Vol.2025-MUS-142 No.3
2025/3/6

超球面上の確率的表現学習を用いたマルチモーダル音楽情報検索

中塚 貴之[1,a)]　濱崎 雅弘[1,b)]　後藤 真孝[1,c)]

概要：本稿では，音楽情報（楽曲，画像，テキスト）について，そのうちの一つ，もしくは複数の組み合わせをクエリとして，そのクエリに合った音楽情報を検索するマルチモーダル音楽情報検索のための，確率的表現学習に基づく手法について述べる．具体的には，音楽情報の各データについて，本来のペアの確率分布が互いに近づき，ランダムなペアの確率分布が互いに遠ざかるように，それぞれのデータのモダリティに対応した三つのエンコーダ（楽曲エンコーダ，画像エンコーダ，テキストエンコーダ）を学習する．このような学習を実現するために，我々は確率的対照学習と超球面上の最適輸送（Spherical Sliced-Wasserstein; SSW）の組み合わせに基づく新たな損失関数を設計した．マルチモーダル音楽情報検索について，提案手法の有効性をベンチマークデータセットおよびプライベートデータセットを用いた定量的実験によって示した．またプライベートデータセットを用いた定性分析を実施し，同じ音楽カテゴリタグが付された音楽情報が超球面上において近くに配置されていることを確認した．

1. はじめに

楽曲と映像 [43]，楽曲とテキスト [16] といった，音楽情報についてのマルチモーダル表現学習は，音楽情報検索において重要な技術である．これまでの表現学習に基づく音楽情報検索の手法は，学習済みのエンコーダを用いて音楽情報の各データを高次元潜在空間における単一のベクトルとして表現し，そのベクトルを利用した検索を行う決定論的なアプローチを取ってきた．しかし，音楽情報の各データを単一のベクトルで表現することには，いくつか課題がある．例えば，ジャケット画像と複数トラックの楽曲の組み合わせといった一対多の関係や，異なるアーティストが歌う異なる楽曲だが，楽曲タイトルが同じといった紛らわしい関係が存在しており，そういった関係について単一のベクトルで表現することは困難である．これらの課題を解決するアプローチとして，それぞれのデータを単一のベクトルではなく確率分布として表現する確率的表現学習が注目されている [7, 26, 30]．

確率的表現学習（図 1）は，それぞれのデータについて，複雑で多様な表現が可能な高次元潜在空間における確率分布として表現する方法である．この方法を実現するために

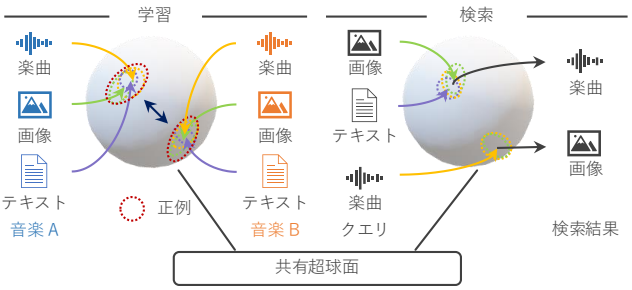

図 1　超球面上における確率的表現学習．（**左図**）共有超球面上において，正例となる楽曲，画像，テキストの本来のペアの確率分布が近くに配置され，負例となる無関係なペアの確率分布が遠くに配置されるように，それらのモダリティに対応したエンコーダをそれぞれ学習する．（**右図**）このように学習したエンコーダを，マルチモーダル音楽情報検索に活用する．音楽情報について，そのうちの一つ（例えば，楽曲），もしくは複数の組み合わせ（例えば，画像とテキスト）をクエリとして，それぞれのクエリに合った画像や楽曲を，確率分布間の距離に基づいて検索することができる．

は，データ一つひとつについて，最適な確率分布を推定できるようにエンコーダを学習する必要がある．このとき重要となるのが，エンコーダの学習に用いる損失関数を適切に設計することである．確率的表現学習における損失関数の設計には大きく三つのアプローチがあるが，本稿では新たに四つ目のアプローチ [29] を提案する．

一つ目は，Probability Product Kernel [18] を用いて，二つの分布間の類似度を計算するアプローチである．これは，確率的単語埋め込み [47]，顔認識 [41]，画像分類 [23] といっ

[1] 産業技術総合研究所
National Institute of Advanced Industrial Science and Technology (AIST)
a) takayuki.nakatsuka@aist.go.jp
b) masahiro.hamasaki@aist.go.jp
c) m.goto@aist.go.jp

たタスクで活用されている．二つ目は，Hedged Instance Embeddings（HIB）[31] を用いたアプローチである．このアプローチは，それぞれの分布からランダムに取得したサンプル点（ベクトル）の距離（match probability）を損失関数の計算に用いており，テキストと画像のクロスモーダル検索 [7,26] や，映像の自己教師有り表現学習 [33] といったタスクで活用されている．三つ目は，決定論的なアプローチで使用される損失関数におけるベクトル変数を，単に確率分布で置き換えるアプローチである．例えば，Gauss 分布 [5,30,38] や von Mises-Fisher（vMF）分布 [22,27,39] といったパラメトリックな分布から，reparameterization trick [9,21] を用いて取得したサンプル点の距離に基づいて，分布間の類似度を計算している．

これらのアプローチは，テキストと画像のクロスモーダル検索 [7,26] やマルチモーダル画像検索 [30] といった検索タスクに活用されている．Chun ら [7] は，確率的表現学習を初めてクロスモーダル検索に応用した Probabilistic Cross-Modal Embedding（PCME）を提案した．Li ら [26] は，データセットに含まれるテキストと画像の意味的相関スコアを計算する Average Semantic Precision（ASP）と，その ASP を最適化する Differentiable ASP Approximation を提案した．Neculai ら [30] は，テキストと画像を組み合わせたクエリを用いた画像検索手法である Multimodal Probabilistic Composer（MPC）を提案した．しかしこれらの手法は，それぞれの分布からランダムに取得されたサンプル点の組み合わせについて距離（類似度）を計算するアプローチであり，分布間の距離（類似度）の計算として必ずしも最適なアプローチではない（図 2 における左図）．そのため，これらのアプローチでは分布の特性を失う可能性があり，結果的に検索タスクにおける性能の低下に繋がることが考えられる．

本研究ではこの問題に取り組むために，マルチモーダル音楽情報検索を対象として，対照学習と最適輸送のそれぞれに基づく新しい二つの損失関数を提案し，それらを組み合わせた損失関数をエンコーダの学習に活用する．対照学習は，複数のモダリティのデータについて，それらの潜在表現が同一の高次元潜在空間を共有するように，それぞれのモダリティに対応したエンコーダを共同で学習する手法である [37,49]．この対照学習について，確率的表現学習の文脈では，分布間のユークリッド距離ではなく，分布間の角距離に基づいた損失関数を利用することが，検索タスクにおける性能を向上するために重要であると示されている [39]．加えて，単一のスカラー値で分布の分散を表す vMF 分布は，Gauss 分布に比べて分布の分散推定が簡潔になり，検索タスクにおける性能が向上するといった利点がある [27]．これらの知見に基づき，我々はそれぞれのデータの潜在表現として vMF 分布（超球面上の確率分布）を採用し，Probabilistic Contrastive Learning [22] を複数のモダリティのデータに拡張したマルチモーダル確率的対照損失関数を提案する．また，最適輸送 [46] は分布間の距離の計算方法として頑健でかつ有用である（図 2 における右図）．したがって，高次元潜在空間において正例のペアの分布を近づけるエンコーダの学習において最適輸送を取り入れることで，より正確な表現学習が可能になる．そこで我々は，このような最適輸送の利点をマルチモーダル音楽情報検索タスクにおいて活用するため，超球面上の分布の最適輸送を高速に計算可能な Spherical Sliced-Wasserstein（SSW）p -距離 [4] に基づく損失関数を提案する．

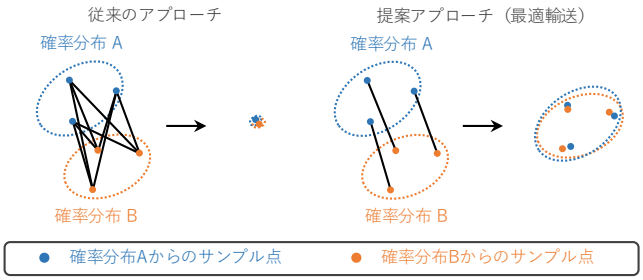

図 2 最適輸送の利点．（**左図**）従来のアプローチ（ [7,22,30] など）は確率分布 A と B の形状を一致させるために，それぞれの分布から得られたサンプル点についてランダムな組み合わせの距離を計算している．つまり，これらのアプローチは部分最適化を計算しているため，必ずしも二つの分布が一致するとは限らない．さらに，部分最適化の結果として，ある一点にサンプル点が集まってしまう可能性があり，その場合は分布としての性質が失われてしまう．（**右図**）最適輸送は，分布の形状を一致させるために，最適なサンプル点の組み合わせを探索する全体最適化を計算している．そのため提案アプローチでは，分布としての性質を損なわずに，分布を一致させることができる．

提案する損失関数を用いることで，図 1 に示すように，音楽情報の各データを vMF 分布として表現することが可能なエンコーダを学習することができる．エンコーダの学習時には，楽曲，その楽曲の代表画像（ジャケット画像，サムネイル画像など），その楽曲のメタデータから生成されたテキストで構成できるペア（つまり，本来のペア）を正例とし，無関係なペア（異なる曲，異なる音楽ジャンル，異なるアーティストなど）を負例として扱う．そして学習済みのエンコーダを用いて，音楽情報の各データを vMF 分布として表現することで，分布間の距離に基づくマルチモーダル音楽情報検索が実現できる．

こうして音楽情報の各データを潜在空間内における確率分布で表現すると，それらを複数組み合わせたクエリで検索したいときには単にそれらの確率分布を混合した分布を用いて検索できるため，確率的表現はマルチモーダル音楽情報検索にとって大きな利点となる．ベンチマークデータセットである YT8M-MusicVideo データセットとプライベートデータセットである AS5M データセットを用いて定量的評価と定性的分析を行い，提案手法の有効性を検証する．

2. 数学的背景

2.1 タスクの定式化

本研究では，楽曲エンコーダ，画像エンコーダ，テキストエンコーダのそれぞれの入力として，楽曲のメルスペクトログラム，RGB 画像，トークン化されたテキストを用いる．これらの入力は，先行研究 [37, 49] に基づいて決定した．

$\mathbf{A} = \{\mathbf{a}_n \in \mathbb{R}^{D^{\mathbf{a}}}\}_{n=1}^{N}$ をスペクトログラムの集合，$\mathbf{I} = \{\mathbf{i}_n \in \mathbb{R}^{D^{\mathbf{i}}}\}_{n=1}^{N}$ を画像の集合，$\mathbf{T} = \{\mathbf{t}_n \in \mathbb{R}^{D^{\mathbf{t}}}\}_{n=1}^{N}$ をトークン化されたテキストの集合とする．$\mathbf{a}, \mathbf{i}, \mathbf{t}$ のそれぞれに付された添字は対応しており，楽曲 $\mathbf{a}_n$ と添字 n が等しい場合，$\mathbf{i}_n$ はその楽曲の代表画像，$\mathbf{t}_n$ はその楽曲のメタデータから生成されたテキストのトークンを表している．また，$D^{\mathbf{a}}$ はスペクトログラムの次元数，$D^{\mathbf{i}}$ は画像の次元数，$D^{\mathbf{t}}$ はトークン化されたテキストの次元数である．N はデータセットに含まれる楽曲数である．

次に，$\mathbf{Z^A} = \{\mathbf{z}_n^{\mathbf{a}} \in \mathbb{R}^d\}_{n=1}^{N}$ をスペクトログラムの潜在変数の集合，$\mathbf{Z^I} = \{\mathbf{z}_n^{\mathbf{i}} \in \mathbb{R}^d\}_{n=1}^{N}$ を画像の潜在変数の集合，$\mathbf{Z^T} = \{\mathbf{z}_n^{\mathbf{t}} \in \mathbb{R}^d\}_{n=1}^{N}$ をトークン化されたテキストの潜在変数の集合とする．ここで，d はそれぞれの潜在変数の次元数を表す．

我々は，確率分布 $p(\mathbf{z}_n^{\mathbf{a}}|\mathbf{a}_n)$, $p(\mathbf{z}_n^{\mathbf{i}}|\mathbf{i}_n)$, $p(\mathbf{z}_n^{\mathbf{t}}|\mathbf{t}_n)$ が，共有超球面 $S_{\text{shared}}^{d-1} = \{||\mathbf{z}_n|| = 1\}$ 上において互いに近づくように，$\mathbf{A}$ を $\mathbf{Z^A}$ として表現する楽曲エンコーダ $f_{\mathbf{A}}$，$\mathbf{I}$ を $\mathbf{Z^I}$ として表現する画像エンコーダ $f_{\mathbf{I}}$，$\mathbf{T}$ を $\mathbf{Z^T}$ として表現するテキストエンコーダ $f_{\mathbf{T}}$ の三つを同時に学習する．

2.2 確率的対照学習

対照学習は確立された深層学習手法の一つであり，その手法は近年の様々な研究において幅広く利用されている [25]．特に，N-pairs loss [42], InfoNCE loss [32], MoCo [14] といった，N 個のペア（例えば，1 個の正例と $N-1$ 個の負例）に基づいて損失関数を計算する手法は，マルチモーダル表現学習において有用な手法として採用されている [16, 37, 43, 49]．しかし，これらの対照損失関数は，単一のベクトルを入力とした利用を想定して設計されており，確率分布を直接扱うことは考慮されていない．

近年，Kirchhoff ら [22] は，確率分布に対し対照学習を実施できるように，InfoNCE を再設計した損失関数である MCInfoNCE を提案した．MCInfoNCE $\mathcal{L}_{MC}$ は次式で定義される．

$$\mathcal{L}_{MC} = -\frac{1}{m}\sum_{j=1}^{m}\sum_{l=1}^{L}\log\frac{e^{\text{sim}(\mathbf{z}_j^l,\mathbf{z}_+^l)/\tau}}{e^{\text{sim}(\mathbf{z}_j^l,\mathbf{z}_+^l)/\tau}+\sum_{\mathbf{z}_-}e^{\text{sim}(\mathbf{z}_j^l,\mathbf{z}_-^l)/\tau}}, \tag{1}$$

ここで，m はミニバッチサイズ，L はそれぞれの確率分布から取得するサンプル点の数，τ は損失関数のスケールを制御する，温度スケーリングと呼ばれるハイパーパラメータである．また，$\mathbf{z}_n \sim p(\mathbf{z}_n)$ を基準としたときに，$\mathbf{z}_+ \sim p(\mathbf{z}_+|\mathbf{z}_n)$ は正例となるデータの確率分布から取得したサンプル点，$\mathbf{z}_- \sim p(\mathbf{z}_-|\mathbf{z}_n)$ は負例となるデータの確率分布から取得したサンプル点である．そして，$\text{sim}(\cdot,\cdot)$ は cosine 類似度といった類似度を計算する関数である．MCInfoNCE は，それぞれの確率分布に対し Monte-Carlo サンプリングを適用することでサンプル点を取得し，そのサンプル点を損失関数の計算に利用している．ただし，MCInfoNCE は単一のモダリティのデータを対象として設計された損失関数である．そのため提案手法では，その対象を複数のモダリティのデータに拡張し，MCInfoNCE を再設計する（詳細を 3.1 節で述べる）．

2.3 最適輸送

最適輸送を用いた応用研究は，コンピュータビジョン分野 [2, 12, 13] で近年注目されている．しかし，分布間の最適輸送距離を直接計算することは，計算コストが非常に高いことが知られている [24]．この問題を解決するアプローチとして，確率分布を特定の多様体（例えば，平面や円周）に射影し，その多様体上で最適輸送距離を計算する方法がある [4, 24]．

本研究では，最適輸送の計算手法である Spherical Sliced-Wasserstein（SSW）p -距離 [4] を用いる．SSW p -距離は，超球面上の確率分布間の最適輸送距離を高効率に計算できる手法であり，我々はこの手法を確率的表現学習において先駆的に利用する．

2.3.1 Spherical Sliced-Wasserstein の定義

SSW p -距離（$p \geq 1$）は，有限な p 次モーメントを持つ，超球面 S^{d-1} 上で絶対連続である二つの確率測度 $\mu, \nu \in \mathcal{P}_{p,ac}(S^{d-1})$ について次式で定義される．

$$SSW_p(\mu,\nu) = \int_{\mathbb{V}_{d,2}} W_p\left(\mu \circ P^{U^{-1}}, \nu \circ P^{U^{-1}}\right) \, \mathrm{d}\sigma, \tag{2}$$

ここで，$\mathbb{V}_{d,2} = \{U \in \mathbb{R}^{d\times 2},\ U^\top U = I_2\}$ は Stiefel 多様体 [3], σ は $\mathbb{V}_{d,2}$ 上の一様分布, P^U は S^{d-1} 上の点 $\mathbf{z} \in S^{d-1}$ を U によって導かれる S^{d-1} の大円 S^1 上に射影する関数, W_p は S^1 上の最適輸送距離 [10, 36] である．P^U について，ほとんど至るところの $\mathbf{z} \in S^{d-1}$ で, $P^U(\mathbf{z}) = U^\top\mathbf{z}/||U^\top\mathbf{z}||_2$ と計算ができる [4]．また U の選択による影響を受けないようにするために，Bonet ら [4] はランダムに生成した複数の U のそれぞれについて SSW p -距離を計算し，それらの平均値を最適輸送距離として用いている．本研究でも，Bonet らと同様に最適輸送距離を計算する．

2.3.2 大円上の最適輸送距離

ここでは，SSW p -距離として最も簡単な $p = 1$ の場合を考える．これは，超球面上の確率分布を大円上に射影し，その大円上における最適輸送距離を計算することに相当す

る．超球面 S^{d-1} からある大円上に射影された二つの確率測度 $\mu', \nu' \in \mathcal{P}(S^1)$ の間の最適輸送距離 W_1 は次式で定義される．

$$W_1(\mu', \nu') = \int_0^1 |F_{\mu'}(t) - F_{\nu'}(t) - \mathrm{LevMed}(F_{\mu'} - F_{\nu'})| \, \mathrm{d}t, \tag{3}$$

ここで，$F_{\mu'}, F_{\nu'}$ はそれぞれ μ', ν' の累積分布関数，LevMed$(\cdot)$ は次式で定義される level median 関数 [17] である．

$$\mathrm{LevMed}(f) = \min \left\{ \arg\min_{\alpha \in \mathbb{R}} \int_0^1 |f(t) - \alpha| \mathrm{d}t \right\}, \tag{4}$$

ただし，α は shift パラメータである．式 (2)，式 (3)，式 (4) によって，超球面上の確率分布間の最適輸送である SSW_1 を計算できる．特筆すべき点は，$p = 1$ の場合に限り，式 (3) における積分は，単に S^1 上に射影したサンプル点をソートするだけで計算ができるということである．つまり，図 2 における右図のように，最適輸送の計算に必要であるサンプル点の最適な組み合わせは，全ての組み合わせを調べることなく，ソートのみで算出することができるため，高速に計算することができる．提案手法では，この SSW_1 を損失関数として用いる．本稿の付録に，SSW_1 の計算手順と疑似コードを記載する．

3. 提案手法

本研究では，マルチモーダル音楽情報検索を対象として，マルチモーダル確率的対照損失関数（3.1 節）と最適輸送の計算手法である SSW に基づく損失関数（3.2 節）の二つの新しい損失関数を提案する [29]．これらの二つの損失関数は，エンコーダの学習においてそれぞれ異なる働きをする．マルチモーダル確率的対照損失関数は，$S^{d-1}_{\mathrm{shared}}$ 上の無関係なペアの確率分布が遠くに配置されるように働きかけることで，結果的に正例のペアの確率分布が近くに配置される．一方，SSW に基づく損失関数は，直接，正例のペアの確率分布が近くに配置されるように働きかける（無関係なペアの確率分布は扱わない）．そのため，これらの異なる働きをする二つの損失関数を組み合わせることで，それぞれの損失関数の利点をエンコーダの学習に活かすことができる．学習済みのエンコーダをマルチモーダル音楽情報検索に応用する方法は，3.3 節で述べる．

確率的表現学習における標準的なアプローチは，潜在空間における確率分布について，Gauss 分布 [5, 30, 38] や vMF 分布 [22, 27, 39] といったパラメトリックな分布を仮定する方法である．提案手法では先行研究に倣い，音楽情報の各データについて，次式のような $S^{d-1}_{\mathrm{shared}}$ 上の vMF 分布を仮定する．

$$p(\mathbf{z}^{\mathbf{a}}_n | \mathbf{a}_n) = \mathrm{vMF}(\mathbf{z}^{\mathbf{a}}_n; \mu(\mathbf{a}_n), \kappa(\mathbf{a}_n)), \tag{5}$$

$$p(\mathbf{z}^{\mathbf{i}}_n | \mathbf{i}_n) = \mathrm{vMF}(\mathbf{z}^{\mathbf{i}}_n; \mu(\mathbf{i}_n), \kappa(\mathbf{i}_n)), \tag{6}$$

$$p(\mathbf{z}^{\mathbf{t}}_n | \mathbf{t}_n) = \mathrm{vMF}(\mathbf{z}^{\mathbf{t}}_n; \mu(\mathbf{t}_n), \kappa(\mathbf{t}_n)), \tag{7}$$

ここでは，2.1 節において定義した変数を用いている．このアプローチでは，それぞれの vMF 分布について，中心方向を定める単位ベクトル $\mu(\cdot)$ と，その方向への分布の集中度を示すパラメータである $\kappa(\cdot)$ を適切に推定できるエンコーダを学習する必要がある．

また学習時には，それぞれの分布から rejection-sampling reparameterization trick [9] を用いて L 個のサンプル点を取得し，損失関数の計算に利用する．提案する損失関数（3.1 節および 3.2 節）で用いる表記には次を使用する．

$$\zeta_n \sim \mathrm{vMF}(\mathbf{z}^{*}_n; \mu(*_n), \kappa(*_n)), \tag{8}$$

$$\eta_n \sim \mathrm{vMF}(\mathbf{z}^{\star}_n; \mu(\star_n), \kappa(\star_n)), \tag{9}$$

ここで，ζ_n，η_n（$*, \star \in \{\mathbf{a}, \mathbf{i}, \mathbf{t}\}, * \neq \star$）は，それぞれ vMF 分布から取得した L 個のサンプル点である．

3.1 マルチモーダル確率的対照損失関数

対照学習は複数のモダリティのデータについて，潜在空間で共通の表現が可能な手法である [16, 37, 43, 49]．そこで我々は，単一のモダリティのデータを対象とした先行研究である MCInfoNCE（式 (1)）を拡張し，複数のモダリティのデータについて全ての組み合わせを考慮したマルチモーダル確率的対照損失関数 $\mathcal{L}_C$ を新たに設計する．提案するマルチモーダル確率的対照損失関数 $\mathcal{L}_C$ は次式で定義される．

$$\mathcal{L}_C = -\frac{1}{m} \sum_{<\zeta, \eta>} \sum_{j=1}^{m} \log \frac{e^{\mathrm{sim}(\zeta_j, \eta_+)/\tau}}{\sum_{k=1}^{m} e^{\mathrm{sim}(\zeta_j, \eta_k)/\tau}}, \tag{10}$$

ここで，m はミニバッチの大きさ，τ は温度スケーリングである．また，ζ_j を基準としたときに，η_+ は正例となるデータの確率分布である．そして $\mathrm{sim}(\cdot, \cdot)$ は，二つの分布のそれぞれから取得した L 個のサンプル点を用いて，それらの分布の類似度を計算する関数であり，次式で定義される．

$$\begin{aligned} \mathrm{sim}(\zeta_j, \eta_k) &\simeq \mathrm{sim}\left(\left\{ \mathbf{z}^{*,l}_j \right\}_{l=1}^{L}, \left\{ \mathbf{z}^{\star,l}_k \right\}_{l=1}^{L} \right) \\ &= \frac{1}{L} \sum_{l=1}^{L} \frac{{\mathbf{z}^{*,l}_j}^{\top} \mathbf{z}^{\star,l}_k}{||\mathbf{z}^{*,l}_j|| ||\mathbf{z}^{\star,l}_k||}. \end{aligned} \tag{11}$$

このマルチモーダル確率的対照損失関数 $\mathcal{L}_C$ は，無関係なペアの分布が遠くに配置されるように働きかける．

3.2 Spherical Sliced-Wasserstein に基づく損失関数

式 (2)，式 (3)，式 (4) より導出できる SSW_1 を用い

て，SSW に基づく損失関数 $\mathcal{L}_S$ を次のように定式化する．

$$\mathcal{L}_S = \frac{1}{m} \sum_{<\zeta,\eta>} \sum_{j=1}^{m} SSW_1(\zeta_j, \eta_j). \tag{12}$$

この式 (12) は，S^{d-1}_{shared} 上にある正例のペア ζ_j，η_j について，まず，それぞれから取得した L 個のサンプル点を S^1 に射影し，それらの射影した点を対応づけるために S^1 上でソートし，S^1 上の最適輸送距離 W_1 を計算することで，SSW_1 を求めている．この SSW に基づく損失関数 $\mathcal{L}_S$ は，正例のペアの分布を近くに配置するように働きかける．

提案する二つの損失関数 $\mathcal{L}_C$，$\mathcal{L}_S$ の双方の利点を活かすために，次式で定義する損失関数 $\mathcal{L}$ をエンコーダの学習に用いる．

$$\mathcal{L} = \mathcal{L}_C + \lambda_S \mathcal{L}_S, \tag{13}$$

ここで，λ_S は重みである．

3.3 確率的表現を用いたマルチモーダル音楽情報検索

学習済みのエンコーダを用いることで，音楽情報の各データを S^{d-1}_{shared} 上における確率分布として表現することが可能になる．これを利用して，S^{d-1}_{shared} 上の確率分布間の距離に基づくマルチモーダル音楽情報検索を行うことができる．

まず音楽情報（楽曲，画像，テキスト）について，そのうちの一つをクエリとして用いる場合，そのクエリの確率分布から得られるサンプル点と，検索対象となるデータセット内の各データについての確率分布から得られるサンプル点のそれぞれについて Fréchet 平均を計算する．また提案手法では，潜在表現として確率分布を用いているため，複数のクエリについてそれぞれの確率分布を混合することで，あたかも一つのクエリであるかのように利用できる．つまり，音楽情報について複数の組み合わせをクエリとして用いることが可能であり，その場合も複数のモダリティのデータの混合分布から得られるサンプル点と，検索対象となるデータセット内の各データについての確率分布から得られるサンプル点のそれぞれについて Fréchet 平均を計算する．そしてマルチモーダル音楽情報検索を行う際には，それらの Fréchet 平均について cosine 類似度を計算する．このとき，ある確率分布のペアの類似度が高い場合，それらの確率分布を潜在表現として持つ，クエリと検索対象のそれぞれの音楽情報のペアが合っていることを示している．そのため，得られた類似度について降順に並び替え，類似度が上位の潜在表現を持つ音楽情報を検索結果として提示する．

4. 実験と結果

本章では，まず，確率的表現学習によって得られた確率分布について，本来のペアが S^{d-1}_{shared} においてどの程度近くに配置されているかを定量的に評価するための比較実験について述べる．次に，提案手法について定性的な分析を行うため，音楽情報の各データについて，学習された表現の性質について調べる．最後に，提案手法を用いて学習したエンコーダを利用して，音楽情報をクエリとした際の検索結果例を示す．

4.1 実験設定

4.1.1 データセット

実験では，特性の異なる二つのデータセットを用いた．テスト用のデータセットの大きさは，先行研究 [35, 43] の設定に従った．

YT8M-MusicVideo データセット [43] は，YouTube-8M データセット [1] から，“music video” とタグ付けされた動画を対象として抽出されたデータセットである．このデータセット内で入手可能な 60,785 の YouTube チャンネルから得られる 73,113 曲（一曲あたり平均 4 分程度，48 kHz のサンプリング周波数）と，そのサムネイル画像（縦横比 16:9 の RGB 画像），およびメタデータ（楽曲名，チャンネル名，アップロード年月日）を用いた．このデータセットをさらに，学習用データセット（64,001 曲），検証用データセット（7,112 曲），テスト用データセット (2,000 曲) にランダムに分割した．このとき，同一の YouTube チャンネルの楽曲が，分割された複数のデータセットにわたって含まれないようにした．評価のために，三つのランダムなシード値を与えて学習したエンコーダのそれぞれについて，定量評価指標に基づき検索性能を計測した．

AS5M データセットは，視聴用の楽曲（それぞれ 30 秒，44.1 kHz ののサンプリング周波数），それらの楽曲のジャケット画像，それらの楽曲のメタデータ（楽曲名，アーティスト名，アルバム名，音楽ジャンル，公開年月日）で構成される非公開のデータセットである．このデータセットには，174,629 組のアーティストによる 5,920,828 曲と 1,115,668 枚のジャケット画像が含まれている（つまり，各画像が平均しておよそ 5.3 曲に関連付いている）．このデータセットをさらに，学習用データセット，検証用データセット，テスト用データセットの比率がそれぞれ 8:1:1 になるようにランダムに分割した．このとき，同一のアーティストやジャケット画像が，分割された複数のデータセットにわたって含まれないようにした．評価のために，テスト用データセットからランダムに選択した 2,000 曲で構成される小データセットを 10 個作成し，学習したエンコーダをそれぞれの小データセットにおいて定量評価指標に基づき検索性能を計測した．

4.1.2 実装詳細

4.1.2.1 エンコーダの設計

楽曲エンコーダは，Contrastive Language-Audio Pre-training (CLAP) [49] の音声モデルをバックボーンネットワークとして用いた．また，画像エンコーダとテキスト

エンコーダは，Contrastive Language-Image Pretraining (CLIP) [37] の画像モデルと言語モデルのそれぞれをバックボーンネットワークとして用いた．エンコーダを学習する際には，まず，それぞれのバックボーンネットワークについて Transformers [48] で公開されている事前学習済みパラメータ（CLAP のパラメータとして “laion/clap-htsat-fused” を利用し，CLIP のパラメータとして “openai/vit_base_patch16_224” を利用した）を設定し，学習中はエンコーダの projection 層を更新した．

4.1.2.2 楽曲

各楽曲は，Transformers [48] で公開されている CLAP の特徴抽出器を用いてメルスペクトログラムへ変換し，それを楽曲エンコーダの入力として用いた．楽曲エンコーダの学習時には，データ拡張として，frequency masking と time masking [28]，および時間領域におけるランダムな切り出し [44] を行った．

4.1.2.3 画像

各 RGB 画像は，224 px × 224 px の大きさに変更し，それを画像エンコーダの入力として用いた．画像エンコーダの学習時には，データ拡張として，random resized crop（scale=[0.08, 1.0], ratio=[0.75, 1.33]），random horizontal flip（probability=0.5），および random erasing（probability=0.2） [50] を適用した．

4.1.2.4 テキスト

各メタデータについて，まず，keyword-to-caption [49] と呼ばれるデータ生成手法[*1]を用いて，テキストに変換し，得られたテキストをトークン化した．このとき，CLIP [37] における実験条件に合わせて，トークンの最大長は 77 とした．テキストエンコーダの学習時には，データ拡張として，keyword-to-caption を用いてメタデータからテキストを生成する際に，メタデータを dropout（probability=0.05） [40] した．

4.1.2.5 学習条件

それぞれの実験条件において NVIDIA A100 GPU を 16 台用いて計算を行った．このとき，それぞれの GPU で楽曲，画像，テキストを 64 組ずつ処理する分散学習を行った．実装には，PyTorch [34] を用いた．エンコーダの学習時には，Adam [20] を用い，その学習率は 1.0×10^{-4} と設定した．潜在変数の次元数は，MPC [30] の実験条件に合わせて，$d = 512$ に設定した．また，vMF 分布のパラメータである κ は， MCInfoNCE [22] の実験条件を参考にして，尖度の高い分布が得られるように $\kappa(\cdot) \in (64, 128)$ と設定した．各分布から取得するサンプル点の数は，経験的に $L = 16$ と設定した．マルチモーダル確率的対照損失関数 $\mathcal{L}_C$ について，温度スケーリング τ の値は，MoCo [14] で元々用いられていた $\tau = 0.07$ と設定した．また，SSW に基づく損失関数 $\mathcal{L}_S$ について，Bonet ら [4] の方法に倣い，ランダムな U について SSW_1 距離を 100 回繰り返し計算し，その平均値を用いた．4.4 節における分析に基づき，重み λ_S を 1.0 に設定した．

4.1.3 定量評価指標

検索タスクにおける標準的な評価指標である，平均逆順位（MRR） [8]，再現率（R@k），中央順位（MR） [43] を用いた比較実験を実施した．MRR は，それぞれのクエリについて検索結果の中から最初に正解のデータが見つかる順位を記録し，テストデータセットのすべてのクエリについてその順位の逆数を平均した値である．MRR の値が高いほど，より高い精度で正解のデータを検索できていることを示している．R@k は，それぞれのクエリについて検索結果の上位 k 番目までに正解のデータが含まれているか真偽を判定し，テストデータセットのすべてのクエリについてその真の割合を算出した値であり，表において百分率で表記する．R@k の値が高いほど，正解のデータを上位で検索できていることを示している．MR は，それぞれのクエリについて検索結果の中から最初に正解のデータが見つかる順位を記録し，テストデータセットのすべてのクエリについてその順位の中央値を算出した値である．MR の値が小さいほど，正解のデータが上位に集中していることを示している．

4.2 実験条件

提案手法（損失関数 $\mathcal{L}$）と，その比較手法として確率的表現学習を用いたテキストと画像のクロスモーダル検索手法である **PCME** [7] および **MPC** [30] のそれぞれについてその検索性能を評価した．提案手法は，確率的表現学習をマルチモーダル音楽情報検索に適用した先駆的な手法であるため，楽曲とその他のモダリティのデータ（画像，テキスト）の検索については，適切な比較手法が存在しない．そのため，3.1 節で定義したマルチモーダル確率的対照損失関数 $\mathcal{L}_C$ のみを用いた**ベースライン**を比較手法に加えた．

4.3 結果

表 1–3 は YT8M-MusicVideo データセット，表 4–6 は AS5M データセットについて，それぞれの手法の検索タスクにおける定量評価指標の結果を示している．これらの結果が示すように，提案手法は，YT8M-MusicVideo データセットと AS5M データセットの双方における全ての検索タスクにおいて，テキストと画像のクロスモーダル検索手法である PCME [7] および MPC [30] の検索性能を上回っ

[*1] 検索タスクにおいて，定型文とメタデータを用いてテキストを生成し利用することが，学習・検索の両方で効果的であることが知られている [37]．そのため実験では，YT8M-MusicVideo データセットにおいて，定型文として “title” is a music video uploaded by “channel name” on “upload date.” を用いた．また，AS5M データセットにおいて，定型文として “song title” is a(n) “music genre” song by “artist name”, released on “release date.” “song title” is collected to “collection name.” を用いた．

表 1 YT8M-MusicVideo データセットを用いたマルチモーダル画像検索タスクにおける性能比較.

	楽曲 → 画像			テキスト → 画像			楽曲 & テキスト → 画像		
手法	MRR ↑	R@1 (%) ↑	MR ↓	MRR ↑	R@1 (%) ↑	MR ↓	MRR ↑	R@1 (%) ↑	MR ↓
PCME	–	–	–	0.025 ± 0.003	0.73 ± 0.08	369	–	–	–
MPC	–	–	–	0.014 ± 0.001	0.2 ± 0.11	425	–	–	–
ベースライン	0.024 ± 0.001	$\mathbf{0.73 \pm 0.09}$	272	0.048 ± 0.001	1.92 ± 0.12	166	0.044 ± 0.001	1.55 ± 0.11	166
提案手法	$\mathbf{0.028 \pm 0.001}$	0.65 ± 0.08	**247**	$\mathbf{0.115 \pm 0.0}$	$\mathbf{6.68 \pm 0.1}$	**92**	$\mathbf{0.119 \pm 0.002}$	$\mathbf{6.8 \pm 0.29}$	**72**

表 2 YT8M-MusicVideo データセットを用いたマルチモーダルテキスト検索タスクにおける性能比較.

	楽曲 → テキスト			画像 → テキスト			楽曲 & 画像 → テキスト		
手法	MRR ↑	R@1 (%) ↑	MR ↓	MRR ↑	R@1 (%) ↑	MR ↓	MRR ↑	R@1 (%) ↑	MR ↓
PCME	–	–	–	0.023 ± 0.002	0.73 ± 0.16	372	–	–	–
MPC	–	–	–	0.013 ± 0.001	0.13 ± 0.05	427	–	–	–
ベースライン	0.026 ± 0.001	0.6 ± 0.18	226	0.046 ± 0.001	1.47 ± 0.1	167	0.054 ± 0.002	1.83 ± 0.3	131
提案手法	$\mathbf{0.039 \pm 0.001}$	$\mathbf{1.17 \pm 0.09}$	**180**	$\mathbf{0.118 \pm 0.002}$	$\mathbf{6.87 \pm 0.21}$	**89**	$\mathbf{0.139 \pm 0.002}$	$\mathbf{7.97 \pm 0.46}$	**55**

表 3 YT8M-MusicVideo データセットを用いたマルチモーダル楽曲検索タスクにおける性能比較.

	画像 → 楽曲			テキスト → 楽曲			画像 & テキスト → 楽曲		
手法	MRR ↑	R@1 (%) ↑	MR ↓	MRR ↑	R@1 (%) ↑	MR ↓	MRR ↑	R@1 (%) ↑	MR ↓
ベースライン	0.021 ± 0.001	0.52 ± 0.05	263	0.028 ± 0.001	0.68 ± 0.08	219	0.032 ± 0.002	0.83 ± 0.26	191
提案手法	$\mathbf{0.027 \pm 0.001}$	$\mathbf{0.58 \pm 0.06}$	**235**	$\mathbf{0.041 \pm 0.003}$	$\mathbf{1.25 \pm 0.37}$	**173**	$\mathbf{0.05 \pm 0.002}$	$\mathbf{1.75 \pm 0.25}$	**141**

表 4 AS5M データセットを用いたマルチモーダル画像検索タスクにおける性能比較.

	楽曲 → 画像			テキスト → 画像			楽曲 & テキスト → 画像		
手法	MRR ↑	R@1 (%) ↑	MR ↓	MRR ↑	R@1 (%) ↑	MR ↓	MRR ↑	R@1 (%) ↑	MR ↓
PCME	–	–	–	0.069 ± 0.004	2.82 ± 0.34	131	–	–	–
MPC	–	–	–	0.026 ± 0.002	0.62 ± 0.15	240	–	–	–
ベースライン	0.046 ± 0.002	1.37 ± 0.19	141	0.125 ± 0.005	6.21 ± 0.56	50	0.1 ± 0.004	4.39 ± 0.53	60
提案手法	$\mathbf{0.074 \pm 0.004}$	$\mathbf{2.94 \pm 0.46}$	**94**	$\mathbf{0.539 \pm 0.005}$	$\mathbf{45.37 \pm 0.65}$	**2**	$\mathbf{0.508 \pm 0.008}$	$\mathbf{41.35 \pm 1.12}$	**2**

表 5 AS5M データセットを用いたマルチモーダルテキスト検索タスクにおける性能比較.

	楽曲 → テキスト			画像 → テキスト			楽曲 & 画像 → テキスト		
手法	MRR ↑	R@1 (%) ↑	MR ↓	MRR ↑	R@1 (%) ↑	MR ↓	MRR ↑	R@1 (%) ↑	MR ↓
PCME	–	–	–	0.067 ± 0.003	2.73 ± 0.27	131	–	–	–
MPC	–	–	–	0.025 ± 0.002	0.57 ± 0.13	239	–	–	–
ベースライン	0.062 ± 0.002	1.93 ± 0.27	82	0.126 ± 0.006	5.99 ± 0.59	47	0.146 ± 0.007	6.96 ± 0.76	30
提案手法	$\mathbf{0.113 \pm 0.004}$	$\mathbf{4.99 \pm 0.37}$	**46**	$\mathbf{0.541 \pm 0.007}$	$\mathbf{44.21 \pm 0.99}$	**2**	$\mathbf{0.58 \pm 0.009}$	$\mathbf{47.75 \pm 1.19}$	**2**

表 6 AS5M データセットを用いたマルチモーダル楽曲検索タスクにおける性能比較.

	画像 → 楽曲			テキスト → 楽曲			画像 & テキスト → 楽曲		
手法	MRR ↑	R@1 (%) ↑	MR ↓	MRR ↑	R@1 (%) ↑	MR ↓	MRR ↑	R@1 (%) ↑	MR ↓
ベースライン	0.045 ± 0.002	1.32 ± 0.2	138	0.067 ± 0.003	2.11 ± 0.24	77	0.069 ± 0.003	2.43 ± 0.32	74
提案手法	$\mathbf{0.072 \pm 0.004}$	$\mathbf{2.62 \pm 0.33}$	**92**	$\mathbf{0.115 \pm 0.005}$	$\mathbf{4.86 \pm 0.47}$	**44**	$\mathbf{0.126 \pm 0.006}$	$\mathbf{5.54 \pm 0.62}$	**37**

た．さらに，提案手法は，MCInfoNCE [22] を拡張した手法である確率的対照損失関数のみを用いたベースラインと比較しても，ほぼすべての検索タスクにおいて検索性能が向上している．これらの結果から，最適輸送に基づく損失関数である $\mathcal{L}_S$ が，検索タスクにおける検索性能向上に寄与していることが言える．また，音楽情報の各データを二

表 7 YT8M-MusicVideo データセットを用いた重み λ_S による性能比較.

		楽曲 & テキスト → 画像			楽曲 & 画像 → テキスト			画像 & テキスト → 楽曲		
手法	λ_S	MRR ↑	R@1 (%) ↑	MR ↓	MRR ↑	R@1 (%) ↑	MR ↓	MRR ↑	R@1 (%) ↑	MR ↓
提案手法	0.01	0.105 ± 0.002	5.6 ± 0.19	78	0.129 ± 0.002	7.9 ± 0.56	62	0.044 ± 0.002	1.7 ± 0.23	151
提案手法	0.1	0.11 ± 0.003	5.8 ± 0.25	74	0.135 ± 0.002	7.6 ± 0.44	57	0.044 ± 0.003	1.4 ± 0.22	148
提案手法	1.0	$\mathbf{0.119 \pm 0.002}$	$\mathbf{6.8 \pm 0.29}$	**72**	$\mathbf{0.139 \pm 0.002}$	$\mathbf{7.97 \pm 0.46}$	**55**	$\mathbf{0.05 \pm 0.002}$	$\mathbf{1.75 \pm 0.25}$	**141**
提案手法	10	0.071 ± 0.001	4.37 ± 0.23	118	0.087 ± 0.003	5.87 ± 0.42	99	0.035 ± 0.003	1.08 ± 0.19	166
提案手法	100	0.077 ± 0.001	3.72 ± 0.13	122	0.082 ± 0.001	3.82 ± 0.19	112	0.017 ± 0.001	0.3 ± 0.11	329

表 8 AS5M データセットを用いた重み λ_S による性能比較.

		楽曲 & テキスト → 画像			楽曲 & 画像 → テキスト			画像 & テキスト → 楽曲		
手法	λ_S	MRR ↑	R@1 (%) ↑	MR ↓	MRR ↑	R@1 (%) ↑	MR ↓	MRR ↑	R@1 (%) ↑	MR ↓
提案手法	0.01	0.504 ± 0.006	41.08 ± 0.95	**2**	0.573 ± 0.008	47.0 ± 0.93	**2**	0.116 ± 0.005	4.8 ± 0.35	41
提案手法	0.1	0.502 ± 0.005	40.78 ± 0.76	**2**	0.573 ± 0.009	46.88 ± 1.02	**2**	0.117 ± 0.005	4.9 ± 0.38	40
提案手法	1.0	$\mathbf{0.508 \pm 0.008}$	$\mathbf{41.35 \pm 1.12}$	**2**	$\mathbf{0.58 \pm 0.009}$	$\mathbf{47.75 \pm 1.19}$	**2**	$\mathbf{0.126 \pm 0.006}$	$\mathbf{5.54 \pm 0.62}$	**37**
提案手法	10	0.447 ± 0.006	35.04 ± 0.79	3	0.533 ± 0.007	42.5 ± 1.01	**2**	0.109 ± 0.004	4.5 ± 0.39	44
提案手法	100	0.397 ± 0.006	30.42 ± 0.71	5	0.457 ± 0.005	35.08 ± 0.61	3	0.065 ± 0.002	2.08 ± 0.27	82

表 9 AS5M データセットを用いた極端な λ_S の値におけるマルチモーダル検索性能比較.

		楽曲 & テキスト → 画像			楽曲 & 画像 → テキスト			画像 & テキスト → 楽曲		
手法	λ_S	MRR ↑	R@1 (%) ↑	MR ↓	MRR ↑	R@1 (%) ↑	MR ↓	MRR ↑	R@1 (%) ↑	MR ↓
ベースライン ($\mathcal{L}_C$ のみ)	0	0.1 ± 0.004	4.39 ± 0.53	60	0.146 ± 0.007	6.96 ± 0.76	30	0.069 ± 0.003	2.43 ± 0.32	74
提案手法 ($\lambda_S \to 0$)	1.0×10^{-12}	0.102 ± 0.006	4.48 ± 0.57	58	0.148 ± 0.011	6.94 ± 0.77	29	0.07 ± 0.004	2.43 ± 0.35	73
提案手法	1.0	0.508 ± 0.008	41.35 ± 1.12	2	0.58 ± 0.009	47.75 ± 1.19	2	0.126 ± 0.006	5.54 ± 0.62	37
$\mathcal{L}_S$ のみ ($\lambda_S \to \infty$)	-	0.023 ± 0.004	0.88 ± 0.41	482	0.013 ± 0.001	0.3 ± 0.13	602	0.005 ± 0.0	0.05 ± 0.02	941

つ組み合わせたクエリを用いた場合，一つのみをクエリとして用いた場合と比較して，ほぼすべての検索タスクにおいて優れた検索性能を示している．提案手法は，潜在表現として確率分布を用いているため，複数の確率分布を簡易に混合することが可能であり，マルチモーダル音楽情報検索において有益である．

一方，データセット間に見られる検索性能の差については，それらのデータセット規模の違いによって説明できる．提案手法で用いている transformer モデルには，スケーリング則 [19] に従って性能が向上するといった性質が，様々なモダリティのデータ [6, 11, 15] や転移学習 [19] で確認されている．実際，YT8M-MusicVideo データセットは，AS5M データセットに比べて二桁ほどデータセット規模が小さく，特に楽曲検索タスクではその検索性能が下がる傾向がみられた．また，提案手法では CLAP の音声モデルを利用しているが，このモデルは LAION-Audio-630k データセット [49] で学習されており，CLIP の画像・言語モデルの学習で使用されたデータセットほどの規模はないため，検索性能に差が生じたことが考えられる．

再現率（R@k）について，より詳細な結果については付録（A.1.3 節）に記載する．

4.4 重み λ_S の分析

本研究の貢献は，マルチモーダル音楽情報検索を対象として，マルチモーダル確率的対照損失関数 $\mathcal{L}_C$ と最適輸送の計算手法である SSW に基づく損失関数 $\mathcal{L}_S$ の異なる二つの損失関数を新たに提案し，それらを組み合わせた点にある．提案手法の損失関数である式（13）について，重み λ_S を極端に大きい値や小さい値に設定してしまうと，うまく機能しない．これは，3 章で述べたように，それぞれの損失関数が異なる働きを持っているためである．

この重み λ_S について適切な値を探すために，比較実験を実施した．この比較実験は，$\lambda_S = \{0.01, 0.1, 1.0, 10.0, 100.0\}$ について実施した．実験結果を表 7，8 に示す．これらの結果が示すように，最適な λ_S の値は 1.0 であることがわかる．また λ_S の値が大きい場合，検索タスクにおける検索性能が低くなる結果が観察された．

この結果についてさらに検証するために，重み λ_S について極端な値を用いて実験した．この追加実験では，$\lambda_S \to 0$ を模した $\lambda_S = 1.0 \times 10^{-12}$ と，$\lambda_S \to \infty$ を模した $\mathcal{L} = \mathcal{L}_S$ について実施した．実験結果を表 9 に示す．この結果が示す通り，$\lambda_S \to 0$ の場合，重み λ_S が非常に小さく，SSW に基づく損失関数の働きの恩恵を受けられないため，検索タスクにおける検索性能はほとんどベースラインと同じで

あった．また，$\lambda_S \to \infty$ の場合，SSW に基づく損失関数のみ，つまり確率分布について正例のペアのみ損失関数の計算に用いるため，表現学習がうまくいかず，検索タスクにおける検索性能は大きく減少した．これらの結果は，これら二つの損失関数を適切な重みで組み合わせることが検索性能の向上において極めて重要であることを示している．

4.5 学習にかかる実行時間

表 10 に示すように，SSW に基づく損失関数の計算は高速である．提案手法では，SSW に基づく損失関数をマルチモーダル確率的対照損失関数に加えて計算しているにもかかわらず，ベースラインであるマルチモーダル確率的対照損失関数のみを計算した場合と比較して，学習時間はほぼ同じであった．よって，マルチモーダル確率的対照損失関数と SSW に基づく損失関数の双方を用いた提案手法は，マルチモーダル音楽情報検索において実用的である．

4.6 定性的分析

定性的分析では，楽曲，画像，テキストの潜在表現の性質について調べるために，音楽ジャンルごとに可視化した．AS5M データセットにおけるテスト用の小データセットから，データセットに含まれる割合の高い上位 10 種類のジャンルの楽曲（Pop, Hip-Hop/Rap, Dance, New Age, Jazz, Electronic, Classical, Rock, Alternative, Soundtrack）である 12,180 曲分の楽曲，画像，テキストを用いた．音楽情報の各データについて，それぞれの確率分布からのサンプル点の Fréchet 平均を計算し，t-SNE [45] を用いて二次元平面に射影した．図 3 に示すように，音楽情報の各データについて，ジャンルごとにまとまりを形成する傾向が見られたことから，同一ジャンルに含まれる楽曲，画像，テキストが関連付いていることを確認した．

4.7 マルチモーダル音楽情報検索の結果例

確率的表現学習の利点は，3.3 で述べたように，複数のデータの組み合わせをクエリとして用いることができる点である．この利点を活かして，実際に YT8M-MusicVideo データセットを用いたマルチモーダル音楽情報検索を行った[*2]．提案手法とベースラインの双方について，複数のデータの組み合わせたクエリに合った（つまり，S^{d-1}_{shared} 上でそのクエリの分布に最も近い分布である）上位三件の音楽情報を検索結果例として掲載している．実際に検索結果を視聴したところ，提案手法の方がベースラインよりも質的に優れた検索結果を得ることを確認した．

5. おわりに

本稿では，音楽情報の各データを確率的表現として活用した，マルチモーダル音楽情報検索のための手法を提案した．以下に，本研究の貢献をまとめる．一つ目は，確率的表現学習を用いたマルチモーダル音楽情報検索を実現するために，次の技術的貢献をした点である．まず，確率的表現学習において単一のモダリティのデータのみで使用されていた von Mises-Fisher (vMF) 分布を，マルチモーダル音楽情報検索のために拡張した．そして，超球面上の最適輸送を効率的に計算できる Spherical Sliced-Wasserstein (SSW) [4] p -距離を，マルチモーダル音楽情報検索タスクのために初めて導入した．さらに，確率的対照学習に基づく損失関数 $\mathcal{L}_C$ と，最適輸送に基づく損失関数 $\mathcal{L}_S$ の二つの損失関数を新たに設計した．二つ目は，提案手法である確率的対照学習に基づく損失関数 $\mathcal{L}_C$ と最適輸送に基づく損失関数 $\mathcal{L}_S$ の組み合わせが，マルチモーダル音楽情報検索において精度の高い検索が可能となることを定量的に示した点である．三つ目は，定性的分析を実施し，同一音楽の楽曲・画像・テキストが共有超球面 S^{d-1}_{shared} 上で近くに配置されていることを確認した点である．この分析結果も，提案手法がマルチモーダル音楽情報検索において有用であることを示している．

これら一連の貢献は，モダリティを問わず多様な検索タスクに対して適用可能な，より汎用性の高いアプローチにつながる可能性がある．今後，より挑戦的な検索タスクにおいても，提案手法が実用的な解決方法を提供できることを期待している．

表 10 学習にかかる実行時間の比較．

データセット	手法	実行時間（秒/組）
YT8M-MusicVideo	ベースライン	0.1050 ± 0.008
	提案手法	0.1096 ± 0.009
AS5M	ベースライン	0.05624 ± 0.0002
	提案手法	0.05629 ± 0.0003

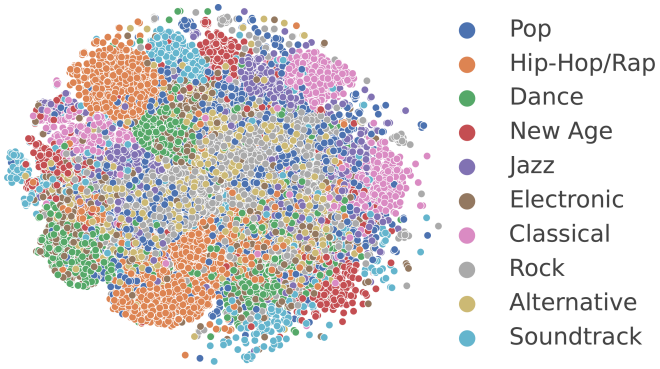

図 3 音楽情報の各データにおける潜在表現の可視化．AS5M データセットにおけるテストデータセットの小データセット群に含まれる楽曲，画像，テキストの潜在表現を t-SNE [45] を用いてジャンルごとに可視化した．

[*2] https://t39nakatsuka.github.io/ISMIR2024-demo/Demo.html

謝辞

本研究の一部は，JST CREST JPMJCR20D4 および JSPS 科研費 22K18017 の助成を受けたものです.

参考文献

[1] Abu-El-Haija, S., Kothari, N., Lee, J., Natsev, P., Toderici, G., Varadarajan, B. and Vijayanarasimhan, S.: YouTube-8M: A large-scale video classification benchmark, *arXiv preprint arXiv:1609.08675* (2016).

[2] Arjovsky, M., Chintala, S. and Bottou, L.: Wasserstein generative adversarial networks, *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 214–223 (2017).

[3] Bendokat, T., Zimmermann, R. and Absil, P.-A.: A Grassmann manifold handbook: Basic geometry and computational aspects, *arXiv preprint arXiv:2011.13699* (2020).

[4] Bonet, C., Berg, P., Courty, N., Septier, F., Drumetz, L. and Pham, M.-T.: Spherical sliced-wasserstein, *Proceedings of the International Conference on Learning Representations (ICLR)* (2023).

[5] Chang, J., Lan, Z., Cheng, C. and Wei, Y.: Data uncertainty learning in face recognition, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5710–5719 (2020).

[6] Cherti, M., Beaumont, R., Wightman, R., Wortsman, M., Ilharco, G., Gordon, C., Schuhmann, C., Schmidt, L. and Jitsev, J.: Reproducible scaling laws for contrastive language-image learning, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2818–2829 (2023).

[7] Chun, S., Oh, S. J., De Rezende, R. S., Kalantidis, Y. and Larlus, D.: Probabilistic embeddings for cross-modal retrieval, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8415–8424 (2021).

[8] Craswell, N.: Mean Reciprocal Rank, *Encyclopedia of Database Systems*, Springer US (2009).

[9] Davidson, T. R., Falorsi, L., De Cao, N., Kipf, T. and Tomczak, J. M.: Hyperspherical Variational Auto-Encoders, *Proceeding of the Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 856–865 (2018).

[10] Delon, J., Salomon, J. and Sobolevski, A.: Fast transport optimization for Monge costs on the circle, *SIAM J. Appl. Math.*, Vol. 70, No. 7, pp. 2239–2258 (2010).

[11] Droppo, J. and Elibol, O.: Scaling laws for acoustic models, *Proceedings of Interspeech 2021*, pp. 2576–2580 (2021).

[12] Frogner, C., Zhang, C., Mobahi, H., Araya, M. and Poggio, T. A.: Learning with a Wasserstein loss, *Proceeding of the Advances in Neural Information Processing Systems (NIPS)*, Vol. 28 (2015).

[13] Garg, D., Wang, Y., Hariharan, B., Campbell, M., Weinberger, K. Q. and Chao, W.-L.: Wasserstein distances for stereo disparity estimation, *Proceeding of the Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 33, pp. 22517–22529 (2020).

[14] He, K., Fan, H., Wu, Y., Xie, S. and Girshick, R.: Momentum Contrast for Unsupervised Visual Representation Learning, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9729–9738 (2020).

[15] Hernandez, D., Kaplan, J., Henighan, T. and McCandlish, S.: Scaling laws for transfer, *arXiv preprint arXiv:2102.01293* (2021).

[16] Huang, Q., Jansen, A., Lee, J., Ganti, R., Li, J. Y. and Ellis, D. P.: MuLan: A joint embedding of music audio and natural language, *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pp. 559–566 (2022).

[17] Hundrieser, S., Klatt, M. and Munk, A.: *The Statistics of Circular Optimal Transport*, pp. 57–82, Springer Nature Singapore (2022).

[18] Jebara, T., Kondor, R. and Howard, A.: Probability product kernels, *J. Mach. Learn. Res.*, Vol. 5, pp. 819–844 (2004).

[19] Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J. and Amodei, D.: Scaling laws for neural language models, *arXiv preprint arXiv:2001.08361* (2020).

[20] Kingma, D. P. and Ba, J.: Adam: A Method for Stochastic Optimization, *Proceedings of the International Conference on Learning Representations (ICLR)* (2015).

[21] Kingma, D. P. and Welling, M.: Auto-encoding variational bayes, *Proceeding of the International Conference on Learning Representations (ICLR)* (2014).

[22] Kirchhof, M., Kasneci, E. and Oh, S. J.: Probabilistic Contrastive Learning Recovers the Correct Aleatoric Uncertainty of Ambiguous Inputs, *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 17085–17104 (2023).

[23] Kirchhof, M., Roth, K., Akata, Z. and Kasneci, E.: A Non-isotropic Probabilistic Take on Proxy-based Deep Metric Learning, *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 435–454 (2022).

[24] Kolouri, S., Nadjahi, K., Simsekli, U., Badeau, R. and Rohde, G.: Generalized sliced wasserstein distances, *Proceeding of the Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 32 (2019).

[25] Le-Khac, P. H., Healy, G. and Smeaton, A. F.: Contrastive representation learning: A framework and review, *IEEE Access*, Vol. 8, pp. 193907–193934 (2020).

[26] Li, H., Song, J., Gao, L., Zeng, P., Zhang, H. and Li, G.: A Differentiable Semantic Metric Approximation in Probabilistic Embedding for Cross-Modal Retrieval, *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 35, pp. 11934–11946 (2022).

[27] Li, S., Xu, J., Xu, X., Shen, P., Li, S. and Hooi, B.: Spherical confidence learning for face recognition, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15629–15637 (2021).

[28] Luo, Y. and Mesgarani, N.: Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation, *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, Vol. 27, No. 8, pp. 1256–1266 (2019).

[29] Nakatsuka, T., Hamasaki, M. and Goto, M.: Harnessing the Power of Distributions: Probabilistic Representation Learning on Hypersphere for Multimodal Music Information Retrieval, *Proceedings of the the 25th International Society for Music Information Retrieval Conference (ISMIR)* (2024).

[30] Neculai, A., Chen, Y. and Akata, Z.: Probabilistic Compositional Embeddings for Multimodal Image Retrieval, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.

4547–4557 (2022).
[31] Oh, S. J., Murphy, K., Pan, J., Roth, J., Schroff, F. and Gallagher, A.: Modeling uncertainty with hedged instance embedding, *arXiv preprint arXiv:1810.00319* (2018).
[32] Oord, A. v. d., Li, Y. and Vinyals, O.: Representation Learning with Contrastive Predictive Coding, *arXiv preprint arXiv:1807.03748* (2018).
[33] Park, J., Lee, J., Kim, I.-J. and Sohn, K.: Probabilistic representations for video contrastive learning, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14711–14721 (2022).
[34] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L. et al.: PyTorch: An Imperative Style, High-performance Deep Learning Library, *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, Vol. 32, pp. 8024–8035 (2019).
[35] Prétet, L., Richard, G. and Peeters, G.: Cross-Modal Music-Video Recommendation: A Study of Design Choices, *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, pp. 1–9 (2021).
[36] Rabin, J., Delon, J. and Gousseau, Y.: Transportation distances on the circle, *J. Math. Imaging Vis.*, Vol. 41, No. 1, pp. 147–167 (2011).
[37] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J. et al.: Learning transferable visual models from natural language supervision, *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 8748–8763 (2021).
[38] Roads, B. D. and Love, B. C.: Enriching imagenet with human similarity judgments and psychological embeddings, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3547–3557 (2021).
[39] Scott, T. R., Gallagher, A. C. and Mozer, M. C.: von Mises-Fisher loss: An exploration of embedding geometries for supervised learning, *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10612–10622 (2021).
[40] Sellam, T., Das, D. and Parikh, A. P.: BLEURT: Learning robust metrics for text generation, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 7881–7892 (2020).
[41] Shi, Y. and Jain, A. K.: Probabilistic face embeddings, *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6902–6911 (2019).
[42] Sohn, K.: Improved Deep Metric Learning with Multi-Class N-pair Loss Objective, *Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS)*, Vol. 29, pp. 1857–1865 (2016).
[43] Surís, D., Vondrick, C., Russell, B. and Salamon, J.: It's Time for Artistic Correspondence in Music and Video, *Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition Conference (CVPR)*, pp. 10564–10574 (2022).
[44] Takahashi, R., Matsubara, T. and Uehara, K.: Data augmentation using random image cropping and patching for deep CNNs, *IEEE Trans. Circuits. Syst. Video Technol.*, Vol. 30, No. 9, pp. 2917–2931 (2019).
[45] Van der Maaten, L. and Hinton, G.: Visualizing data using t-SNE, *J. Mach. Learn. Res.*, Vol. 9, No. 11, pp. 2579–2605 (2008).
[46] Villani, C.: *Topics in optimal transportation*, Vol. 58, American Mathematical Soc. (2021).
[47] Vilnis, L. and McCallum, A.: Word representations via gaussian embedding, *Proceedings of the International Conference on Learning Representations (ICLR)*, pp. 1–12 (2014).
[48] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., Drame, M., Lhoest, Q. and Rush, A. M.: Transformers: State-of-the-Art Natural Language Processing, *Proceedings of the Empirical Methods in Natural Language Processing: System Demonstrations (EMNLP-SD)*, pp. 38–45 (2020).
[49] Wu, Y., Chen, K., Zhang, T., Hui, Y., Berg-Kirkpatrick, T. and Dubnov, S.: Large-scale Contrastive Language-Audio Pretraining with Feature Fusion and Keyword-to-Caption Augmentation, *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5 (2023).
[50] Zhong, Z., Zheng, L., Kang, G., Li, S. and Yang, Y.: Random erasing data augmentation, *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pp. 13001–13008 (2020).

付　　録

A.1 Spherical Sliced-Wasserstein の計算手順と疑似コード

A.1.1 SSW のアルゴリズム

提案手法では，SSW p -距離について，$p = 1$ の場合（つまり，SSW_1）を用いた．アルゴリズム 1 は，SSW_1 の計算手順を示している．

Algorithm 1 SSW_1 の計算手順

入力： $\zeta_n \sim p(\mathbf{z}_n^*|*_n)$, $\eta_n \sim p(\mathbf{z}_n^\star|\star_n)$ $(*, \star \in \{\mathbf{a}, \mathbf{i}, \mathbf{t}\}, * \neq \star)$
ステップ 1： $E \ni e_{ij} \sim \mathcal{N}(0,1)$ である行列 $E \in \mathbb{R}^{d\times 2}$ を作成
ステップ 2： 行列 E に QR 分解を適用し U を計算：$U = \mathrm{QR}(E)$
ステップ 3： サンプル点 $z^\zeta \in \zeta_n$, $z^\eta \in \eta_n$ を大円 S^1 上に射影：$\hat{z}^\zeta = \frac{U^\top z^\zeta}{||U^\top z^\zeta||_2}$, $\hat{z}^\eta = \frac{U^\top z^\eta}{||U^\top z^\eta||_2}$
ステップ 4： atan2 関数を用いて，大円 S^1 上における座標を計算：$\tilde{z}^\zeta = \frac{\mathrm{atan2}(-y_{\hat{z}^\zeta}, -x_{\hat{z}^\zeta})+\pi}{2\pi}$, $\tilde{z}^\eta = \frac{\mathrm{atan2}(-y_{\hat{z}^\eta}, -x_{\hat{z}^\eta})+\pi}{2\pi}$（ただし，$\hat{z}^\zeta = (x_{\hat{z}^\zeta}, y_{\hat{z}^\zeta})$, $\hat{z}^\eta = (x_{\hat{z}^\eta}, y_{\hat{z}^\eta})$）
ステップ 5： 式（3）を用いて $W_1(\sum \delta_{\tilde{z}^\zeta}, \sum \delta_{\tilde{z}^\eta})$ を計算
ステップ 6： W_1 の計算を $\mathcal{T}$ 回繰り返し，平均を算出：$SSW_1(\zeta, \eta) \approx \frac{1}{\mathcal{T}} \sum^{\mathcal{T}} W_1(\sum \delta_{\tilde{z}^\zeta}, \sum \delta_{\tilde{z}^\eta})$
出力： $SSW_1(\zeta, \eta)$

A.1.2 SSW に基づく損失関数の疑似コード

アルゴリズム 2 は，SSW に基づく損失関数における SSW_1 の疑似コードを示している．この疑似コードは，PyTorch [34] の記法に基づいて書かれている．

SSW に基づく損失関数の計算手順は，まず，推定されたパラメータから vMF 分布を生成し，得られた vMF 分

Algorithm 2 SSW に基づく損失関数における SSW_1 を計算する疑似コード

```
# a_mu, a_kappa - 楽曲のミニバッチについて推定された
    vMF 分布のパラメータ.
# i_mu, i_kappa - 画像のミニバッチについて推定された
    vMF 分布のパラメータ.
# t_mu, t_kappa - テキストのミニバッチについて推定され
    たvMF 分布のパラメータ.

# VonMisesFisher - ‘‘torch.distributions.
    Distribution’’を用いたvMF 分布の実装. ‘‘https://
    github.com/nicola-decao/s-vae-pytorch/blob/
    master/hyperspherical_vae/distributions/
    von_mises_fisher.py’’を参照した.
# L - それぞれの確率分布から取得するサンプル点の数.

# SSW_1 - Spherical Sliced-Wasserstein (SSW) 距離
    . ‘‘https://github.com/clbonet/
    Spherical_Sliced-Wasserstein/blob/main/lib/
    sw_sphere.py’’ を参照した.

# SSW に基づく損失関数
SBLossFunction(a_mu, a_kappa, i_mu, i_kappa, t_mu,
     t_kappa):

    # 推定されたパラメーターからvMF 分布を生成し, rejection
        -sampling reparameterization
        trick を用いてサンプル点を取得.
    p_audio = VonMisesFisher(a_mu, a_kappa).rsample
        (L)
    p_image = VonMisesFisher(i_mu, i_kappa).rsample
        (L)
    p_text = VonMisesFisher(t_mu, t_kappa).rsample(
        L)

    # 正例のペア(ここでは, ミニバッチにおいて同じ添字であ
        るデータのペア)の確率分布間のSSW 距離を計算.
    p_distance = (SSW_1(p_audio, p_image) + SSW_1(
        p_image, p_text) + SSW_1(p_text, p_audio))
         / 3

    return p_distance
```

布に rejection-sampling reparameterization trick [9] を適用することで，それぞれの分布から L 個のサンプル点を得る．そして，得られたサンプル点を用いて正例のペアの確率分布間の SSW_1 を計算する．

A.1.3 再現率（R@k）についての詳細な結果

表 A·1–A·6 に，より大きい k における再現率（R@k）の結果を示す．ここでは R@k について，$k = 5, 10, 15$ のそれぞれの値で算出した．これらの結果からわかるように，提案手法はより大きい k についても，R@k について比較手法よりも高い検索性能を示している．

表 A·1 YT8M-MusicVideo データセットを用いたマルチモーダル画像検索タスクにおける R@k 指標の詳細な性能比較.

	楽曲 → 画像			テキスト → 画像			楽曲 & テキスト → 画像		
手法	R@5 (%) ↑	R@10 (%) ↑	R@15 (%) ↑	R@5 (%) ↑	R@10 (%) ↑	R@15 (%) ↑	R@5 (%) ↑	R@10 (%) ↑	R@15 (%) ↑
PCME	–	–	–	2.82 ± 0.55	4.78 ± 0.69	6.73 ± 0.9	–	–	–
MPC	–	–	–	1.32 ± 0.29	2.83 ± 0.13	4.1 ± 0.07	–	–	–
ベースライン	2.5 ± 0.16	4.67 ± 0.31	6.32 ± 0.18	5.85 ± 0.15	9.6 ± 0.45	12.8 ± 0.59	5.52 ± 0.12	8.65 ± 0.41	11.53 ± 0.62
提案手法	$\mathbf{3.45 \pm 0.44}$	$\mathbf{5.85 \pm 0.29}$	$\mathbf{7.82 \pm 0.52}$	$\mathbf{15.13 \pm 0.14}$	$\mathbf{21.18 \pm 0.21}$	$\mathbf{24.75 \pm 0.32}$	$\mathbf{15.45 \pm 0.67}$	$\mathbf{21.15 \pm 0.4}$	$\mathbf{25.37 \pm 0.55}$

表 A·2 YT8M-MusicVideo データセットを用いたマルチモーダルテキスト検索タスクにおける R@k 指標の詳細な性能比較.

	楽曲 → テキスト			画像 → テキスト			楽曲 & 画像 → テキスト		
手法	R@5 (%) ↑	R@10 (%) ↑	R@15 (%) ↑	R@5 (%) ↑	R@10 (%) ↑	R@15 (%) ↑	R@5 (%) ↑	R@10 (%) ↑	R@15 (%) ↑
PCME	–	–	–	2.58 ± 0.15	4.57 ± 0.1	6.0 ± 0.52	–	–	–
MPC	–	–	–	1.2 ± 0.2	2.55 ± 0.04	3.72 ± 0.31	–	–	–
ベースライン	2.93 ± 0.17	5.22 ± 0.16	7.27 ± 0.2	6.02 ± 0.2	9.88 ± 0.42	12.83 ± 0.45	6.73 ± 0.05	11.55 ± 0.29	15.15 ± 0.19
提案手法	$\mathbf{4.68 \pm 0.44}$	$\mathbf{7.73 \pm 0.42}$	$\mathbf{10.4 \pm 0.4}$	$\mathbf{15.28 \pm 0.27}$	$\mathbf{21.32 \pm 0.21}$	$\mathbf{25.18 \pm 0.3}$	$\mathbf{18.35 \pm 0.51}$	$\mathbf{24.62 \pm 0.8}$	$\mathbf{30.47 \pm 0.49}$

表 A·3 YT8M-MusicVideo データセットを用いたマルチモーダル楽曲検索タスクにおける R@k 指標の詳細な性能比較.

	画像 → 楽曲			テキスト → 楽曲			画像 & テキスト → 楽曲		
手法	R@5 (%) ↑	R@10 (%) ↑	R@15 (%) ↑	R@5 (%) ↑	R@10 (%) ↑	R@15 (%) ↑	R@5 (%) ↑	R@10 (%) ↑	R@15 (%) ↑
ベースライン	2.15 ± 0.23	4.27 ± 0.49	5.97 ± 0.49	3.08 ± 0.24	5.4 ± 0.37	7.58 ± 0.51	3.93 ± 0.1	6.98 ± 0.08	9.52 ± 0.2
提案手法	$\mathbf{3.15 \pm 0.21}$	$\mathbf{5.98 \pm 0.39}$	$\mathbf{8.02 \pm 0.45}$	$\mathbf{4.98 \pm 0.12}$	$\mathbf{8.92 \pm 0.09}$	$\mathbf{11.48 \pm 0.14}$	$\mathbf{6.35 \pm 0.2}$	$\mathbf{10.17 \pm 0.37}$	$\mathbf{13.52 \pm 0.36}$

表 A·4 AS5M データセットを用いたマルチモーダル画像検索タスクにおける R@k 指標の詳細な性能比較.

	楽曲 → 画像			テキスト → 画像			楽曲 & テキスト → 画像		
手法	R@5 (%) ↑	R@10 (%) ↑	R@15 (%) ↑	R@5 (%) ↑	R@10 (%) ↑	R@15 (%) ↑	R@5 (%) ↑	R@10 (%) ↑	R@15 (%) ↑
PCME	–	–	–	9.24 ± 0.67	14.33 ± 0.81	18.0 ± 0.68	–	–	–
MPC	–	–	–	2.94 ± 0.36	5.17 ± 0.54	7.45 ± 0.58	–	–	–
ベースライン	5.6 ± 0.39	9.77 ± 0.33	13.24 ± 0.47	17.43 ± 0.71	25.35 ± 0.83	30.66 ± 1.22	13.71 ± 0.74	21.04 ± 0.58	26.12 ± 0.48
提案手法	$\mathbf{9.9 \pm 0.75}$	$\mathbf{15.74 \pm 0.51}$	$\mathbf{20.22 \pm 0.61}$	$\mathbf{65.67 \pm 0.65}$	$\mathbf{72.66 \pm 0.77}$	$\mathbf{76.25 \pm 0.57}$	$\mathbf{63.15 \pm 0.6}$	$\mathbf{70.88 \pm 0.74}$	$\mathbf{75.04 \pm 0.75}$

表 A·5 AS5M データセットを用いたマルチモーダルテキスト検索タスクにおける R@k 指標の詳細な性能比較.

	楽曲 → テキスト			画像 → テキスト			楽曲 & 画像 → テキスト		
手法	R@5 (%) ↑	R@10 (%) ↑	R@15 (%) ↑	R@5 (%) ↑	R@10 (%) ↑	R@15 (%) ↑	R@5 (%) ↑	R@10 (%) ↑	R@15 (%) ↑
PCME	–	–	–	8.9 ± 0.5	13.97 ± 0.75	17.78 ± 0.73	–	–	–
MPC	–	–	–	2.76 ± 0.45	5.23 ± 0.65	7.33 ± 0.7	–	–	–
ベースライン	8.17 ± 0.36	13.6 ± 0.46	18.23 ± 0.7	17.7 ± 0.62	25.6 ± 0.84	31.24 ± 0.57	20.24 ± 0.98	29.88 ± 1.05	37.11 ± 0.99
提案手法	$\mathbf{15.74 \pm 0.58}$	$\mathbf{23.8 \pm 0.65}$	$\mathbf{29.82 \pm 0.99}$	$\mathbf{65.78 \pm 0.62}$	$\mathbf{72.03 \pm 0.68}$	$\mathbf{75.58 \pm 0.7}$	$\mathbf{70.24 \pm 0.79}$	$\mathbf{77.14 \pm 0.49}$	$\mathbf{80.75 \pm 0.64}$

表 A·6 AS5M データセットを用いたマルチモーダル楽曲検索タスクにおける R@k 指標の詳細な性能比較.

	画像 → 楽曲			テキスト → 楽曲			画像 & テキスト → 楽曲		
手法	R@5 (%) ↑	R@10 (%) ↑	R@15 (%) ↑	R@5 (%) ↑	R@10 (%) ↑	R@15 (%) ↑	R@5 (%) ↑	R@10 (%) ↑	R@15 (%) ↑
ベースライン	5.36 ± 0.4	9.58 ± 0.49	13.14 ± 0.52	8.8 ± 0.36	14.89 ± 0.42	19.51 ± 0.82	8.88 ± 0.55	15.2 ± 0.63	20.01 ± 0.89
提案手法	$\mathbf{9.8 \pm 0.55}$	$\mathbf{15.89 \pm 0.57}$	$\mathbf{20.08 \pm 0.78}$	$\mathbf{16.3 \pm 0.71}$	$\mathbf{24.87 \pm 0.78}$	$\mathbf{30.67 \pm 0.95}$	$\mathbf{17.88 \pm 0.75}$	$\mathbf{26.77 \pm 0.99}$	$\mathbf{33.18 \pm 0.96}$