

MELOV: Multimodal Entity Linking with Optimized Visual Features in Latent Space

Xuhui Sui, Ying Zhang*, Yu Zhao, Kehui Song, Baohang Zhou, Xiaojie Yuan
College of Computer Science, VCIP, TMCC, TBI Center, Nankai University, China
{suixuhui, zhaoyu, songkehui, zhoubaohang}@dbis.nankai.edu.cn
{yingzhang, yuanxj}@nankai.edu.cn

Abstract

Multimodal entity linking (MEL), which aligns ambiguous mentions within multimodal contexts to referent entities from multimodal knowledge bases, is essential for many natural language processing applications. Previous MEL methods mainly focus on exploring complex multimodal interaction mechanisms to better capture coherence evidence between mentions and entities by mining complementary information. However, in real-world social media scenarios, vision modality often exhibits low quality, low value, or low relevance to the mention. Integrating such information directly will backfire, leading to a weakened consistency between mentions and their corresponding entities. In this paper, we propose a novel latent space vision feature optimization framework MELOV, which combines inter-modality and intra-modality optimizations to address these challenges. For the inter-modality optimization, we exploit the variational autoencoder to mine shared information and generate text-based visual features. For the intra-modality optimization, we consider the relationships between mentions and build graph convolutional network to aggregate the visual features of semantic similar neighbors. Extensive experiments on three benchmark datasets demonstrate the superiority of our proposed framework.

1 Introduction

Entity linking (EL) is the task of assigning ambiguous mentions in text to their corresponding entities in a knowledge base. As a bridge between unstructured content and structured knowledge bases, EL plays a vital role in many downstream natural language processing applications, such as content analysis (Huang et al., 2018), semantic search (Blanco et al., 2015), question answering (Xiong et al., 2019; Longpre et al., 2021) and so on.

*Corresponding author.



Figure 1: Examples of multimodal entity linking with low-quality, low-value, low-relevance and good visual images. Entities that share the same color as mentions are the corresponding gold entities.

Traditional EL methods (Chisholm and Hachey, 2015; Eshel et al., 2017) mainly focus on addressing the text-based EL by resolving text context. However, with the rapid development of social media, images along with text have become the most common form of web information, which brings new challenges for EL models to effectively integrate visual information and understand multimodal content. Thus, the multimodal entity linking (MEL) task (Moon et al., 2018) has been proposed, which extends the scope from textual EL to heterogeneous formats, i.e. linking mentions with textual context and visual context to referent entities in multimodal knowledge bases.

Multiple previous MEL works (Adjali et al., 2020; Gan et al., 2021; Wang et al., 2022a; Zhang and Huang, 2022; Wang et al., 2023; Xing et al., 2023; Shi et al., 2023; Luo et al., 2023; Zhang et al., 2023) mainly focus on exploring complex multimodal interaction mechanisms to better capture coherence evidence between mentions and entities by mining complementary information. Although

these works have achieved promising performance in the past few years, they usually assume that each mention is associated with high-quality, high-value, and high-relevance visual images, which is unrealistic in real-world social media scenarios. Figure 1 shows several social media examples and MEL still faces the pervasive challenges presented by the following poor visual images:

Low-quality. User-posted images may appear blurry, unclear, or pixelated due to factors such as image compression, low original image quality and network transmission issues.

Low-value. As shown in Figure 1(b), this image is widely used in social media, which only expresses the happy mood of users in most cases. However, this type of images does not enhance the overall comprehension of the text and may even mislead the MEL model into linking “Michael” to a person rather than the brand “Michael Kors”.

Low-relevance. Accompanying images often correspond to the whole context of text, rather than being specifically tied to the mention. Figure 1(c) is such an example whose image is low-relevance to the mention “Michael” in text, which will interfere with MEL models matching “Michael Jordan”.

There is a huge gap between all of these poor visual images and the images of their corresponding entities. Integrating them directly will backfire, leading to a weakened consistency between mentions and their corresponding entities. Some previous works (Zhang et al., 2021) have realized the negative impact of noisy images. They propose a straightforward method to remove these poor images by assessing the correlation score between the mention image and mention text. However, the low-quality and low-relevance images still possess valuable implicit visual cues. Discarding them outright would result in the loss of significant complementary visual information. Thus, while the straightforward approach is effective in handling low-value images, it is still not ideal for dealing with low-quality and low-relevance images.

To address these challenges, in this paper, we propose a novel **Multimodal Entity Linking** framework with **Optimized Visual** features in latent space, **MELOV** in short. Our MELOV contains two perspectives of optimization, i.e. inter-modality and intra-modality. Specifically, for the inter-modality optimization, we exploit variational autoencoder to mine shared semantic information from heterogeneous textual features and generate latent vision-specific features. For the intra-

modality optimization, we observe that similar or related mentions often possess related visual information, which is very useful for optimizing poor images. An example is shown in Figure 1(d), “Michael Jordan” even appears directly in this image. To effectively aggregate the visual information of semantic similar neighbors, we construct mention graphs using correlation scores between mentions and employ graph convolutional network for information propagation within the graphs. Finally, our MELOV adaptively assesses the contributions of original visual features, inter-modality generated visual features, and intra-modality aggregated visual features to the MEL task, and fuses them in varying proportions within the latent space. The optimized visual features can simultaneously handle all types of poor images.

The main contributions of this paper are summarized as follows:

- For the first time, we analyze various types of poor images in multimodal entity linking and propose to optimize visual features in latent space, aiming to not only eliminate the negative effects of poor images but also retain the implicit visual cues of original images.
- We propose a joint optimization framework that incorporates inter-modality generation and intra-modality aggregation, effectively leveraging the shared information from heterogeneous textual features and relevant visual details of semantic similar neighbors.
- We compare our MELOV with state-of-the-art MEL approaches on three public benchmark datasets. Experimental results demonstrate the superiority of our proposed framework.

2 Related Work

2.1 Entity Linking

Traditional entity linking (EL) mainly focuses on text-only corpus, which has been widely explored in recent years. EL methods can be roughly divided into two groups based on their granularity: local-level and global-level. The local-level approaches (Francis-Landau et al., 2016; Cao et al., 2017; Eschel et al., 2017; Gupta et al., 2017; Peters et al., 2019; Wu et al., 2020; Sui et al., 2022, 2023) separately link each mention according to the similarity between mentions with context and entities. The global-level approaches (Le and Titov, 2018; Yang

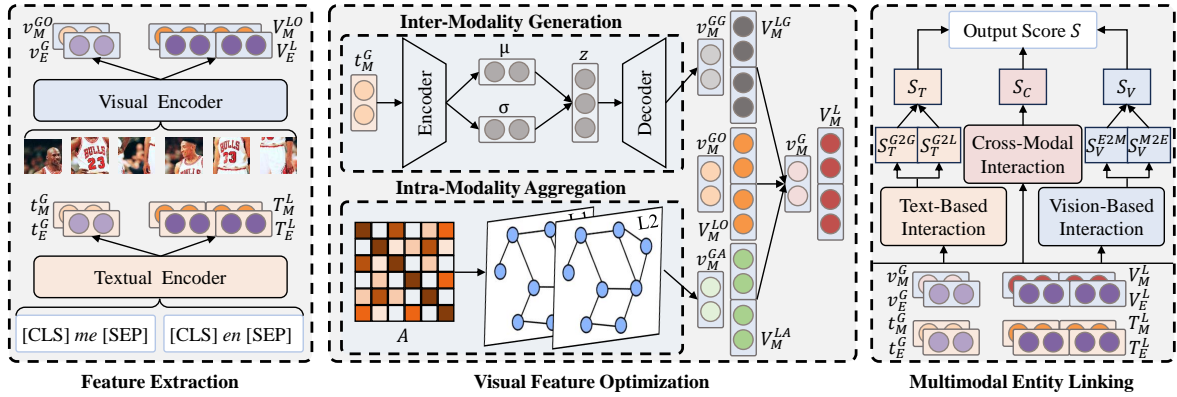


Figure 2: The overall architecture of our proposed MELOV framework, which contains three components: feature extraction, visual feature optimization, and multimodal entity linking.

et al., 2018; Fang et al., 2019; Yang et al., 2019) strive to jointly disambiguate mentions by considering the global coherence of entities within the same document. While these methods have achieved significant progress, they are designed solely for text modality and cannot effectively integrate abundant complementary information from visual images. This limits the performance of EL methods in the surge of multimodal information and motivates researchers to study multimodal entity linking.

2.2 Multimodal Entity Linking

Multimodal entity linking (MEL) is an extension of EL that utilizes additional visual images to aid in disambiguating mentions. Moon et al. (2018) first proposes the task and leverages cross-modal attention to fuse features. Adjali et al. (2020) designs concatenation operation and triplet loss to interact modality. Wang et al. (2022a) proposes a hierarchical multimodal co-attention to mine fine-grained relationships across modalities. Zhang et al. (2023) combines global and bottleneck fusions to integrate multimodal information. Shi et al. (2023) leverages the in-context learning of LLMs to generate target entities. Xing et al. (2023) explicitly models and dynamically selects different alignment relations between mentions and entities. Luo et al. (2023) uses multi-grained interactions and unit-consistent learning to enhance the representations.

In general, all these methods mainly focus on exploring complex multimodal interaction mechanisms to better capture coherence evidence between mentions and entities. However, they overlook the prevalence of poor visual images accompanying text for MEL. Integrating them directly will backfire, leading to a weakened consistency between mentions and their corresponding entities. To re-

move the negative impact of noisy images, Zhang et al. (2021) assesses the correlation scores between the text and image of mentions to filter out poor images. However, many poor images still possess valuable implicit visual cues. Discarding them outright would result in the loss of significant complementary visual information. This motivates us to find another way, that is optimizing visual features in latent space to not only eliminate the negative effects of poor images but also retain the implicit visual cues of original images.

3 Methodology

Figure 2 shows the overall architecture of our MELOV. We first extract both global and local features of textual and visual inputs to obtain global descriptive semantics and preserve fine-grained details of words or image patches. Afterwards, we optimize visual features at two perspectives: inter-modality generation based on the shared information of heterogeneous textual modality and intra-modality aggregation based on the correlation of mentions. We obtain the optimized visual features by adaptively fusing the two perspective features and original features. Finally, we employ multi-grained interactions using optimized visual features as input to derive the similarity matching score for each mention-entity pair to make the ultimate multimodal entity linking decisions.

3.1 Feature Extraction

To extract meaningful textual features, following Luo et al. (2023), we utilize pre-trained BERT (Devlin et al., 2019) as our textual encoder. The inputs of each mention m_i and entity e_i are respectively constructed as: $[CLS] me [SEP]$ and $[CLS]$

en [SEP], where *me* = *mention* [SEP] *sent* and *en* = *title* [SEP] *attr*. The *mention*, *sent*, *title*, and *attr* refer to word-piece tokens of the mention, the sentence where the mention is located, the entity title, and the attributes associated with the entity. Then we feed the inputs into BERT and obtain the hidden states of all word tokens T_{M_i} and T_{E_i} . Finally, we regard the hidden state corresponding to the position of special [CLS] token as global textual features $t_{M_i}^G$ and $t_{E_i}^G$, the entire hidden states as local textual features $T_{M_i}^L$ and $T_{E_i}^L$.

To capture expressive visual features, following Luo et al. (2023), we employ pre-trained Vision Transformer (ViT) (Dosovitskiy et al., 2021) as our visual encoder. Given an image of a mention or entity, we first rescale it into $C \times H \times W$ pixels and reshape it into $H \times W/P^2$ flattened 2D patches. Here, C represents the number of channels, $H \times W$ denotes the image resolution and P is the patch size. Then, we feed them into ViT and obtain the hidden states of all patches. Similar to the textual feature extraction, we regard the hidden state of the [CLS] as global visual features $v_{M_i}^{GO}$ or $v_{E_i}^G$, the whole hidden states as local visual features $V_{M_i}^{LO}$ or $V_{E_i}^L$.

3.2 Visual Feature Optimization

In multimodal entity linking, heterogeneous textual features encompass abundant shared semantic information of visual and textual modalities, such as the characteristics and categories of mentions. On the other hand, visual features of similar or related mentions contain rich vision-specific details associated with the mention. Both the two perspectives are significant for optimizing visual features.

3.2.1 Inter-Modality Generation

For inter-modality optimization, we propose to utilize the cross-reconstruction of variational autoencoder (VAE) (Yi et al., 2023) to mine shared semantic information from heterogeneous textual features and generate vision-specific features in latent space. Specifically, the VAE encoder takes the mention global textual features $t_{M_i}^G$ as input to feed-forward neural networks (FFNNs) to derive latent mean vector μ and standard deviation vector σ : $\mu = t_{M_i}^G W_\mu$ and $\sigma = t_{M_i}^G W_\sigma$, where W_μ and $W_\sigma \in \mathbb{R}^{d_t \times d_z}$ are trainable weights, d_t is the dimension of textual features. The two vectors jointly describe the distribution of latent space $z \sim q(z|t_{M_i}^G) = \mathcal{N}(\mu, \sigma^2)$. We utilize the reparameterization strategy (Kingma and Welling, 2014) to sample the latent variation: $z = \mu + \sigma \odot \epsilon$,

where $\epsilon \sim \mathcal{N}(0, \mathbf{I})$. Then the latent variation z is fed into the VAE decoder to generate global visual features $v_{M_i}^{GG} = zW_d$, where $W_d \in \mathbb{R}^{d_z \times d_v}$ and d_v is the dimension of visual features. Based on the evidence lower bound function (Kingma and Welling, 2014) for VAE, the training loss of our global feature generation is computed as follows:

$$\mathcal{L}_{GG} = \|v_{M_i}^{GO} - v_{M_i}^{GG}\|^2 + \text{KL}(q(z|t_{M_i}^G)||p(z)) \quad (1)$$

By constructing visual features as accurately as possible, our model learns to mine shared semantic information from heterogeneous textual features within latent space, enabling the generation of text-based vision-specific features. Similarly, for local features, we also generate $V_{M_i}^{LG}$ by utilizing VAE and calculate the local loss \mathcal{L}_{LG} . The final loss of our inter-modality generation is calculated as:

$$\mathcal{L}_G = \mathcal{L}_{GG} + \mathcal{L}_{LG} \quad (2)$$

3.2.2 Intra-Modality Aggregation

To effectively leverage intra-modality related visual information of semantic similar neighbors, we propose to construct mention graphs and employ graph convolutional network (GCN) (Kipf and Welling, 2017) for information aggregation. Specifically, considering that the textual modality provides the most reliable information about mentions, we calculate the similarity between mentions by utilizing textual representations. The common approach for this is to calculate the Cosine or Euclidean distance. However, these traditional similarity measures are not sufficient in accurately capturing the local manifold structure and are unable to capture complex relationships such as higher-order statistics. Inspired by Kang et al. (2017); Zhang et al. (2022), we utilize a kernel-driven approach to calculate the similarity for two given mentions m_i and m_j :

$$\text{sim}(m_i, m_j) = \exp\left(-\frac{\|t_{M_i}^G - t_{M_j}^G\|_2^2}{2b^2}\right) \quad (3)$$

where b represents the bandwidth used to regulate the emphasis given to the similarity of small distances relative to large distances. We determine this value by setting it as a fraction of the average distance between mentions. For all mentions in the training set, the similarity matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ is created by assessing the $\text{sim}(m_i, m_j)$ of each pair of mentions, where n is the number of mentions.

In our similarity matrix, all cells are always positive, indicating that all mentions are treated as

similar. However, this will introduce noise from dissimilar mentions. Thus, we employ a learnable threshold to filter out these dissimilar mentions:

$$\mathbf{A}_{ij} = \begin{cases} \mathbf{A}_{ij} & \text{if } \mathbf{A}_{ij} > \tau \\ 0 & \text{if } \mathbf{A}_{ij} \leq \tau \end{cases} \quad (4)$$

where τ is the learnable threshold. We formulate the original global visual features of all mentions $\mathbf{v}_M^{GO} \in \mathbb{R}^{n \times d_v}$ as a graph, with the similarity matrix \mathbf{A} as the graph adjacency matrix. Then, we employ GCN to facilitate the information propagation within the graph, thereby aggregating visual information from semantic similar neighbors. A GCN layer is a nonlinear transformation which maps from \mathbf{v}_M^l to \mathbf{v}_M^{l+1} , defined as:

$$\mathbf{v}_M^{l+1} = \phi(\mathbf{A}\mathbf{v}_M^l\mathbf{W}_g + \mathbf{b}_g) \quad (5)$$

where l is the current layer index, $\mathbf{v}_M^0 = \mathbf{v}_M^{GO}$, $\mathbf{W}_g \in \mathbb{R}^{d_v \times d_v}$ and $\mathbf{b}_g \in \mathbb{R}^{d_v}$, ϕ is a non-linear activation function. After L layers of GCN, we obtain the intra-modality aggregated global visual features \mathbf{v}_M^{GA} . Similarly, for the local features, we formulate \mathbf{V}_M^{LO} as a graph and feed them into another GCN to obtain intra-modality aggregated local visual features \mathbf{V}_M^{LA} .

3.2.3 Information Adaptive Fusion

Now we possess three different types of visual features: original visual features focus on original visual images themselves, inter-modality generated visual features emphasize shared semantic information from heterogeneous textual features, and intra-modality aggregated visual features highlight related visual details of semantic similar neighbors. To take all of them into consideration, we utilize the attention mechanism to adaptively assess their contributions to the MEL task and fuse them in different proportions. Specifically, for global visual features, we put them into different FFNNs and utilize the Sigmoid function to obtain the contribution weights α^O , α^G and α^A . Furthermore, we add a constraint $\alpha^O + \alpha^G + \alpha^A = 1$ by calculating $\alpha^O = \frac{\alpha^O}{\alpha^O + \alpha^G + \alpha^A}$, $\alpha^G = \frac{\alpha^G}{\alpha^O + \alpha^G + \alpha^A}$, $\alpha^A = 1 - \alpha^O - \alpha^G$. The optimized global visual features are obtained by calculating the weighted sum of different global visual features:

$$\mathbf{v}_M^G = \alpha^O \cdot \mathbf{v}_M^{GO} + \alpha^G \cdot \mathbf{v}_M^{GG} + \alpha^A \cdot \mathbf{v}_M^{GA} \quad (6)$$

Similarly, we also utilize the attention mechanism to obtain the optimized local visual features \mathbf{V}_M^L .

Num. of	WikiMEL	RichpediaMEL	WikiDiverse
Sentences	22,070	17,724	7,405
M. in train	18,092	12,463	11,351
M. in valid	2,585	1,780	1,664
M. in test	5,169	3,562	2,078
Entities	109,976	160,935	132,460

Table 1: Overall statistics of WikiMEL, RichpediaMEL and WikiDiverse datasets. M. denotes Mentions.

3.3 Multimodal Entity Linking

To derive matching scores for each mention-entity pair, following Luo et al. (2023), we utilize multi-grained multimodal interaction, which contains text-based, vision-based and cross-modal interactions. Our text-based interaction involves two scores. The global-to-global score measures the global consistency by using dot product: $S_T^{G2G} = \mathbf{t}_M^G \cdot \mathbf{t}_E^G$. And the global-to-local score captures fine-grained context consistency clues between entity global features and mention context vector: $S_T^{G2L} = \mathbf{t}_E^G \cdot h_t$, where the mention context vector h_t is obtained using attention mechanism, mean pooling MP and layer norm LN: $h_t = \text{LN}(\text{MP}(\text{softmax}(\frac{QK^T}{\sqrt{d_t}})V))$, $Q = \mathbf{T}_E^L$, $K = V = \mathbf{T}_M^L$. We obtain the text-based matching score by averaging the two scores: $S_T = (S_T^{G2G} + S_T^{G2L})/2$.

In our vision-based interaction, we employ a dual-gated mechanism to consider both mention and entity views. For entity-to-mention interaction, we first interact entity global and mention local features: $h_{vc} = \text{LN}(\mathbf{v}_E^G + \text{MP}(\mathbf{V}_M^L))\mathbf{W}_{v1}$, where $\mathbf{W}_{v1} \in \mathbb{R}^{d_v \times d_v}$. Then, we interact mention global features with a gate operation: $h_v = \text{LN}(\text{Tanh}(h_{vc}\mathbf{W}_{v2}) \cdot h_{vc} + \mathbf{v}_M^G)$, where $\mathbf{W}_{v2} \in \mathbb{R}^{d_v \times 1}$. After sufficient interaction and fusion, we obtain the matching score from entities to mentions $S_V^{E2M} = h_v \cdot \mathbf{V}_E^G$ and similarly from mentions to entities S_V^{M2E} . The vision-based matching score is the average of them: $S_V = (S_V^{E2M} + S_V^{M2E})/2$.

Our cross-modal interaction measures the consistency between mentions and entities after integrating visual and textual information. We first use FFNNs to map the textual and visual features \mathbf{t}_E^G , \mathbf{t}_M^G , \mathbf{V}_E^L and \mathbf{V}_M^L to shared dense space features h_{et} , h_{mt} , H_{ev} and H_{mv} . To aggregate entity image patch information, we employ correlation scores between textual and visual features as guidance: $\alpha_c^i = \frac{\exp(h_{et} \cdot H_{ev}^i)}{\sum_{i=1}^{n_p} \exp(h_{et} \cdot H_{ev}^i)}$, $h_{ec} = \sum_{i=1}^{n_p} \alpha_c^i \cdot H_{ev}^i$, where n_p is the patch number. Then, we interact the textual features with a gate operation to obtain the entity cross-modal representation $h_e = \text{LN}(\text{Tanh}(h_{et}\mathbf{W}_c) \cdot h_{et} + h_{ec})$, where

Models	WikiMEL				RichpediaMEL				WikiDiverse			
	H@1	H@3	H@5	MRR	H@1	H@3	H@5	MRR	H@1	H@3	H@5	MRR
BLINK	74.66	86.63	90.57	81.72	58.47	81.51	88.09	71.39	57.14	78.04	85.32	69.15
BERT	74.82	86.79	90.47	81.78	59.55	81.12	87.16	71.67	55.77	75.73	83.11	67.38
CLIP	83.23	92.10	94.51	88.23	67.78	85.22	90.04	77.57	61.21	79.63	85.18	71.69
ViLT	72.64	84.51	87.86	79.46	45.85	62.96	69.80	56.63	34.39	51.07	57.83	45.22
METER	72.46	84.41	88.17	79.49	63.96	82.24	87.08	74.15	53.14	70.93	77.59	63.71
DZMNED	78.82	90.02	92.62	84.97	68.16	82.94	87.33	76.63	56.90	75.34	81.41	67.59
JMEL	64.65	79.99	84.34	73.39	48.82	66.77	73.99	60.06	37.38	54.23	61.00	48.19
VELML	76.62	88.75	91.96	83.42	67.71	84.57	89.17	77.19	54.56	74.43	81.15	66.13
GHMFC	76.55	88.40	92.01	83.36	72.92	86.85	90.60	80.76	60.27	79.40	84.74	70.99
MIMIC	87.98	95.07	96.37	91.82	81.02	91.77	94.38	86.95	63.51	81.04	86.43	73.44
MELOV	88.91	95.61	96.58	92.32	84.14	92.81	94.89	88.80	67.32	83.69	87.54	76.57
MELOV w/o inter	88.25	95.21	96.40	91.92	83.72	92.22	94.44	88.74	65.93	82.24	87.20	75.12
MELOV w/o intra	88.52	95.37	96.48	92.03	83.58	92.03	94.40	88.40	64.73	81.94	86.80	75.06
MELOV w/o τ	88.82	95.45	96.51	92.14	83.63	92.71	94.51	88.66	65.28	82.28	87.18	75.36
MELOV (filter)	88.24	95.18	96.37	91.85	80.66	90.34	93.63	86.36	62.75	80.13	86.14	72.87

Table 2: Experimental results on three MEL datasets. We report the results of baselines according to Luo et al. (2023). The results of our MELOV are averaged 5 runs using different random seeds. The highest values are in bold.

$W_c \in \mathbb{R}^{d_c \times d_c}$. Similarly, we obtain the mention representation h_m by replacing h_{et} and H_{ev} with h_{mt} and H_{mv} . We compute the cross-modal matching score using the dot product: $S_C = h_e \cdot h_m$.

The final score is defined as the average of the three scores $S = (S_T + S_V + S_C)/3$. To maximize the score of the corresponding entity among others and ensure to learn good representations in each interaction, we employ an in-batch contrastive training approach that considers all interactions as our loss function. For each mention-entity pair (m_i, e_i) in a batch of B pairs, the training loss is computed as follows:

$$\mathcal{L}_E = \sum_{X \in \{\emptyset, T, V, C\}} -\log\left(\frac{\exp(S_X(m_i, e_i))}{\sum_j^B \exp(S_X(m_i, e_j))}\right) \quad (7)$$

where e_i is the gold entity of the mention m_i . Eventually, the loss of our MELOV is calculated as:

$$\mathcal{L} = \mathcal{L}_E + \lambda \mathcal{L}_G \quad (8)$$

where λ is the hyperparameter to control the loss.

4 Experiments

4.1 Datasets and Baselines

We evaluate our MELOV under three public multimodal entity linking datasets: WikiMEL (Wang et al., 2022a), RichpediaMEL (Wang et al., 2022a), and WikiDiverse (Wang et al., 2022b). Table 1 shows the overall statistics of these three datasets. For fair comparison, following previous works (Wang et al., 2022a; Luo et al., 2023), we utilize a subset KB of Wikidata as our entity set in each dataset. For WikiMEL and RichpediaMEL, the

data is split into 70% training, 10% validation and 20% test sets respectively. As for WikiDiverse, the proportions are 80%, 10% and 10%.

For the quantitative evaluation of our proposed MELOV, we utilize the following competitive methods for comparison. The first type of baselines is the text-based methods that only use the textual modality information, including BLINK (Wu et al., 2020), and BERT (Devlin et al., 2019). Another type is the language-and-vision pre-training models, including CLIP (Radford et al., 2021), ViLT (Kim et al., 2021), and METER (Dou et al., 2022). Furthermore, our baselines also contain state-of-the-art MEL methods, including DZMNED (Moon et al., 2018), JMEL (Adjali et al., 2020), VELML (Zheng et al., 2022), GHMFC (Wang et al., 2022a), and MIMIC (Luo et al., 2023).

4.2 Implementation Details

In our experiments, following (Luo et al., 2023), we initial the weights of BERT and ViT by using the pre-trained CLIP-Vit-Base-Patch32 version. The maximum sequence length of words for textual input is set to 40. All images are rescaled into 224×224 resolution and the patch size P is 32. We use the AdamW (Loshchilov and Hutter, 2019) optimizer with a batch size of 128 for optimizing. We train our MELOV 20 epochs with a learning rate of $1e-5$. Experiments were conducted on a PC with 256 GB RAM, 4 Intel(R) Xeon(R) Gold 6226R CPUs and an NVIDIA GeForce RTX A6000 GPU with 48 GB memory. For evaluation, we utilize the H@k and MRR as our metrics. H@k represents the hit rate of the corresponding entity among the top-k ranked entities. And MRR indicates the mean

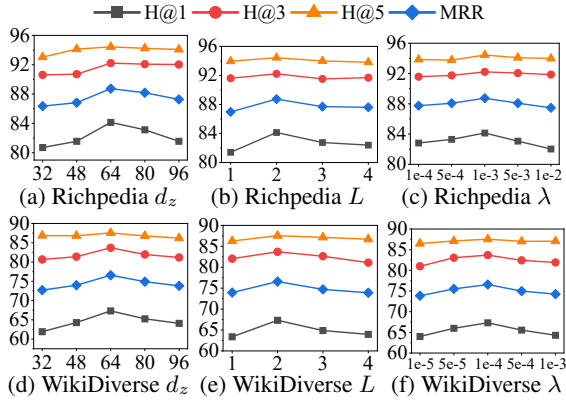


Figure 3: Performance of our MELOV with varying hyperparameters on RichpediaMEL and WikiDiverse.

reciprocal rank of the corresponding entity.

4.3 Overall Performance

Table 2 shows the experimental results of our MELOV in comparison with baselines on three MEL benchmark datasets. We have the following observations. Firstly, the text-based methods BLINK and BERT perform unsatisfactorily since they only rely on textual inputs and ignore visual information. Secondly, some multimodal methods, such as ViLT and JMEL, exhibit inferior performance compared to text-based methods. This observation indicates that shallow modality interaction and naive multimodal fusion do not enhance, and may even degrade, the performance of MEL. Thirdly, most multimodal methods, such as CLIP, GHMFC and MIMIC, perform significantly better than text-based methods. And MIMIC achieves the best results among all baselines, which demonstrates the effectiveness of integrating visual information and the superiority of our chosen multi-grained interaction in our multimodal entity linking module. Finally, our MELOV outperforms all baselines on all three datasets and achieves new state-of-the-art performance. Specifically, MELOV gains 0.93%, 3.12% and 3.81% absolute improvement of H@1 on WikiMEL, RichpediaMEL and WikiDiverse respectively. This suggests the effectiveness of optimizing visual features in latent space and the superiority of our MELOV.

4.4 Ablation Study

To better understand our proposed MELOV, we conduct a series of ablation studies, as also presented in Table 2. We can observe that MELOV w/o inter-modality generation and MELOV w/o intra-modality aggregation both result in a de-

Metrics	WikiMEL		RichpediaMEL		WikiDiverse	
	H@1	MRR	H@1	MRR	H@1	MRR
Dot	88.05	91.92	83.00	87.95	64.70	74.75
Cosine	87.53	91.54	83.12	88.26	65.39	75.17
Euclidean	88.33	92.08	83.25	88.56	65.60	75.42
Kernel	88.91	92.32	84.14	88.80	67.32	76.57

Table 3: Performance with different similarity metrics of our MELOV on three MEL datasets.

crease in performance, confirming the effectiveness of these two optimizations. The performance of MELOV w/o inter drops more on WikiMEL, while MELOV w/o intra drops more on RichpediaMEL and WikiDiverse. This is reasonable since different datasets have distinct optimization requirements. We also find that our MELOV outperforms MELOV w/o τ , highlighting the importance of utilizing such a learnable threshold in the similarity matrix to avoid the noise brought by dissimilar mentions. Finally, we replace the whole optimization process with the filtering way proposed by Zhang et al. (2021). We observe that this leads to a significant performance drop and MELOV (filter) even performs worse than MIMIC on RichpediaMEL and WikiDiverse. This is consistent with our claim that many poor images still possess valuable implicit visual cues. Discarding them outright would result in the loss of significant complementary visual information, ultimately causing a significant degradation in performance. And our method, optimizing visual features in latent space, is a better way to handle poor images.

4.5 Hyperparameter Analysis

To investigate the impact of hyperparameters on the performance of our MELOV, we vary them (the latent dimension size d_z , number of GCN layers L , and final loss hyperparameter λ) on RichpediaMEL and WikiDiverse, which is shown in Figure 3. Firstly, we find that d_z has a significant impact on the final performance by influencing the effect of generating visual features. A small dimension may make our model inadequate to mine shared information from heterogeneous textual features, while a large one may cause redundancy to lead to performance degradation. Secondly, increasing the number of GCN layers does not always lead to improved performance. Excessive message passing and aggregation can potentially exacerbate the issue of data sparsity. Finally, λ , which is used to control the strength of \mathcal{L}_G , also affects the performance. In general, our MELOV obtains best results while setting d_z as 64, L as 2, and λ as 0.001.









Mention	Predictions	Mention	Predictions	Mention	Predictions
 Screenshot of the message YouTube visitors in Turkey used to find.	MELOV  ✓ Turkey Transcontinental country straddling Western Asia ... MIMIC He Zö Co Çp Dat Ee Ft Gç Hh İ İ ğ Kk Ll Mm Nn Oo Qq Pp Rr Ss Şş Tt Uu Vv Ww Yy Zz Turkish Oghuz Turkic language of the Turkish people.	 Taliban launched an attack on a NATO base in Afghanistan, which was repelled soon afterwards.	MELOV  ✓ North Atlantic Treaty Organization Intergovernmental military alliance ... MIMIC  NATO Training Mission-Afghanistan Military organization	 Stephen Colbert has topped a poll to determine the name of an International Space Station module.	MELOV  ✓ International Space Station Modular space station in low ... MIMIC  International Space Station program ongoing space research program...

Figure 4: Case study of our MELOV and MIMIC. There are three cases from the test set of WikiDiverse. Each case contains a mention with multimodal context and the predicted H@1 entities. The symbol ✓ marks the gold entity.

4.6 Analysis of Similarity Calculation

To explore the effect of employing various similarity metrics in computing the GCN similarity matrix on our MELOV performance, we compare our kernel-driven approach with three traditional similarity metrics (i.e. dot product similarity, Cosine similarity and Euclidean distance), as presented in Table 3. Upon observation, Euclidean distance outperforms the other two traditional metrics, indicating its efficacy in capturing numerical disparities among feature nodes. Furthermore, kernel-driven approach demonstrates superior performance in comparison to these traditional metrics, highlighting the effectiveness of our chosen similarity calculation method. This approach can adequately capture local manifold structure and complex relationships such as higher-order statistics.

4.7 Case Study

To conduct a qualitative analysis, Table 4 presents three WikiDiverse cases comparing our MELOV with the current state-of-the-art baseline MIMIC. Notably, all the three cases contain poor mention images. They will mislead MEL models into linking to wrong text-related entities, armed forces related entities and human-related entities respectively, like MIMIC. However, our MELOV effectively leverages the shared information from heterogeneous textual features and relevant visual details of semantic similar neighbors to optimize visual features in latent space to avoid this phenomenon, which helps to make correct decisions.

4.8 Performance on Zero-Shot Setting

With the proliferation of vast amounts of data on the web, new entities are emerging constantly. As shown in the left part of Figure 5, all three datasets contain a large number of unseen samples, whose entities never appear during training. Thus, it is

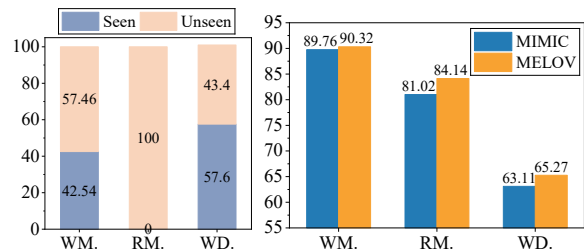


Figure 5: Performance on zero-shot setting. The left part is the proportions of seen and unseen samples in test sets of three datasets. The right part is the H@1 results for unseen samples. WM., RM. and WD. denote WikiMEL, RichpediaMEL and WikiDiverse respectively.

necessary to investigate the performance of MEL models on the zero-shot setting, which mainly focuses on unseen entities. We conduct experiments on these unseen samples and the results are shown in the right part of Figure 5. We can find that our MELOV significantly outperforms the existing state-of-the-art model MIMIC on all three datasets. Through the joint optimization of inter-modality generation and intra-modality aggregation, our MELOV can learn richer and more suitable visual representations, thereby enhancing the generalization ability on new entities.

5 Conclusion

In this paper, we focus on poor images in multimodal entity linking. To avoid the negative effects of poor images while preserving the implicit visual cues of original images, we propose MELOV, a novel joint optimization framework to combine inter-modality generation and intra-modality aggregation to optimize visual features in latent space. This effectively leverages the shared information from heterogeneous textual features and relevant visual details of semantic similar neighbors, allowing for simultaneously handling all types of

poor images. Experimental results on three public multimodal entity linking datasets demonstrate the effectiveness of optimizing visual features in latent space and our proposed MELOV achieves new state-of-the-art performance.

Limitations

Although our MELOV has demonstrated its effectiveness on the multimodal entity linking task, there are still some limitations to be addressed in the future: 1) In this work, for inter-modality generation, we only utilize the variational autoencoder to mine shared semantic information from heterogeneous textual features and generate vision-specific features. However, other state-of-the-art generation frameworks also can be used and may work better. We leave the exploration of other generation frameworks to future work. 2) Another limitation is the resource limitation. We argue that the more resources we leverage, the richer and more suitable the optimized visual features will be. However, for a fair comparison with baselines, we only use the inter-modality heterogeneous textual information and intra-modality related visual information. We will investigate to utilize more resources to optimize visual features in future work.

Acknowledgements

This research is supported by the National Natural Science Foundation of China (No. 62272250, 62302243, U22B2048), the Natural Science Foundation of Tianjin, China (No. 22JCJQC00150, 23JCYBJC01230), and the Fundamental Research Funds for the Central Universities, Nankai University (63241442).

References

- Omar Adjali, Romaric Besançon, Olivier Ferret, Hervé Le Borgne, and Brigitte Grau. 2020. [Multimodal entity linking for tweets](#). In *Advances in Information Retrieval - 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14-17, 2020, Proceedings, Part I*, volume 12035 of *Lecture Notes in Computer Science*, pages 463–478. Springer.
- Roi Blanco, Giuseppe Ottaviano, and Edgar Meij. 2015. [Fast and space-efficient entity linking for queries](#). In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, WSDM 2015, Shanghai, China, February 2-6, 2015*, pages 179–188. ACM.
- Yixin Cao, Lifu Huang, Heng Ji, Xu Chen, and Juanzi Li. 2017. [Bridge text and knowledge by learning multi-prototype entity mention embedding](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1623–1633. Association for Computational Linguistics.
- Andrew Chisholm and Ben Hachey. 2015. [Entity disambiguation with web links](#). *Trans. Assoc. Comput. Linguistics*, 3:145–156.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.
- Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. [An image is worth 16x16 words: Transformers for image recognition at scale](#). In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.
- Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Chenguang Zhu, Pengchuan Zhang, Lu Yuan, Nanyun Peng, Zicheng Liu, and Michael Zeng. 2022. [An empirical study of training end-to-end vision-and-language transformers](#). In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 18145–18155. IEEE.
- Yotam Eshel, Noam Cohen, Kira Radinsky, Shaul Markovitch, Ikuya Yamada, and Omer Levy. 2017. [Named entity disambiguation for noisy text](#). In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017), Vancouver, Canada, August 3-4, 2017*, pages 58–68. Association for Computational Linguistics.
- Zheng Fang, Yanan Cao, Qian Li, Dongjie Zhang, Zhenyu Zhang, and Yanbing Liu. 2019. [Joint entity linking with deep reinforcement learning](#). In *The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*, pages 438–447. ACM.
- Matthew Francis-Landau, Greg Durrett, and Dan Klein. 2016. [Capturing semantic similarity for entity linking with convolutional neural networks](#). In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San*

- Diego California, USA, June 12-17, 2016, pages 1256–1261. The Association for Computational Linguistics.
- Jingru Gan, Jinchang Luo, Haiwei Wang, Shuhui Wang, Wei He, and Qingming Huang. 2021. [Multimodal entity linking: A new dataset and A baseline](#). In *MM '21: ACM Multimedia Conference, Virtual Event, China, October 20 - 24, 2021*, pages 993–1001. ACM.
- Nitish Gupta, Sameer Singh, and Dan Roth. 2017. [Entity linking via joint encoding of types, descriptions, and context](#). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 2681–2690. Association for Computational Linguistics.
- Zhipeng Huang, Bogdan Cautis, Reynold Cheng, Yudian Zheng, Nikos Mamoulis, and Jing Yan. 2018. [Entity-based query recommendation for long-tail queries](#). *ACM Trans. Knowl. Discov. Data*, 12(6):64:1–64:24.
- Zhao Kang, Chong Peng, and Qiang Cheng. 2017. [Kernel-driven similarity learning](#). *Neurocomputing*, 267:210–219.
- Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. [Vilt: Vision-and-language transformer without convolution or region supervision](#). In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 5583–5594. PMLR.
- Diederik P. Kingma and Max Welling. 2014. [Auto-encoding variational bayes](#). In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*.
- Thomas N. Kipf and Max Welling. 2017. [Semi-supervised classification with graph convolutional networks](#). In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.
- Phong Le and Ivan Titov. 2018. [Improving entity linking by modeling latent relations between mentions](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 1595–1604. Association for Computational Linguistics.
- Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. 2021. [Entity-based knowledge conflicts in question answering](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 7052–7063. Association for Computational Linguistics.
- Ilya Loshchilov and Frank Hutter. 2019. [Decoupled weight decay regularization](#). In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.
- Pengfei Luo, Tong Xu, Shiwei Wu, Chen Zhu, Linli Xu, and Enhong Chen. 2023. [Multi-grained multimodal interaction network for entity linking](#). In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2023, Long Beach, CA, USA, August 6-10, 2023*, pages 1583–1594. ACM.
- Seungwhan Moon, Leonardo Neves, and Vitor Carvalho. 2018. [Multimodal named entity disambiguation for noisy social media posts](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 2000–2008. Association for Computational Linguistics.
- Matthew E. Peters, Mark Neumann, Robert L. Logan IV, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A. Smith. 2019. [Knowledge enhanced contextual word representations](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 43–54. Association for Computational Linguistics.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. [Learning transferable visual models from natural language supervision](#). In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.
- Senbao Shi, Zhenran Xu, Baotian Hu, and Min Zhang. 2023. [Generative multimodal entity linking](#). *CoRR*, abs/2306.12725.
- Xuhui Sui, Ying Zhang, Kehui Song, Baohang Zhou, Xiaojie Yuan, and Wensheng Zhang. 2023. [Selecting key views for zero-shot entity linking](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 1303–1312. Association for Computational Linguistics.
- Xuhui Sui, Ying Zhang, Kehui Song, Baohang Zhou, Guoqing Zhao, Xin Wei, and Xiaojie Yuan. 2022. [Improving zero-shot entity linking candidate generation with ultra-fine entity type information](#). In *Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022, Gyeongju, Republic of Korea, October 12-17, 2022*, pages 2429–2437. International Committee on Computational Linguistics.

- Peng Wang, Jiangheng Wu, and Xiaohang Chen. 2022a. [Multimodal entity linking with gated hierarchical fusion and contrastive training](#). In *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*, pages 938–948. ACM.
- Sijia Wang, Alexander Hanbo Li, Henghui Zhu, Sheng Zhang, Pramuditha Perera, Chung-Wei Hang, Jie Ma, William Yang Wang, Zhiguo Wang, Vittorio Castelli, Bing Xiang, and Patrick Ng. 2023. [Benchmarking diverse-modal entity linking with generative models](#). In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 7841–7857. Association for Computational Linguistics.
- Xuwu Wang, Junfeng Tian, Min Gui, Zhixu Li, Rui Wang, Ming Yan, Lihan Chen, and Yanghua Xiao. 2022b. [Wikidiverse: A multimodal entity linking dataset with diversified contextual topics and entity types](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 4785–4797. Association for Computational Linguistics.
- Ledell Wu, Fabio Petroni, Martin Josifoski, Sebastian Riedel, and Luke Zettlemoyer. 2020. [Scalable zero-shot entity linking with dense entity retrieval](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 6397–6407. Association for Computational Linguistics.
- Shangyu Xing, Fei Zhao, Zhen Wu, Chunhui Li, Jianbing Zhang, and Xinyu Dai. 2023. [DRIN: dynamic relation interactive network for multimodal entity linking](#). In *Proceedings of the 31st ACM International Conference on Multimedia, MM 2023, Ottawa, ON, Canada, 29 October 2023- 3 November 2023*, pages 3599–3608. ACM.
- Wenhan Xiong, Mo Yu, Shiyu Chang, Xiaoxiao Guo, and William Yang Wang. 2019. [Improving question answering over incomplete kbs with knowledge-aware reader](#). In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 4258–4264. Association for Computational Linguistics.
- Xiyuan Yang, Xiaotao Gu, Sheng Lin, Siliang Tang, Yueting Zhuang, Fei Wu, Zhigang Chen, Guoping Hu, and Xiang Ren. 2019. [Learning dynamic context augmentation for global entity linking](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 271–281. Association for Computational Linguistics.
- Yi Yang, Ozan Irsoy, and Kazi Shefaet Rahman. 2018. [Collective entity disambiguation with structured gradient tree boosting](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 777–786. Association for Computational Linguistics.
- Jing Yi, Yaochen Zhu, Jiayi Xie, and Zhenzhong Chen. 2023. [Cross-modal variational auto-encoder for content-based micro-video background music recommendation](#). *IEEE Trans. Multim.*, 25:515–528.
- Chaohe Zhang, Xu Chu, Liantao Ma, Yinghao Zhu, Yasha Wang, Jiangtao Wang, and Junfeng Zhao. 2022. [M3care: Learning with missing modalities in multimodal healthcare data](#). In *KDD '22: The 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, August 14 - 18, 2022*, pages 2418–2428. ACM.
- Dongjie Zhang and Longtao Huang. 2022. [Multimodal knowledge learning for named entity disambiguation](#). In *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 3160–3169. Association for Computational Linguistics.
- Gongrui Zhang, Chenghuan Jiang, Zhongheng Guan, and Peng Wang. 2023. [Multimodal entity linking with mixed fusion mechanism](#). In *Database Systems for Advanced Applications - 28th International Conference, DASFAA 2023, Tianjin, China, April 17-20, 2023, Proceedings, Part III*, volume 13945 of *Lecture Notes in Computer Science*, pages 607–622. Springer.
- Li Zhang, Zhixu Li, and Qiang Yang. 2021. [Attention-based multimodal entity linking with high-quality images](#). In *Database Systems for Advanced Applications - 26th International Conference, DASFAA 2021, Taipei, Taiwan, April 11-14, 2021, Proceedings, Part II*, volume 12682 of *Lecture Notes in Computer Science*, pages 533–548. Springer.
- Qiushuo Zheng, Hao Wen, Meng Wang, and Guilin Qi. 2022. [Visual entity linking via multi-modal learning](#). *Data Intell.*, 4(1):1–19.