Utilizing Language-Image Pretraining for Efficient and Robust Bilingual Word Alignment

Tuan Dinh, Jy-yong Sohn, Shashank Rajput, Timothy Ossowski, Yifei Ming, Junjie Hu, Dimitris Papailiopoulos, Kangwook Lee
University of Wisconsin, Madison, USA

Abstract

Word translation without parallel corpora has become feasible, rivaling the performance of supervised methods. Recent findings have shown that the accuracy and robustness of unsupervised word translation (UWT) can be improved by making use of visual observations, which are universal representations across languages. In this work, we investigate the potential of using not only visual observations but also pretrained language-image models for enabling a more efficient and robust UWT. Specifically, we develop a novel UWT method dubbed Word Alignment using Language-Image Pretraining (WALIP), which leverages visual observations via the shared embedding space of images and texts provided by CLIP models (Radford et al., 2021). WALIP has a two-step procedure. First, we retrieve word pairs with high confidences of similarity, computed using our proposed *image-based fingerprints*, which define the initial pivot for the word alignment. Second, we apply our *robust Procrustes algorithm* to estimate the linear mapping between two embedding spaces, which iteratively corrects and refines the estimated alignment. Our extensive experiments show that WALIP improves upon the state-of-the-art performance of bilingual word alignment for a few language pairs across different word embeddings and displays great robustness to the dissimilarity of language pairs or training corpora for two word embeddings.

1 Introduction

Translating words across different languages is one of the long-standing research tasks and a standard building block for general machine translation. Word translation is helpful for various downstream applications, such as sentence translation (Conneau et al., 2017; Hu et al., 2019) or cross-lingual transfer learning in language models (de Vries and Nissim, 2020). Unsupervised word translation (UWT) has recently drawn a great deal of attention (Smith et al., 2017; Artetxe et al., 2017; Conneau et al., 2017; Hoshen and Wolf, 2018; Hartmann et al., 2019), reducing the need for bilingual supervision.

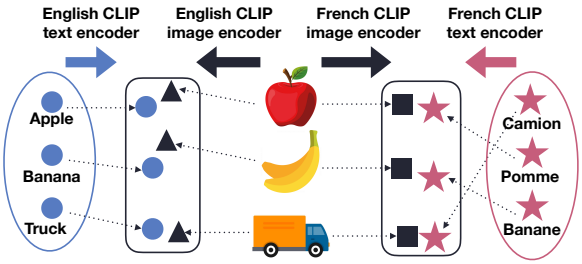


Figure 1: Conceptual visualization of WALIP for unsupervised word translation between English and French. We can connect English and French words in an unsupervised fashion through the shared images. CLIP models (Radford et al., 2021) can be used as human simulators to associate words with images.

Without any prior knowledge of the languages' connection, aligning their words is non-trivial. Most works on UWT rely on the structural similarity between continuous word embedding spaces across languages (Mikolov et al., 2013b; Smith et al., 2017; Ormazabal et al., 2019). They learn a linear mapping between spaces by establishing the parallel vocabulary based on the languages' similarity (Smith et al., 2017; Artetxe et al., 2017) and solving the Procrustes problem (Schönemann, 1966; Gower and Dijksterhuis, 2004) or directly solving Wasserstein-Procrustes to optimize the bilingual assignment matrix (Zhang et al., 2017b; Grave et al., 2019). Recent works focus on learning the mapping by aligning the embedding distributions using adversarial approaches (Zhang et al., 2017a; Conneau et al., 2017) or iterative matching methods (Hoshen and Wolf, 2018), establishing strong baselines for bilingual alignment. Despite achieving high performance, most of these text-only methods may need a large amount of data for good alignments (Sigurdsson et al., 2020) and rely

Email: Tuan Dinh (tuan.dinh@wisc.edu)

on the similarity between language pairs and training corpora, thus not working well for all language pairs or training corpora (Søgaard et al., 2018; Artetxe et al., 2018a; Sigurdsson et al., 2020).

Words can also be connected via the visual world. Visual similarity provides additional prior knowledge for easing word translation (Sigurdsson et al., 2020; Surís et al., 2020). MUVE (Sigurdsson et al., 2020) learns a linear mapping called *adaptlayer* by aligning videos and captions in joint embedding space, shared across languages. Globetrotter (Surís et al., 2020), via contrastive learning, learns the multilingual text embeddings aligned with image embeddings, which can be used for sentence or word translation. While these works demonstrate promise in translation across different types of word embeddings and training corpora, they require extensive joint training for the shared embedding of multiple languages. More importantly, these embeddings are used for translating all words, whereas not every word can be represented by or aligned with images or videos. It is unclear how these embeddings are helpful for non-visual words and whether they properly utilize the words' topological information (Mikolov et al., 2013b).

Our contibutions. In this work, we propose WALIP (**W**ord **A**lignment with **L**anguage-**I**mage **P**retraining) as a new unsupervised word alignment method that leverages the visual information and the word/image embeddings developed by CLIP (Radford et al., 2021), reducing the cost of aligning embedding distributions. Fig. 1 shows an example inspiring WALIP. Consider a conversation between a French and an English speaker. To translate `apple` to French, the English speaker can show an image of `apple`, and the French speaker can easily understand and provides its French translation as `pomme`. Similarly, they can pair more words that describe simple objects, which help them to translate more complex words/sentences. This observation inspires us to leverage visual information as the pivot for matching words in different languages. To do so, we utilize language-image pretraining models (Radford et al., 2021) to simulate the human ability of text-image correlation. Using a set of diverse images, we construct an image-based vector representation for a word where each coordinate measures the similarity between the word and one image from the set. We use this representation, called a *fingerprint*, to identify initial word pairs. As most images are restricted to non-abstract nouns, for the second step, we rely on the topological similarity of languages' static word embeddings (Bojanowski et al., 2016; Pennington et al., 2014; Mikolov et al., 2013c,a) for the full alignment via solving their linear mapping (Ormazabal et al., 2019). We introduce a robust Procrustes algorithm to correct the mismatched pairs.

Via extensive experiments, we show that WALIP is highly effective in bilingual alignment. We achieve comparable or better performance than the state-of-the-art (SoTA) baselines and close the gap to supervised methods. For instance, on the Dictionary benchmark (Sigurdsson et al., 2020) with HowToWorld-based word embedding (Miech et al., 2019), we achieve the SoTA performance on all evaluated pairs (English→{French, Korean, Japanese}), achieving significant improvement in accuracy (6.7%, 2.5%, and 4.5%, *cf.* Table 2) over the previous SoTA (Sigurdsson et al., 2020). Our method also shows excellent robustness to the dissimilarity of language pairs and word embeddings. We empirically prove the effectiveness of our method through various ablation studies.

2 Problem Setup

We formally describe the target problem of unsupervised word alignment and provide the preliminaries required for solving this problem.

2.1 Unsupervised Word Alignment

Suppose we have access to the dictionaries of two languages (say source language A and target language B), denoted by $A_{\text{dict}} = \{a_1, \cdots, a_{n_a}\}$ and $B_{\text{dict}} = \{b_1, \cdots, b_{n_b}\}$, where n_a and n_b are the number of words in the dictionaries. Our work focuses on the word alignment (translation) problem: finding the mapping from the source dictionary A_{dict} to the target dictionary B_{dict}. This mapping between dictionaries can be represented by an equivalent index mapping $\pi : [n_a] \to [n_b]$, i.e., we consider the word a_i in the source language is mapped (aligned) to the word $b_{\pi(i)}$ in the source language, for $i \in [n_a]$. Here, $[n] = \{1, 2, \cdots, n\}$ is defined as the set of positive integers up to n ($n > 0$). Note that we focus on *unsupervised* word alignment where no sample word pairs $(a_i, b_{\pi(i)})$ are given to the algorithm. Together with two dictionaries and pretrained word vectors (Bojanowski et al., 2016; Pennington et al., 2014), we assume the access to two ingredients: (1) a large-scale image dataset with d images, denoted

by $G = \{g_1, \cdots, g_d\}$ and (2) pre-trained monolingual CLIP model for each language.

2.2 Preliminaries

We review some conventional techniques suggested for aligning two sets of vectors (e.g., word embeddings) in an unsupervised manner.

Procrustes problem Assuming the source and target word embeddings can be aligned with a linear mapping (Søgaard et al., 2018; Ormazabal et al., 2019), most existing works consider the unsupervised word alignment as a Procrustes problem (Gower and Dijksterhuis, 2004). Formally speaking, let $X, Y \in \mathbb{R}^{n \times d}$ be the matrices containing the $d-$dimensional embeddings for n words in the source and target languages. The Procrustes algorithm aims to find $W \in \mathbb{R}^{d \times d}$ such that $\|XW - Y\|_F$ is minimized. The recent attempt (Xing et al., 2015) finds that the orthogonal mapping W improves the translation performance, with the optimal mapping defined as

$$W^* = \underset{W \in \mathcal{O}_d}{\texttt{argmin}} \|XW - Y\|_F = \texttt{SVD}(Y^T X)$$

where $\mathcal{O}_d$ is the set of $d \times d$ orthogonal matrices and SVD is the singular value decomposition.

CSLS The Cross-domain Similarity Local Scaling (CSLS) score (Conneau et al., 2017) was proposed to robustly measure the similarity between two words' embeddings. Given two sets of embeddings $X = \{x_i\}_{i \in [m]}$ and $Y = \{y_i\}_{i \in [n]}$ and a pre-defined integer K representing the number of neighbors to count, the similarity of x_i and y_j measured by CSLS is defined as

$$\text{CSLS}(x_i, y_j) = 2\cos(x_i, y_j) - r_Y(x_i) - r_X(y_j)$$

where $\cos(\cdot, \cdot)$ is the cosine similarity and $r_Y(x_i) = \frac{1}{K} \sum_{y_j \in \mathcal{N}_Y(x_i)} \cos(x_i, y_j)$ is the average cosine similarity of x_i and $\mathcal{N}_Y(x_i)$, the K nearest neighbors of x_i among the elements of Y. Intuitively, CSLS performs cross-domain normalization to address the hub phenomenon (Radovanovic et al., 2010) of K-NN in high-dimensional spaces, where some vectors are nearest to many vectors (in dense areas) while others are isolated.

3 WALIP

We now describe the WALIP method. We first provide the high-level idea of our algorithm and then specify each stage in our algorithm.

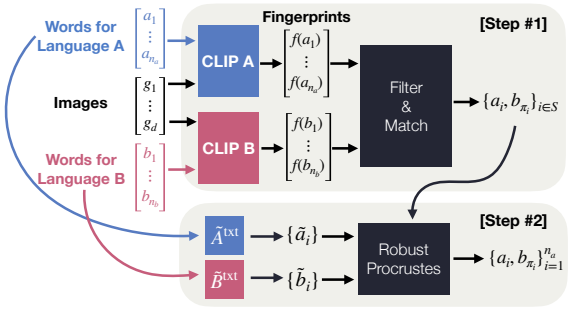


Figure 2: WALIP for translating n_a words $\{a_1, \cdots, a_{n_a}\}$ in language A and n_b words $\{b_1, \cdots, b_{n_b}\}$ in language B. Suppose we have access to three ingredients: (1) a set of d images $\{g_i\}_{i=1}^d$, (2) the CLIP model for each language, and (3) static word embeddings for each language, denoted by $\tilde{A}^{\text{txt}}$ and $\tilde{B}^{\text{txt}}$. In step 1, for each word a_i, we build a fingerprint $f(a_i)$ defined in equation 1, and build $f(b_i)$ for words b_i as well. Then we match the words whose fingerprints share a high similarity, thus having an initial mapping $\pi : [n_a] \to [n_b]$ that pairs a_i and b_{π_i} for $i \in S \subseteq [n_a]$. In step 2, we use non-contextual word vectors and initially matched pairs to solve the mapping between all words. Here, we apply the robust Procrustes algorithm for better alignment.

3.1 High-level Idea and the Pipeline

Our idea is to enable effective and robust word alignment by using (1) the similarity of words in different languages with respect to the visual representation and (2) the structure of word embedding spaces. Using shared images, we can connect words in two languages using language-image pre-training models (Radford et al., 2021). However, this method is mostly applicable to visual words, which can be described by a set of images, such as object nouns or non-abstract nouns. To further map non-visual words, we use the topological information of words. Motivated by existing studies (Mikolov et al., 2013b; Ormazabal et al., 2019) showing the existence of a linear association between static word vectors of two languages, we learn a linear mapping via robust Procrustes with initially identified word pairs.

Fig. 2 illustrates WALIP used for aligning words $\{a_i\}$ in language A and words $\{b_i\}$ in language B. WALIP consists of two steps. First, we select pairs $\{a_i, b_{\pi_i}\}$ having similar visual meanings; this can be done by using each word's fingerprint, defined as the similarity of the word and an image set measured by encoders of CLIP models. Secondly, we iteratively align two word embeddings using identified pairs by finding the linear mapping with our

robust matching algorithm. The second step utilizes the topological similarity (i.e., the *degree of isomorphism*) between two vector spaces (Vulić et al., 2020). Next, we describe two steps in detail.

3.2 Step 1: Pairing up Visually Similar Words using Language-Image Association

To pair up words via images, we construct a new representation of each word based on its association with a set of images. Using this representation, we measure the semantic similarity between words and select pairs that have high similarity scores.

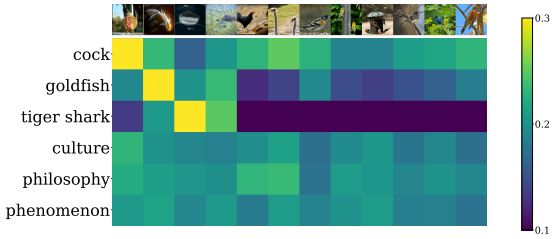
3.2.1 Image-based Fingerprints

We denote the image/text encoder of the CLIP model for language A as A^{img} and A^{txt}, respectively. Similarly, we define B^{img} and B^{txt} for language B. The critical advantage of the CLIP model is the access to the shared embedding space where an image g_i and the corresponding word (a_i or b_i) are aligned. The key idea of our method is to utilize this embedding space for each source/target language to find the bilingual word mapping. Specifically, given d images $\{g_1, \cdots, g_d\}$, we define a $d-$dimensional vector (called a **fingerprint**) for each word, representing how this word is aligned with each image on the embedding space provided by CLIP. We couple words a_i and b_i if their fingerprints are highly similar.

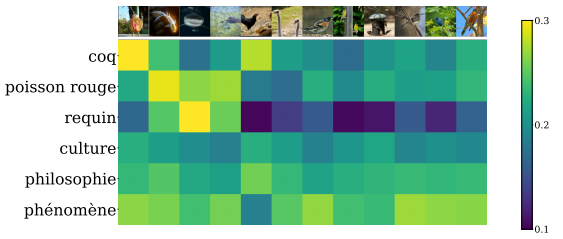
We first define the fingerprint of each word $a_i \in A_{\text{dict}}$ in the source language as a d-dimensional vector $f(a_i) = [f^a_{i,1}, \cdots, f^a_{i,d}]$ where $f^a_{i,j} = \text{sim}(A^{\text{txt}}(a_i), A^{\text{img}}(g_j))$ is the similarity between the embedding of the i-th word and the embedding of the j-th image. Similarly, we define the fingerprint of each word $b_i \in B_{\text{dict}}$ in the target language as $f(b_i) = [f^b_{i,1}, \cdots, f^b_{i,d}]$ where $f^b_{i,j} = \text{sim}(B^{\text{txt}}(b_i), B^{\text{img}}(g_j))$. This fingerprint represents how a word is similar to each image, according to the embedding space of pre-trained CLIP models. We denote the fingerprint of the i-th word in the dictionary of language $l \in \{a, b\}$ as

$$f(l_i) = [f^l_{i,1}, \cdots, f^l_{i,d}]. \quad (1)$$

Fig. 3a and Fig. 3b show examples of English and French fingerprints. For this example, we measure the similarity of each word with 12 images from ImageNet (Deng et al., 2009), resulting in a 12-dim vector. Values of each coordinate (per row) indicate how likely the word matches the image.



(a) Fingerprints for English words



(b) Fingerprints for French words

Figure 3: Illustration of image-based fingerprints defined in 1 for English words (a) and their corresponding translations in French (b). The similarity between each word (inserted in a simple template of CLIP such as "`This is a photo of []`") and all images serves as the fingerprint (each row). The top three rows are of visual words, and the three bottom rows are of abstract words. For visual words, their fingerprints are more distinguishable, and they share similar patterns to the fingerprints of translated words.

The top three rows of each figure are fingerprints for visual words (`cock, goldfish, tiger shark`), and the three bottom rows are the ones for abstract words (`culture, philosophy, phenomenon`). Compared to the visual words, abstract words have more uniform fingerprints (similar values for most coordinates). That is, the fingerprints of abstract words are *not* distinguishable. Also, we can see that all two words in each English-French pair of the first three rows {(`cock, coq`), (`goldfish, poisson rouge`), (`tiger shark, requin`)} share highly similar fingerprints.

3.2.2 Identifying Pivot Pairs

Based on the observation in the previous section, we identify the pairs with similar meanings based on the fingerprints. Consider two visual words a_i, b_j in different languages with similar meanings (e.g., $a_i =$ "`tiger shark`" and $b_j =$ "`requin`" in the example of Fig. 3). For a given set of images, the fingerprints of the two words would be similar, i.e., $f(a_i) \approx f(b_j)$, as shown in Fig. 3. This observation allows us to use the similarity of fingerprints for word translation.

Algorithm 1 `Visual-Word-Filtering`

Input: Fingerprints $\mathcal{F} = \{f(l_i)\}_{l\in\{a,b\}, i\in[n_l]}$
Output: Updated fingerprints $\mathcal{F}$
for $l \in \{a, b\}$ **do**
  $f^{(l)}_{i,j} \leftarrow j$-th element of $f(l_i)$, for $j \in [d]$
  $f^{(l)}_{i,\max} \leftarrow \max_j f^{(l)}_{i,j}$ for $i \in [n_l]$
  $S_l \leftarrow \{i : f^{(l)}_{i,\max} \geq \text{median}_i(f^{(l)}_{i,\max})\}$
  for $i \in S_l$ **do**
    $\bar{q} \leftarrow$ `0.9-th quantile of` $\{f^{(l)}_{i,k}\}_{k=1}^{d}$
    $f^{(l)}_{i,j} \leftarrow f^{(l)}_{i,j} \cdot \mathbf{1}_{\{f^{(l)}_{i,j} \geq \bar{q}\}}$
    $f^{(l)}_{i,j} \leftarrow f^{(l)}_{i,j} / |f^{(l)}_{i,j}|$

Keeping only visually aligned words We first filter only words that are highly correlated to and represented by a set of images. For the i-th word l_i in language $l \in \{a, b\}$, we compute the maximum similarity value $f^{(l)}_{i,\max} = \max_j f^{(l)}_{i,j}$ within the corresponding fingerprint $f(l_i) = [f^{(l)}_{i,1}, \cdots, f^{(l)}_{i,d}]$. Then, for each language $l \in \{a, b\}$, we only focus on the words S_l having the maximum similarity beyond the median. We sparsify fingerprints by keeping only values greater than the 0.9^{th}-quantile and normalize the fingerprint. This revised fingerprint allows us to focus on images highly similar to the given word. See Algo. 1 for the pseudocode.

Selecting pairs with high similarity scores For source words $\{a_i\}_{i\in S_a}$ and target words $\{b_j\}_{j\in S_b}$, we measure the similarity of fingerprints $f(a_i)$ and $f(b_j)$, using CSLS (Sec. 2.2). Recall that our final goal is to find a mapping $\pi : [n_a] \to [n_b]$ indicating that the word a_i is translated to $b_{\pi(i)}$, and we want to map a_i to b_j having similar fingerprints. Based on the similarity score $c_{i,j} = \text{CSLS}(f(a_i), f(b_j))$ for $i \in S_a$ and $j \in S_b$, we set $\pi(i) = \arg\max_j c_{i,j}$. This gives us an initial set of word pairs, where two words in each pair are more likely visual words and share highly similar fingerprint patterns. See Algo. 2 for the pseudocode.

3.3 Step 2: Iterative Robust Procrustes for Learning Linear Mapping

Given the initial pairs in Step 1, we now learn the linear mapping between two embedding spaces. We note that the linear mapping observation (Ormazabal et al., 2019) has been only shown for static word embeddings (Bojanowski et al., 2016; Pennington et al., 2014; Mikolov et al., 2013c). Hence,

Algorithm 2 `Matching-Filtering`

Input: Fingerprints $\mathcal{F} = \{f(l_i)\}_{l\in\{a,b\}, i\in[n_l]}$, Threshold quantile q
Output: Word index mapping $\pi : [n_a] \to [n_b]$
$c_{i,j} \leftarrow \text{CSLS}(f(a_i), f(b_j))$ for $i \in [n_a], j \in [n_b]$
$\bar{c} \leftarrow$ `q-th quantile of` $\{c_{i,j}\}$
$\pi \leftarrow$ empty mapping from $[n_a]$ to $[n_b]$
for $i \in [n_a]$ **do**
  $j^\star \leftarrow \arg\max_j c_{i,j}$
  if $c_{i,j^\star} \geq \bar{c}$ **then**
    $\pi(i) \leftarrow j^\star$

Algorithm 3 `Robust-Procrustes`

Input: Vectors $X, Y \in \mathbb{R}^{n\times d}$
Output: Linear mapping $W^* \in \mathbb{R}^{d\times d}$
Set $\epsilon = 0.001$, $M = 5$
Initial mapping $W_0 =$ `Procrustes`(X, Y)
for $m \in \{1, \cdots, M\}$ **do**
  $\alpha_i \leftarrow \frac{1}{\|y_i - W_{k-1}x_i\|^2+\epsilon}$ for $i \in [n]$
  $\alpha_i \leftarrow \alpha_i / \max_j \alpha_j$
  $D \leftarrow$ `Diag`$(\alpha_1^{1/2}, \ldots, \alpha_n^{1/2})$
  $W_m \leftarrow$ `Procrustes`(DX, DY)
$W^* \leftarrow W_M$

we use non-contextualized static word embeddings for learning the mapping. As the mapping is estimated in an unsupervised manner, there are likely many mismatched pairs. It is essential to eliminate these mismatched pairs and learn the ground-truth mapping from the correct pairs. To do so, we introduce our robust matching algorithm.

3.3.1 Correcting Mismatches with Error-Weighting Robust Procrustes

Our idea is to dynamically assign weights to each pair: small weights for incorrect pairs and high weights for correct ones. This idea has the same theme as robust Procrustes algorithms (Groenen et al., 2005) where they focus on denoising. How do we determine the correctness? As the linear mapping W aims to map a word vector x to a word vector y, we can measure the error of the pair by the residual $r(x, y) = \|y - Wx\|_2$. Since the pair is likely to be correct when the residual is small, we assign $\alpha(x, y) = 1/r(x,y)$ as the weight of the pair. Using these weighted pairs, we can apply Procrustes to solve the linear mapping W. We repeat this process a few times to obtain a stable W. Algo. 3 presents our algorithm.

3.3.2 Iterative Learning of the Mapping

Given a set of initial pairs, we apply robust Procrustes on the corresponding static word vectors to estimate the linear map W. Then we translate the whole set of source word vectors to align with the target vectors. For each word in the target, we keep only source words corresponding to the highest scores to generate a set of candidate pairs. We measure the CSLS score of each pair and select pairs beyond the similarity threshold c as the initial pairs for the next iteration. Also, we measure the loss of each translation as the average Euclidean distance between paired vectors of target and translated words. We repeat these two steps until the convergence (loss does not change).

We note that while these steps are similar to the refinement procedure in previous works (Conneau et al., 2017), our iterative learning uses the novel robust Procrustes algorithm and an adaptive scheme for hyper-parameters. In particular, as the number of iterations increases, we gradually reduce the learning threshold (to increase the number of high-score pairs) and reduce the number of candidates (to reduce the number of low-score pairs) because W gradually becomes better and closer to the true W^*. We observe that the initial pairs are crucial for the success of our iterative mapping, as shown in the ablation study (Fig. 5, Sec. 4.4). Algo. 4 presents our proposed WALIP algorithm.

Algorithm 4 `WALIP`

Input: Source dictionary $A_{\text{dict}} = \{a_1, \cdots, a_{n_a}\}$, target dictionary $B_{\text{dict}} = \{b_1, \cdots, b_{n_b}\}$, CLIP models $(A^{\text{txt}}, A^{\text{img}})$, $(B^{\text{txt}}, B^{\text{img}})$, set of images $G = \{g_1, \cdots, g_d\}$, word vectors T_A for A_{dict}, T_B for B_{dict}, number of alignment steps K, threshold quantile q.

Output : $\pi : [n_a] \to [n_b]$ such that $a_{\pi(i)} \equiv b_i$

for language $l \in \{a, b\}$ **do**
  $f(l_i) \leftarrow$ fingerprint in (1) for $i \in [n_l]$
$\mathcal{F} \leftarrow \{f(l_i)\}_{l \in \{a,b\}, i \in [n_l]}$
$\mathcal{F} \leftarrow$ `Visual-Word-Filtering`$(\mathcal{F})$
$\pi_0 \leftarrow$ `Matching-Filtering`$(\mathcal{F}, q)$
for $k \in \{1, \cdots, K\}$ **do**
  $s^A_{k-1} \leftarrow \{i \in [n_a] : \pi_{k-1}(a_i) \in B_{\text{dict}}\}$
  $s^B_{k-1} \leftarrow \{j \in [n_b] : \exists a_i \text{ s.t. } \pi_{k-1}(a_i) = b_j\}$
  $T'_A \leftarrow T_A[s^A_{k-1}], \quad T'_B \leftarrow T_B[s^B_{k-1}]$
  $W^* \leftarrow$ `Robust-Procrustes`(T'_A, T'_B)
  $T_A \leftarrow T_A W^*$
  $\epsilon = \|T_A - T_B\| / (\|T_A\| + \|T_B\|)$
  $q \leftarrow \min\{0.9, \max\{0.1, \epsilon\}\}$
  $\pi_k \leftarrow$ `Matching-Filtering`$(\{T_A, T_B\}, q)$
$\pi \leftarrow$ `Matching-Filtering`$(\{T_A, T_B\}, 0)$

3.4 Advantages of WALIP

Computing efficiency WALIP is computationally efficient, especially in comparison with MUVE and GLOBETROTTER that require intensive training for aligning visual and text embeddings. With pretrained CLIPs, our first step requires no extra training for pivot pair matching, while the second step involves only a few matrix computations.

Robustness to language dissimilarity Using the same set of images and well-trained CLIPs for each language, the fingerprints of the meaning-similar words are intuitively similar across the languages because they all represent the same visual correlation to the same image set. Therefore, using fingerprints indeed improves the robustness of pivot matching, especially in the pairs whose two languages come from different families. This may not be the case for methods using static word embeddings (Søgaard et al., 2018).

Furthermore, the *modular two-stage design* of WALIP eases the integration of similarly functioning models, such as different pretrained language-image models or algorithms for solving linear mapping. Also, our image-based fingerprint provides an *interpretable representation* of words.

4 Experiments

We first evaluate the bilingual alignment across language pairs (Sec. 4.2), highlighting the performance of dissimilar language pairs. Sec. 4.3 shows the robustness when two languages use two static word embeddings trained on different corpora. Sec. 4.4 provides various ablation studies: zero-shot cross-transfer with English-trained CLIPs (Fig. 4), different schemes for initial matching (Fig. 5), robust Procrustes (Fig. 6), and size-varying image sets (Fig. 7).

4.1 Settings

Static Word Embeddings We use two embeddings: HowToWorld (HTW)-based Word2Vec (Miech et al., 2019; Sigurdsson et al., 2020) that trains Word2Vec (Mikolov et al., 2013c) on HTW video datasets and Wiki-based Fasttext (Bojanowski et al., 2016) that trains Fasttext on the Wikipedia corpus.

Evaluation Benchmark We use the *Dictionary* datasets (Sigurdsson et al., 2020) which are test sets of MUSE bilingual dictionaries (Conneau et al., 2017). Each dictionary provides a set of matched pairs in two languages where each word in the source language can have multiple translations in the target language.

Evaluation Metrics Our metric is *recall@n* used in (Sigurdsson et al., 2020) for $n = 1, 10$: given n retrieved words for a given query, the retrieval is correct if at least one of n words is the correct translation of the query. *Recall@n* presents the fraction of source words that are correctly translated. By default, we report *recall@1*, equivalent to *precision@1*, and the matching accuracy used in (Conneau et al., 2017).

Baselines We describe what baselines we have compared in this paper. **CLIP-NN** is a simple baseline that performs double 1-nearest neighbor (1-NN) on CLIP embeddings: Given a source word, we perform the 1-NN to find the nearest image (using source CLIP) and then perform the 1-NN on the target CLIP to find the nearest target word. For *recall@n*, we perform the similar double k-NN where $k = \lceil \sqrt{n} \rceil$. **MUSE** (Conneau et al., 2017) is a text-only method that learns the cross-lingual linear mapping via adversarially aligning embeddings' distributions and iterative refinement with Procrustes. **MUVE** (Sigurdsson et al., 2020) replaces the linear layer learned in the first stage of MUSE with the *AdaptLayer* learned by jointly training the embeddings of videos and captions, shared across languages. **Globetrotter** (Surís et al., 2020) uses image-caption pairs to align the text embeddings of multiple languages to image embeddings using a contrastive objective. **Variants of WALIP.** We provide variants of WALIP for the ablation by replacing each module in our method with similarly functioning methods. For step 1, we replace fingerprints with clip-based text embeddings (*clip-text*) or replace the entire step with simple methods, such as *substring matching* that chooses pairs of words sharing long common substrings or *character mapping* that uses frequency analysis (or letter counting) (Ycart, 2012) to map two languages' character sets before applying substring matching. For step 2, we replace the static word embeddings with clip-text or fingerprints.

Implementation Details Here, we provide the details of implementing our algorithms. **CLIP models:** We use publicly available pretrained CLIPs for English[1], Russian[2], Korean[3], and Japanese.[4] For other languages, we finetune English CLIP models on Multi30K (Elliott et al., 2016, 2017) and MS-COCO datasets (Lin et al., 2014; Scaiella et al., 2019; C., 2020) with translated captions for each target language. Specifically, we finetune each model for 20 epochs using the `NCEInfo` loss (Oord et al., 2018) without changing the architectures of the original CLIP's encoders. We use Adam optimizer (Kingma and Ba, 2014) ($\beta_1, \beta_2 = 0.9, 0.98$) with a learning rate of `1e-7` and cosine annealing scheduler (Loshchilov and Hutter, 2016). **Image datasets.** We use 3000 images from ImageNet (Deng et al., 2009). We find that high-resolution images provide the best initial mappings among tested image data. **Prompts for words in CLIPs:** As for the input of CLIPs, we convert every single word to a full sentence. We use the prompt templates suggested in (Radford et al., 2021), and we apply prompt-ensemble (Radford et al., 2021) for the best embedding. In particular, we use a set of (two to seven) prompts for each word and average these text embeddings as the word embedding. **Hyper-parameters:** We use 20–40 iterations of robust Procrustes in step 2. We observe that the losses on pairs of similar languages (e.g., English-French) converge after a few iterations. In Algo. 4, together with the adaptive scheme for the quantile threshold, we also test the simple scheme with alternatively changing q from a set of discrete values $\{0.1, 0.3, 0.5, 0.7\}$ that alternatively change. To generate the matching candidates for each word in Algo. 2, we initially select all words in the top $k = \{10, 5, 3, 1\}$ highest scores and gradually decreases k as the loss converges.

4.2 Bilingual Word Alignment Accuracy

Table 1 and Table 2 show our evaluation of bilingual alignment on different language pairs, using Wiki-based and HTW-based embeddings, respectively. The benchmark is the Dictionary dataset.

Wiki-based embeddings Table 1 shows the evaluation using Fasttext embeddings. We note that we provide two versions of MUSE: one trained on the given dictionaries and one trained with extra vocab-

[1] https://github.com/openai/CLIP
[2] https://github.com/sberbank-ai/ru-clip
[3] https://github.com/jaketae/koclip
[4] https://huggingface.co/rinna/japanese-clip-vit-b-16

Table 1: Comparing bilingual alignment methods on Wiki-based word embedding. We report Recall@1 on the Dictionary dataset. WALIP achieves SoTA performance in many language pairs, close to the supervised method. *MUVE (Sigurdsson et al., 2020) does not report results in this setting.*

	Method	En→Ko	En→Ru	En→Fr	En→It	En→Es	En→De	Es→De	It→Fr
Text-only	(Upper bound) Supervision	69.1	85.5	93.5	92.1	93.3	92.5	91.5	95.1
	MUSE (extra vocabularies)	59.3	**83.0**	**92.5**	**91.6**	**93.0**	**92.5**	**89.1**	**94.5**
	MUSE	2.8	65.9	84.5	84.9	85.1	73.6	83.0	92.3
	WALIP (substr. match. - st.#1)	0.2	0.0	92.0	90.3	92.0	92.1	88.7	94.3
	WALIP (char. map. - st.#1)	0.2	5.0	90.9	0.1	0.1	0.3	0.5	0.5
Text-Image	CLIP-NN	2.5	9.4	1.3	10.5	8.2	7.1	7.3	6.5
	GLOBETROTTER	0.1	4.0	52.3	50.1	46.4	46.8	38.3	49.3
	WALIP (clip-text - st.#1)	0.3	0.0	58.9	79.4	56.2	50.8	46.5	52.5
	WALIP (clip-text - st.#2)	0.2	15.7	59.3	59.1	59.1	52.3	46.8	52.1
	WALIP (fingerprint - st.#2)	0.2	0.5	31.3	39.0	32.6	31.3	34.7	43.3
	WALIP	**62.3**	82.7	**92.6**	90.7	92.2	**92.6**	**89.2**	**94.5**

Table 2: Comparing bilingual alignment methods on HTW-based embedding. We report *Recall@n* on Dictionary dataset (Sigurdsson et al., 2020). WALIP outperforms baselines across all language pairs.

Method	En→Fr		En→Ko		En→Ja	
	R@1	R@10	R@1	R@10	R@1	R@10
(Up.) Sup.	57.9	80.1	41.8	72.1	41.1	68.3
MUSE (extra.)	26.3	42.3	11.8	23.9	11.6	23.5
MUVE	28.9	45.7	17.7	33.4	15.1	31.2
WALIP (sub.)	35.5	56.0	0.0	0.2	0.3	2.1
WALIP	**35.6**	**56.2**	**20.2**	**42.4**	**19.6**	**41.0**

Table 3: Comparing algorithms with the dissimilarity of word embeddings for source and target languages (HTW and Wiki). We report *Recall@1* on En→Fr translation evaluated on Dictionary dataset. WALIP outperforms other baselines across two settings.

Method	Wiki-HTW	HTW-Wiki
MUSE	0.3	0.3
VecMap	0.1	0.1
MUVE	32.6	41.2
WALIP	**34.3**	**60.0**

ularies (full available vocabularies). Among unsupervised methods, WALIP achieves comparable or even the best performances in most cases. In particular, we achieve SoTA on five pairs, especially for En→ Ko, where we significantly outperform the main baselines with large margins as well as MUSE with extra vocabularies. Moreover, we consistently outperform GLOBETROTTER – the closest method utilizing image information in translation. We note that MUVE only reports recall@10 for En→Fr as 82.4, which is far below our recall (97.5). It's worth mentioning that most baselines (except CLIP-NN) require intensive training for aligning two embedding spaces before applying Procrustes. Also, the performance gaps between our method and the supervised method are relatively small.

HTW-based embeddings Following (Sigurdsson et al., 2020), we apply algorithms on three available language pairs (En→Fr, En→Ko, and En→Ja) and report *recall@n*, shown in Table 2. We compare WALIP with best-performing baselines in the previous section: MUSE, MUVE, and the WALIP variant with substring matching. Here, we report results of MUSE and MUVE from (Sigurdsson et al., 2020). WALIP consistently outperforms other unsupervised baselines with large margins, achieving the SoTA performances for all pairs, with the *recall@1* gaps to the second-best method (MUVE) being 6.7, 2.8, and 4.5 for En→Fr, En→Ko, and En→Ja. We draw a similar conclusion in terms of *recall@10*. We also see that substring matching does not work well on En→{Ko, Ja}.

Robustness on dissimilar language pairs For both types of embeddings, we see that WALIP works comparably well regardless of whether or not two languages in the pair have the same language family. Specifically, most baselines do not work well on En→Ko and En→Ja (except the MUSE with extra vocabularies). Moreover, while our variant with substring matching method works well on English-family pairs when combined with our robust Procrustes, they do not perform on En→{Ko, Ja, Ru} as these pairs come from different families (with different alphabets). This also shows the importance of image-based fingerprints.

4.3 Robustness to the Dissimilarity of Static Word Embeddings

Following (Sigurdsson et al., 2020), we evaluate WALIP in the scenario where the two static word embeddings used in Step 2 come from different training corpora: one trained on the Wiki corpus and one trained on the HowToWorld corpus. We add an additional baseline *VecMap* (Artetxe et al., 2017) that shows more robustness than MUSE against the training initialization. Table 3 presents

our comparison with MUSE, VecMap, and MUVE on the En→Fr. [5] We can see that WALIP and MUVE are more robust against the dissimilarity compared to MUSE and VecMap, shown in significant performance gaps. Furthermore, WALIP outperforms MUVE on both settings. For instance, when the English embedding uses the HTW corpus and the French embedding uses the Wiki corpus, we achieve 60% in terms of *recall@1*, showing a gap of nearly 20% compared to MUVE.

4.4 Ablation Study

We perform ablation studies using the Fasttext embedding and the Dictionary dataset.

Can we reuse the English-trained CLIP model for different languages? Large-scale language models exhibit the strong ability of cross-lingual zero-shot transfer (Hu et al., 2020). This experiment investigates whether a CLIP model pretrained on English can be used for other languages in our method. Intuitively, this zero-shot transfer is probably doable only when new languages use the same alphabet (and the same tokenizers). To verify, we directly use the English-trained CLIP model to obtain the fingerprints for all languages (zero-shot cross-transfer). Fig. 4 compares WALIP models using finetuned CLIPs and WALIP models using English CLIPs. First, we see that using English CLIPs leads to the drops in initial matching accuracies, which measure the precision of mapping on selected pairs. However, these drops only affect the languages dissimilar to English and do not significantly affect the final translation performance on languages similar (or close) to English. For instance, we observe similar performances between the two settings in En→{Fr, It, De, Es} while recalls are nearly 0 for En→{Ko, Ru} when only English CLIPs are used. Hence, we show that English CLIPs can be transferred to use for English-related languages in our WALIP framework.

Effectiveness of image-based fingerprints We investigate the effect of fingerprints on the final translation performance. Fig. 5 compares variants of WALIP using different initial mapping methods, namely random matching (red), clip-text embeddings (olive), substring matching (green), and image-based fingerprints (ours, dark blue). We also report the scores in Table 1. WALIP models with fingerprints are shown to perform the best among

[5] MUVE provides only results of the English-French pair.

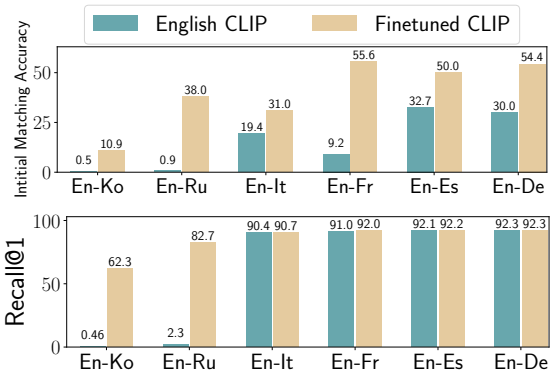


Figure 4: Zero-shot cross-lingual transfer with English CLIP for translation. When replacing finetuned CLIPs (yellow) with English CLIPs (dark green), all pairs' initial matching accuracies drop. However, (bottom) the final recall scores are only nearly 0 for dissimilar pairs (En→{Ko, Ru}) while remaining mostly the same for English-related languages (It, Fr, Es, De).

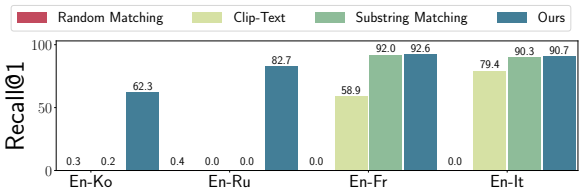


Figure 5: WALIP with different variants of step 1. Compared to image-based fingerprints (dark blue), other methods or embeddings for the initial mapping results in lower recall scores, especially for dissimilar language pairs (En→Ko and En→Ru). These results indicate the importance of initial mapping in our method.

variants across all pairs, especially when two languages are highly dissimilar (En→{Ko, Ru}).

Effectiveness of robust Procrustes Fig. 6 shows the comparison between robust Procrustes and standard Procrustes algorithms, given the same initial mapping. Robust Procrustes indeed helps improve over the normal Procrustes, especially when two languages are dissimilar. For instance, on En→Ko, using robust Procrustes increases the final *recall@1* of translation by nearly 23%.

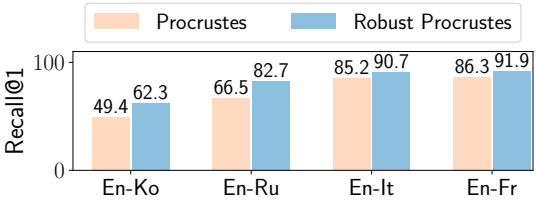


Figure 6: Investigating the effect of robust Procrustes. Robust Procrustes helps improve the translation across different language pairs. The effect is more significant on "difficult" pairs, such as English-Russian.

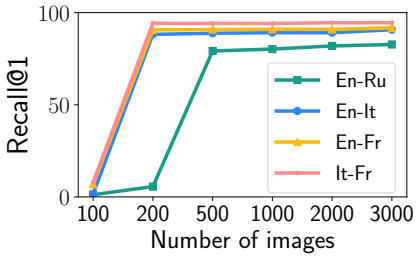


Figure 7: Varying the size of image datasets used for fingerprint construction. The performances improve as the number of images increases from 100 to 500 and remain mostly unchanged. Hence, WALIP may need only a small number of images.

The effect of image datasets We vary the size of image sets (ImageNet) in building fingerprints. Fig. 7 shows our evaluations of different language pairs. As the number of images increases, the translation performance increases or converges, observed over all pairs. WALIP needs up to 1000 images to achieve good performance across all evaluated pairs. As the languages become more dissimilar, we may need more images to distinguish them.

5 Related Works

Unsupervised Word Translation (UWT) Word translation has been an active topic (Mikolov et al., 2013b; Xing et al., 2015; Artetxe et al., 2017), and translating words without supervision has drawn much attention in the last decade (Conneau et al., 2017; Zhang et al., 2017a,b; Artetxe et al., 2017, 2018b). Most existing works on UWT make use of the observation that languages exhibit a similar structure in their continuous embedding spaces (Mikolov et al., 2013b; Ormazabal et al., 2019). They exploit this finding by learning a linear mapping between the source and the target word embedding space. In doing so, early works (Smith et al., 2017; Artetxe et al., 2017) establish the parallel vocabulary based on the languages' similarity and estimate the mapping by solving the Procrustes problem (Schönemann, 1966; Gower and Dijksterhuis, 2004). Other works study the problem under the assignment problem and directly solve Wasserstein-Procrustes to optimize the one-to-one assignment matrix (Zhang et al., 2017b; Grave et al., 2019) or hyperalignment for multiple languages (Alaux et al., 2018; Taitelbaum et al., 2019). Recent works focus on learning the mapping between two word-embedding spaces via aligning their distributions, using adversarial approaches (Zhang et al., 2017a; Conneau et al., 2017) or iterative matching methods (Hoshen and Wolf, 2018). Using this technique, MUSE (Conneau et al., 2017) establishes a strong baseline for bilingual alignment. However, Søgaard et al. (2018) finds that these approaches are sensitive to initialization and do not work for some types of highly dissimilar and low-resource languages. They also involve intensive training or optimizing pairwise languages. WALIP, in contrast, aligns the two embedding spaces without the need for intensive training (with the pretrained CLIP models), significantly saving the computing.

Visual information has been used for improving machine translation (Hewitt et al., 2018; Zhou et al., 2018; Kiros et al., 2018; Yang et al., 2020). Focusing on word translation, MUVE (Sigurdsson et al., 2020) uses captioned instructional videos to learn a joint video-text embedding space for pairs of languages using shared parameters. MUVE trains a linear layer called adaptlayer that maps source word vectors into the target embedding space for translation. This method trains encoders on videos, requiring more computation than WALIP, which works on images instead of video and can work with off-the-shelf image captioning models. Globetrotter (Surís et al., 2020) learns the multilingual text embeddings aligned with image embeddings via contrastive learning. The learned text embeddings are used for multilingual sentence translation and refined for word translation. While both works show great robustness to the choice of word embeddings and training corpora, they require extensive training for learning the embedding. Also, though not every word can be represented by and aligned with images or videos, both works incorporate visual information to train new joint embedding for which all words use for translation. Hence, it is unclear how the embeddings are helpful for non-visual words and if they properly utilize the words' topological information. We, instead, combine both the visual and topological information.

Vision-Language Models Utilizing large-scale pre-trained vision-language models for multimodal downstream tasks has become an emerging paradigm with remarkable performance (Uppal et al., 2022). In general, two types of architectures exist single-stream models like VisualBERT (Li et al., 2019) and ViLT (Kim et al., 2021) that concatenate text and visual features and feed into a single transformer-based encoder; dual-stream models

such as CLIP (Radford et al., 2021), ALIGN (Jia et al., 2021), and FILIP (Yao et al., 2021) that use separate encoders for text and image and optimize with contrastive objectives to align semantically similar features in a different modality. In particular, CLIP enjoys popularity due to its simplicity, inference efficiency, and strong performance due to pretraining on a huge dataset collected from the internet. CLIP-like models inspire numerous follow-up works (Zhang et al., 2021; Li et al., 2022; Zhou et al., 2022), aiming to improve data efficiency and better adaptation to downstream tasks. Moreover, multiple pretrained CLIP models are publicly available for different languages, making it an ideal choice in this work.

Symbol Grounding Symbol grounding (Harnad, 1990) is the problem of associating symbols (words of a language) to a common, meaningful entity. (Alaux et al., 2018) align word embeddings of multiple languages to a common space by jointly computing the permutation and mapping matrices. Visual grounding techniques for translation ground words into visual representations (images or videos), which are common across languages. Our proposed algorithm falls in this category, and the closest existing work is the previously discussed method – MUVE (Sigurdsson et al., 2020).

6 Conclusion

We propose WALIP, a novel unsupervised bilingual word alignment method using pretrained CLIP models. WALIP first leverages the visual similarity between words as the auxiliary for matching initial and simple word pairs, via the image-based fingerprint representation computed by language-image pretraining models. Then WALIP uses these initial pairs as pivots to learn the linear transformation between two static word embeddings. We introduce a robust Procrustes algorithm based on error-weighting to robustly estimate the linear mapping. WALIP is computationally efficient as we reduce the need for training to aligning two embeddings' distributions thanks to the aid of visual information and pretrained CLIP models. WALIP achieves the SoTA alignment performances on several language pairs across word embedding types, especially for pairs in which two languages are highly dissimilar. WALIP displays the robustness to the dissimilarity of static word embeddings' training corpora.

References

Jean Alaux, Edouard Grave, Marco Cuturi, and Armand Joulin. 2018. Unsupervised hyper-alignment for multilingual word embeddings. In *International Conference on Learning Representations*.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2017. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 451–462.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018a. Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018b. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 789–798, Melbourne, Australia. Association for Computational Linguistics.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

García C. 2020. MS-COCO-ES.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

Wietse de Vries and Malvina Nissim. 2020. As good as new. how to successfully recycle english gpt-2 to make models for other languages. *arXiv preprint arXiv:2012.05628*.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

Desmond Elliott, Stella Frank, Loïc Barrault, Fethi Bougares, and Lucia Specia. 2017. Findings of the second shared task on multimodal machine translation and multilingual image description. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 215–233, Copenhagen, Denmark. Association for Computational Linguistics.

Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. 2016. Multi30k: Multilingual english-german image descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, pages 70–74. Association for Computational Linguistics.

John C Gower and Garmt B Dijksterhuis. 2004. *Procrustes problems*, volume 30. OUP Oxford.

Edouard Grave, Armand Joulin, and Quentin Berthet. 2019. Unsupervised alignment of embeddings with wasserstein procrustes. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1880–1890. PMLR.

Patrick JF Groenen, Patrizia Giaquinto, and Henk AL Kiers. 2005. An improved majorization algorithm for robust procrustes analysis. In *New developments in classification and data analysis*, pages 151–158. Springer.

Stevan Harnad. 1990. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1-3):335–346.

Mareike Hartmann, Yova Kementchedjhieva, and Anders Søgaard. 2019. Comparing unsupervised word translation methods step by step. *Advances in Neural Information Processing Systems*, 32.

John Hewitt, Daphne Ippolito, Brendan Callahan, Reno Kriz, Derry Tanti Wijaya, and Chris Callison-Burch. 2018. Learning translations via images with a massively multilingual image dataset. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2566–2576.

Yedid Hoshen and Lior Wolf. 2018. Non-adversarial unsupervised word translation. *arXiv preprint arXiv:1801.06126*.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *International Conference on Machine Learning*, pages 4411–4421. PMLR.

Junjie Hu, Mengzhou Xia, Graham Neubig, and Jaime Carbonell. 2019. Domain adaptation of neural machine translation by lexicon induction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2989–3001, Florence, Italy. Association for Computational Linguistics.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Jamie Kiros, William Chan, and Geoffrey Hinton. 2018. Illustrative language understanding: Large-scale visual grounding with image search. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 922–933.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. 2022. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. In *International Conference on Learning Representations (ICLR)*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Ilya Loshchilov and Frank Hutter. 2016. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*.

Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. 2019. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2630–2640.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Tomas Mikolov, Quoc V Le, and Ilya Sutskever. 2013b. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013c. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Aitor Ormazabal, Mikel Artetxe, Gorka Labaka, Aitor Soroa, and Eneko Agirre. 2019. Analyzing the limitations of cross-lingual word embedding mappings. *arXiv preprint arXiv:1906.05407*.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR.

Milos Radovanovic, Alexandros Nanopoulos, and Mirjana Ivanovic. 2010. Hubs in space: Popular nearest neighbors in high-dimensional data. *Journal of Machine Learning Research*, 11(sept):2487–2531.

Antonio Scaiella, Danilo Croce, and Roberto Basili. 2019. Large scale datasets for image and video captioning in italian. *Italian Journal of Computational Linguistics*, 2(5):49–60.

Peter H Schönemann. 1966. A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31(1):1–10.

Gunnar A Sigurdsson, Jean-Baptiste Alayrac, Aida Nematzadeh, Lucas Smaira, Mateusz Malinowski, Joao Carreira, Phil Blunsom, and Andrew Zisserman. 2020. Visual grounding in video for unsupervised word translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10850–10859.

Samuel L Smith, David HP Turban, Steven Hamblin, and Nils Y Hammerla. 2017. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. *arXiv preprint arXiv:1702.03859*.

Anders Søgaard, Sebastian Ruder, and Ivan Vulić. 2018. On the limitations of unsupervised bilingual dictionary induction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 778–788, Melbourne, Australia. Association for Computational Linguistics.

Dídac Surís, Dave Epstein, and Carl Vondrick. 2020. Globetrotter: Unsupervised multilingual translation from visual alignment. *arXiv preprint arXiv:2012.04631*.

Hagai Taitelbaum, Gal Chechik, and Jacob Goldberger. 2019. A multi-pairwise extension of Procrustes analysis for multilingual word translation. 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)*, pages 3560–3565, Hong Kong, China. Association for Computational Linguistics.

Shagun Uppal, Sarthak Bhagat, Devamanyu Hazarika, Navonil Majumder, Soujanya Poria, Roger Zimmermann, and Amir Zadeh. 2022. Multimodal research in vision and language: A review of current and emerging trends. *Information Fusion*, 77:149–171.

Ivan Vulić, Sebastian Ruder, and Anders Søgaard. 2020. Are all good word vector spaces isomorphic? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3178–3192, Online. Association for Computational Linguistics.

Chao Xing, Dong Wang, Chao Liu, and Yiye Lin. 2015. Normalized word embedding and orthogonal transform for bilingual word translation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1006–1011.

Ziyan Yang, Leticia Pinto-Alva, Franck Dernoncourt, and Vicente Ordonez. 2020. Using visual feature space as a pivot across languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3673–3678.

Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. 2021. Filip: Fine-grained interactive language-image pre-training. *arXiv preprint arXiv:2111.07783*.

Bernard Ycart. 2012. Letter counting: a stem cell for cryptology, quantitative linguistics, and statistics. *arXiv preprint arXiv:1211.6847*.

Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. 2017a. Adversarial training for unsupervised bilingual lexicon induction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1959–1970.

Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. 2017b. Earth mover's distance minimization for unsupervised bilingual lexicon induction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1934–1945.

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

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2022. Conditional prompt learning for vision-language models. In *The IEEE / CVF Computer Vision and Pattern Recognition Conference (CVPR)*.

Mingyang Zhou, Runxiang Cheng, Yong Jae Lee, and Zhou Yu. 2018. A visual attention grounding neural model for multimodal machine translation. *arXiv preprint arXiv:1808.08266*.