Vol.2024-MUS-139 No.11
2024/3/9

メモリ機構付き深層距離学習に基づく音楽と画像の双方向検索

中塚 貴之[1,a)]　濱崎 雅弘[1,b)]　後藤 真孝[1,c)]

概要：本稿では，音楽をクエリとしてその音楽に合った画像の検索（またはその逆）という双方向の検索を実現するための，深層距離学習に基づく手法について述べる．音楽と画像の適切なペアとして音楽とその代表画像（例えば，ジャケット画像やサムネイル画像）を利用し，まず，音楽と画像の共有埋め込み空間において，音楽とその代表画像の埋め込み特徴量が互いに近づき，音楽と画像のランダムなペアの埋め込み特徴量が互いに遠ざかるように，音楽エンコーダと画像エンコーダをそれぞれ学習する．さらに我々は，学習時の音楽と画像の埋め込み特徴量を保存し，学習イテレーションをまたいで保持するメモリ機構を提案し，それをエンコーダの学習に利用する．具体的には，それぞれのエンコーダの学習が進んだ段階で，メモリに保存した音楽と画像の埋め込み特徴量から，エンコーダのバッチ学習に有益な埋め込み特徴量を取り出し，損失関数の計算において使用する．このような学習を行うために，我々は音楽と画像のペアを 78,325 件含むデータセットを新たに構築した．このデータセットを用いて，検索タスクにおける性能評価指標である平均逆順位・再現率・中央順位に基づく比較実験を実施し，提案機構の有効性を示した．また，同じデータセットを用いた定性分析では，音楽と画像の共有埋め込み空間において，関連するカテゴリタグが付された音楽と画像が近くに埋め込まれていることを確認した．

1. はじめに

ジャケット画像やサムネイル画像といった音楽の代表画像は，その音楽に合わせてデザインされている [5, 36, 40]．その目的として，消費者に音楽を広く販促すること，そして音楽の鑑賞体験を向上させることが挙げられている．これらの主張を裏付けるように，有名なデザイナーである Vlad Sepetov 氏は，インタビュー[*1]において，音楽に合った画像の制作に対する信念を述べている．またこれまでに，多くのデザイナーが音楽に合った画像を制作してきた結果として，異なる音楽ジャンルでは画像のデザインが大きく異なることが分析によって明らかになっている [25]．実際に，ジャケット画像から得られる画像特徴量は，音楽のジャンル分類に利用できることが示されてきた [17]．本研究では，この音楽と画像の密接な関係を利用して，音楽と画像を双方向に検索するタスクに取り組む（図 1）．

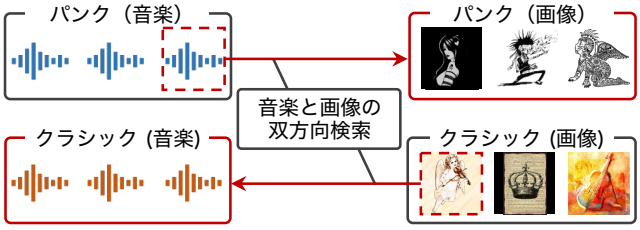

図 1: 音楽と画像を双方向に検索するタスクの概念図．本研究では，（1）音楽をクエリとしてその音楽に合った画像を検索，および（2）画像をクエリとしてその画像に合った音楽を検索する二つのタスクについて取り組む．

音楽と画像の双方向の検索を実現できる技術は，様々な音楽情報検索アプリケーションとして利用できる．この技術の応用例として，新たに音楽を作ったミュージシャンが，画像素材サイト等で取り扱われている権利が明確な画像の中から，その音楽に合った画像を検索できるアプリケーションや，画像に合った音楽を複数検索して，プレイリストを作成するアプリケーションが挙げられる．さらにこの技術は，膨大な数に及ぶ音楽と画像について，それらの潜在的な関係性を明らかにできる可能性がある．

音楽をクエリとした画像検索や画像をクエリとした音楽検索に関する先行研究 [2, 15, 18, 22, 23, 27, 30–32, 37, 47, 51, 56, 57] では，ムードタグや感情タグ，テキストといったメ

[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

[*1] https://www.irishtimes.com/culture/music/the-art-of-the-sleeve-every-album-cover-tells-a-story-1.2821084 （2024 年 2 月 6 日にアクセス）

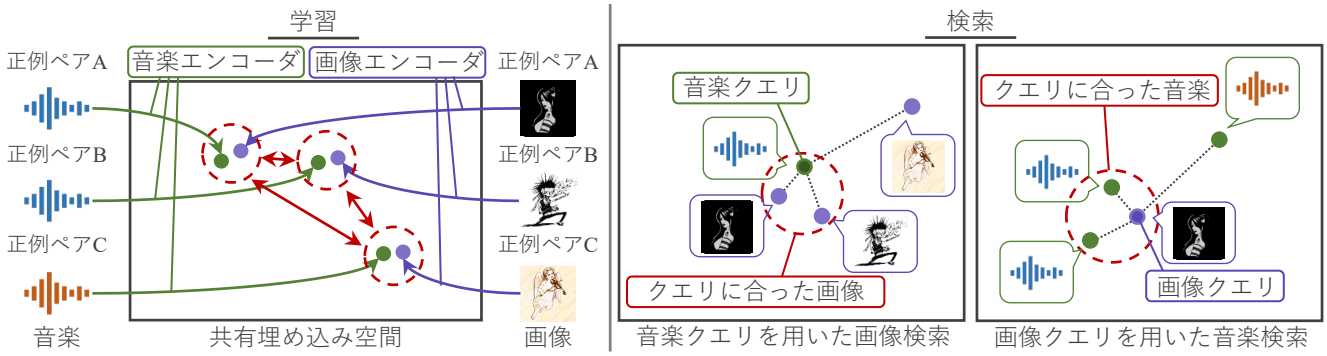

図 2: 提案手法の概要図．（左図）我々は，共有埋め込み空間において，音楽とその代表画像（つまり，正例のペア）の埋め込み特徴量が互いに近づき，音楽と画像のランダムな（負例の）ペアの埋め込み特徴量が互いに遠ざかるようにエンコーダを学習する．（右図）音楽をクエリとしてその音楽に合った画像（またはその逆）を，学習したエンコーダを用いて得られる音楽と画像の埋め込み特徴量の類似度に基づいて検索することが可能になる．

タデータを使用するアプローチが取られてきた．しかし，それらのメタデータが必ずしも音楽と画像に付されていないといった問題や，メタデータが様々なデータセットや音楽サービスの間で一貫していないといった問題がある．そのため，このアプローチで大規模なデータセットを取り扱う場合，全ての音楽と画像に一貫してメタデータを付す必要があり，その作業者に大変な負担を強いる可能性がある．また，音楽と画像に付したいメタデータによっては，作業者に専門的な音楽知識が求められる．さらにこのアプローチは，マイナーなタグが付された音楽や画像を検索し難くするといった問題があることが報告されている [10,41]．これらの問題を避けるために，本研究ではメタデータを一切使用せず，音楽と画像のみを活用するアプローチを試みる．

このアプローチで，音楽と画像の双方向の検索を実現するために，我々は深層距離学習に基づく手法を提案する．提案手法の概要を図 2 に示す．我々は，音楽と画像の適切なペア（つまり，音楽とその代表画像）を正例のペア，音楽と画像のランダムなペアを負例のペアと仮定し，音楽と画像をそれぞれ共有埋め込み空間に埋め込む二つのエンコーダ（音楽エンコーダと画像エンコーダ）を学習する．つまり，共有埋め込み空間において，音楽とその代表画像の埋め込み特徴量が互いに近づき，音楽と画像のランダムなペアの埋め込み特徴量が互いに遠ざかるように，音楽エンコーダと画像エンコーダをそれぞれ学習する．これらのエンコーダを適切に学習することができれば，音楽をクエリとしてその音楽に合った画像（またはその逆）を，それらの埋め込み特徴量の類似度に基づいて検索することが可能になる．

深層距離学習を用いてエンコーダを適切に学習するために必要なことは，エンコーダの学習にとって有益なペアを構成することである [33,43,46]．しかし前述の仮定において，音楽とその代表画像は唯一の正例のペアであるため，メタデータを使わないアプローチと比較して正例のペアが非常に少ないといった問題がある．この問題を解決するために，我々は先行研究のメモリ機構 [44,53] の仕組みを拡張した新しいメモリ機構 SCFEM [24] を提案する．提案機構は，学習時の音楽と画像の埋め込み特徴量をメモリに保存し，学習イテレーションをまたいで保持する仕組みを備えている．提案機構を用いることで，新たに正例のペアを構成することができる．具体的には，音楽の埋め込み特徴量と，メモリに保存されたその音楽およびその代表画像の埋め込み特徴量で構成されるそれぞれのペアが正例のペアであると仮定する（画像についても同様）ことで，正例のペアを増やすことが可能となる．

このような学習を行うために必要な，音楽とその代表画像を含むデータセットは一般には公開されていないため，我々は音楽と画像のペア[*2]を 78,325 件含むデータセット（以後，MCA データセットと呼ぶ）を構築した．我々はこの MCA データセットを用いて，平均逆順位 [4]・再現率・中央順位 [38] に基づく比較実験を実施し，提案機構の有効性を検証する．また，学習によって得られた埋め込み特徴量が，音楽と画像の共有埋め込み空間においてどのように分布しているか調べるために，MCA データセットを用いて分析する．

2. 関連研究

2.1 深層距離学習に基づく情報検索

音楽と画像の双方向検索を実現した深層距離学習に基づく手法として，感情タグをエンコーダの学習に使用した [51] や [57] が挙げられる．しかし，これらの手法は依然として，メタデータを使用するアプローチを取っている．はじめに（1 章）で述べたように，メタデータを使用するアプローチは様々な問題があり，特に，大規模なデータセットを用いた実験を困難にしている．一方で，音楽と動画を

[*2] 30 秒の試聴用の音楽とそのジャケット画像．その詳細は 4.1.1 項で述べる．

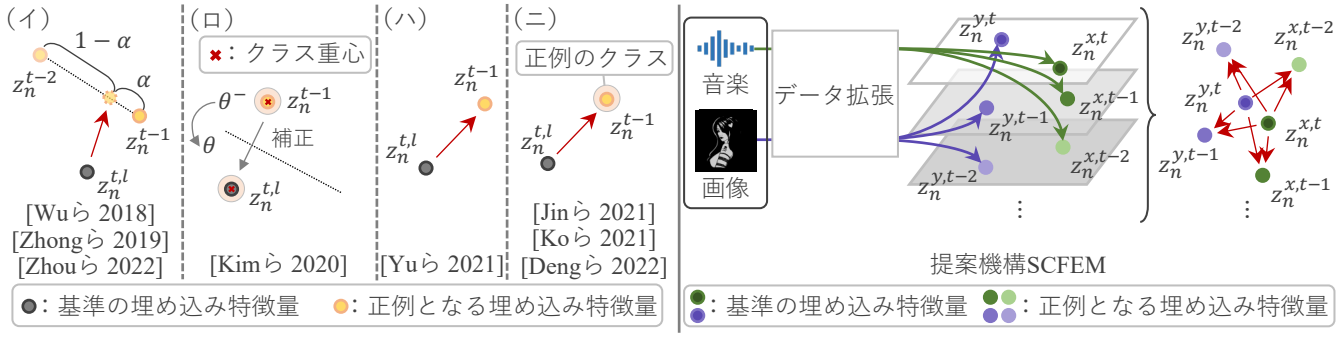

図 3: 正例のペアのみに着目したメモリ機構の図式．（左図）本研究で取り組むタスクにおいて，先行研究のメモリ機構では，正例のペアを多くとも一つしか構成することができない．（右図）それに対して，提案機構である SCFEM [24] を用いることで，複数の正例のペアを構成することが可能となる．我々の SCFEM は，先行研究の XBM [44] に基づいているが，XBM は学習エポックごとにメモリを初期化するため，メモリに保存された埋め込み特徴量から正例のペアを構成することはできない．

対象とした研究では，メタデータを使用しないアプローチが提案されている [10, 29, 38, 52]．Hong らは，類似した音楽（または類似した動画）の埋め込み特徴量が互いに近づくような新たな損失関数を設計し，エンコーダを学習する CBVMR を提案した [10]．Yi らは，変分オートエンコーダを用いて音楽，動画，およびテキストの潜在表現を学習する CMVAE を提案した [52]．Prétet らは，CBVMR で提案された特徴抽出モジュールの有効性を検証するために，そのモジュールを他の代表的なモジュールと入れ替える比較実験を実施した [29]．Surís らは，音楽と動画のための Transformer に基づくエンコーダを提案した [38]．しかし，これらの先行研究において正例のペアを増やすアプローチは検討されていない．本研究では，新たな正例のペアを構成し，増やすことができるメモリ機構を用いて，音楽と画像を双方向に検索するタスクに取り組む．

2.2 メモリ機構

メモリ機構は，エンコーダの学習にとって有益な埋め込み特徴量を再利用することを目的として，学習時の埋め込み特徴量をメモリに保存し，学習イテレーションをまたいで保持する仕組みである [6, 11, 12, 14, 44, 48, 53, 59, 60]．メモリ機構の有効性は，多様なコンピュータビジョンのタスクにおいて実証されてきた [8, 11, 14, 16, 42, 44, 48–50, 53, 59]．さらにメモリ機構は，双方向検索タスク（例えば，ソースコードとバイナリコード [54] や RGB 画像と赤外線画像 [19]，料理画像と調理レシピ [34]）に取り組んだ研究においてもその活用例が報告されている．本研究では，音楽と画像の双方向検索タスクにおいて，メモリ機構の有効性を検証する．

これまでに提案されてきたメモリ機構は，メモリに保存された埋め込み特徴量の取り扱い方法によって，次の四種類に分類することができる．図 3 の（イ）–（ニ）に，それぞれの方法の図式を示す．

（イ）それぞれの埋め込み特徴量の移動平均を計算し，それを新たな埋め込み特徴量として用いる方法 [48, 59, 60]
（ロ）同じクラス（カテゴリ）に属する複数の埋め込み特徴量から重心を計算し，その計算結果とエンコーダの現在のネットワークパラメータを用いた際の同クラスに属する埋め込み特徴量との差を埋めるように学習する方法 [12]
（ハ）それぞれの埋め込み特徴量を直接用いる方法 [44, 53]
（ニ）同じクラス（カテゴリ）に属する複数の埋め込み特徴量から代表値（平均，重心など）を計算し，それを新たな埋め込み特徴量として用いる方法 [6, 11, 14]

しかしこれらのメモリ機構は，分類タスクのように，複数の正例のペアが元々存在するタスクへの適用を前提としているため，正例のペアを増やすことを目的としていない．そのため本研究で取り組むタスクにおいて，先行研究のメモリ機構 [44, 48, 53, 59, 60] をそのまま利用するメリットは小さく，クラスを用いるメモリ機構 [6, 11, 12, 14] は利用できない．また，これらのメモリ機構は画像のみといった，一種類のデータのみで利用されてきたため，複数種類のデータへの利用は検証されていない．そこで本研究では，図 3 の右に示すように，音楽と画像の二種類の埋め込み特徴量をメモリに保存し，さらにその二種類の埋め込み特徴量を最大限活用することで，正例のペアを増やすメモリ機構を提案する．

3. 提案手法

本章では，深層距離学習に基づく提案手法について述べる．まず，本研究が取り組むタスクを定式化する（3.1 節）．次に，エンコーダを学習するためのフレームワークについて述べる（3.2 節）．最後に学習したエンコーダを用いて，音楽と画像を双方向に検索する手法について述べる（3.3 節）．

3.1 タスクの定式化

我々は音楽エンコーダの入力として，先行研究 [20,45,58] に倣い，音楽の複素スペクトログラムを用いる．また，画像エンコーダの入力として RGB 画像を用いる．$\mathbf{X} = \{\mathbf{x}_n \in \mathbb{R}^{D^{\mathbf{x}}}\}_{n=1}^{N}$ を複素スペクトログラムの集合，$\mathbf{Y} = \{\mathbf{y}_n \in \mathbb{R}^{D^{\mathbf{y}}}\}_{n=1}^{N}$ を RGB 画像の集合とする．$\mathbf{x}$ と $\mathbf{y}$ に付された添字はそれぞれ対応しており，添字が等しい場合，つまり $\mathbf{x}_n$ と $\mathbf{y}_n$ は，音楽とその代表画像を表している．また，$D^{\mathbf{x}}$ は複素スペクトログラムの次元数，$D^{\mathbf{y}}$ は RGB 画像の次元数である．N はデータセットに含まれる音楽と画像のペアの件数である．

次に，$\mathbf{Z}^{\mathbf{X}} = \{\mathbf{z}_n^{\mathbf{x}} \in \mathbb{R}^{D^{\mathbf{z}}}\}_{n=1}^{N}$ を複素スペクトログラムの埋め込み特徴量の集合，$\mathbf{Z}^{\mathbf{Y}} = \{\mathbf{z}_n^{\mathbf{y}} \in \mathbb{R}^{D^{\mathbf{z}}}\}_{n=1}^{N}$ を RGB 画像の埋め込み特徴量の集合とする．ここで，$D^{\mathbf{z}}$ はそれぞれの埋め込み特徴量の次元数である．また，$D^{\mathbf{z}}$ 次元のユークリッド空間 $\mathcal{S}$ を音楽と画像の共有埋め込み空間と呼ぶ．

我々は，$\mathbf{z}_n^{\mathbf{x}}$ と $\mathbf{z}_n^{\mathbf{y}}$ が共有埋め込み空間 $\mathcal{S}$ において互いに近づくように，$\mathbf{X}$ を $\mathbf{Z}^{\mathbf{X}}$ に埋め込む音楽エンコーダ $f_{\mathrm{M}}(\cdot;\boldsymbol{\theta})$ と，$\mathbf{Y}$ を $\mathbf{Z}^{\mathbf{Y}}$ に埋め込む画像エンコーダ $f_{\mathrm{I}}(\cdot;\boldsymbol{\phi})$ を学習する（つまり，$\mathbf{x}_n \overset{f_{\mathrm{M}}}{\longmapsto} \mathbf{z}_n^{\mathbf{x}}$ と $\mathbf{y}_n \overset{f_{\mathrm{I}}}{\longmapsto} \mathbf{z}_n^{\mathbf{y}}$ を学習する）．ここで $\boldsymbol{\theta}$ と $\boldsymbol{\phi}$ は，それぞれのエンコーダのネットワークパラメータである．

3.2 学習フレームワーク

まず，深層距離学習の基本的な学習フレームワークについて述べる（3.2.1 項）．そして，我々が提案するメモリ機構である SCFEM [24] について述べる（3.2.2 項）．提案機構の図式を図 4 に示す．

3.2.1 深層距離学習を用いた音楽と画像の埋め込み

深層距離学習は，正例のペアを互いに近づけ，負例のペアを互いに遠ざける学習手法である．実際に，この学習手法は音楽と動画を双方向に検索するタスク [10,29,38,52] において利用されている．

この深層距離学習について，エンコーダの学習で使用される一般の損失関数 $\mathcal{L}(\boldsymbol{B})$ を解析する GPW フレームワーク [43] が提案されている．GPW フレームワークでは次式のように，損失関数を解析するための式 $\mathcal{F}(\boldsymbol{B})$ を提供している：

$$\begin{aligned}\mathcal{F}(\boldsymbol{B}) &= \frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{m}\left.\frac{\partial\mathcal{L}(\boldsymbol{B})}{\partial B_{ij}}\right|_{l} B_{ij} \\ &= \frac{1}{m}\sum_{i=1}^{m}\left(\sum_{(\mathbf{x}_i,\mathbf{y}_j)\in\mathcal{N}} w_{ij}^{B} B_{ij} - \sum_{(\mathbf{x}_i,\mathbf{y}_j)\in\mathcal{P}} w_{ij}^{B} B_{ij}\right). \quad (1)\end{aligned}$$

ここで，m はミニバッチ数，$\mathcal{P}$ と $\mathcal{N}$ はそれぞれ正例のペアの集合および負例のペアの集合，$w_{ij}^{B} = \left|\left.\frac{\partial\mathcal{L}(\boldsymbol{B})}{\partial B_{ij}}\right|_{l}\right|$ は l 番目の学習イテレーションにおける重み，そして $\boldsymbol{B}$ は類似度行列であり，その行列の (i,j) 番目の要素は $\mathbf{z}_i^{\mathbf{x}}$ と $\mathbf{z}_j^{\mathbf{y}}$ のコサイン類似度（つまり，$B_{ij} = \mathrm{sim}(\mathbf{z}_i^{\mathbf{x}}, \mathbf{z}_j^{\mathbf{y}}) = \mathbf{z}_i^{\mathbf{x}\top}\mathbf{z}_j^{\mathbf{y}} / |\mathbf{z}_i^{\mathbf{x}}||\mathbf{z}_j^{\mathbf{y}}|$）

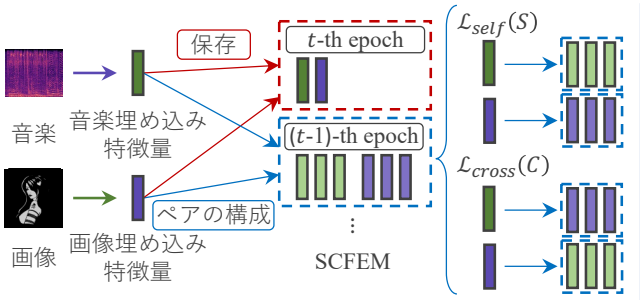

図 4: 任意の音楽（または画像）を基準にして，メモリに保存された埋め込み特徴量とペアを構成する仕組みの図式．次の学習イテレーションに進む前に，音楽（または画像）の埋め込み特徴量をメモリに保存する．また，その埋め込み特徴量を基準にして，メモリに保存された音楽と画像の埋め込み特徴量とのペアを構成し，それらのペアの損失関数を計算する．

で定義される．式 (1) は，エンコーダのネットワークパラメータを最適化する際にどのような要素が関係するかを表している．つまりエンコーダの学習にとって重要なことは，構成し得るペアの数を制御するミニバッチ数 m，B_{ij} に割り当てられた重み w_{ij}^{B}，そしてエンコーダの学習に有益な正例のペアの集合 $\mathcal{P}$ および負例のペアの集合 $\mathcal{N}$ を適切に設定することであるとわかる．

本研究では音楽と画像の二種類のデータを取り扱うため，先行研究 [10,29,38,52] に倣い，二つのタイプのペア（任意の音楽を基準にしたペア，および任意の画像を基準にしたペア）を構成する．この二つのタイプのペアを用いる場合，式 (1) は次のように書き換えることができる：

$$\begin{aligned}\mathcal{F}(\boldsymbol{B}) = \frac{1}{m}\sum_{i=1}^{m}&\left(\sum_{(\mathbf{x}_i,\mathbf{y}_j)\in\mathcal{N}} w_{ij}^{B} B_{ij} + \sum_{(\mathbf{y}_i,\mathbf{x}_j)\in\mathcal{N}} \hat{w}_{ij}^{B} \hat{B}_{ij}\right) \\ &- \frac{1}{m}\sum_{i=1}^{m}\left(\sum_{(\mathbf{x}_i,\mathbf{y}_j)\in\mathcal{P}} w_{ij}^{B} B_{ij} + \sum_{(\mathbf{y}_i,\mathbf{x}_j)\in\mathcal{P}} \hat{w}_{ij}^{B} \hat{B}_{ij}\right). \quad (2)\end{aligned}$$

ここで，$\hat{w}_{ij}^{B} = \left|\left.\frac{\partial\mathcal{L}(\boldsymbol{B})}{\partial \hat{B}_{ij}}\right|_{l}\right|$，$\hat{B}_{ij} = \mathrm{sim}(\mathbf{z}_i^{\mathbf{y}}, \mathbf{z}_j^{\mathbf{x}})$ である．本研究では，ペアの集合 $\mathcal{P}$，$\mathcal{N}$ に着目することで，エンコーダのネットワークパラメータを最適化することを目指す．

我々は損失関数として，先行研究 [38] に倣い，InfoNCE [26] を利用する．InfoNCE は，任意の音楽（または画像）を基準にして，m 個の画像（または音楽）から一個の正例のペアと $m-1$ 個の負例のペアを構成する．InfoNCE を用いた損失関数 $\mathcal{L}_{batch}(\boldsymbol{B})$ は次式で表される：

$$\begin{aligned}\mathcal{L}_{batch}(\boldsymbol{B}) = &-\frac{1}{m}\sum_{i=1}^{m}\log\frac{e^{B_{i+}/\tau}}{\sum_{j=1}^{m} e^{B_{ij}/\tau}} \\ &-\frac{1}{m}\sum_{i=1}^{m}\log\frac{e^{\hat{B}_{i+}/\tau}}{\sum_{j=1}^{m} e^{\hat{B}_{ij}/\tau}}. \quad (3)\end{aligned}$$

ここで 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り，式 (3) 中の $+$ は，任意の音楽（または画像）を基準にした際，その音楽（または画像）と正例のペアとなり得る画像（または音楽）を表す．式 (3) の右辺の各項は，それぞれ式 (2) で考慮した二つのタイプのペアを表す．また，InfoNCE を式 (2) を用いて解析すると，重み w_{ij}^B を次式のように導出できる：

$$w_{ij}^B = \begin{cases} \frac{1}{\tau} - \chi_{ij}^B & ((\mathbf{x}_i, \mathbf{y}_j) \in \mathcal{P}), \\ \chi_{ij}^B & ((\mathbf{x}_i, \mathbf{y}_j) \in \mathcal{N}). \end{cases} \tag{4}$$

ここで，$\chi_{ij}^B = e^{B_{ij}/\tau} / \{\tau(e^{B_{i+}/\tau} + \sum_{(\mathbf{x}_i,\mathbf{y}_j)\in\mathcal{N}} e^{B_{ij}/\tau})\}$ である．重み $\hat{w}_{ij}^B$ は w_{ij}^B と同様に導出できる．

エンコーダの最適なネットワークパラメータ $\boldsymbol{\theta}^*$，$\boldsymbol{\phi}^*$ は，次式の損失関数 $\mathcal{L}_{batch}$ を最小化することで推定できる：

$$\boldsymbol{\theta}^*, \boldsymbol{\phi}^* = \arg\min_{\boldsymbol{\theta},\boldsymbol{\phi}} \mathcal{L}_{batch}. \tag{5}$$

InfoNCE は，前述のように一個の正例のペアと $m-1$ 個の負例のペアで構成できる損失関数であるため，本研究で取り組むタスクのように，正例のペアが非常に少ないケースにおいて有効である．一方で，正例のペアを一個のみ使用するため，正例のペアについて，式 (4) の重みが小さくなる問題がある．つまり，InfoNCE は本質的に，ランダムなペアを互いに遠ざけるように機能する損失関数である．この問題を解決するために，本研究では正例のペアを増やすことが可能なメモリ機構 SCFEM [24] を提案する．

3.2.2 SCFEM

我々は，任意の音楽（または画像）の埋め込み特徴量が学習中に大きく変化しない “slow drift” 現象 [44] を利用して，新しいメモリ機構である SCFEM [24] を提案する．提案機構は，深層距離学習の学習フレームワークにおけるモジュールとして簡単に組み込むことができる．また提案機構を用いてエンコーダを学習する際，ミニバッチ数よりもはるかに多いデータを少ない計算資源で実行することが可能となる．

まず，$\boldsymbol{M}^{\mathbf{x}}, \boldsymbol{M}^{\mathbf{y}} \in \mathbb{R}^{N\times D^{\mathbf{z}}\times E}$ をそれぞれ音楽メモリおよび画像メモリと呼ぶ．ここで E は，メモリに保存できる学習エポック数である．提案機構では，学習の始めにメモリ $\boldsymbol{M}^{\mathbf{x}}$ と $\boldsymbol{M}^{\mathbf{y}}$ を初期化する．このメモリ機構は，エンコーダの事前学習が完了した段階で使い始める．次の学習イテレーションに進む前に，エンコーダを用いて得られた音楽と画像の埋め込み特徴量をそれぞれメモリ $\boldsymbol{M}^{\mathbf{x}}$ と $\boldsymbol{M}^{\mathbf{y}}$ に保存する．メモリに保存できる埋め込み特徴量の最大数を超えるタイミングがきたら，メモリに保存したタイミングが最も早い埋め込み特徴量を，新たに得られた埋め込み特徴量といれかえる．

提案機構の重要な側面として，任意の音楽（または画像）について，音楽メモリ $\boldsymbol{M}^{\mathbf{x}}$ および画像メモリ $\boldsymbol{M}^{\mathbf{y}}$ の両方に，その音楽（または画像）と正例のペアとなり得る埋め込み特徴量が保存されていることが挙げられる．例えば，任意の音楽を基準にすると，その代表画像の埋め込み特徴量だけでなく，その音楽の過去の埋め込み特徴量とのペアも正例のペアとみなすことができる．音楽の埋め込み特徴量と音楽メモリ $\boldsymbol{M}^{\mathbf{x}}$ に保存された音楽の埋め込み特徴量で構成されるペア，および画像の埋め込み特徴量と画像メモリ $\boldsymbol{M}^{\mathbf{y}}$ に保存された画像の埋め込み特徴量で構成されるペアについての損失関数を $\mathcal{L}_{self}$ とすると，損失関数 $\mathcal{L}_{self}$ は式 (3) を基に次のように書ける：

$$\begin{aligned} \mathcal{L}_{self}(\boldsymbol{S}) = &- \frac{1}{m}\sum_{i=1}^{m}\sum_{e=0}^{E-1} \log \frac{w_e e^{S_{i+}^{\mathbf{x}}/\tau}}{\sum_{j=1}^{N} e^{S_{ij}^{\mathbf{x}}/\tau}} \\ &- \frac{1}{m}\sum_{i=1}^{m}\sum_{e=0}^{E-1} \log \frac{w_e e^{S_{i+}^{\mathbf{y}}/\tau}}{\sum_{j=1}^{N} e^{S_{ij}^{\mathbf{y}}/\tau}}. \end{aligned} \tag{6}$$

ここで，$\boldsymbol{S}$ は類似度行列であり，その行列の (i,j) 番目の要素は，ある学習イテレーションにおける埋め込み特徴量と，メモリに保存された埋め込み特徴量のコサイン類似度（つまり，$S_{ij}^{\mathbf{x}} = \mathrm{sim}(\mathbf{z}_i^{\mathbf{x},t}, \mathbf{z}_j^{\mathbf{x},t-e} \in \boldsymbol{M}^{\mathbf{x}})$ と $S_{ij}^{\mathbf{y}} = \mathrm{sim}(\mathbf{z}_i^{\mathbf{y},t}, \mathbf{z}_j^{\mathbf{y},t-e} \in \boldsymbol{M}^{\mathbf{y}})$）で定義される．また，$t$ は現在の学習エポック数で，$e \in \{0, 1, \cdots, E-1\}$ は現在の学習エポック数 t と埋め込み特徴量をメモリに保存した際の学習エポック数の差を表す．そして，$\{w_e\}_{e=0}^{E-1}$ は重みの集合である．同様に，音楽の埋め込み特徴量と画像メモリ $\boldsymbol{M}^{\mathbf{y}}$ に保存された画像の埋め込み特徴量で構成されるペア，および画像の埋め込み特徴量と音楽メモリ $\boldsymbol{M}^{\mathbf{x}}$ に保存された音楽の埋め込み特徴量で構成されるペアについての損失関数を $\mathcal{L}_{cross}$ とすると，損失関数 $\mathcal{L}_{cross}$ は次のように書ける：

$$\begin{aligned} \mathcal{L}_{cross}(\boldsymbol{C}) = &- \frac{1}{m}\sum_{i=1}^{m}\sum_{e=0}^{E-1} \log \frac{w_e e^{C_{i+}/\tau}}{\sum_{j=1}^{N} e^{C_{ij}/\tau}} \\ &- \frac{1}{m}\sum_{i=1}^{m}\sum_{e=0}^{E-1} \log \frac{w_e e^{\hat{C}_{i+}/\tau}}{\sum_{j=1}^{N} e^{\hat{C}_{ij}/\tau}}. \end{aligned} \tag{7}$$

ここで，$\boldsymbol{C}$ は類似度行列であり，その行列の (i,j) 番目の要素は，ある学習イテレーションにおける埋め込み特徴量と，メモリに保存された埋め込み特徴量のコサイン類似度（つまり，$C_{ij} = \mathrm{sim}(\mathbf{z}_i^{\mathbf{x},t}, \mathbf{z}_j^{\mathbf{y},t-e} \in \boldsymbol{M}^{\mathbf{y}})$ と $\hat{C}_{ij} = \mathrm{sim}(\mathbf{z}_i^{\mathbf{y},t}, \mathbf{z}_j^{\mathbf{x},t-e} \in \boldsymbol{M}^{\mathbf{x}})$）で定義される．損失関数 $\mathcal{L}_{self}$ と $\mathcal{L}_{cross}$ について，GPW フレームワークを用いた詳細な分析は付録（A.1 節）に示す．

最後に，$\mathcal{L}_{batch}$ と $\mathcal{L}_{self}$，$\mathcal{L}_{cross}$ を組み合わせて，エンコーダの最適なネットワークパラメータ $\boldsymbol{\theta}^*$， $\boldsymbol{\phi}^*$ を次式で求めることができる：

$$\boldsymbol{\theta}^*, \boldsymbol{\phi}^* = \arg\min_{\boldsymbol{\theta},\boldsymbol{\phi}} \left(\mathcal{L}_{batch} + \lambda_{self}\mathcal{L}_{self} + \lambda_{cross}\mathcal{L}_{cross}\right). \tag{8}$$

ここで，λ_{self} と λ_{cross} は，損失関数にかかる重みである．

3.3 音楽と画像の双方向検索

エンコーダの学習が完了したら，音楽と画像の任意のペアについて，それらの類似度を次の手順で推定できる．

(1) 音楽の複素スペクトログラムを計算する．
(2) 学習したエンコーダを用いて音楽と画像の埋め込み特徴量を得る．
(3) 得られた埋め込み特徴量の類似度を計算する．

類似度が高いほど，その音楽と画像のペアは合っているとみなすことができる．

4. 実験と結果

本章では，提案機構の有効性を検証するための実験方法とその結果について述べる．我々は，音楽クエリを用いた画像検索タスクおよび画像クエリを用いた音楽検索タスクの二つのタスクについて，検索タスクにおける性能評価指に基づく比較実験を実施した．また，学習によって得られた埋め込み特徴量の定性分析を実施した．

4.1 実験設定

4.1.1 データセット

我々は実験のために，新たに MCA データセットを構築した．MCA データセットは，40,151 名（グループ）のアーティストの，78,325 件の音楽と画像のペアで構成されている．このデータセットは 250 以上の音楽ジャンルを含んでいる．音楽は試聴用音楽であり，それぞれの長さは 30 秒，サンプリングレートは 44.1 kHz である．画像はジャケット画像であり，正方形の RGB 画像である．MCA データセットの構築方法は，先行研究 [1, 38, 52] と同様に，インターネット上の音楽サービスが公開している試聴用音楽を収集した．したがって試聴用音楽は，元の（フル尺の）音楽からその音楽サービスが試聴用に選択した区間である．また，試聴用音楽を収集する際に，同時にそのジャケット画像を取得した．収集対象の音楽は，アルバム曲ではなくシングル曲に限定した．つまり音楽と画像の正例のペアは，それぞれについてただ一つのみである．我々は MCA データセットを，訓練データセット（62,659 件），検証データセット（7,833 件），およびテストデータセット（7,833 件）の三つに分割し，実験を実施した．

4.1.2 実装詳細

4.1.2.1 音楽情報処理

音楽の複素スペクトログラムは，nnAudio [3] を用いて，STFT [7] を計算して求めた．その計算の際，出力する周波数ビンの数 F は 1,025 とした．また Hann 窓を用い，その窓幅は 512 とした．そして，複素スペクトログラムから $2 \times F \times 256$ 次元のテンソルを切り出して使用した．これはおよそ 3 秒の長さに相当する．音楽エンコーダは，切り出した複素スペクトログラムを 256 次元の共有埋め込み空間に埋め込むように設計した．音楽エンコーダを学習する際，音楽のデータ拡張として複素スペクトログラムから切り出す箇所をランダムに選んだ．テスト時には，音楽全体の埋め込み特徴量を求めるために，その音楽の複素スペクトログラムから複数のテンソルを切り出し，それらの埋め込み特徴量の平均を用いた．具体的には，複素スペクトログラムの先頭から $2 \times F \times 256$ 次元のテンソルを，窓幅の半分 (つまり，128) をずらしながら繰り返し切り出し，それらのテンソルの埋め込み特徴量の平均を求めて使用した．

4.1.2.2 画像情報処理

画像は 256 px × 256 px の大きさとなるように拡大・縮小した．画像エンコーダは，拡大・縮小した画像を 256 次元の共有埋め込み空間に埋め込むように設計した．画像エンコーダを学習する際，画像のデータ拡張としてアフィン変換を適用した．具体的には，回転（$[-25°, 25°]$），並行移動（$[0.15, 0.15]$），拡大・縮小（$[0.75, 1.25]$）をランダムに適用した．

4.1.2.3 エンコーダの設計

我々は，HRFormer [55] をバックボーンネットワークとして利用した．HRFormer を利用する際，その最終層を，クラス確率を出力する層から，埋め込み層に変更した．

4.1.2.4 学習条件

我々は，PyTorch [28] を用いて実装した．学習を始める前に，エンコーダのネットワークパラメータは初期化した．エンコーダのネットワークパラメータ最適化には，Adam [13] を用いた．その学習率は，1.0×10^{-4} と設定した．損失関数の重みを，$\mathcal{L}_{self}$ と $\mathcal{L}_{cross}$ の値の大きさが等しくなるように，それぞれ経験的に $\lambda_{self} = 0.3$ および $\lambda_{cross} = 0.2$ と設定した．温度スケーリング τ の値は，MoCo [26] に倣い，0.07 と設定した．メモリ機構を用いる際は，それを用いる前にエンコーダを学習する必要がある．先行研究 [44] に倣い，3.2.1 節の式 (3) のみを用いて，エンコーダを事前に 5 万回以上のイテレーションで学習した．8 台の NVIDIA A100 を用いて，三日間エンコーダを学習し続けた．

4.1.3 ランキングに基づく評価指標

我々は，検索タスクにおいて標準的な性能評価指標である平均逆順位 [4]・再現率・中央順位 [38] を用いて比較実験を実施した．平均逆順位を MRR，再現率を R@k，中央順位を MR とそれぞれ表記する．MRR および R@k は値が高いほど，MR は値が低いほど性能が良いことを表す．また R@k は，音楽（または画像）クエリを用いた検索結果の上位 k 番目までにその代表画像（または音楽）が含まれていれば正解と判定したときに，テストデータセットのすべてのクエリにおける正解判定の割合を表しており，図表において百分率で表記する．

4.2 実験条件

提案機構の有効性を検証するために，我々は次に示す手

表 1: MCA データセットのテストデータセットを用いた比較実験における各評価指標（MRR，R@k，および MR）の値．

	音楽クエリを用いた画像検索				画像クエリを用いた音楽検索			
	MRR	R@50	R@100	MR	MRR	R@50	R@100	MR
ランダム	1.22×10^{-3}	0.64	1.28	3917	1.22×10^{-3}	0.64	1.28	3917
CBVMR [10]	1.34×10^{-3}	0.75	1.52	3686	1.27×10^{-3}	0.61	1.39	3656
ベースライン [55]	3.37×10^{-3}	2.09	4.05	1957	3.42×10^{-3}	2.08	4.06	1926
+ データ拡張	3.82×10^{-3}	2.84	5.57	1614	3.81×10^{-3}	2.56	5.22	1626
+XBM [44]	4.23×10^{-3}	2.78	5.86	1594	5.04×10^{-3}	3.22	6.09	1600
+SCFEM（提案機構）	$\mathbf{1.14 \times 10^{-2}}$	**7.45**	**12.3**	**1066**	$\mathbf{9.75 \times 10^{-3}}$	**7.06**	**11.8**	**1059**

法と比較した．

- **ベースライン：** HRFormer [55] を音楽エンコーダおよび画像エンコーダのバックボーンネットワークとして使用した．ただし，データ拡張およびメモリ機構は使用しなかった．
- **ベースライン + データ拡張：** ベースラインに加え，4.1.2 節で述べたデータ拡張を使用した．ただし，メモリ機構は使用しなかった．
- **ベースライン + データ拡張 +XBM [44]：**ベースライン + データ拡張に加え，メモリ機構として XBM [44] を使用した．本研究において XBM は，SCFEM において $E = 1$ とした場合と等価である．
- **ベースライン + データ拡張 +SCFEM：**ベースライン + データ拡張に加え，提案機構である SCFEM を使用した．この比較実験において，SCFEM のパラメータを $E = 2$ および $w_0 = w_1 = 1.0$ と設定した．

また参考までに，次に示す手法の結果を併記する．

- **ランダム：**クエリに対してランダムな検索結果を返した．
- **CBVMR：**本研究のタスクに関連のある CBVMR [10] を使用した．ただし，CBVMR は音楽と動画の双方向検索を目的とした手法のため，特に画像情報処理の方法が提案手法と異なる．本実験では，動画レベル特徴量のかわりに，同 [10] で用いられている，画像フレームレベルの特徴量を主成分分析し，それらを白色化した値を用いた．

4.3 結果

表 1 に，音楽クエリを用いた画像検索タスクと，画像クエリを用いた音楽検索タスクのそれぞれにおける，MRR，R@k，および MR の値を示す．提案機構はベースライン手法と比較して，音楽クエリを用いた画像検索タスクにおいて，MRR について 2.70 ～ 3.38 倍，R@50 について 2.62 ～ 3.56 倍，そして MR について 528 ～ 891 の性能改善を達成した．同様に，画像クエリを用いた音楽検索タスクにおいて，MRR について 1.93 ～ 2.85 倍，R@50 について 2.19 ～ 3.39 倍，そして MR について 541 ～ 867 の性能改善を達成した．図 5 に，それぞれのタスクについて，任意の k における R@k の経験的累積分布関数を示す．この結果から，提案機構はほとんど全ての R@k について，ベースライン手法より優れていることがわかる．したがって，これらの検索タスクにおいて，多くの有益なペアを構成できるようにした提案機構が有効であったこと示している．

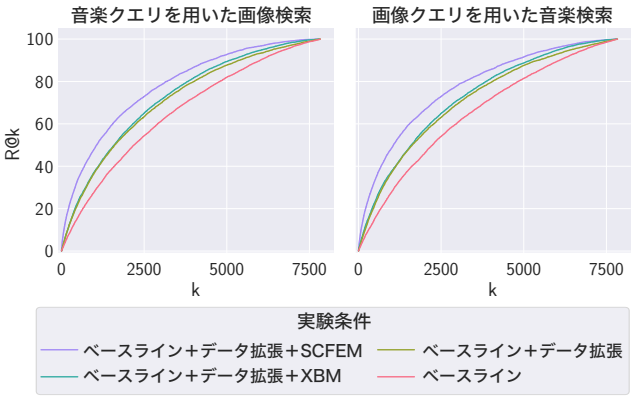

図 5: 任意の k における R @ k の経験的累積分布関数．

さらに，表 1 の結果から，データ拡張を用いた方が検索タスクにおける性能が良いことがわかる．先行研究 [10, 29, 38, 52] では，データ拡張なしで音楽や画像の特徴量を用いているが，この結果はデータ拡張の重要性を示唆している．

4.4 提案手法を構成する各要素の比較実験

我々は，エンコーダのバックボーンネットワークや SCFEM のパラメータといった，提案手法を構成する各要素について比較実験を実施した．また，エンコーダを事前に学習する必要性を検証した．

4.4.1 バックボーンネットワーク

バックボーンネットワークの選択によって，検索タスクにおける性能が大きく変化するため，いくつかの有名なニューラルネットワークモデルをバックボーンネットワークとして用いて比較した．比較実験には，畳み込みニューラルネットワークに基づくモデル [9,35,39] と Transformer に基づくモデル [21,55] を用いた．この実験において，$\tau = 1.0$ を用いた．比較実験の結果を図 6 に示す．この結果から，

表 2: メモリの大きさと重みによる性能比較.

	音楽クエリを用いた画像検索				画像クエリを用いた音楽検索			
	MRR	R@50	R@100	MR	MRR	R@50	R@100	MR
$E = 1$ (XBM [44])	4.23×10^{-3}	2.78	5.86	1594	5.04×10^{-3}	3.22	6.09	1600
$E = 2, w_1 = 1.0$	1.14×10^{-2}	7.45	12.3	1066	9.75×10^{-3}	7.06	11.8	1059
$E = 3, w_1 = w_2 = 0.5$	1.10×10^{-2}	8.00	12.9	1014	9.49×10^{-3}	6.92	12.2	1002
$E = 3, w_1 = 0.6, w_2 = 0.4$	1.13×10^{-2}	**8.04**	12.9	1034	1.05×10^{-2}	**7.50**	12.3	1010
$E = 3, w_1 = 0.7, w_2 = 0.3$	1.11×10^{-2}	7.53	12.7	1014	**1.09** $\times 10^{-2}$	7.49	12.0	**982**
$E = 3, w_1 = 0.8, w_2 = 0.2$	**1.26** $\times 10^{-2}$	7.78	**13.0**	1024	**1.09** $\times 10^{-2}$	7.34	**12.5**	1022
$E = 3, w_1 = 0.9, w_2 = 0.1$	1.10×10^{-2}	7.91	12.9	**1009**	1.07×10^{-2}	7.47	12.2	1010

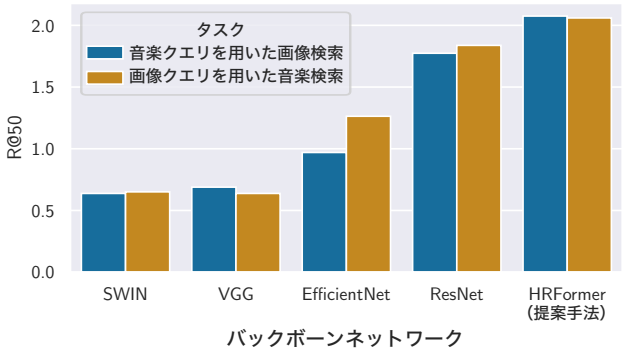

図 6: バックボーンネットワークの性能比較.

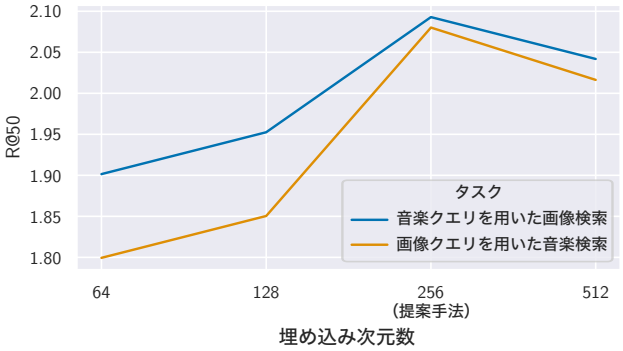

図 7: 埋め込み特徴量の次元数の性能比較.

提案手法においてバックボーンネットワークとして用いた HRFormer [55] が適切な選択であったことがわかる.

4.4.2 メモリの大きさと重み

提案機構である SCFEM は，そのメモリの大きさを柔軟に決めることができる．そのため，メモリの大きさと，損失関数にかかる重みについて比較実験を実施した．全ての実験条件において w_0 は 1.0 で固定した．結果を表 2 に示す．この結果から，メモリの大きさを増やすことで，検索タスクにおける性能をさらに改善することが可能であることを示した．ただし，メモリの大きさを増やした際には，それに合わせて重みを調整する必要がある.

4.4.3 埋め込み特徴量の次元数

埋め込み特徴量の次元数 $D^{\mathbf{z}}$ の違いによる，検索タスクにおける性能への影響を調べるために，$D^{\mathbf{z}}$ がそれぞれ 64, 128, 256, 512 の場合における R@50 の性能を比較した．結果を図 7 に示す．この結果から，4.3 節における比較実験で使用した，$D^{\mathbf{z}}$ が 256 のときが最も良いことがわかる.

4.4.4 エンコーダの事前学習の必要性

エンコーダの学習の初期段階において，そのネットワークパラメータは大きく更新されることがあるため，埋め込み特徴量も大きく変わる可能性がある．そのため，SCFEM を含むメモリ機構を用いる際は，3.2 節で述べたようにエンコーダを事前に学習する必要がある．図 8 に，SCFEM をエンコーダの事前学習なしで使用したケースとベースライン [55] のそれぞれについて，検証データセットにおける損失関数の値の推移を示す．この結果から，SCFEM をエンコーダの事前学習なしで使用したケースでは，ベースラインと比べて，エンコーダの学習の進みが遅いことがわかる．つまり，SCFEM をエンコーダの学習の初期段階から使用すると，その学習に悪影響を与えることを実験的に明らかにした．したがって，SCFEM を用いる際は，エンコーダを事前に学習する必要がある.

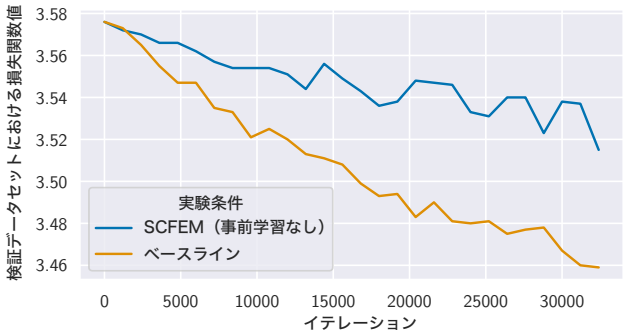

図 8: 検証データセットにおける損失関数の値の推移.

4.5 埋め込み特徴量の定性分析

学習したエンコーダを用いて得られる埋め込み特徴量の性質を調べるために，我々は定性分析を実施した．MCA データセットのテストデータセットの中から，メタル・ジャ

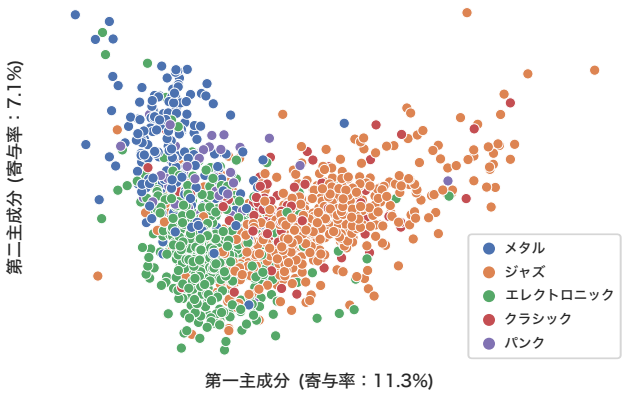

図 9: 音楽と画像の埋め込み特徴量の主成分分析.

ズ・クラシック・エレクトロニック・パンクの 5 種類の音楽ジャンルタグ[*3]が付された 686 件の音楽と画像のペアを選択し，それらの埋め込み特徴量について主成分分析を実施した．その結果を図 9 に示す．主成分分析の結果から，共有埋め込み空間において，同じ音楽ジャンルタグが付された音楽と画像の埋め込み特徴量は，比較的近くに埋め込まれていることがわかる．さらに興味深いことに，メタルとパンクの音楽ジャンルタグが付された音楽と画像の埋め込み特徴量は，互いに近くに埋め込まれていることがわかる．これは，メタルとパンクが，他のジャンルと比べて近い音楽ジャンルであることに起因している．この結果は．1 章で述べたように，音楽とその代表画像が密接に関連していることを裏付けている．

5. 最後に

本稿では，深層距離学習のフレームワークにモジュールとして簡単に組み込むことができる提案機構 SCFEM について述べた．本研究の貢献は次のとおりである．第一に，SCFEM はより多くの埋め込み特徴量をメモリに保存する仕組みを備えており，その仕組みは有益なペアを構成するために活用できる．音楽と画像を双方向に検索するタスクにおいて，SCFEM の有効性を示した．第二に，SCFEM は，検索タスクにおける標準的な性能評価指標において，他の手法と比較して性能が高いことを示した．さらに SCFEM のメモリの大きさを増やすことで，より性能が改善されることを示した．第三に，定性分析によって，音楽と画像の共有埋め込み空間において，同じ音楽ジャンルタグが付された音楽と画像が近くに埋め込まれていることを明らかにした．さらに，メタルやパンクといった，関連のある音楽ジャンルが付された音楽と画像も近くに埋め込まれていることを発見した．

SCFEM の仕組みには汎用性があるため，音楽と画像を双方向に検索するタスクに限らず，様々なタスクへの応用が可能である．我々は，この提案機構が広範なタスクにおいてその有効性が検証されることを期待している．

[*3] 音楽ジャンルタグのようなメタデータは，エンコーダの学習において使用していない．

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] Chao, J., Wang, H., Zhou, W., Zhang, W. and Yu, Y.: TuneSensor: A Semantic-Driven Music Recommendation Service for Digital Photo Albums, *Proceedings of the International Semantic Web Conference (ISWC)* (2011).
[3] Cheuk, K. W., Anderson, H., Agres, K. and Herremans, D.: nnAudio: An on-the-fly GPU Audio to Spectrogram Conversion Toolbox Using 1D Convolutional Neural Networks, *IEEE Access*, Vol. 8, pp. 161981–162003 (2020).
[4] Craswell, N.: *Mean Reciprocal Rank*, p. 1703, Springer US (2009).
[5] Cunningham, S. J. and Nichols, D. M.: How people find videos, *Proceedings of the ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL)*, pp. 201–210 (2008).
[6] Deng, Z., Zhong, Y., Guo, S. and Huang, W.: InsCLR: Improving Instance Retrieval with Self-Supervision, *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pp. 516–524 (2022).
[7] Gabor, D.: Theory of communication, *Journal of the Institution of Electrical Engineers*, Vol. 94, No. 73 (1947).
[8] 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).
[9] He, K., Zhang, X., Ren, S. and Sun, J.: Deep Residual Learning for Image Recognition, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778 (2016).
[10] Hong, S., Im, W. and Yang, H. S.: CBVMR: Content-Based Video-Music Retrieval Using Soft Intra-Modal Structure Constraint, *Proceedings of the ACM International Conference on Multimedia Retrieval (ICMR)*, pp. 353–361 (2018).
[11] Jin, Z., Gong, T., Yu, D., Chu, Q., Wang, J., Wang, C. and Shao, J.: Mining Contextual Information Beyond Image for Semantic Segmentation, *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7231–7241 (2021).
[12] Kim, Y., Park, W. and Shin, J.: BroadFace: Looking at Tens of Thousands of People at Once for Face Recognition, *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 536–552 (2020).
[13] Kingma, D. P. and Ba, J.: Adam: A Method for Stochastic Optimization, *Proceedings of the International Conference for Learning Representations (ICLR)*, pp. 1–13 (2015).
[14] Ko, B., Gu, G. and Kim, H.-G.: Learning with Memory-based Virtual Classes for Deep Metric Learning, *Pro-*

ceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 11792–11801 (2021).

[15] Li, B. and Kumar, A.: Query by Video: Cross-Modal Music Retrieval, *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pp. 604–611 (2019).

[16] Li, S., Chen, D., Liu, B., Yu, N. and Zhao, R.: Memory-Based Neighbourhood Embedding for Visual Recognition, *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6102–6111 (2019).

[17] Libeks, J. and Turnbull, D.: You Can Judge an Artist by an Album Cover: Using Images for Music Annotation, *IEEE MultiMedia*, Vol. 18, No. 4, pp. 30–37 (2011).

[18] Liu, C.-L. and Chen, Y.-C.: Background Music Recommendation Based on Latent Factors and Moods, *Knowledge-Based Systems*, Vol. 159, pp. 158–170 (2018).

[19] Liu, J., Sun, Y., Zhu, F., Pei, H., Yang, Y. and Li, W.: Learning Memory-Augmented Unidirectional Metrics for Cross-Modality Person Re-Identification, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19366–19375 (2022).

[20] Liu, Y., Zhang, H., Zhang, X. and Yang, L.: Supervised Speech Enhancement with Real Spectrum Approximation, *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5746–5750 (2019).

[21] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S. and Guo, B.: Swin Transformer: Hierarchical Vision Transformer using Shifted Windows, *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10012–10022 (2021).

[22] Mazumder, P., Singh, P., Parida, K. K. and Namboodiri, V. P.: Avgzslnet: Audio-visual generalized zero-shot learning by reconstructing label features from multi-modal embeddings, *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 3090–3099 (2021).

[23] Mercea, O.-B., Riesch, L., Koepke, A. and Akata, Z.: Audio-visual Generalised Zero-shot Learning with Cross-modal Attention and Language, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10553–10563 (2022).

[24] Nakatsuka, T., Hamasaki, M. and Goto, M.: Content-Based Music-Image Retrieval Using Self-and Cross-Modal Feature Embedding Memory, *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 2174–2184 (2023).

[25] Negus, K.: *Producing pop: Culture and conflict in the popular music industry*, Edward Arnold (2011).

[26] Oord, A. v. d., Li, Y. and Vinyals, O.: Representation Learning with Contrastive Predictive Coding, *arXiv preprint arXiv:1807.03748* (2018).

[27] Parida, K., Matiyali, N., Guha, T. and Sharma, G.: Coordinated joint multimodal embeddings for generalized audio-visual zero-shot classification and retrieval of videos, *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 3251–3260 (2020).

[28] 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 Advances in Neural Information Processing Systems (NeurIPS)*, pp. 8024–8035 (2019).

[29] 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).

[30] Sasaki, S., Hirai, T., Ohya, H. and Morishima, S.: Affective Music Recommendation System Reflecting the Mood of Input Image, *Proceedings of the International Conference on Culture and Computing (ICCC)*, pp. 153–154 (2013).

[31] Sasaki, S., Hirai, T., Ohya, H. and Morishima, S.: Affective Music Recommendation System Based on the Mood of Input Video, *Proceedings of the International Conference on Multimedia Modeling (MMM)*, pp. 299–302 (2015).

[32] Schindler, A. and Rauber, A.: Harnessing Music-Related Visual Stereotypes for Music Information Retrieval, *ACM Transactions on Intelligent Systems and Technology (TIST)*, Vol. 8, No. 2, pp. 1–21 (2016).

[33] Schroff, F., Kalenichenko, D. and Philbin, J.: Facenet: A Unified Embedding for Face Recognition and Clustering, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 815–823 (2015).

[34] Shukor, M., Couairon, G., Grechka, A. and Cord, M.: Transformer Decoders with MultiModal Regularization for Cross-Modal Food Retrieval, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4567–4578 (2022).

[35] Simonyan, K. and Zisserman, A.: Very Deep Convolutional Networks for Large-scale Image Recognition, *Proceedings of the International Conference on Learning Representations (ICLR)*, pp. 1–14 (2015).

[36] Steve, J. and Sorger, M.: Covering Music: A Brief History and Analysis of Album Cover Design, *Journal of Popular Music Studies*, Vol. 11, No. 1, pp. 68–102 (1999).

[37] Surís, D., Duarte, A., Salvador, A., Torres, J. and Giró-i Nieto, X.: Cross-Modal Embeddings for Video and Audio Retrieval, *Proceedings of the European Conference on Computer Vision Workshops (ECCVW)* (2018).

[38] Surís, D., Vondrick, C., Russell, B. and Salamon, J.: It's Time for Artistic Correspondence in Music and Video, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10564–10574 (2022).

[39] Tan, M. and Le, Q.: EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks, *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 6105–6114 (2019).

[40] Vad, M.: The Album Cover, *Journal of Popular Music Studies*, Vol. 33, No. 3, pp. 11–15 (2021).

[41] Van den Oord, A., Dieleman, S. and Schrauwen, B.: Deep Content-Based Music Recommendation, *Proceedings of the Advances in Neural Information Processing Systems (NIPS)*, pp. 2643–2651 (2013).

[42] Vinyals, O., Blundell, C., Lillicrap, T., Wierstra, D. et al.: Matching Networks for One Shot Learning, *Proceedings of the Advances in Neural Information Processing Systems (NIPS)*, pp. 3630–3638 (2016).

[43] Wang, X., Han, X., Huang, W., Dong, D. and Scott, M. R.: Multi-Similarity Loss with General Pair Weighting for Deep Metric Learning, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5022–5030 (2019).

[44] Wang, X., Zhang, H., Huang, W. and Scott, M. R.:

Cross-Batch Memory for Embedding Learning, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6388–6397 (2020).
[45] Wang, Y. and Wang, D.: A Deep Neural Network for Time-Domain Signal Reconstruction, *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4390–4394 (2015).
[46] Wu, C.-Y., Manmatha, R., Smola, A. J. and Krahenbuhl, P.: Sampling matters in deep embedding learning, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2840–2848 (2017).
[47] Wu, X., Qiao, Y., Wang, X. and Tang, X.: Bridging Music and Image via Cross-Modal Ranking Analysis, *IEEE Transactions on Multimedia (TOM)*, Vol. 18, No. 7, pp. 1305–1318 (2016).
[48] Wu, Z., Efros, A. A. and Yu, S. X.: Improving Generalization via Scalable Neighborhood Component Analysis, *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 685–701 (2018).
[49] Wu, Z., Xiong, Y., Yu, S. X. and Lin, D.: Unsupervised Feature Learning via Non-Parametric Instance Discrimination, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3733–3742 (2018).
[50] Xiao, T., Li, S., Wang, B., Lin, L. and Wang, X.: Joint Detection and Identification Feature Learning for Person Search, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3415–3424 (2017).
[51] Xing, B., Zhang, K., Zhang, L., Wu, X., Dou, J. and Sun, S.: Image–Music Synesthesia-Aware Learning Based on Emotional Similarity Recognition, *IEEE Access*, Vol. 7, pp. 136378–136390 (2019).
[52] Yi, J., Zhu, Y., Xie, J. and Chen, Z.: Cross-modal Variational Auto-encoder for Content-based Micro-video Background Music Recommendation, *IEEE Transactions on Multimedia (TOM)* (2021).
[53] Yu, S., Wang, C., Mao, Q., Li, Y. and Wu, J.: Cross-Epoch Learning for Weakly Supervised Anomaly Detection in Surveillance Videos, *IEEE Signal Processing Letters*, Vol. 28, pp. 2137–2141 (2021).
[54] Yu, Z., Zheng, W., Wang, J., Tang, Q., Nie, S. and Wu, S.: CodeCMR: Cross-modal retrieval for function-level binary source code matching, *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, pp. 3872–3883 (2020).
[55] Yuan, Y., Fu, R., Huang, L., Lin, W., Zhang, C., Chen, X. and Wang, J.: HRFormer: High-Resolution Transformer for Dense Prediction, *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)* (2021).
[56] Zeng, D., Yu, Y. and Oyama, K.: Audio-Visual Embedding for Cross-Modal Music Video Retrieval through Supervised Deep CCA, *Proceedings of the IEEE International Symposium on Multimedia (ISM)*, pp. 143–150 (2018).
[57] Zhao, S., Li, Y., Yao, X., Nie, W., Xu, P., Yang, J. and Keutzer, K.: Emotion-Based End-to-End Matching Between Image and Music in Valence-Arousal Space, *Proceedings of the ACM International Conference on Multimedia (MM)*, pp. 2945–2954 (2020).
[58] Zheng, N. and Zhang, X.-L.: Phase-Aware Speech Enhancement Based on Deep Neural Networks, *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, Vol. 27, No. 1, pp. 63–76 (2018).
[59] Zhong, Z., Zheng, L., Luo, Z., Li, S. and Yang, Y.: Invariance Matters: Exemplar Memory for Domain Adaptive Person Re-identification, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 598–607 (2019).
[60] Zhou, T., Zhang, M., Zhao, F. and Li, J.: Regional semantic contrast and aggregation for weakly supervised semantic segmentation, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4299–4309 (2022).

付　　録

A.1 損失関数 $\mathcal{L}_{self}$ と $\mathcal{L}_{cross}$ の分析

損失関数を解析するための式（2）を用いると，$\mathcal{L}_{self}$ の場合は次のように書ける：

$$\mathcal{F}(\boldsymbol{S}) = \frac{1}{m}\sum_{i=1}^{m}\sum_{e=0}^{E-1}\left(\sum_{(\mathbf{x}_i^t,\mathbf{x}_j^{t-e})\in\mathcal{N}} w_{ij}^{\mathbf{x}} S_{ij}^{\mathbf{x}} + \sum_{(\mathbf{y}_i^t,\mathbf{y}_j^{t-e})\in\mathcal{N}} w_{ij}^{\mathbf{y}} S_{ij}^{\mathbf{y}}\right) - \frac{1}{m}\sum_{i=1}^{m}\sum_{e=0}^{E-1}\left(\sum_{(\mathbf{x}_i^t,\mathbf{x}_j^{t-e})\in\mathcal{P}} w_{ij}^{\mathbf{x}} S_{ij}^{\mathbf{x}} + \sum_{(\mathbf{y}_i^t,\mathbf{y}_j^{t-e})\in\mathcal{P}} w_{ij}^{\mathbf{y}} S_{ij}^{\mathbf{y}}\right). \tag{A.1}$$

ここで，$w_{ij}^{\mathbf{x}} = \left|\frac{\partial\mathcal{L}(\boldsymbol{S})}{\partial S_{ij}^{\mathbf{x}}}\right|_l\Big|$ と $w_{ij}^{\mathbf{y}} = \left|\frac{\partial\mathcal{L}(\boldsymbol{S})}{\partial S_{ij}^{\mathbf{y}}}\right|_l\Big|$ は l 番目のイテレーションにおける重みである．また $\boldsymbol{S}$ は類似度行列であり，その行列の (i,j) 番目の要素は，ある学習イテレーションにおける音楽（または画像）の埋め込み特徴量と，メモリに保存された音楽（または画像）の埋め込み特徴量のコサイン類似度で定義される．それぞれ次のように書ける：

$$w_{ij}^{\mathbf{x}} = \begin{cases} \frac{1}{\tau} - \chi_{ij}^{\mathbf{x}} & \left((\mathbf{x}_i^t,\mathbf{x}_j^{t-e})\in\mathcal{P}\right), \\ \chi_{ij}^{\mathbf{x}} & \left((\mathbf{x}_i^t,\mathbf{x}_j^{t-e})\in\mathcal{N}\right), \end{cases} \tag{A.2}$$

$$\chi_{ij}^{\mathbf{x}} = \frac{1}{\tau}\frac{e^{S_{ij}^{\mathbf{x}}/\tau}}{e^{S_{i+}^{\mathbf{x}}/\tau} + \sum_{(\mathbf{x}_i^t,\mathbf{x}_j^{t-e})\in\mathcal{N}} e^{S_{ij}^{\mathbf{x}}/\tau}}, \tag{A.3}$$

$$S_{ij}^{\mathbf{x}} = \frac{\mathbf{z}_i^{\mathbf{x},t\top}\mathbf{z}_j^{\mathbf{x},t-e}}{|\mathbf{z}_i^{\mathbf{x},t}||\mathbf{z}_j^{\mathbf{x},t-e}|},\ \mathbf{z}_j^{\mathbf{x},t-e}\in\boldsymbol{M}^{\mathbf{x}}, \tag{A.4}$$

$$w_{ij}^{\mathbf{y}} = \begin{cases} \frac{1}{\tau} - \chi_{ij}^{\mathbf{y}} & \left((\mathbf{y}_i^t,\mathbf{y}_j^{t-e})\in\mathcal{P}\right), \\ \chi_{ij}^{\mathbf{y}} & \left((\mathbf{y}_i^t,\mathbf{y}_j^{t-e})\in\mathcal{N}\right), \end{cases} \tag{A.5}$$

$$\chi_{ij}^{\mathbf{y}} = \frac{1}{\tau}\frac{e^{S_{ij}^{\mathbf{y}}/\tau}}{e^{S_{i+}^{\mathbf{y}}/\tau} + \sum_{(\mathbf{y}_i^t,\mathbf{y}_j^{t-e})\in\mathcal{N}} e^{S_{ij}^{\mathbf{y}}/\tau}}, \tag{A.6}$$

$$S_{ij}^{\mathbf{y}} = \frac{\mathbf{z}_i^{\mathbf{y},t\top}\mathbf{z}_j^{\mathbf{y},t-e}}{|\mathbf{z}_i^{\mathbf{y},t}||\mathbf{z}_j^{\mathbf{y},t-e}|},\ \mathbf{z}_j^{\mathbf{y},t-e}\in\boldsymbol{M}^{\mathbf{y}}. \tag{A.7}$$

同様に，$\mathcal{L}_{cross}$ の場合は次のように書ける：

$$\mathcal{F}(\boldsymbol{C}) = \frac{1}{m}\sum_{i=1}^{m}\sum_{e=0}^{E-1}\left(\sum_{(\mathbf{x}_i^t,\mathbf{y}_j^{t-e})\in\mathcal{N}} w_{ij}^C C_{ij} + \sum_{(\mathbf{y}_i^t,\mathbf{x}_j^{t-e})\in\mathcal{N}} \hat{w}_{ij}^C \hat{C}_{ij}\right) - \frac{1}{m}\sum_{i=1}^{m}\sum_{e=0}^{E-1}\left(\sum_{(\mathbf{x}_i^t,\mathbf{y}_j^{t-e})\in\mathcal{P}} w_{ij}^C C_{ij} + \sum_{(\mathbf{y}_i^t,\mathbf{x}_j^{t-e})\in\mathcal{P}} \hat{w}_{ij}^C \hat{C}_{ij}\right). \tag{A.8}$$

ここで，$w_{ij}^C = \left|\frac{\partial\mathcal{L}(\boldsymbol{C})}{\partial C_{ij}}\right|_l$ と $\hat{w}_{ij}^C = \left|\frac{\partial\mathcal{L}(\boldsymbol{C})}{\partial \hat{C}_{ij}}\right|_l$ は l 番目のイテレーションにおける重みである．また $\boldsymbol{S}$ は類似度行列であり，その行列の (i,j) 番目の要素は，ある学習イテレーションにおける音楽（または画像）の埋め込み特徴量と，メモリに保存された画像（または音楽）の埋め込み特徴量のコサイン類似度で定義される．それぞれ次のように書ける：

$$w_{ij}^C = \begin{cases} \frac{1}{\tau} - \chi_{ij}^C & \left((\mathbf{x}_i^t,\mathbf{y}_j^{t-e})\in\mathcal{P}\right), \\ \chi_{ij}^C & \left((\mathbf{x}_i^t,\mathbf{y}_j^{t-e})\in\mathcal{N}\right), \end{cases} \tag{A.9}$$

$$\chi_{ij}^C = \frac{1}{\tau}\frac{e^{C_{ij}/\tau}}{e^{C_{i+}/\tau} + \sum_{(\mathbf{x}_i^t,\mathbf{y}_j^{t-e})\in\mathcal{N}} e^{C_{ij}/\tau}}, \tag{A.10}$$

$$C_{ij} = \frac{\mathbf{z}_i^{\mathbf{x},t\mathsf{T}}\mathbf{z}_j^{\mathbf{y},t-e}}{|\mathbf{z}_i^{\mathbf{x},t}||\mathbf{z}_j^{\mathbf{y},t-e}|},\ \mathbf{z}_j^{\mathbf{y},t-e}\in\boldsymbol{M}^{\mathbf{y}}, \tag{A.11}$$

$$\hat{w}_{ij}^C = \begin{cases} \frac{1}{\tau} - \hat{\chi}_{ij}^C & \left((\mathbf{y}_i^t,\mathbf{x}_j^{t-e})\in\mathcal{P}\right), \\ \hat{\chi}_{ij}^C & \left((\mathbf{y}_i^t,\mathbf{x}_j^{t-e})\in\mathcal{N}\right), \end{cases} \tag{A.12}$$

$$\hat{\chi}_{ij}^C = \frac{1}{\tau}\frac{e^{\hat{C}_{ij}/\tau}}{e^{\hat{C}_{i+}/\tau} + \sum_{(\mathbf{y}_i^t,\mathbf{x}_j^{t-e})\in\mathcal{N}} e^{\hat{C}_{ij}/\tau}}, \tag{A.13}$$

$$\hat{C}_{ij} = \frac{\mathbf{z}_i^{\mathbf{y},t\mathsf{T}}\mathbf{z}_j^{\mathbf{x},t-e}}{|\mathbf{z}_i^{\mathbf{y},t}||\mathbf{z}_j^{\mathbf{x},t-e}|},\ \mathbf{z}_j^{\mathbf{x},t-e}\in\boldsymbol{M}^{\mathbf{x}}. \tag{A.14}$$

A.2 SCFEM のアルゴリズム

我々が提案した SCFEM の疑似コードをアルゴリズム 1 に示す．この疑似コードは PyTorch [28] で書かれている．複数エポックにわたって音楽と画像のそれぞれの埋め込み特徴量を保存する SCFEM は，XBM [44] に基づいて実装している．XBM との違いは，その仕組みのほかに，メモリの初期化のタイミングにある．XBM は学習エポックごとにメモリを初期化するのに対し，SCFEM は学習開始時の一回のみメモリを初期化する．SCFEM は，音楽とその代表画像について，メモリに保存された埋め込み特徴量を使って，最大で $2E-1 (E \geq 1)$ の正例のペアを構成することができる．また SCFEM は，XBM と同様に，既存の深層距離学習の学習フレームワークにモジュールとして簡単に組み込むことができる．

Algorithm 1 SCFEM の疑似コード．

```
# 音楽エンコーダf_M と画像エンコーダ f_I を
# T エポック事前学習する

# メモリM を初期化する
M = Memory()

# 損失関数を定義する
Loss = LossFunction()

for epoch in range(T+1, EPOCH_END):
   for x, y, ids in dataloader: # x: 音楽, y: 画像,
         ids: 音楽とその代表画像に与えるID

      # 音楽と画像を共有埋め込み空間に埋め込む
      mus_emb = f_M(x)
      img_emb = f_I(y)

      # 埋め込み特徴量をメモリに保存する
      # メモリに保存できる埋め込み特徴量の最大数を超える
          タイミングがきたら,メモリに保存したタイミング
          が最も早い埋め込み特徴量を,新たに得られた埋め
          込み特徴量といれかえる
      M.update(mus_emb.detach(), img_emb.detach(),
          ids)

      # ミニバッチの類似度行列を計算する
      sim_1 = torch.matmul(torch.t(mus_emb),
          img_emb)
      sim_2 = torch.matmul(torch.t(img_emb),
          mus_emb)

      # 損失関数を計算する
      loss = Loss(sim_1, ids) + Loss(sim_2, ids)

      # メモリM から埋め込み特徴量を取り出す
      mus_emb_M, img_emb_M, ids_M = M.get()

      # 音楽(または画像)の埋め込み特徴量と,メモリに保存
          された音楽(または画像)の埋め込み特徴量のコサ
          イン類似度行列を計算する
      sim_s1 = torch.matmul(torch.t(mus_emb),
          mus_emb_M)
      sim_s2 = torch.matmul(torch.t(img_emb),
          img_emb_M)

      # 音楽(または画像)の埋め込み特徴量と,メモリに保存
          された画像(または音楽)の埋め込み特徴量のコサ
          イン類似度行列を計算する
      sim_c1 = torch.matmul(torch.t(mus_emb),
          img_emb_M)
      sim_c2 = torch.matmul(torch.t(img_emb),
          mus_emb_M)

      # 全ての損失関数の値を合計する
      loss += Loss(sim_s1, ids_M)
              + Loss(sim_s2, ids_M)
              + Loss(sim_c1, ids_M)
              + Loss(sim_c2, ids_M)

      # エンコーダのネットワークパラメータを更新する
      loss.backward()
      optimizer.step()
```