Unstructured Object Matching using Co-Salient Region Segmentation

Ioana-Sabina Stoian Ionut-Catalin Sandu Daniel Voinea Alin-Ionut Popa
{ioanasas, sanion, dvoinea, popaaln}@amazon.com
Amazon, Romania

Abstract

Unstructured object matching is a less-explored and very challenging topic in the scientific literature. This includes matching scenarios where the context, appearance and the geometrical integrity of the objects to be matched changes drastically from one image to another (e.g. a pair of pyjamas which in one image is folded and in the other is worn by a person), making it impossible to determine a transformation which aligns the matched regions. Traditional approaches like keypoint-based feature matching perform poorly on this use case due to the high complexity in terms of viewpoint, scene context variety, background variations or high degrees of freedom concerning structural configurations. In this paper we propose a deep learning framework consisting of a twins based matching approach leveraging a co-salient region segmentation task and a cosine-similarity based region descriptor pairing technique. The importance of our proposed framework is demonstrated on a novel use case consisting of image pairs with various objects used by children. Additionally, we evaluate on Human3.6M and Market-1501, two datasets with humans depicting various appearances and kinematic configurations captured under different backgrounds.

1. Introduction

Determining correspondences between pairs of images is a challenging and intensely explored task in the computer vision community. Its impact is directly visible in subdomains such as optical flow [20], camera calibration [11], stereo reconstruction [35], structure from motion [29] and even semantic region correspondence [7]. There are a multitude of factors which contribute to the difficulty of such a task. One is the phenomenon of *scene-shift* where we have the same scene, however with totally different viewpoint, illumination, background objects, thus creating context confusion and ambiguity across both images. Another key factor which is difficult to overcome is caused by the structure of the searched object as it might change drastically across the viewpoints. Traditionally, the most commonly used approach is to search for a set of 2D keypoint correspondences across the queried image set which are aligned via an inferred transformation matrix [4, 5, 26].

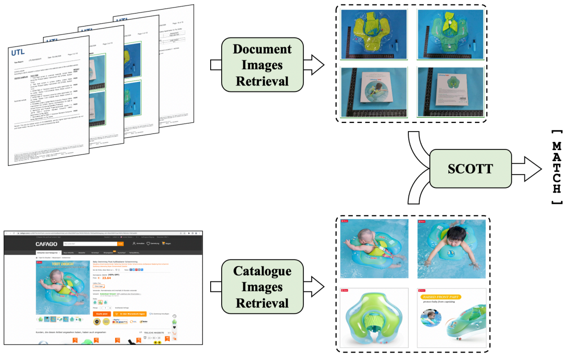

Figure 1. **Sample use-case for SCOTT framework.** Given a catalogue product with an available image set describing the product and a compliance document describing the showcased product using text and visual content, our proposed **SCOTT** framework predicts whether there exists at least one match in the image pairs set determined by the cartesian product between the catalogue image set and the document image set. In this particular case, we have an inflatable pillow which in the catalogue set is illustrated as inflated with different backgrounds and in the compliance document is deflated and wrinkled.

One of the major limitations of such approaches is when the matching object does not possess a structured geometrical configuration to allow for such a heuristic to be applied. Such examples are displayed in figure 2. This use-case is often met in compliance check situations or image content search applications as it can be seen in figure 1. Big stores ensure that prior to making a product available on a specific marketplace, it passes all the necessary compliance checks, thus ensuring whether the object depicted in the attached compliance document is the same as the object showcased in the catalogue.

In this paper we address the unstructured object matching topic with two major contributions:

- a framework for unstructured object matching, entitled

SCOTT (*i.e.* un**S**tru**C**tured **O**bjec**T** ma**T**ching) based on segmentation of the potential common regions of interest followed by descriptor pairing using cosine-similarity of the retrieved regions,

- an unstructured object matching problem with an evaluation dataset, entitled **TIC** (*i.e.* **T**oys **I**n **C**atalogue), with matching and non-matching pairs of objects used by children (*e.g.* toys, sleepwear, accessories).

The motivation behind our proposed pipeline is to provide a matching algorithm which *(i)* can localize the potentially similar regions from a semantic perspective using a co-salient region segmentation pipeline and *(ii)*, determines whether the retrieved regions represent the same object without imposing a transformation flow from one image to another. Thus, our work situates at the boundary between the co-salient object detection literature which retrieves the same class of regions across multiple regions, and the object matching literature which retrieves correspondence flows of 2D keypoints across the queried images for the objects of interest. To the best of our knowledge, we are the first to propose a dataset with an evaluation setup for this particular matching use-case with ground-truth segmentation masks for matching regions.

2. Related Work

In the following we discuss the most relevant recent approaches with respect to our proposed novel problem and carefully position our work regarding to their claims.

2.1. Keypoint Based Matching.

A consistent body of literature related to image similarity [5, 9, 13, 16, 17, 26, 27, 33, 37] is written around the idea of keypoint correspondences for image matching as they are robust to the major challenges involving object appearance. However, the major limitation of these approaches is that they require a rigid geometrical structure of the searched objects across the queried images. Usually, these methods are suitable as proxy tasks for structure from motion or triangulation-based approaches required for 3D based understanding of indoor and outdoor scenes. One relevant work with respect to our proposed approach is [23]. The authors perform a k-nn based search in the keypoint matching space to retrieve the most similar images. The matching is performed via a weakly supervised attention-based keypoint matching. These approaches are unsuitable for our use-case as we are dealing with unstructured objects, with undefined geometrical correspondence from one image to another.

2.2. Contrastive Twins Representation.

Another relevant research thread [3, 6, 32] with respect to our task is via image-level embedding comparisons. One such approach is the work of [22]. The authors leverage a twins architecture to obtain image-level embeddings which are mapped using a contrastive loss function [10]. The main advantage of such approaches is that via the contrastive learning the image embedding space is constrained to act as a densely represented clustering of the targeted classes. In [8] the authors propose a similar full image embedding; however, the embedding space representation is constrained via triplet-loss. Also, a key aspect is that their method requires the searched object to belong to a specific semantic category. The main difference with respect to our pipeline is that we provide a specific embedding for the salient foreground region of both images independent with regard to the semantic class of the matched object.

2.3. Datasets

The majority of the datasets [19, 28, 30] used for the matching problems are designed around the idea of structure from motion or 3D reconstruction, which assumes that the objects depicted in the pictures have the same geometrical structure. One such example is the PhotoTourism dataset [30] which contains iconic buildings from across the world captured under different photographic scenarios. Another group of datasets [1, 18, 19, 25] is dedicated towards developing the scientific topic of co-salient region detection. This translates to the idea of recognizing similar semantic regions across a set of images. In [18], the authors proposed a dataset with image sets grouped according to semantic category of the foreground object depicted in them. Additionally, for each category of images, they provide a foreground segmentation masks for the objects of interest.

We propose a dataset for our targeted problem consisting of matching and non-matching image pairs with various objects used by children. The depicted matching objects from our dataset do not have a geometric structure across the matching images to allow for an alignment transformation across both structures and they do not necessarily belong to the same semantic class (*e.g.* clothes, toys, paintings). This is different than the previous matching datasets as we are targeting class agnostic matching scenarios of objects which do not necessarily possess a geometric structure. Our work is placed at the boundary between these two domains, object correspondence and semantically meaningful regions retrieval. We aim at retrieving the semantically similar objects from a pair of images and decide whether the identity of the highlighted objects from both images matches or not. Thus, this topic created the necessity of introducing such a dataset.

2.4. Co-saliency Detection.

Co-salient region detection aims at identifying the common image regions across the set of analysed images. Usually, these approaches [7, 14, 31, 34] are widely used in the context of unsupervised video based segmentation, by focusing on the consistent temporal foreground regions. One such example is the work of [21]. They present a twins based architecture with a cross-attention based mechanism for highlighting the most relevant image regions across the analysed frames. In [2] the authors propose a self-distillation transformer architecture to learn the salient object of an image without the use of any form of supervision. The work of [14] is similar to our approach. The authors propose a common-foreground segmentation pipeline followed by a contrastive loss approach for the foreground / background region embeddings. Basically, they constrain the model to have a clustered embedded space representation of the foreground regions and to make the background embeddings as dissimilar as possible. However, they only retrieve similar semantic regions, without making any inference regarding the identity of the retrieved objects. For example, they claim to segment image regions corresponding to horses, however, they cannot infer whether the same horse is depicted in an image set.

Our approach is different as we build a common image pair embedding by paying attention at both foreground regions obtained using a co-segmentation task of salient regions. This recovers the regions with the same semantic meaning, such as the general class of objects. Next, we apply a binary classifier on top of a region descriptor pairing heuristic leveraged by cosine-similarity to decide if the highlighted semantic regions are actually the same objects or not.

3. TIC Dataset

In order to provide results for our proposed use case, we collected a dataset with pairs of images depicting matching and non-matching objects. We downloaded publicly available images of catalogue products, in particular children activity related products (*e.g.* fashion items such as pyjamas, hoodies or toys such as Lego sets, plush objects, painting sets, toys, R/C cars, dolls). This use-case is frequently found in compliance related automatic inspections. One such example is when analysing test reports proving the safety of one product with respect to the targeted age category or the legislation compatibility regarding a certain target market (see figure 1). In such a scenario, the first prerequisite condition is to validate whether the pictures within the analysed test report depict the same product as the one presented in the catalogue image set. Another use case is when needing to decide whether a product is depicted in two different catalogues.

In total, we collected 16, 313 images of children related products. These images correspond to a total of 2, 653 unique products. A brief statistic of the distribution of images with respect to the unique products they are describing is illustrated in table 1. Objects designed for children are very diverse in terms of structure, scene context and appearance when advertised in catalogues or online due to marketing reasons, thus they make an excellent candidate for our use-case.

Statistic	**Min**	**Max**	**Median**	**Mean ± Std**
Value	2	28	6	6.19 ± 8.9

Table 1. **Distribution of images in TIC per unique product item.** For each unique product item, we were able to retrieve on average 6.19 images. However, the pool of retrieved images varies quite significantly from 2 images up to 28 images per product. Some of the images depict the same product, however, they might highlight product brand description, or similar products from the same category.

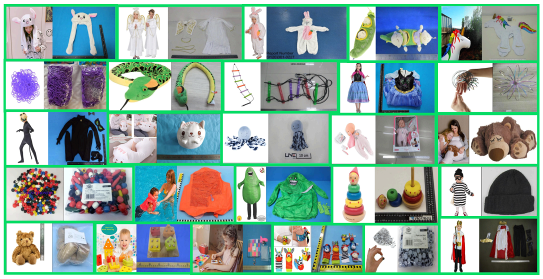

Figure 2. **Sample of matching pairs from TIC dataset.** Each pair is highlighted by a green box. The dataset contains mainly photos of objects designed for children such as toy sets, plush items or sleepwear. Please notice the structural difference between the objects in the illustrated matching pairs as well as the background variety of their corresponding images.

3.1. Annotation Protocol

We created an annotation plugin interface where for each pair of images corresponding to a certain product, we annotated them as matching or not matching class labels. This step was required due to the fact that not all the images are depictions of the same object as some may refer to specification tables or logos of the company producing the object. Additionally, the annotators were asked to draw a segmentation mask on each image for the matching case.

As a result of the labelling process, we obtained a total of 13, 172 pairs of matching and non-matching elements. Out of these, there are 6, 586 (*i.e.* 50%) matching pairs and 6, 586 (*i.e.* 50%) non-matching pairs (check figure 2 for matching pairs). There is also a number of 1, 753 pairs which were ambiguous from a matching perspective and we did not include those in the final dataset. These are image pairs where the depicted objects belong to the same semantic category, however, it is difficult even for humans to decide whether it is the same object or not. One such example is a Lego product assembled in totally different configura-

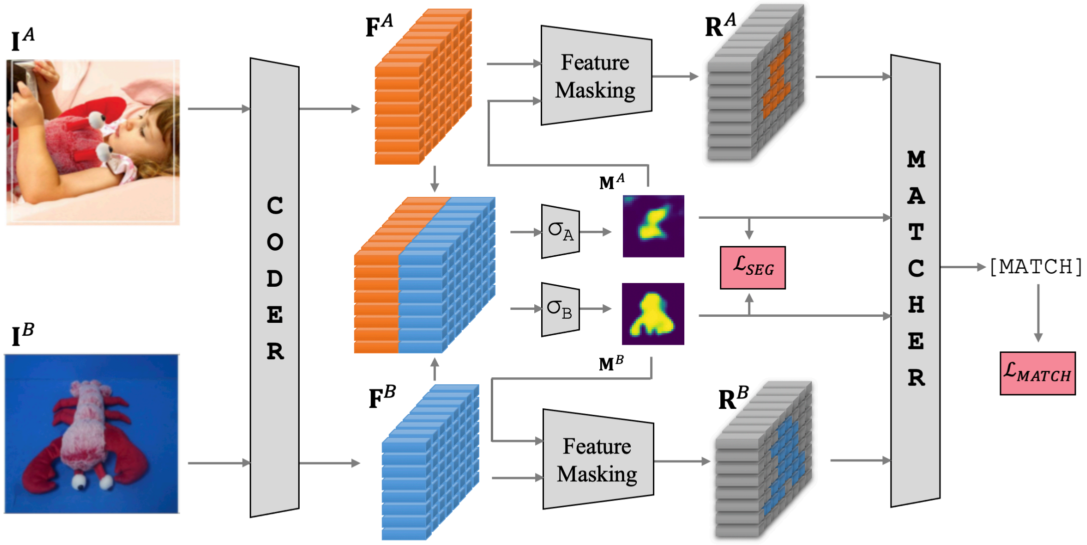

Figure 3. **Detailed overview of our proposed unstructured object matching framework, SCOTT.** Given a pair of images, $\mathbf{I}^A$ and $\mathbf{I}^B$, firstly, we obtain a pair of full image embeddings, $\mathbf{F}^A$ and $\mathbf{F}^B$, by leveraging the **CODER** backbone. The potential co-salient image regions from both images, $\mathbf{R}^A$ and $\mathbf{R}^B$, are filtered using the segmentation masks $\mathbf{M}^A$ and $\mathbf{M}^B$, respectively. Lastly, the co-salient regions are provided as input to the **MATCHER** head to predict the final classification score. The entire pipeline is trained using a multi-task loss composed of a segmentation loss, $\mathcal{L}_{\text{SEG}}$ over the σ_A and σ_B heads, as well as a binary cross-entropy loss, $\mathcal{L}_{\text{CE}}$ on top of **MATCHER**.

tions, which might represent the same brick set.

4. Unstructured Object Matching

We now introduce our proposed framework for unstructured object matching across an image pair, **SCOTT**. We describe in detail the components of the model from a methodological perspective.

Given a pair of images $(\mathbf{I}^A, \mathbf{I}^B)$, where $\mathbf{I}^A, \mathbf{I}^B \in \mathbb{R}^{w_0 \times h_0 \times 3}$, we want to assign a similarity score $\tilde{S} \in [0,1]$ corresponding to the case of matching or non-matching objects, respectively. Thus, for each pair of images $(\mathbf{I}^A, \mathbf{I}^B)$ we have an attached pair of ground truth figure-ground masks, $\mathbf{M}^A, \mathbf{M}^B \in \mathbb{R}^{w \times h}$ and target class label $S \in \{0,1\}$ corresponding to `[NON-MATCH]` or `[MATCH]`, respectively. The first step is to obtain a pair of feature maps describing the most relevant information from both inputted images, $\mathbf{I}^A$ and $\mathbf{I}^B$. At the same time, we are interested in building the pair of maps in a co-dependent manner. For this purpose, we build a **CO**dependent enco**DER** (*i.e.* **CODER**) inspired from the architecture of U-net [24]. It is composed of `CONV`, `RELU` and `MAX-POOL` blocks which compress the spatial image information in a depth-wise manner. We denote each such block with ψ_i, where i represents the iteration index. Different from [24], we considered concatenating the signal from both images prior to passing it to the next encoding block.

$$\mathbf{F}_i^A = \begin{cases} \psi_i^A(\mathbf{I}^A), & \text{if } i = 1 \\ \psi_i^A([\mathbf{F}_{i-1}^A, \mathbf{F}_{i-1}^B]), & \text{otherwise} \end{cases}$$

After $i = 4$ iterations of information down sampling, we end up with $\mathbf{F}^A = \mathbf{F}_4^A$, where $\mathbf{F}^A \in \mathbb{R}^{w \times h \times d}$ with $w = \frac{w_0}{32}$ and $h = \frac{h_0}{32}$. In practice, the best results where obtained with $d = 512$. The operation is similar with respect to $\mathbf{I}_B$ and ψ^B obtaining $\mathbf{F}_B \in \mathbb{R}^{w \times h \times d}$. Basically, $\mathbf{F}^A$ and $\mathbf{F}^B$ contain the encoded information of $\mathbf{I}^A$ and $\mathbf{I}^B$ in a correlated manner. This is a mixed signal strategy, which implicitly constrains the feature maps to be consistent. A visualization of **CODER** and how it operates is illustrated in figure 4.

4.1. Co-salient Image Region Segmentation

Once $\mathbf{F}^A$ and $\mathbf{F}^B$ are computed, we are interested in determining the co-salient regions from both images. This translates to recovering the potentially common objects. For this purpose, we constructed two segmentation heads, σ_A and σ_B, operating on the concatenated feature map information, $\mathbf{F}^A \| \mathbf{F}^B$, where $\|$ is the concatenation operator defined as $\| : (\mathbb{R}^{w \times h \times d}, \mathbb{R}^{w \times h \times d}) \rightarrow \mathbb{R}^{w \times h \times 2d}$. The segmentation heads are inspired from [12]. The intuition behind using the concatenated information is to learn an implicit correlation between the co-salient regions of both images by spatially overlapping the common information. The heads σ_A and σ_B output two segmentation masks $\tilde{\mathbf{M}}^A \in \mathbb{R}^{w \times h}$

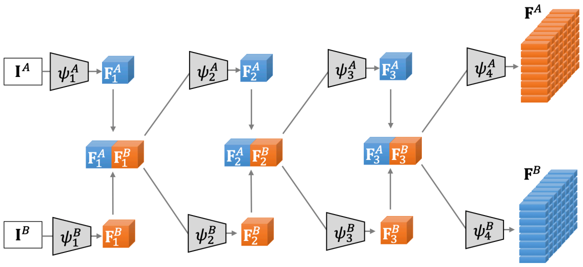

Figure 4. **CODER.** At each encoding step i, the feature representations, $\mathbf{F}_i^A$ and $\mathbf{F}_i^B$, from each image are merged and passed to the next processing encoding modules, ψ_{i+1}^A and ψ_{i+1}^B. The idea is to get a joint processing backbone which highlights the common descriptors from both images.

and $\tilde{\mathbf{M}}^B \in \mathbb{R}^{w \times h}$ corresponding to $\mathbf{I}^A$ and $\mathbf{I}^B$, respectively. They contain foreground and background segmentation masks of the co-salient objects from both images. The foreground regions provided by $\mathbf{M}^A$ and $\mathbf{M}^B$ is used to mask the irrelevant (background) features from $\mathbf{F}^A$ and $\mathbf{F}^B$, respectively.

$$\mathbf{R}^A = \{\mathbf{F}_{ij}^A \in \mathbb{R}^d \mid \mathbf{M}_{ij}^A > \alpha, i = \overline{1..w}, j = \overline{1..h}\}$$
$$\mathbf{R}^B = \{\mathbf{F}_{ij}^B \in \mathbb{R}^d \mid \mathbf{M}_{ij}^B > \alpha, i = \overline{1..w}, j = \overline{1..h}\}$$

In practice, parameter α is validated. Intuitively, $\mathbf{R}^A \in \mathbb{R}^{P \times d}$ and $\mathbf{R}^B \in \mathbb{R}^{Q \times d}$ contain the foreground feature information extracted from $\mathbf{F}^A$ and $\mathbf{F}^B$, respectively.

4.2. Saliency Guided Co-attention Based Matching

In this subsection we define the final component inside our pipeline, the **MATCHER**. Its role is to make the final inference over the highlighted co-salient feature regions with respect to the unstructured object matching task. A detailed view of it can be visualised in figure 5. Having the co-salient descriptors $\mathbf{R}^A$ and $\mathbf{R}^B$, the only remaining thing is to decide whether they represent the same object or not. Inspired from [21], we build a cosine similarity matrix, $\mathbf{S} \in \mathbb{R}^{P \times Q}$, between the two foreground synthesized descriptors. We used the cosine similarity as opposed to the L1 or L2 distance, as it is more robust with respect to the magnitude of the involved feature vectors. Also, the cosine similarity measurement operates better in a high dimensional feature space, such as our architectural design where $d = 512$.

By leveraging the similarity information from matrix $\mathbf{S}$ we are able to pair the descriptors $\mathbf{R}^A$ and $\mathbf{R}^B$ as follows,

$$\begin{aligned}\mathbf{R}^{A \cup B} =& \mathbf{S} \odot (\mathbf{R}^A \times \mathbf{R}^B) \\ =& \{\mathbf{S}_{ij} \cdot [\mathbf{R}_i^A, \mathbf{R}_j^B] \mid i = \overline{1..P}, j = \overline{1..Q}\}\end{aligned}$$

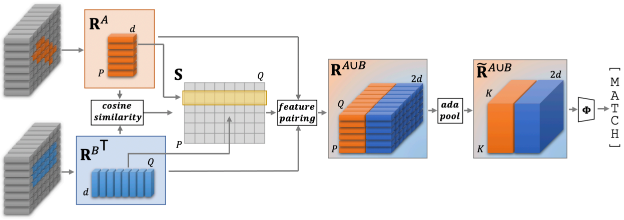

Figure 5. **MATCHER.** The segmented feature regions $\mathbf{R}^A$ and $\mathbf{R}^B$ are compared using a cosine-similarity metric to produce a similarity matrix $\mathbf{S}$. This is used to generate a weighted tensor, $\mathbf{R}^{A \cup B} \in \mathbb{R}^{P \times Q \times 2d}$, with all the feature combinations between $\mathbf{R}^A$ and $\mathbf{R}^B$. Given that P and Q are different for every pair, we apply `[adaptive-pool]` to bring its dimensionality to $\mathbb{R}^{K \times K \times 2d}$. Lastly, a CNN head, Φ is applied to obtain the final classification score.

where $\mathbf{R}^{A \cup B} \in \mathbb{R}^{P \times Q \times 2d}$, $\times$ represents the cartesian product for two sets and $\odot$ represents the Hadamard product. The necessity of $\mathbf{S}$ is to emphasize the similar pairs and ignore dissimilar ones.

In the current format, the information encoded in $\mathbf{R}^{A \cup B}$ is different for every pair of images $\mathbf{I}_A$ and $\mathbf{I}_B$ as the segmented regions can be very different. To standardize its dimensionality for learning purposes, we apply an adaptive max pooling layer to bring the dimensionality of $\mathbf{R}^{A \cup B}$ from $\mathbb{R}^{P \times Q \times 2d}$ to $\mathbb{R}^{K \times K \times 2d}$. In our experimental setup, the best results were obtained with $K = 64$. Lastly, we apply a classification head, Φ, over the pooled set of features, $\tilde{\mathbf{R}}^{A \cup B}$, to provide us with the matching score between images $\mathbf{I}_A$ and $\mathbf{I}_B$.

4.3. Training of SCOTT.

The entire **SCOTT** ensemble is trained by propagating gradients through both **CODER** and **MATCHER** modules. The model is penalized by a binary cross-entropy loss, $\mathcal{L}_{\text{CE}}$ on the predicted match score, $\tilde{S}$, and a cross-entropy image segmentation loss, $\mathcal{L}_{\text{SEG}}$ applied over the predicted segmentation masks, $\tilde{\mathbf{M}}^A$ and $\tilde{\mathbf{M}}^B$. The training procedure is applied in 2 steps. Firstly, the **CODER** and the segmentation heads σ_A and σ_B, respectively, are trained using the matching pairs only. Secondly, the entire pipeline is trained on all pairs using the steps mentioned in algorithm 1, without backpropagating gradients through σ_A and σ_B for non-matching pairs.

$$\mathcal{L} = \mathcal{L}_{\text{CE}}(S, \tilde{S}) + \mathcal{L}_{\text{SEG}}(\mathbf{M}^A, \tilde{\mathbf{M}}^A) + \mathcal{L}_{\text{SEG}}(\mathbf{M}^B, \tilde{\mathbf{M}}^B)$$

5. Experiments

Experiments are performed on our proposed dataset TIC, Human3.6M [15] and Market-1501 [36]. In the following

Algorithm 1: MATCHER

```
Input: $\mathbf{R}^A \in \mathbb{R}^{P\times d}$ and $\mathbf{R}^B \in \mathbb{R}^{Q\times d}$
Output: $\tilde{S} \in \{0,1\}$
$\mathbf{S} \leftarrow \texttt{cosine-similarity}(\mathbf{R}^A, \mathbf{R}^B)$
$\mathbf{R}^{A\cup B} \leftarrow \{\mathbf{S}_{ij} \cdot [\mathbf{R}_i^A, \mathbf{R}_j^B] \mid i = \overline{1..P}, j = \overline{1..Q}\}$
$\tilde{\mathbf{R}}^{A\cup B} \leftarrow \texttt{adaptive-pool}(\mathbf{R}^{A\cup B})$
$\tilde{S} \leftarrow \Phi(\tilde{\mathbf{R}}^{A\cup B})$
```

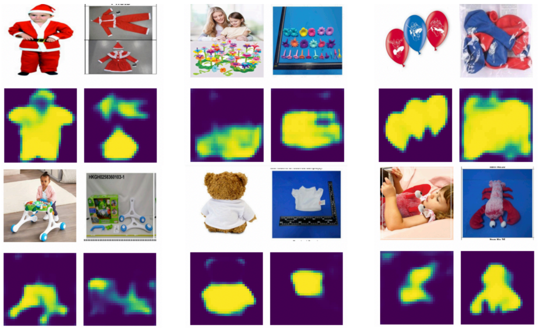

Figure 6. **Sample matches and segmentations from TIC test set.** Notice the appearance variation and the differences in terms of context and geometrical composition of the searched objects.

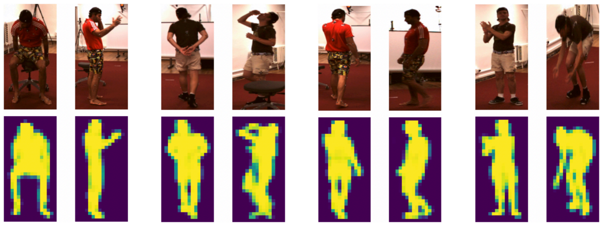

Figure 7. **Sample matches and segmentations of humans from H3.6M.** Notice that the posture configuration is very different in all the pairs, and for some of them the human is facing front and back in the matching pair.

we describe the experimental setup used for each dataset and the most important results obtained with respect to our proposed task.

5.1. TIC Dataset.

In the case of our own proposed dataset, TIC, we considered the following data split: $3,293$ pairs ($1,647$ matching) used for *test*, 658 pairs (329 matching) used for *validation*, and $9,221$ pairs ($4,610$ matching) used for *train*.

To show the performance of our method, we tested against 2 additional baselines, ORB feature matching [26] and SuperGlue deep matching algorithm [27]. Both of the methods are focused around retrieving image keypoints pairs that are similar across the corresponding image pair they are related to. For each external method, we validated the hyperparameters such as minimum number of keypoint correspondences required for an object match or the score threshold for matching keypoint descriptors. Comparison results are available in table 2. Note that we are illustrating a novel matching use-case and we compensate the lack of external method evaluation with different ablation studies within our framework, thus emphasizing the complexity of the proposed task. We report the precision, recall and F1-score for the image pairs corresponding to the matching class.

Method		Precision	Recall	F1-Score
ORB matching [26]		0.52	0.95	0.67
SuperGlue [27]		0.66	0.71	0.68
SCOTT w image embd.	CODER	**0.83**	0.32	0.46
	ResNet50	0.51	**0.99**	0.67
SCOTT w co-salient reg. embd.	CODER	0.75	0.84	0.79
	ResNet50	0.75	0.87	0.80
SCOTT w MATCHER	**CODER**	0.79	0.83	**0.81**
	ResNet50	0.81	0.80	0.80

Table 2. **TIC Test Set Results.** We report the precision, recall and F1-score for the matching class. Our method is compared against two relevant baselines, ORB feature matching [26] and SuperGlue [27]. Additionally, we illustrate the performance using *ResNet50* instead of **CODER** as well as 2 alternatives for feature embedding integration: **SCOTT** with image embedding which uses the entire $\mathbf{F}^A$ and $\mathbf{F}^B$ descriptors via a standard pooling operation and **SCOTT** with segmented embedding which uses standard pooling operation over the $\mathbf{R}^A$ and $\mathbf{R}^B$ region descriptors. The highest score is obtained with **SCOTT** framework using the proposed **CODER** backbone and **MATCHER** head.

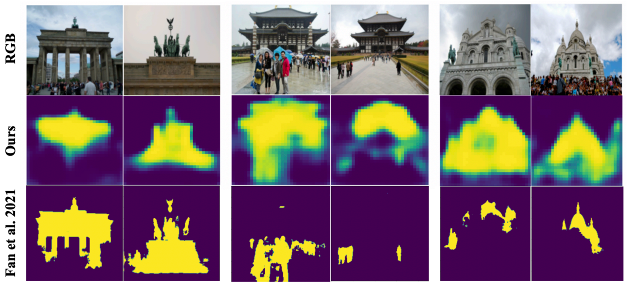

Figure 8. **Sample co-salient image region segmentation on PhotoTourism dataset.** *(First row)* RGB content, *(second row)* segmentation obtained with our method and *(third row)* segmentation obtained using [7]. Notice that the images distribution is different with respect to the ones used for training the algorithms. We aim to illustrate the generalization property of the compared methods on novel data distributions. Our model is able to pick up similar images on outdoor scenes, with building objects, a totally different context, object and appearance distribution from **TIC**.

In table 2 we perform several ablation studies with different components of our method to highlight the impact of

Method	IoU Image A	IoU Image B	IoU Both Images
ResNet50	0.45	0.37	0.41
CODER	**0.52**	**0.55**	**0.53**
GCoNet [7]	0.33	0.31	0.32

Table 3. **Segmentation Evaluation on PhotoTourism Dataset.** We report the performance on a subset of PhotoTourism using **CODER** backbone, *ResNet50* twin architecture backbone and [7]. Neither method was trained on PhotoTourism as we want to illustrate the robustness of the co-saliency segmentation method on novel, unseen data distribution.

Similarity Metric	Precision	Recall	F1-Score
Cosine	**0.79**	**0.83**	**0.81**
L1 distance	0.74	0.82	0.77
L2 distance	0.76	0.81	0.79

Table 4. **Ablation study on TIC with different similarity metrics required for building matrix S.** We illustrate the importance of the cosine-similarity metric used for weighting the feature pairing matrix $\mathbf{R}^{A \cup B}$. The cosine similarity leads to improved performance metrics when compared to the L1 and L2 distance measures, respectively.

both **CODER** backbone and **MATCHER** head. To prove the importance of **CODER**, we experimented by plugging in a twins *ResNet50* backbone inside **SCOTT**. Also, we use 3 different embedding merging strategies to demonstrate the importance of **MATCHER**: *(a)* pooling the entire image embedding from each $\mathbf{F}^A$ and $\mathbf{F}^B$, thus incorporating in a single descriptor the entire image information (*i.e*. both foreground and background regions), *(b)* pooling the foreground regions encoded in $\mathbf{R}^A$ and $\mathbf{R}^B$, thus using a descriptor which synthesizes the foreground information of each image individually and *(c)* the **MATCHER** foreground information merging strategy of both images. The best performance is obtained with **MATCHER** (see table 2 line 9) as it optimally combines the embeddings of the foreground regions of $\mathbf{I}_A$ and $\mathbf{I}_B$ via similarity matrix $\mathbf{S}$. The poor performance obtained using the entire image embedding is caused by the ambiguity induced by the background regions which can be similar or radically different for matching cases. Table 4 provides a comparative overview of the performance associated with the three metrics for the matrix $\mathbf{S}$. For this experiment, we used the **SCOTT** framework with **CODER** backbone and **MATCHER** classification head. The best performance is obtained using the cosine-similarity as it treats the region descriptors as vectors and computes their alignment, being robust to their magnitude and dimensionality.

We illustrate in figure 8 the visual performance of the segmentation module, σ_A and σ_B, on PhotoTourism [30] using our proposed **SCOTT** framework trained solely on TIC. In table 3 we quantitatively compare our work against [7]. We labelled $1,000$ random pairs with different building landmarks to evaluate the performance of the segmentation method. Neither method was trained on this dataset as we want to illustrate the generalization property for the task of co-salient region detection. Our method is able to pick up the common potential objects given the context, illumination, foreground or viewpoint changes. With this experiment we want to emphasize that with a simple feature map concatenation approach, we can highlight the common semantically meaningful objects, without the usage of advanced feature merging heuristics. In figure 9, left, we illustrate the F1-score distribution across TIC test set with respect to the image ratios of the segmented regions, $\mathbf{R}^A$ and $\mathbf{R}^B$. The highest scores occur when the segmented regions are sufficiently large in both images (*i.e.* $\approx 30-60\%$). In the right side of the figure, we illustrate a spatial distribution of segmented regions ratios. The majority are relatively small with respect to the image size, $\approx 10-40\%$. In figure 6 we illustrate sample matching and segmentation results of our method on TIC test set proving robustness to structure, semantics of the matched objects, background context as well as appearance. Also, in figure 10 we illustrate situations where our proposed **SCOTT** is unable to correctly match or segment the correct objects. Usually, these situations occur when there are multiple matching objects in the same image pair, as is the case of the middle pair of images or when there are objects with the same semantics, and very similar appearance representation.

Method	Precision	Recall	F1-Score
Human3.6M			
SuperGlue [27]	0.49	**0.96**	0.65
SCOTT w image embedding	0.55	0.88	0.67
SCOTT w co-salient region embedding	0.68	0.93	0.78
SCOTT w MATCHER	**0.76**	0.86	**0.8**
Market-1501			
SCOTT trained on Human3.6M	0.6	0.66	0.63
SCOTT fine-tuned on Market-1501	**0.65**	**0.97**	**0.78**

Table 5. **H3.6M and Market-1501 Results.** In this setup, we used only **CODER** backbone and varied the classification head used. The best performance on H3.6M is obtained using **SCOTT** with **MATCHER**, with an improvement of 0.02 F1-score over co-salient region embedding. For Market-1501 we compared the best performing model trained on H3.6M against the same model pretrained on H3.6M and fine-tuned on Market-1501 using binary cross-entropy loss.

5.2. Human3.6M Dataset.

To illustrate the versatility of our framework, we test against the task of human matching on the Human3.6m [15] dataset. The RGB data was collected using 4 different RGB cameras. The actions are performed by 10 different human subjects. The viewpoint and action variety made this

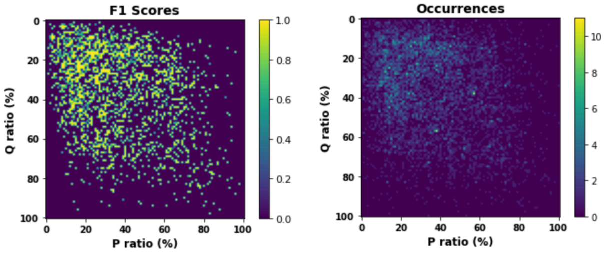

Figure 9. **Visual correlation between F1-scores and image ratios of the segmented pairs over TIC test set.** *(Left)* We illustrate the F1-score distribution across the P, Q image ratios. *(Right)* Spatial histogram of segmented regions (*i.e.* $\mathbf{R}^A$ and $\mathbf{R}^B$) ratios with respect to size of images (*i.e.* $\mathbf{I}^A$ and $\mathbf{I}^B$).

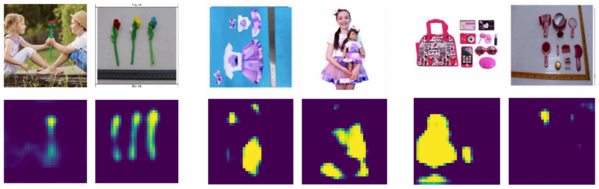

Figure 10. **Failure cases of SCOTT on TIC dataset.** Our algorithm has difficulty in retrieving the correct match when there are multiple objects under the same semantic category present in the image which have a highly similar appearance. Such is the case of the leftmost example, where in the left image we have two kids holding a red toy tulip and in the right side image, we have 3 identical toy tulips, however, of different colors.

dataset perfectly suitable for our proposed framework. The image pair sampling is performed as follows: subjects 1, 5, 6 and 7 for train with 34,889 pairs, subjects 8, 9 and 11 for validation with 2,349 pairs, and subjects 2, 3 and 4 for test with 24,420 pairs. Numerical results with our approach are available in table 5. We illustrate matching results obtained with *(a)* the embedding of the entire image, *(b)* the embedding of the segmented region and *(c)* the proposed **MATCHER** embedding strategy. The best performing method is **MATCHER**, however the improvement is minimal with respect to the matching based on the segmented region embedding. The matching using [27] performs the worst, as it gets confused by the similar background, thus creating many false positives. In figure 7 we illustrate sample matches together with segmentation results. The method is able to cope with front / back views of and different human pose configurations.

5.3. Market-1501

We extend the experiments on human matching on the **Market-1501** [36] dataset. It contains 32,668 images of 1,501 different persons taken from 6 disjoint cameras. For our experimental setup, we used the train set of 750 persons and split it in 3 sets: train with 550 subjects, validation with 50 subjects and test with 150 subjects. We have no available segmentation masks so we use the pretrained segmentation weights of **SCOTT** from Human3.6M. We compared the performance of the **MATCHER** head by using the pretrained weights from Human3.6M and a fine-tuned model on Market-1501 train set. Performance details are available in table 5. We observe an increase of 0.15 f1-score by fine-tuning the **MATCHER** head. This is obtained using the segmentation heads σ_A and σ_B trained on Human3.6M. Visual results on Market-1501 are illustrated in figure 11.

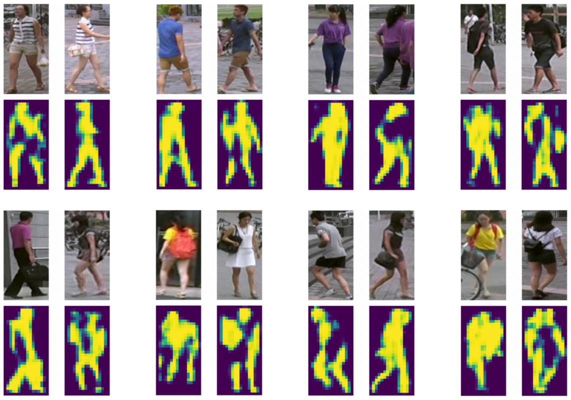

Figure 11. **Sample segmentation and matching results on Market-1501 dataset.** On the first 2 rows we illustrate matching and figure-ground segmentation masks and on the third and fourth rows non-matching pairs. The segmentations are obtained with using **SCOTT** which was trained on Human3.6M only.

6. Conclusions

We presented a novel use-case in the context of object similarity, namely unstructured object matching. We proposed TIC dataset containing pairs of images depicting objects designed for children with 13,172 pairs of matching / non-matching objects. Additionally, we proposed a framework based on co-salient region detection and classification using cosine-similarity descriptor pairing of recovered regions. The tackled problem is hard as we are dealing with scenarios where the searched object does not possess a rigid geometrical structure, thus making it impossible to determine a correspondence flow between the images. We demonstrated the effectiveness of our proposed approach on TIC, H36M and Market-1501, and emphasizing the difficulty of the proposed use case.

Acknowledgements. We thank Javier Romero and Adrian-Mihai Alecu for their valuable assistance, outstanding feedback and insights provided over time.

References

[1] Dhruv Batra, Adarsh Kowdle, Devi Parikh, Jiebo Luo, and Tsuhan Chen. icoseg: Interactive co-segmentation with intelligent scribble guidance. In *CVPR*, 2010. 2

[2] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. *arXiv preprint arXiv:2104.14294*, 2021. 3

[3] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *CVPR*, 2021. 2

[4] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *CVPR Workshops*, 2018. 1

[5] Mihai Dusmanu, Ignacio Rocco, Tomas Pajdla, Marc Pollefeys, Josef Sivic, Akihiko Torii, and Torsten Sattler. D2-net: A trainable cnn for joint description and detection of local features. In *CVPR*, 2019. 1, 2

[6] Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. With a little help from my friends: Nearest-neighbor contrastive learning of visual representations. *arXiv preprint arXiv:2104.14548*, 2021. 2

[7] Qi Fan, Deng-Ping Fan, Huazhu Fu, Chi-Keung Tang, Ling Shao, and Yu-Wing Tai. Group collaborative learning for co-salient object detection. In *CVPR*, 2021. 1, 2, 6, 7

[8] Weifeng Ge. Deep metric learning with hierarchical triplet loss. In *ECCV*, 2018. 2

[9] Hugo Germain, Vincent Lepetit, and Guillaume Bourmaud. Visual correspondence hallucination: Towards geometric reasoning. *arXiv preprint arXiv:2106.09711*, 2021. 2

[10] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *CVPR*, 2006. 2

[11] Richard Hartley and Andrew Zisserman. Multiple view geometry in computer vision. *Robotica*, 19(2):233–236, 2001. 1

[12] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017. 4

[13] Yihui He, Rui Yan, Katerina Fragkiadaki, and Shoou-I Yu. Epipolar transformer for multi-view human pose estimation. In *CVPR Workshops*, 2020. 2

[14] Kuang-Jui Hsu, Yen-Yu Lin, Yung-Yu Chuang, et al. Co-attention cnns for unsupervised object co-segmentation. In *IJCAI*, 2018. 2, 3

[15] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *PAMI*, 2014. 5, 7

[16] Wei Jiang, Eduard Trulls, Jan Hosang, Andrea Tagliasacchi, and Kwang Moo Yi. COTR: Correspondence Transformer for Matching Across Images. In *ICCV*, 2021. 2

[17] Yuhe Jin, Dmytro Mishkin, Anastasiia Mishchuk, Jiri Matas, Pascal Fua, Kwang Moo Yi, and Eduard Trulls. Image matching across wide baselines: From paper to practice. *IJCV*, 2021. 2

[18] Teemu Kinnunen, Joni-Kristian Kamarainen, Lasse Lensu, Jukka Lankinen, and Heikki Käviäinen. Making visual object categorization more challenging: Randomized caltech-101 data set. In *ICPR*, 2010. 2

[19] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *International Workshop - 3dRR)*, 2013. 2

[20] Ce Liu, Jenny Yuen, and Antonio Torralba. Sift flow: Dense correspondence across scenes and its applications. *PAMI*, 33(5):978–994, 2010. 1

[21] Xiankai Lu, Wenguan Wang, Chao Ma, Jianbing Shen, Ling Shao, and Fatih Porikli. See more, know more: Unsupervised video object segmentation with co-attention siamese networks. In *CVPR*, 2019. 3, 5

[22] Iaroslav Melekhov, Juho Kannala, and Esa Rahtu. Siamese network features for image matching. In *ICPR*, 2016. 2

[23] Hyeonwoo Noh, Andre Araujo, Jack Sim, Tobias Weyand, and Bohyung Han. Large-scale image retrieval with attentive deep local features. In *ICCV*, 2017. 2

[24] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 4

[25] Michael Rubinstein, Armand Joulin, Johannes Kopf, and Ce Liu. Unsupervised joint object discovery and segmentation in internet images. In *CVPR*, 2013. 2

[26] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. Orb: An efficient alternative to sift or surf. In *ICCV*, 2011. 1, 2, 6

[27] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *CVPR*, 2020. 2, 6, 7, 8

[28] Torsten Sattler, Will Maddern, Carl Toft, Akihiko Torii, Lars Hammarstrand, Erik Stenborg, Daniel Safari, Masatoshi Okutomi, Marc Pollefeys, Josef Sivic, et al. Benchmarking 6dof outdoor visual localization in changing conditions. In *CVPR*, 2018. 2

[29] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, 2016. 1

[30] Noah Snavely, Steven M Seitz, and Richard Szeliski. Photo tourism: exploring photo collections in 3d. In *ACM Siggraph*. 2006. 2, 7

[31] Guolei Sun, Wenguan Wang, Jifeng Dai, and Luc Van Gool. Mining cross-image semantics for weakly supervised semantic segmentation. In *ECCV*, 2020. 2

[32] Jue Wang, Gedas Bertasius, Du Tran, and Lorenzo Torresani. Long-short temporal contrastive learning of video transformers. *arXiv preprint arXiv:2106.09212*, 2021. 2

[33] Olivia Wiles, Sebastien Ehrhardt, and Andrew Zisserman. Co-attention for conditioned image matching. In *CVPR*, 2021. 2

[34] Cheng-Kun Yang, Yung-Yu Chuang, and Yen-Yu Lin. Unsupervised point cloud object co-segmentation by co-contrastive learning and mutual attention sampling. In *ICCV*, 2021. 2

[35] AA Zakharov and AE Barinov. An algorithm for 3d-object reconstruction from video using stereo correspondences. *PAMI*, 25(1):117–121, 2015. 1

[36] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *ICCV*, 2015. 5, 8

[37] Qunjie Zhou, Torsten Sattler, and Laura Leal-Taixe. Patch2pix: Epipolar-guided pixel-level correspondences. In *CVPR*, 2021. 2