DOI: https://doi.org/10.62985/j.huit_ojs.vol26.no2E.419

COMPARATIVE ANALYSIS OF LOSS FUNCTIONS FOR IMAGE-TEXT MATCHING UNDER NOISY CORRESPONDENCE

Tam T. Ngo[1,*], Anh V. Nguyen[2], Hoa N. Nguyen[1]

[1]*VNU University of Engineering and Technology, Hanoi, Vietnam*

[2]*HCM University of Foreign Languages and Information Technology, Vietnam*

*Email: *23020568@vnu.edu.vn*

Received: 19 January 2026; Revised: 3 April 2026; Accepted: 22 April 2026

ABSTRACT

Image–Text Matching (ITM) plays an important role in vision–language applications such as cross-modal retrieval. However, real-world datasets often contain noisy correspondence, where image–text pairs are incorrectly aligned or only partially related, which can degrade model performance. In this paper, we conduct a comparative analysis of common loss functions for ITM under noisy conditions, including Triplet Loss and InfoNCE. We further introduce a new objective, Similarity-based Negative Log-Likelihood (SNLL), which formulates image–text alignment as a probabilistic binary classification over all pairwise similarities. Experiments on the MS-COCO dataset under different noise levels show that while all methods perform similarly on clean data, SNLL achieves more stable training and higher retrieval performance when noise increases, demonstrating stronger robustness to noisy correspondence.

Keywords: Text-to-Image, Cross-Modality, Noisy Correspondence, Contrastive objectives, Similarity-based Negative Log-Likelihood.

1. INTRODUCTION

Image-Text Matching (ITM) refers to the task of retrieving the most relevant images or captions for a given textual or visual query. This task has wide-ranging applications, including cross-modal retrieval, where paired image-caption datasets such as MS-COCO [1] are employed, as well as zero-shot classification, in which class labels are expressed as text prompts. Due to its central role in both visual understanding and language processing, ITM has become a cornerstone task in vision-language research.

However, a significant challenge facing modern ITM systems is the presence of Noisy Correspondence (NC) [2]. In web-crawled image–text pairs, these correspondences are often incorrectly aligned or only partially related, introducing substantial noise. This noise severely compromises contrastive learning objectives and frequently leads to degraded performance in downstream retrieval tasks.

Motivated by this challenge, this paper provides a comparative study of several state-of-the-art loss functions for ITM. Specifically, we analyze Triplet Loss [3], InfoNCE [4], and an effective alternative we introduce, Similarity-based Negative Likelihood Loss (SNLL), aiming to gain a deeper understanding of their behaviors when training deep ITM models. We posit that a better comprehension of how these loss functions handle various noise conditions is crucial for developing more robust ITM architectures. To evaluate their effectiveness under varying noise conditions, we conduct extensive experiments on the MS-COCO benchmark. Our main contributions are summarized as follows:

1. We provide a *comparative survey* of the predominant loss functions (Triplet Loss, InfoNCE) utilized in Image-Text Matching (ITM).

2. We propose the *Similarity-based NLL (SNLL)*, a novel loss formulation designed to improve learning capacity in the presence of noisy correspondence.

3. We perform *comprehensive empirical evaluations* on MS-COCO, assessing the performance of Triplet Loss, InfoNCE, and SNLL under multiple noise scenarios to highlight their respective strengths and weaknesses.

The remainder of this paper is organized as follows. In *§*2, we describe our main problem and introduce some related work. *§*3 details our proposed Similarity-based NLL (SNLL) loss function. In *§*4 reports the experimental results and comparative analysis. Finally, *§*5 concludes the paper and outlines future research directions.

2. RELATED WORK

Image–Text Matching is a fundamental task in vision–language understanding that aims to project images and textual descriptions into a shared embedding space, where semantically aligned image–text pairs lie close together while mismatched pairs are pushed apart. Early methods [5] relied on global CNN–RNN embeddings optimized with triplet ranking losses, which, despite their simplicity, could not capture fine-grained region–word relationships. To address this limitation, attention-based local alignment approaches were introduced, and later, multi-layer vision–language Transformers further enhanced cross-modal interactions at the cost of increased computation. More recently, large-scale contrastive pretraining methods such as CLIP [6] and ALIGN [7], along with modern extensions like SigLIP [8] and BLIP [9]/BLIP-2 [10], have pushed ITM beyond retrieval, enabling stronger generalization and broader multimodal capabilities.

While model architecture has evolved significantly, the optimization objective remains a key determinant of ITM performance. Triplet Ranking Loss remains widely adopted but is sensitive to margin selection and highly vulnerable to noisy or mismatched pairs. Contrastive objectives such as InfoNCE offer scalable softmax-based alignment but depend on temperature tuning and struggle to distinguish hard negatives from false or duplicated negatives. These limitations indicate that both ranking-based and contrastive losses can be unstable when trained on imperfect real-world data.

A key challenge in practical ITM is Noisy Correspondence (NC), where image–text pairs are mislabeled, weakly aligned, or only partially relevant. Noisy Correspondence Learning (NCL) was first formalized by Huang et al., who highlighted how misaligned pairs in web-scale datasets can significantly distort similarity learning. Unlike traditional noisy-label settings that focus on category-level annotation errors, NC stems from cross-modal misalignment within paired data, making the problem inherently more complex. While recent studies have explored debiased contrastive learning, noise-aware caption analysis, and general noise-robust strategies, these efforts primarily target single-modality noise and often overlook cross-modal mismatches, which are central to ITM. Despite its practical importance, systematic evaluations of foundational loss functions under NC remain scarce, leaving open questions about their intrinsic robustness-thus motivating the investigation presented in this work.

3. PROPOSED METHOD

3.1. Preliminaries

3.1.1. Clip-based Network Architecture

Fig. 1 depicts a CLIP network architecture , which encodes images and text separately and

produces feature embeddings for each modality. Specifically,

Image Encoder: An input image $I \in R^{HxWxC}$ is first divided into $M = (H/S) \times (W/S)$ non-overlapping patches of size $S \times S$. Each patch is linearly projected into a D-dimensional embedding, and learnable [CLS] token is prepended to form the sequence $\{p_{\text{CLS}}, p_1, \ldots, p_M\}$ This sequence passes through P Transformer blocks, yielding the set of hidden representations $\{f_{CLS}^I, f_1^I, \ldots, f_M^I\}$. In deterministic models f_{CLS}^I serves as the global image feature.

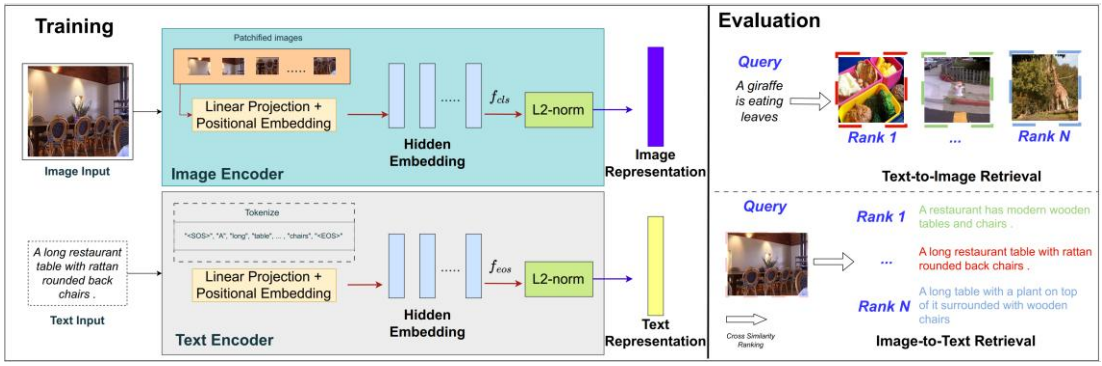

Fig. 1. Overview of training and evaluation ITM model.

Text Encoder. A text sequence T is tokenized into $\{t_{SOS}, t_1, \ldots, t_L, t_{EOS}\}$ where L is the number of tokens. After inserting start/end tokens, the sequence is embedded and processed by Q Transformer blocks. The outputs$\{f_{SOS}^T, f_1^T, \ldots, f_L^T, f_{EOS}^T\}$ are D-dimensional; f_{EOS}^T represents the global text embedding in standard deterministic pipelines.

3.1.2. Loss Functions

To effectively align image and text embeddings in a shared representation space, we employ metric-learning objectives that encourage matched pairs to be close while pushing apart mismatched pairs.

Triplet Loss. Given an image embedding f^I, a matched text embedding f_+^T, and a mismatched text embedding f_-^T, the triplet loss enforces that the similarity between (f^I, f_+^T) is greater than that of (f^I, f_-^T) by a margin α. Using cosine distance d(u,v) = $1 - \frac{u^T v}{\|u\|\|v\|}$, the loss is defined as

$$\mathcal{L}_{triplet} = \max(0, (f^I, f_+^T) - d(f^I, f_-^T) + \alpha) \quad (1)$$

This formulation encourages the model to learn discriminative embeddings that separate positive and negative samples.

InfoNCE Loss. For a mini-batch of N image-text pairs $\{(f_i^v, f_j^t)\}_{i=1}^N$, we compute the similarity-based probability between image i and text j as

$$p_{i,j} = \frac{\exp\left(\frac{sim(f_i^v, f_j^t)}{\tau}\right)}{\sum_{k=1}^N \exp\left(\frac{sim(f_i^v, f_k^t)}{\tau}\right)} \quad (2)$$

where sim $(\cdot, \cdot)$ denotes cosine similarity and τ is a temperature hyperparameter. Based on this, the InfoNCE loss encourages matched pairs to have higher similarity than mismatched ones, and is defined as the negative log-likelihood of correct matches:

$$\mathcal{L}_{nce} = -\frac{1}{N}\sum_{i=1}^{N}\log p_{i,i} \tag{3}$$

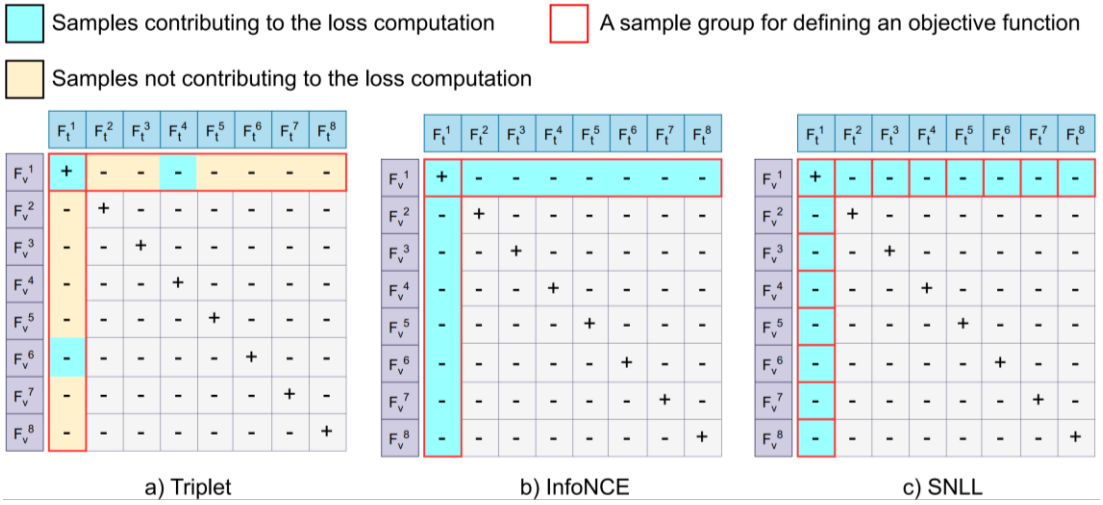

Fig. 2. Comparison of loss function supervision scopes in image-text matching: Triplet, InfoNCE, and SNLL. Each matrix illustrates which sample pairs contribute to the objective, highlighting the progression from sparse to dense and probabilistic supervision.

3.2. Similarity-based NLL Loss Function

In image-text matching (ITM), learning discriminative cross-modal representations requires loss functions that can effectively distinguish between matched and mismatched pairs. Traditional objectives such as triplet loss enforce relative ranking between positive and negative samples, while InfoNCE loss maximizes the likelihood of correct matches via a softmax over batch similarities. However, both approaches rely on sampling strategies that may overlook fine-grained pairwise interactions.

To address this limitation, we propose a modification of the Negative Log-Likelihood (NLL) loss [11], termed Similarity-based NLL (SNLL). This formulation recasts NLL into a binary classification objective defined over all image-text pairs. Specifically, given a mini-batch of N image-text pairs $\{(f_i^v, f_j^t)\}_{i=1}^{N}$, we compute the similarity score between image i and text j as

$$s_{i,j} = sim\left(f_i^v, f_j^t\right)$$

Unlike NCInfo, which focuses only on diagonal entries $p_{i,i}$ (Equation 2), SNLL considers all pairwise combinations. Binary labels $y_{i,j} \in \{0, 1\}$ are assigned to indicate whether image i and text j form a matched pair. The final loss is defined as a weighted negative log-likelihood over all pairs:

$$L_{\{SNLL\}(I_i,T_j)} = -y_{\{i,j\}}\log\sigma\left(\frac{s_{i,j}}{\tau}\right) - \left(1-y_{i,j}\right) log\ \sigma\left(1-\frac{s_{i,j}}{\tau}\right) \tag{4}$$

where τ is the temperature parameter. Following prior work [12], we set $\tau = 5$, since the optimal logit range for NLL lies within $[-5, 5]$, whereas similarity values are bounded in $[-1, 1]$.

As illustrated in Fig. 2, the theoretical distinction between our proposed SNLL and existing contrastive objectives lies in the scope and granularity of supervision. Triplet loss operates on sparse triplets-one anchor, one positive, and one negative—enforcing relative ranking but relying heavily on sampling quality. InfoNCE loss improves batch efficiency by applying softmax over diagonal entries, yet it only supervises matched pairs and ignores off-diagonal mismatches. In contrast, SNLL reformulates contrastive learning as a probabilistic binary classification task

over all pairs, using sigmoid-weighted negative log-likelihood to softly separate matched and mismatched samples. This enables dense supervision, smooth optimization, and better tolerance to label noise.

In the remainder of this paper, we analyze the impact of noisy supervision on ITM models trained with different loss functions. Our objective is to elucidate and visualize how common objectives-such as Triplet Loss [3], InfoNCE Loss [4], and the proposed Similarity-based NLL Loss-respond to noise in the training data, including mislabeled pairs and semantically ambiguous matches. Through controlled experiments across multiple benchmarks, we aim to reveal the sensitivity of each formulation to noise perturbations and to identify which losses exhibit greater robustness and generalization. This analysis provides empirical insights into the behavior of contrastive learning under imperfect data conditions and informs the design of more noise-tolerant training strategies for multimodal retrieval tasks.

4. EXPERIMENTS AND EVALUATION

4.1. Implementation Details

Datasets and Evaluation Metrics. We validate our model on the well-known dataset MS COCO Caption is used with a split of 113,287 training images, 5,000 validation images. According to this dataset, previous works [13, 14, 15] use two kinds of evaluation metrics as COCO-5k and COCO-1k. For COCO-5K retrieval we use all 5,000 test images against 25,000 captions; COCO-1K averages recall over five disjoint 1,000-image subsets. On EC, we report mean average precision (mAP) and cumulative matching curve rank@1 accuracy metrics (R@1) and R-Precision to emphasize precision under noisy negatives. We also compute RSUM which is the sum of R@1, R@5, and R@10 for each retrieval direction on COCO-1k. To be consistent with other papers, we will present the modality-averaged scores of retrievals.

Training Settings. Our method standardizes input images to a resolution of 256×128 pixels and constrains text sequences to 77 tokens. For data augmentation, we employ techniques such as Random Horizontal Flipping, Random Cropping with Padding exclusively for images, and Random Erasing applied to both images and texts following [16, 17]. The model architecture incorporates a pre-trained CLIP-ViT-B/16 as the image encoder and a CLIP Text Transformer as the text encoder . Throughout the training process, we utilize the Adam optimizer for 15 epochs for others with a batch size of 128 and an initial learning rate of 1×10^{-5}. Our experiments are conducted on a single NVIDIA A100 GPU node. For noise scenarios, we follow to mix up a number of sample pairs in proportions of 20%, 50%, and 80%.

Baseline. In experiments, we adopt Triplet Loss and InfoNCE Loss as baseline objectives for training image-text matching models. Triplet Loss serves as a classical metric-learning approach that enforces relative ranking between positive and negative pairs, while InfoNCE provides a probabilistic formulation that maximizes the likelihood of correct matches within a batch. Together, these baselines establish strong reference points for evaluating the effectiveness of our proposed Similarity-based NLL Loss (SNLL), allowing us to systematically compare their sensitivity to noisy supervision and their robustness across different datasets.

4.2. Results

In this section, we present the results of our comparative evaluation across different noise levels. We evaluate the performance of Triplet Loss, InfoNCE, and our proposed SNLL on the MS-COCO dataset (113k training images, 5k/1k test splits) across three noise scenarios: clean data (0% noise), moderate noise (50%), and high noise (80%). Key metrics include Recall@1 (R@1), Recall@5 (R@5), Recall@10 (R@10), and RSUM (sum of R@K across both text-to-image and image-to-text directions), averaged over modalities.

Table 1 summarizes the results. Under clean conditions (0% noise), all methods achieve

comparable performance, with RSUM scores around 538; SNLL slightly outperforms baselines (RSUM 539.59 vs. 538.34 for InfoNCE and 536.82 for Triplet), gaining 0.3-1% in R@1 due to its dense supervision mechanism. As noise increases to 50%, divergences emerge sharply: Triplet collapses (RSUM drops 66% to 179.79; R@1 on COCO-1k falls from 75.65% to 11.48%), while InfoNCE degrades moderately (RSUM 504.70, -6%). SNLL maintains superior stability (RSUM 522.98, -3%), exceeding InfoNCE by 18 RSUM points and preserving 70.73% R@1 on COCO-1k.

At 80% noise, the gap widens further. Triplet fails catastrophically (RSUM 24.15; R@1 3.32%), rendering it unusable. InfoNCE sustains functionality but suffers substantial decline (RSUM 458.87, -15%). In contrast, SNLL retains the highest scores (RSUM 494.94, -8%; COCO-1k R@1 63.47% vs. InfoNCE 53.94%), demonstrating 18% relative improvement in top 1 recall.

Table 1. Performance comparison under different noise rates on benchmarks

Method	Noise	COCO-1k			COCO-5k			RSUM
		R@1	R@5	R@10	R@1	R@5	R@10	
InfoNCE	0%	76.10	94.99	98.08	56.88	82.43	89.61	538.34
Triplet		75.65	94.73	98.04	56.03	81.44	89.50	536.82
SNLL		76.56	95.01	98.22	57.36	82.71	89.64	539.59
InfoNCE	50%	68.97	91.75	96.63	43.78	72.34	81.28	504.70
Triplet		11.48	32.88	45.53	3.65	12.68	19.92	179.79
SNLL		70.73	92.43	96.83	51.33	77.22	88.46	522.98
InfoNCE	80%	53.94	83.50	91.99	31.36	59.14	70.91	458.87
Triplet		3.32	6.20	8.14	0.81	2.11	3.56	24.15
SNLL		63.47	88.94	95.05	41.89	69.32	79.60	494.94

4.3. Visualization Analysis

The results reveal clear differences in robustness under noisy correspondence. Triplet Loss proves most vulnerable as each update hinges on a single positive-negative pair; corruption in either generates contradictory gradients that rapidly destabilize the embedding space. Hard-negative mining exacerbates this, triggering early collapse as evidenced by Fig. 3 (RSUM plummets post-epoch 5).

InfoNCE exhibits greater resilience via batch-level softmax aggregation, mitigating individual noisy samples. Yet its diagonal-only supervision induces bias as noise accumulates, while prolonged training causes softmax over-sharpening-excessive confidence in surviving clean positives-yielding the gradual RSUM drift observed in Fig. 3 (538→505).

By contrast, SNLL leverages dense pairwise supervision across all image-text pairs, with each contributing independently via sigmoid probabilities. This dilutes noisy influences and circumvents softmax over-confidence, sustaining stable dynamics and superior RSUM.

5. CONCLUSION

In this paper, a comparative study of loss functions for image-text matching under noisy correspondence was conducted, focusing on Triplet Loss, InfoNCE, and the proposed Similarity-based NLL (SNLL). The experiments on MS-COCO across multiple noise levels demonstrate that while all methods perform similarly on clean data, Triplet Loss rapidly collapses as noise increases, and InfoNCE exhibits gradual performance degradation over prolonged training due to its reliance on diagonal positive pairs. In contrast, SNLL consistently achieves the highest recall and RSUM scores, maintaining stable learning dynamics even under severe noise, thanks to its dense pairwise supervision and probabilistic formulation.

Fig. 3. R-sum Progress under 50% Noise on COCO-1k: Comparison of Triplet Loss, InfoNCE, and SNLL (ours) in terms of RSUM - the sum of R@1, R@5, and R@10 across both retrieval directions. SNLL demonstrates superior stability and peak performance, while Triplet Loss collapses early and InfoNCE shows moderate degradation over time.

These findings highlight that the choice of loss function is critical for robust image-text matching in real-world scenarios, where web-crawled datasets often contain substantial misalignment and ambiguous correspondences. By treating all image–text pairs in a batch as supervised instances, SNLL effectively dilutes the impact of corrupted samples and mitigates over-confident softmax behavior. Future work will extend this analysis to larger vision-language backbones and additional benchmarks and explore integrating SNLL with noise-aware sampling or correspondence refinement strategies to further enhance robustness in large-scale cross-modal retrieval systems.

REFERENCES

[1] X. Chen *et al*., "Microsoft COCO Captions: Data Collection and Evaluation Server," *arXiv preprint arXiv*:1504.00325, 2015. doi: https://doi.org/10.48550/arXiv.1504.00325.

[2] Z. Huang *et al*., "Learning with Noisy Correspondence for Cross-modal Matching," in *Advances in Neural Information Processing Systems*, vol. 34, 2021, pp. 29406–29419. doi: https://doi.org/10.48550/arXiv.2105.03805.

[3] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A Unified Embedding for Face Recognition and Clustering," in 2015 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 815–823. doi: https://doi.org/10.1109/CVPR.2015.7298682.

[4] A. van den Oord, Y. Li, and O. Vinyals, "Representation Learning With Contrastive Predictive Coding," *arXiv preprint arXiv:1807.03748*, 2018. doi: https://doi.org/10.48550/arXiv.1807.03748.

[5] D. H. Pham, A. D. Nguyen, and H. N. Nguyen, "GAN-based Data Augmentation and Pseudo-label Refinement With Holistic Features for Unsupervised Domain Adaptation Person Re-identification," *Knowledge-Based Systems*, vol. 298, p. 111471, 2024. doi:

https://doi.org/10.1016/j.knosys.2024.111471.

[6] A. Radford *et al*., "Learning Transferable Visual Models From Natural Language Supervision," in *International Conference on Machine Learning (ICML)*, PMLR, 2021, pp. 8748–8763. doi: https://doi.org/10.48550/arXiv.2103.00020.

[7] C. Jia *et al*., "Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision," *in International Conference on Machine Learning (ICML)*, PMLR, 2021, pp. 4904–4916. doi: https://doi.org/10.48550/arXiv.2102.05918.

[8] X. Zhai *et al*., "Sigmoid Loss for Language Image Pre-training," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 11975–11986. doi: https://doi.org/10.1109/ICCV51070.2023.01098.

[9] J. Li *et al*., "BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation," *in International Conference on Machine Learning (ICML)*, PMLR, 2022, pp. 12888–12900. doi: https://doi.org/10.48550/arXiv.2201.12086.

[10] Junnan Li *et al*. "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models". In: *International conference on machine learning*. PMLR. 2023, pp. 19730–19742. doi: https://doi.org/10.48550/arXiv.2301.12597.

[11] R. A. Fisher, "On the Mathematical Foundations of Theoretical Statistics," *Philosophical Transactions of the Royal Society of London*. Series A, vol. 222, pp. 309–368, 1922. doi: https://doi.org/10.1098/rsta.1922.0009.

[12] A. D. Nguyen *et al*., "Impact Analysis of Different Effective Loss Functions by Using Deep Convolutional Neural Network for Face Recognition," in *From Born-Physical to Born-Virtual: Augmenting Intelligence in Digital Libraries*, Y.-H. Tseng, M. Katsurai, and H. N. Nguyen, Eds. Cham: Springer, 2022, pp. 101–111. doi: https://doi.org/10.1007/978-3-031-21756-2_8.

[13] S. Chun *et al*., "Probabilistic Embeddings for Cross-Modal Retrieval," in 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 8411–8420. doi: https://doi.org/10.48550/arXiv.2101.05068.

[14] Kun Zhang *et al*. "Negative-Aware Attention Framework for Image-Text Matching". In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022, pp. 15640–15649. doi: https://doi.org/10.1109/CVPR52688.2022.01521

[15] Y. Song and M. Soleymani, "Polysemous Visual-Semantic Embedding for Cross-Modal Retrieval," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 1979–1988. doi: https://doi.org/10.1109/CVPR.2019.00208.

[16] A. D. Nguyen and H. N. Nguyen, "Enhancing Text-Based Person Retrieval by Combining Fused Representation and Reciprocal Learning With Adaptive Loss Refinement," *IEEE Transactions on Image Processing*, vol. 34, pp. 5147–5157, 2025. doi: https://doi.org/10.1109/TIP.2025.3594880.

[17] A. D. Nguyen, H.-Y. Kim, and H. N., "TALIU: A Novel Decoder and Augmentation Strategy for Boosting Tampered Document Image Detection," *IEEE Access*, pp. 1–1, 2025. doi: https://doi.org/10.1109/ACCESS.2025.3560360.