中性運動解纏於異質性人體運動風格轉換

鈕愷夏
國立臺灣大學
台北, 臺灣
hsiasiasiasia0829@gmail.com

沈奕超
東京大學
東京, 日本
ichaoshen@g.ecc.u-tokyo.ac.jp

陳炳宇
國立臺灣大學
台北, 臺灣
robin@ntu.edu.tw

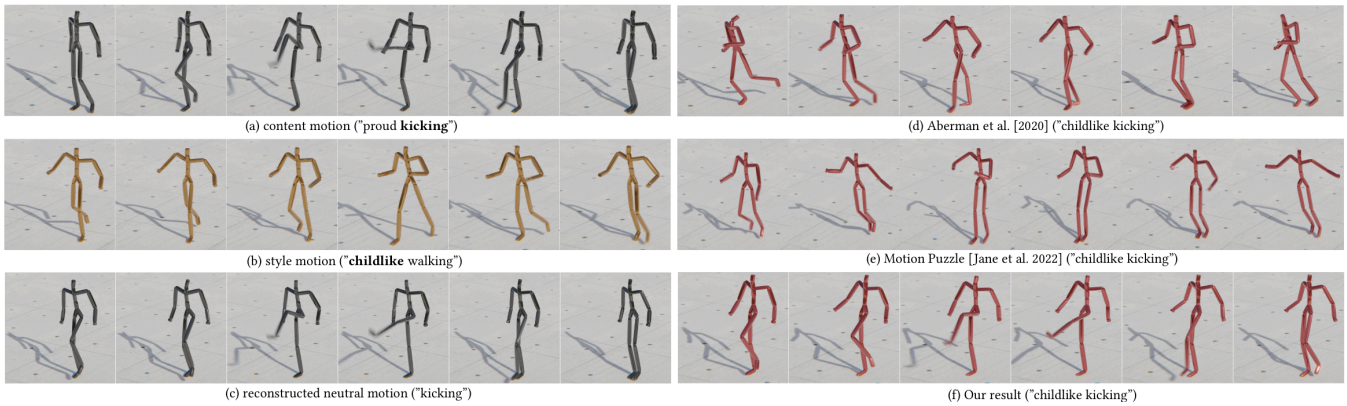


圖 1: 給定（a）"自信的踢腿" 作為輸入的內容運動，以及（b）"幼稚的行走" 作為輸入的風格運動，（d）[1] 往往也會將風格運動（即 "行走"）的內容轉移到內容運動中，而未保留所需的 "踢腿" 內容，而（e）[15] 則無法保留所需的內容，僅將風格轉移到上半身。相比之下，我們的方法通過重建中性運動並生成令人滿意的轉換結果，更好地解纏了內容和風格運動。

摘要

人體運動風格轉換是遊戲設計師和動畫師能應用的重要技術，可以增強運動片段中的角色表現力。然而，目前的數據驅動方法要麼需要配對的運動數據，要麼會由於模型對於運動的解纏不良造成產生的結果不具有所需的內容和風格。因此本研究提出了一種新穎的中性運動解纏（NMD）模型，可以產生高質量的人體運動風格轉換結果，特別是在異質性運動之間。我們的框架通過對抗式的訓練過程，同時學習重建中性運動和輸入運動來分解輸入運動的內容和風格。此外，我們提出了一種新穎的分群損失，進一步增強了內容和風格潛在空間的解纏。我們通過全面的實驗、消融研究和用戶研究評估了我們的方法。結果表明，特別是在異質性運動之間轉換風格時，我們的方法可以生成比現有的運動風格轉換更好的結果。

關鍵詞

運動風格轉換, 運動合成, 異質性運動, 深度學習, 對比式學習

1 介紹

風格在人體運動中扮演著至關重要的角色，反映了情感、個性和意圖等各種方面。對於遊戲設計師或動畫師來說，為人類運動添加微妙的風格可以增強原始運動的表現力和多樣性。然而，首先從專業演員那裡捕捉到所有風格的運動是不切實際的。除此之外，使用有限的運動風格資料，手動將所需的風格從一個運動片段轉移到另一個運動片段是極具挑戰性且耗時的。因此，準確且自動的運動風格轉換方法的需求不斷增加，以滿足對高品質動畫消費日益增長的需求。

通常，以往的方法 [1, 2, 6, 13, 15, 25, 27] 會通過從一個風格運動片段中提取風格，並將其應用到包含內容動作片段的新的運動片段中，來生成風格化的運動。然而，儘管最近基於數據驅動的方法利用神經網絡來模擬多樣化的風格取得了一些進展，但它們在將風格轉移到具有異質行為的兩個運動片段（如「踢腿」和「揮拳」）之間時仍然面臨困難。主要原因是現有的方法難以準確地將運動片段中的風格和內容分離，因此風格運動片段中的內容和風格都會被轉移到內容運動片段中。因此，最終的運動結果往往無法被識別為所期望的具有指定風格的運動。當在包含同質性內容的兩個運動片段之間轉移風格時（如「行走」和「跑步」），這個問題較不明顯。相比之下，當將「孩童般」的風格從「孩童般的行走」運動（見Figure 1(b)）轉移到「自信的踢腿」運動（見Figure 1(a)）時，過去的方法可能無法生成「孩童般的踢腿」，因為「行走」的內容沒有被分離，導致不合理的腿部運動，類似於「行走」（見Figure 1(d)），或者僅將風格轉移到某些身體部位（例如僅轉移到上半身，如Figure 1(e)所示）。

為了解決這些挑戰，我們提出了一種名為中性運動解纏 *(Neutral Motion Disentanglement, NMD)* 的模型，這是一種新穎的基於神經網絡的方法，用於在未配對的異質性運動之間轉移風格。我們的方法通過從給定的運動片段中恢復中性運動來解纏風格和內容，從而減少在異質性運動之間轉移風格時複製內容的影響。為了進一步對各種內容和風格進行聚類，我們引入了對比損失（contrastive loss），學習到更有效分離它們的表示。

如Figure 2所示，我們的 NMD 模型包含兩個分別編碼內容和風格的獨立編碼器，還包括一個主解碼器和一個中性解碼

CGW '24, July 11–12, 2024, Taipei City, Taiwan

器。主解碼器旨在恢復輸入運動，而中性解碼器僅使用編碼的內容編碼作為輸入來恢復中性運動。我們的架構受到之前一種運動預測方法 [4] 的啓發，但有兩個不同之處：首先，我們將完整的運動片段餵入內容和風格編碼器，而不是部分片段，以提取完整運動片段的風格。其次，我們從頭生成一個新的運動片段，而不是生成可以與前一部分連接的後半片段。在訓練過程中，我們將相同的運動片段餵入內容和風格編碼器。我們使用重建和聚類目標來優化所有編碼器和解碼器的參數。重建目標幫助我們重建輸入運動和中性運動，進而促進內容和風格的解纏。同時，另一個目標鼓勵不同內容和風格的聚類。

我們在各種內容和風格上測試了我們的方法，並進行了不同方法之間的定量評估和消融研究，以證明其有效性。結果顯示，我們的方法在運動風格轉移結果方面優於以往的方法，特別是在需要在異質性運動之間轉移風格的情況下。

2 相關工作

2.1 圖像風格轉移

最早嘗試在圖像之間轉移風格的研究可以追溯到圖像類比技術 [9]。該技術通過給定的一對圖像（A 和 A'）描述的視覺關係轉移到新圖像（B），從而創建一個類比（B'）。後來，這一概念被以各種方式擴展，如結合深度神經網絡 [19] 和擴散模型 [22]。

另一方面，[8] 提出了一種稱為神經圖像風格轉移的方法，消除了從圖像對中指定視覺關係的需求。他們使用預訓練的分類網路來計算圖像之間的風格相似性。這一發展促使了進一步的研究，以從不同方面改進風格轉移，例如更快的風格轉移 [17]、使用實例歸一化（Instance Normalization, IN）層提高質量 [24] 和自適應實例歸一化（Adaptive Instance Normalization, AdaIN）層 [14]。IN 和 AdaIN 模塊的使用已成為圖像風格轉移中的常見做法。例如，[28] 應用了 AdaIN 層並提出了一種稱為 CAST 的架構，利用 CycleGAN 生成風格化圖像。他們還通過 MOCO 模塊引入對比學習，以確保風格化圖像的潛在編碼與輸入目標風格圖像的潛在編碼相似。

我們的聚類損失受到 [28] 和 [11] 在圖像風格轉移中提出的語義損失的啓發。這種額外的損失使我們能夠更好地聚類內容和風格的潛在編碼，並解決異質性運動風格轉移的挑戰。

2.2 運動風格轉移

在計算機動畫中，人類運動風格轉移一直是一個長期存在的挑戰。該領域的早期嘗試依賴於手工設計的特徵 [2, 3, 20, 25, 27]。然而，這些人工設計的特徵難以描述不精確的風格定義，導致轉移結果不讓人滿意。

為了解決這個問題，利用機器學習和深度學習方法從範例運動中直接提取風格的數據驅動方法引起了極大關注 [1, 6, 12, 13, 15, 23, 26]。其中，Aberman 等人 [1] 使用 3D 運動和 2D 影片作為輸入的風格運動，實現了高質量的無配對運動風格轉移。Jang 等人 [15] 提出的 "motion puzzle"，可以控制個別身體部位的運動風格，顯著擴大了風格化運動的範圍。然而，這些方法在轉移異質性運動之間的風格時仍然面臨困難，因為內容和風格的解纏不夠滿意，例如從 "自豪的踢腿" 到 "孩子般的行走"。MOCHA [16] 在實時將目標角色的運動風格和身體比例轉移到輸入來源運動中。然而，它需要一個無風格且中性的來源運動，這限制了它的實用性。相反，我們的方法通過一個新穎的中性運動重建目標，將異質性運動的內容與風格解纏開來。此外，我們增強了內容和風格潛在空間的聚類。這兩個特點使我們的方法能夠生成更高質量的轉移結果。

2.3 行為解纏模型

Blattmann 等人 [5] 提出了一種行為表示法，該表示法代表人體運動動量與姿態的獨立性。這種行為表示法可以用於改變任意姿勢下的人體運動行為，或者用於預測未來的人體運動 [4]。學習這種表示法的關鍵在於使用輔助解碼器來防止行為編碼與運動內容混合在一起，反之亦然。受到這些方法的啓發，我們在我們的模型中加入了一個中性解碼器，用於重建一個不帶任何運動風格的中性運動。這種方法能夠更好地解纏內容和風格，從而改善轉移結果。

3 方法

3.1 概覽與運動資料表示

我們方法的輸入是一個具有所需內容 $\hat{c}$ 的內容運動片段 $\mathbf{M}^{\hat{c}}$，和一個具有所需風格 $\hat{s}$ 的風格運動片段 $\mathbf{M}_{\hat{s}}$。我們的方法不依賴任何其他文本描述作為輸入。我們的目標是生成一個新的運動片段 $\mathbf{M}_{\hat{s}}^{\hat{c}}$，其中顯示了所需風格 $\hat{s}$，但保留了內容運動片段中的內容 $\hat{c}$。例如，當我們想通過從 "沮喪的揮拳" 運動片段轉移到 "憤怒的踢腿" 運動片段來獲得一個 "沮喪的踢腿" 的運動片段時，則 $\hat{s}$ = "沮喪"，$\hat{c}$ = "踢腿"。我們將每個運動 $\mathbf{M} \in \mathbb{R}^{T \times n} = [m_0, m_1, ..., m_{T-1}]$ 表示為每幀的旋轉信息 m 和幀數 T 的串聯，其中每幀的姿勢定義為總共 $n = 31$ 個關節連接的四元數的串聯。

3.2 NMD: 中性運動解纏模型

為了實現運動風格轉移，以往的基於深度學習的方法將問題視為在整個人體 [1] 或身體部位 [15] 之間進行翻譯的問題。雖然這些方法可以成功地在具有相似行為的運動之間轉移風格，但它們無法在異質運動上達到相同的質量。造成這種情況的主要原因是內容-風格潛在空間之間以及內容和風格潛在空間內部的解纏不佳。首先，這些方法不需要生成一個中性運動，因此在轉移風格時很難解纏內容和風格。其次，內容空間內解纏不佳使得網絡難以區分具有不同內容的運動片段，因此在轉移過程中部分內容資訊會直接被複製。同樣的情況也會在風格空間內解纏不佳時發生。為了解決這些問題，我們提出了一種中性運動解纏（NMD）模型，旨在學習如何在運動片段之間轉移風格的同時重建中性運動。此外，我們提出使用一種新穎的聚類損失來改善內容和風格潛在空間內部的解纏。

我們的模型受到 [4] 的啓發，該研究介紹了一個框架，用於將行為從輸入運動中解纏出來，並將其轉移到進行中的運動中以實現運動預測。如Figure 2所示，我們的模型由四個組件組成：一個內容編碼器 E_C，一個風格編碼器 E_S，一個主解碼器 G_M 和一個中性解碼器 G_N。給定一個輸入運動片段 $\mathbf{M}$，內容編碼器 E_C 和風格編碼器 E_S 將其編碼為內容編碼 $z_c = E_C(\mathbf{M})$ 和風格編碼 $z_s = E_S(\mathbf{M})$。我們使用具有 128 隱層維度的單層 LSTM [10] 作為內容和風格編碼器的架構。

主解碼器 G_M 的目標是使用內容編碼 z_c 和風格編碼 z_s 重建輸入運動片段 $\mathbf{M}$。另一方面，中性解碼器 G_N 的目標是僅使用內容編碼 z_c 來重建中性運動片段，以防止內容編碼器保留任何風格信息。總體而言，在訓練過程中的生成過程可以表示為：

$$\tilde{\mathbf{M}} = G_M(E_C(\mathbf{M}), E_S(\mathbf{M})) \tag{1}$$

$$\tilde{\mathbf{M}}_{\text{neutral}} = G_N(E_C(\mathbf{M})), \tag{2}$$

其中，$\tilde{\mathbf{M}}$ 和 $\tilde{\mathbf{M}}_{\text{neutral}}$ 分別表示重建的耦合運動和中性運動。我們在主解碼器和中性解碼器的單層 GRU [7] 輸出上近一步使用

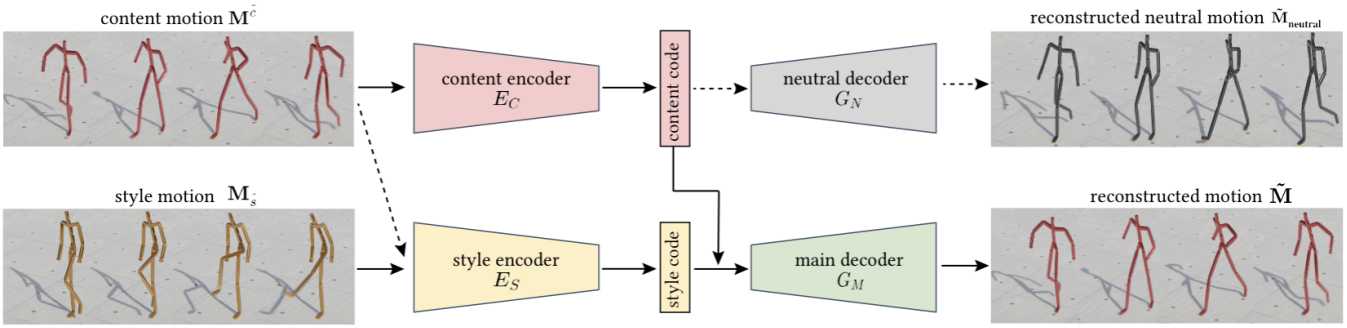


圖 2: 我們的中性運動解纏（NMD）模型框架如下所示。在訓練過程中，我們將相同的運動用於內容運動和風格運動，即 $\mathbf{M}_{\hat{s}} = \mathbf{M}^{\hat{c}}$。我們訓練內容編碼器 E_C、風格編碼器 E_S 和中性解碼器 G_N，僅使用內容編碼來重建中性運動 $\tilde{\mathbf{M}}_{\text{neutral}}$。同時，我們還訓練主解碼器 G_M，以重建原始運動 $\tilde{\mathbf{M}} = \mathbf{M}^{\hat{c}} = \mathbf{M}_{\hat{s}}$。在推斷過程中，僅使用兩個編碼器 E_C、E_S 和主解碼器 G_M 來執行運動風格轉移。（⇢ 表示訓練流程，→ 表示推斷流程）。

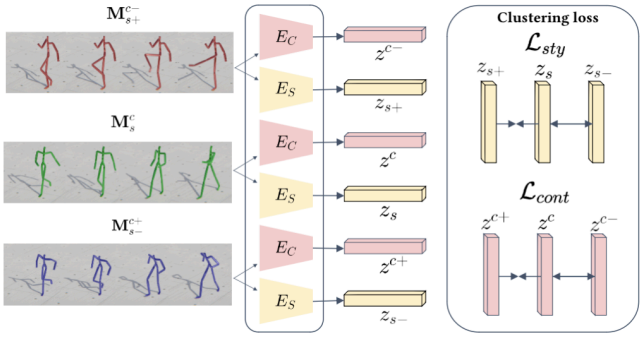


圖 3: 為了計算給定的樣本運動三元組 $(\mathbf{M}_s^c, \mathbf{M}_{s+}^{c-}, \mathbf{M}_{s-}^{c+})$ 的風格聚類損失（這裡我們省略了 $[i]$），我們確保相同風格的潛在向量保持彼此接近，而不同風格的潛在向量則保持彼此較遠。此外，我們對內容聚類損失的計算方法也是相同的。

單個線性轉換層。在應用時（→ 在Figure 2中），我們從主解碼器 G_M 獲得轉移結果，而中性解碼器將不會被使用。

3.3 訓練流程與損失

在訓練過程中，我們的目標包括同時重建原始運動和中性運動。為了實現這些目標，我們使用以下損失函數：

$$\mathcal{L}_{\text{all}} = \mathcal{L}_{\text{recon}} + \beta \left(\mathcal{L}_{\text{sty}} + \mathcal{L}_{\text{cont}}\right), \tag{3}$$

在我們的所有實驗中，我們設置 $\beta = 1.0$。

在每個訓練迭代中，我們通過交替兩個階段來優化所有編碼器和解碼器的參數：耦合運動重建階段和中性運動重建階段。

3.3.1 中性運動重建階段. 在這個階段，目標是優化中性解碼器的參數，強迫它嘗試僅從內容編碼重建輸入的運動片段 $\mathbf{M}$：

$$\mathcal{L}_{\text{neutral}} = \|G_N(E_C(\mathbf{M})) - \mathbf{M}\|_2. \tag{4}$$

為了最小化Equation (4)，我們不提供額外的中性運動片段，而是使用輸入的運動片段作為目標。這是因為並不總是獲得每個運動的中性版本。與其強迫網絡重建特定的中性運動，我們採用對抗訓練策略，細節將在下一階段介紹。

3.3.2 耦合運動重建階段. 在這個階段，整體目標是優化編碼器和主解碼器的參數，以鼓勵它使用內容編碼和風格編碼準確的重建輸入運動 $\mathbf{M}$：

$$\begin{aligned}\mathcal{L}_{\text{recon}} = &\|G_M(E_C(\mathbf{M}), E_S(\mathbf{M})) - \mathbf{M}\|_2 \\ &- D_{\text{KL}}(E_C(\mathbf{M}) \| \mathcal{N}(0, I)) - \mathcal{L}_{\text{neutral}},\end{aligned} \tag{5}$$

其中 D_{KL} 是 KL 散度的損失函數，$\mathcal{N}(0, I)$ 是多變量正態分佈。

正如之前提到的，包含 $-\mathcal{L}_{\text{neutral}}$ 的目的是防止 G_N 僅從 $E_C(\mathbf{M})$ 重建 $\mathbf{M}$。它還鼓勵濾除可能包含在內容編碼 $E_C(\mathbf{M})$ 中的任何風格信息。

聚類損失. 儘管上述的對抗訓練策略和損失函數鼓勵內容和風格潛在空間的解纏，但它們並未提供對內容和風格潛在空間內部的解纏的指導。因此，我們引入了一種聚類損失來增強不同內容和風格之間的分離。如Figure 3所示，給定一個樞紐運動 $\mathbf{M}_s^c$，其風格為 s，內容為 c，我們隨機選擇 K 個三元組 $(\mathbf{M}_s^c, \mathbf{M}_{s+}^{c-}[i], \mathbf{M}_{s-}^{c+}[i])_{i=1}^K$，其中每個三元組包括具有相同風格但不同內容的運動 $(\mathbf{M}_{s+}^{c-})$ 和具有相同內容但不同風格的運動 $(\mathbf{M}_{s-}^{c+})$。對比表示通過將樞紐風格編碼 $E_S(\mathbf{M}_s^c)$ 與正面風格編碼 $E_S(\mathbf{M}_{s+}^{c-})[i]$ 相對於負面風格編碼 $E_S(\mathbf{M}_{s-}^{c+})[i]$ 的餘弦距離最小化來學習運動的風格編碼。我們定義了風格對比損失來訓練我們的風格編碼器 E_S：

$$\mathcal{L}_{\text{sty}} = \frac{1}{K}\sum_{i=1}^{K} -\left(E_S(\mathbf{M}_s) \cdot E_S(\mathbf{M}_{s+}^{c-}[i])\right) + \left(E_S(\mathbf{M}_s) \cdot E_S(\mathbf{M}_{s-}^{c+}[i])\right), \tag{6}$$

在這裡，$\cdot$ 表示兩個向量的內積。

對於內容編碼，我們採用了與Equation (6)相同的對比損失形式：

$$\mathcal{L}_{\text{cont}} = \frac{1}{K}\sum_{i=1}^{K} -\left(E_C(\mathbf{M}_s^c) \cdot E_C(\mathbf{M}_{s-}^{c+}[i])\right) + \left(E_C(\mathbf{M}_s^c) \cdot E_C(\mathbf{M}_{s+}^{c-}[i])\right). \tag{7}$$

4 實驗與評估

4.1 實作細節與資料集

我們在 PyTorch [21] 中實現了我們的 NMD 模型，並使用 Adam [18] 優化器進行了參數更新。我們將中性解碼器的初始學習率設置為 0.005，其餘部分為 0.001，並使用 StepLR 逐漸

降低學習率。我們使用 NVIDIA GeForce GTX 4070Ti GPU 對中性解碼器和主解碼器進行了 1,000 個 epoch 的訓練。在這項工作中，我們使用了兩個不同的數據集，一個是由 [26] 提出的數據集 X，另一個是由 [1] 提出的數據集 A。訓練數據集 X 大約需要 2 小時，而訓練數據集 A 大約需要 2.5 小時。我們還使用這兩個數據集重新訓練了 [1] 和 [15] 提出的模型，以生成定性和定量評估中的結果。

4.1.1 資料前處理. 數據集 X 包含 8 種不同的風格和 6 種不同類型的內容。另一方面，數據集 A 包含多達 16 種不同的風格和具有不同行為的多個運動序列。然而，原始的數據集 A 沒有根據行為進行分類。因此，我們對數據集進行了額外的分段處理，最終將數據標記為 8 種不同的內容。在訓練和測試兩個數據集中，運動序列被分成了長度為 $T = 32$ 幀的短重疊段，重疊時間為 $\frac{T}{4}$。此外，我們從數據集 X 中刪除了標記為“過渡”的運動，並將“行走”和“奔跑”視為兩個數據集中相同的內容。總的來說，對於數據集 X，我們使用了 1,500 個運動進行訓練，500 個運動進行測試。對於數據集 A，我們使用了 12,000 個運動進行訓練，6,000 個運動進行測試。

此外，我們通過多種方式豐富了訓練數據集。首先，我們通過鏡像化 3D 骨架來使原始運動片段的數量翻倍。其次，我們通過將具有較少數據的內容運動的重疊時間從 $\frac{T}{4}$ 縮短到 $\frac{T}{8}$ 來解決數據偏差問題。用於用戶研究和定量評估的測試數據集都是在訓練期間未使用的未見的運動片段。

4.2 定量分析

在Figure 4中，我們將我們的方法獲得的轉移結果與 [1] 進行了比較。我們將內容、風格和生成的運動骨架分別以黑色、黃色和紅色可視化。在Figure 4(c) 左側，我們可以看到 [1] 獲得的轉移結果中只包含來自風格運動的內容（即“行走”），而缺少來自內容運動（即“踢腿”）的內容。相反，我們的方法表現更好，生成了一個包含所需“踢腿”運動的轉移結果，其風格為“自大”，其特點是肩膀向後傾斜，腿部運動從後向前。同樣，在Figure 4(c) 右側，我們可以看到 [1] 獲得的轉移結果也只包含“行走”，而未反映來自內容運動（即“揮拳”）的內容。我們假設他們的結果對“行走”的傾象可能是由於他們無法有效地解決原始數據集中的數據偏差（數據集 X 中有 49% 的運動具有“行走”內容）。最後，[1] 獲得的轉移結果總是存在手部關節運動不自然的問題，這些都需要額外的後處理才能去除這些瑕疵。相比之下，我們的方法再不做後處理時生成的運動更穩定。

此外，我們在Figure 1中比較了我們的方法和 Motion Puzzle [15] 獲得的結果。雖然上半身的風格被轉移到了內容運動中，但他們的方法似乎也將風格運動的內容（“行走”）轉移到了結果中。這可以通過轉移結果中的腿部運動觀察到，它們似乎在跳躍，兩條腿一起向前和向後移動，而不是“踢腿”或“行走”。在附帶的影片中，我們進一步展示了我們的方法在異質性運動之間生成了令人滿意的轉移結果，並且對關節晃動具有較高的抵抗力。

4.2.1 使用者研究. 我們進行了一項使用者研究，以展示我們的方法相對於 [1] 能夠生成更好的轉移結果。我們從大學招募了 15 名參與者，這些參與者之前沒有從事過計算機動畫研究。按照 [1] 提出的用戶研究程序，我們要求每個參與者回答一份問卷，評估轉移結果的真實性、內容保留和風格轉移性。我們專注於使用數據集 X ([26]) 來評估轉移結果，該數據集包含了“行走”、“踢腿”、“揮拳”和“跳躍”等異質性運動。

(a) Realism.

	Ours	Aberman *et al.* [1]
Realism	86.6 %	13.4 %

(b) Content preservation and style transfer.

	Ours	Aberman *et al.* [1]
Content preservation	81.7 %	18.3 %
Style Transfer	56.4 %	43.6 %

表 1: 使用者研究結果關於真實性，內容保留和風格轉移性。(a) 真實性：相較於 [1]，我們的方法產生了更合理的結果。(b) 內容保留和風格轉移性：我們的方法更好地保留了內容運動的內容和風格運動的風格。我們額外標注了每個指標的最佳結果。

真實性. 在這部分，我們旨在評估不同運動的真實性和可信度。對於每個問題，參與者被呈現一個參考運動和兩個候選運動，每個候選運動都執行相同類型的內容和風格（例如“生氣的踢腿”）。參考運動取自原始的運動資料集，而兩個候選運動分別來自 [1] 和我們的方法。然後，參與者被要求在考慮參考運動的情況下，在兩個候選運動中選擇最真實的一個。

對於真實性問題，我們收到了 168 個答案。Table 1(a) 中的結果表明，我們的方法產生的運動比 [1] 更為真實。這種改進是由於更好的內容和風格解纏和聚類。此外，我們的方法產生的結果沒有明顯的關節振動，從而生成更真實的運動。

內容保留與風格轉移性. 我們在內容保留和風格轉移性方面比較了我們的方法和 [1] 所得到的轉移結果。在每個問題中，參與者被呈現四個運動：一個內容運動，一個風格運動，一個由 [1] 得到的轉移結果，以及我們方法得到的另一個轉移結果。他們被問到兩個問題：“哪個運動的內容更類似於內容運動”，以及“哪個運動的風格更類似於風格運動”。

對於每個問題，我們分別收到了 168 個答案。如Table 1(b) 所示，我們的方法產生的結果在內容保留和風格轉移方面獲得了更高的評分。與明顯更優秀的內容保留性相比，我們的方法和 [1] 之間在風格轉移比例上的差異很小。這是因為 [1] 所得到的轉移結果因為解纏不準確，通常類似於直接複製風格運動中的所有資訊（包括內容）。這種結果可能會讓參與者覺得風格已經轉移了。

4.3 定量分析

在本節中，我們評估了我們的方法的性能，並使用 [15] 提出的兩個度量標準與 [1] 進行比較：內容識別準確率（CRA）和風格識別準確率（SRA）。我們使用這些指標來衡量兩種方法在內容保留和風格轉移方面的程度。為了計算這些指標，我們將資料集 X 和 A 依照 9 : 1 的比例分成訓練集和測試集。然後，我們針對每個資料集使用其對應的訓練集，訓練內容分類器和風格分類器。資料集 X 的內容分類器和風格分類器的準確率分別為 100.00% 和 94.00%，資料集 A 的準確率分別為 88.34% 和 87.58%。然後，我們使用這些分類器來計算 CRA 和 SRA，其中較高的 CRA 表示在內容保留方面表現更好，較高的 SRA 表示風格轉移更優秀。我們進行了多次試驗，以獲取每個指標的平均準確率值。對於每次試驗，我們隨機抽取了 275 個具有隨機

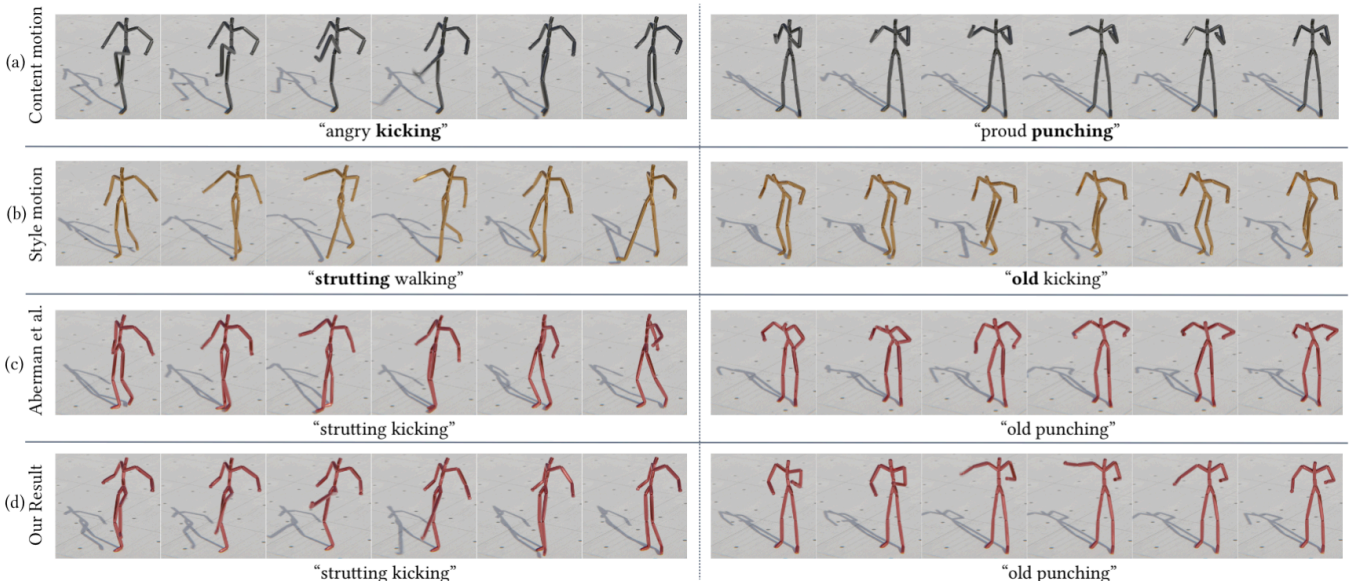


圖 4: 我們的方法和 [1] 所獲得的轉移結果比較。從左到右，內容和風格運動組合是 (i)"生氣的踢腿" 轉換為 "自大的行走"，以及 (ii)"自信的揮拳" 轉換為 "年邁的踢腿"。

(a) Dataset X ([26])

Methods	CRA↑(%)	SRA↑ (%)
Real MoCap motion	100.00	94.00
Aberman *et al.* [1]	7.242 ±1.443	67.238 ±5.314
Motion Puzzle	1.72 ±1.566	37.71 ±7.117
Ours	95.172±2.251	39.656 ±6.096

(b) Dataset A ([1])

Methods	CRA↑ (%)	SRA↑ (%)
Real MoCap motion	88.34	87.58
Aberman *et al.* [1]	24.75 ±0.458	58.073 ±0.765
Motion Puzzle	7.023 ±1.002	38.547 ±0.786
Ours	82.97 ±0.248	34.077 ±0.751

表 2: Xia 和 Aberman 資料集的定量評估。我們對每種方法產生的轉移結果使用 275 和 3,003 個隨機抽樣的異質性運動對計算內容辨識準確率（CRA）和風格辨識準確率（SRA）。表格報告了 CRA 和 SRA 在五次試驗中的平均（± 標準差）值。我們同樣突出顯示了每個指標的最佳結果。

內容的運動對，而不考慮 "行走" 作為內容運動，並為數據集 A 選擇了 3,003 個運動對。請注意，所有抽樣的運動對都具有**異質性內容**，這與先前的工作中的評估不同。然後，我們使用訓練好的分類器來預測我們的方法、Aberman 等人的方法 [1] 和 Motion Puzzle [15] 產生的轉移結果的內容和風格。

我們在 Table 2 中呈現了兩個資料集的結果。對於資料集 X，我們觀察到我們的方法具有顯著更高的 CRA 準確率，這表明轉移結果保留了來自內容運動的期望內容。同時，我們觀察

Dataset X	CRA↑(%)	SRA↑(%)
Ours	95.172 ±2.251	39.656 ±6.096
Ours (w/o neutral decoder)	71.48 ±7.321	31.36 ±4.607
Ours (w/o clustering loss)	4.482 ±1.968	25.518 ±8.214

表 3: 在資料集 X 上進行消融研究。我們對每種方法在每次試驗中產生的 275 個樣本進行內容辨識準確率（CRA）和風格辨識準確率（SRA）的計算。表格報告了每個指標在五次試驗中的平均值（± 標準差）。我們突出顯示了每個指標的最佳結果。

到在資料集 A 上 CRA 的優越性能較小。這可能是因為資料集 A 中的運動的內容和風格比資料集 X 中的運動更耦合。例如，"殭屍" 風格保持手在空中而不是利用動量變化來進行表現，這與情緒風格的情況相反。此外，我們觀察到對於這兩個資料集，[1] 得到的結果在 SRA 上表現較好，但 CRA 較低。這是因為他們的方法不僅將期望的風格轉移到了內容運動，而且也轉移了內容。這項觀察結果與使用者研究結果一致，即由 [1] 產生的結果在內容保留方面僅獲得了 18.3% 的評分。同時，根據我們的使用者研究，我們的結果在風格轉移方面獲得了更高的評分 (56.4%)，而 [1] 產生的結果為 (43.6%)。這項發現表明，SRA 指標並不完全與人類評分一致，而且我們的結果仍然可以有效地進行風格轉移。

總的來說，定量評估結果表明，我們的方法透過大量增強內容保留程度，並最小化轉移風格的損失，產生了更令人滿意的結果。

4.4 消融研究

我們進行了消融研究以評估中性解碼器和聚類損失的影響，並將結果呈現在表格中。

中性解碼器. 我們透過移除中性解碼器並在Equation (5)中排除 $-\mathcal{L}_{\text{neutral}}$ 此項，調查了中性解碼器的影響。如Table 3所示，CRA 和 SRA 得分都顯著下降，顯示風格和內容資訊被混合在一起。如Figure 5(c) 所示，由於解纏效果較差，傳遞結果中包含的風格運動的期望風格較少。

聚類損失. 為了評估我們模型中聚類損失的影響，我們只使用 L_{recon} 來優化我們模型的參數。我們的結果表明，沒有聚類損失，傳遞的運動將風格運動中的內容和風格與內容運動中的一小部分內容資訊（即肩部運動）混合在一起。因此，產生的傳遞結果令人不滿意，如Figure 5(d) 所示。

4.5 潛在空間視覺化

使用主成分分析（PCA）將兩個資料集中的運動產生的內容和風格編碼投影到 3D 空間。我們的目標是展示我們的方法如何有效地解開不同的內容和風格。

內容編碼. 我們在Figure 6中視覺化了由我們的內容編碼器 E_C 產生的內容編碼的 3D 投影，其中每個樣本都使用與其內容標籤對應的顏色進行視覺化。我們可以觀察到對於兩個資料集，使用聚類損失的不同內容的樣本更容易分離。

風格編碼. 在Figure 7中，我們展示了由我們的風格編碼器 E_S 生成風格編碼的 3D 投影，其中每個樣本都使用與其風格標籤對應的顏色進行可視化。我們的方法展現出更好的風格編碼聚類能力，使得相似的風格在潛空間中相對接近，例如 “年邁的” 和 “沮喪的” 。當使用聚類損失進行最佳化時，這種效果尤其明顯。然而，對於資料集 A，我們觀察到即使使用了聚類損失，聚類效果仍然不夠令人滿意。這可能是因為資料集 A 中包含的 姿勢風格類型（例如 “殭屍” ）更難區分。儘管如此，我們的方法仍然增強了風格編碼的聚類效果，從而產生了更好的轉移結果。

5 限制與未來研究

姿勢風格運動解纏. 我們使用的數據集中包含的運動風格大致可分為抽象（例如 “生氣”）和姿勢（例如 “殭屍”）風格。我們的方法可以有效地將內容和抽象風格分離，但從內容中分離姿勢風格是具有挑戰性的。未來，一個潛在的解決方案是將實例歸一化等技術（例如 AdaIN [14]）與我們的方法結合使用。

中性運動解纏品質. 雖然我們的方法通過中性運動重建目標生成了保留內容比以往更好的轉換結果，但重建的中性運動並不總是完全與輸入運動片段中呈現的風格分離開來。如Figure 8所示，與來自原始運動數據集相同主題的 “中性的行走” 運動片段（Figure 8(c)）相比，重建的 “中性的行走” （Figure 8(b)）仍然與輸入 “老年的行走” 運動片段（Figure 8(a)）相似。

6 結論

在這項工作中，我們提出了一種新穎的中性運動分離（NMD）模型，可以在異質性運動之間轉換人類運動風格。我們的模型具有以下兩個新穎特點。首先，它包括一種獨特的網絡架構，具有中性解碼器和交替訓練流程，以增強內容和風格的分離。其次，我們引入了一種新穎的聚類損失，增強了潛在空間中內容和風格的聚類。我們進行了全面的定性和定量評估，結果顯示，與以前的工作相比，我們的方法在轉換異質性運動風格時生成了更加滿意的轉換結果。

REFERENCES

[1] Kfir Aberman, Yijia Weng, Dani Lischinski, Daniel Cohen-Or, and Baoquan Chen. 2020. Unpaired motion style transfer from video to animation. *ACM Trans. Graph.* 39, 4 (2020), 64–1.

[2] Kenji Amaya, Armin Bruderlin, and Tom Calvert. 1996. Emotion from motion. In *Proc. Graphics Interface*, Vol. 96. Toronto, Canada, 222–229.

[3] Andreas Aristidou, Qiong Zeng, Efstathios Stavrakis, KangKang Yin, Daniel Cohen-Or, Yiorgos Chrysanthou, and Baoquan Chen. 2017. Emotion control of unstructured dance movements. In *Proc. SCA*. 1–10.

[4] German Barquero, Sergio Escalera, and Cristina Palmero. 2023. Belfusion: Latent diffusion for behavior-driven human motion prediction. In *Proc. ICCV*. 2317–2327.

[5] Andreas Blattmann, Timo Milbich, Michael Dorkenwald, and Bjorn Ommer. 2021. Behavior-driven synthesis of human dynamics. In *Proc. CVPR*. 12236–12246.

[6] Matthew Brand and Aaron Hertzmann. 2000. Style machines. In *Proc. SIGGRAPH (SIGGRAPH '00)*. ACM Press/ Addison-Wesley Publishing Co., USA, 183—192. https://doi.org/10.1145/344779.344865

[7] Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259* (2014).

[8] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. 2016. Image style transfer using convolutional neural networks. In *Proc. CVPR*. 2414–2423.

[9] Aaron Hertzmann, Charles E. Jacobs, Nuria Oliver, Brian Curless, and David H. Salesin. 2001. Image analogies. In *Proc. SIGGRAPH*. Association for Computing Machinery, New York, NY, USA, 327—340. https://doi.org/10.1145/383259.383295

[10] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Comput.* 9, 8 (nov 1997), 1735—1780. https://doi.org/10.1162/neco.1997.9.8.1735

[11] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. 2018. Cycada: Cycle-consistent adversarial domain adaptation. In *Proc. ICML*. Pmlr, 1989–1998.

[12] Daniel Holden, Ikhsanul Habibie, Ikuo Kusajima, and Taku Komura. 2017. Fast neural style transfer for motion data. *IEEE Computer Graphics and Applications* 37, 4 (2017), 42–49.

[13] Eugene Hsu, Kari Pulli, and Jovan Popović. 2005. Style translation for human motion. *ACM Trans. Graph.* 24, 3 (jul 2005), 1082—1089. https://doi.org/10.1145/1073204.1073315

[14] Xun Huang and Serge Belongie. 2017. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proc. ICCV*. 1501–1510.

[15] Deok-Kyeong Jang, Soomin Park, and Sung-Hee Lee. 2022. Motion puzzle: Arbitrary motion style transfer by body part. *ACM Trans. Graph.* 41, 3 (2022), 1–16.

[16] Deok-Kyeong Jang, Yuting Ye, Jungdam Won, and Sung-Hee Lee. 2023. MOCHA: Real-Time Motion Characterization via Context Matching. In *Proc. SIGGRAPH Asia (SA '23)*. Association for Computing Machinery, New York, NY, USA, Article 7, 11 pages. https://doi.org/10.1145/3610548.3618252

[17] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. 2016. Perceptual losses for real-time style transfer and super-resolution. In *Proc. ECCV*. Springer, 694–711.

[18] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[19] Jing Liao, Yuan Yao, Lu Yuan, Gang Hua, and Sing Bing Kang. 2017. Visual attribute transfer through deep image analogy. *ACM Trans. Graph.* 36, 4 (2017), 1–15.

[20] C Karen Liu, Aaron Hertzmann, and Zoran Popović. 2005. Learning physics-based motion style with nonlinear inverse optimization. *ACM Trans. Graph.* 24, 3 (2005), 1071–1081.

[21] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Proc. NeurIPS*. Curran Associates, Inc., 8024–8035.

[22] Adéla Šubrtová, Michal Lukáč, Jan Čech, David Futschik, Eli Shechtman, and Daniel Sỳkora. 2023. Diffusion Image Analogies. In *Proc. SIGGRAPH*. 1–10.

[23] Tianxin Tao, Xiaohang Zhan, Zhongquan Chen, and Michiel van de Panne. 2022. Style-ERD: Responsive and coherent online motion style transfer. In *Proc. CVPR*. 6593–6603.

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

[25] Munetoshi Unuma, Ken Anjyo, and Ryozo Takeuchi. 1995. Fourier principles for emotion-based human figure animation. In *Proc. SIGGRAPH*. 91–96.

[26] Shihong Xia, Congyi Wang, Jinxiang Chai, and Jessica Hodgins. 2015. Realtime style transfer for unlabeled heterogeneous human motion. *ACM Trans. Graph.* 34, 4 (2015), 1–10.

[27] M Ersin Yumer and Niloy J Mitra. 2016. Spectral style transfer for human motion between independent actions. *ACM Trans. Graph.* 35, 4 (2016), 1–8.
[28] Yuxin Zhang, Fan Tang, Weiming Dong, Haibin Huang, Chongyang Ma, Tong-Yee Lee, and Changsheng Xu. 2022. Domain enhanced arbitrary image style transfer via contrastive learning. In *Proc. SIGGRAPH*. 1–8.

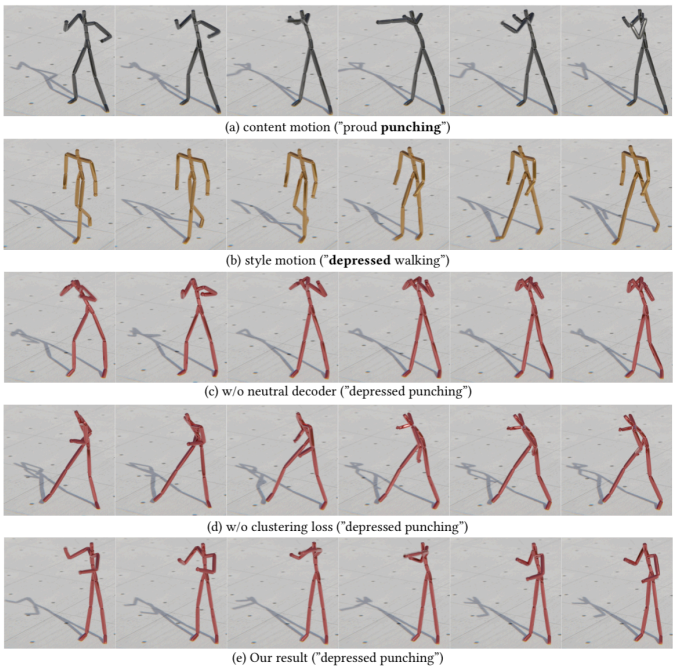


圖 5: 消融研究將（a）內容運動（“自信的揮拳”）轉換為（b）風格運動（“沮喪的行走”）的結果。(c) 在不使用中性解碼器的情況下，我們的模型無法將 “沮喪” 轉移至 “踢腿” 的內容。(d) 在不使用聚類損失的情況下，轉換結果中的內容與風格運動（“行走”）中的內容相似，而不是所需的內容（“揮拳”）。(e) 我們的完整方法生成了類似 “踢腿” 的運動，但以 “沮喪” 的風格呈現。

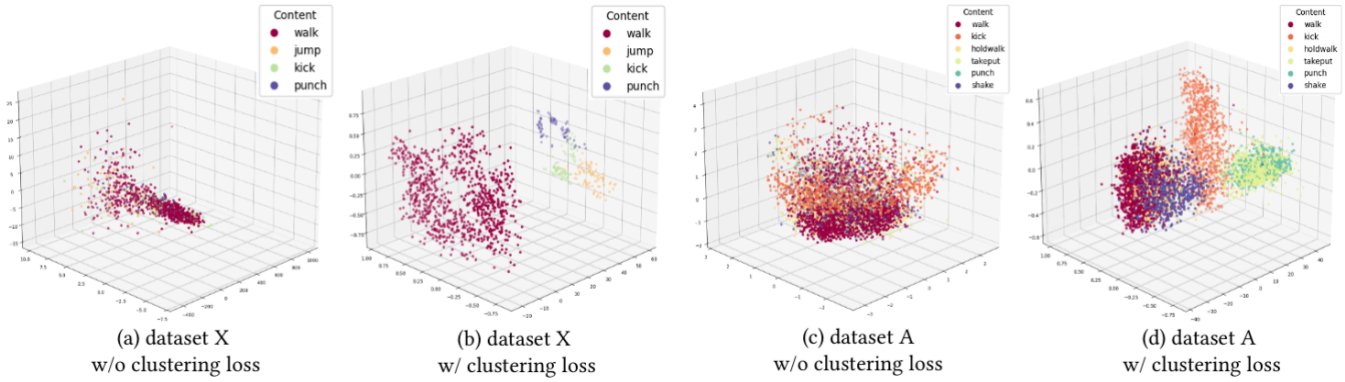
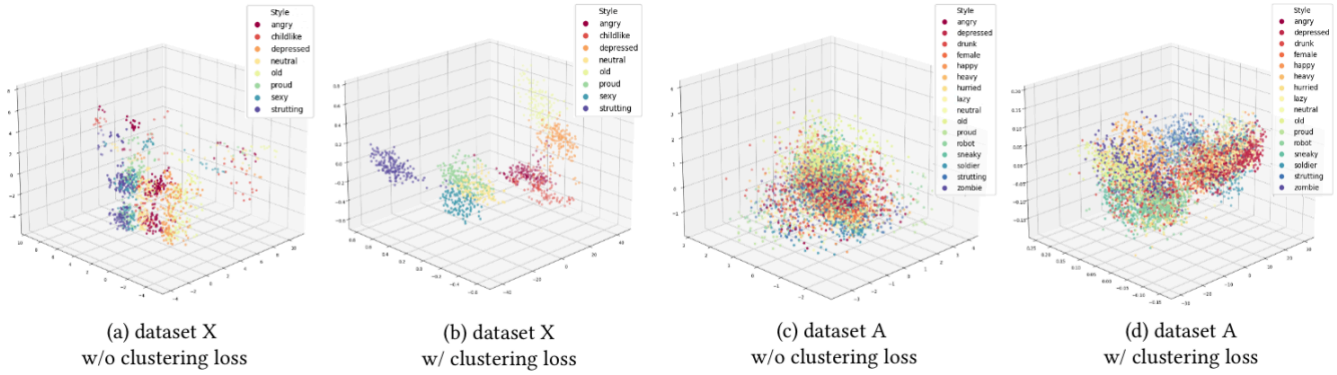


圖 6: 數據集 X 和數據集 A 的投影內容編碼。當使用聚類損失進行優化時，內容編碼之間的分離效果更好。



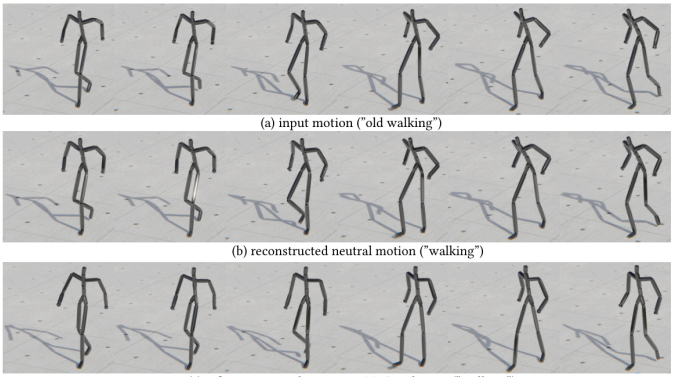
(a) dataset X w/o clustering loss

(b) dataset X w/ clustering loss

(c) dataset A w/o clustering loss

(d) dataset A w/ clustering loss

圖 7: 數據集 X 和數據集 A 的投影風格編碼。使用聚類損失進行優化時，風格編碼之間的分離效果更好。



(a) input motion ("old walking")

(b) reconstructed neutral motion ("walking")

(c) reference neutral motion in MoCap dataset ("walking")

圖 8: 我們的中性解碼器重建的（b）中性 "行走" 運動片段與（a）輸入的 "年邁的行走" 運動片段相比，仍然更加相似，而不是與（c）來自原始運動數據集中的中性 "行走" 相似。這表明輸入運動片段中的風格尚未被完全分離。