

# Supplementary Materials: Semantic Distillation from Neighborhood for Composed Image Retrieval

In this supplementary material, we provide more details of the implementation and analysis of the proposed SADN model, which are unable to be elaborated in the main submission due to the space limitations.

## 1 DISTRIBUTION VISUALIZATION OF FEATURE EMBEDDINGS

To investigate the effect of SADN in semantic alignments across query and target domains, we utilize the tSNE tool [1] to visualize the semantic embeddings randomly sampling from categories in FashionIQ and CIRR in Figure 1, where yellow dots represent the query features and blue dots represent target features. The comparisons between the models without the neighbor-aware aggregation module and our SADN show the refined query features output by SADN and target features are mixed together and no distinct gaps could be detected, which demonstrates that SADN could effectively distill distribution characteristics from the target domain to be fused on the query representations. This improvement could be ascribed to the residual connection on the query features and adaptive weight aggregation based on semantics to maintain the primary semantic from the query and strengthen the correlated component from the neighbor features in the target domain. Note that the distribution impacts on CIRR are different from those on FashionIQ and show clustering effects, which are attributed to the diverse visual objects and various types of modifiers of CIRR. The model with the neighborhood aggregation decreases the modality gap between the mixed-modal representations and visual representations.

## 2 FURTHER ANALYSIS

To explore the specific design of the neighbor-aware aggregation in the proposed SADN, we deploy a group of variants of the model to further explore the distance measurement in the divergence-based measurement module, as:

- SADN w Manhattan Dis: Eq. 7 is replaced by  $u_i^j = \sqrt{\sum^d (N_i^j - Q_i)^2}$ ,
- SADN w Euclidean Dis: Eq. 7 is replaced by  $u_i^j = \sum^d (|N_i^j - Q_i|)$ ,
- SADN w Cosine Dis: Eq. 7 is replaced by  $u_i^j = \frac{\hat{Q}_i^T \cdot N_i^j}{\|\hat{Q}_i\| \|N_i^j\|}$ ,
- DBC w Target: use target representations  $T_i$  to substitute for the  $Q_i$  in Eq. 7.

From Table 1, it can be seen that the model with Manhattan distance measurement has an obvious recall rate drop as this measurement could excessively magnify the numerical differences between the neighbor representations and the query representations rather than explore the latent semantic divergences between these two representations. In terms of Euclidean distance and cosine distance,

their performances are not as competitive as the proposed Mahalanobis distance, and one possible reason could be they do not take the correlations between the dimensions into consideration and they are affected by the disentanglement of features. As for the divergence-based correction measuring the distances between the neighborhood representations and target features instead of distances between neighbor representations and query features, there is no significant improvement by “DBC w Target”. Besides, note that the target guidance is unable during testing and validating, we argue that the proposed distances between neighbor representations and query features is a cost-effective choice for accurately sensing the semantic divergences based on the embedding distances.

**Table 1: Ablation experiments of different measurements of divergences. Best results are marked in bold.**

Models	R@10			Mean
	Dress	Shirt	Toptee	
SADN w Manhattan Dis	38.82	43.82	46.91	43.18
SADN w Euclidean Dis	39.51	43.62	47.16	43.44
SADN w Cosine Dis	39.46	<b>43.87</b>	47.01	43.45
DBC w Target	39.91	43.48	47.88	43.76
SADN	<b>40.01</b>	43.67	<b>48.04</b>	<b>43.91</b>

## 3 CASE STUDY

Figure 3 shows examples based on the ranking of the similarity scores by our SADN and “SADN w/o NAA” on FashionIQ and CIRR. The ground-truth target images are marked in red boxes. After integrating the neighbor-aware aggregation, the rankings of the positive images are improved and the issues of the unbalanced concentrations on reference images and modification texts are reduced. As the cases in the FashionIQ show, some false negative samples also have high ranks. Based on our proposal, they are regarded as neighbor instances and are aggregated on the raw query features adaptively based on the semantic correlations and divergence corrections, which could be conducive to extracting the semantics of users’ queries reflected in the target domain. As seen in the second example in Figure 3, though the ground-truth ranks top in the primitive ranking results, our SADN elevates the rankings of other correlated candidate images depicting long-sleeved shirts with horizontal striped patterns and no buttons to be in line with the users’ requirements. In terms of CIRR dataset, due to the subset configuration, more visually similar images, e.g., pictures with one dog with an open mouth in the fourth example, may confuse the model without the NAA module, whereas the whole SADN could capture the accurate semantic distinctions between the hard negative instances and the positive target image. This improvement could be ascribed to the divergence-based correction to enhance the awareness of the semantic differences.

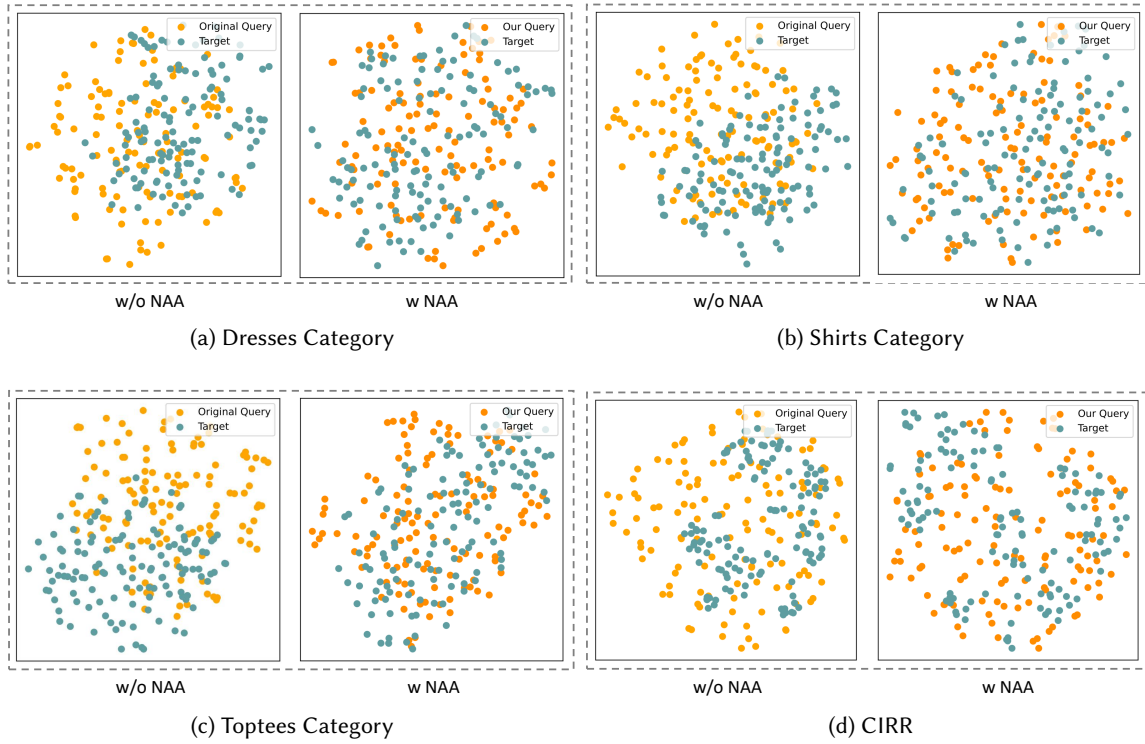


Figure 1: Distribution Comparison in FashionIQ and CIRR datasets. “w/o NAA” means the model without the neighbor-aware aggregation module. “w NAA” refers to SADN after aggregating the neighbor-aware aggregation module.

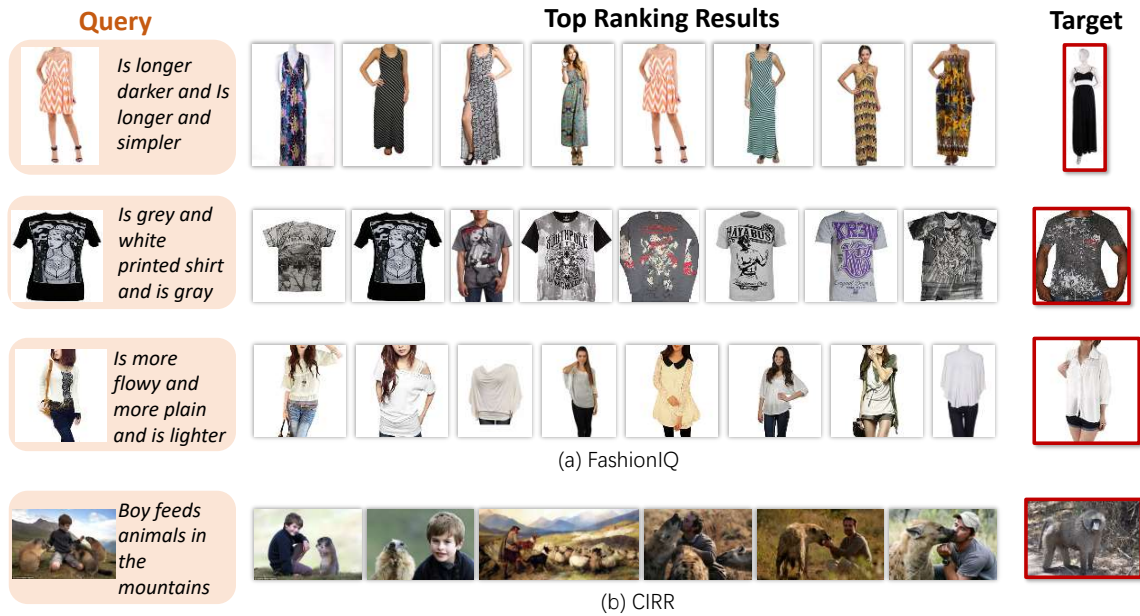


Figure 2: Failure cases of the proposed SADN on FashionIQ and CIRR datasets.



**Figure 3: Case study of the proposed SADN on FashionIQ and CIRR datasets. “SADN w/o NAA” means the model without the neighbor-aware aggregation module.**

Figure 2 also displays some failure cases on FashionIQ and CIRR datasets, where target images have fallen to low rankings. The top retrieval results usually are related to partial semantics in the hybrid-modal queries or could be regarded as false negatives given the queries of the reference image and text modifications. Besides, the annotations of textual modifiers in the query sometimes are vague and could not direct to the target examples (as seen in the example in the CIRR), which may affect the evaluation metrics.

Apart from the noisy annotations, the performances could be limited by the primitive compositor to some extent. A more effective way to combine the reference image and modifiers to obtain the representative query features could be explored in the future.

## REFERENCES

- [1] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing Data using t-SNE. *Journal of Machine Learning Research* 9, 86 (2008), 2579–2605.