Unsupervised Image Captioning Based on Instance Segmentation for Person Re-Identification

Margarita N. Favorskaya [1], Maxim V. Savkov [1]

[1] Reshetnev Siberian State University of Science and Technology, Institute of Informatics and Telecommunications, 31, Krasnoyarsky Rabochy ave., Krasnoyarsk, 660037 Russian Federation – favorskaya@sibsau.ru, msavkov2017@gmail.com

Commission II, WG II/8

Keywords: Image Captioning, Person Re-Identification, Unsupervised Learning, Deep Learning.

Abstract

The invention of vision-language models such as CLIP (Contrastive Language-Image Pre-training) has had a positive impact on various image classification tasks, including the task of person re-identification (Re-ID), which aims to detect a person of interest using multiple non-overlapping cameras. This allows to consider the person re-identification as a multi-task problem involving both visual and textual descriptors. However, CLIP-based models mainly suffer from coarse-grained alignment issues and use a supervised learning strategy, which is undesirable for real-time person Re-ID task. The multi-modal problem statement based on preliminary instance segmentation on a person helps to achieve fine-grained alignment of visual-text descriptors. We propose an unsupervised image captioning method based on the CutLER detector, where visual features are extracted only from the object of interest without considering background data. The experiments were conducted using human images selected from the MSCOCO dataset. More than 8,000 images were processed. Experimental results with CutLER pre-segmentation showed an improvement in the caption generation accuracy as measured by BLEU1-BLEU4, METEOR, ROUGE, CIDEr, and SPICE metrics.

1. Introduction

The problem of person re-identification (Re-ID) is widely applied in the fields of video surveillance, human behaviour analysis and intelligent security. The methods used depend strongly on short-term or long-term observation, which determines the learning strategy. The short-term observation means that a person record may not be present in the retrieval database. The person of interest may appear in any location at different time instances and be captured by different cameras. In this case, information about a person should be collected and processed "on the fly". The query person can be represented by an image, video sequence or textual description. In this study, we assume that the input information is an image.

We are interested in developing an unsupervised learning strategy that is relevant for person Re-ID in a crowd with possible overlapping of visual projections. This is a very challenging task that requires all possible semantic clues, attribute details and different modalities. Recent advances in deep language models have led to the idea of applying vision-language models to the Re-ID task. Visual language model encodes images to obtain features and then generates textual descriptions that clarify the contextual information in the image. This approach is closely related to the fundamental concept of image captioning. The Re-ID problem in the short-term statement can be considered as a special case of the application of the vision-language models.

First CNN-based Re-ID models such as OSNet (Zhou et al., 2019) were limited by the volume of training data and were prone to overtraining. Since then, architectures have been improved, new modules and blocks have been proposed. The AANet model (Tay et al., 2019) integrated key attributes such as hair, upper clothing colour, lower clothing colour, dress, lower clothing, and gender into an unified learning framework using discriminative image-level attention attribute sub-maps. Consecutive batch dropblock network (Tan et al., 2021) provided an effective attentive feature learning strategy to extract global and local features. In (Rao et al., 2021), a counterfactual attention learning method was proposed to improve attention learning based on causal inference. Recently, transformer-based models such as the end-to-end part-aware transformer (Li et al., 2021) and the disentangled representation learning network (Jia et al., 2023) leverage the strengths of both CNNs and transformers to further improve the performance of the models on occluded visual projections.

Currently, multimodal approaches that combine visual and textual data have become very popular in image retrieval. The CLIP model (Radford et al., 2021) maps images and text into a shared latent space, allowing the model to match features between the two modalities. CLIP (Contrastive Language-Image Pre-training) model demonstrates advances in cross-modal retrieval, classification, and generation. However, it has some limitations related to the extraction of context-specific details and the need for quality and relevance of image-text pairs during training.

Visual human representation is characterized by limited, small-sized visual features with varying colours and predictable relationships. This means that Re-ID visual-textual descriptor is problem-oriented and smaller in size compared to the base CLIP model. Another proposal is to use instance segmentation to obtain a visual projection of a person, which significantly improve recognition accuracy. We tested the proposed method on human images selected from the MSCOCO dataset, demonstrating its effectiveness and generalization capability.

The rest of the paper is as follows. Section 2 provides a brief review of Re-ID image capture methods. Section 3 introduces the proposed method. Experimental results are included in Section 4, and Section 5 concludes the paper.

2. Related Work

Text-image person re-identification aims at extracting fine-grained information from images and texts and comparing their

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLVIII-2-W9-2025-87-2025 | © Author(s) 2025. CC BY 4.0 License.

correspondences. Since 2017 (Li et al., 2017), the text-image models have undergone a difficult development path. Some studies are based on the CLIP model and its modifications, while others use different approaches. Contrastive learning is a core concept underlying the CLIP model (Tian et al., 2020). This model assumes a contrastive relationship between image and text or text and image and selects the image-text pair with the highest similarity. A CLIP-based fine-grained information mining framework called CFine (Yan et al., 2023) transferred multi-modal knowledge to extract intra-modal discriminative clues and inter-modal correspondences shared across modalities. The CLIP-ReID model was proposed in (Li et al., 2023) with a two-stage strategy for training textual tokens. Pure contrastive learning and feature completion contrastive learning were introduced for creating text-image person re-identification (iTIReID) descriptor (Du et al., 2024). The modal fusion approach proposed in (Li et al., 2025) combined the visual features extracted by the ResNet-50 model with the text representations based on the Transformer Decoder so that the text features can dynamically guide the visual features. This model was built upon the CLIP-ReID framework. In (Ren et al., 2025), a mixed semantic clustering expert model was developed using semantic space via the CLIP text encoder.

In (Wang et al., 2024a), a mutual benefit mechanism between human visual features and high-level semantic textual information was proposed for sports event scenes. The visual feature extraction module used a multi-granularity network based on the ResNet-50 model, which effectively integrated global and local information across multiple granularities. The end-to-end text recognition module included a feature pyramid network as the base structure, a region proposal network for generating text proposals, Fast-RCNN for bounding box regression, and a mask branch responsible for text instance segmentation and recognition.

Another issue is the development of unsupervised learning ReID. While the early unsupervised method Re-ID mainly learns invariant components such as dictionary, metric or saliency, cross-camera label estimation is one of the popular approaches in deep unsupervised methods. The performance of unsupervised Re-ID has improved significantly in recent years and continues to improve thanks to powerful attention schemes, image domain generation, cross-dataset learning, and so on. In the literature, the unsupervised person Re-ID problem is interpreted as unsupervised domain adaptation (UDA) and fully unsupervised (USL). The main idea of UDA is to fine-tune a model that is first trained on a fully labeled source domain and then transferred to an unlabeled target domain. The UDA methods often use pre-trained GANs or reinforcement learning models and then perform self-training on the target domain. Compared with the UDA methods, the USL methods have no restrictions on the source domain or the target domain, which makes this solution more attractive. Conventional approaches, such as those based on metric learning, are unable to effectively address problems associated with unsupervised Re-ID. Currently, the existing USL methods generally adopt clustering, including contractive learning, pseudo-label generation, inter-samples links, and so on. However, these methods do not allow generating higher quality pseudo-labels using clustering. Supplementing fine-grained local foreground information to the global information helps reduce pseudo-label noise.

For example, an unsupervised Re-ID method called multiple pseudo labels joint training predicted the presence of multiple pseudo-labels, improving the accuracy of pedestrian Re-ID in multi-camera systems (Tang et al., 2021). An adaptive information augmentation method based on k-nearest neighbour algorithm for unsupervised Re-ID problem (Wang et al., 2024b) obtained global and local features using dual branch structure, while an adaptive foreground enhancement module improved the robustness of features and the accuracy of pseudo-labels, and reduced the noise level of labels.

Image captioning remains a crucial task for text-image Re-ID. Image captions, generated from visual features and linguistic elements, must be both meaningful and grammatically correct. Recent image captioning models are mainly based on encoding-decoder architectures, which are typically based on CNN-LSTM, CNN-RNN or CNN-Transformer (Vaswani et al., 2017) models, and attention mechanisms. Transformer-based models generate high-quality captions but require higher resource requirements and computational costs. At the same time, CNN-LSTM and CNN-RNN models remain popular due to their simplicity and lower computational complexity.

Most of vision-language models require a lot of manual annotation. Unsupervised learning is a promising solution for image captioning. The unsupervised image captioning model was first proposed in (Feng et al., 2019). This model required an image set, a corpus for generating fluent sentences describing the image, and an object detection model. The image and caption were then projected into a common latent space to reconstruct each other to improve the generalization ability of the model. The fast unsupervised image captioning model proposed in (Yang et al., 2023) had encoder-decoder architecture with a pre-training module to train the decoder module followed by an MLP layer to output the text description. It is worth noting that research in the field of unsupervised image captioning for Re-ID has been insufficient. There methods require further development.

3. Proposed method

Transformation of representative features of visual content to language content is a challenging task, which can be solve differently. The earlier captioning models had a gap between local content and structural information, when high-level representations of the last layers of CNN were used as conditional language elements. Another approach employed additional probability distributions of common words. More appropriate way is exploiting unlabeled data, when a pseudo caption for each image is generated for initializing the image captioning model by the pseudo image-sentence pairs as it was done in (Feng et al., 2019).

The simplicity and compactness of representation of the global CNN features are the main advantages of the CNN-based image captioning model. However, specific and fine-grained descriptions are sources of detailed contextual information that can be extracted using an attention mechanism or graph model. We simplified the common task of image captioning by adding an instance segmentation module as a front-end module to the image captioning model. Based on our previous experiments (Favorskaya and Savkov, 2024), the unsupervised two-stage CutLER detector (Wang et al., 2023) was chosen. Thus, our image captioning model, whose architecture is shown in Figure 1, is close to the unsupervised image captioning model developed in (Feng et al., 2019).

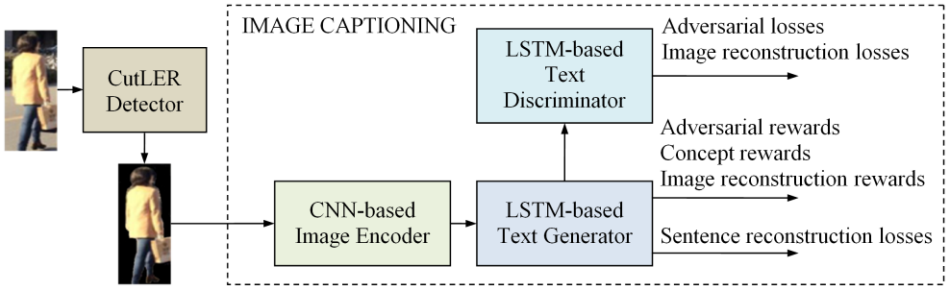

Figure 1. The architecture of an unsupervised image captioning model with additional unsupervised segmentation.

Most Re-ID models process images containing significant background information, which can introduce noise into feature representations. To mitigate this issue, we applied the CutLER detector to perform unsupervised instance segmentation, ensuring that only the person in the foreground is retained. CutLER's work consists of three steps: creating multiple initial coarse masks for the image using the MaskCut module, training detectors based on the coarse masks (using conditional random fields to smooth object boundaries and remove small background artifacts), and creating the final mask. To better extract visual features, we further clean the background using the detected mask as shown in Figure 1.

Although the use of the CutLER model significantly reduces the amount of background noise and improves the clarity of extracted visual features, it also introduces a dependency on the quality of instance masks. Inaccurate segmentation may lead to partial occlusion or truncation of the person shape, which in turn negatively affects the text generation pipeline.

The input data for image captioning are real images of persons and a corpus of corresponding sentences describing the person, as cues for unsupervised learning of the Text Generator. Visual and textual data do not correlate with each other. CNN-based Image Encoder is often built on Inception-ResNet-v4 or ResNet-50 models. Here, we used the R2-Inception-V4 model.

LSTM-based Text Generator includes 10 LSTM (long short-term memory) cells and tries to describe the content of an image in a natural sentence. Each LSTM cell outputs a probability distribution over all the words in the vocabulary and previously generated words. LSTM uses three types of gates: input, forget, and output ones, to achieve the storage and update of context information in long-term dependent tasks. These three gates are used to control whether the current cell value should be forgotten, whether the input data should be read, and whether the new cell value should be output. The generated word is selected from the dictionary in accordance with the obtained probability distribution:

$$\begin{aligned} x_{-1} &= FC\left(f_{image}\right) \\ x_t &= \mathbf{W}_e s_t \ , t \in \{0,\ldots,n-1\} \\ \left[p_{t+1} h_{t+1}^g\right] &= LSTM^g\left(x_t, h_t^g\right),\ t \in \{-1,\ldots,n-1\} \\ s_t &\sim p_t,\ t \in \{1,\ldots,n\} \end{aligned} \quad , \qquad (1)$$

where x_{-1} = initial value of the LSTM-based network
FC = fully-connected layer
f_{image} = visual features in FC layer
x_t = the LSTM input at the t-th time step
s_t = a vector representation of the generated word at the t-th time step
$\mathbf{W}_e$ = the word embedding matrix
n = length of the generated sentence
h_t^g = LSTM hidden state in Text Generator at the t-th time step
p_t = probability over the dictionary at the t-th time step
$LSTM^g$ = LSTM in Text Generator
~ = sampling operation

The generated word vector s_t is sampled from the probability distribution p_t, forming the tokens of the beginning and end of the sentence.

The LSTM-based discriminator tries to distinguish whether a partial sentence is a real sentence from the corpus or generated by the model:

$$\left[q_t, h_t^d\right] = LSTM^d\left(x_t, h_{t-1}^d\right),\ t \in \{1,\ldots,n\}, \qquad (2)$$

where q_t = probability of the generated partial sentence at the t-th time step
h_t^d = LSTM hidden state in Text Discriminator at the t-th time step
$LSTM^d$ = LSTM in Text Discriminator

Text Discriminator estimates the probability that the sentence $S = [s_1, ..., s_t]$ provided by the Text Generator will be considered as a partial sentence with the first t words in the sentence $\hat{S}$ from the corpus.

Since the paired image-sentence data are not available, the objective functions that enable unsupervised image captioning are special rewards, namely adversarial, concept and image reconstruction rewards. The adversarial reward is assigned at each time-step and is calculated as the logarithm of the probability q_t. Increasing the value of the adversarial reward means that the generator gradually learns to generate plausible sentences by following grammar rules. However, the plausible sentences generated in this way may not match the input image. Image captions are relevant to an image when the captioning model incorporates visual concepts into the generated sentence. This is done by introducing a concept reward, which is estimated by the confidence score of that visual concept. Since visual concept detectors can reliably detect a limited number of object concepts, the third type of reward, image reconstruction rewards, plays the role of an assessor of objective generalization ability. However, since methods of image reconstruction from sentences are not sufficiently developed, it is advisable to

evaluate not the reconstructed image, but its features. The objective function of sentence reconstruction is defined as the cross-entropy loss.

In contrast to the original unsupervised image captioning model proposed by Feng (Feng et al., 2019), which uses object detectors and external caption datasets, our method introduces an explicit instance segmentation step using CutLER. Unlike prior approaches that rely on fixed region proposals or scene-based visual context, our pipeline ensures that only the person shape is used for caption generation. This assumption results in more relevant and precise textual descriptions, particularly suited for the person Re-ID task, where visual context beyond the person may introduce noise.

4. Experimental Results

We selected the MSCOCO (Microsoft Common Objects in Context) dataset (Lin et al., 2014) for the experiment as one of the most widespread and diverse datasets used for training and evaluating computer vision systems. In the full version, it contains more than 200 thousand images (of which about 118 thousand are allocated for the training sample, 5 thousand for the validation sample, and another 5 thousand for the test sample). In addition, MSCOCO is notable for the fact that the scenes in it are as close to everyday as possible (kitchen utensils, people, street scenes, animals, etc.). The developed thematic and visual diversity helps COCO-trained models to better generalize to real conditions. Additionally, as a text corpus, we will use a large collection of descriptions collected from Shutterstock to train the model to generate detailed phrases based on visual features.

To pre-train the proposed model, it is necessary to annotate a number of images of the dataset. Annotations were made for 12% of the dataset. The image captioning model was trained using an unsupervised procedure similar to the framework described by Feng (Feng et al., 2019). Pseudo-captions were initialized using a corpus of text descriptions describing clothing, colours, and human appearance. Training was performed for 40 epochs with a batch size of 32 and a learning rate of 0.0002. The model was optimized using a combination of an adversarial loss, concept recognition score, and feature reconstruction reward. Pseudo-captions were generated by projecting image embeddings into a shared latent space and decoding them using an LSTM-based generator trained on the external corpus.

The task of Re-ID requires annotation of clothing and colour: 3 types of clothing above the waist and 3 types of clothing below the waist were selected. An example of annotation is phrases like: “A man in a blue jacket and black trousers”, “A woman in a white T-shirt and gray shorts” and so on. Figure 2 shows examples of original and segmented images and the generated text annotations. In Figure 2, Concepts (keywords) describe the segmented images, and Models generate captions.

To evaluate the quality of generated text descriptions, five metrics were selected: BLEU, METEOR, ROUGE, CIDEr, and SPICE. Bilingual Evaluation Understanding (BLEU) is one of the earliest and most widely used metrics for evaluating the quality of machine translations and automatically generated descriptions (Papineni et al., 2002). BLEU is based on comparing the *n*-grams of the result with one or more reference texts; the higher the match, the better the model retains context and vocabulary. The different variants of BLEU reflect the depth of the analysis: BLEU-1 (unigrams) shows the degree of coincidence of individual words, but does not take into account their sequence. Higher orders of *n*-grams, such as BLEU-2, BLEU-3, and BLEU-4, already allow us to judge the correct formation of phrases and take into account the sentence structure, since they check the matches of bigrams, trigrams, and tetragrams, respectively.

Metric for Evaluation of Translation with Explicit ORdering (METEOR) expands the analysis by taking into account synonyms, morphological forms and word order, which makes it more sensitive to the lexical and grammatical features of the text (Agarwal and Lavie, 2008).

Recall-Oriented Understanding for Gisting Evaluation (ROUGE) is a set of metrics, most often including ROUGE-N and ROUGE-L, focused on the measurement of completeness (recall) (Lin, 2004). They assess the extent to which *n*-grams or the longest common sequence overlap with the reference description, which initially made ROUGE particularly useful in the task of automatic text abstraction, but later found wide application in the evaluation of image captioning systems.

Another important metric, Consensus-based Image Description Evaluation (CIDEr), is used primarily to evaluate the quality of visual descriptions (Vedantam et al., 2015). It relies on the TF-IDF representation of *n*-grams, matching the resulting signature with several reference ones at once, and encourages consistency of terms. Finally, Semantic Propositional Image Caption Evaluation (SPICE) focuses on checking semantic relationships: the metric builds a scene graph and compares objects, attributes, and relationships in the generated description with the same elements in the reference one (Anderson et al., 2016).

Thus, each of the listed metrics evaluates certain aspects of the quality of the generated texts, complementing each other and providing a comprehensive verification of the compliance of automatic descriptions with reference ones. Table 1 shows the averaged accuracy values for the selected MSCOCO images with and without using CutLER. Here, B1, B2, B3, and B4 are the BLEU-1, BLEU-2, BLEU-3, and BLEU-4 metrics, respectively. M, R, C, and S are the METEOR, ROUGE, CIDEr, and SPICE metrics, respectively.

Method	B1	B2	B3	B4	M	R	C	S
Without CutLER	41.0	22.5	11.2	5.6	12.4	28.7	28.6	8.1
With CutLER	41.6	23.8	11.9	5.8	13.2	30.3	28.9	8.1

Table 1. Averaged accuracy values for the selected MSCOCO images with and without CutLER

It can be seen from Table 1 that when CutLER (the foreground automatic segmentation module) is added, all text metrics (BLEU1-BLEU4, METEOR, ROUGE-L, CIDEr and SPICE) become higher. The main reason is that by clipping the background, the captioning model “looks” only at the human shape and highlights relevant attributes better (clothing colour, accessories, etc.). As a result, the descriptions are more accurate and consistent, which is reflected in improved performance.

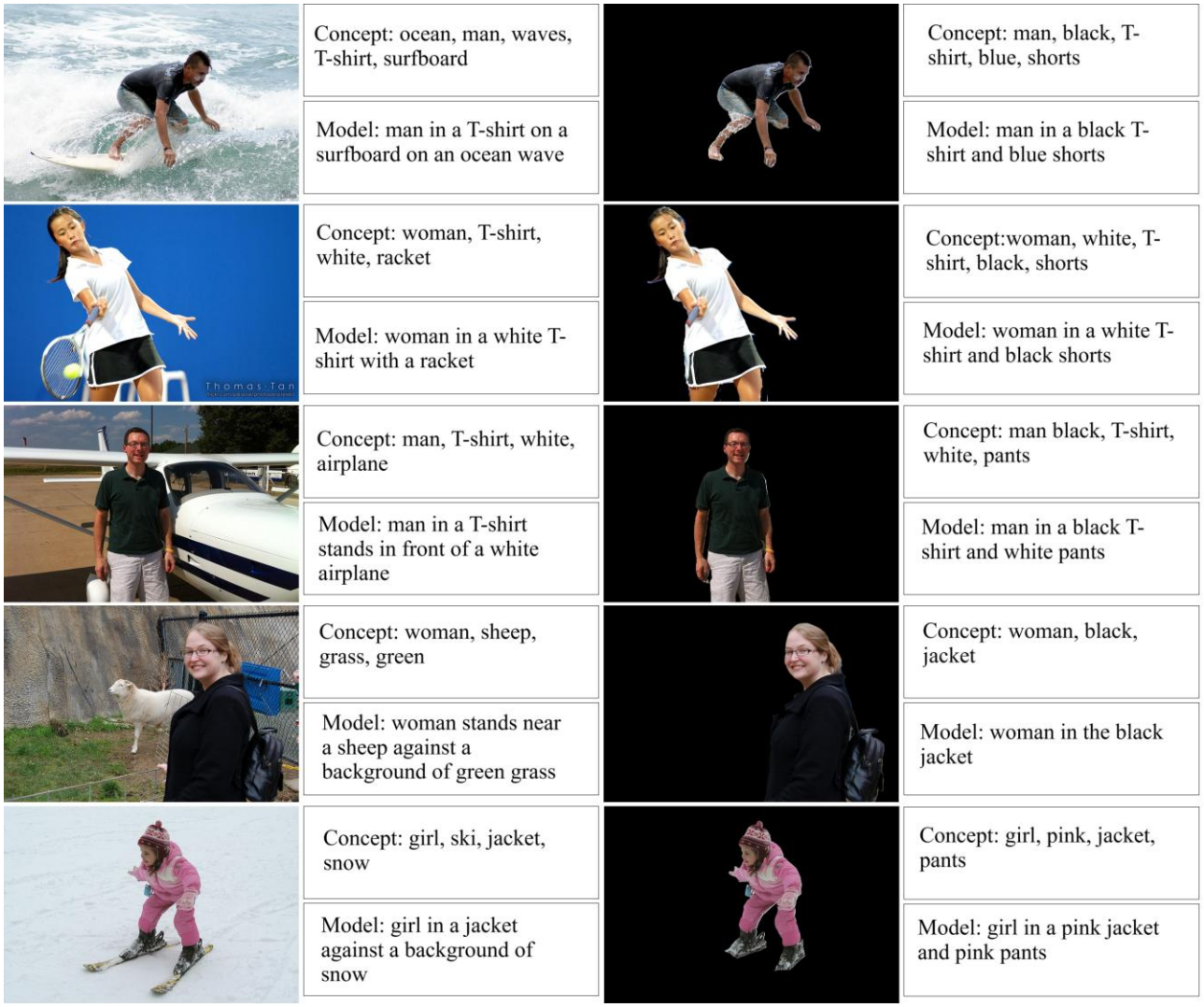

Figure 2. Examples of MSCOCO images and text annotations: **a** original images, **b** generated text annotations without using CutLER, **c** segmented images using CutLER, **d** generated text annotations with using CutLER.

5. Conclusions

The proposed method, combining unsupervised segmentation with the generation of text descriptions, demonstrates an improvement in the quality of automatic image caption in an experiment on the MSCOCO dataset. The inclusion of the target object selection stage significantly reduces the influence of background areas and contributes to a more accurate identification of key visual characteristics of the scene. This is reflected in the increased performance of BLEU, METEOR, ROUGE-L, CIDEr and SPICE metrics compared to the model without segmentation, which indicates the effectiveness of the proposed approach in forming detailed and consistent descriptions in the absence of annotated image–text pairs.

Despite the improvements obtained using foreground segmentation and unsupervised learning, the model’s dependence on segmentation quality and the lack of precise text-image alignment remain challenges. Further improvements could include the inclusion of attention-based refinement mechanisms or self-supervised fine-tuning of visual embeddings to improve robustness across different scenes.

References

Agarwal, A., Lavie, A., 2008: METEOR, m-bleu and m-ter: Evaluation metrics for high correlation with human rankings of machine translation output. *Third Workshop on Statistical Machine Translation*, ACL, Columbus, Ohio, pp. 115-118.

Anderson, P., Fernando, B., Johnson, M., Gould, S., 2016: SPICE: Semantic propositional image caption evaluation. *In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds) Computer Vision – ECCV 2016*, LNCS, vol. 9909, Springer, Cham, pp. 382-398.

Du, G., Gong, T., Zhang, L., 2024: Contrastive completing learning for practical text–image person ReID: Robuster and cheaper. Expert Systems With Applications 248, 123399.1-123399.15.

Favorskaya, M.N., Savkov, M.V., 2024: Study on unsupervised instance segmentation models for person re-identification. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci., XLVIII-*

2/W5-2024, International Workshop on “Photogrammetric Data Analysis” – PDA24, Moscow, Russia, pp. 41-48.

Feng, Y., Ma, L., Liu, W., Luo, J., 2019: Unsupervised image captioning. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Long Beach, CA, USA, pp. 4125-4134.

Jia, M., Cheng, X., Lu, S., Zhang, J., 2023: Learning disentangled representation implicitly via transformer for occluded person re-identification. *IEEE Trans. Multimed*. 25, 1294-1305.

Li, S., Xiao, T., Li, H., Zhou, B., Yue, D., Wang, X., 2017: Person search with natural language description. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Honolulu, HI, USA, pp. 1970-1979.

Li, Y., He, J., Zhang, T., Liu, X., Zhang, Y., Wu, F., 2021: Diverse part discovery: Occluded person re-identification with part-aware transformer. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Nashville, TN, USA, pp. 2898-2907.

Li, S., Sun, L., Qingli Li, Q., 2023: CLIP-ReID: Exploiting vision-language model for image re-identification without concrete text labels. *AAAI'23/IAAI'23/EAAI'23*, Washington, D.C., USA, 156, pp. 1405-1413.

Li, X., Guo, H., Zhang, M., Fu, B., 2025: Image–text person re-identification with transformer-based modal fusion. *Electronics* 14, 525.1-525.15.

Lin, C.-Y., 2004: Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out*, ACL, Barcelona, Spain, pp. 74-81.

Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C.L., Dollar, P., 2014: Microsoft COCO: Common objects in context. *In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds) Computer Vision – ECCV 2014. ECCV 2014. LNCS*, vol. 8693, pp. 740-755. Springer, Cham.

Papineni, K., Roukos, S., Ward, T., Zhu, W.-J., 2002: BLEU: a method for automatic evaluation of machine translation. *40th Annual Meeting of the Association for Computational Linguistics*, ACL, Philadelphia, PA, USA, pp. 311-318.

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I., 2021: Learning transferable visual models from natural language supervision. *The 38th International Conference on Machine Learning (ICML 2021)*, Virtual, 139, pp. 8748-8763.

Rao, Y., Chen, G., Lu, J., Zhou, J., 2021: Counterfactual attention learning for fine-grained visual categorization and re-identification. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, IEEE, Montreal, QC, Canada, pp. 1025-1034.

Ren, K., Hu, C., Xi, H., Li, Y., Fan, J., Liu, L., 2025: MoSCE-ReID: Mixture of semantic clustering experts for person re-identification. *Neurocomputing* 626, 129587.1-129587.13.

Tang, Q., Cao, G., Jo, K.H., 2021: Fully unsupervised person re-identification via multiple pseudo labels joint training. *IEEE Access 2021*, 9, 165120-165131.

Tan, H., Liu, X., Bian, Y., Wang, H., Yin, B., 2021: Incomplete descriptor mining with elastic loss for person re-identification. *IEEE Trans. Circuits Syst. Video Technol*. 32, 160-171.

Tay, C.-P., Roy, S., Yap, K.-H., 2019: AANet: Attribute attention network for person re-identifications. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Long Beach, CA, USA, pp. 7134-7143.

Tian, Y., Krishnan, D., Isola, P., 2020: Contrastive multiview coding. *16th European Conference on Computer Vision – ECCV 2020*, Glasgow, UK, Part XI, pp. 776-794.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I., 2017: Attention is all you need. *31st Conference on Neural Information Processing Systems (NIPS 2017)*, Long Beach, CA, USA, pp.1-11.

Vedantam, R., Lawrence Zitnick, C., Parikh, D., 2015: CIDEr: Consensus-based image description evaluation. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Boston, MA, USA, pp. 4566-4575.

Wang, X., Girdhar, R., Yu, S.X., Misra, I., 2023. Cut and learn for unsupervised object detection and instance segmentation. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* IEEE, Vancouver, Canada, 3124-3134.

Wang, R., Zhu, Y., Wan, Z., Chen, H., Zhu, Z., Zhou, W., Han, C., Ding, Y., 2024a: A person re-identification method for sports event scenes incorporating textual information mining. *IET Image Proc*. 18(7), 1681-1693.

Wang, Q., Huang, Z., Fan, H., Fu, S., Tang, Y., 2024b: Unsupervised person re-identification based on adaptive information supplementation and foreground enhancement. *IET Image Process.* 18, 4680-4694.

Yang, R., Cui, X., Qin, Q., Deng, Z., Lan, R., Luo, X., 2023: Fast RF-UIC: A fast unsupervised image captioning model. *Displays* 79, 102490.1-102490.8.

Yan, S., Dong, N., Zhang, L., Tang, J., 2023: CLIP-driven fine-grained text-image person re-identification. *IEEE Trans. Image Process*. 32, 6032–6046.

Zhou, K., Yang, Y., Cavallaro, A., Xiang, T., 2019: Omni-scale feature learning for person re-identification. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, IEEE, Seoul, Republic of Korea, pp. 3702–3712.