

# Visual Named Entity Linking: A New Dataset and A Baseline

Wenxiang Sun, Yixing Fan, Jiafeng Guo\*, Ruqing Zhang, Xueqi Cheng

CAS Key Lab of Network Data Science and Technology, Institute of Computing Technology,  
Chinese Academy of Sciences, Beijing, China  
University of Chinese Academy of Sciences, Beijing, China  
{sunwenxiang20s,fanyixing,guojiafeng,zhangruqing,cxq}@ict.ac.cn

## Abstract

Visual Entity Linking (VEL) is a task to link regions of images with their corresponding entities in Knowledge Bases (KBs), which is beneficial for many computer vision tasks such as image retrieval, image caption, and visual question answering. While existing tasks in VEL either rely on textual data to complement a multi-modal linking or only link objects with general entities, which fails to perform named entity linking on large amounts of image data. In this paper, we consider a purely Visual-based Named Entity Linking (VNEL) task, where the input only consists of an image. The task is to identify objects of interest (i.e., visual entity mentions) in images and link them to corresponding named entities in KBs. Since each entity often contains rich visual and textual information in KBs, we thus propose three different sub-tasks, i.e., visual to visual entity linking (V2VEL), visual to textual entity linking (V2TEL), and visual to visual-textual entity linking (V2VTEL). In addition, we present a high-quality human-annotated visual person linking dataset, named WIKIPerson. Based on WIKIPerson, we establish a series of baseline algorithms for the solution of each sub-task, and conduct experiments to verify the quality of the proposed datasets and the effectiveness of baseline methods. We envision this work to be helpful for soliciting more works regarding VNEL in the future. The codes and datasets are publicly available at <https://github.com/ict-bigdatalab/VNEL>.

## 1 Introduction

An in-depth understanding of visual content in an image is fundamental for many computer vision tasks. VEL (Tilak et al., 2017; Maigret et al., 2016) is a task to put the image understanding to the entity-level. For example, given an image of the debate between Trump and Hillary, the goal of VEL

\*Corresponding authors.

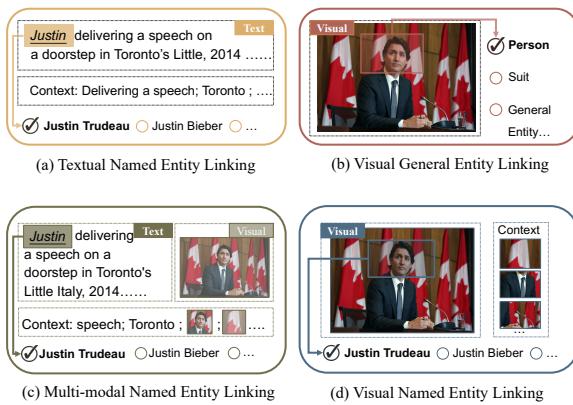


Figure 1: Different categories of Entity Linking. VNEL is a task to identify images individually without any text input and link visual mentions to specific named entities in KBs.

is not only to recognize the region of Trump and Hillary, but also to link them to the correct entity in KBs (e.g., Wikidata (Vrandečić and Krötzsch, 2014), DBpedia (Auer et al., 2007), or YAGO (Fabian et al., 2007)). Just as the significance of textual entity linking for many NLP tasks such as Information Extraction and Information Retrieval (Sevgili et al., 2022), visual tasks, such as image retrieval (Datta et al., 2008) and image caption (Tariq and Foroosh, 2017), would also benefit from entity-level fine-grained comprehension of images.

In recent years, VEL has been given increasing attention. Early works (Tilak et al., 2017; Weegar et al., 2014) try to link objects in images with general entities, e.g., ‘Person’ and ‘Suit’, in KBs as is described in Figure 1(b). Apparently, these works are restricted to the coarse-level entity linking and fail to distinguish objects within the same class. Besides, there are also some works that make use of deep image understanding to link objects with named entities in KBs (Müller-Budack et al., 2021; Zheng et al., 2022; Dost et al., 2020; Gan et al., 2021). However, they generally require detailed entity mention information in text, which plays a vital role via multi-modal entity linking as shown

Dataset	Multi-modal	Entity-aware	Entity-labeled	Modality	KB	Source	Lang	Size
AIDA(Hoffart et al., 2011)		✓	✓	$T^m \rightarrow T^e$	Wikipedia	News social media	en	1K docs
Flicker30K(Young et al., 2014)	✓		✓			News	en	30k images
BreakingNews(Ramisa et al., 2017)	✓		✓			News	en	100k images
SnapCaptionsKB(Moon et al., 2018)	✓	✓	✓	$T^m + V \rightarrow T^e$	Freebase	Social Media	en	12K captions
WIKIDiverse(Wang et al., 2022)	✓	✓	✓	$T^m + V \rightarrow T^e, V^e$	Wikipedia	News	en	8K captions
WIKIPerson		✓	✓	$V^m \rightarrow V^e$ $V^m \rightarrow T^e$ $V^m \rightarrow V^e, T^e$	Wikipedia	News	en	50k Images

Table 1: The public related dataset of WIKIPerson.  $T^m$ ,  $T^e$ ,  $V^m$ ,  $V^e$ , and  $V$  represent textual mention, textual entity, visual mention, visual entity, and visual information, respectively.

in Figure 1(c). We argue that all the above tasks fail to process the named entity linking well for images without any text annotations, which is often the case in social media platforms.

In this work, we consider a purely Visual-based Named Entity Linking (VNEL) task, which is described in Figure 1(d). Given an image without textual description, the goal is to link the visual mention in the image with the whole image as the context to the corresponding named entity in KBs. Considering the format of entity in KBs, such as textual descriptions, images, and other structured attributes, we further introduce three sub-tasks according to the type of entity context, i.e., the visual to visual entity linking (V2VEL), visual to textual entity linking (V2TEL), and visual to visual-textual entity linking (V2VTEL). We believe these tasks could put forward higher requirements and more detailed granularity for image understanding, cross-modal alignment, and multi-modal fusion.

Following the definition of VNEL, currently public available EL datasets may not fit for our research, as they either only focus on textual modality or lack of detailed annotations for entity information in each image. As a result, we release a new dataset called WIKIPerson. The WIKIPerson is a high-quality human-annotated visual person linking dataset based on Wikipedia. Unlike previously commonly-used datasets in EL, the mention in WIKIPerson is only an image containing the PERSON entity with its bounding box. The corresponding label identifies an unique entity in Wikipedia. For each entity in Wikipedia, we provide textual descriptions as well as images to satisfy the need of three sub-tasks.

In the experiments, we benchmark a series of baseline models on WIKIPerson under both zero-shot and fine-tuned settings. In detail, we adopt a universal contrastive learning framework to learn a robust and effective representation for both mentions and entities. Experimental results show that

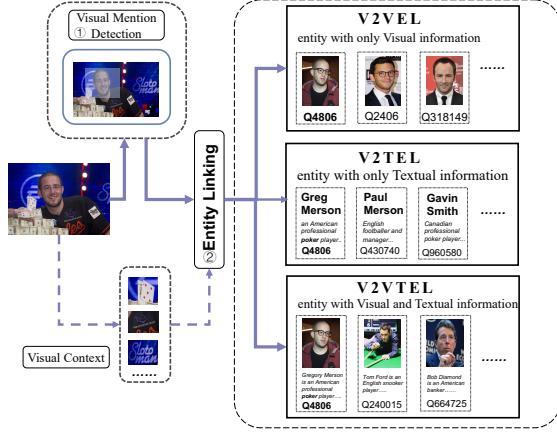


Figure 2: VNEL with its three sub-tasks.

existing models are able to obtain a reasonably good performance on different VNEL tasks, but there is still a large room for further enhancements.

## 2 The Visual Named Entity Linking Task

This section first presents a formal definition of the task. Then we introduce the complete building procedure of the human-annotated dataset, which covers a wide variety of Wikipedia person entities for further research. Finally, an in-depth data analysis will be elaborated on in detail.

### 2.1 Definition of VNEL and Three Sub-tasks

VNEL takes an image as input and extracts bounding boxes around objects, and then links them to entities in KBs. More precisely, given an image  $I$ , all visual mentions  $V^m$ , which are regions of the image, are firstly recognized with a bounding box. Then, all visual mentions  $V^m$  are linked with the corresponding entity  $e$  in knowledge base  $E$ . The visualized process of the VNEL task is shown in Figure 2, which often consists of two stages, namely the visual mention detection stage and the visual entity linking stage. In this work, we follow existing works (Mulang' et al., 2020; Sil et al., 2018) to pay attention to the visual entity linking

stage.

Generally, each entity  $e_i \in E$  is often characterized with rich textual and visual descriptions, and each modality of the description can provide sufficient information for visual entity linking. To make the task more clearly presented, we further decompose the VNEL task into three sub-tasks according to the type of description used for the entity. In the first place, only the visual description  $V_{e_i}$  of the entity can be used in the visual entity linking stage, which we denote as the V2VEL sub-task. The core of V2VEL is to match two visual objects. It is worth noting that entities in KB may contain more than one image. To simply this, we take the first image of  $e_i$  as  $V_{e_i}$ , and leave the multiple images per entity as the future work. In the second place, only the textual description  $T_{e_i}$  of the entity is used in the visual entity linking stage, which we denote as the V2TEL sub-task. The V2TEL task aims to evaluate the ability in image-text matching, central to cross-modal entity linking. Finally, both the visual description and the textual description  $(V_{e_i}, T_{e_i})$  of the entity could be employed to link the visual mention, which we denote as the V2VTEL sub-task. The V2VTEL task could leverage both textual and visual modality to complement each other in linking visual mentions.

Formally, let  $e_i$  represent the  $i^{th}$  entity in KB with corresponding visual description  $V_{e_i}$  or textual description  $T_{e_i}$  and the whole image can be seen as visual context  $V^c$ . As a result, three sub-tasks of the VNEL can be formulated as the following respectively:

$$\begin{aligned} e^*(m) &= \arg \max_{V \rightarrow V} \Phi^\alpha(V^m, V_{e_i} | V^c), \\ e^*(m) &= \arg \max_{V \rightarrow T} \Phi^\beta(V^m, T_{e_i} | V^c), \\ e^*(m) &= \arg \max_{V \rightarrow V+T} \Phi^\gamma(V^m, (V_{e_i}, T_{e_i}) | V^c), \end{aligned}$$

where  $\Phi$  represents the value of the score function between a mention and an entity.

## 2.2 Dataset Setups of WIKIPerson

To facilitate research on VNEL, we introduce WIKIPerson, a benchmark dataset designed for linking person in images with named entities in KB. The dataset building process is shown in Figure 3, which consists of three main steps. We firstly select the data source to build the input image collection, and then filter and clean the collection to obtain a high-quality dataset. Finally, we annotate

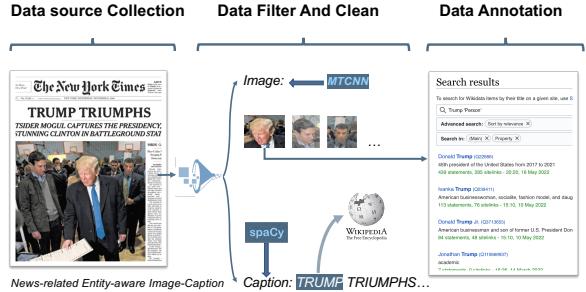


Figure 3: The procedure of building WIKIPerson.

each image by several experienced annotators. In the following, we will describe each step in detail.

### 2.2.1 Data Source Collection

For the source of data, we follow existing works (Ramisa et al., 2017; Tran et al., 2020; Liu et al., 2020; Biten et al., 2019) to use News collections, since the content of images in News collection often contains many named entities at a higher degree of specificity, e.g., specific people, which convey key information regarding the events presented in the images. In this paper, we choose VisualNews<sup>1</sup>, which has the largest data scale with 1.2 million image-text pairs among them as the original data source. In addition, VisualNews covers diverse news topics, consisting of more than one million images accompanied by news articles, image captions, author information, and other metadata. All these additional metadata could help us in the subsequent entity annotation procedure. However, only images and annotated mentions with bounding boxes are available in all VNEL sub-tasks.

For the knowledge base, we employ the commonly-used Wikipedia as back-end, consisting of a wide range and abundant information of entities. Specifically, we crawl the first image of each entity from wiki commons as the visual description and the text information from Wikipedia as the textual description, respectively.

### 2.2.2 Data Filter and Clean

In this work, we pay our attention to PERSON mentions in images since person is the most common named entity, and leave the research on other entity types for future work. For this purpose, we keep only images with PERSON mentions from the news collection, and remove non-PERSON entities from the KB. Specifically, for each image-caption pair in the news collection, we take Spacy to analyze the text caption and filter out the corre-

<sup>1</sup><https://github.com/FuxiaoLiu/VisualNews-Repository>

Input Images	Entity In KB (Visual and Textual Info)
Bounding box: [92, 36, 129, 89] WikiId: Q2808	Elton John (Q2808) English singer and pianist
Bounding box: [65, 31, 95, 81] WikiId: Q459830	Steven Gerrard (Q459830) English association football player and manager
Bounding box: [124, 48, 189, 147] WikiId: Q607	Michael Bloomberg (Q607) American businessman and politician, 108 <sup>th</sup> mayor of NY

Figure 4: Examples of the WIKIPerson dataset. **Left:** An image and its mention’s bounding box with WikiId, which represents an unique entity in Wikipedia. **Right:** The ground truth entity in KB with both visual and textual information.

sponding data without any PERSON entities. Moreover, we leverage the MTCNN model (Zhang et al., 2016), which is the state-of-the-art face detection model, to check the number of PERSON mentions in each image. Then, we select images with the number of person mentions less than 4 to reduce the complexity of the task. Lastly, we remove repeated and blurred images to keep the quality of the dataset.

### 2.2.3 Data Annotation

The primary goal of WIKIPerson is to link the PERSON mention in the image to the correct Wikipedia entity. As a consequence, the annotators need to identify the person mention and label each mention with the corresponding Wikipedia entity in the form of a Wikidata id.<sup>2</sup>

In the earlier step, Spacy is used to identify the caption of origin image-text pairs to extract possible PERSON entities. MTCNN is adopted to recognize the faces, supplying bounding boxes in the picture. So the annotators only need to check the faces in the bounding box and choose the corresponding entity from the results generated by searching the keywords of PERSON entities detected in the caption. In this way, we can largely reduce the labor in labeling the entity of each mention. Mentions that do not have corresponding entities in Wikipedia will be filtered in the procedure.

In the process of data annotation, we designed end-to-end labeling web demos to facilitate manual annotation. The provided information on the website includes news images, captions, news content, and possible candidate entities with pictures

	#Image	#E <sub>cov</sub>	#M <sub>avg</sub> <sup>I</sup>	#KB
WIKIPerson	48K	13k	1.08	120K

Table 2: Statistics of WIKIPerson. #E<sub>cov</sub> and #M<sub>avg</sub><sup>I</sup> denotes number of covered entities and average number of mentions per image, respectively.

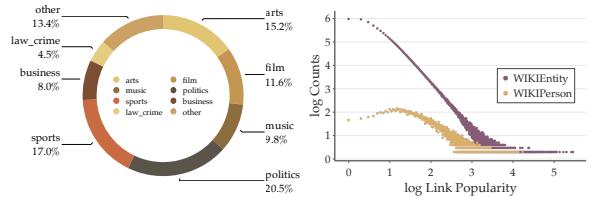


Figure 5: **Left:** Topic distribution of entities in WIKIPerson. **Right:** Link popularity distribution between entities in WIKIPerson and the whole Wikipedia.

and descriptions to help the annotator make judgments. All annotators have linguistic knowledge and are instructed with detailed annotation principles. The annotators need to link the mention with each bounding box to the correct entity in Wikipedia. Finally, after the labeling, we can get the dataset full of the image which comprises several mentions with each bounding box and corresponding entity WikiId.<sup>3</sup>

## 2.3 Dataset Analysis

### 2.3.1 Basic Statistics

Table 2 shows the statistics of the WIKIPerson in detail. The dataset contains a total of 48k different news images, covering 13k out of 120K (i.e. |E| ≈ 120K) PERSON named entities, each of which corresponds to a celebrity in Wikipedia. Many entities appear many times in the data, which ensures that entities can be fully learned. Unlike many datasets in traditional EL, the image of the PERSON named entity usually focuses on a single person in the news except for the scene such as group photo, debate, etc. As a result, the average amount of the mention per image is about 1.08 and only about 3k images contain more than one mention.

### 2.3.2 Entity Distribution

The WIKIPerson comprises diverse PERSON named entity types such as politicians, singers, actresses, sports players, and so on, from different news agencies. These entities do not belong to a

<sup>2</sup><https://en.wikipedia.org/wiki/Wikidata#Concept>

<sup>3</sup>More examples from WIKIPerson are shown in Appendix.

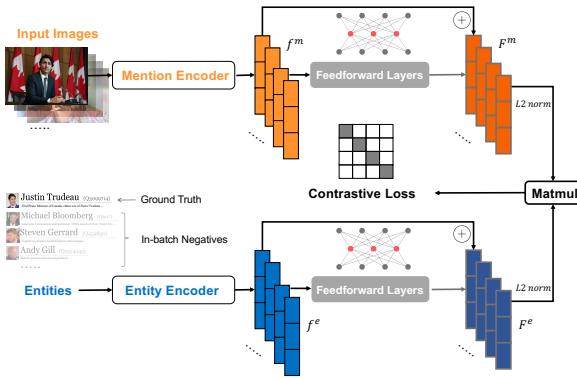


Figure 6: The overall framework of different baselines.

single analogy but are widely distributed in different topics, occupations, skin colors, and multiple age stages. The detailed information is shown in left of Figure 5. It can be observed that in addition to the common politician in the news, the dataset also includes artistic, sports, entertainment, and even criminal topics, which greatly increases the richness of image information. The diversity makes the task could pay attention to the alignment between the background information of the picture, e.g., visual context and entity’s meta info in KBs.

Moreover, considering the difference in entities’ popularity, we analyzed the link-popularity of the entities in the WIKIPerson compared to that in the whole Wikipedia. As shown in the right of Figure 5, both covered entities and the whole Wikipedia entities conform to the long-tailed distribution, which ensures that the dataset will not be biased because of some significantly popular entities. Generally speaking, celebrities are likely to be reported in news articles, which causes the entity in our dataset to be more prevalent than in the whole Wikipedia. To the best of our knowledge, WIKIPerson is the first diverse human-annotated PERSON-entity-aware dataset with high research value.

### 3 Baseline Methods

Generally, the VNEL task is to link mentions in the input image with the corresponding entities from a large-scale KB. Typically, the existing VNEL system is often implemented as a two-stage process, i.e., the candidate retrieval stage and the entity disambiguation stage, to balance the efficiency and the effectiveness. In this work, we implement a fast end-to-end linking directly from a large-scale collection by employing an efficient model.

We take a widely-used bi-encoder contrastive learning framework to learn robust and effective representations of both visual mentions and entities. Given a visual mention  $V^m$  and a candidate entity  $e_i$ , which is accompanied by visual description  $V_{e_i}$  and/or textual description  $T_{e_i}$ , the framework aims to produce a relevance score between the mention and the entity. The overall structure of the framework is shown in Figure 6, which consists of two major components, namely the mention encoder and the entity encoder. These two encoders aim to extract features as embeddings  $f^m$  for the input image and  $f^e$  for the entity. For each encoder, we directly take existing pre-trained models as the implementation. Inspired by existing works (Gao et al., 2021; Zhang et al., 2021b) in applying pre-trained model, we add a feed-forward layer to transform the vector generated from the encoder to the task-oriented embedding space. After that, a residual connection (He et al., 2016) is added to obtain  $F^m$  and  $F^{e_i}$ , followed by using L2 norm and dot-production to calculate the similarity score.

$$\begin{aligned} f^m &= \text{Encoder}^m(V^m), f^{e_i} = \text{Encoder}^e(e_i) \\ F^m &= f^m + \text{ReLU}(f^m \mathbf{W}_1^m) \mathbf{W}_2^m \\ F^{e_i} &= f^{e_i} + \text{ReLU}(f^{e_i} \mathbf{W}_1^e) \mathbf{W}_2^e \\ e^*(m) &= \arg \max_{e_i \in E} F^m \cdot F^{e_i} \end{aligned}$$

Where  $\mathbf{W}_1^m$  and  $\mathbf{W}_2^m$  are learnable parameters for mention representation learning, and  $\mathbf{W}_1^e$  and  $\mathbf{W}_2^e$  are learnable parameters for entity representation learning.

Since each sub-task of VNEL have different types of inputs, we thus implement each baseline with different encoders:

- **V2VEL Encoders:** We adopt ResNet (Szegedy et al., 2017) in a single-modal way following (Schroff et al., 2015), which has been pre-trained on the vggface2 (Cao et al., 2018) to extract visual features. Here, both mention and entity encoder use ResNet and share the parameters.
- **V2TEL Encoders:** We directly take CLIP (Radford et al., 2021), which has been pre-trained with a large-scale image-text dataset, to implement the mention encoder and the entity encoder. For entity encoder, we apply two types of textual information about entity, i.e., entity name (CLIP\_N) and entity name with description (CLIP\_N\_D), to study the influence of the entity’s meta info.

	Sub-Task	Model	Recall				MRR		
			R@1	R@3	R@5	R@10	MRR@3	MRR@5	MRR@10
zero-shot	V2VEL	ResNet	0.3097	0.4053	0.4479	0.5076	0.3518	0.3616	0.3695
		CLIP_N	0.4393	0.5673	0.6145	0.6724	0.4964	0.5071	0.5149
	V2TEL	CLIP_N_D	0.4586	0.5872	0.6323	0.6827	0.5158	0.5260	0.5328
		ResNet+CLIP_N	0.5644	0.6665	0.6981	0.7309	0.6101	0.6174	0.6217
	V2VTEL	ResNet+CLIP_N_D	0.5892	<b>0.6859</b>	<b>0.7102</b>	<b>0.7440</b>	<b>0.6327</b>	<b>0.6383</b>	<b>0.6429</b>
		CLIP_N + ResNet	0.5667	0.6618	0.6893	0.7072	0.6095	0.6158	0.6184
		CLIP_N_D + ResNet	<b>0.5895</b>	0.6794	0.7066	0.7235	0.6302	0.6365	0.6389
fine-tune	V2VEL	ResNet	0.4212	0.5530	0.5832	0.6428	0.4701	0.4821	0.4899
		CLIP_N	0.5527	0.6860	0.7250	0.7756	0.6126	0.6215	0.6285
	V2TEL	CLIP_N_D	0.5946	0.7180	0.7550	0.8022	0.6550	0.6634	0.6697
		ResNet+CLIP_N	0.7171	0.8115	0.8385	0.8634	0.7600	0.7661	0.7696
	V2VTEL	ResNet+CLIP_N_D	0.7301	<b>0.8242</b>	<b>0.8512</b>	<b>0.8798</b>	0.7714	0.7776	<b>0.7815</b>
		CLIP_N + ResNet	0.7180	0.7921	0.8082	0.8177	0.7502	0.7539	0.7552
		CLIP_N_D + ResNet	<b>0.7370</b>	0.8178	0.8347	0.8445	<b>0.7739</b>	<b>0.7778</b>	0.7792

Table 3: Experimental results of baselines among three sub-tasks under both zero-shot and fine-tuned settings.

- **V2VTEL Encoders:** We combine encoders of V2VEL and V2TEL to implement the V2VTEL. Specifically, we take a simple but effective strategy that uses one model to recall Top-K results and the other to re-rank. For example, ResNet + CLIP means recall with the ResNet first and re-rank Top-K results with CLIP again. We also test different combinations about the order of V2VEL encoders and V2TEL encoders, whose results are listed in Section 4.1.<sup>4</sup>

In the training step, the contrastive loss function of a single mention-entity sample is defined as:

$$\mathcal{L}(V^m, e_i) = -\log \left[ \frac{\exp(\Phi(V^m, e_i^+)/\tau)}{\sum_{k \neq i}^- \exp(\Phi(V^m, e_k^-)/\tau)} \right]$$

where  $e_i^+$  represents the ground truth positive entity of  $V_m$  and  $e_k^-$  denotes the  $k^{th}$  candidate of  $V^m$  in the batch, which is all negative samples.  $\tau$  is the temperature coefficient that helps control the softmax's smoothness(Jang et al., 2016).

## 4 Experiments

During experiments, we split images in WIKIPerson into train, dev, and test set with the ratio of 6:2:2. Besides, to avoid the bias of popular entities affecting the evaluation, each named entity appears at most once in test set. For evaluation, we report two widely-used metrics of Top-k retrieve: Recall@K (K=1, 3, 5, 10) and Mean Reciprocal

<sup>4</sup>The detailed analysis of this strategy is displayed in the Appendix due to the page limitation.

Rank (MRR@K, K=3, 5, 10).<sup>5</sup>

### 4.1 Results

All results are summarized in Table 3. Since all the encoders we adopt are pre-trained and can be directly applied in each task, we thus report both zero-shot and fine-tuned performances to show the effectiveness of all baselines.

**Zero-shot v.s. fine-tune.** In zero-shot, we directly use the embedding generated from the encoder as the feature. As we can see, ResNet has achieved a reasonable good performance for R@10 (i.e., 0.5076), which demonstrates the effectiveness of the pre-trained model. Moreover, we can see that the CLIP, which is pre-trained with about 400M image-caption pairs, has achieved better performances against ResNet with either CLIP\_N or CLIP\_N\_D across all metrics. When combining ResNet with CLIP, we observe a distinct improvement for all combination, which demonstrate the effectiveness in combining both visual description and textual description in VNEL. While comparing the zero-shot with fine-tuned baselines, all models have obtained significant improvements, e.g., an average improvement of MRR@10 is 0.13. The improvements verify the quality of dataset and demonstrates that the WIKIPerson could significantly boost the ability of visual named entity linking.

**Sub-tasks of VNEL.** We focus on the below part of table 3 where all models are fine-tuned on WIKIPerson.

- 1) The V2VEL sub-task: As the most funda-

<sup>5</sup>The description about the evaluation metrics can be found in the Appendix.

	Input Images	Linking Results (Top-3)		
(a)		ResNet:  Tiger Woods (Q10993) American golfer	Jason Heyward (Q163399) American Professional baseball player	David Ortiz (Q348000) Dominican-American Hall of Fame baseball player
		CLIP_N_D:  Anirban Lahiri (Q869101) professional golfer	Tiger Woods (Q10993) American golfer	Louis Oosthuizen (Q5614499) Professional golfer
(b)		ResNet:  Andy Gill (Q2524045) British guitarist and record producer	Bill Clinton (Q1124) 42nd president of the United States from 1993 to 2001	Andy Zaltzman (Q4791518) British comedian and author
		CLIP_N_D:  Bill Clinton (Q1124) 42nd president of the United States from 1993 to 2001	Chelsea Clinton (Q2289671) American author and global health advocate	Philippe Reines (Q7184879) Aide to Hillary Clinton
(c)		ResNet:  Celeste Ng (Q1876362) American novelist	Chihiro Fukushima (Q87575) Japanese Olympic sprinter	Priscilla Chan (Q20900227) American pediatrician and philanthropist
		CLIP_N_D:  Lydia Ko (Q6707639) South Korea-born New Zealand golfer	Minjee Lee (Q6867753) Australian professional golfer	Briana Provencher (Q6015992) US Olympic sailing contestant

Figure 7: The qualitative case studies of Top-3 predicted entities. The result with a green border is the ground truth entity of the input image.

mental part concerning VNEL, the ResNet extracts features for both visual mentions and visual descriptions of entities, and matches them in visual feature space. However, it obtains generally low absolute numbers in different evaluation metrics, e.g., 0.4212 on R@1, which leaves a large room for improvement. A possible reason is that the image of an entity in KB are often earlier pictures which show very different state (e.g., age and occasion) with entities appeared in news articles.

2) The V2TEL sub-task: CLIP obtains higher performance compared to ResNet by matching the visual mention with textual descriptions of the entity. Besides, these results show that the cross-modal matching between the image and the text is very powerful in linking images with entities. Moreover, by comparing the two different types of textual information about the entity, we can see that entity description could provide useful information in distinguishing disambiguate entities since CLIP\_N\_D outperforms CLIP\_N over all metrics.

3) The V2VTEL sub-task: By combining the textual information and visual information of each entity, the performance could be further boosted. For example, the relative improvement of ResNet+CLIP\_N\_D over ResNet and CLIP\_N\_D against R@1 is about 73% and 23%, respectively. These results verify that both textual and visual modality of the entity could complement each other

in linking visual mentions with named entities. Moreover, as for different order of combination between ResNet and CLIP, we can see that each method could obtain a relatively close performance, which confirms the effectiveness of the strategy in combining the V2VEL method and the V2TEL method.

## 4.2 Qualitative Analysis

To better understand baseline methods among different sub-tasks, we show several cases in Figure 7. The input image is on the left, and the top 3 predicted results are partitioned into two rows corresponding to different baselines. The entity with a green border is the ground truth entity.

The first case of Figure 7 is a picture of a famous American golfer named Tiger Woods. ResNet could identify the correct entity, and other returned results have a similar face to the input image. CLIP\_N\_D also returned the ground truth entity at the second position in the top-3 results, and all three candidates are professional golfers. This shows that only text descriptions may unable to disambiguate between the correct entity and irrelevant entities.

Analogously, the second case is an image about Bill Clinton speaking at his foundation. ResNet links it to the entity "Andy Gill", which looks very similar to Clinton. While CLIP\_N\_D correctly predicts the ground truth entity in the first position,

and all returned entities are related to Clinton. This verifies that CLIP\_N\_D can learn high-level association between image mention and entity meta-info.

The last case is an image of a famous Chinese tennis sports player named Li Na. We can see that the image has complicated backgrounds, and both ResNet and CLIP\_N\_D cannot link the mention with the ground truth entity in top-3 returned results. This motivates the need for focused research on building effective VNEL models.

From all the above cases, it is clearly presented that ResNet pays more attention to the pixel-level matching, and CLIP learns high-level semantic connection between mentions and entities. However, the dynamic nature of the input images highlights the difficulty of the task, especially for entities with outdated pictures. We believe this work could pave the way for better visual entity linking.

## 5 Related Work

**Entity Linking.** There is extensive research on EL, which serves as a classic NLP task. With the help of large-scale pre-train language models (Devlin et al., 2018; Liu et al., 2019), several recent deep learning methods (Mulgat et al., 2020; Yamada et al., 2019; De Cao et al., 2020) achieve 90%+ accuracy on AIDA (Hoffart et al., 2011), which is a commonly used high-quality robust EL dataset. However, as mentioned in (Cao et al., 2020), it seems that the current methods have already torched the task ceiling. As a result, many more challenging EL-related tasks are formulated. For example, zero-shot entity linking (Logeswaran et al., 2019; Wu et al., 2019), engaging other features like global coherence across all entities in a document, NIL prediction, joining MD and ED steps together, or providing completely end-to-end solutions to address emerging entities is rapidly evolving (Sevgili et al., 2022).

**Multi-modal Entity Linking.** Recently, Multi-modal Entity Linking(MEL) (Moon et al., 2018) task has also been proposed for consideration. Given a text with images attached, MEL uses both textual and visual information to map an ambiguous mention in the text to an entity in the KBs. (Moon et al., 2018) proves that image information helps identify the mention in social media for the fuzzy and short text. Furthermore, (Adjali et al., 2020) transfer the scene to Twitter and perform MEL on Twitter users. (Zhang et al., 2021a) proposes an attention-based structure to eliminate dis-

tracting information from irrelevant images and builds a multi-source Social Media multi-modal dataset. (Wang et al., 2022) builds a multi-modal Entity Linking Dataset with Diversified Contextual Topics and Entity Types. However, for all those works, the text input plays a vital part, and the visual input only serves as a complementary role to the text.

**Multi-modal Dataset.** At the same time, our work is also related to the multi-modal image-text datasets, which is also a hot issue in recent years. Flickr30k (Young et al., 2014) annotates 30k image-caption pairs from Flickr with five descriptive sentences per image, such as "a man is wearing a tie." In addition, MSCOCO caption (Chen et al., 2015) scale up the size with over one and a half million captions describing over 330000 images. However, the caption in all these datasets is descriptive sentences and non-entity aware. As a result, some work has started to build a news-related dataset for entity-aware image caption tasks. For example, (Ramisa et al., 2017) focus on the news website and have crawled 100k image-caption pairs. (Biten et al., 2019; Liu et al., 2020) expand the size of the dataset. Nevertheless, the detailed entity information is neither annotated nor linked to the KBs.

## 6 Conclusion and Future Work

To tackle the limitation that previous visual entity linking either rely on textual data to complement a multi-modal linking or only link objects with general entities, we introduce a purely Visual-based Named Entity Linking task, where the input only contains the image. The goal of this task is to identify objects of interest in images and link them to corresponding named entities in KBs. Considering the rich multi-modal contexts of each entity in KBs, we propose three different sub-tasks, i.e. the V2VEL sub-task, the V2TEL sub-task, and the V2VTEL sub-task. Moreover, we build a high-quality human-annotated visual person linking dataset, named WIKIPerson, which aims at recognizing persons in images and linking them to Wikipedia. Based on WIKIPerson, we introduce several baseline algorithms for each sub-task. According to the experimental results, the WIKIPerson is a challenging dataset worth further explorations. In the future, we intend to build a larger scale VNEL dataset with diverse types and adopt more advanced models to achieve higher accuracy.

## Limitations

**Low extensibility of the entity information.** In the V2VEL sub-task, each entity in the KB can have more than one attached image. However, in our paper, only the first image is selected for convenience, which will inevitably omit additional information. At the same time, in the V2TEL sub-task, we only use the short descriptive sentences of the entity. How to integrate longer unstructured text information is also a problem worth exploring.

## Ethics Statement

We collected data based on open-source datasets and databases. These data have been strictly manually reviewed and do not contain any pictures that are sexual or violate politics. We are authorized by the relevant authority in our university to hire employees from the laboratory to build the platform and carry out the annotations. All employees are adults and ethical. On average, they were paid £5–£10/hour.

## Acknowledgements

This work was funded by the National Natural Science Foundation of China (NSFC) under Grants No. 61902381 and 62006218, the Youth Innovation Promotion Association CAS under Grants No. 2021100 and 20144310, the Young Elite Scientist Sponsorship Program by CAST under Grants No. YES20200121, and the Lenovo-CAS Joint Lab Youth Scientist Project.

## References

- Omar Adjali, Romaric Besançon, Olivier Ferret, Hervé Le Borgne, and Brigitte Grau. 2020. Multimodal entity linking for tweets. In *European Conference on Information Retrieval*, pages 463–478. Springer.
- Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary G. Ives. 2007. Dbpedia: A nucleus for a web of open data. In *The Semantic Web, 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007 + ASWC 2007, Busan, Korea, November 11-15, 2007*, volume 4825 of *Lecture Notes in Computer Science*, pages 722–735. Springer.
- A. F. Biten, L. Gomez, M. Rusinol, and D. Karatzas. 2019. Good news, everyone! context driven entity-aware captioning for news images. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- N. D. Cao, G. Izacard, S. Riedel, and F. Petroni. 2020. Autoregressive entity retrieval.
- Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. 2018. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 67–74. IEEE.
- Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.
- Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z Wang. 2008. Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys (Csur)*, 40(2):1–60.
- Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni. 2020. Autoregressive entity retrieval. *arXiv preprint arXiv:2010.00904*.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Shahi Dost, Luciano Serafini, Marco Rospocher, Lamberto Ballan, and Alessandro Sperduti. 2020. Vt-linker: Visual-textual-knowledge entity linker. In *ECAI 2020*, pages 2897–2898. IOS Press.
- M Fabian, Kasneci Gjergji, WEIKUM Gerhard, et al. 2007. Yago: A core of semantic knowledge unifying wordnet and wikipedia. In *16th International world wide web conference, WWW*, pages 697–706.
- Jingru Gan, Jinchang Luo, Haiwei Wang, Shuhui Wang, Wei He, and Qingming Huang. 2021. Multimodal entity linking: a new dataset and a baseline. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 993–1001.
- Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. 2021. Clip-adapter: Better vision-language models with feature adapters. *arXiv preprint arXiv:2110.04544*.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.
- Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordino, Hagen Fürstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum. 2011. Robust disambiguation of named entities in text. In *Proceedings of the 2011 conference on empirical methods in natural language processing*, pages 782–792.

- Eric Jang, Shixiang Gu, and Ben Poole. 2016. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*.
- F. Liu, Y. Wang, T. Wang, and V. Ordonez. 2020. Visualnews : Benchmark and challenges in entity-aware image captioning.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
- Lajanugen Logeswaran, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, Jacob Devlin, and Honglak Lee. 2019. Zero-shot entity linking by reading entity descriptions. *arXiv preprint arXiv:1906.07348*.
- Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.
- Cédric Maigrot, Vincent Claveau, Ewa Kijak, and Ronan Sicre. 2016. Mediaeval 2016: A multimodal system for the verifying multimedia use task. In *MediaEval 2016: "Verifying Multimedia Use" task*.
- S. Moon, L. Neves, and V. Carvalho. 2018. Multimodal named entity disambiguation for noisy social media posts. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.
- Isaiah Onando Mulang', Kuldeep Singh, Chaitali Prabhu, Abhishek Nadgeri, Johannes Hoffart, and Jens Lehmann. 2020. Evaluating the impact of knowledge graph context on entity disambiguation models. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 2157–2160.
- Eric Müller-Budack, Jonas Theiner, Sebastian Diering, Maximilian Idahl, Sherzod Hakimov, and Ralph Ewerth. 2021. Multimodal news analytics using measures of cross-modal entity and context consistency. *International Journal of Multimedia Information Retrieval*, 10(2):111–125.
- Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.
- A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, and J. Clark. 2021. Learning transferable visual models from natural language supervision.
- A. Ramisa, Fei Yan, Francesc Moreno-Noguer, and K. Mikolajczyk. 2017. Breakingnews: Article annotation by image and text processing. *IEEE Transactions on Pattern Analysis Machine Intelligence*, PP(99):1–1.
- Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823.
- Özge Sevgili, Artem Shelmanov, Mikhail Y. Arkhipov, Alexander Panchenko, and Chris Biemann. 2022. Neural entity linking: A survey of models based on deep learning. *Semantic Web*, 13(3):527–570.
- Avirup Sil, Gourab Kundu, Radu Florian, and Wael Hamza. 2018. Neural cross-lingual entity linking. In *Thirty-Second AAAI Conference on Artificial Intelligence*.
- Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. 2017. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-first AAAI conference on artificial intelligence*.
- Amara Tariq and Hassan Foroosh. 2017. A context-driven extractive framework for generating realistic image descriptions. *IEEE Trans. Image Process.*, 26(2):619–632.
- Neha Tilak, Sunil Gandhi, and Tim Oates. 2017. Visual entity linking. In *2017 International Joint Conference on Neural Networks, IJCNN 2017, Anchorage, AK, USA, May 14-19, 2017*, pages 665–672. IEEE.
- Alasdair Tran, Alexander Mathews, and Lexing Xie. 2020. Transform and tell: Entity-aware news image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13035–13045.
- Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.
- X. Wang, J. Tian, M. Gui, Z. Li, R. Wang, M. Yan, L. Chen, and Y. Xiao. 2022. Wikidiverse: A multimodal entity linking dataset with diversified contextual topics and entity types. *arXiv e-prints*.
- Rebecka Weegar, Linus Hammarlund, Agnes Tegen, Magnus Oskarsson, Kalle Åström, and Pierre Nugues. 2014. Visual entity linking: A preliminary study. In *Workshops at the Twenty-Eighth AAAI Conference on Artificial Intelligence*.
- Ledell Wu, Fabio Petroni, Martin Josifoski, Sebastian Riedel, and Luke Zettlemoyer. 2019. Scalable zero-shot entity linking with dense entity retrieval. *arXiv preprint arXiv:1911.03814*.
- Ikuya Yamada, Koki Washio, Hiroyuki Shindo, and Yuji Matsumoto. 2019. Global entity disambiguation with pretrained contextualized embeddings of words and entities. *arXiv preprint arXiv:1909.00426*.
- P. Young, A. Lai, M. Hodosh, and J. Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Nlp.cs.illinois.edu*.

Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. 2016. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters*, 23(10):1499–1503.

L. Zhang, Z. Li, and Q. Yang. 2021a. *Attention-Based Multimodal Entity Linking with High-Quality Images*. Database Systems for Advanced Applications.

Renrui Zhang, Rongyao Fang, Peng Gao, Wei Zhang, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. 2021b. Tip-adapter: Training-free clip-adapter for better vision-language modeling. *arXiv preprint arXiv:2111.03930*.

Qiushuo Zheng, Hao Wen, Meng Wang, and Guilin Qi. 2022. Visual entity linking via multi-modal learning. *Data Intell.*, 4(1):1–19.

## A Baselines Details

**Parameters Setting.** In the architecture, we set the number of layers in the feed-forward as 2 and the dimensions are [512\*1024, 1024\*512] both for mention and entity in the two models. The initial learning rate is set to 2e-4 for ResNet and 2e-6 for CLIP. Images are all resized to  $224 \times 224$  pixels according to the common size and textual information is truncated to 77 words. The batch sizes for ResNet and clip are both set to 64. All the methods are implemented in Pytorch (Paszke et al., 2019) and optimized by the AdamW (Loshchilov and Hutter, 2017) algorithm.

**Experimental setup.** We train our models on two NVIDIA Tesla V100 GPU. We train each model with much to 20 epochs. For inference, we use Faiss<sup>6</sup> to achieve fast recall in large-scale embedding space with about 500ms per instance.

## B Evaluation Metrics

All evaluation and empirical analysis are reported by two widely-used metrics of Top-k retrieve: Recall and Mean Reciprocal Rank (MRR). The final result is the average score among all the cases.

$$\text{Recall@K} = \frac{1}{Q} \sum_{i=1}^Q \mathbf{1}_{qk_i}(gt_i)$$

$$\text{MRR@K} = \frac{1}{Q} \sum_{i=1}^Q \frac{1}{\text{rank}_i}$$

where  $\mathbf{1}_A(x)$  denotes a 0,1 valued indicator function.  $qk_i$ ,  $gt_i$  are the Top-k result and the ground truth of query i. MRR is a measure to evaluate systems that "Where is the first relevant item". For a single query, the reciprocal rank is  $\frac{1}{\text{rank}}$  where rank is the position of the highest-ranked answer. If no correct answer was returned in the query, then the reciprocal rank is 0.

## C More Examples from WIKIPerson

To demonstrate more details of our dataset, we pick two examples from our dataset. (Figure 8, Figure 9).

## D Detailed Analysis

According to the experimental results, the re-ranking strategy improves performance to a certain degree. So we conduct a detailed analysis of the



Figure 8: The images of Taylor Swift (Q26876, a famous American singer-songwriter) in WIKIPerson.



Figure 9: The images of Indra Nooyi (Q264913, Indian American business executive and former CEO of PepsiCo) in WIKIPerson.

strategy to help understand the reason and provide some insights for future model designs.

Firstly, we analyze the effect of re-ranking sequence length, which is the main factor affecting the result. Specifically, we conduct research on the re-ranking sequence length. Then we plot the Recall@1 for ResNet + CLIP\_N\_D and CLIP\_N\_D + ResNet in Figure 10. From the results, we can see that both two methods achieve high performance as the re-ranking length increases at the beginning. Then it starts to decrease slightly. It can be simply inferred that when the re-ranking length continues to grow to the size of the  $|E|$ , the re-ranking model can be equal to the single ResNet or CLIP\_N\_D. Besides, these two models have different Inflection Points and speeds of the downturn. CLIP\_N\_D

<sup>6</sup><https://github.com/facebookresearch/faiss>

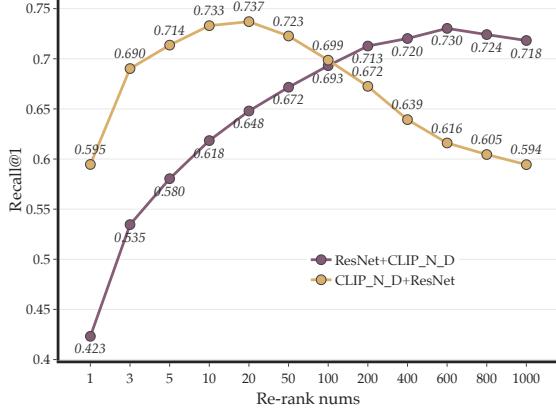


Figure 10: The Recall@1 of the models with different re-rank size.

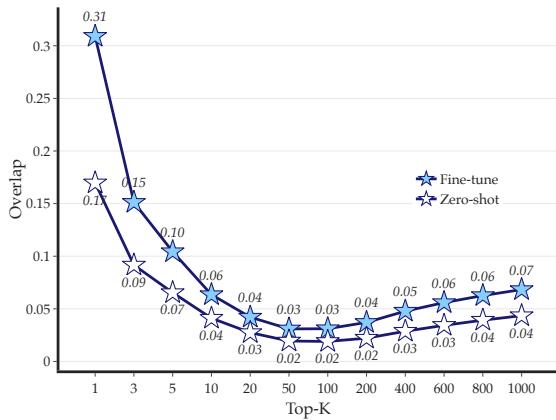


Figure 11: Overlap of Top-k result between CLIP\_N\_D and ResNet in zero-shot and fine-tune.

+ ResNet reaches its peak at lower re-rank length and decent sharply while ResNet + CLIP\_N\_D increases until re-rank length equals 600 and decent slowly. The reason for the phenomenon is that CLIP\_N\_D outperforms ResNet. As a result, a larger re-rank size is necessary for ResNet to guarantee to recall the ground truth.

Secondly, we notice that the Top-k results of CLIP\_N\_D and ResNet differ greatly. As a result, we plot the precise overlap between ResNet and CLIP\_N\_D's Top-k result in Figure 11.

The origin and fine-tune model have the same trend: with the increase of the K, the overlap decreases first and increases later. When k nears 50, the overlap minimum. For fine-tune model, it has a higher overlap than the zero-shot. The overlap starts from 30.1%, which means only the 30.1% of entities are identical among the Top-1 results between the two models even though they have comparable performance. Then it drops to 15% sharply. When k equals  $|E|$ , the overlap will reach

100%. Smaller coverage with high and comparable model performance ensures that using one model to re-ranking based on the recall of the other model could improve performance significantly.