style Disentanglement

Given N image samples $\{I_i\}_{i=1}^N$, we assume two representations of content and style : $\{C_i\}_{i=1}^N$, and $\{S_i\}_{i=1}^N$ respectively. We propose two metrics to evaluate properties previously investigated in vector latent space disentanglement [19, 20]: *uncorrelation*, and *informativeness*. Then, we discuss two properties of the representations: *utility* and *interpretability*, which will highlight advantages and disadvantages of content-style disentanglement.

3.1. Distance Correlation (DC)

Disentangled representations separate content and style into independent latent spaces [25]. Independent content and style variables must satisfy $p(C, S) = p(C)p(S)$, however directly measuring independence between spatial C and vector S with existing metrics is not feasible. Since independent representations must be uncorrelated [11], we use the *empirical Distance Correlation* (DC) [51] to measure the correlation between tensors of arbitrary dimensionality. Note that DC is bounded in the $[0, 1]$ range, while differently from other correlation-independence metrics, such as the kernel target alignment [15] and the Hilbert-Schmidt independence criterion [23], it has the advantage of not requiring any pre-defined kernels.

For N samples, consider two N-row matrices T_1 and T_2. In general, T_1 (or T_2) have different column dimension as they are formed by concatenating images I_i, content features C_i or style features S_i. For I_i and C_i we first concatenate the channels and then row-scan to form a vector; S_i is already a vector. DC is then defined as:

$$DC(T_1, T_2) = \frac{dCov(T_1, T_2)}{\sqrt{dCov(T_1, T_1)dCov(T_2, T_2)}}, \quad (1)$$

with:

$$dCov(X, Y) = \sqrt{\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{A_{i,j}B_{i,j}}{N^2}}. \quad (2)$$

[1] Note that our metrics are generic, thus can be readily applied to disentanglement methods that encode both C and S as vector representations, too, such as [22].

Here, $dCov$ is the distance covariance between any two N-row matrices X and Y. A and B are distance matrices for X and Y, respectively. In particular, each matrix element $a_{i,j}$ of A is the Euclidean distance between two samples $\|X^i - Y^j\|$, after subtracting the mean of row i and column j, as well as the matrix mean. B is similarly calculated. We estimate disentanglement between C and S using their distance correlation, $DC(C, S)$, with lower values (closer to 0) indicating higher disentanglement.

However, notice that C and S can be uncorrelated ($DC(C, S) = 0$) either when they encode unrelated information or, more critically, when one encodes *all* the information whilst the other *nothing*, *i.e.* noise. The latter case happens with posterior collapse, and it indicates full entanglement. As a result, $DC(C, S)$ is not enough to assess the disentanglement degree between C and S, and it needs a complementary metric to measure their information content, or informativeness.

3.2. Information Over Bias (IOB)

To explicitly measure the amount of information that C and S encode, we propose the *Information Over Bias* (IOB) metric. An important role of IOB is to detect posterior collapse, when C and S are disentangled but one is not informative about the input.

Given features $z \in \{C, S\}$ produced from N image samples at inference, we aim to measure the amount of information encoded in C or S representations. That is, we train a decoder G_θ, modeled as a neural network with parameters θ, to reconstruct images I, given the latent representations z predicted by the disentanglement framework.

Thus, we define IOB as the expectation over the test images of the ratio:

$$\begin{aligned} IOB(I, z) &= \mathbb{E}_i\left[\frac{\text{MSE}(I_i, G_\theta(\mathbb{1}))}{\text{MSE}(I_i, G_\theta(z_i))}\right] \\ &= \frac{1}{N}\sum_{i=1}^{N}\left(\frac{\frac{1}{K}\sum_{k=1}^{K}||I_i^k - G_\theta(\mathbb{1})||^2}{\frac{1}{K}\sum_{k=1}^{K}||I_i^k - \tilde{I}_i{}^k||^2 + \varepsilon}\right), \end{aligned} \quad (3)$$

where I and $\tilde{I}$ are an image and its reconstruction obtained through G_θ; $i = 1 \dots N$, $k = 1 \dots K$ are indices iterating on the test images and the image pixels, respectively; ε is a small value that prevents division with zero. We justify the above definition of IOB by observing that a post-hoc minimization of the MSE between $\tilde{I}$ and I is equivalent to maximizing the log likelihood (see our analysis in Appendix A). Notice that the ratio aims at ruling out from IOB both data biases (due to common structure, colors, pose, etc. across the images of the dataset) and architectural biases that one could introduce in the design of G_θ. In particular, this is done by computing the ratio between the MSE obtained after training G_θ to reconstruct the images from their *in-*

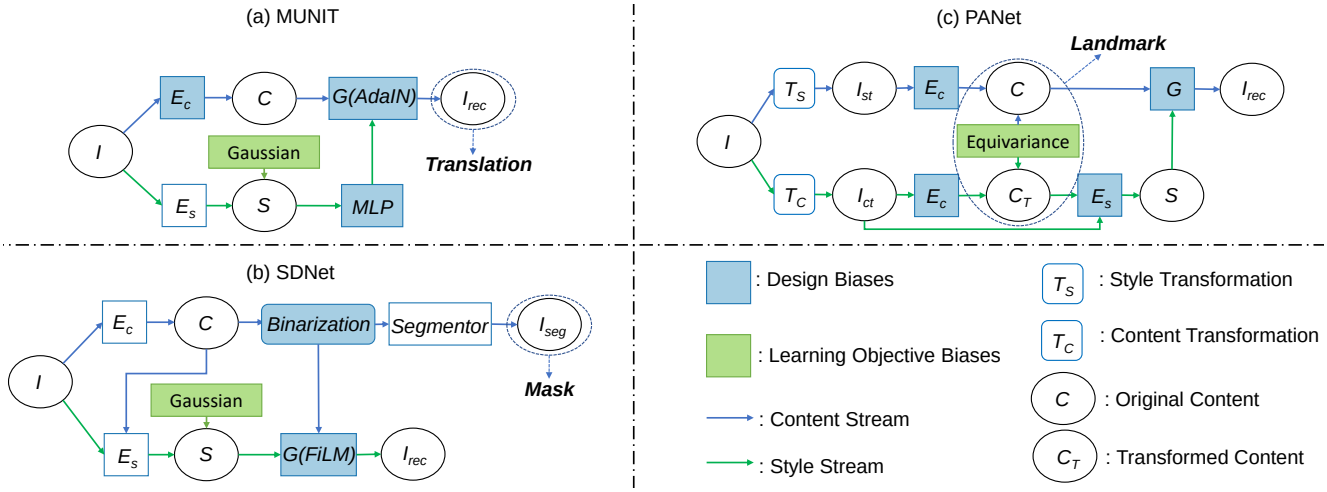

Figure 2: Model schematics. a) MUNIT: E_c uses Instance Normalization to remove style from content; E_s uses global pooling. b) SDNet: content is represented as binary images, and is used for segmentation. Style features minimize the KL divergence with a Gaussian distribution. c) PANet: training encourages content C and style S to be equivariant to intensity and spatial transformations.

formative representation z (*i.e.* $\text{MSE}(I_i, G_\theta(z_i))$, and after training G_θ from an *uninformative* constant tensor $\mathbb{1}$ (*i.e.* $\text{MSE}(I_i, G_\theta(\mathbb{1}))$). In the latter case, G_θ will only learn the dataset bias it can model, given θ. Hence, high values of IOB can be associated with higher information inside the representation z, while the lower bound $IOB = 1$ means that no information of the images I is encoded in z.[2]

3.3. Utility and Interpretability

As discussed in the previous sections, we can use DC and IOB to measure the degree of disentanglement between latent representations. However, among the primary goals of disentanglement, there is improving the task performance (utility) and the representation interpretability. Hence, we investigate the relationship between $C - S$ disentanglement and these two notions.

In particular, we measure utility by quantifying performance on a downstream task, which for disentangled representations is typically image translation [29, 35] to translate image content from one domain to another. We also consider tasks using content *e.g.* to extract segmentations [10] or landmarks [40], and therefore assess how effectively it can be used in downstream tasks. We detail performance metrics for each application in Section 5.

Assessing interpretability is not trivial. Here, we assume that interpretability implies semantic representations. Previously, vector representations were considered semantic if a portion of the latent space corresponded to specific data variations [12]. In disentangled representations, the style semantics were qualitatively evaluated with latent traversals of individual dimensions [10]. Thus, we consider a style interpretable if images produced by linear traversals in the style latent space are realistic and smoothly change intensity. In spatial representations, such data variation should be confined to individual objects: thus, a semantic content should split distinct objects into separate channels of C. We empirically evaluate this with qualitative visuals wherever possible.

[2]Optimising G_θ with stochastic gradient descent can introduce noise and slightly alter the measure. For example, IOB may, in practice, even be slightly smaller than 1. Thus, we average results across multiple runs and initializations of G_θ, which contributes to the computational load of estimating IOB.

4. Applications

Many applications disentangle content from style [6, 22] or other attributes, such as pose, geometry, and motion [16, 27, 54, 56], to boost model performance in vision tasks. For our analysis we select and discuss three recent approaches from diverse applications such as image translation (MUNIT [29]), semantic segmentation (SDNet [10]), and pose estimation (PANet [40]). All resemble auto-encoders, mapping input images to disentangled features, as shown in Fig. 2, but use several biases at various degrees as we detail below. Our objective is to elucidate how different biases affect disentanglement using these known models, and their chosen biases, as exemplars.

We report detailed model descriptions and summarize *design* and *learning* biases in Appendix B. Here we describe how each bias is enforced.

In particular, for: **a) MUNIT** we consider ablations without Instance Normalization (IN) [53], AdaIN layers, or removing style Latent Regression (LR) loss (for fairness, we

Table 1: Comparative evaluation of MUNIT variants using the proposed metrics. We use FID and $LPIPS$ to measure translation quality and diversity between SYNTHIA [46] and Cityscapes [14] samples. Results are in "mean ±std" format. Arrows (↑, ↓) indicate the direction of each metric improvement.

		Design Bias		**Learning Bias**
Metric	Original Model	w/o AdaIN	w/o Instance Normalization (IN)	w/o Latent Regression (LR)
$DC(C,S)$ (↓)	0.44 ±0.06	0.43 ±0.01	0.66 ±0.03	**0.40** ±0.08
$DC(I,C)$ (↑)	0.57 ±0.07	0.58 ±0.08	**0.73** ±0.03	0.57 ±0.08
$DC(I,S)$ (↑)	0.70 ±0.02	0.56 ±0.03	0.63 ±0.05	**0.73** ±0.03
$IOB(I,C)$ (↑)	4.36 ±0.38	4.85 ±0.10	**5.01** ±0.12	4.34 ±0.58
$IOB(I,S)$ (↑)	1.31 ±0.04	1.17 ±0.04	1.28 ±0.06	**1.46** ±0.05
FID (↓)	73.48 ±8.35	**52.48** ±5.03	71.4 ±4.86	104.51 ±4.21
$LPIPS$ (↑)	0.08 ±0.01	0.06 ±0.01	**0.10** ±0.01	0.09 ±0.01

do not remove LR of the content as it is fundamental for the functioning of the model); **b) SDNet** we identify content Binarization, Gaussian approximation, LR and the FiLM-based [43] decoder as the main biases that affect content-style disentanglement; we investigate their impact on the representations and how they affect the main task of the model (semantic segmentation); **c) PANet** we remove the Gaussian approximation and replace its specific content-style conditioning with AdaIN. Note that we analyse PANet performance in its main task, i.e. pose estimation.

Why these models? We choose these models as they are well-known SOTA in the corresponding domains and cover the cases of: **a)** no supervision and weak C constraints (MUNIT), **b)** no supervision with strong C constraints (PANet), and **c)** supervision with strong C constraints (SDNet).

5. Experiments

For each model, we analyze the effect that design choices and learning objectives have on disentanglement and task performance, and we evaluate utility and interpretability of the learned representations. We use the implementations provided by the authors, ablating only the components needed for our analysis. In all tables, arrows (↑, ↓) indicate direction of metric improvement; best results are in bold. Numbers are the average of 5 different runs with different weight initialization. Data description and detailed learning setup can be found in Appendices C-E. Note that the evaluated models process different types of data and domains, *i.e.* color and intensity images.

5.1. Image-to-Image Translation

Setup. We consider the original MUNIT and three variants: **i)** we replace the AdaIN modules of the decoder with simple style concatenations, reducing the restrictions on the re-combination of C and S. **ii)** We remove the LR loss, responsible for the style Gaussianity. **iii)** We remove IN from the content encoder, to confirm that it helps to focus on the content only, discarding the original style [28]. As [29] we evaluate quality and diversity of the translated images using the Fréchet Inception Distance (FID) [24] and LPIPS [58].

Results. Table 1 reports the results of the ablations on the SYNTHIA [46] and Cityscapes [14] datasets. Replacing AdaIN (**w/o AdaIN**) with simple concatenation does not affect the level of C-S disentanglement, but it leads to a 0.14 absolute decrease of $IOB(I,S)$ and $DC(I,S)$, indicating that the style becomes less informative and less correlated with the input. Here, we observe an information shift to the content (lower $IOB(I,S)$, higher $IOB(I,C)$) which leads to better translation quality, but also the worst diversity ($LPIPS = 0.06$). We believe that this variant is worse than the original model, which had more balanced quality/diversity scores.

By removing the LR learning bias (**w/o LR**), we observe that the style is significantly more correlated to the input image. If the style distribution is no longer Gaussian, the style has more degrees of freedom to encode non-relevant information, which contributes to higher $IOB(I,S)$ and higher C-S disentanglement. This ablation leads to a significant translation quality decrease, while contrary to the analysis in [29], the diversity is not negatively affected.

Finally, by removing IN (**w/o IN**) we expect a more entangled content that will also encode some style information. Our expectations are confirmed by the decrease of C-S disentanglement ($DC(C,S) = 0.66$), and a more informative content (which is also more correlated to the input image). Interestingly, relaxing the content constraints for a task that does not require a strictly semantic content (*e.g.* image segmentation), leads to the best quality/diversity balance. Note that we define the best balance as achieving the highest average ranking in FID and $LPIPS$ (*e.g.* the "w/o IN" model variant is the 1^{st} in $LPIPS$ and 2^{nd} in FID, thus the best overall model).

Summary. Our experiments show there is a trade-off be-

Table 2: Comparative evaluation of SDNet variants using the proposed metrics. We use the $Dice$ score to measure semantic segmentation performance on the ACDC [5] dataset with 1.5% annotation masks. Results are in "mean $\pm$std" format. Arrows ($\uparrow, \downarrow$) indicate the direction of each metric improvement.

		Design Bias		**Learning Bias**
Metric	Original Model	SPADE	w/o Binarization	w/o KL Divergence and Latent Regression (LR)
$DC(C,S)$ ($\downarrow$)	0.49 $\pm$0.02	0.52 $\pm$0.01	**0.44** $\pm$0.00	0.64 $\pm$0.03
$DC(I,C)$ ($\uparrow$)	0.94 $\pm$0.01	0.93 $\pm$0.01	**0.98** $\pm$0.02	0.94 $\pm$0.01
$DC(I,S)$ ($\uparrow$)	0.43 $\pm$0.02	0.45 $\pm$0.01	0.44 $\pm$0.01	**0.66** $\pm$0.00
$IOB(I,C)$ ($\uparrow$)	4.71 $\pm$0.26	5.09 $\pm$0.00	**5.89** $\pm$0.22	4.84 $\pm$0.23
$IOB(I,S)$ ($\uparrow$)	1.00 $\pm$0.01	1.00 $\pm$0.04	0.98 $\pm$0.04	1.00 $\pm$0.04
$Dice$ ($\uparrow$)	0.62 $\pm$0.02	**0.75** $\pm$0.02	0.63 $\pm$0.04	0.61 $\pm$0.04

tween the translation quality/diversity and disentanglement, for the considered translation task.[3] Our metrics indicate that a partially disentangled C-S space coupled with a near-Gaussian style latent space leads to the best quality/diversity performance. For MUNIT this is achieved by removing the IN design bias.

5.2. Medical Segmentation

Setup. In SDNet, content binarization and style Gaussianity are the key representation constraints. We evaluate their implications on performance, together with decoder design implications, by: **i)** removing content thresholding (w/o Binarization), **ii)** removing style Gaussianity (w/o KL divergence and LR), and **iv)** considering a new decoder, obtained replacing the FiLM style conditioning with *SPADE* [41]. SPADE can be less restrictive, allowing the style to encode more image-related information, such as textures, rather than just intensity values (see Appendix D.1). We also assess model performance with the Dice Score [18, 50].

Results. Table 5 reports the ablation results on the ACDC [5] dataset. We highlight that when using all the available annotations (fully supervised learning), all SDNet variants achieve a similar accuracy (see Appendix D.2 for more details), suggesting that strong learning biases, such as supervised segmentation costs, make disentanglement less important. Thus, we consider the semi-supervised training case with minimal supervision, using only the 1.5% of available labelled data.

Broadly speaking, the style encodes little information in all SDNet variants, probably because all medical images in ACDC have similar styles (data bias), and reconstructing using an average style is enough to have low $IOB(I,S)$.

However, C-S disentanglement is still important to obtain a good content representation. For example, it is evident that intermediate levels of disentanglement (**SPADE**) lead to the best segmentation performance. In this variant, disentanglement decreases compared to the original model, as some style information is probably leaked to the content (higher $DC(C,S)$ and $IOB(I,C)$). On the other hand, removing C binarization (**w/o Binarization**) also leads to a more informative content; since the correlation between C and S decreases, we assume that the extra information encoded in C is not part of the style.

Lastly, removing the Gaussian prior constraints from the style latent space (**w/o KL and LR**) leads to the lowest degree of disentanglement as there is no information bottleneck on S, while also to a slight decrease of the Dice score.

Summary. We find disentanglement to have minimum effect on task performance when training with strong learning signals (*i.e.* supervised costs). In the semi-supervised setting, a higher (but not full) degree of disentanglement leads to better performance, while the amount of information in C alone is not enough to achieve adequate segmentation performance.

5.3. Pose Estimation

Setup. Together with the original PANet model, we consider four possible variants, relaxing design biases on both content and style, and learning biases. In detail: **i)** we experiment with a different conditioning mechanism to re-entangle style and content, that consists of the use of AdaIN, rather than just multiplying each style vector with a separate content channel (introducing a bias on S, similar to MUNIT). **ii)** We consider the case where, instead of learning a different style for each channel of the content, we extract a global style vector, predicted by an MLP (relaxing the tight 1:1 correspondence between C and S channels). **iii)** We also consider the case where each content part is not approximated by Gaussian distributions. Since we cannot use the original decoder to combine C and S, we reintroduce the style using AdaIN. **iv)** Finally, we evaluated the effect of the equivariance constraint, by removing it from the cost function.

Results. Table 3 reports results of the ablations on the

[3] Note that the effect of C-S disentanglement on task performance also depends on the data bias. An indicative example is the "edges-to-shoes" where the translation is between zero-style and normal images.

Table 3: Comparative evaluation of PANet variants using the proposed metrics. We use SIM to measure the performance in terms of pose estimation from landmarks on the DeepFashion [37] dataset. Results are in "mean ±std" format. Arrows (↑, ↓) indicate the direction of each metric improvement.

		Design Bias			**Learning Bias**
Metric	Original Model	AdaIN	MLP	AdaIN w/o Gaussian	w/o Equivariance
$DC(C,S)$ (↓)	0.65 ±0.01	0.36 ±0.02	0.69 ±0.03	**0.25** ±0.01	0.76 ±0.08
$DC(I,C)$ (↑)	0.59 ±0.01	0.56 ±0.01	0.58 ±0.02	0.53 ±0.01	**0.60** ±0.02
$DC(I,S)$ (↑)	**0.83** ±0.01	0.81 ±0.01	0.82 ±0.03	0.38 ±0.06	0.82 ±0.01
$IOB(I,C)$ (↑)	1.50 ±0.08	1.52 ±0.08	1.49 ±0.06	**1.53** ±0.06	1.50 ±0.08
$IOB(I,S)$ (↑)	1.09 ±0.04	1.10 ±0.15	**1.21** ±0.09	1.12 ±0.09	1.13 ±0.06
SIM (↑)	**0.71** ±0.02	0.64 ±0.01	0.68 ±0.01	0.58 ±0.00	0.47 ±0.04

DeepFashion [37] dataset. We assess model performance using SIM [7] to measure the similarity between the predicted and ground truth landmarks visualized as heatmaps.

Whilst the original model is the best to predict landmarks, it only achieves an average degree of disentanglement (see $DC(C,S)$). Using an **AdaIN**-based decoder always improves disentanglement as it has a strong inductive bias on the re-entangled representation (see $DC(C,S)$ for AdaIN, and **AdaIN w/o Gaussian**), but it leads to worse landmark detection – the representation adapts tightly to the strongly-biased decoder, and the content loses transferability to other tasks, and interpretability (see Fig. 4, and also Fig. 9 in the Appendix).

Using an **MLP** to encode style relaxes the specific conditioning between C and S (a design bias) and reduces disentanglement. In fact, there is an information shift from C to S, as indicated by the higher $IOB(I,S)$, and we observe a high $DC(C,S)$. Here, a moderate decrease of disentanglement shows slightly lower task performance.

Finally, the equivariance cost is the most important factor for disentanglement: removing it (**w/o Equivariance**) leads to the most entangled representation (high $DC(C,S)$), and accuracy decrease in landmark detection.

Summary. Overall, higher partial entanglement leads to better landmark detection. Again, balance is the key to improve the auxiliary tasks. In PANet, the partial disentanglement is achieved by carefully balancing the design biases used to extract the style and to reintroduce it inside the content, while decoding. Relaxing such biases with AdaIN or MLP makes landmark detection worse.

5.4. Discussion

We now discuss the relationship between C-S disentanglement and inductive biases, task performance, interpretability of the latent representations. We emphasize that our metrics are uncorrelated from each other, as depicted in Fig. 3, and thus complementary (see Appendix F for the metrics correlation per model).

Do biases affect C-S disentanglement? Results in Section 5 illustrate that learning and design biases critically affect disentanglement.

So, Yes. But, no evaluation can specifically characterize the relative importance of each one, since this depends on both data and task. In MUNIT, disentanglement is mainly encouraged by the content-related design and learning biases. In fact, IN is key to removing style information from the content, and the model cannot be successfully trained without LR of the content. Disentanglement in SDNet is susceptible to both types of biases. Using a SPADE decoder or removing content thresholding leads to more entanglement, while making the style Gaussian through learning constraints restricts its informativeness and encourages disentanglement. Similarly, PANet disentanglement is affected both by designing the content as Gaussian, and by the equivariance of C and S *w.r.t.* spatial or intensity transformations, respectively.

Is there a "sweet spot" between C-S disentanglement and performance? Yes. We find a clear sweet spot between C-S disentanglement and downstream task performance. In particular, we observe that lower disentan-

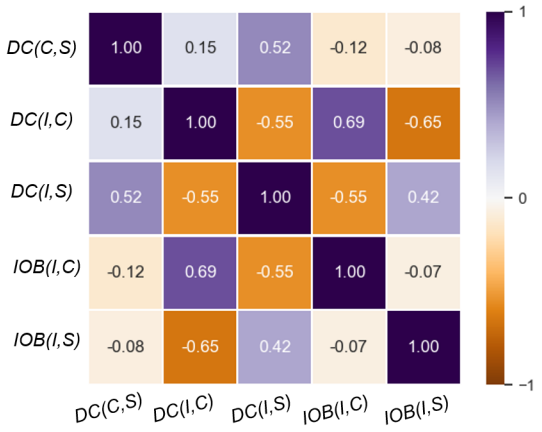

Figure 3: Pearson correlation coefficients of the proposed metrics across all models visualized as a heatmap. Values close to 1 and -1 indicate a strong correlation.

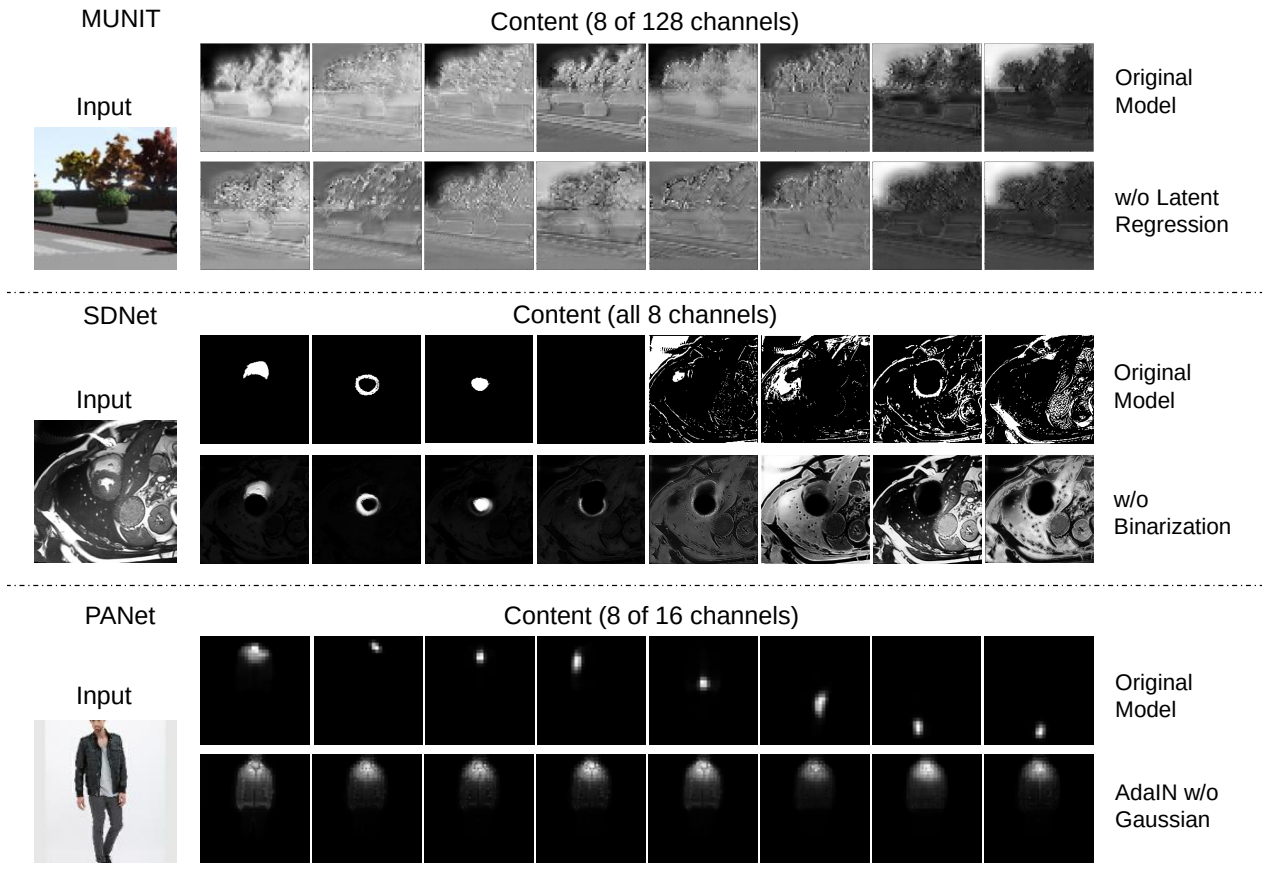

Figure 4: Content *interpretability* of each original model (top row) and a variant with the most correlated C and S (bottom row). Removing content-related design biases from SDNet and PANet leads to less interpretable representations (same objects/joints appear in different channels), such as the unconstrained content representations of MUNIT.

glement that is based on relaxing the content constraints (*e.g.* removing IN), and not based on removing biases that enforce style priors, such as the Gaussian distribution and C-S equivariance, leads to better performance.

How interpretable is content representation? Interpretability is hard to quantify without metrics. Here, we consider the content interpretable if distinct objects appear in different channels. We qualitatively analyze content interpretability in Fig. 4 (see Appendix G for more visuals).

Content interpretability varies a lot with different *design biases* of the model, while learning biases don't seem to affect it. Missing restrictive design bottlenecks on C, MUNIT spreads the content across channels. Instead, SDNet and PANet original models factorize C to have different objects, or parts, in different channels. In SDNet, a semantic content is encouraged by applying softmax and Binarization: this forces pixels to activate only in specific channels, and the model starts grouping together related structures. Similarly, approximating the body parts as 2D Gaussians in PANet enforces an information bottleneck on each channel of C. Removing the content constraints from SDNet and PANet results in spreading the spatial information across all channels, and in subsequent loss of interpretability.

6. Conclusion

We have proposed a set of metrics to evaluate the degree of disentanglement between image content and style. Our extensive experiments on three state-of-the-art models show how design and learning biases affect disentanglement. Our metrics are complementary to each other and, when used together, can quantitatively assess the informativeness and the uncorrelation of the latent variables. Our findings suggest that even though disentanglement enables the implementation of certain tasks, partially (dis)entangled representations can lead to better performance than fully disentangled ones. Additionally, our analysis suggests that strict design constraints on the content representations lead to increased interpretability, which can be exploited in post-hoc tasks. Using the proposed metrics will enable the design of better (or selection of) models that achieve a "sweet spot" of disentanglement and performance, rather than pursuing very high (or low) disentanglement.

References

[1] Alessandro Achille and Stefano Soatto. Emergence of invariance and disentanglement in deep representations. *The Journal of Machine Learning Research*, 19(1):1947–1980, 2018. 1

[2] Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. Deep variational information bottleneck. *arXiv preprint arXiv:1612.00410*, 2016. 12

[3] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013. 1

[4] Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*, 2013. 12

[5] Olivier Bernard, Alain Lalande, and Clement Zotti *et al.* Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved? *IEEE Transactions on Medical Imaging (TMI)*, 37(11):2514–2525, 2018. 6, 12, 13

[6] Diane Bouchacourt, Ryota Tomioka, and Sebastian Nowozin. Multi-level variational autoencoder: Learning disentangled representations from grouped observations. In *Proc. AAAI Conference on Artificial Intelligence (AAAI)*, 2018. 4

[7] Zoya Bylinskii, Tilke Judd, Aude Oliva, Antonio Torralba, and Frédo Durand. What do different evaluation metrics tell us about saliency models? *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 41(3):740–757, 2019. 7

[8] Xiaobin Chang, Tao Xiang, and Timothy M Hospedales. Scalable and effective deep cca via soft decorrelation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1488–1497, 2018. 2

[9] James Charles, Tomas Pfister, Derek Magee, David Hogg, and Andrew Zisserman. Domain adaptation for upper body pose tracking in signed TV broadcasts. In *Proc. British Machine Vision Conference (BMVC)*, 2013. 2

[10] Agisilaos Chartsias, Thomas Joyce, Giorgos Papanastasiou, Scott Semple, Michelle Williams, David E. Newby, Rohan Dharmakumar, and Sotirios A. Tsaftaris. Disentangled representation learning in cardiac image analysis. *Medical Image Analysis*, 58, 2019. 2, 4, 12, 13

[11] Ricky T. Q. Chen, Xuechen Li, Roger Grosse, and David Duvenaud. Isolating sources of disentanglement in VAEs. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, pages 2615–2625, 2018. 1, 2, 3

[12] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, pages 2172–2180, 2016. 4

[13] Taco S. Cohen and Max Welling. Learning the irreducible representations of commutative lie groups. In *Proc. International Conference on Machine Learning (ICML)*, pages 1755–1763, 2014. 1

[14] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The Cityscapes dataset for semantic urban scene understanding. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3213–3223, 2016. 5, 12

[15] Nello Cristianini, John Shawe-Taylor, Andre Elisseeff, and Jaz S Kandola. On kernel-target alignment. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, pages 367–373, 2002. 3

[16] Emily Denton and Vighnesh Birodkar. Unsupervised learning of disentangled representations from video. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, pages 4414–4423, 2017. 4

[17] Guillaume Desjardins, Aaron Courville, and Yoshua Bengio. Disentangling factors of variation via generative entangling. In *arXiv preprint arXiv:1210.5474*, 2012. 1

[18] Lee R Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302, 1945. 6

[19] Kien Do and Truyen Tran. Theory and evaluation metrics for learning disentangled representations. In *International Conference on Learning Representations (ICLR)*, 2020. 1, 3

[20] Cian Eastwood and Christopher K. I. Williams. A framework for the quantitative evaluation of disentangled representations. In *International Conference on Learning Representations (ICLR)*, 2018. 1, 2, 3

[21] Patrick Esser, Johannes Haux, and Bjorn Ommer. Unsupervised robust disentangling of latent characteristics for image synthesis. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, pages 2699–2709, 2019. 1

[22] Aviv Gabbay and Yedid Hoshen. Demystifying inter-class disentanglement. In *International Conference on Learning Representations (ICLR)*, 2020. 3, 4

[23] Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In *Proc. International conference on algorithmic learning theory (ALT)*, pages 63–77. Springer, 2005. 3

[24] 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 *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, pages 6626–6637, 2017. 5

[25] Irina Higgins, David Amos, David Pfau, Sebastien Racaniere, Loic Matthey, Danilo Rezende, and Alexander Lerchner. Towards a definition of disentangled representations. *arXiv preprint arXiv:1812.02230*, 2018. 1, 3

[26] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. β-VAE: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations (ICLR)*, 2017. 2

[27] Jun-Ting Hsieh, Bingbin Liu, De-An Huang, Li F Fei-Fei, and Juan Carlos Niebles. Learning to decompose and disentangle representations for video prediction. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, pages 517–526, 2018. 4

[28] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, pages 1501–1510, 2017. 2, 5, 11

[29] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *Proc. European Conference on Computer Vision (ECCV)*, pages 179–196, 2018. 2, 4, 5, 11, 12

[30] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4401–4410, 2019. 1, 2

[31] Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In *Proc. International Conference on Machine Learning (ICML)*, pages 2649–2658, 2018. 2

[32] Diederik P Kingma and Max Welling. Auto-encoding variational Bayes. *International Conference on Learning Representations (ICLR)*, 2014. 12

[33] Tejas D. Kulkarni, William F. Whitney, Pushmeet Kohli, and Josh Tenenbaum. Deep convolutional inverse graphics network. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, page 2539–2547, 2015. 1

[34] Abhishek Kumar, Prasanna Sattigeri, and Avinash Balakrishnan. Variational inference of disentangled latent concepts from unlabeled observations. In *International Conference on Learning Representations (ICLR)*, 2018. 1, 2

[35] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *Proc. European Conference on Computer Vision (ECCV)*, pages 36–52, 2018. 2, 4

[36] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, pages 700–708, 2017. 2

[37] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. DeepFashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1096–1104, 2016. 7, 13

[38] Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Rätsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *International Conference on Learning Representations Workshops (ICLRW)*, 2019. 1

[39] Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Rätsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. A commentary on the unsupervised learning of disentangled representations. In *Proc. AAAI Conference on Artificial Intelligence (AAAI)*, pages 13681–13684, 2020. 1

[40] 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, 4, 12, 14

[41] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2337–2346, 2019. 6, 13

[42] Esser Patrick, Ekaterina Sutter, and Björn Ommer. A variational U-Net for conditional appearance and shape generation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8857–8866, 2018. 1

[43] Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron C. Courville. FiLM: Visual reasoning with a general conditioning layer. In *Proc. AAAI Conference on Artificial Intelligence (AAAI)*, 2018. 5

[44] Scott Reed, Kihyuk Sohn, Yuting Zhang, and Honglak Lee. Learning to disentangle factors of variation with manifold interaction. In *Proc. International Conference on Machine Learning (ICML)*, pages 1431–1439, 2014. 1

[45] Karl Ridgeway and Michael C. Mozer. Learning deep disentangled embeddings with the F-statistic loss. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, page 185–194, 2018. 1, 2

[46] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M. Lopez. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3234–3243, 2016. 5, 12

[47] Rui Shu, Yining Chen, Abhishek Kumar, Stefano Ermon, and Ben Poole. Weakly supervised disentanglement with guarantees. In *International Conference on Learning Representations (ICLR)*, 2020. 2

[48] N. Siddharth, B. Paige, J.-W. van de Meent, A. Desmaison, N. D. Goodman, P. Kohli, F. Wood, and P. Torr. Learning disentangled representations with semi-supervised deep generative models. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 1

[49] Zengjie Song, Oluwasanmi Koyejo, and Jiangshe Zhang. Toward a controllable disentanglement network. *arXiv preprint arXiv:2001.08572*, 2020. 2

[50] Thorvald Sørensen. A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on danish commons. *Royal Danish Academy of Sciences and Letters*, 5(4):1–34, 1948. 6

[51] Gábor J Székely, Maria L Rizzo, Nail K Bakirov, et al. Measuring and testing dependence by correlation of distances. *The annals of statistics*, 35(6):2769–2794, 2007. 3

[52] Frederik Träuble, Elliot Creager, Niki Kilbertus, Anirudh Goyal, Francesco Locatello, Bernhard Schölkopf, and Stefan Bauer. Is independence all you need? on the generalization of representations learned from correlated data. *arXiv preprint arXiv:2006.07886*, 2020. 1

[53] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016. 4

[54] Ruben Villegas, Jimei Yang, Seunghoon Hong, Xunyu Lin, and Honglak Lee. Decomposing motion and content for nat-

ural video sequence prediction. In *International Conference on Learning Representations (ICLR)*, 2017. 4

[55] Yijun Xiao and William Yang Wang. Disentangled representation learning with Wasserstein total correlation. *arXiv preprint arXiv:1912.12818*, 2019. 1, 2

[56] Xianglei Xing, Tian Han, Ruiqi Gao, Song-Chun Zhu, and Ying Nian Wu. Unsupervised disentangling of appearance and geometry by deformable generator network. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10354–10363, 2019. 4

[57] Jimei Yang, Scott E. Reed, Ming-Hsuan Yang, and Honglak Lee. Weakly-supervised disentangling with recurrent transformations for 3D view synthesis. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, pages 1099–1107, 2015. 1

[58] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 586–595, 2018. 5

[59] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, pages 465–476, 2017. 11

A. Maximizing Likelihood by Minimizing Mean Square Error

Let y denote a generic pixel in an image I, and $\tilde{y}$ the respective pixel in the reconstructed image $\tilde{I}$, obtained trough a learned decoding function.

If we assume the reconstruction error, denoted as ε, to be normally distributed (*i.e.* $\varepsilon \sim \mathcal{N}\left(0, \sigma^2\right)$), then, the predicted value $\tilde{y}$ is normally distributed around the true value y, thus $\tilde{y} \sim \mathcal{N}\left(y, \sigma^2\right)$. Based on this assumption, the probability density function can be defined as:

$$f\left(\tilde{y}|y, \sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(\tilde{y}-y)^2}{2\sigma^2}}. \tag{4}$$

Given a set of observations, *e.g.* the pixels of the image, we maximize the likelihood $\mathcal{L}$ as the product of the probability densities of the observations:

$$\mathcal{L} = \prod_{i=1}^{n} f\left(y_i|\tilde{y}_i, \sigma^2\right) = \left(2\pi\sigma^2\right)^{-n/2} e^{-\frac{\sum_{i=1}^{n}\left(y_i-\tilde{y}_i\right)^2}{2\sigma^2}}. \tag{5}$$

Assuming the variance of the error to be independent from the input variables, optimizing the latter formula is equivalent to optimize:

$$\log\left(\frac{\mathcal{L}}{\left(2\pi\sigma^2\right)^{-n/2}}\right) = -\frac{\sum_{i=1}^{n}\left(y_i-\tilde{y}_i\right)^2}{2\sigma^2}. \tag{6}$$

Thus, maximizing the original likelihood function is equivalent to minimizing $\sum_{i=1}^{n}\left(y_i-\tilde{y}_i\right)^2$, that is the scaled Mean Squared Error (MSE). Thus, by training the decoder to minimize MSE, we train it to maximize the Mutual Information (MI) between z and I.

After training the decoder G_θ (see Section 3.2), computing MSE equivalent to directly measuring the MI. There is a relationship between likelihood and MSE (shown below), but the likelihood acts as a lower bound to MI.

Relationship MSE - likelihood: Note that if we divide both parts of the equation by n and then we multiply by $-2\sigma^2$, we obtain:

$$\sum_{i=1}^{n}\frac{\left(y_i-\tilde{y}_i\right)^2}{n} = -\frac{2\sigma^2}{n}\cdot\log\frac{\mathcal{L}}{\left(2\pi\sigma^2\right)^{-n/2}}, \tag{7}$$

that is:

$$MSE = -\frac{2\sigma^2}{n}\log(\mathcal{L}) - \sigma^2\log(2\pi\sigma^2). \tag{8}$$

Since we assume homoscedastic distributions, *i.e.* fixed σ^2, Equation 8 can be expressed as:

$$MSE = -\frac{a}{n}\log(\mathcal{L}) - b, \tag{9}$$

where a and b are positive constants.

B. Detailed Application Description

Table 4 summarizes the *design* and *learning biases* of the methods presented in Section 4. Note that the biases are reported as modules, without indicating the way they are used in our experiments (*e.g.* AdaIN is reported without specifying that it is removed from the original MUNIT, but is added to PANet as a variant).

B.1. MUNIT for Image-to-Image Translation

Multimodal Unsupervised Image-to-image Translation (MUNIT) [29] does not impose strict constraints on the learned representations, and achieves disentanglement with both design and learning biases.

The basic assumption is that multi-domain images (a necessary *data bias*), share common content information, but differ in style. A content encoder maps images to multi-channel feature maps, by removing style with IN layers [28] (*design bias*). A second encoder extracts global style information with fully connected layers and global pooling. Finally, style and content are combined in a decoder with AdaIN modules [28] (*design bias*).

Disentanglement is additionally promoted with a bidirectional reconstruction loss [59] that enables style transfer. In order to learn a smooth representation manifold, two LR losses (*learning bias*) are applied on content and style extracted from input images: content LR penalizes the distance to the content extracted from reconstructed images,

whereas style LR encourages encoded style distributions to match their Gaussian priors. Finally, adversarial learning encourages realistic synthetic images.

B.2. SDNet for Medical Image Segmentation

SDNet [10] is a semi-supervised framework that disentangles medical images in anatomical features (content) and imaging-specific characteristics (style). Similarly to other models, SDNet uses separate content and style encoders, but here a segmentation network is applied on the content features trained with supervised objectives and annotated images (*data bias*).

However, in contrast to MUNIT, SDNet does not impose a design bias on the encoder, but rather on the content which is represented as multi-channel binary maps of the same resolution as the input (*design bias*).

This is obtained with a softmax and a thresholding function with the straight-through operator [4], such that any style is removed from the content. To encourage style features to encode residual information (and not content), a loss enforces the style representation to approximate a standard Gaussian, following the VAE formulation [32] (*learning bias*). In this setup, any information encoded in style comes at a cost, and thus encoding redundant information is prevented [2]. Furthermore, a LR loss of the style is employed to prevent posterior collapse of the decoder (*learning bias*).

Finally, style and content are combined to reconstruct the input image by applying a series of convolutional layers with feature-wise linear modulation (FiLM) conditioning. Similarly to AdaIN, FiLM modules are restrictive, allowing the style only to normalize the conditioned feature maps, and thus further discouraging the style from encoding content information (*design bias*).

B.3. PANet for Pose Estimation

For the pose estimation task, we consider a dual-stream autoencoder denoted as PANet [40]. PANet consists of two branches that decouple pose (content) and appearance (style) but employs heavily entangled encoders-decoders.

The content is represented as a multi-channel feature map, where each channel corresponds to a specific body part (since the number of parts are fixed, this imposes a strong *data bias*). A Gaussian distribution is applied to each feature map to remove any style information, whilst also preserving the spatial correspondence (*design bias*).

The corresponding style information is extracted from the encoder features using average pooling (*design bias*). More critically, style vectors do not correspond to global image style, since they are applied to specific content parts during decoding (*design bias*).

Finally, disentanglement is encouraged with a transformation equivariance loss (*learning bias*). This ensures that the spatial transformations, such as translations and rotations, affect only the content, while the intensity ones, such as the color and texture information, affect only the style.

Table 4: Overview of the *design* and *learning biases* that are investigated in the context of the three investigated vision tasks: a) image-to-image translation (MUNIT), b) medical segmentation (SDNet), and c) pose estimation (PANet).

		MUNIT	SDNet	PANet
Design Bias	AdaIN	√		√
	Instance Normalization	√		
	SPADE		√	
	Binarization		√	
	MLP			√
Learning Bias	Latent Regression	√	√	
	KL Divergence		√	
	Equivariance			√

C. SYNTHIA-Cityscapes Description and MUNIT Training Setup

Data. We use SYNTHIA [46], which consists of over $20,000$ rendered images and corresponding pixel-level semantic annotations, where 13 classes of objects are labeled for aiding segmentation and scene understanding problems. We also use Cityscapes [14], which contains a set of diverse street scene stereo video sequences and over $5,000$ frames of high-quality semantic annotations, where 30 classes of instances are labeled in the segmentation masks.

Training setup. MUNIT achieves unsupervised multimodal image-to-image translation by minimizing the following loss function:

$$\mathcal{L}_{total} = \mathcal{L}_{GAN} + \lambda_1 \mathcal{L}_{rec} + \lambda_2 \mathcal{L}_{c-rec} + \lambda_3 \mathcal{L}_{s-rec}, \quad (10)$$

where $\mathcal{L}_{rec}$ is the image reconstruction loss, $\mathcal{L}_{c-rec}$ and $\mathcal{L}_{s-rec}$ denote the content and style reconstruction losses, and $\lambda_1 = 10$, $\lambda_2 = 1$ are the hyperparameters used by the authors in [29].

D. ACDC Description and SDNet Training Setup

Data. We use data from the Automatic Cardiac Diagnosis Challenge (ACDC) [5], which contains cardiac cine-MR images acquired from different MR scanners and resolution on 100 patients. Images were resampled to 1.37 mm^2/pixel resolution and cropped to 224×224 pixels. Manual segmentations are provided for the left ventricular cavity, the myocardium and right ventricle in the end-systolic and end-diastolic cardiac phases. In total there are 1920 images with

Table 5: Comparative evaluation of SDNet [10] variants on the ACDC [5] dataset with 100% annotation masks, using the proposed metrics. The *Dice Score* metric is used to measure the performance in terms of semantic segmentation. Arrows (↑, ↓) indicate the direction of each metric improvement.

		Design Bias		**Learning Bias**
SDNet	Original Model	SPADE	w/o Binarization	w/o KL Divergence and Latent Regression
$DC(C, S)$ (↓)	0.48	0.59	**0.43**	0.57
$DC(I, C)$ (↑)	0.97	0.94	**0.97**	0.95
$DC(I, S)$ (↑)	0.44	**0.57**	0.44	0.53
$IOB(I, C)$ (↑)	5.66	5.63	**6.21**	3.86
$IOB(I, S)$ (↑)	0.99	**1.02**	1.00	0.96
Dice Score (↑)	0.82	0.83	0.82	0.81

manual segmentations and 23,530 images with no segmentations.

Training setup. SDNet is trained by minimizing the following loss function:

$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_{KL} + \lambda_2 \mathcal{L}_{seg} + \lambda_3 \mathcal{L}_{rec} + \lambda_4 \mathcal{L}_{z_{rec}}, \quad (11)$$

where $\mathcal{L}_{KL}$ is the KL Divergence measured between the sampled and the predicted style vectors, $\mathcal{L}_{rec}$ is the image reconstruction loss, $\mathcal{L}_{seg}$ is the anatomy segmentation loss, and $\mathcal{L}_{z_{rec}}$ is the LR loss between the sampled and the re-encoded style vector. $\lambda_1 = 0.01, \lambda_2 = 10, \lambda_3 = 1$, and $\lambda_4 = 1$ are the hyperparameters used by the authors in [10].

D.1. SPADE Decoder

As described in Section 4 of the original manuscript, SDNet relies on a FiLM-based decoder to combine the content and style information and reconstruct the input image. The key characteristic of FiLM is that it gradually adds style information to the content-based reconstruction process. Additionally, an alternative approach for combining the content and style features is investigated, by using a SPADE decoder [41] to further expose the design bias added by the decoder architecture.

A SPADE block receives the content channels and projects them onto an embedding space using two convolutional layers to produce the modulation parameters (tensors) γ and β. These parameters are then used to scale (γ) and shift (β) the normalized activations of the style representation. We utilize multiple SPADE blocks to fuse content and style information at different levels of granularity during decoding. A schematic of the utilized SPADE decoder in the context of SDNet is depicted in Fig. 5.

D.2. Medical Segmentation (100% Annotations)

In Section 5.2, we present the results of the SDNet model variants trained with minimal supervision, using only the 1.5% of the provided ACDC [5] annotations. Here, we provide the results for the same experiment but using the 100% of the provided annotations. From the results reported in Table 5, it can be seen that when using strong inductive biases, such as the supervised losses in this experiment, the degree of disentanglement does not significantly affect the segmentation performance (utility).

E. DeepFashion Description and PANet Training Setup

Data. We use DeepFashion [37], a large-scale dataset with over 800,000 diverse images of people in different poses and clothing, that also has annotations of body joints. We only used full-body images, specifically 32,032 images for training and 7,997 images for testing.

Training setup. PANet is trained in an unsupervised

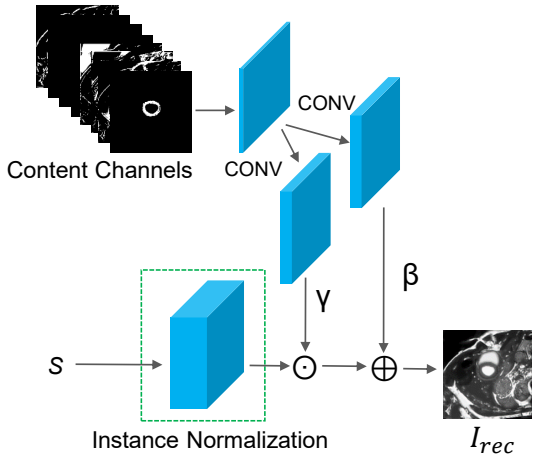

Figure 5: Detailed representation of the SPADE decoder [41] in the context of SDNET [10]. Style is denoted as S, while CONV represents the convolution operation. Note that γ and β parameters are applied to the normalized style activations through element-wise multiplication and addition, respectively.

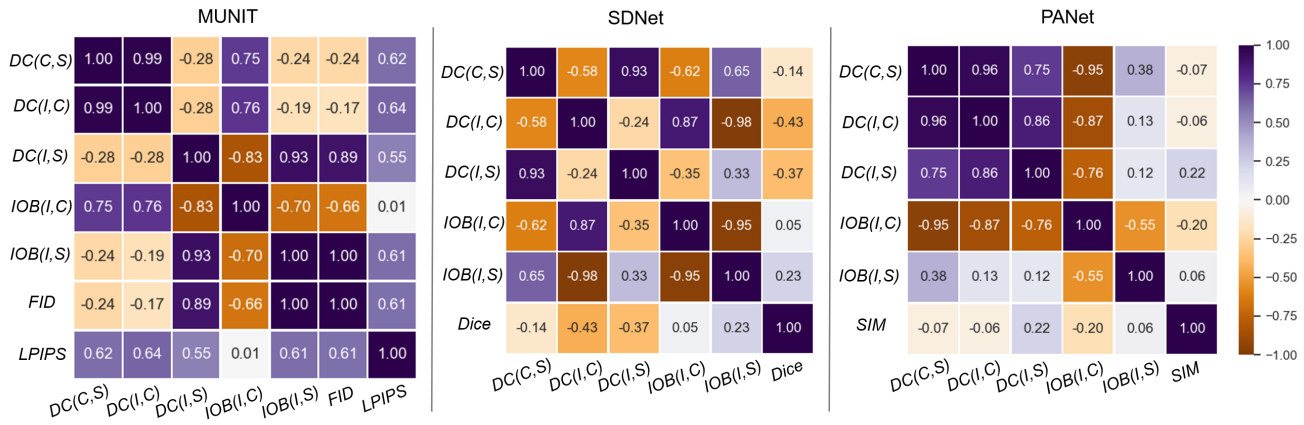

Figure 6: Pearson correlation of the proposed metrics across all applications/models visualized as heatmap. Values close to 1 and -1 indicate strong correlation.

way with the following loss function:

$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_{rec} + \lambda_2 \mathcal{L}_{equiv}, \tag{12}$$

where $\mathcal{L}_{rec}$ is the mean absolute error between the reconstructed and the input image. $\mathcal{L}_{equiv}$ is an equivariance cost, that ensures that the mean and covariance of the parts coordinates don't change after some style transformation. Based on the implementation details presented in [40], we set $\lambda_1 = \lambda_2 = 1$.

F. Metrics Correlation and Disentanglement-performance Trade-off

As noted in Section 5.4, we report that the proposed metrics are uncorrelated with each other. Here, we present the Pearson correlation computed between disentanglement and performance metrics for each of the investigated models. Intuitively, contrary to the desired low (or no) correlation between disentanglement metrics across all models (see Fig. 3), we would expect that the performance metric(s) of each application would be correlated with at least one DC or IOB variant. In fact, this correlation can be exploited to find the "sweet spot" between disentanglement and performance. Fig. 6 confirms our intuition for all investigated models, highlighting the strong correlation of FID and LPIPS in the MUNIT scenario, which is the only model that utilizes both C and S directly in the main task, *i.e.* I2I translation, and not in any parallel one.

G. Qualitative Evaluation

We visualize the content and style representations in order to reason about their interpretability. We consider the content semantic if distinct objects appear in different channels, whereas the style is semantic when images reconstructed while traversing the style manifold between two points have smooth appearance changes, and are realistic.

As an extension of the samples presented and discussed in Section 5.4, here we provide visualizations for all model variants. In particular, Figs. 7 and 8 depict several channels of content, as well as style traversals for different MUNIT and SDNet model variants, respectively. However, Fig. 9 presents solely content representations, as PANet does not assume a prior distribution on the style latent vector, thus style traversals are not possible. When interpolating between two style vectors, the originally proposed MUNIT produces realistic images, and smooth appearance changes. Instead, removing the LR constraint affects the image quality. Similarly, the original SDNet presents high image quality and smooth transitions, while removing the content Binarization leads to low intensity (style) diversity.

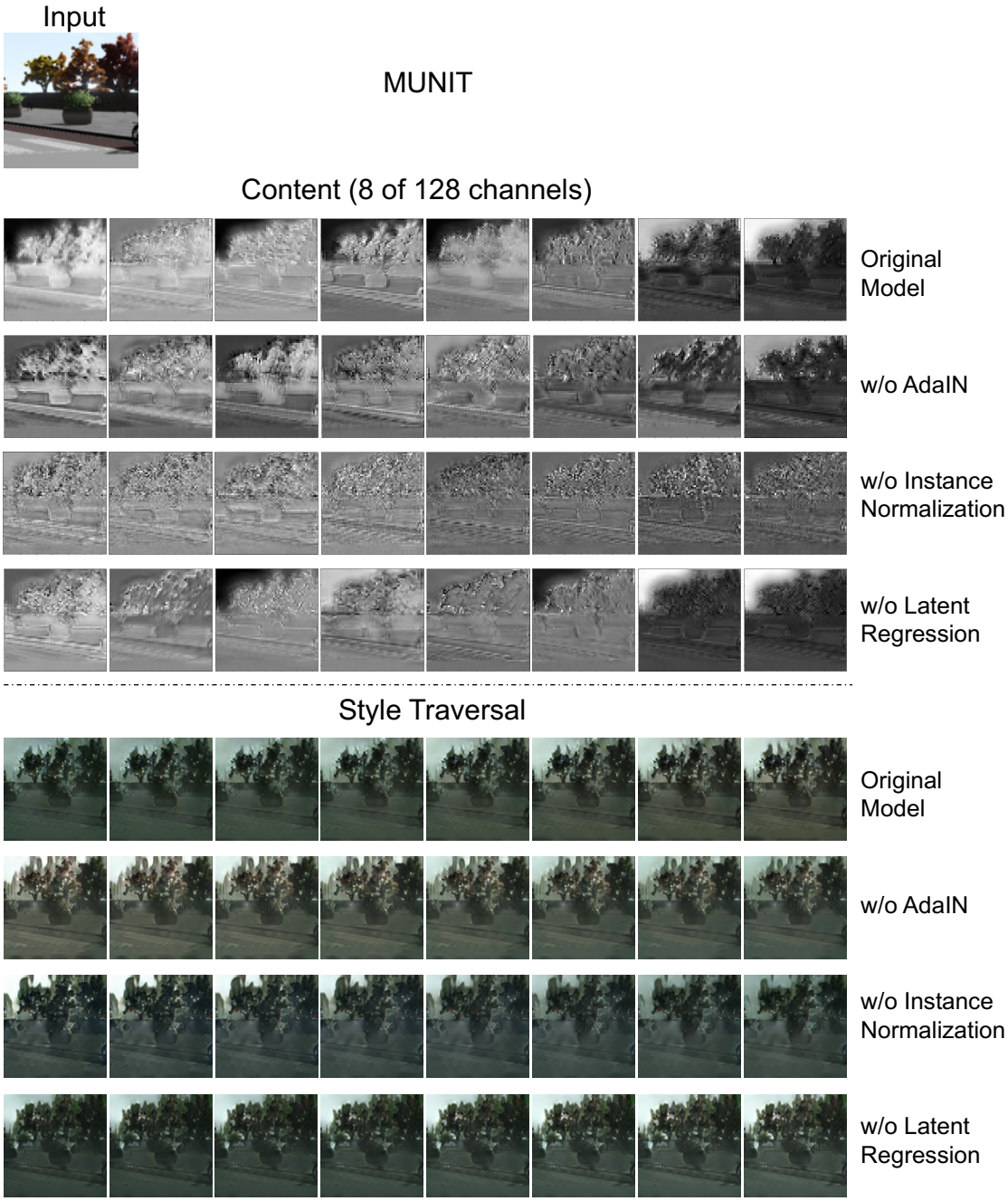

Figure 7: MUNIT: Qualitative examples to assess the interpretability of the content and style representations of the investigated model variants for different biases. For each variant, we show 8 channels of the content and 8 indicative style traversals. The input image is depicted at the top left of the figure.

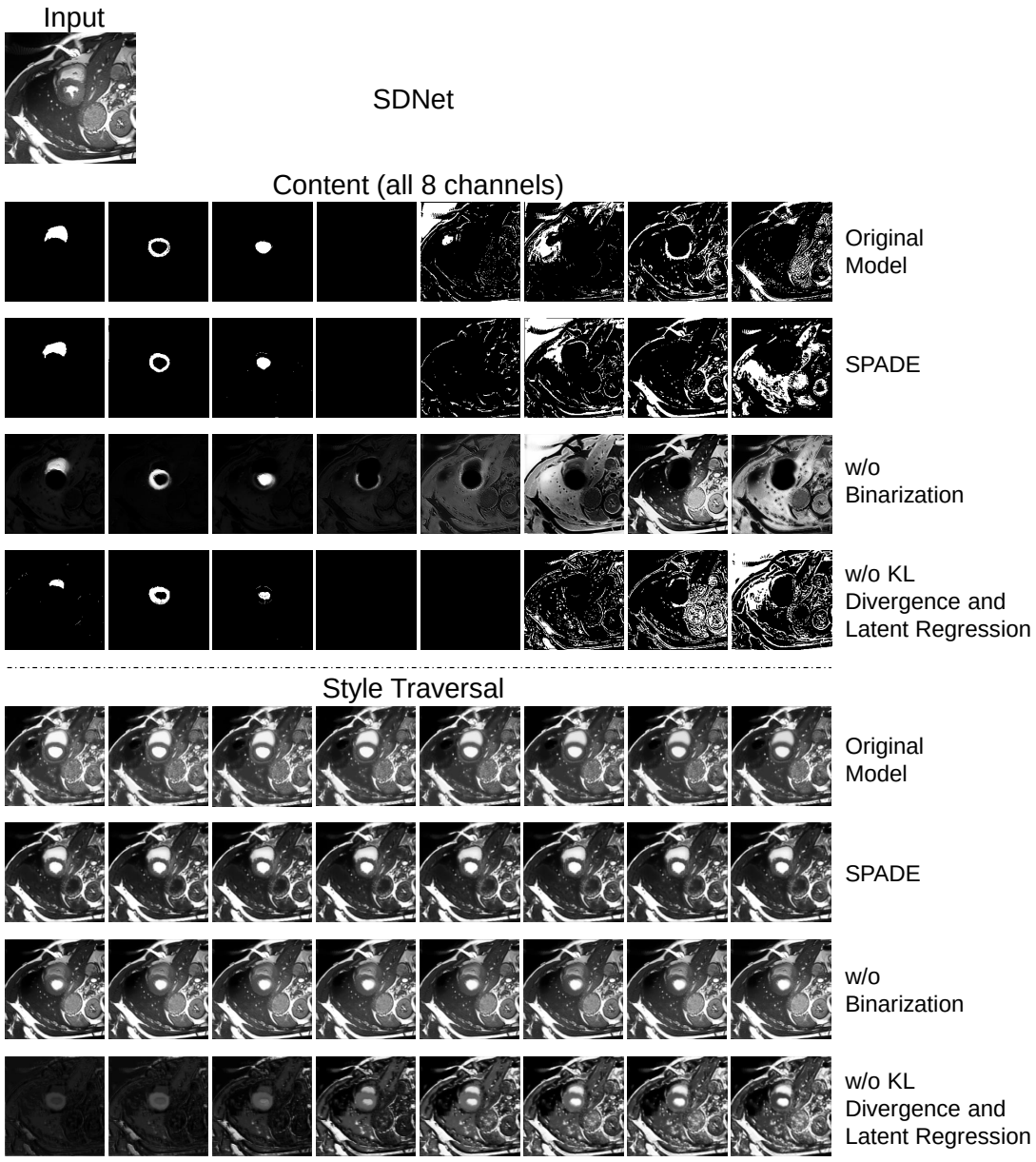

Figure 8: SDNet: Qualitative examples to assess the interpretability of the content and style representations of the investigated model variants for different biases. For each variant, we show 8 channels of the content and 8 indicative style traversals. The input image is depicted at the top left of the figure.

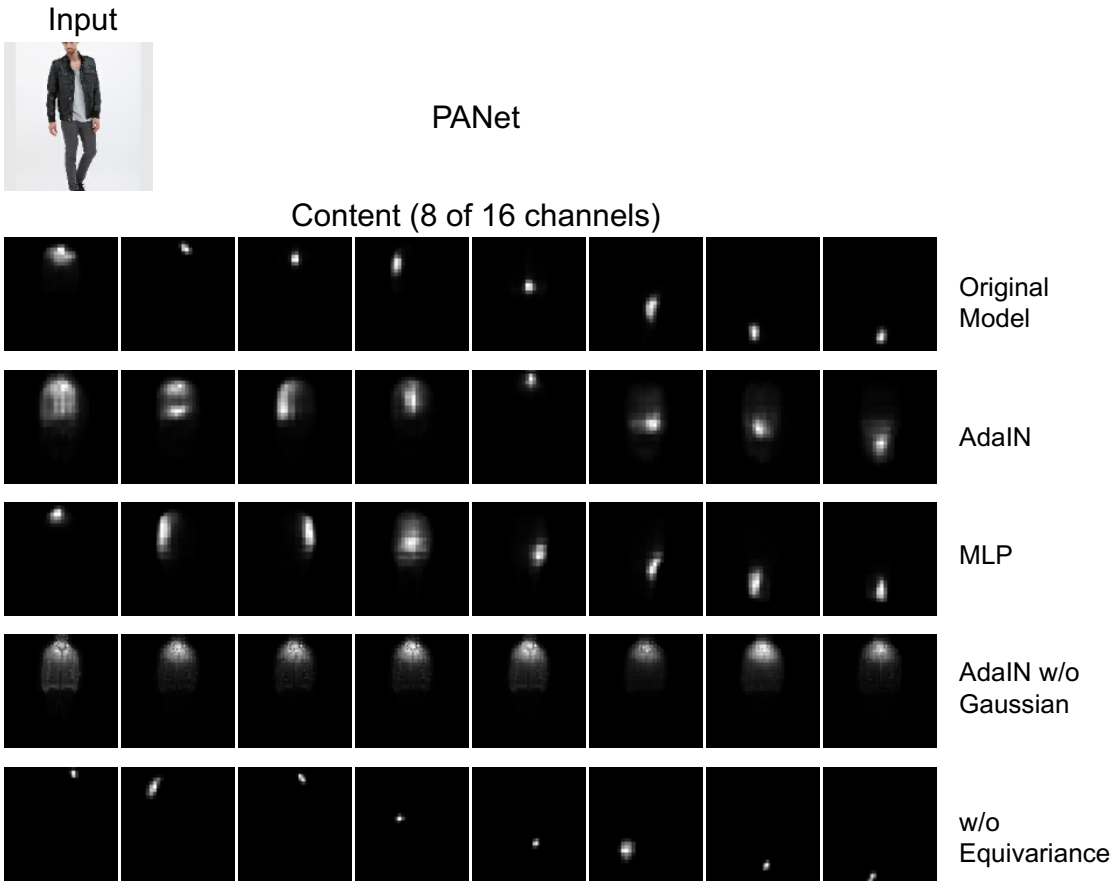

Figure 9: PANet: Qualitative examples to assess the interpretability of the content and style representations of the investigated model variants for different biases. For each variant, we show 8 channels of the content. Note that since PANet does not assume a prior distribution on the style, no style are shown. The input image is depicted at the top left of the figure.