

# Weakly Supervised Face Naming with Symmetry-Enhanced Contrastive Loss

Tingyu Qu<sup>1\*</sup>, Tinne Tuytelaars<sup>2</sup>, Marie-Francine Moens<sup>1</sup>

<sup>1</sup> Department of Computer Science, KU Leuven

<sup>2</sup> Department of Electrical Engineering, KU Leuven

{tingyu.qu, tinne.tuytelaars, sien.moens}@kuleuven.be

## Abstract

We revisit the weakly supervised cross-modal face-name alignment task; that is, given an image and a caption, we label the faces in the image with the names occurring in the caption. Whereas past approaches have learned the latent alignment between names and faces by uncertainty reasoning over a set of images and their respective captions, in this paper, we rely on appropriate loss functions to learn the alignments in a neural network setting and propose **SECLA** and **SECLA-B**. **SECLA** is a **Symmetry-Enhanced Contrastive Learning-based Alignment** model that can effectively maximize the similarity scores between corresponding faces and names in a weakly supervised fashion. A variation of the model, **SECLA-B**, learns to align names and faces as humans do, that is, learning from easy to hard cases to further increase the performance of **SECLA**. More specifically, **SECLA-B** applies a two-stage learning framework: (1) Training the model on an easy subset with a few names and faces in each image-caption pair. (2) Leveraging the known pairs of names and faces from the easy cases using a bootstrapping strategy with additional loss to prevent forgetting and learning new alignments at the same time. We achieve state-of-the-art results for both the augmented Labeled Faces in the Wild dataset and the Celebrity Together dataset. In addition, we believe that our methods can be adapted to other multimodal news understanding tasks.

## 1. Introduction

With the rapid development of online media, news in multimedia format (e.g., text, image, and video) has become mainstream, which has driven the need for multimodal news understanding.

In multimodal news presentations, images and their respective captions enrich the content of news article text. Unlike in the conventional image captioning task [10, 23, 26],

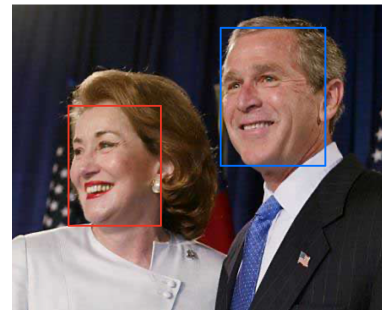


Figure 1. News Image with Caption: President **George W. Bush** attends a fundraiser for former Secretary of Transportation **Elizabeth Dole** while in Greensboro, North Carolina, July 25, 2002. Dole is running for the U.S. Senate seat to be vacated in November by Senator **Jesse Helms**.

where the contents of the image and its caption are often semantically well aligned, a news image-caption pair usually has weak semantic alignment, which makes the automated understanding of news images a challenging task.

Frequently, nonfamous people are shown in news images, for which no separate face recognizer is trained. However, in such cases, weak supervision of the caption helps in assigning names to faces. Finding which name in the caption corresponds to which face in the image is referred to as the face-name alignment task. Interpreting the image and the caption in Figure 1, we will assign the face in the blue bounding box to the name George W. Bush, the face in the red box to the name Elizabeth Dole and "NOFACE" (i.e., names without corresponding faces) will be assigned to the name "Jesse Helms".

In this work, we revisit the weakly supervised face-name alignment task. The task is tailored to real-world multimodal news understanding by disambiguating names and faces of different persons in a weakly supervised setting, where explicit annotations of faces are lacking. This task is different from face recognition tasks [18], which usually require carefully curated labels for training. With the significant number of news articles published every day, it is almost impossible to maintain such a database for face

\*We acknowledge the support from China Scholarship Council, CELSA/19/018, and MACCHINA project (KU Leuven C14/18/065).

recognition.

Specifically, to address the weakly supervised face-name alignment problem, we propose **SECLA**, a **Symmetry-Enhanced Contrastive Learning Alignment** model consisting of pretrained feature extractors for feature extraction and multilayer perceptron (MLP) projectors for mapping features of faces and names onto the same embedding space. Given projected features, SECLA is trained with symmetry-enhanced loss functions, including a bidirectional contrastive loss function and an agreement loss function. The design of the symmetry-enhanced loss function is motivated by the assumption that each face corresponds to at least one name (i.e., dense alignment), which is often violated when null links are present, that is, names with no corresponding faces or faces with no corresponding names, as pointed out by [12, 13].

SECLA achieves better performance than the previous state-of-the-art (SOTA) model [17]. Moreover, for easier subsets that contain few faces and names in each image-caption pair, SECLA can achieve (near-)perfect alignment results, which motivates us to propose **SECLA-B**, a model that learns with a two-stage learning strategy and bootstrapping. We first train a SECLA model on an easy subset and then add the rest of the data into the second stage of training with loss functions that can utilize the known pairs of faces and names and their prototypes. SECLA-B further improves the SECLA performance. Thanks to the simple architectures, our proposed methods have great potential to be adapted to other multimodal news understanding tasks.

In summary, our contributions are threefold. (1) We propose a contrastive learning-based model, called SECLA, that can accurately learn the alignment between faces and names in image caption pairs in a weakly supervised fashion. (2) To address violations of the dense alignment assumption, we propose a novel symmetry-enhanced contrastive loss function that learns to disambiguate when the model is unsure about the correct alignment. The SECLA model with symmetry-enhanced losses achieves SOTA performance for the task. We achieve (near-)perfect performance for the easy dataset with few faces and names in each image-caption pair. (3) We propose a two-stage learning strategy with bootstrapping to utilize the (near-)perfect performance of SECLA on the easy subset. The model trained with this strategy, called SECLA-B, further improves the performance of SECLA.

## 2. Related Works

Past studies on face-name alignment can be divided into two categories:

**Without external data.** Most face-name alignment studies have focused on exploring the relationship of image-caption pairs in a dataset without the use of external data mined from the web. These methods are typically built with

an uncertainty reasoning framework. Berg *et al.* [1, 2] proposed methods based on modified k-means clustering for captioned news images. Guillaumin *et al.* [11] adopted a graph-based approach for retrieving all faces of the same person and aligning names to faces in each image-caption pair. Su *et al.* [19, 20] focused on an easier problem where only people with names occurring more than a certain number of times were studied and achieved satisfactory results. Pham *et al.* [17] extended [1] with a probabilistic approach under the EM framework. Given a good initialization of clusters with image-caption pairs, their method can make alignments in both directions, namely, name-to-face and face-to-name, using uncertainty reasoning. Their method achieved SOTA results on the Labeled Faces in the Wild (LFW) dataset. In this paper, we study the weakly supervised face-name alignment problem, as in [17]. Different from past works without external data, we do not rely on uncertainty reasoning; instead, we learn alignments in a neural network setting by designing appropriate loss functions that help disambiguate possible alignment issues in a weakly supervised setting.

**With external data.** Another line of research has focused on designing systems that utilize external data (mostly web data) to augment internal data, especially non-celebrities data that occur only a few times. Zhang *et al.* [27] explored the syntactic structure of captions for initializing the ranking of names and applied web mining to acquire extra data for rare celebrities. Chen *et al.* [5, 6] designed graph-based alignment methods to satisfy predefined constraints and used a Google image search for external data. Recent work by Tian *et al.* [22] proposed a system that integrates deep multimodal analysis, cross-modal correlation learning, and web mining. Their proposed system is similar to that in [27].

Apart from the face-name alignment task, our work is also closely related to a similar task, that is, **image-sentence matching**. Karpathy *et al.* [13] discussed the violation of the dense alignment assumption and proposed an alignment model based on hinge loss. Hessel *et al.* [12] studied an unsupervised image-sentence matching problem. They proposed a dense similarity measure for calculating the similarity between image-sentence pairs and a hinge loss function with a top-k selection process to mitigate the violation of the dense alignment assumption [12, 13]. We adopt a similar dense similarity measure, as detailed in Sec. 3.3.

Finally, our work is related to **visual grounding**. Wang *et al.* [25] adopted a contrastive loss to learn the alignment between objects and phrases in a weakly supervised fashion. Cui *et al.* [7] proposed the human-centric visual grounding task called Who’s Waldo, where the aim is to link anonymized names in the caption to the bounding boxes of people in the image. The authors added a NONAME constant to deal with null links.

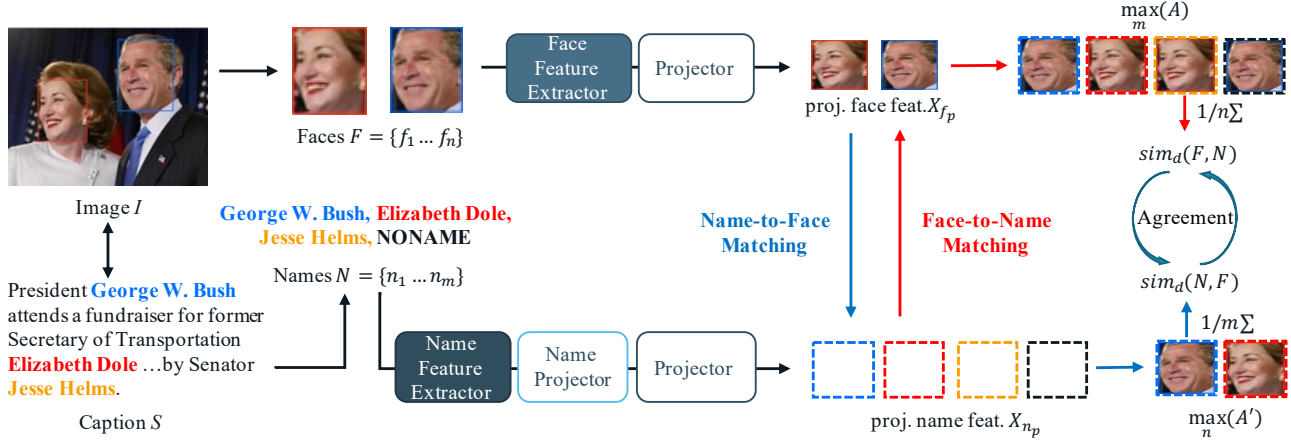


Figure 2. The architecture of the proposed SECLA model. We first use feature extractors (FaceNet for faces and BERT for names) to obtain the features of faces  $f_1, \dots, f_n$  in image  $I$  and names  $n_1, \dots, n_m$  (with one NONAME added) in caption  $S$ . The features are passed through projectors to a common embedding space. Then, we learn the alignment of faces and names in both directions, that is, the name-to-face and face-to-name alignments, with a contrastive loss. To further enhance the symmetry from the two directions, we apply an agreement loss to the similarity matrices that result from the computation of dense similarity measures.

### 3. Methodology

In this section, we formally describe the face-name alignment problem and introduce the proposed SECLA model. We then present the framework for training SECLA and a detailed explanation of two-stage learning with a bootstrapping strategy for SECLA-B. We conclude the section with the alignment scheme for the problem.

#### 3.1. Problem Formulation

Given a news image  $I$  with faces  $F = \{f_1, f_2, \dots, f_n\}$  and the corresponding caption  $S$  with names  $N = \{n_1, n_2, \dots, n_m\}$ , the goal of the face-name alignment task is to find the links between corresponding face  $f_i$  and name  $n_j$ . We observe two types of links: (1) Normal links: these are links with correctly detected faces and names, which appear in images and the corresponding captions. (2) Null links: these are links containing faces without names ("NONAME") or names without faces ("NOFACE").

#### 3.2. Model Architecture

As shown in Figure 2, our SECLA model consists of two modules: (1) a feature extractor module for extracting features from face  $f_i$  and name  $n_j$  and (2) a projector module for projecting features onto a common embedding space.

The feature extractor module includes (1) a pretrained FaceNet [18] to extract embedding  $X_f^i \in \mathbb{R}^{d_f}$  for face  $f_i$  and (2) a pretrained BERT [8] to produce name embedding  $X_n^j \in \mathbb{R}^{d_n}$  for name  $n_j$ .

The projector module consists of two types of projectors, (1) name projectors and (2) common projectors. We first pass the name embedding  $X_n$  through a name projec-

tor  $g_n(\cdot) : \mathbb{R}^{d_n} \rightarrow \mathbb{R}^{d_f}$  to obtain the projected embedding  $X_{n'}$ , which has the same dimension as face embedding  $X_f$ . Then, we use two common projectors  $g_c(\cdot) : \mathbb{R}^{d_f} \rightarrow \mathbb{R}^{d_p}$  to project  $X_{n'}$  and  $X_f$  onto the same embedding space with a lower dimension  $d_p$ . We denote  $X_{n_p} \in \mathbb{R}^{d_p}$  and  $X_{f_p} \in \mathbb{R}^{d_p}$  as the projected features for  $X_{n'}$  and  $X_f$ , respectively. We calculate the similarity score between  $f_i$  and  $n_j$  as  $\text{sim}(f_i, n_j) = (X_{f_p}^i)^T \cdot X_{n_p}^j$  for model training and subsequent inferencing.

#### 3.3. Training the SECLA model

To train our SECLA model, we propose the following loss functions that operate on projected features  $X_{n_p}^j$  and  $X_{f_p}^i$ : (1) contrastive loss that maximizes the dense similarities between corresponding images and captions and (2) agreement loss that forces the symmetry from the similarity matrices computed from two directions.

**Contrastive Loss.** Inspired by previous works in caption ranking [9] and visual grounding [25], we apply the following contrastive loss computed from the face-to-name direction:

$$\mathcal{L}_{f,n} = -\log \frac{e^{\text{sim}_d(F,N)}}{\sum_{F_k \in \text{batch}} e^{\text{sim}_d(F_k,N)}} \quad (1)$$

The dense similarity measure  $\text{sim}_d(F, N)$  between image  $I$  and caption  $S$  is defined as:

$$\text{sim}_d(F, N) = \frac{1}{n} \sum_{i=1}^n \max_j A_{i,j} \quad (2)$$

where  $A_{i,j} = \text{sim}(f_i, n_j)$ ,  $i = 1, 2, \dots, n$ ,  $j = 1, 2, \dots, m$

Similar to the face-to-name contrastive loss  $\mathcal{L}_{f,n}$ , we define name-to-face contrastive loss as:

$$\mathcal{L}_{n,f} = -\log \frac{e^{\text{sim}_d(N,F)}}{\sum_{N_k \in \text{batch}} e^{\text{sim}_d(N_k,F)}} \quad (3)$$

with  $\text{sim}_d(N, F) = \frac{1}{m} \sum_{j=1}^m \max_i A'_{j,i}$ , where  $A'_{j,i} = \text{sim}(n_j, f_i)$ ,  $j = 1, 2, \dots, m$ ,  $i = 1, 2, \dots, n$ .

In a weakly supervised setting, the connection between faces and names within the same image-caption pair remains unknown during training, and we must rely on dense similarity measures to learn a meaningful embedding space. However, as illustrated in [12, 13], the dense alignment assumption can be violated when null links exist. Unlike the phrase-region matching process in visual grounding, null links are extremely common in our task. As long as the numbers of faces and names in one image-caption pair are not equal or there exist noncelebrity faces, null links can be expected. Even with only one face in an image, we can have no corresponding name in the caption, and vice versa. To this end, unlike in visual grounding [25], where the impact of the more symmetrical contrastive loss function is limited, the symmetrical loss functions in our task are important. Consequently, we compute the contrastive loss functions in two directions, that is, face-to-name and name-to-face, to symmetrize the computation of the similarity scores.

**Agreement Loss.** The bidirectional contrastive loss can mitigate violations of the dense alignment assumption. However, this is not sufficient for our task. Especially in the early stage of training, with randomly initialized projectors, it is difficult for the model to differentiate between names or to guide the gradient flow in the right direction.

To further enhance the symmetry of our bidirectional model, we design an agreement loss for additional regularization during training, which has been proven effective in word alignment models [4, 16]. Our agreement loss forces the dense similarity scores computed from the face-to-name and name-to-face directions to be as close as possible. Without it, we could have more aligned face-name pairs differ from two directions due to the  $\max(\cdot)$  operator. Specifically, for a mini batch of  $B$  samples, each containing  $m$  names and  $n$  faces, we obtain a tensor  $Z$  of size  $B \times B \times n \times m$  to represent all combinations of  $B$  images and  $B$  captions. By computing dense similarity measures from two directions, we obtain matrices  $D_{n,f}$  and  $D_{f,n}$  of size  $B \times B$ , whose diagonal elements correspond to the image-caption pairs. Then, we define the agreement loss as:

$$\mathcal{L}_{\text{agree}} = \text{MSE}(\text{diag}(D_{n,f}), \text{diag}(D_{f,n})) \quad (4)$$

where each entry of the similarity matrices  $D_{n,f}$  and  $D_{f,n}$  results from the computation of  $\text{sim}_d(F_k, N_k)$  and  $\text{sim}_d(N_k, F_k)$  for  $k$ th image-caption pair in the batch, re-

spectively.  $\text{diag}(\cdot)$  denotes the diagonal elements of a matrix.

**Total Loss** We combine Equations 1, 3, and 4 to obtain the final loss function:

$$\mathcal{L} = \mathcal{L}_{f,n} + \mathcal{L}_{n,f} + \alpha \mathcal{L}_{\text{agree}} \quad (5)$$

where  $\alpha$  is a hyperparameter.

### 3.4. Two-Stage Training with Bootstrapping

We extend SECLA with a two-stage training strategy to adapt it to the need for learning new faces and names from the stream of news in a real-world scenario. For stage1 training, we train a SECLA model (stage1 model) on an easy subset ( $\mathcal{D}_{\text{easy}}$ ) of the full dataset ( $\mathcal{D}_{\text{all}}$ ), which contains one or two potential links in each pair. We denote the set of names that appear in stage1 training without repetition as  $N_{\text{unique}}$ . Next, we present the different strategies for stage 2 training in this part.

We first present a two-stage training strategy using simple heuristics in Algorithm 1. After stage1 training, we split the rest of dataset  $\mathcal{D}_{\text{rest}}$  into two subsets, namely,  $\mathcal{D}_{\text{match}}$  and  $\mathcal{D}_{\text{unmatch}}$ , where  $\mathcal{D}_{\text{match}}$  contains all the pairs matched by the stage1 model\*, and  $\mathcal{D}_{\text{unmatch}}$  contains all the unmatched pairs. Then, we fine-tune the stage1 model on  $\mathcal{D}_{\text{unmatch}}$  with the same loss used for stage1 training.

---

#### Algorithm 1: Two-stage Training using Simple Heuristics

---

**Input :** Easy subset  $\mathcal{D}_{\text{easy}}$ , complementary subset  $\mathcal{D}_{\text{rest}}$ , parameter  $\theta_o$  of stage1 model  $g_{\text{stage1}}(\cdot)$  trained on  $\mathcal{D}_{\text{easy}}$ , Set of unique names  $N_{\text{unique}}$  in  $\mathcal{D}_{\text{easy}}$

**Output:** Stage 2 model  $g_{\text{stage2}}(\cdot)$

**Initialize:**  $\mathcal{D}_{\text{match}} = \emptyset$  and contains pairs matched in  $\mathcal{D}_{\text{rest}}$  using the stage1 model

**for** Each image-caption pair  $\{I, S\} \in \mathcal{D}_{\text{rest}}$  contains  $\{f_i, n_j\}$ ,  $i = 1, \dots, n$ ,  $j = 1, \dots, m$  **do**

**if**  $n_j \in N_{\text{unique}}$  **then**

$f_{\text{aligned}} = \arg \max_{f_i \in F} (g_{\text{stage1}}(f_i, n_j))$

$\mathcal{D}_{\text{match}} \leftarrow \mathcal{D}_{\text{match}} \cup \{f_{\text{aligned}}, n_j\}$

**else**

        Continue

Let  $\mathcal{D}_{\text{unmatch}} = \mathcal{D}_{\text{rest}} \setminus \mathcal{D}_{\text{match}}$

$g_{\text{stage2}}(\cdot) \leftarrow \text{Fine-tune}(\mathcal{L}, g_{\text{stage1}}(\cdot), \mathcal{D}_{\text{unmatch}})$

---

Using simple heuristics, the trained model can obtain reasonably good performance, as shown below, given a stage1 model with perfect alignment results. However, (1) a stage1 model with 100% correctness cannot always be

\*The matching is based on inference by computing  $\text{sim}(f_i, n_j)$  with projected features from  $g_{\text{stage1}}$  for face  $f_i$  and name  $n_j$ . We match the names in  $N_{\text{unique}}$  with the faces.

expected, (2) Even with a stage1 model that achieves perfect alignment results for  $D_{easy}$ , we can still make mistakes while matching new faces in  $D_{rest}$  with known names, leading to errors in  $D_{unmatch}$  for fine-tuning the model. (3) As more new names and faces are added during training, the model tends to forget the known names and faces.

To tackle the problems of Algorithm 1, we propose a two-stage training strategy combined with bootstrapping, as shown in Algorithm 2; it includes new loss functions to prevent forgetting and learns better alignments.

---

**Algorithm 2:** Two-stage Training with Bootstrapping (SECLA-B)

---

**Input :** Easy subset  $\mathcal{D}_{easy}$ , Full dataset  $\mathcal{D}_{all}$ , Stage1 model  $g_{stage1}(\cdot)$  trained on  $\mathcal{D}_{easy}$  with parameter  $\theta_o$ , set of unique names and the corresponding faces seen in stage1  $\{N_{unique}, F_{seen}\} \subset \mathcal{D}_{easy}$

**Output:** Stage2 model  $g_{stage2}(\cdot)$

**for** Each batch with B image-caption pairs **do**

**Initialize** batch of matched names and faces

$B_{match} = \{\}$ , batch of unmatched names and faces  $B_{unmatch} = \{\}$ , model  $g(\cdot)$  with  $\theta_0$  of  $g_{stage1}(\cdot)$

**for** Each pair  $\{I, S\} \in$  batch of  $\mathcal{D}_{all}$ , containing  $\{f_i, n_j\}, i = 1, \dots, n, j = 1, \dots, m$  **do**

**if**  $n_j \in N_{unique}$  **then**

$f_{aligned} = \arg \max_{f_i \in F} (g(n_j, f_i))$   
             $B_{match} \leftarrow B_{match} \cup \{f_{aligned}, n_j\}$

**else**

$B_{unmatch} \leftarrow B_{unmatch} \cup \{f_i, n_j\}$

**if**  $len(B_{match}) > 0$  **then**

$g_{stage2}(\cdot) \leftarrow \text{Train}(\mathcal{L}, g(\cdot), B_{unmatch}) + \text{Train}(\mathcal{L}_{stage2}, g(\cdot), B_{match})$

**else**

$g_{stage2}(\cdot) \leftarrow \text{Train}(\mathcal{L}, g(\cdot), B_{unmatch})$

---

After matching the faces using  $N_{unique}$  and the stage1 model, we consider the remaining unmatched pairs in  $B_{unmatch}$  that are not seen in stage1 training. Naturally, we can treat these pairs the same way as the pairs in  $D_{easy}$  in stage1 training. Hence, we apply the same loss used for stage1 training, as presented in Equation 5, for  $B_{unmatch}$ .

For  $B_{match}$ , it is desired that our model can maximize the dense similarity measures between the set of matched faces  $F'_i$  and the set of matched names  $N'_j$  from the same sample. Inspired by Li *et al.* [15], we propose a contrastive loss to pull matched faces closer to their respective names in the same sample but push them away from the retrieved prototype faces in other samples. Specifically, given  $k$  samples of image-caption pairs with matched name sets

$\{N'_1, N'_2, \dots, N'_k\}$ , matched face sets  $\{F'_1, F'_2, \dots, F'_k\}$  and retrieved prototype sets  $\{P_1, P_2, \dots, P_k\}$ , we have:

$$\mathcal{L}_{f,n,p} = -\log \frac{e^{sim_d(F'_i, N'_i)}}{e^{sim_d(F'_i, N'_i)} + \sum_{P_j \in P \setminus \{P_i\}} e^{sim_d(P_j, N'_i)}} \quad (6)$$

$$\mathcal{L}_{n,f,p} = -\log \frac{e^{sim_d(N'_i, F'_i)}}{e^{sim_d(N'_i, F'_i)} + \sum_{P_j \in P \setminus \{P_i\}} e^{sim_d(N'_i, P_j)}} \quad (7)$$

We force the symmetry from the two directions as:

$$\mathcal{L}_{f,n,p}^B = \mathcal{L}_{f,n,p} + \mathcal{L}_{n,f,p} \quad (8)$$

In addition, to learn more clustered representations of the faces, it is desired that we can pull the representations of matched faces closer to their prototypes in the same sample and push them away from other prototypes. We design a contrastive loss that operates on the face-prototype level as:

$$\mathcal{L}_{f,p} = -\log \frac{e^{sim_d(F'_i, P_i)}}{\sum_{P_j \in P} e^{sim_d(F'_i, P_j)}} \quad (9)$$

$$\mathcal{L}_{p,f} = -\log \frac{e^{sim_d(P_i, F'_i)}}{\sum_{P_j \in P} e^{sim_d(P_j, F'_i)}} \quad (10)$$

We denote the bidirectional loss as:

$$\mathcal{L}_{f,p}^B = \mathcal{L}_{f,p} + \mathcal{L}_{p,f} \quad (11)$$

Finally, we denote the additional loss for  $B_{match}$  as:

$$\mathcal{L}_{stage2} = \mathcal{L}_{f,n,p}^B + \mathcal{L}_{f,p}^B \quad (12)$$

In practice, given the known names and faces from stage1 training, different types of prototype faces can be selected: (1) randomly sampled faces, (2) average faces, (3) medoid faces with the minimum average distance to other faces of the same people, and (4) faces matched by the model because they have the highest similarity score with a name. However, due to factors such as orientation, blurry conditions or lighting conditions, face representations can vary greatly even for the same person. Depending on the data distribution, the model performance is affected by different types of prototype faces. We discuss the effect of different choices of prototype faces in Sec. 5.3.

### 3.5. Alignment Scheme

We align faces with names according to the similarity scores from  $sim(f_i, n_j) = (X_{f_p}^i)^T \cdot X_{n_p}^j$ . Specifically, the alignment scheme consists of three parts:

- For both training and inference, we add additional NONAME to all the name lists without NONAME.<sup>†</sup> Then, NONAME is treated as an extra name. The NONAME embedding is fixed during training.

<sup>†</sup>Since "NOFACE" links are not included in the Celebrity Together (CelebTo) dataset, we do not add constant NOFACE embeddings to the face list during inference.

- For face  $f_i$ , if  $\arg \max_m A_{i,j} = n_m$ , we align  $f_i$  to  $n_m$ .
- If "NOFACE" links are considered, after aligning all the faces, we align the remaining names to NOFACE.

The alignment scheme is applied during inference.

## 4. Experimental Setup

### 4.1. Datasets

We conduct extensive experiments on the same LFW dataset used in [17]. Following [17], we extract names with the named entity recognizer from OpenNLP package<sup>‡</sup> and detect bounding boxes for faces with the OpenCV implementation of [24]. We obtain 10976 pairs of news images and captions, covering 23928 links, from LFW. On average, 1.97 names and 1.32 faces appear in each pair.

We also use the Celebrity Together (CelebTo) dataset [28]. CelebTo has the same numbers of faces and names in each pair. It does not include "NOFACE" links and no captions are provided. It contains 193523 images with 545519 faces and 2533 identities. The average number of faces in each image is 2.82, and that of NONAME is 1.16.

### 4.2. Implementation Details

We adopt Inception ResNet (V1) [21] pre-trained on VG-Face2 [3] as the face feature extractor and BERT-based [8] as the name feature extractor. The name representation is the average embedding from all the tokens of the summation of the last four layers of BERT. We fix the embedding for "NONAME" as that for the "[UNK]" token. We construct a one-layer MLP as the name projector  $g_n(\cdot) : \mathbb{R}^{768} \rightarrow \mathbb{R}^{512}$  and a three-layer MLP as the common projector  $g_c(\cdot) : \mathbb{R}^{512} \rightarrow \mathbb{R}^{128}$ . We use ReLU as the activation function. We set the hyperparameter  $\alpha=0.15$  in Equation 5 and the threshold in the hinge loss to 0.2 [12]. We train our model using the Adam optimizer [14] with a learning rate  $=3 \times 10^{-4}$ . We train SECLA with a batch size of 20 for 30 epochs and 3 epochs for LFW and CelebTo, respectively. For SECLA-B, the stage1 model is trained on the subset ( $D_{one}$ ) with one name and one face and the subset ( $D_{2name}$ ) with two names and two faces for LFW and CelebTo, respectively. We exclude NONAME from both sets. Stage1 training takes 15 epochs and 5 epochs, and stage2 training takes 20 epochs and 2 epochs for LFW and CelebTo, respectively. The models are trained on one NVIDIA TITAN Xp GPU.

### 4.3. Evaluation Metrics

Due to the existence of null links in the LFW dataset that might refer to different names or faces, we do not use accuracy [1, 2] to evaluate aligned links. Following [17],

<sup>‡</sup><http://opennlp.sourceforge.net/>

we adopt precision, recall, and F1 score as the evaluation metrics, where

$$\text{Precision} = \frac{\# \text{link}_{\text{correct}}}{\# \text{link}_{\text{found}}}, \text{Recall} = \frac{\# \text{link}_{\text{correct}}}{\# \text{link}_{\text{gt}}}, \text{F1} = \frac{2 \text{Recall} \cdot \text{Precision}}{\text{Recall} + \text{Precision}} \quad (13)$$

For CelebTo, we use accuracy for evaluation.

## 5. Results

### 5.1. Overall Performance

In Table 1, we compare different alignment models on the LFW and CelebTo datasets. For LFW, we choose the best performance reported in [17] in terms of F1 score and recall for comparison. We also implement models trained with the loss proposed in [25], which is equivalent to the unidirectional version of our contrastive loss, as well as the learning strategy proposed in [12]. For CelebTo, we consider the face-name alignment task as a metric learning problem by excluding all the context information (e.g., syntactic structures of the captions, co-occurrence of faces in images) from training, which was heavily relied on in [17].

Our SECLA model outperforms the previous SOTA and other approaches from related fields, as shown in Table 1. In particular, we find that for LFW, SECLA improves the results of the model in [17] by a large margin in terms of recall and F1 but only achieves slightly better precision (74.90% vs 76.96%). This is because SECLA always identifies all the links within an image-caption pair, in contrast to [17], which only finds parts of the links above a certain threshold. According to Equation 13, we calculate precision as the number of correct links divided by the number of links found, so it is expected that we cannot achieve high precision in that case. Moreover, the models trained with unidirectional loss  $\mathcal{L}_{f,n}$  and  $\mathcal{L}_{n,f}$  achieve very poor performance with F1 scores of 43.61% and 39.34%, respectively, on LFW and accuracy of 44.77% and 41.29%, respectively, on CelebTo. We randomly initialize our projectors so that the semantic meaning of face features and name features are not close enough, leading to the initialized scores for different names and faces falling into a small interval near zero (mostly [-0.1, 0.1]). In that case, without proper symmetry-enhanced regularization, instead of optimizing the similarity scores for matched pairs toward the positive side, the model randomly optimizes the scores. The model in [12] achieves relatively better results on LFW but performs much worse on CelebTo. With the strong constraint of margins in the decision boundary, the hinge loss used in [12] makes the lack of semantic similarities of faces and names at the early stage of training a big problem. Our contrastive loss functions in Equations 1 and 3 do not have this problem and thus achieve significantly better results.

Adding agreement loss increases the performance of SECLA from F1=78.14% to 80.83% on LFW, and accuracy=81.79% to 87.46% on CelebTo. The agreement loss

Model	Contrastive	Symmetry	Precision	LFW Recall	F1	CelebTo Accuracy
Pham <i>et al.</i> [17] (using $P(f n)$ )			74.90%	70.56%	72.66%	/
Pham <i>et al.</i> [17] (using $P(n f)$ )			69.99%	72.73%	71.33%	/
Wang <i>et al.</i> [25] ( $\mathcal{L}_{f,n}$ )	✓		41.52%	45.92%	43.61%	44.77%
Wang <i>et al.</i> [25] ( $\mathcal{L}_{n,f}$ )	✓		37.46%	41.42%	39.34%	41.29%
Hesssel <i>et al.</i> [12] (w/o TK)		✓	61.35%	67.84%	64.43%	45.70%
Hesssel <i>et al.</i> [12] (TK)		✓	63.51%	70.23%	66.70%	40.31%
SECLA- $\alpha\mathcal{L}_{agree}$ (ours)	✓	✓	74.40%	82.28%	78.14%	81.79%
SECLA (ours)	✓	✓	76.96%	85.11%	80.83%	87.46%
SECLA-B (ours)	✓	✓	<b>77.94%</b>	<b>86.19%</b>	<b>81.86%</b>	<b>88.36%</b>

Table 1. Performance Comparison of the Alignment Models on the LFW and CelebTo Datasets. We compare our results with the best results from [17], the results of the model trained with the loss function in [25] (which is equivalent to our single-directional loss function), and the results of the model trained using the dense similarity measure with hinge loss and the top k selection strategy (TK), as in [12].

works better when more faces and names appear.

Model	Precision	LFW Recall	F1	CelebTo Accuracy
Pipeline	69.94%	77.34%	73.46%	67.50%
SECLA	76.96%	85.11%	80.83%	87.46%
SECLA-B	<b>77.94%</b>	<b>86.19%</b>	<b>81.86%</b>	<b>88.36%</b>

Table 2. Performance comparison of SECLA-B on LFW and CelebTo. The Pipeline model is trained with Algorithm 1.

We present the performance comparison of SECLA-B with the baselines in Table 2. We take SECLA and the model learned with Algorithm 1 (denoted as Pipeline) as baselines. As shown in Table 2, we achieve good results with Pipeline; i.e., F1 = 73.46% on LFW and accuracy = 67.50% on CelebTo. However, the performance of Pipeline is much worse than that of SECLA. Based on Algorithm 1, the errors made by the stage1 model accumulate during fine-tuning. Specifically, after stage1 training on an easy subset, we split the rest of the dataset into  $D_{match}$  and  $D_{unmatch}$  and fine-tune the stage1 model on  $D_{unmatch}$ .  $D_{match}$  contains pairs of matched faces and names, which are not guaranteed to be correctly aligned. Even if SECLA achieves perfect alignment results for stage1 training, for a new face of a known name, the model could still make mistakes. Incorrect matches made by the stage1 model result in incorrectly matched faces and names in  $D_{unmatch}$ , which is used for fine-tuning the stage1 model. Hence, the error accumulates and causes performance degradation. Despite the errors made by the stage1 model, SECLA-B further improves the performance of the strong SECLA model.

## 5.2. Case Study for SECLA

### Experiments on $D_{easy}$

In this part, we apply the SECLA model on easy subset  $D_{easy}$ , which is also the stage1 training of SECLA-B. The performance of the stage1 model is vital for SECLA-B since

Model	Add NONAME	Accuracy
SECLA	✓	<b>96.26%</b>
SECLA		90.52%
SECLA- $\alpha\mathcal{L}_{agree}$	✓	95.27%

Table 3. Performance of SECLA on  $D_{2name}$  subset of CelebTo.

we have to rely on it for matching faces in stage2 training.

For LFW, we extract a subset with one face in each pair ( $D_{one}$ ) as  $D_{easy}$ . We exclude NONAME and NOFACE to decrease the level of ambiguity. The best alignment results achieved by [17] are precision = 85.41%, recall = 91.17% and F1 = 88.19%. Aligning one name to one face is our task in its easiest form, yet the model in [17] confuses common names with NONAME in approximately 10% of the cases. For comparison, our SECLA model achieves perfect alignment results by correctly aligning all the links.

In CelebTo, each image contains at least two faces. Thus, we extract a subset with two faces in each pair ( $D_{2name}$ ) as  $D_{easy}$ ; it contains 53334 samples with 2437 unique names. Although this is a more difficult task, where we have to align two names to two faces in each sample, the majority of the names occur more than 10 times in  $D_{2name}$ , making the task slightly easier.

In addition, we conduct ablations of SECLA on  $D_{2name}$ . As presented in Table 3, without agreement loss, the accuracy decreases from 96.26% to 95.27%. Since most names occur frequently in  $D_{2name}$ , which is favorable for the contrastive loss to differentiate between pairs, the symmetry constraint imposed by agreement loss indeed helps the model to learn more accurate alignments. Adding NONAME in the name list during training, we observe a large improvement in accuracy from 90.52% to 96.26%.

### Experiments with More Ambiguities

For cases with more ambiguities, we conduct experiments on a subset of LFW with at least 2 faces. This subset contains 3029 image-caption pairs with 8960 links. Al-

though the number of links is not very high compared to that of image-caption pairs, there are 4177 unique names within the 8960 links. Most of the names only appear once or twice. Furthermore, we obtain 776 incorrect links, which can mislead the model to learn incorrect alignments.

We present the results in Table 4. The model in [17] struggles in this more ambiguous task, i.e.,  $F1 = 47.48\%$ . The performance of the unidirectional model varies greatly ( $F1=65.45\%$  for name-to-face and  $F1=47.87\%$  for face-to-name). With and without agreement loss, our bidirectional model achieves very positive results, i.e.,  $F1=66.89\%$  and  $F1=65.97\%$ , respectively. The challenge of this task is that there are too many unique names compared to the number of links, which is not beneficial for the contrastive loss. However, our model still outperforms the previous SOTA [17] by a large margin, and the agreement loss is proven to be useful. We also observe improvements in performance by adding the agreement loss to unidirectional models.

Model	Precision	Recall	F1
Pham <i>et al.</i> [17]	44.05%	51.51%	47.48%
SECLA	<b>64.00%</b>	<b>70.06%</b>	<b>66.89%</b>
SECLA - $\alpha\mathcal{L}_{agree}$	63.11%	69.10%	65.97%
SECLA - $\mathcal{L}_{n,f}$	61.80%	67.66%	64.59%
SECLA - $\mathcal{L}_{f,n}$	63.06%	69.04%	65.91%
SECLA - $\alpha\mathcal{L}_{agree} - \mathcal{L}_{n,f}$	45.79%	50.13%	47.87%
SECLA - $\alpha\mathcal{L}_{agree} - \mathcal{L}_{f,n}$	62.61%	68.55%	65.45%

Table 4. Performance of SECLA on LFW with at least 2 faces.

### 5.3. Case Study for SECLA-B

#### Choices of Prototype Faces

As illustrated in Sec. 3.4, we consider four types of prototype faces, namely, (1) a randomly selected face (random face), (2) an average face (avg. face), (3) a face with a minimum average distance to the other faces (medoid face), and (4) a face with the highest similarity score to the respective name (matched face). We present the best performance achieved by SECLA-B with different choices of prototype faces in Table 5. For LFW, different choices of prototype faces have limited effects on the performance of SECLA-B. The main reason for that is that most of the names in the subset ( $D_{one}$ ) used for stage1 training occur only once or twice. In that case, the effect of different choices of prototypes is limited. For CelebTo, we observe that using the average face causes significant degradation in performance compared to using other types of prototype faces or even SECLA (87.46% accuracy on CelebTo). The cause of this is also related to the data distribution. As stated before, only 2437 unique names can be found in 53334 samples of  $D_{2name}$ . For most identities, due to variations in faces, the average face calculated from hundreds of faces is no longer a good representative. In general, CelebTo favors prototype

faces with more diversity (random face/matched face).

Prototype Type	Precision	LFW		CelebTo Accuracy
		Recall	F1	
random face	77.65%	85.87%	81.55%	<b>88.36%</b>
avg. face	<b>77.94%</b>	<b>86.19%</b>	<b>81.86%</b>	81.19%
medoid face	77.87%	86.12%	81.79%	86.07%
matched face	77.74%	85.97%	81.61%	87.94%

Table 5. Best performance achieved by SECLA-B with different choices of prototype faces for stage 2 training.

#### Ablation Studies of SECLA-B

In this part, we present ablation studies of SECLA-B. We show the results of SECLA-B trained with matched faces, and these results are satisfactory for both datasets. We also explore the effect of adding the NOFACE to the matched face list, which is not considered for SECLA. Unlike in SECLA, where no matching is realized, incorrectly matched faces could lead to errors that accumulate during stage2 training of SECLA-B.<sup>§</sup> The additional NOFACE could potentially mitigate this problem. The results in Table 6 show that adding NOFACE is effective, especially for CelebTo, where the stage1 model only correctly aligns 96.26% of the links. Overall, we achieve the best performance with both  $\mathcal{L}_{n,f,p}^B$  and  $\mathcal{L}_{f,p}^B$  used during stage2 training. We observe that  $\mathcal{L}_{n,f,p}^B$  contributes the most to the SECLA-B performance. And  $\mathcal{L}_{f,p}^B$  alone is not enough to achieve satisfactory results when the performance of the stage1 model is limited.

Model	NOFACE	Precision	LFW		CelebTo Accuracy
			Recall	F1	
SECLA-B	✓	<b>77.74%</b>	<b>85.97%</b>	<b>81.61%</b>	<b>87.94%</b>
SECLA-B		77.25%	85.43%	81.13%	86.74%
SECLA-B- $\mathcal{L}_{n,f,p}^B$	✓	77.47%	85.67%	81.37%	80.90%
SECLA-B- $\mathcal{L}_{f,p}^B$	✓	77.62%	85.84%	81.52%	87.11%

Table 6. Ablation study of SECLA-B with matched faces.

## 6. Conclusion

We present a bidirectional contrastive learning framework (SECLA) enhanced with forced symmetry to address the weakly supervised face-name alignment problem without any annotated training examples or external data mined from the web. The method shows satisfactory performance on LFW and CelebTo. Moreover, we present the SECLA-B model trained with a two-stage learning strategy and bootstrapping; it further improves the performance of SECLA. Our methods can be adapted to other multimodal news understanding tasks, including multimodal summarization and news image captioning, which we leave for future research.

<sup>§</sup>Our experiments also show that the additional NOFACE have little effect on the performance of SECLA. We fix the embedding of NOFACE as  $[F_{n,1}, F_{n,2}, \dots, F_{n,512}]$  with  $F_{n,i} \sim N(0, 1)$ .

## References

- [1] Tamara Berg, Alexander Berg, Jaety Edwards, and David Forsyth. Who's in the picture. In L. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems*, volume 17. MIT Press, 2005.
- [2] T.L. Berg, A.C. Berg, J. Edwards, M. Maire, R. White, Yee-Whye Teh, E. Learned-Miller, and D.A. Forsyth. Names and faces in the news. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, volume 2, pages II–II, 2004.
- [3] Qiong Cao, Li Shen, Weidi Xie, Omkar M. Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *FG*, pages 67–74. IEEE Computer Society, 2018.
- [4] Chi Chen, Maosong Sun, and Yang Liu. Mask-align: Self-supervised neural word alignment. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4781–4791, Online, Aug. 2021. Association for Computational Linguistics.
- [5] Zhineng Chen, Bailan Feng, Chong-Wah Ngo, Caiyan Jia, and Xiangsheng Huang. Improving automatic name-face association using celebrity images on the web. In Alexander G. Hauptmann, Chong-Wah Ngo, Xiangyang Xue, Yu-Gang Jiang, Cees Snoek, and Nuno Vasconcelos, editors, *ICMR*, pages 623–626. ACM, 2015.
- [6] Zhineng Chen, Wei Zhang, Bin Deng, Hongtao Xie, and Xiaoyan Gu. Name-face association with web facial image supervision. *Multim. Syst.*, 25(1):1–20, 2019.
- [7] Yuqing Cui, Apoorv Khandelwal, Yoav Artzi, Noah Snaveley, and Hadar Averbuch-Elor. Who's waldo? linking people across text and images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1374–1384, October 2021.
- [8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.
- [9] Hao Fang, Saurabh Gupta, Forrest N. Iandola, Rupesh Kumar Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C. Platt, C. Lawrence Zitnick, and Geoffrey Zweig. From captions to visual concepts and back. In *CVPR*, pages 1473–1482. IEEE Computer Society, 2015.
- [10] Ali Farhadi, Seyyed Mohammad Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David A. Forsyth. Every picture tells a story: Generating sentences from images. In Kostas Daniilidis, Petros Maragos, and Nikos Paragios, editors, *ECCV (4)*, volume 6314 of *Lecture Notes in Computer Science*, pages 15–29. Springer, 2010.
- [11] Matthieu Guillaumin, Thomas Mensink, Jakob J. Verbeek, and Cordelia Schmid. Automatic face naming with caption-based supervision. In *CVPR*. IEEE Computer Society, 2008.
- [12] Jack Hessel, Lillian Lee, and David Mimno. Unsupervised discovery of multimodal links in multi-image, multi-sentence documents. 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 2034–2045, Hong Kong, China, Nov. 2019. Association for Computational Linguistics.
- [13] Andrej Karpathy, Armand Joulin, and Fei-Fei Li. Deep fragment embeddings for bidirectional image sentence mapping. In Zoubin Ghahramani, Max Welling, Corinna Cortes, Neil D. Lawrence, and Kilian Q. Weinberger, editors, *NIPS*, pages 1889–1897, 2014.
- [14] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *ICLR (Poster)*, 2015.
- [15] Junnan Li, Pan Zhou, Caiming Xiong, and Steven Hoi. Prototypical contrastive learning of unsupervised representations. In *International Conference on Learning Representations*, 2021.
- [16] Percy Liang, Ben Taskar, and Dan Klein. Alignment by agreement. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 104–111, New York City, USA, June 2006. Association for Computational Linguistics.
- [17] Phi The Pham, Marie-Francine Moens, and Tinne Tuytelaars. Cross-media alignment of names and faces. *IEEE Transactions on Multimedia*, 12(1):13–27, 2010.
- [18] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.
- [19] Xueping Su, Jinye Peng, Xiaoyi Feng, and Jun Wu. Labeling faces with names based on the name semantic network. *Multimedia Tools Appl.*, 75(11):6445–6462, 2016.
- [20] Xueping Su, Jinye Peng, Xiaoyi Feng, Jun Wu, Jianping Fan, and Li Cui. Cross-modality based celebrity face naming for news image collections. *Multimedia Tools Appl.*, 73(3):1643–1661, 2014.
- [21] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence, AAAI '17*, pages 4278–4284. AAAI Press, 2017.
- [22] Yong Tian, Lian Zhou, Yuejie Zhang, Tao Zhang, and Weiguo Fan. Deep cross-modal face naming for people news retrieval. *IEEE Transactions on Knowledge and Data Engineering*, 33(5):1891–1905, 2021.
- [23] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *CVPR*, pages 3156–3164. IEEE Computer Society, 2015.
- [24] Paul Viola and Michael Jones. Robust real-time object detection. *International Journal of Computer Vision*, 57(2):137–154, 2004.

- [25] Qinxin Wang, Hao Tan, Sheng Shen, Michael W. Mahoney, and Zhewei Yao. Maf: Multimodal alignment framework for weakly-supervised phrase grounding. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *EMNLP (1)*, pages 2030–2038. Association for Computational Linguistics, 2020.
- [26] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In Francis R. Bach and David M. Blei, editors, *ICML*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 2048–2057. JMLR.org, 2015.
- [27] Yuejie Zhang, Wei Wu, Yang Li, Cheng Jin, Xiangyang Xue, and Jianping Fan. Automatic name-face alignment to enable cross-media news retrieval. In Francesca Rossi, editor, *IJCAI*, pages 2768–2775. IJCAI/AAAI, 2013.
- [28] Yujie Zhong, Relja Arandjelovic, and Andrew Zisserman. Compact deep aggregation for set retrieval. In Laura Leal-Taixé and Stefan Roth, editors, *ECCV Workshops (4)*, volume 11132 of *Lecture Notes in Computer Science*, pages 413–430. Springer, 2018.