Learning Joint Visual Semantic Matching Embeddings for Language-guided Retrieval

Yanbei Chen*[1] and Loris Bazzani✉[2]

[1] Queen Mary University of London
yanbei.chen@qmul.ac.uk
[2] Amazon
bazzanil@amazon.com

Abstract. Interactive image retrieval is an emerging research topic with the objective of integrating inputs from multiple modalities as query for retrieval, e.g., textual feedback from users to guide, modify or refine image retrieval. In this work, we study the problem of composing images and textual modifications for language-guided retrieval in the context of fashion applications. We propose a unified Joint Visual Semantic Matching (JVSM) model that learns image-text compositional embeddings by jointly associating visual and textual modalities in a *shared* discriminative embedding space via compositional losses. JVSM has been designed with *versatility* and *flexibility* in mind, being able to perform multiple image and text tasks in a *single* model, such as text-image matching and language-guided retrieval. We show the effectiveness of our approach in the fashion domain, where it is difficult to express keyword-based queries given the complex specificity of fashion terms. Our experiments on three datasets (Fashion-200k, UT-Zap50k, and Fashion-iq) show that JVSM achieves state-of-the-art results on language-guided retrieval and additionally we show its capabilities to perform image and text retrieval.

Motivations

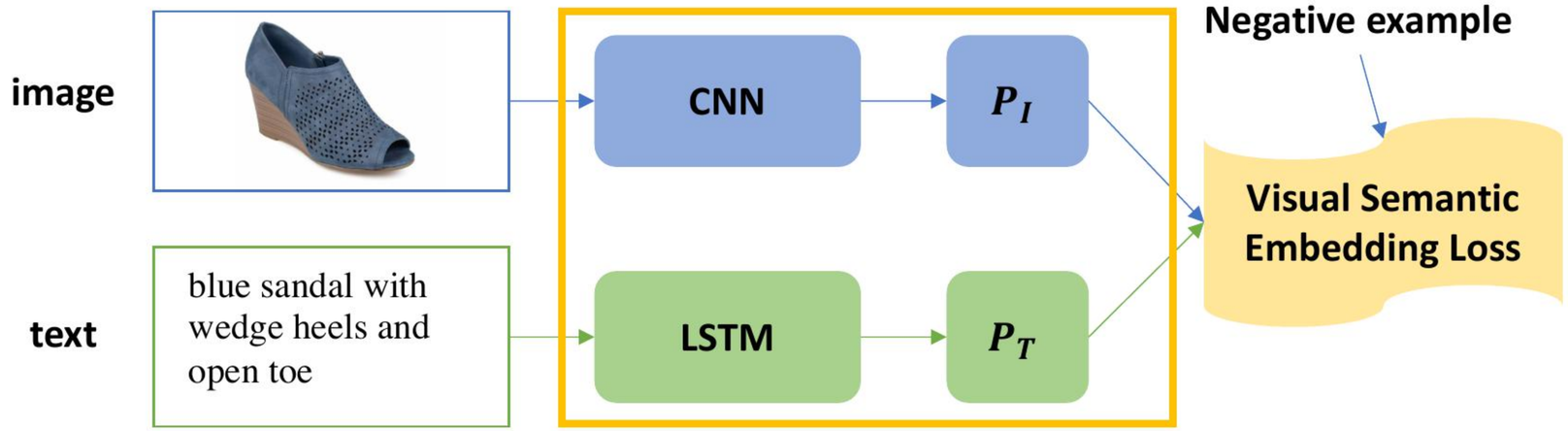
- User-friendly retrieval interfaces should entail the flexibility to ingest **various forms of information** such as images and textual descriptions/modifications
- Core technology for **improving the online shopping experience** via shopping assistants

Language-guided Retrieval: Aim to **modify** and **narrow down** the retrieval results given textual user feedback

Proposal

- Joint Visual Semantic Matching (JVSM) is a **simple** and **effective** model with composite loss functions
- JVSM is **flexible**, it…
  - learns a visual semantic embedding space shared by **image and text**
  - learns the mapping functions that allow to **compose** image and modified text for refining image retrieval results
- JVSM can be trained with **privileged information** (available only at training time)
- JVSM can perform **language-guided retrieval tasks** as well as **image-text matching tasks**

Visio-linguistic Embeddings - Model

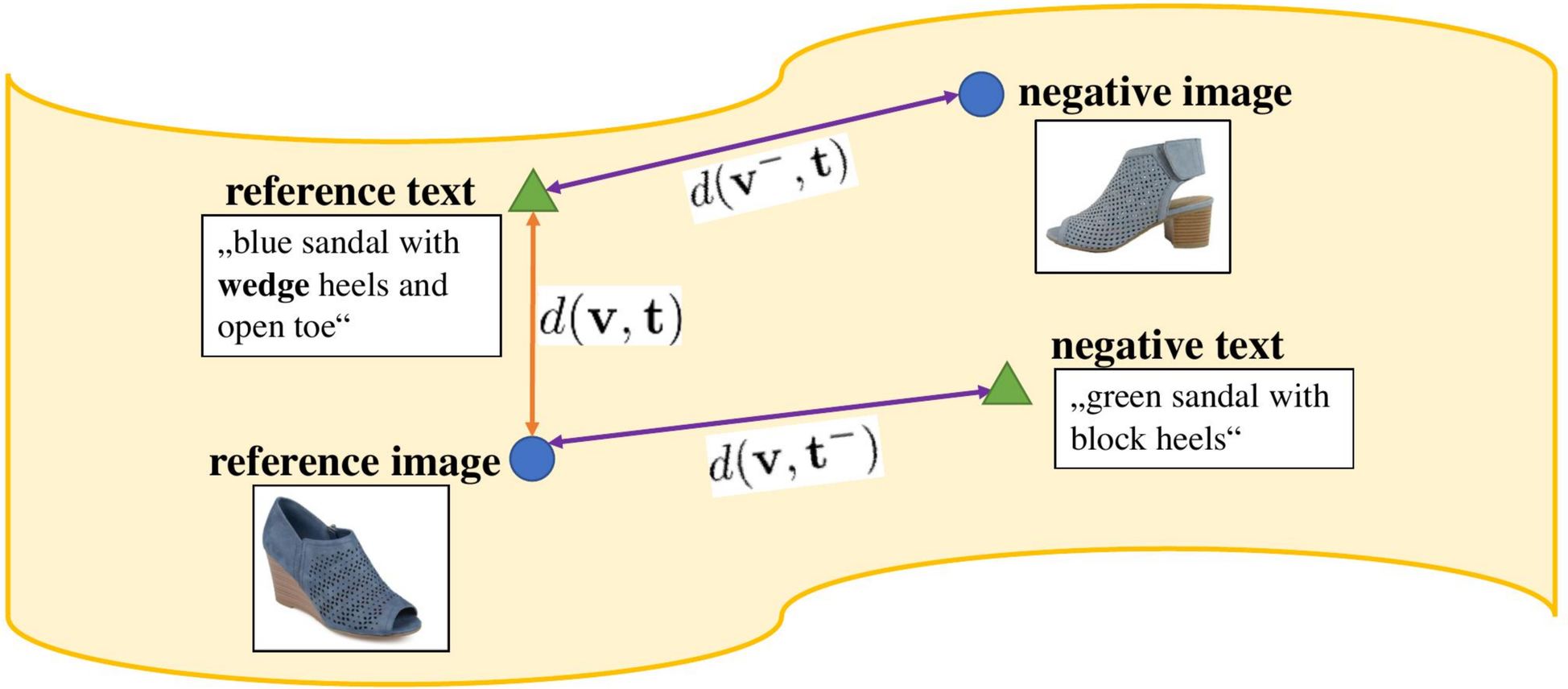


Visual Semantic Embedding Loss [VSE]:

$$L_{vse} = [d(\mathbf{v}, \mathbf{t}) - d(\mathbf{v}, \mathbf{t}^-) + m]_+ + [d(\mathbf{v}, \mathbf{t}) - d(\mathbf{v}^-, \mathbf{t}) + m]_+$$

[VSE] Liwei Wang, Yin Li, and Svetlana Lazebnik. Learning deep structure-preserving image-text embeddings. CVPR 2016.

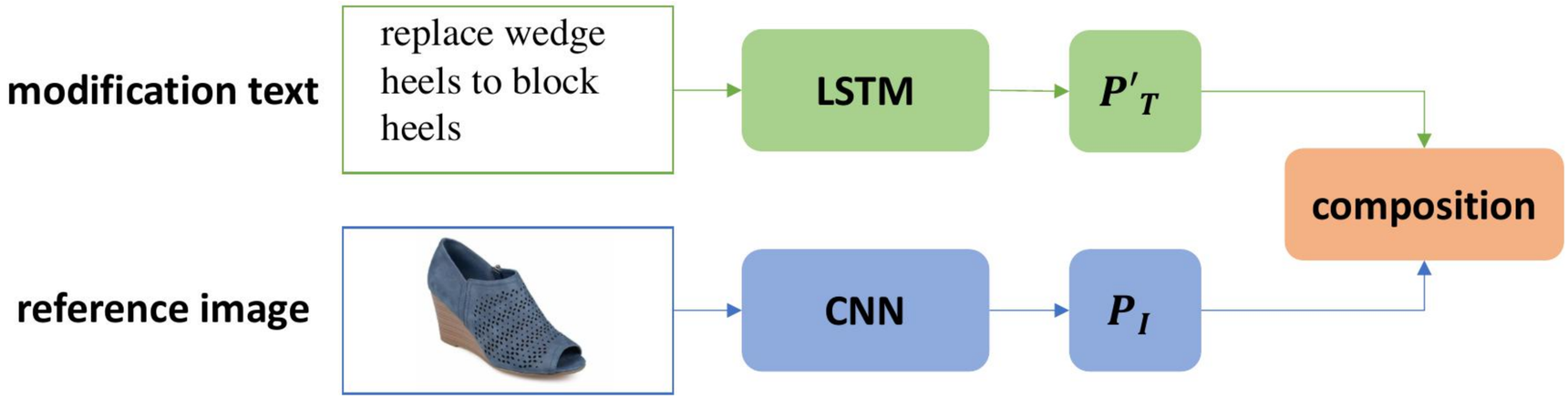
Visio-linguistic Embeddings - Intuition



$$L_{vse} = [d(\mathbf{v}, \mathbf{t}) - d(\mathbf{v}, \mathbf{t}^-) + m]_+ + [d(\mathbf{v}, \mathbf{t}) - d(\mathbf{v}^-, \mathbf{t}) + m]_+$$

Liwei Wang, Yin Li, and Svetlana Lazebnik. Learning deep structure-preserving image-text embeddings. CVPR 2016.

Compositional Embeddings - Model



[TIRG] as learnable composition model (gating + residual):

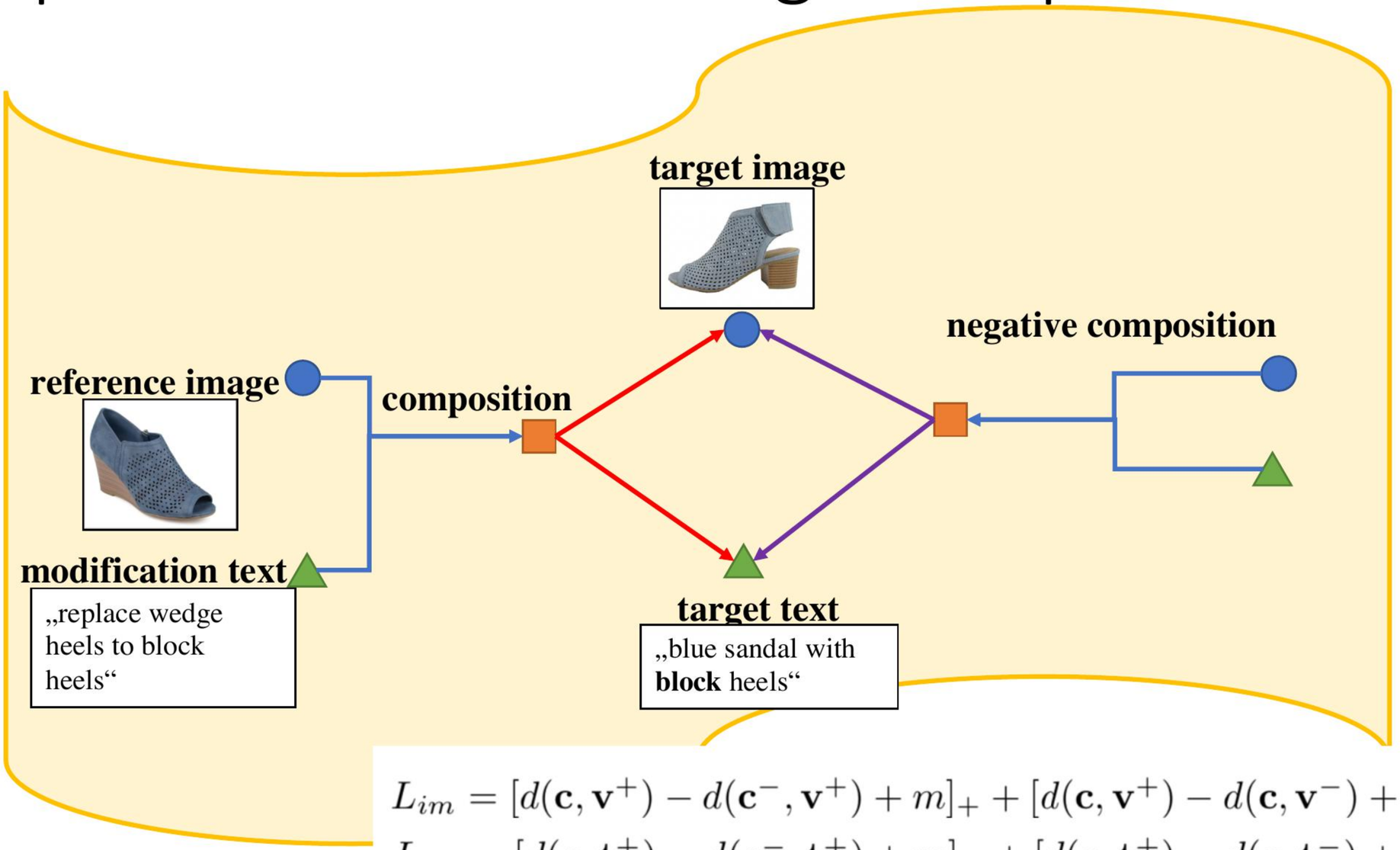
$$\phi_{xt}^{rg} = w_g f_{\text{gate}}(\phi_x, \phi_t) + w_r f_{\text{res}}(\phi_x, \phi_t)$$

$$f_{\text{gate}}(\phi_x, \phi_t) = \sigma(W_{g2} * \text{RELU}(W_{g1} * [\phi_x, \phi_t]) \odot \phi_x$$

$$f_{\text{res}}(\phi_x, \phi_t) = W_{r2} * \text{RELU}(W_{r1} * ([\phi_x, \phi_t]))$$

[TIRG] Nam Vo, Lu Jiang, Chen Sun, Kevin Murphy, Li-Jia Li, Li Fei-Fei, and James Hays. Composing text and image for image retrieval - an empirical odyssey. CVPR 2019

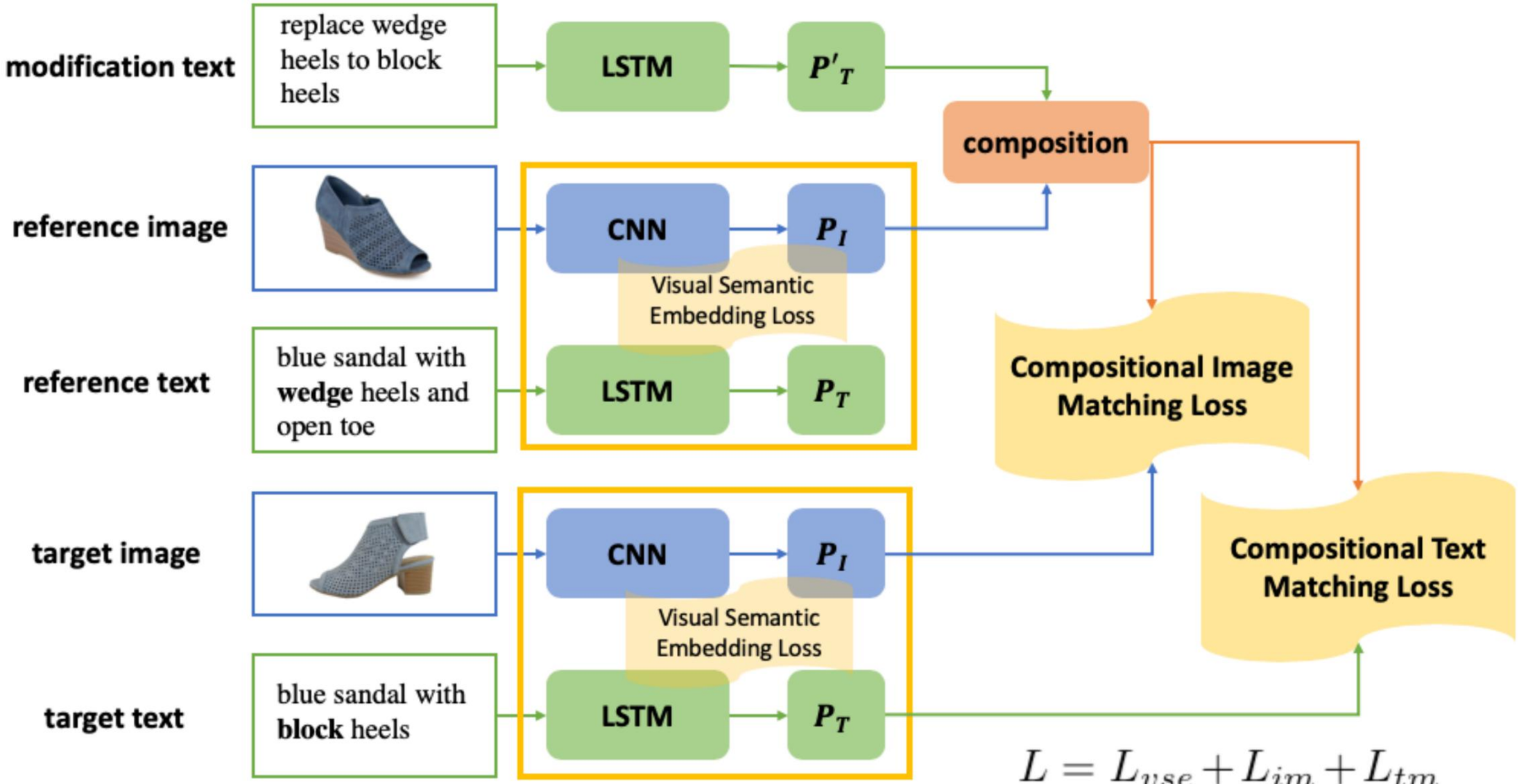
Compositional Embeddings - Proposed Loss



$$L_{im} = [d(\mathbf{c}, \mathbf{v}^+) - d(\mathbf{c}^-, \mathbf{v}^+) + m]_+ + [d(\mathbf{c}, \mathbf{v}^+) - d(\mathbf{c}, \mathbf{v}^-) + m]_+$$

$$L_{tm} = [d(\mathbf{c}, \mathbf{t}^+) - d(\mathbf{c}^-, \mathbf{t}^+) + m]_+ + [d(\mathbf{c}, \mathbf{t}^+) - d(\mathbf{c}, \mathbf{t}^-) + m]_+$$

Proposed Model



Results

Fashion-200k

Method	R@1	R@10	R@50
Han et al. [11]	6.3	19.9	38.3
Show and Tell [34]	12.3	40.2	61.8
Relationship [28]	13.0	40.5	62.4
FiLM [25]	12.9	39.5	61.9
TIRG [36]	14.1	42.5	63.8
TIRG* [36]	15.1	41.9	62.0
JVSM (ours)	**19.0**	**52.1**	**70.0**

UT-Zap50k

Method	R@1	R@10	R@50
TIRG* [36]	4.5	25.4	56.4
JVSM (ours)	**10.6**	**37.1**	**63.5**

Fashion-Iq

Method	Dress		Shirt		Toptee	
	R@10	R@50	R@10	R@50	R@10	R@50
TIRG* [36]	7.3	18.1	10.1	21.8	10.5	23.8
1-turn [10]	7.7	23.9	5.0	17.3	5.2	17.3
JVSM (ours)	**10.7**	**25.9**	**12.0**	**27.1**	**13.0**	**26.9**

Qualitative Results (1)

Please join our Q&A live session.
Thanks!