CONTENT-STYLE LEARNING FROM UNALIGNED DOMAINS: IDENTIFIABILITY UNDER UNKNOWN LATENT DIMENSIONS

Sagar Shrestha and Xiao Fu
School of Electrical Engineering and Computer Science
Oregon State University
Corvallis, OR 97331, USA
{shressag,xiao.fu}@oregonstate.edu

ABSTRACT

Understanding identifiability of latent content and style variables from unaligned multi-domain data is essential for tasks such as domain translation and data generation. Existing works on content-style identification were often developed under somewhat stringent conditions, e.g., that all latent components are mutually independent and that the dimensions of the content and style variables are known. We introduce a new analytical framework via cross-domain *latent distribution matching* (LDM), which establishes content-style identifiability under substantially more relaxed conditions. Specifically, we show that restrictive assumptions such as component-wise independence of the latent variables can be removed. Most notably, we prove that prior knowledge of the content and style dimensions is not necessary for ensuring identifiability, if sparsity constraints are properly imposed onto the learned latent representations. Bypassing the knowledge of the exact latent dimension has been a longstanding aspiration in unsupervised representation learning—our analysis is the first to underpin its theoretical and practical viability. On the implementation side, we recast the LDM formulation into a regularized multi-domain GAN loss with coupled latent variables. We show that the reformulation is equivalent to LDM under mild conditions—yet requiring considerably less computational resource. Experiments corroborate with our theoretical claims.

1 INTRODUCTION

In multi-domain learning, "domains" are typically characterized by a distinct "style" that sets their data apart from others (Choi et al., 2020). Take handwritten digits as an example: writing styles of different persons can define different domains. Shared information across all domains, such as the identities of the digits in this case, is termed as "content". Learning content and style representations from multi-domain data facilitates many important applications, e.g., domain translation (Huang et al., 2018), image synthesis (Choi et al., 2020), and self-supervised representation learning (Von Kügelgen et al., 2021; Lyu et al., 2022; Daunhawer et al., 2023); see more in Huang et al. (2018); Lee et al. (2020); Choi et al. (2020); Wang et al. (2016); Yang et al. (2020); Wu et al. (2019).

Recent advances showed that understanding the *identfiability* of the latent content and style components from multi-domain data allows to design more reliable, predicable, and trustworthy learning systems (Hyvarinen et al., 2019; Lyu et al., 2022; Xie et al., 2023; Kong et al., 2022; Shrestha & Fu, 2024; Gresele et al., 2020; Gulrajani & Hashimoto, 2022). A number of works studied content/style identifiability when the multi-domain data have *sample-to-sample cross-domain alignment* according to shared contents. Specifically, identifiability was established for sample-aligned multi-domain settings under the assumption that multi-domain data are linear and nonlinear mixtures of latent content and style components, in the context of canonical correlation analysis (CCA), multiview analysis and self-supervised learning (SSL); see Ibrahim et al. (2021); Sørensen et al. (2021); Wang & Isola (2020); Von Kügelgen et al. (2021); Lyu et al. (2022); Karakasis & Sidiropoulos (2023); Daunhawer et al. (2023)

When cross-domain samples are *unaligned*, it becomes significantly more challenging to establish identifiability of the content and style components. The recent works in Xie et al. (2023); Sturma et al. (2023); Kong et al. (2022); Timilsina et al. (2024) made meaningful progresses towards this goal. These works considered mixture models of content and style for each domain, similar to those in Lyu et al. (2022); Von Kügelgen et al. (2021); Ibrahim et al. (2021); Sørensen et al. (2021); Karakasis & Sidiropoulos (2023); Daunhawer et al. (2023), but without cross-domain alignment. The new results in (Xie et al., 2023; Sturma et al., 2023; Kong et al., 2022; Timilsina et al., 2024) provide theory-backed solutions to a suite of timely and important applications, e.g., cross-language retrieval, multimodal single cell data alignment, causal representation learning, and image data translation and generation.

Challenges. The content-style identifiability results in existing unaligned multi-domain learning works are intriguing and insightful, but some challenges remain. First, the conditions used in their proofs have a number of restrictions, which limits the proof's applicability in many cases. For example, Sturma et al. (2023); Timilsina et al. (2024) assume that the all data reside in a linear subspace, which is over-simplification of reality; Xie et al. (2023); Kong et al. (2022) assume that the content and style variables are component-wise independent and that a large number of domains exist—both can be hard to fulfil. Second, the existing identifiability analyses in unaligned multi-domain learning (Xie et al., 2023; Kong et al., 2022; Sturma et al., 2023; Timilsina et al., 2024) (as well as those in aligned multi-domain learning) all need to know the dimensions of the content and style variables, which are not available in practice. Selecting these dimensions often involves extensive trial and error.

Contributions. In this work, we advance the analytical and computational aspects of content-style learning from unaligned multi-domain data. Our detailed contributions are as follows:

(i) Enhanced Identifiability of Content and Style: We propose a content-style identification criterion via constrained *latent distribution matching* (LDM). We show that the identifiability conditions under LDM are much more relaxed relative to those in existing works. Specifically, our results hold for nonlinear mixture models, as opposed to the linear ones used in Sturma et al. (2023); Timilsina et al. (2024). Unlike Xie et al. (2023); Kong et al. (2022); Sturma et al. (2023), no elementwise mutual independence assumption is needed in our proof. More importantly, our result holds for as few as *two* domains (whereas Xie et al. (2023); Kong et al. (2022) needs the existence of a large number of domains). The new results widens the applicability of content-style identifiable models in a substantial way.

(ii) Content-Style Identifiability under Unknown Latent Dimensions: We consider the scenario where the latent content and style dimensions are unknown—which is the case in practical settings. Note that existing works determine the content and style dimensions often by heuristics, e.g., trial-and-error. However, wrongly selected latent dimensions can largely degrade the performance of some tasks; e.g., an over-estimated style dimension hinders the diversity of data in generation tasks (see Sec. 6). We show that, by imposing proper sparsity constraints onto the LDM formulation, the content-style identifiability is retained even without knowing the exact latent dimensions. To our knowledge, this result is the first of the kind in the context of nonlinear mixture identification.

(iii) Efficient Implementation: We prove that the LDM formulation is equivalent to a sparsity-constrained, latent variable-coupled muti-domain GAN loss, under reasonable conditions. Directly realizing the LDM formulation would impose multiple complex modules, including the DM and content-style separation modules, in the learned latent domain. Simultaneously learning the latent space and optimizing these modules can be computationally involved. The GAN-based formulation circumvents such complicated operations and thus substantially simplifies the implementation.

For theory validation, we perform experiments over a series of image translation and generation tasks.

Notation. Please see Appendix A.1 for detailed notation designation. A particular remark is that $\mathbb{P}_{\boldsymbol{x}}$ and $p_{\boldsymbol{x}}(\cdot)$ represent the probability measure of $\boldsymbol{x}$ and the probability density function (PDF) of $\boldsymbol{x}$, respectively. The "push forward" notation $[\boldsymbol{f}]_{\#\mathbb{P}_{\boldsymbol{x}}}$ means the distribution of $\boldsymbol{f}(\boldsymbol{x})$.

2 BACKGROUND

Content-Style Modeling in Multi-Domain Analysis. Consider the case where the data are acquired over N domains $\mathcal{X}^{(n)} \subseteq \mathbb{R}^d$, where $n = 1, \ldots, N$. We assume that any sample from domain n can be represented as a function (or, a nonlinear mixture) of content and style components, i.e.,

$$\boldsymbol{c} \sim \mathbb{P}_{\boldsymbol{c}},\ \boldsymbol{s}^{(n)} \sim \mathbb{P}_{\boldsymbol{s}^{(n)}},\ \boldsymbol{x}^{(n)} = \boldsymbol{g}(\boldsymbol{c}, \boldsymbol{s}^{(n)}), \tag{1}$$

where $\mathbb{P}_{\boldsymbol{s}^{(n)}}$ and $\mathbb{P}_{\boldsymbol{c}}$ are distributions of the style components in nth domain and the content components, respectively. Let $\mathcal{C} \subseteq \mathbb{R}^{d_{\rm C}}$ and $\mathcal{S}^{(n)} \subseteq \mathbb{R}^{d_{\rm S}}$ be the open set supports of $\mathbb{P}_{\boldsymbol{c}}$ and $\mathbb{P}_{\boldsymbol{s}^{(n)}}$. Then, we define $\mathcal{X}^{(n)} = \{\boldsymbol{g}(\boldsymbol{c}, \boldsymbol{s}^{(n)}) | \boldsymbol{c}, \boldsymbol{s}^{(n)} \in \mathcal{C} \times \mathcal{S}^{(n)}\} \subseteq \mathbb{R}^d$ as the support of $\boldsymbol{x}^{(n)} \sim \mathbb{P}_{\boldsymbol{x}^{(n)}}$. Let $\mathcal{X} = \cup_{n=1}^N \mathcal{X}^{(n)} \subseteq \mathbb{R}^d$ and $\mathcal{S} = \cup_{n=1}^N \mathcal{S}^{(n)} \subseteq \mathbb{R}^{d_{\rm S}}$ represent the whole data space and the whole style space, respectively. We assume that the nonlinear function $\boldsymbol{g} : \mathcal{C} \times \mathcal{S} \to \mathcal{X}$ is a differentiable *bijective* function. This is a common assumption in latent component identification works, e.g., Von Kügelgen et al. (2021); Hyvarinen et al. (2019); Khemakhem et al. (2020), which basically says that every data sample has an associated unique representation in a latent domain. A remark is that although $\mathcal{X} \subseteq \mathbb{R}^d$ and d might be greater than $d_{\rm S} + d_{\rm C}$, the bijective property can hold as $\mathcal{X}$ resides within a low dimensional manifold (Von Kügelgen et al., 2021).

The model in (1) is widely adopted (explicitly or implicitly) in multi-domain analysis; see examples from Huang et al. (2018); Lee et al. (2020); Choi et al. (2020); Wang et al. (2016); Yang et al. (2020); Wu et al. (2019). This model makes sense when the "domains" are participated using distinguishable semantic meaning; e.g., in Fig. 1, "style" includes the writing manners (handwritten/printed) and display background colors (black/gray). Under the model in (1), learning $\boldsymbol{g}$ (and its inverse $\boldsymbol{f}$) as well as the latent components $\boldsymbol{c}$ and $\boldsymbol{s}^{(n)}$ is the key to facilitate a number of important applications.

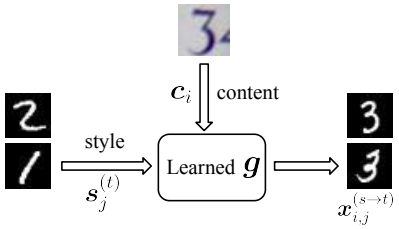


Figure 1: Cross-domain translation from source domain s to target domain t.

Application: Cross-Domain Translation. Learning content and style components from a sample in the source domain $(\boldsymbol{c}_i, \boldsymbol{s}_i^{(s)}) = \boldsymbol{f}(\boldsymbol{x}_i^{(s)})$ and a sample from the target domain $(\boldsymbol{c}_j, \boldsymbol{s}_j^{(t)}) = \boldsymbol{f}(\boldsymbol{x}_j^{(t)})$ can assist translate $\boldsymbol{x}_i^{(s)}$ to its corresponding representation in the target domain. This can be realized by generating a new sample $\boldsymbol{x}_{i,j}^{(s\to t)} = \boldsymbol{g}(\boldsymbol{c}_i, \boldsymbol{s}_j^{(t)})$; see Fig. 1 for illustration and Lyu et al. (2022); Huang et al. (2018); Wang et al. (2016).

Application: Data Generation. If $\boldsymbol{c}$ and $\boldsymbol{s}^{(n)}$ can be learned from the samples, then one can also learn the distributions $\mathbb{P}_{\boldsymbol{c}}$ and $\mathbb{P}_{\boldsymbol{s}^{(n)}}$ using off-the-shelf distribution learning tools, e.g., GAN (Goodfellow et al., 2014). This way, one can draw samples from the distributions, i.e., $\boldsymbol{c}_{\rm new} \sim \mathbb{P}_{\boldsymbol{c}},\ \boldsymbol{s}_{\rm new}^{(n)} \sim \mathbb{P}_{\boldsymbol{s}^{(n)}}$ and generate new samples $\boldsymbol{x}_{\rm new}^{(n)} = \boldsymbol{g}(\boldsymbol{c}_{\rm new}, \boldsymbol{s}_{\rm new}^{(n)})$ with intended styles.

Other Applications. We should mention that the content-style modeling is also a critical perspective for understanding representation learning paradigms, e.g., the SSL frameworks (Von Kügelgen et al., 2021; Lyu et al., 2022; Daunhawer et al., 2023; Wang & Isola, 2020).

Content-Style Identifiability. In recent years, the identifiability of $\boldsymbol{f}$, $\boldsymbol{c}$ and $\boldsymbol{s}^{(n)}$ started drawing attention, due to its usefulness in building more reliable/predictable systems.

Aligned Domains: Results from Self-Supervised Learning (SSL). The works (Von Kügelgen et al., 2021; Daunhawer et al., 2023; Lyu et al., 2022; Karakasis & Sidiropoulos, 2023) studied content identifiability in the context of representation learning, in particular, SSL and multiview learning. It was shown that when $N = 2$, if content-shared pairs $\{\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}\}$ are available, then enforcing $\boldsymbol{f}(\boldsymbol{x}^{(1)}) = \boldsymbol{f}(\boldsymbol{x}^{(2)}),\ \forall$ content-shared pairs $(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)})$ can provably learn $\boldsymbol{c}$, under reasonable conditions. The learning criterion can be realized by various loss functions, e.g., Euclidean fitting-based (Lyu et al., 2022; Karakasis & Sidiropoulos, 2023) and contrastive loss-based (Von Kügelgen et al., 2021; Daunhawer et al., 2023) criteria. The identifiability of the style components was also considered under similar aligned domain settings; see (Lyu et al., 2022; Eastwood et al., 2023).

Unaligned Domains: Progresses and Challenges. Aligned samples are readily available in applications such as data-augmented SSL (Von Kügelgen et al., 2021; Daunhawer et al., 2023; Lyu et al.,

2022). However, in other applications such as image style translation and image generation, aligned samples are hard to acquire (Zhu et al., 2017). For unaligned multi-domain data, the identfiiability issue of content and style has also been recently addressed. For example, the work of Sturma et al. (2023) extended the linear ICA model to unaligned multi-domain settings, in the context of causal learning. The work of Timilsina et al. (2024) took a similar linear mixture model but showed content-style identifiability under more relaxed conditions. The work of Xie et al. (2023); Kong et al. (2022) proved content-style identifiability under a more realistic nonlinear mixture model similar to that in (1). However, the main result there relies on a number of somewhat stringent conditions. That is, two notable assumptions in Xie et al. (2023); Kong et al. (2022) boil down to (i) that all components in $\boldsymbol{z} = (\boldsymbol{c}, \boldsymbol{s}^{(n)})$ are elementwise statistically independent given the domain index n; and (ii) that there exist at least $2d_{\mathrm{S}} + 1$ domains. These conditions can be hard to fulfil. See more detailed discussions on existing results in Appendix B.

The Dimension Knowledge Challenge. Notably, all the existing works in this domain (under both aligned and unaligned settings) assume that the dimensions of $\boldsymbol{c}$ and $\boldsymbol{s}^{(n)}$ are known. However, in mixture model learning, such knowledge is hard to acquire (especially in the nonlinear mixture case). As we will show, using wrongly selected d_{C} and d_{S} can be rather detrimental to content-style learning tasks—e.g., an over-estimated style dimension could lead to a serious lack of diversity in generated new samples. Consequently, the dimensions are often selected by extensive trial and error in practice.

3 Main Result

In this work, we revisit content-style learning from a latent distribution matching (LDM) viewpoint. Recall that $\boldsymbol{c}$ and $\boldsymbol{s}^{(n)}$ represent the content and the style of the nth domain, respectively. We assume:

Assumption 3.1 (Block Independence)**.** The block variables $\boldsymbol{c} \in \mathbb{R}^{d_{\mathrm{C}}}$ and $\{\boldsymbol{s}^{(n)} \in \mathbb{R}^{d_{\mathrm{S}}}\}_{n=1}^{N}$ are statistically independent, i.e., $p(\boldsymbol{c}, \boldsymbol{s}^{(1)}, \ldots, \boldsymbol{s}^{(N)}) = p_{\boldsymbol{c}}(\boldsymbol{c}) \prod_{n=1}^{N} p_{\boldsymbol{s}^{(n)}}(\boldsymbol{s}^{(n)})$.

The assumption was used in various multi-domain models (Lyu et al., 2022; Eastwood et al., 2023; Wang et al., 2016; Choi et al., 2020; Timilsina et al., 2024). It makes sense when the styles can be combined with contents in an "arbitrary" way without affecting the contents (e.g., the writing style of digits can change freely without affecting the identity of the digits). Next, we will use this assumption to build our learning criterion. We propose the following learning criterion:

$$\begin{aligned} &\text{find } \boldsymbol{f} : \mathcal{X} \to \mathbb{R}^{d_{\mathrm{C}}+d_{\mathrm{S}}} \text{ injective} \\ &\text{s.t. } [\boldsymbol{f}_{\mathrm{C}}]_{\#\mathbb{P}_{\boldsymbol{x}^{(i)}}} = [\boldsymbol{f}_{\mathrm{C}}]_{\#\mathbb{P}_{\boldsymbol{x}^{(j)}}},\ i \neq j, \forall i, j \in [N], \quad \text{(distribution matching)} \qquad (2a) \\ &\quad [\boldsymbol{f}_{\mathrm{S}}]_{\#\mathbb{P}_{\boldsymbol{x}^{(n)}}} \perp\!\!\!\perp [\boldsymbol{f}_{\mathrm{C}}]_{\#\mathbb{P}_{\boldsymbol{x}^{(n)}}}, \forall n \in [N], \quad \text{(block-indep. enforcing)} \qquad (2b) \end{aligned}$$

where $\boldsymbol{f}_{\mathrm{C}}(\boldsymbol{x}^{(n)}) \in \mathbb{R}^{d_{\mathrm{C}}}$ represents the first d_{C} outputs of $\boldsymbol{f}$ that are designated to represent the content components, $\boldsymbol{f}_{\mathrm{S}}(\boldsymbol{x}^{(n)}) \in \mathbb{R}^{d_{\mathrm{S}}}$ represents the learned style from domain n, Eq. (2a) matches the distributions of $\boldsymbol{f}_{\mathrm{C}}(\boldsymbol{x}^{(i)})$ and $\boldsymbol{f}_{\mathrm{C}}(\boldsymbol{x}^{(j)})$—i.e., the learned contents from domains i and j, respectively—and Eq. (2b) imposes a block independence constraint on the learned content $\boldsymbol{f}_{\mathrm{C}}(\boldsymbol{x}^{(n)})$ and style $\boldsymbol{f}_{\mathrm{S}}(\boldsymbol{x}^{(n)})$ from each domain following Assumption (3.1).

3.1 Warm Up: Enhanced Identifiability with Known Latent Dimensions

We first show that the content-style identifiability under (1) and known d_{C} and d_{S} can be substantially enhanced relative to existing works. We will remove the need for the dimension knowledge in the next subsection. To establish identifiability via solving Problem (2), we make the following assumption:

Assumption 3.2 (Domain Variability)**.** Let $\mathcal{A} \subseteq \mathcal{Z} := \mathcal{C} \times \mathcal{S}$ be any measurable set that satisfies (i) $\mathbb{P}_{\boldsymbol{z}^{(n)}}[\mathcal{A}] > 0$ for any $n \in [N]$ and (ii) $\mathcal{A}$ cannot be expressed as $\mathcal{B} \times \mathcal{S}$ for any set $\mathcal{B} \subset \mathcal{C}$. Then, there exists a pair of $i_{\mathcal{A}}, j_{\mathcal{A}} \in [N]$ such that the following holds:

$$\mathbb{P}_{\boldsymbol{z}^{(i_{\mathcal{A}})}}[\mathcal{A}] \neq \mathbb{P}_{\boldsymbol{z}^{(j_{\mathcal{A}})}}[\mathcal{A}], . \qquad (3)$$

Note that for any $\mathcal{A}$, we only need one pair of $(i_{\mathcal{A}}, j_{\mathcal{A}})$ to satisfy the condition, and the pair can change over different $\mathcal{A}$'s. Essentially, Eq. (3) requires that the styles have sufficiently diverse distributions. This assumption is a standard characterization for the distributional diversity of the domains in the literature; see Xie et al. (2023); Kong et al. (2022) and its variant Timilsina et al. (2024).

Under Assumptions 3.1 and (3.2), denote $\widehat{\boldsymbol{f}}$ as a solution to Problem (2). Then, we have:

Theorem 3.3 (Identifiability under Known Latent Dimensions)**.** *Under Eq. (1), suppose that Assumptions 3.1 and 3.2 hold, and that the $\widehat{\boldsymbol{f}}$ is differentiable. Then, we have $\widehat{\boldsymbol{f}}_{\mathrm{C}}(\boldsymbol{x}^{(n)}) = \boldsymbol{\gamma}(\boldsymbol{c})$ and $\widehat{\boldsymbol{f}}_{\mathrm{S}}(\boldsymbol{x}^{(n)}) = \boldsymbol{\delta}(\boldsymbol{s}^{(n)}), \forall n \in [N]$, where $\boldsymbol{\gamma} : \mathcal{C} \to \mathbb{R}^{d_{\mathrm{C}}}$ and $\boldsymbol{\delta} : \mathcal{S} \to \mathbb{R}^{d_{\mathrm{S}}}$ are injective functions.*

The proof of Theorem 3.3 is in Appendix C. Theorem 3.3 purports that the solution of Problem (2) identifies the model (1)—including the content/style components and the inverse mapping of the generative function $\boldsymbol{g}$ (up to $\boldsymbol{\gamma}$ and $\boldsymbol{\delta}$). Theorem 3.3 uses conditions that are significantly more relaxed relative to those in existing works Xie et al. (2023); Sturma et al. (2023); Kong et al. (2022); Timilsina et al. (2024). First, instead of assuming the elements of $\boldsymbol{z}^{(n)} = (\boldsymbol{c}, \boldsymbol{s}^{(n)})$ are statistically independent as in Xie et al. (2023); Sturma et al. (2023); Kong et al. (2022), our proof is based on the assumption that the content and styles are block independent (cf. Assumption 3.1). This block-independence assumption, which is the key for style identifiability, is similar to those in Lyu et al. (2022) and Timilsina et al. (2024)—but the former assumes aligned domains and the latter can only work under linear mixture models (see Theorem B.2 in Appendix B.2). Second, Theorem 3.3 does not need the existence of $N = 2d_{\mathrm{S}} + 1$ domains as in Xie et al. (2023); Kong et al. (2022) (see Theorem B.3 in Appendix B.3)—our result can hold over as few as $N = 2$ domains. As a result, our Theorem 3.3 applies to a considerably wider range of cases relative to those in existing works.

3.2 Identifiability without Dimension Knowledge

Theorem 3.3 still uses the knowledge of d_{C} and d_{S}. In this subsection, we propose a modifed learning criterion that does not use the exact dimension information. To proceed, let $\widehat{d}_{\mathrm{C}}$ and $\widehat{d}_{\mathrm{S}}$ denote the user-specified latent dimensions for $\boldsymbol{f}$, i.e., $\boldsymbol{f} : \mathcal{X} \to \mathbb{R}^{\widehat{d}_{\mathrm{C}}+\widehat{d}_{\mathrm{S}}}$, $\boldsymbol{f}_{\mathrm{C}} : \mathcal{X} \to \mathbb{R}^{\widehat{d}_{\mathrm{C}}}$ and $\boldsymbol{f}_{\mathrm{S}} : \mathcal{X} \to \mathbb{R}^{\widehat{d}_{\mathrm{S}}}$. Note that these dimensions need not to be exact. We consider the following learning criterion:

$$\underset{\boldsymbol{f}:\text{ injective}}{\text{minimize}} \quad \sum_{n=1}^{N} \mathbb{E}\left[\left\|\boldsymbol{f}_{\mathrm{S}}\left(\boldsymbol{x}^{(n)}\right)\right\|_0\right] \tag{4a}$$

$$\text{subject to } [\boldsymbol{f}_{\mathrm{C}}]_{\#\mathbb{P}_{\boldsymbol{x}^{(i)}}} = [\boldsymbol{f}_{\mathrm{C}}]_{\#\mathbb{P}_{\boldsymbol{x}^{(j)}}}, \forall i, j \in [N], \tag{4b}$$

$$[\boldsymbol{f}_{\mathrm{S}}]_{\#\mathbb{P}_{\boldsymbol{x}^{(n)}}} \perp\!\!\!\perp [\boldsymbol{f}_{\mathrm{C}}]_{\#\mathbb{P}_{\boldsymbol{x}^{(n)}}}, \forall n \in [N], \tag{4c}$$

Problem (4) minimizes the "effective dimension" of the extracted style component, while satisfying the distribution matching and independence constraints. The idea is to use excessive $\widehat{d}_{\mathrm{C}}$ and $\widehat{d}_{\mathrm{S}}$ so that one has enough dimensions to represent the content and style information. Note that trivial solutions could occur when using over-estimated $\widehat{d}_{\mathrm{C}}$ and $\widehat{d}_{\mathrm{S}}$. For instance, when $\boldsymbol{f}_{\mathrm{C}}$ is a constant function, $\boldsymbol{f}_{\mathrm{S}}$ can still be an injective function of $\boldsymbol{x}^{(n)}$ given large enough $\widehat{d}_{\mathrm{S}}$. This pathological solution satisfies both constraints (4b) and (4c). We use the sparsity objective in (4a) to "squeeze out" the redundant dimensions in $\boldsymbol{f}_{\mathrm{S}}$. This prevents the content information from "leaking" into the learned $\boldsymbol{f}_{\mathrm{S}}$. We formalize this intuition in the following theorem:

Theorem 3.4 (Identifiability without Dimension Knowledge)**.** *Assume that the conditions in Theorem 3.3 hold. Let $\widehat{\boldsymbol{f}}$ represent a solution of Problem (4) and $\widehat{\boldsymbol{f}}$ is differentiable. Assume the following conditions hold: (a) $\widehat{d}_{\mathrm{C}} \geq d_{\mathrm{C}}$ and $\widehat{d}_{\mathrm{S}} \geq d_{\mathrm{S}}$. (b) $0 < p_{\boldsymbol{z}^{(n)}}(\boldsymbol{z}) < \infty, \forall \boldsymbol{z} \in \mathcal{Z} = \mathcal{C} \times \mathcal{S}, \forall n \in [N]$. Then, there exists injective functions $\boldsymbol{\gamma} : \mathcal{C} \to \mathbb{R}^{\widehat{d}_{\mathrm{C}}}$ and $\boldsymbol{\delta} : \mathcal{S} \to \mathbb{R}^{\widehat{d}_{\mathrm{S}}}, \forall n \in [N]$ such that $\widehat{\boldsymbol{f}}_{\mathrm{C}}(\boldsymbol{x}^{(n)}) = \boldsymbol{\gamma}(\boldsymbol{c})$ and $\widehat{\boldsymbol{f}}_{\mathrm{S}}(\boldsymbol{x}^{(n)}) = \boldsymbol{\delta}(\boldsymbol{s}^{(n)}), \forall n \in [N]$.*

The proof of Theorem 3.4 is in Appendix D. Theorem 3.4 means that using Problem (4), there is no need to know d_{S} or d_{C} in advance. Also, note that no extra assumptions on $\boldsymbol{c}$ and $\boldsymbol{s}^{(n)}$ are needed on top of those in Theorem 3.3. Hence, the identifiability result has significant practical implications for content-style identification, where the latent dimension in practice is always hard to acquire.

4 Implementation: Sparsity-Regularized Multi-Domain GAN

At first glance, a conceptually straightforward realization of the learning criterion in Problem (2) could take the following form:

$$\underset{\boldsymbol{f}:\text{ injective}}{\text{minimize}} \sum_{i=1}^{N}\sum_{j>i}^{N} \mathcal{L}_{\text{DM}}(\boldsymbol{f}_{\text{C}}(\boldsymbol{x}^{(i)}), \boldsymbol{f}_{\text{C}}(\boldsymbol{x}^{(j)})) + \lambda \sum_{i=1}^{N} \mathcal{L}_{\text{indep}}(\boldsymbol{f}_{\text{C}}(\boldsymbol{x}^{(i)}), \boldsymbol{f}_{\text{S}}(\boldsymbol{x}^{(i)})), \tag{5}$$

where the first term and the second term promotes the distribution matching (DM) constraint (2a) and the independence constraint (2b), respectively. Similarly, Problem (4) can be implemented in a straightforward manner by adding a sparsity regularization term to Problem (5).

Remark 4.1. Problem (5) is potentially viable but can be costly: Both the LDM modules and the block independence regularization on the learned components often needs rather nontrivial optimization (see (Lyu et al., 2022)). Enforcing $\boldsymbol{f}$ to be injective also needs extra regularization, e.g., autoencoder type regularization (Lyu et al., 2022; Zhu et al., 2017) and entropy-type regularization (Von Kügelgen et al., 2021; Daunhawer et al., 2023).

In light of Remark 4.1, instead of using Problem (5), we reformulate Problems (2) and (4) as follows:

$$\min_{\boldsymbol{q},\boldsymbol{e}_{\text{C}},\boldsymbol{e}_{\text{S}}} \max_{\boldsymbol{d}^{(n)}} \sum_{n=1}^{N} \mathbb{E}\left[\log\left(\boldsymbol{d}^{(n)}\left(\boldsymbol{x}^{(n)}\right)\right) + \log\left(1-\boldsymbol{d}^{(n)}\left(\boldsymbol{q}\left(\boldsymbol{e}_{\text{C}}(\boldsymbol{r}_{\text{C}}), \boldsymbol{e}_{\text{S}}^{(n)}(\boldsymbol{r}_{\text{S}}^{(n)})\right)\right)\right)\right] \tag{6a}$$

$$\text{subject to } \boldsymbol{e}_{\text{S}}^{(n)}(\boldsymbol{r}_{\text{S}}^{(n)}) \text{ has minimal } \|\boldsymbol{e}_{\text{S}}^{(n)}(\boldsymbol{r}_{\text{S}}^{(n)})\|_0,\ \forall \boldsymbol{r}_{\text{S}}^{(n)}. \tag{6b}$$

The above approximates Problems (2) and (4) when the constraint (6b) is absent and active, respectively. In practice, the sparsity constraint can be approximated using sparsity regularization terms (e.g., ℓ_1 norm) easily. Denote $\widehat{d}_{\text{C}}$ and $\widehat{d}_{\text{S}}$ are the estimates of d_{C} and d_{S}, respectively. The idea is to learn invertible nonlinear mappings $\boldsymbol{e}_{\text{C}}$ and $\boldsymbol{e}_{\text{S}}^{(n)}$ that transform independent Gaussian variables (i.e., $\boldsymbol{r}_{\text{C}}$ and $\boldsymbol{r}_{\text{S}}^{(n)}$) to represent content $\boldsymbol{c}$ and style $\boldsymbol{s}^{(n)}$, respectively. Generate $\boldsymbol{r}_{\text{C}} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{\widehat{d}_{\text{C}}})$ and construct an invertible $\boldsymbol{e}_{\text{C}}$ such that $\boldsymbol{e}_{\text{C}}(\boldsymbol{r}_{\text{C}}) \in \mathbb{R}^{\widehat{d}_{\text{C}}}$. Similarly, construct invertible $\boldsymbol{e}_{\text{S}}^{(n)}$ such that $\boldsymbol{e}_{\text{S}}^{(n)}(\boldsymbol{r}_{\text{S}}^{(n)}) \in \mathbb{R}^{\widehat{d}_{\text{S}}}$ with $\boldsymbol{r}_{\text{S}}^{(n)} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{\widehat{d}_{\text{S}}})$. Then, the content and style are mixed by $\boldsymbol{q}$ to match the distribution of $\boldsymbol{x}^{(n)}$ using a logistic loss (i.e., GAN-type DM). In other words, the formulation looks for $\boldsymbol{e}_{\text{C}}$, $\boldsymbol{e}_{\text{S}}^{(n)}$ and $\boldsymbol{q}$ such that $\mathbb{P}_{\boldsymbol{x}^{(n)}} = \mathbb{P}_{\boldsymbol{q}^{(n)}}$, $\boldsymbol{q}^{(n)} = \boldsymbol{q}(\boldsymbol{e}_{\text{C}}(\boldsymbol{r}_{\text{C}}), \boldsymbol{e}_{\text{S}}^{(n)}(\boldsymbol{r}_{\text{S}}^{(n)}))$, $\forall n \in [N]$. This way, instead of directly learning $\boldsymbol{f}$, we learn the generative process $\boldsymbol{g}$ using $\boldsymbol{q}$. Our next theorem shows that $\boldsymbol{q}$ is indeed the inverse of $\boldsymbol{f}$ (up to some ambiguities).

To proceed, denote $\widehat{\mathcal{C}}$ and $\widehat{\mathcal{S}}^{(n)}$ as the sets representing the range of $\widehat{\boldsymbol{e}}_{\text{C}}$ and $\widehat{\boldsymbol{e}}_{\text{S}}^{(n)}$, respectively. Then, the effective domain of $\widehat{\boldsymbol{q}}$ is $\widehat{\mathcal{C}} \times \widehat{\mathcal{S}}$ where $\widehat{\mathcal{S}} = \cup_n \widehat{\mathcal{S}}^{(n)}$. We show that:

Theorem 4.2. *Let $(\widehat{\boldsymbol{q}}, \widehat{\boldsymbol{e}}_{\text{C}}, \widehat{\boldsymbol{e}}_{\text{S}}^{(n)}, \widehat{\boldsymbol{d}})$ be any differentiable optimal solution of Problem (6). Let $\mathcal{C}$ and $\mathcal{S}$ be simply connected open sets. Let $0 < p_{\boldsymbol{z}^{(n)}}(\boldsymbol{z}) < \infty, \forall \boldsymbol{z} \in \mathcal{Z} = \mathcal{C} \times \mathcal{S}$. Under the assumptions in Theorem 3.3, we have the following:*

(a) If $\widehat{d}_{\text{C}} = d_{\text{C}}$ and $\widehat{d}_{\text{S}} = d_{\text{S}}$ and (6b) is absent, then $\widehat{\boldsymbol{q}} : \widehat{\mathcal{C}} \times \widehat{\mathcal{S}} \to \mathcal{X}$ is bijective and $\widehat{\boldsymbol{f}} = \widehat{\boldsymbol{q}}^{-1}$ is also a solution of Problem (2).

(b) If $\widehat{d}_{\text{C}} > d_{\text{C}}$, $\widehat{d}_{\text{S}} > d_{\text{S}}$ and $\widehat{\boldsymbol{q}} : \widehat{\mathcal{C}} \times \widehat{\mathcal{S}} \to \mathcal{X}$ is bijective, $\widehat{\boldsymbol{f}} = \widehat{\boldsymbol{q}}^{-1}$ is also a solution of Problem (4)[1].

Problem (6) has a number of practical advantages over the direct implementation in Problem (5). Particularly, it avoids complex operations in the latent domain. In LDM, performing DM on $\boldsymbol{f}_{\text{C}}(\boldsymbol{x}^{(i)})$ and $\boldsymbol{f}_{\text{C}}(\boldsymbol{x}^{(j)})$ poses quite a nontrivial optimization process. This is because both of the inputs to the DM modules (i.e., $\boldsymbol{f}_{\text{C}}(\boldsymbol{x}^{(i)})$ for all $i \in [N]$) change from iteration to iteration—yet the DM module (e.g., GAN and Wasserstein distance-based DM (Goodfellow et al., 2014; Arjovsky et al., 2017)) itself often involves complex optimization with its own parameters updated on the fly. The new formulation performs GAN-based DM in the *data domain* and keeps one input (the real data) to every GAN module fixed. This reduces a lot of agony in optimization parameter tuning. Problem (6) also does not need any explicit constraint/regularization to enforce the block independence of $\boldsymbol{f}_{\text{C}}$ and

[1] Problem (4) requires $\widehat{\boldsymbol{f}}$ to be injective. Here, although the $\widehat{\boldsymbol{f}}$ learned by Problem (6) seems to be bijective (due to $\widehat{\boldsymbol{f}} = \widehat{\boldsymbol{q}}^{-1}$) instead of only injective, the bijectivity is w.r.t. the domains $\mathcal{X} \to \widehat{\mathcal{C}} \times \widehat{\mathcal{S}}$. The function is indeed only injective when considered w.r.t. $\mathcal{X} \to \mathbb{R}^{\widehat{d}_{\text{C}}+\widehat{d}_{\text{S}}}$; see Sec. A.2 "Injection, bijection, and surjection".

Table 1: Evaluation of the data generation task. Standard deviation reported using $\pm$ for style diversity

Method	FID (↓)			Style Diversity (↑)			Training time, hours (↓)		
	AFHQ	CelebA-HQ	CelebA-7	AFHQ	CelebA-HQ	CelebA-7	AFHQ	CelebA-HQ	CelebA-7
`Transitional-cGAN`	38.00	8.12	70.45	–	–	–	29.65	32.56	12.53
`StyleGAN-ADA`	8.17	5.89	72.10	–	–	–	28.46	32.26	11.55
`I-GAN`	6.28	5.91	**5.18**	0.16 ± 0.02	0.07 ± 0.03	0.07 ± 0.03	29.53	28.76	12.51
Proposed	**6.19**	**5.70**	5.27	$\mathbf{0.50 \pm 0.03}$	$\mathbf{0.36 \pm 0.04}$	$\mathbf{0.26 \pm 0.06}$	27.36	27.78	12.28

$\boldsymbol{f}_{\mathrm{S}}$ (which could be resource consuming (Lyu et al., 2022; Gretton et al., 2007)), as $\boldsymbol{e}_{\mathrm{C}}$ and $\boldsymbol{e}_{\mathrm{S}}^{(n)}$ are constructed to be block independent.

Another quite interesting observation is that, the proof of Theorem 4.2 (a) shows that the bijectivity constraint on $\boldsymbol{q}$ is automatically fulfilled when an additional condition (i.e., that $\mathcal{C}$ and $\mathcal{S}$ are simply connected) is met. This means that the LDM formulation would need extra modules, e.g., $\mathbb{E}\|\boldsymbol{r} \circ \boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{x}\|^2$, to impose injectivity constraints, even when d_{S} and d_{C} are known. When d_{C} and d_{S} are unknown, solving Problem (6) *per se* does not ensure $\boldsymbol{q}$ to be bijective. Nonetheless, we observed that not explicitly enforcing bijectivity in implementations does not affect the performance in practice. Similar phenomenon was observed in nICA implementations; see, e.g., (Hyvarinen & Morioka, 2017; Hyvarinen et al., 2019).

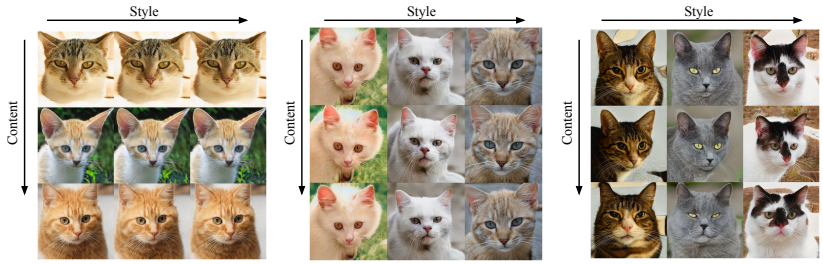
5 Related Works

Nonlinear ICA. Learning content and style components from a nonlinear mixture model is reminiscent of *nonlinear independent analysis* (nICA) (Hyvärinen & Pajunen, 1999; Hyvarinen & Morioka, 2017; Hyvarinen et al., 2019). Most nICA works were developed under single domain settings, with some recent generalizations to multiple views/domains (Gresele et al., 2020; Hyvarinen et al., 2019). Nonetheless, nICA requires that all the latent variables are (conditionally) independent. This is considered a somewhat restrictive assumption in content-style learning.

Content-Style Models in Aligned Multi-Domain Learning. Aligned multi-domain content-style learning is a key technique in data-augmented SSL and representation learning. There, it was shown that elementwise (conditional) independence is not needed, if the goal is to isolate content from style (Von Kügelgen et al., 2021; Lyu et al., 2022; Karakasis & Sidiropoulos, 2023). It was further shown that block independence (similar to Assumption 3.1) is the key to identify the style (Lyu et al., 2022; Daunhawer et al., 2023). However, all these works require cross-domain data alignment.

Content-Style Identification in Unaligned Multi-Domain Learning. Identifiability of unaligned multi-domain learning was studied in the context of various applications, e.g., image translation (Shrestha & Fu, 2024), data synthesis (Xie et al., 2023), cross-domain information retrieval Timilsina et al. (2024), and domain adaptation (Kong et al., 2022; Gulrajani & Hashimoto, 2022; Timilsina et al., 2024). In applications, content-style disentanglement has been applied in various tasks, such as (Hong et al., 2024; Huang et al., 2022; Dai et al., 2023). However, only a handful of works (Kong et al., 2022; Xie et al., 2023; Timilsina et al., 2024) have investigated the identifiability aspects. The work (Kong et al., 2022) postulated a similar content-style model as in (Xie et al., 2023) and came up with identifiability conditions similar to those in (Xie et al., 2023). The mostly related work to ours is (Xie et al., 2023), as both works are interested in content-style identification under (1). Our implementation in Problem (6) partially recovers the marginal distribution matching criterion in (Xie et al., 2023), despite the fact that our learning criteria started with an LDM perspective. Nonetheless, our method enjoys much less restrictive model assumptions for content-style identifiability. Our multi-domain GAN also admits more relaxed neural architecture (see Appendix G).

Content-Style Learning without Knowing Latent Dimensions. The SSL work (Von Kügelgen et al., 2021) presented a proof that essentially established that the content can be learned without knowing the exact dimension d_{C}. However, their result was under the assumption that the domains are *aligned*. In addition, the proof could not hold when style learning is also involved. Our proof solved these challenges. The work in (Xie et al., 2023) used a mask-based formulation to remove the requirement of knowing d_{C} and d_{S}. The mask-based formulation has the flavor of sparsity promoting as in our proposed method. However, they still need to know $d_{\mathrm{C}} + d_{\mathrm{S}}$, which is unlikely available in practice. In addition, the mask-based method in (Xie et al., 2023) did not have theoretical supports.



(a) I-GAN (Xie et al., 2023) (b) Proposed (w/o sparsity) (c) Proposed

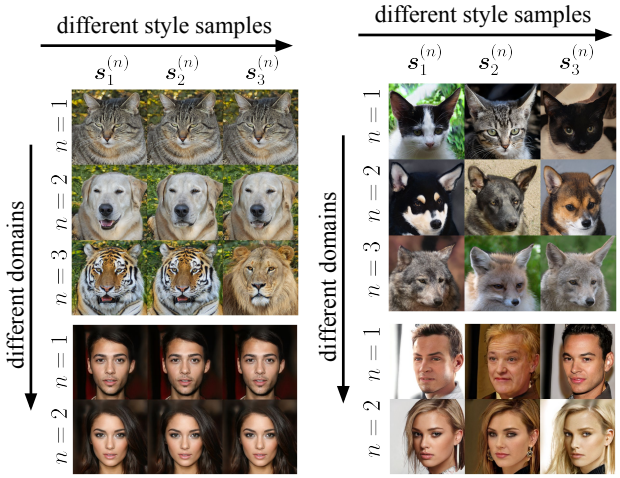
Figure 2: Samples generated by learning content (pose of cat) and style (type of cat) from AFHQ.

6 Numerical Examples

Multi-Domain Data Generation. For the data generation task, we validate our theoretical claims using three real world datasets: animal faces (AFHQ) (Choi et al., 2020), CelebA-HQ (Karras et al., 2018), and CelebA (Liu et al., 2015) with 3, 2, and 7 domains, respectively (see Appendix G.6).

The baselines here are I-GAN (Xie et al., 2023), StyleGAN-ADA (Karras et al., 2020) and Transitional-cGAN (Shahbazi et al., 2021).

Following Xie et al. (2023), we use StyleGAN2-ADA (Karras et al., 2020) to represent our generative function $\boldsymbol{q}$ in (6a). We set $\widehat{d}_{\mathrm{C}} = 384$ and $\widehat{d}_{\mathrm{S}} = 128$ in all the experiments. We use an ℓ_1 regularization term $\lambda \| \boldsymbol{e}_{\mathrm{S}}^{(n)}(\boldsymbol{r}_{\mathrm{S}}^{(n)}) \|_1$ to approximate the sparsity constraint in (6b). Note that other sparsity-promoting regularization (such as the ℓ_p function with $p < 1$) can also be easily used under our framework, which shows similar effectiveness (see Appendix H.4). We find that the algorithm is not very sensitive to the choice λ as any positive λ encourages sparsity of $\boldsymbol{e}_{\mathrm{S}}^{(n)}(\boldsymbol{r}_{\mathrm{S}}^{(n)})$. We use $\lambda = 0.3$ for all the experiments. More detailed experimental settings are in Appendix G.



(a) I-GAN (Xie et al., 2023) (b) Proposed

Figure 3: Samples generated by combining the same content $\bar{\boldsymbol{c}}$ with $\boldsymbol{s}^{(n)}$ for various n's in AFHQ and CelebA-HQ.

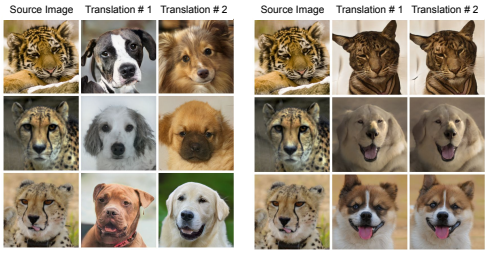
Fig. 2 shows the qualitative results for content-style identification using various methods for the cat domain ($n = 1$) of AFHQ. For each row, we fix the content part $\boldsymbol{c} = \boldsymbol{e}_{\mathrm{C}}(\boldsymbol{r}_{\mathrm{C}})$ (i.e., pose of the cat) and randomly sample different styles $\boldsymbol{s}^{(1)} = \boldsymbol{e}_{\mathrm{S}}^{(1)}(\boldsymbol{r}_{\mathrm{S}}^{(1)})$ where $\boldsymbol{r}_{\mathrm{S}}^{(1)} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{\widehat{d}_{\mathrm{S}}})$ to generate the images $\boldsymbol{x}^{(1)} = \boldsymbol{q}(\boldsymbol{c}, \boldsymbol{s}^{(1)})$. This way, the samples $\boldsymbol{s}_i^{(1)}$ for $i = 1, 2, \ldots$ correspond to various types of cats. Fig. 2 (a) shows that the I-GAN appears to generate the same type of cat even when repeatedly sampled from their learned distribution of $\boldsymbol{s}^{(1)}$. This suggests that the style components are not extracted properly. Fig. 2 (b) shows the result of using the proposed method without any sparsity regularization. As explained earlier, this can lead to learning constant content part with all information captured by the style part. Fig. 2 (b) corroborates with the intuition since we see little to no pose variation in the sampled contents (i.e., the three rows). Fig. 2 (c) shows the result of proposed method, i.e., Problem (6). One can see that both content and style parts demonstrate sufficient diversity, indicating well learned content and style distributions. Appendix H shows similar results for other domains and datasets.

Fig. 3 shows the generated samples of $\boldsymbol{x}^{(n)}$ for different n's using models learned from the AFHQ and CelebA-HQ datasets, which correspond to different species of animals (cat, dog, and tiger) and different genders of people, respectively. The top three rows in each figure correspond to the three

n's (i.e., three domains) of the AFHQ dataset, whereas the bottom two rows correspond to two n's of the CelebA-HQ dataset.

For the jth row associated with each dataset, we sample three different styles $\boldsymbol{s}_i^{(n_j)}, i \in \{1, 2, 3\}$ and combine it with a fixed content $\bar{\boldsymbol{c}}$ to generate the image $\boldsymbol{x}_i^{(n_j)} = \boldsymbol{q}(\bar{\boldsymbol{c}}, \boldsymbol{s}_i^{(n_j)})$ in the jth row and ith column.

Both the baseline `I-GAN` and our method can combine a fixed $\bar{\boldsymbol{c}}$ with $\boldsymbol{s}_i^{(n_j)}$ for different i to create content (pose)-consistent new data (see all the rows). However, one can see that the baseline `I-GAN` was not able to sample different styles in each domain. It seems that every domain n always repeatedly samples the same style components $\bar{\boldsymbol{s}}^{(n)}$ as the same images always appear in the same row. The proposed method can generate quite diverse style samples in all the domains. Additional results are in the Appendix H.



(a) Proposed (b) `I-GAN (Gen)`

Figure 4: Translation by combining content (pose) randomly sampled styles from the dog domain.

Figure 5: Guided translation by combining content (first column) with style (second column) of the images.

Table 1 shows the FID (Heusel et al., 2017), style diversity scores, and training time of the different methods. We use LPIPS distance (Zhang et al., 2018) between pairs of images with the same $\bar{\boldsymbol{c}}$ and different style samples from $\boldsymbol{s}^{(n)}$ to measure the style diversity. The diversity scores are averaged over 6,000 images across all domains, where every 10 images contain the same content with different styles. Note that the baselines `StyleGAN-ADA` and `Transitional-cGAN` do not learn content-style models, and thus the style diversity scores of theirs are not reported. One can see that the FID scores of the methods are similar, meaning that all methods generate realistic looking images. However, the style diversity of the proposed method is 3 to 5 times higher than the baseline over all datasets. The conditional generative models (`Transitional-cGAN` and `StyleGAN-ADA`) sometimes encountered convergence issues on specific datasets as reflected by their FID scores. Finally, the training time of all the methods are in the similar range, the proposed method being slightly faster for AFHQ and CelebA-HQ datasets.

Multi-Domain Translation.

Existing methods use a dedicated system for multi-domain translation (Choi et al., 2018; 2020; Yang et al., 2023). However, since a multi-domain generative model can already disentangle content and style (cf. Theorems 4.2 of this work), one can simply use the generative model for domain translation.

Given an image $\boldsymbol{x}^{(i)}$ in the source domain i, in order to extract the corresponding content $\boldsymbol{c}$ or style $\boldsymbol{s}^{(i)}$, one can simply solve $(\widehat{\boldsymbol{c}}, \widehat{\boldsymbol{s}}) = \arg\min_{\boldsymbol{c},\boldsymbol{s}} \mathsf{div}(\boldsymbol{q}(\boldsymbol{c}, \boldsymbol{s}), \boldsymbol{x}^{(i)})$, where div is some distance metric/divergence measure. There exists many approaches to solving the problem, often referred to as GAN inversion (Xia et al., 2022). In our case, we simply use the Adam optimizer for this inversion step. For div, we use a pre-trained VGG16 (Simonyan & Zisserman, 2014) neural network. More details are in Appendix G. To generate the desired translation, the GAN inversion-extracted content can be combined with a randomly sampled style $\boldsymbol{s}^{(t)} = \boldsymbol{e}_{\mathrm{S}}^{(t)}(\boldsymbol{r}_{\mathrm{S}}^{(t)}), \boldsymbol{r}_{\mathrm{S}}^{(t)} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{\widehat{d}_{\mathrm{S}}})$ from the target domain t. Additionally, one can also extract style from an image in target domain and combine it with extracted content from the source domain for guided translation.

Table 2: Quantitative evaluation of all methods for the translated images.

Method	FID (↓)		Style Diversity (↑)		Training time (hours)		
	AFHQ	CelebA-HQ	AFHQ	CelebA-HQ	Generation	Translation	Total
StarGANv2	16.83	**13.67**	0.45 ± 0.03	$\mathbf{0.45 \pm 0.03}$	–	50.83	50.83
SmoothGAN	53.68	29.69	0.14 ± 0.04	0.09 ± 0.03	–	30.90	30.90
I-GAN (Tr)	19.57	15.26	0.46 ± 0.03	0.29 ± 0.05	29.53	67.46	96.99
Proposed	**13.74**	16.61	$\mathbf{0.53 \pm 0.03}$	0.41 ± 0.03	27.36	–	**27.36**

The baselines used are the method in (Xie et al., 2023), StarGANv2 (Choi et al., 2020), and SmoothGAN (Liu et al., 2021). Note that (Xie et al., 2023) proposed a separate system for the domain translation that uses its pre-trained multi-domain generative model to train a separate translation model (see Appendix F). However, since the aforementioned GAN inversion procedure is also applicable to their generative model as it extracts content and style, we use two versions of their system, namely, I-GAN (Gen) for the method based on GAN inversion and I-GAN (Tr) for the separate translation system proposed in (Xie et al., 2023).

Fig. 4 (a) and (b) show the result of translation from wild domain ($n = 3$) to dog domain ($n = 2$) using randomly sampled style components. The content $\widehat{\boldsymbol{c}}_i^{(3)}, i \in [3]$ extracted for samples in the wild domain is combined with randomly sampled styles $\boldsymbol{s}_j^{(2)}, j \in [2]$ in the dog domain to synthesize the translated images. Our translations in each row contain the same content (i.e., pose of wild) as the input source image, but different styles (i.e., dog species). However, I-GAN (Gen) seems to produce unrealistic samples in some cases (first row). Their style diversity also appears to be limited.

Fig. 5 shows results of guided-translation for all methods for all pairs of domains in the AFHQ domain. Content extracted from the images in the first column is combined with the style from the second column. One can see that the proposed method preserves the style information better than the baselines.

Further experiments on multi-domain translation are presented in Appendix H.2.

Table 2 shows that the image quality (see FID) and diversity (see style diversity) of the translated images are competitive or better than the baselines (see qualitative results in Fig. 9 and 10 of Appendix H.). One can also see that the training time (on a single Tesla V100 GPU) of proposed method is at least 22 and 69 hours shorter than the competitive baselines StarGANv2 and I-GAN (Tr), respectively.

7 Conclusion

We revisited the problem of content-style identification from unaligned multi-domain data, which is a key step for provable domain translation and data generation. We offered a LDM perspective. This new viewpoint enabled us to prove identifiability results that enjoy considerably more relaxed conditions compared to those in previous research. Most importantly, we proved that content and style can be identified without knowing the exact dimension of the latent components. To our knowledge, this stands as the first dimension-agnostic identifiability result for content-style learning. We showed that the LDM formulation is equivalent to a latent domain-coupled multi-domain GAN loss, and the latter features a simpler implementation in practice. We validated our theorems using image translation and generation tasks.

Limitations. Our work focused on sufficient conditions for content-style identifiability, yet the necessary conditions were not fully understood—which is also of great interest. Additionally, our model considers that the domains are in the range of the same generating function. The applicability is limited to homogeneous multi-domain data, e.g., images with the same resolution. An interesting extension is to consider heterogeneous multi-domain models that can deal with very different types of data (e.g., text and audio). Additionally, our work is also limited to continuous data modalities like images, audio, etc. Discrete data modalities like text will require extension of both theory and implementation. This challenge presents another important future work. Finally, another limitation of our work is that the proposed method is based on the GAN framework which is known to be unstable during training. Therefore, novel implementation methods based on more stable distribution matching modules such as flow matching are also of interest as a future work.

ACKNOWLEDGMENT

This work was supported in part by the National Science Foundation (NSF) CAREER Award ECCS-2144889, and in part by Army Research Office (ARO) under Project ARO W911NF-21-1-0227.

REFERENCES

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *Proc. International Conference on Machine Learning (ICML)*, pp. 214–223. PMLR, 2017.

Neal L Carothers. *Real analysis*. Cambridge University Press, 2000.

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 *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8789–8797, 2018.

Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. StarGAN v2: Diverse image synthesis for multiple domains. In *Proc. IEEE Computer Vision and Pattern Recognition (CVPR)*, pp. 8188–8197, 2020.

Pierre Comon. Independent component analysis, a new concept? *Signal processing*, 36(3):287–314, 1994.

Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999.

Gang Dai, Yifan Zhang, Qingfeng Wang, Qing Du, Zhuliang Yu, Zhuoman Liu, and Shuangping Huang. Disentangling writer and character styles for handwriting generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5977–5986, 2023.

George Darmois. Analyse des liaisons de probabilité. In *Proc. International Statistic Conferences*, pp. 231, 1951.

Imant Daunhawer, Alice Bizeul, Emanuele Palumbo, Alexander Marx, and Julia E Vogt. Identifiability results for multimodal contrastive learning. *arXiv preprint arXiv:2303.09166*, 2023.

Cian Eastwood, Julius von Kügelgen, Linus Ericsson, Diane Bouchacourt, Pascal Vincent, Bernhard Schölkopf, and Mark Ibrahim. Self-supervised disentanglement by leveraging structure in data augmentations. *arXiv preprint arXiv:2311.08815*, 2023.

Ryszard Engelking. *Dimension theory*, volume 19. North-Holland Publishing Company Amsterdam, 1978.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2014.

Luigi Gresele, Paul K Rubenstein, Arash Mehrjou, Francesco Locatello, and Bernhard Schölkopf. The incomplete Rosetta stone problem: Identifiability results for multi-view nonlinear ICA. In *Proc. of Uncertainty in Artificial Intelligence (UAI)*, pp. 217–227, 2020.

Arthur Gretton, Kenji Fukumizu, Choon Teo, Le Song, Bernhard Schölkopf, and Alex Smola. A kernel statistical test of independence. *Advances in neural information processing systems (NeuIPS)*, 20, 2007.

Ishaan Gulrajani and Tatsunori Hashimoto. Identifiability conditions for domain adaptation. In *Proc. International Conference on Machine Learning (ICML)*, pp. 7982–7997, 2022.

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. *Advances in Neural Information Processing Systems (NeurIPS)*, 30, 2017.

Ziming Hong, Zhenyi Wang, Li Shen, Yu Yao, Zhuo Huang, Shiming Chen, Chuanwu Yang, Mingming Gong, and Tongliang Liu. Improving non-transferable representation learning by harnessing content and style. In *The Twelfth International Conference on Learning Representations*, 2024.

Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *Proc. of European Conference on Computer Vision (ECCV)*, pp. 172–189, 2018.

Zhuo Huang, Xiaobo Xia, Li Shen, Bo Han, Mingming Gong, Chen Gong, and Tongliang Liu. Harnessing out-of-distribution examples via augmenting content and style. *arXiv preprint arXiv:2207.03162*, 2022.

Aapo Hyvarinen and Hiroshi Morioka. Nonlinear ICA of temporally dependent stationary sources. In *Proc. Artificial Intelligence and Statistics (AISTATS)*, pp. 460–469, 2017.

Aapo Hyvärinen and Petteri Pajunen. Nonlinear independent component analysis: Existence and uniqueness results. *Neural networks*, 12(3):429–439, 1999.

Aapo Hyvarinen, Hiroaki Sasaki, and Richard Turner. Nonlinear ica using auxiliary variables and generalized contrastive learning. In *Proceedings of International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 859–868. PMLR, 2019.

Mohamed Salah Ibrahim, Ahmed S Zamzam, Aritra Konar, and Nicholas D Sidiropoulos. Cell-edge detection via selective cooperation and generalized canonical correlation. *IEEE Transactions on Wireless Communications*, 20(11):7431–7444, 2021.

Paris A. Karakasis and Nicholas D. Sidiropoulos. Revisiting deep generalized canonical correlation analysis. *IEEE Transactions on Signal Processing*, 71:4392–4406, 2023. doi: 10.1109/TSP.2023.3333212.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *Proc. International Conference on Learning Representations (ICLR)*, 2018.

Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:12104–12114, 2020.

Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. Variational autoencoders and nonlinear ica: A unifying framework. In *Proc. International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 2207–2217. PMLR, 2020.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proc. International Conference on Learning Representations (ICLR)*, 2015.

Lingjing Kong, Shaoan Xie, Weiran Yao, Yujia Zheng, Guangyi Chen, Petar Stojanov, Victor Akinwande, and Kun Zhang. Partial disentanglement for domain adaptation. In *Proc. International Conference on Machine Learning (ICML)*, pp. 11455–11472, 2022.

Hsin-Ying Lee, Hung-Yu Tseng, Qi Mao, Jia-Bin Huang, Yu-Ding Lu, Maneesh Singh, and Ming-Hsuan Yang. DRIT++: Diverse image-to-image translation via disentangled representations. *International Journal of Computer Vision*, 128:2402–2417, 2020.

John M Lee and John M Lee. *Smooth manifolds*. Springer, 2012.

Yahui Liu, Enver Sangineto, Yajing Chen, Linchao Bao, Haoxian Zhang, Nicu Sebe, Bruno Lepri, Wei Wang, and Marco De Nadai. Smoothing the disentangled latent style space for unsupervised image-to-image translation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10785–10794, 2021.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proc. International Conference on Computer Vision (ICCV)*, December 2015.

Qi Lyu, Xiao Fu, Weiran Wang, and Songtao Lu. Understanding latent correlation-based multiview learning and self-supervision: An identifiability perspective. In *Proc. International Conference on Learning Representations (ICLR)*, 2022.

Walter Rudin. *Principles of mathematical analysis*, volume 3. McGraw-hill New York, 1976.

Mohamad Shahbazi, Martin Danelljan, Danda Pani Paudel, and Luc Van Gool. Collapse by conditioning: Training class-conditional gans with limited data. In *Proc. International Conference on Learning Representations (ICLR)*, 2021.

Sagar Shrestha and Xiao Fu. Towards identifiable unsupervised domain translation: A diversified distribution matching approach. In *Proc. International Conference on Learning Representations (ICLR)*, 2024.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Mikael Sørensen, Charilaos I Kanatsoulis, and Nicholas D Sidiropoulos. Generalized canonical correlation analysis: A subspace intersection approach. *IEEE Transactions on Signal Processing*, 69:2452–2467, 2021.

Nils Sturma, Chandler Squires, Mathias Drton, and Caroline Uhler. Unpaired multi-domain causal representation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, 2023.

Subash Timilsina, Sagar Shrestha, and Xiao Fu. Identifiable shared component analysis from unaligned multimodal mixtures. *https://arxiv.org/abs/2409.19422*, 2024.

Julius Von Kügelgen, Yash Sharma, Luigi Gresele, Wieland Brendel, Bernhard Schölkopf, Michel Besserve, and Francesco Locatello. Self-supervised learning with data augmentations provably isolates content from style. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 16451–16467, 2021.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *Proc. International Conference on Machine Learning (ICML)*, pp. 9929–9939. PMLR, 2020.

Weiran Wang, Xinchen Yan, Honglak Lee, and Karen Livescu. Deep variational canonical correlation analysis. *arXiv preprint arXiv:1610.03454*, 2016.

Wayne Wu, Kaidi Cao, Cheng Li, Chen Qian, and Chen Change Loy. TransGaGa: Geometry-aware unsupervised image-to-image translation. In *Proc. IEEE Computer Vision and Pattern Recognition (CVPR)*, pp. 8012–8021, 2019.

Weihao Xia, Yulun Zhang, Yujiu Yang, Jing-Hao Xue, Bolei Zhou, and Ming-Hsuan Yang. GAN inversion: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3): 3121–3138, 2022.

Shaoan Xie, Lingjing Kong, Mingming Gong, and Kun Zhang. Multi-domain image generation and translation with identifiability guarantees. In *Proc. International Conference on Learning Representations (ICLR)*, 2023.

Heran Yang, Jian Sun, Aaron Carass, Can Zhao, Junghoon Lee, Jerry L Prince, and Zongben Xu. Unsupervised MR-to-CT synthesis using structure-constrained CycleGAN. *IEEE Transactions on Medical Imaging*, 39(12):4249–4261, 2020.

Shuai Yang, Liming Jiang, Ziwei Liu, and Chen Change Loy. GP-UNIT: Generative prior for versatile unsupervised image-to-image translation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

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) (CVPR)*, pp. 586–595, 2018.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proc. IEEE Computer Vision and Pattern Recognition (CVPR)*, pp. 2223–2232, 2017.

Roland S Zimmermann, Yash Sharma, Steffen Schneider, Matthias Bethge, and Wieland Brendel. Contrastive learning inverts the data generating process. In *Proc. International Conference on Machine Learning (ICML)*, pp. 12979–12990, 2021.