3.7 Experiments and Discussions

In this section, we examine our design using objective measures. We adopt the ANMRR accuracy metric mentioned in Sec. 2.1.3. Then, the simulation environment and conditions are described in Sec. 3.7.1. The simulation results are summarized in Sec. 3.7.2. Finally, we modify the proposed method to reduce its complexity. Also, our scheme is compared against with two other schemes in Sec. 3.7.1.

3.7.1 Simulation Environment

In our previous work [35][36], we have conducted a preliminary experiment to evaluate the proposed method against a 1050-image database. The results show that the multi-instance user perception weighting method is promising, and the pruning concept always improves the query accuracy in our method; also, the pseudo-image concept improves the accuracy in many cases.

We then extend the evaluation process to a much larger scale in [37][38]. The database consists of 18433 images including 256 test (ground-truth) images, 194 people (party) photos, 200 flower pictures, 200 undersea pictures, 200 outdoor scenery pictures, and 17383 images from the Corel gallery.

We collect 38 sets of outdoor scenic images as the ground truth. They are similar in terms of low-level descriptions. We prepare the ground-truth images as follows: each set of ground-truth images is taken on the same spot with slightly different camera pan and tilt angles by hand. The size of a ground-truth set varies from 4 to 16. Images in each set are perceptually similar. However, by examining the low-level features, we observe that the feature values can be quite different. There are several possible causes. One is that these pictures are taken by hands. They are inevitably somewhat shifted and blurred. The other is that different shots have slightly different focus and shutter speed. Another is that photos with shooting angle variation may have different background lighting, which may change the shade of each picture.

Our experiments simulate a typical image query scenario. A user first chooses one or a few "similar" input images to start a query. The matching process returns an ordered list of results; we call it the positive-only query result. If the result is not perfect; that is, not all ground-truth images occupy the highest ranks, or simply $NMRR \neq 0$, then the highest ranked non-ground-truth image is assigned as the negative feedback item. Then, we repeat the query process with both positive and negative images and produce the positive-and-negative query result. If the positive-only result is perfect, both $NMRR_{positive-only}$ and $NMRR_{positive-and-negative}$ are set to zero. Since the smallest ground truth set has only four images, we simulate the conditions of one to three positive images per query. All possible combinations of images in all ground truth sets are tested to derive the $ANMRR$ values.

Two multi-scale schemes are tested: spatial and SNR. The spatial scaling factor (for both width and height) for each down-sampled image is defined as follows: the n-th scale factor (for the n-th pseudo image) is $\alpha - 0.1(n-1)$, where $n = 1, 2$. We perform experiments at $\alpha = 0.9, 0.8, 0.7, 0.6, 0.5$ to look for the best parameter values that lead to the best ANMRR. The SNR-scaled images are generated by applying JPEG compression with a quality factor of $\beta - 0.1(n-1)$ for the n-th scaled version. The test values are $\beta = 0.7, 0.6, 0.5, 0.4, 0.3$.

To examine the effectiveness of our method, we simulate another two weighting schemes under the same assumptions. The first scheme is a variation derived from the MARS system. In this scheme, the distance metric $d_j(f_1, f_2)$ for each feature F_j is normalized as follows:

$$d'_j(f_1, f_2) = \frac{D_j(f_1, f_2) - \mu_j}{3\sigma_j},$$

where μ_j and σ_j^2 are the mean and variance of the distances of F_j in the database. This step ensures that under normal distribution assumption about 99% of the distance values are within the range of $[-1, +1]$. The second parameter-shifting step guarantees that these 99% values are within $[0, 1]$:

$$d''_j(f_1, f_2) = \frac{d'_j(f_1, f_2) + 1}{2}.$$

The final step clamps all calculated distance values between zero and one.

The original MARS system adopts a 5-level relevance feedback. To make it comparable with our simulation environment, we reduce the relevance feedback levels to three: relevant ($Score_l = +1$), no opinion ($Score_l = 0$), and not relevant ($Score_l = -1$). The weighting process is similar to that in the original MARS. Assume the overall query result list is RT, and the result list of feature F_j is RT_j. To calculate the weight w_j, we first initialize $W_j = 0$, and then update W_j as follows:

$$W_j = W_j + Score_l, \text{ for each item } l \text{ which appears in both } RT \text{ and } RT_j.$$

After all W_j have been updated, we compute the weighting factor for each feature F_j as

$$w_j = \frac{W_j}{\sum_{\forall j} W_j}.$$

A final remark about this MARS-like scheme is the RT list. According to the original proposal[15], it is an iterative procudure that leads to the "optimal" RT. The original proposal selects $P_{fd} = 3$ as the maximum number of iterations and shows good convergence in general. In our simulation, we set $P_{fd} = 5$. This is called Scheme A in the rest of this section.

The second scheme we simulate has the same basic structure as our proposed scheme in Sec. 3.3. However, it adopts the same distance normalization in MARS. This is Scheme B. We simulate this scheme for two reasons. One is to compare with a MARS-like scheme to see the effects of different weighting estimation procedures. The other is to compare it with our scheme to see the effects of different distance normalization methods. Our scheme is labeled as Scheme C. We will discuss another scheme in Sec. 3.8, which is similar to Scheme C, and we call it Scheme D.

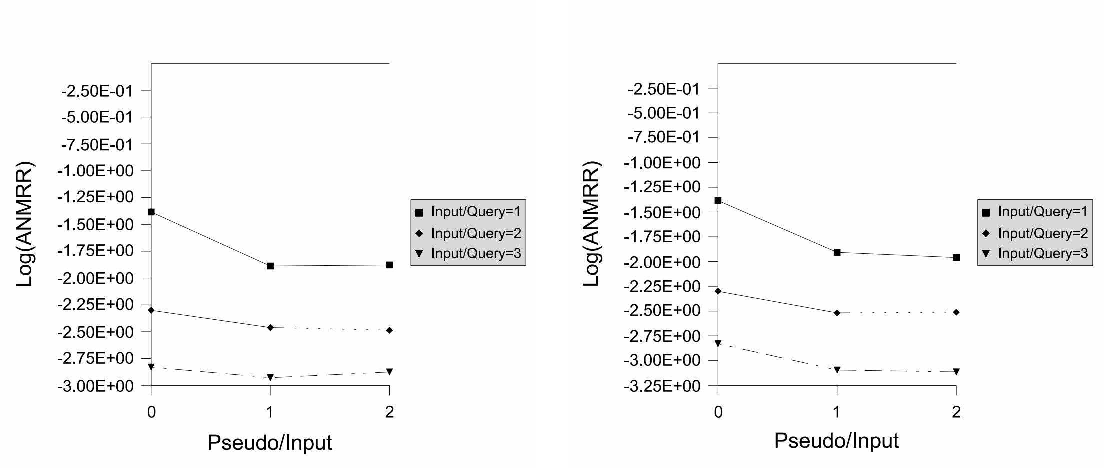
3.7.2 Simulation Results

In this section, we show the simulation of query accuracies, under the environments defined in Sec. 3.7.1. We list the results of the four schemes, **but the discriptions**

about Scheme D are postponed to Sec. 3.8 for clarity. In Sec. 3.7.2, we examine the effects of multi-instance and pseudo-images. Then, we put all simulation statistics together, to compare the efficiency of different matching schemes. The simulation results are summarized in Tables 3.1, 3.2, 3.3, and 3.4. The bold-faced numbers are the winners among all tests with the same query parameters, and the underlined numbers are the poorest performers. The row of *Input/Query* means the number of positive input images selected by the user in a query. The row of "Pseudo/Input" means the number of *pseudo images* created from each (user selected) input image. The first column (Scheme A) is the MARS-like scheme, and the second and third columns (Scheme B and C) are our schemes with different normalization formulas. To see more clearly the differences among various methods and parameters, the ANMRR values are shown in log scale. In the following paragraphs, we will examine these results and discuss the performance of the aforementioned methods.

Multi-instance and Pseudo-images

Here we examine the effect of multi-instances and pseudo-images. Two multi-scale schemes are shown in Fig. 3.10: spatial and SNR scaled pseudo image generation schemes. Figure 3.10(a) is spatial down-sampling with a spatial scaling factor of $\alpha = 0.7$. We examine the effect of different pseudo/input image ratios. Under the same pseudo/input ratio, the more the input images (user provided), the better the query accuracy. For the same number of input images, pseudo images can improve the accuracy, especially when the input images is one or two. However, when input (query) images are higher in number, the addition of pseudo images may lower the matching accuracy. Figure 3.10(b) shows the results of using SNR-scaled pseudo images. The noisy versions (pseudo images) are generated by a scaling factor of $\beta = 0.4$. The general trend of Fig. 3.10(b) is similar to that of Fig. 3.10(a). However, the average ANMRR is better in SNR multi-resolution approach. The other scaling factor values have been tested but the results are less favored.



(a) Spatial-scaled with $\alpha = 0.7$. (b) SNR-scaled with $\beta = 0.4$.

Figure 3.10: Two example simulation results.

Observations on Positive-only Query Results

We first look at the results of positive-only spatial-scaled experiments (Table 3.1). The cases shown here are the spatial scaled pseudo images with the best scaling factors. For each method, multiple input images (all the cases where Pseudo/Input=0) improve the query accuracy. This shows that more "positive" query information would result in better query precision, regardless which scheme is in use. Next, we examine the effect of pseudo images. Our scheme with one pseudo image has the best accuracy in all Input/Query cases. However, the pseudo images do not improve the other two methods as much. Even worse, more pseudo images would degrade the query accuracy. The simulation results also show that increasing pseudo images does not always improve accuracy. Under our current scheme, one pseudo image per input image is the best.

For each query parameter set (Input/Query and Pseudo/Input), we compare the results of different estimation methods. Comparing the MARS-like method (Scheme A) with the Gaussian-normalized method (Scheme B), we observe that when input images are few, the MARS-like scheme wins. In contrast, Scheme B wins when more input images are provided. Since these two methods use the same distance normalization procedure, the difference comes from the weight computing procedures. When few images are available for estimation, iterative training would provide a better guess on the user perception. When more images are provided by a user, ranking-list-based MARS-like scheme does not provide as precise guess as distance-based Gaussian-normalized scheme. Comparing the Gaussian-normalized scheme (Scheme B) with our scheme (Scheme C), the former wins when input images are few and loses when more images are provided. The two methods use the same distance definition and the estimation procedure, so the difference comes from the distance normalization procedures. It is reasonable that the Gaussian-normalized scheme wins for few inputs cases, because the distance metrics are optimized according to the data distribution. This implicitly provides clustering information of the database, and thus produces better results than our method. However, feature distributions in a database may not be the same as the distance distribution viewed from the user perception for a particular query. This may explain why our method wins when more input images are provided. Our user perception (intention) estimation is based only on the user provided information (not the entire database).

We conduct the same analysis on the SNR-scaled case (Table 3.2), and similar conclusions can be drawn. However, there are two noticeable differences. The first one is that in several test cases, Pseudo/Input=2 outperforms Pseudo/Input=1 in the SNR-scaled case. The second is that in most cases, the SNR-scaled pseudo images outperforms the spatial-scaled ones.

Observations on Positive-and-Negative Query Results

Next, we look into the Positive-and-Negative Query cases. As mentioned in Sec. 3.7.1, the simulation is done using the typical query scenario. For each positive-and-negative query, there is zero or one negative image depending on whether the positive-only query is a perfect match or not. Similar to what we did in Sec. 3.7.2, we first examine the simulation results of positive-and-negative feedback with spatial-

Table 3.1: Best $log(ANMRR)$ of spatial-scaled pseudo images (positive-only).

	Scheme A	Scheme B	Scheme C	Scheme D
Input/Query=1				
Pseudo/Input=0	**-2.23**	**-2.23**	-1.38	-1.38
Pseudo/Input=1	**-2.22**	-2.18	-1.94	-1.94
Pseudo/Input=2	**-2.19**	-2.15	-1.93	-1.93
Input/Query=2				
Pseudo/Input=0	-2.40	**-2.45**	-2.30	-2.30
Pseudo/Input=1	-2.40	-2.45	**-2.49**	-2.51
Pseudo/Input=2	-2.33	-2.37	**-2.49**	-2.50
Input/Query=3				
Pseudo/Input=0	-2.40	-2.48	**-2.83**	-2.83
Pseudo/Input=1	-2.41	-2.48	**-2.93**	-2.92
Pseudo/Input=2	-2.33	-2.38	**-2.93**	-2.92

Table 3.2: Best $log(ANMRR)$ of SNR-scaled pseudo images (positive-only).

	Scheme A	Scheme B	Scheme C	Scheme D
Input/Query=1				
Pseudo/Input=0	**-2.23**	**-2.23**	-1.38	-1.38
Pseudo/Input=1	**-2.19**	-2.18	-1.95	-1.95
Pseudo/Input=2	**-2.18**	**-2.18**	-2.00	-1.99
Input/Query=2				
Pseudo/Input=0	-2.40	**-2.45**	-2.30	-2.30
Pseudo/Input=1	-2.45	-2.49	**-2.52**	-2.52
Pseudo/Input=2	-2.47	-2.49	**-2.52**	-2.52
Input/Query=3				
Pseudo/Input=0	-2.40	-2.48	**-2.83**	-2.83
Pseudo/Input=1	-2.54	-2.63	**-3.09**	-3.07
Pseudo/Input=2	-2.63	-2.71	**-3.11**	-3.10

scaled pseudo images (Table 3.3). For all schemes, multiple input images improve the accuracy. Effects of pseudo images are similar to that of the positive-only results. Our method seems to be able to utilize pseudo images better for improving the accuracy. For the other two schemes, pseudo images do not provide significant improvements. The ANMRR values show that the Pseudo/Input=1 cases give the most significant improvement. Additional pseudo images offer much less improvement if any.

Comparing Scheme A (MARS-like scheme) with Scheme B (Gaussian-normalized scheme), we found that the Gaussian-normalized scheme wins in most cases. Our explanation is that in our proposed procedure, the negative feedback does not participate in weights estimation. Since the negative instances may be too diverse to be useful in weights estimation, their role are more appropriate when used in pruning. The simulation results seems to prove this concept. Comparing Scheme C (our scheme) with Scheme B (Gaussian-normalized scheme), the results show that ours wins when sufficient input images are available. The ANMRR values show that the best accuracy is the Input/Query=3 case in our scheme. The reason is that the pruning distance relies on the estimated distance function. Thus, the more precise distance function would lead to a lower “mis-pruning” probability.

The ANMRR values shown in Table 3.4 for the SNR-scaled pseudo images lead to similar conclusions as before. First, multiple input instances improve query accuracy. Second, our method benefits more from the pseudo images. Third, the Gaussian-normalized scheme (Scheme B) wins in almost all cases when comparing to the MARS-like scheme (Scheme A). Fourth, our method (Scheme C) performs better than the Gaussian-normalized when more input images are available. Finally, our method has a significant performance improvement at Input/Query=3, which indicates a good potential of our approach for even more input images.

Table 3.3: Best $log(ANMRR)$ of spatial-scaled pseudo images (positive-and-negative).

	Scheme A	Scheme B	Scheme C	Scheme D
Input/Query=1				
Pseudo/Input=0	-1.97	**-2.12**	-1.39	-1.39
Pseudo/Input=1	-2.07	**-2.21**	-1.97	-1.97
Pseudo/Input=2	-2.06	**-2.22**	-2.00	-1.99
Input/Query=2				
Pseudo/Input=0	**-2.67**	-2.61	-2.40	-2.40
Pseudo/Input=1	-2.65	**-2.71**	-2.63	-2.65
Pseudo/Input=2	**-2.64**	-2.62	-2.60	-2.65
Input/Query=3				
Pseudo/Input=0	-2.72	-2.76	**-3.09**	-3.09
Pseudo/Input=1	-2.73	-2.76	**-3.22**	-3.21
Pseudo/Input=2	-2.62	-2.63	**-3.21**	-3.23

Table 3.4: Best $log(ANMRR)$ of SNR-scaled pseudo images (positive-and-negative).

	Scheme A	Scheme B	Scheme C	Scheme D
Input/Query=1				
Pseudo/Input=0	-1.97	**-2.12**	-1.39	-1.39
Pseudo/Input=1	-1.95	**-2.17**	-2.03	-2.03
Pseudo/Input=2	-1.94	**-2.17**	-2.04	-2.06
Input/Query=2				
Pseudo/Input=0	**-2.67**	-2.61	-2.40	-2.40
Pseudo/Input=1	-2.60	**-2.71**	-2.67	-2.67
Pseudo/Input=2	-2.63	**-2.73**	-2.66	-2.66
Input/Query=3				
Pseudo/Input=0	-2.72	-2.76	**-3.09**	-3.09
Pseudo/Input=1	-2.89	-2.96	**-3.47**	-3.49
Pseudo/Input=2	-2.99	-3.05	**-3.49**	-3.51

Observations on Different Feedback Schemes

In Sec. 3.7.2 and Sec. 3.7.2, we discuss the effects of different weights estimation methods in each specific scheme. In this section, we will discuss the general effect of negative instances and the generation of pseudo images.

Negative instances are important, because they tell us about the “undesired” image properties (or image feature values). That is, the user does not want pictures similar to a negative image. However, the negative images do not provide information about a particular feature whether it is good for matching purpose or not. Two negative images can be close or far away, but positive images should always be close together on the user preferred features. The simulation results show that negative feedback improves query accuracy in many cases, especially when enough positive instances are given. If the number of input instances is small, only our method can consistently improve the accuracy using the negative instances.

Although both multi-scale schemes that generate pseudo images can enhance the query accuracy (especially for our method), we notice that the SNR multi-scaled images not only produces better performance than the spatially scaled ones, they also have consistently improved results. This may be due to the fact that the spatial-scaled images suffer from the aliasing effect when pictures are down-sampled and thus image features are distorted more than those of the SNR-scaled ones. Overall, Scheme C significantly improves the query accuracy by combining multi-instance and pseudo-image concepts.

3.8 Another Distance Measure

In Sec. 3.3, we assume the matching function produces distances that satisfy triangular inequality. This may not be necessary because a CBIR system may adopt non-linear operations either in the extraction or in the matching process. In this section, we define the scatter number by a statistical approach, which does not rely on the geometry theorems and eliminate the sinusoidal operations. Its calculation is thus simpler.

The assumptions and the conjectures are the same as described in 3.3, except that we do not assume the distances satisfy the triangular inequality. To measure the sparseness of a set of feature points, firstly we define a value $scatt_{ij}$ which represents how far the instance q_i is away from the rest of the query instances:

$$scatt_{ij} = \mu_{ij} + \sigma_{ij},$$

where

$$\begin{aligned} \mu_{ij} &= \frac{1}{n-1} \sum_{k=1,k\neq i}^{n} d_j(f_{ij}, f_{kj}) \\ \sigma_{ij}^2 &= \frac{1}{n-1} \sum_{k=1,k\neq i}^{n} (d_j(f_{ij}, f_{kj}))^2 - \mu_{ij}^2. \end{aligned}$$

The $scatt_{ij}$ is similar to the average distance, except that it includes the variation information. The second term (standard deviation) is added into this measure because experiments indicate that an “inconsistent” feature (large standard deviation) is less important.

Then we express the scatter number in a conservative way: calculate the closeness between the given instance and any other point in the set and pick up the maximum; that is,

$$s_j = \max_{\forall i} scatt_{ij}.$$

Note that in this method, the normalization factor is also canceled in each term of the weighted distance function.

This weighting scheme is called scheme D in the previous section. The simulation results are shown in Table 3.5. By comparing to scheme C in Table 3.1 to 3.4, we may see that these two schemes have similar accuracy but scheme D may have the computational advantage. The bold-faced values represent the better ANMRR in scheme D; while the underlined values represent the better ANMRR in scheme C. Some more details of this distance can be found in our previous report [37].

Table 3.5: Best $Log(ANMRR)$ of Scheme D for all Query Schemes.

	Spatial scaling without negative	SNR scaling without negative	Spatial scaling with negative	SNR scaling with negative
Input/Query=1				
Pseudo/Input=0	-1.38	-1.38	-1.39	-1.39
Pseudo/Input=1	-1.94	-1.95	-1.97	-2.03
Pseudo/Input=2	-1.93	-1.99	-1.99	**-2.06**
Input/Query=2				
Pseudo/Input=0	-2.30	-2.30	-2.40	-2.40
Pseudo/Input=1	**-2.51**	-2.52	**-2.65**	-2.67
Pseudo/Input=2	**-2.50**	-2.52	**-2.65**	-2.66
Input/Query=3				
Pseudo/Input=0	-2.83	-2.83	-3.09	-3.09
Pseudo/Input=1	-2.92	-3.07	-3.21	**-3.49**
Pseudo/Input=2	-2.92	-3.10	**-3.23**	**-3.51**

3.9 ANMRR and Precision/Recall Comparisons

In this section, we show a comparison between the ANMRR and the Precision/Recall indexes. Table 3.6 and Table 3.7 list the $log(ANMRR)$ values of Schemes A (MARS-like), B (Gaussian-normalized), and D (our scheme), without and with negative feedback respectively. Here we use the configuration of SNR-scaled pseudo image generation with $\beta = 0.4$. In the precision and recall plots, the horizontal axis denotes the number of retrieved images $|A(q)|$. Roughly, a value of ANMRR corresponds to one precision and one recall curve. As described in Sec. 2.1.3, recall values are not meaningful when the number of relevant images is greater than that of the retrieved images. Since our largest ground-truth set contains 16 images, we plot precision and recall curves with more than 16 retrieved images.

First, we examine the positive-only case (Table 3.6). For Scheme A with input/query=1, ANMRR shows that more pseudo images degrade the query accuracy. The corresponding precision/recall curves are shown in the top-row of Fig. 3.11. The precision values are fairly close (the difference is around 10^{-4}) and almost cannot tell by the plotted curves. By examining the recall curves, we see that no-pseudo case is better than 1-pseudo and 2-pseudo cases; and 1-pseudo and 2-pseudo have close recall rates. The result is consistent to the $ANMRR$ index ($log(ANMRR)$ -2.23, -2.13, and -2.10). If we go through Fig. 3.11 to Fig. 3.13, we can find similar comparison results.

Second, we examine the positive-and-negative case (Table 3.7). We use this case to illustrate the effectiveness of different schemes. Take input/query=1 and pseudo/input=0 as an example, the precision curves (Fig. 3.14) show that Schemes A and B have close performance and are better than Scheme D. The recall curves also show the same trend. In both precision and recall plots, Scheme A is slightly better than Scheme B. However, ANMRR values of the two schemes show the opposite result. This may be due to the definition of these two indexes. Because ANMRR includes the "rank", a high-rank match may cover the performance loss cased by

Table 3.6: $log(ANMRR)$ of SNR-scaled ($\beta = 0.4$) pseudo images (positive-only).

	Scheme A	Scheme B	Scheme D
Input/Query=1			
Pseudo/Input=0	-2.23	-2.23	-1.38
Pseudo/Input=1	-2.13	-2.12	-1.91
Pseudo/Input=2	-2.10	-2.09	-1.98
Input/Query=2			
Pseudo/Input=0	-2.40	-2.45	-2.30
Pseudo/Input=1	-2.45	-2.49	-2.52
Pseudo/Input=2	-2.47	-2.49	-2.52
Input/Query=3			
Pseudo/Input=0	-2.40	-2.48	-2.83
Pseudo/Input=1	-2.53	-2.63	-3.06
Pseudo/Input=2	-2.63	-2.71	-3.10

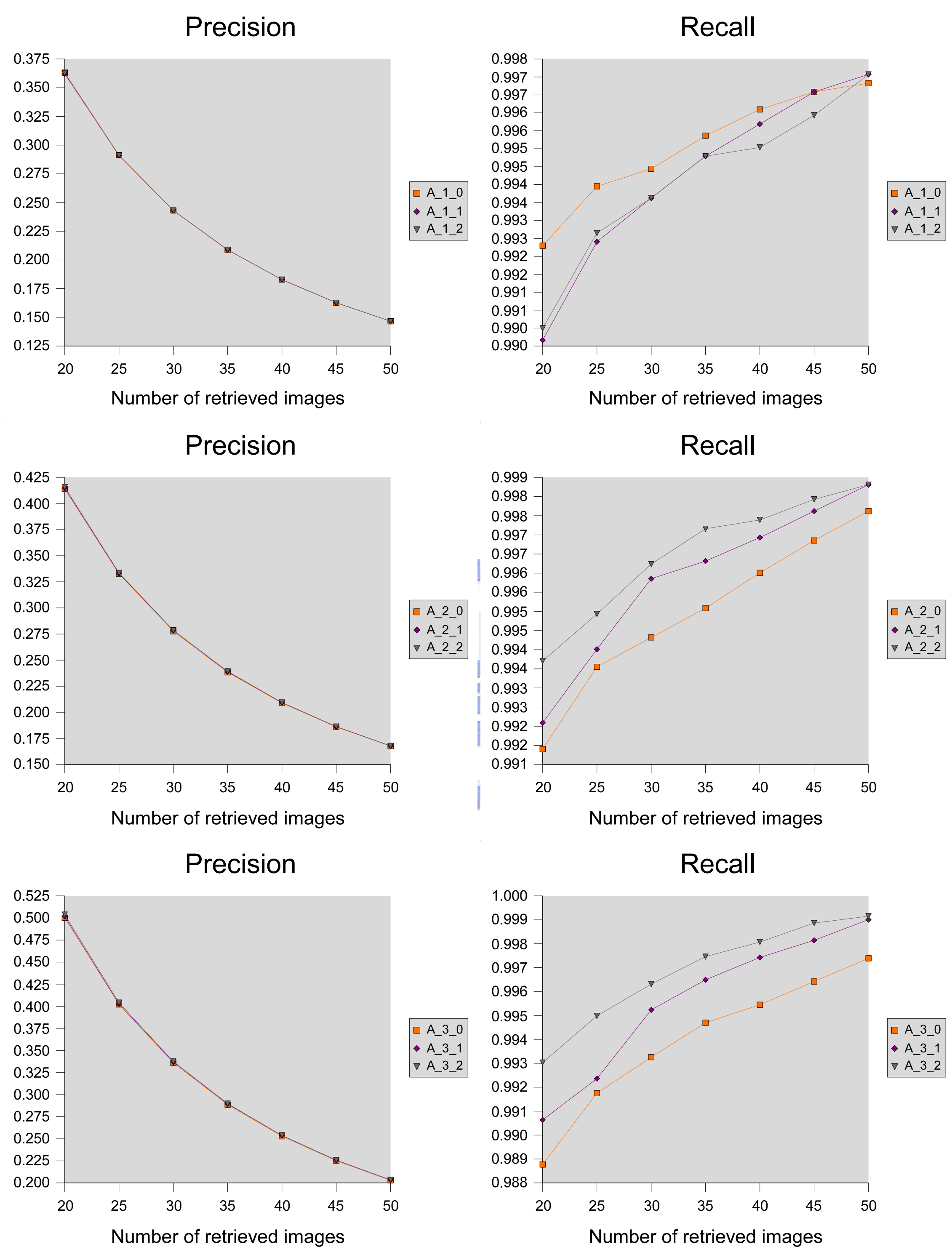


Figure 3.11: Precision and recall curves of positive-only query of Scheme A. The curves are labeled as A_[input/query]_[pseudo/input].

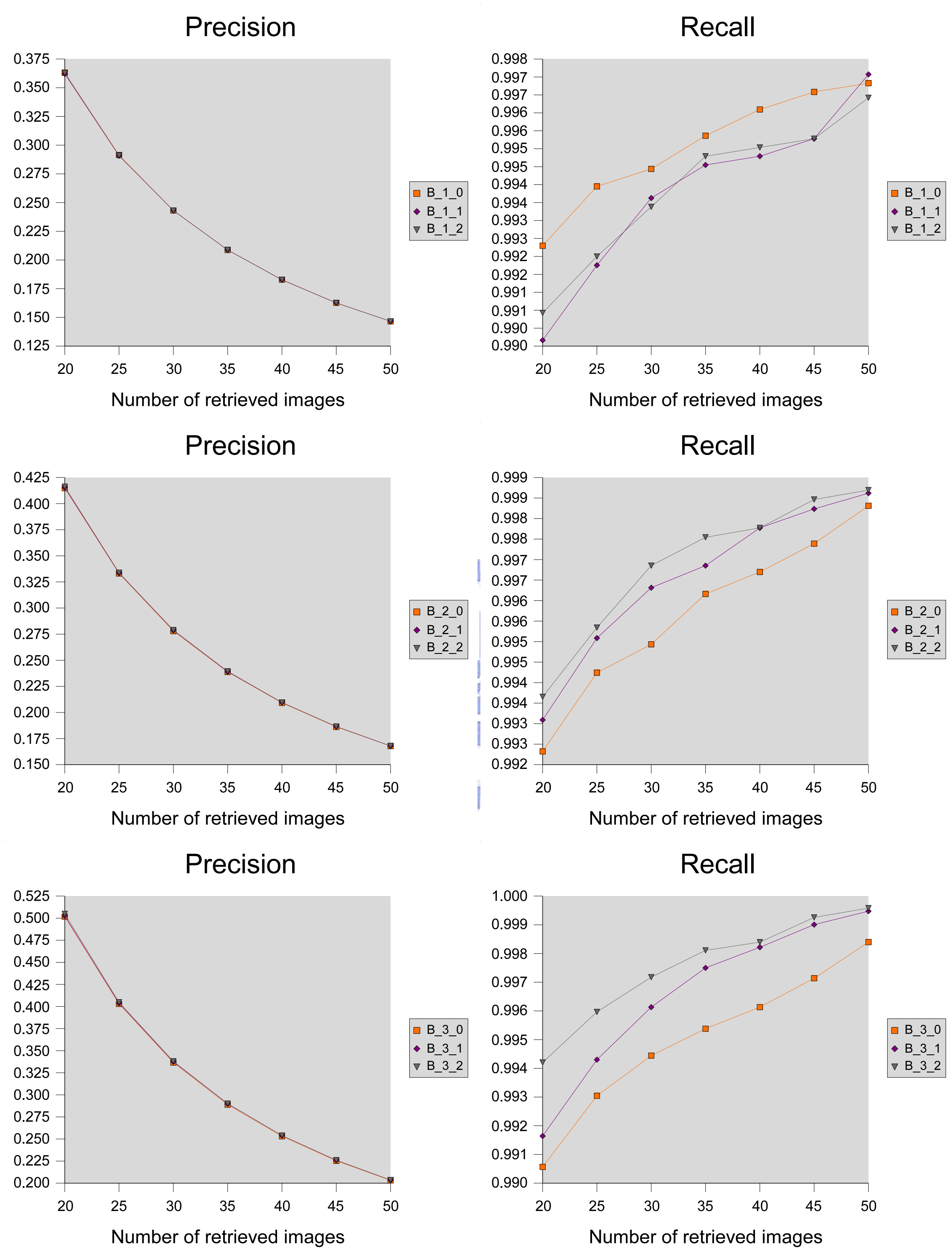


Figure 3.12: Precision and recall curves of positive-only query of Scheme B. The curves are labeled as B_[input/query]_[pseudo/input].

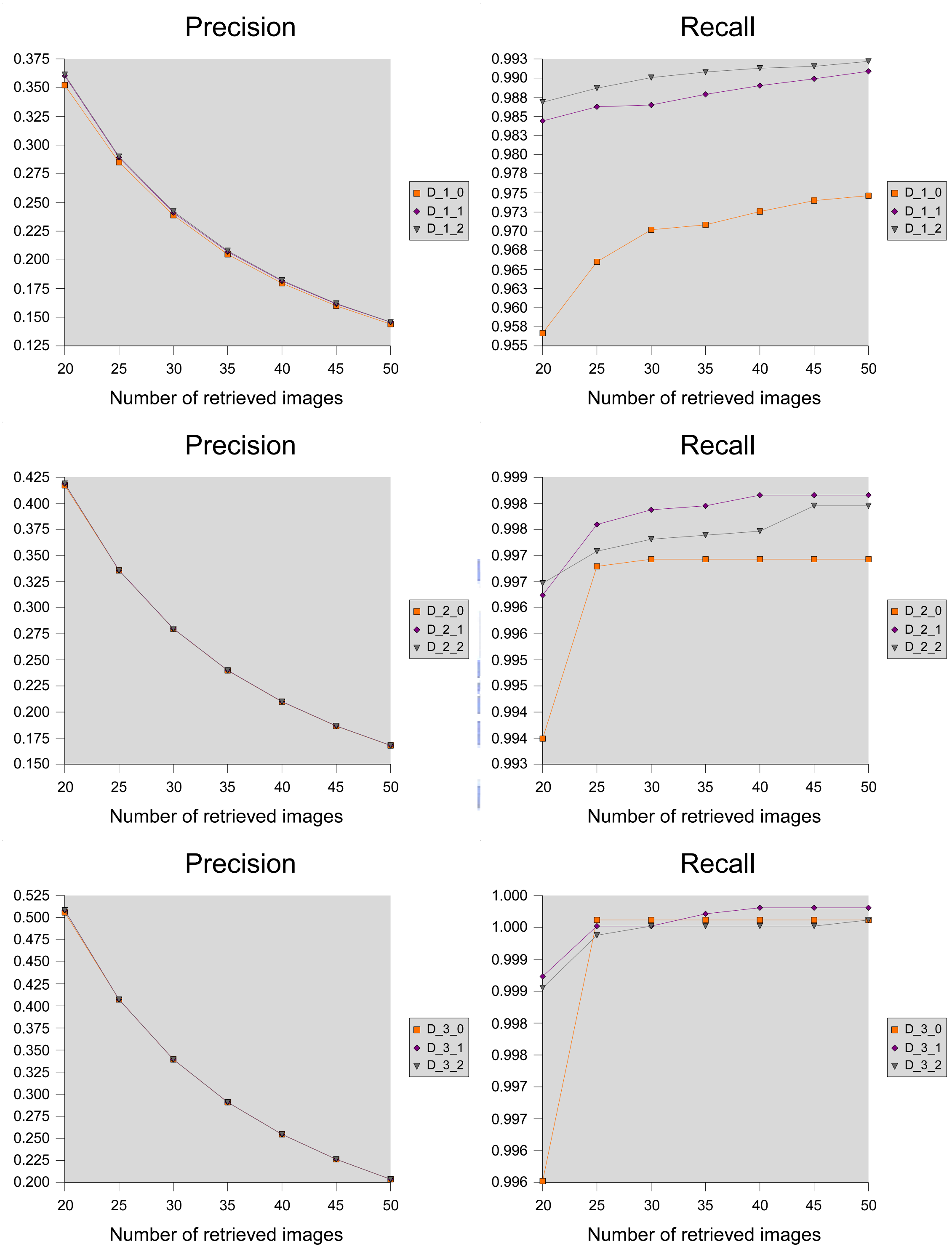


Figure 3.13: Precision and recall curves of positive-only query of Scheme D. The curves are labeled as D_[input/query]_[pseudo/input].

Table 3.7: $log(ANMRR)$ of SNR-scaled ($\beta = 0.4$) pseudo images (positive-and-negative).

	Scheme A	Scheme B	Scheme D
Input/Query=1			
Pseudo/Input=0	-1.97	-2.12	-1.39
Pseudo/Input=1	-1.88	-2.12	-1.97
Pseudo/Input=2	-1.89	-2.11	-2.02
Input/Query=2			
Pseudo/Input=0	-2.67	-2.61	-2.40
Pseudo/Input=1	-2.60	-2.71	-2.67
Pseudo/Input=2	-2.61	-2.73	-2.64
Input/Query=3			
Pseudo/Input=0	-2.72	-2.76	-3.09
Pseudo/Input=1	-2.89	-2.96	-3.49
Pseudo/Input=2	-2.99	-3.05	-3.41

a miss. Since Schemes A and B have close accuracy in this case, it is possible that precision/recall and ANMRR show different comparison results. If we examine Fig. 3.14 to Fig. 3.16 and focus on the performance improvement of Scheme D, we can see that:

- The precision of Scheme A, B, and D are generally close.
- The recall curves of Scheme D are better than the others, when the input instances are higher.

The above trend is consistent to what ANMRR indexes show.

To sum up, although ANMRR and precision/recall may show occasionally different performance results, they are generally consistent, especially when the performance improvement is significant.

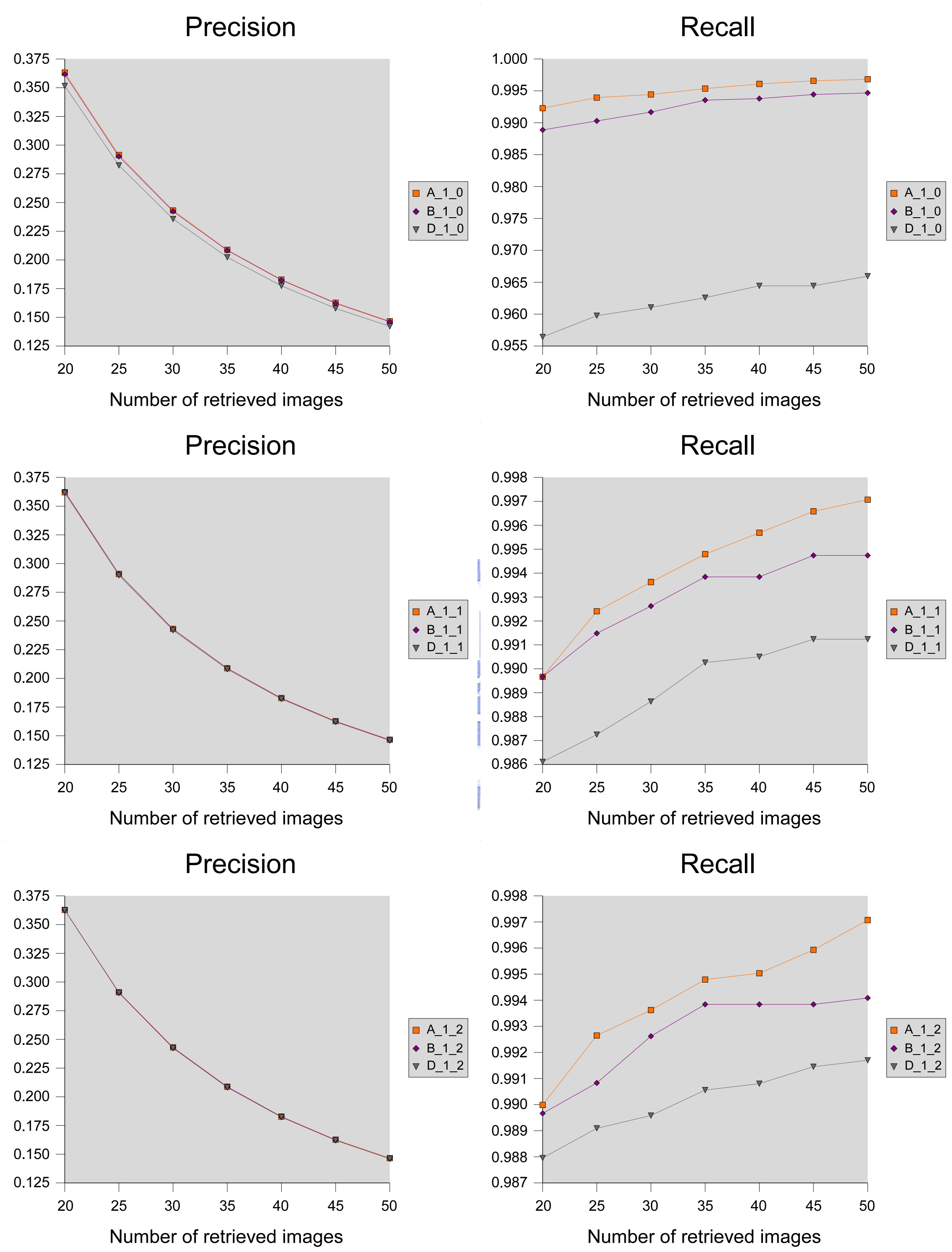


Figure 3.14: Precision and recall curves of positive-and-negative query of input/query=1. The curves are labeled as [scheme]_1_[pseudo/input].

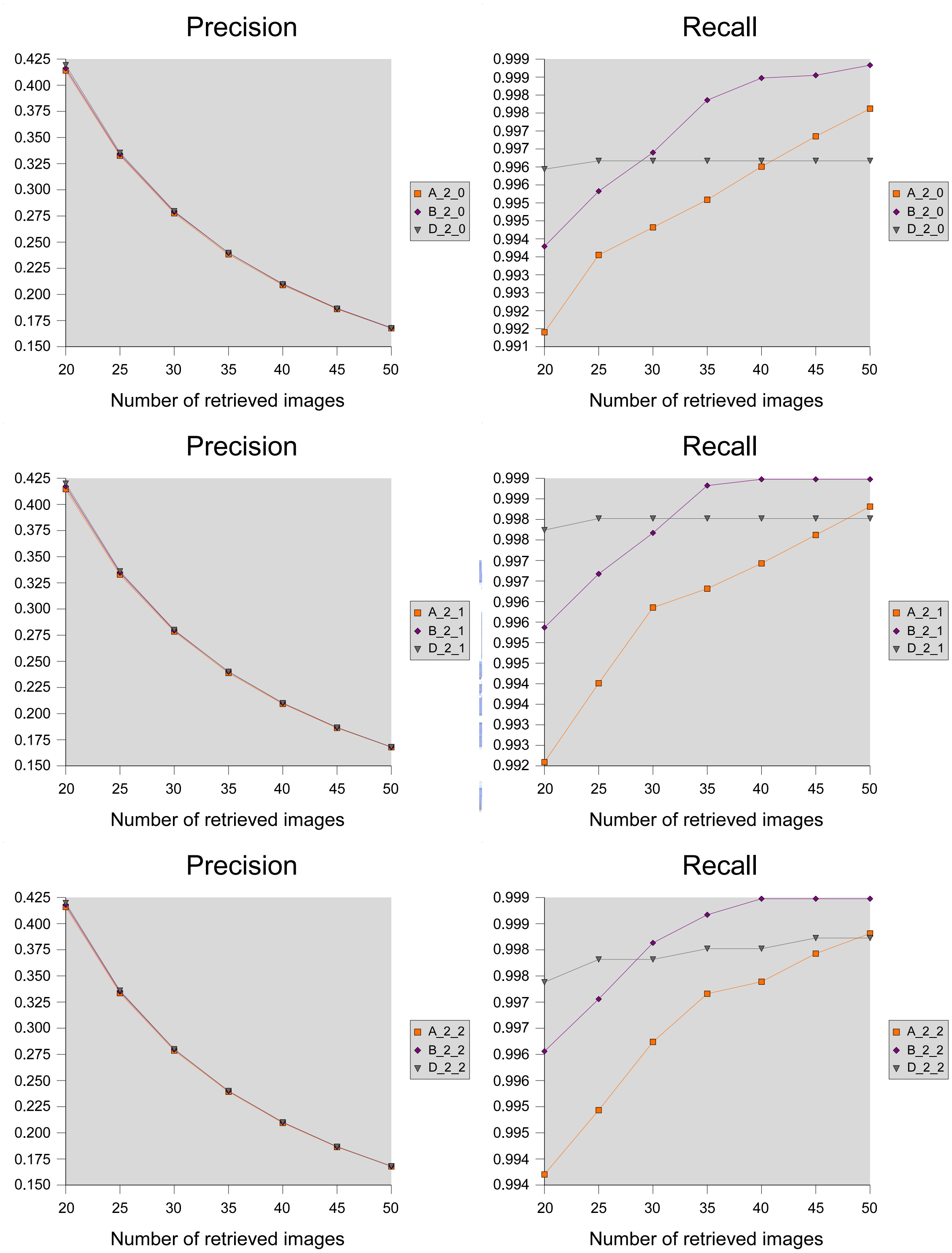


Figure 3.15: Precision and recall curves of positive-and-negative query of input/query=2. The curves are labeled as [scheme]_2_[pseudo/input].

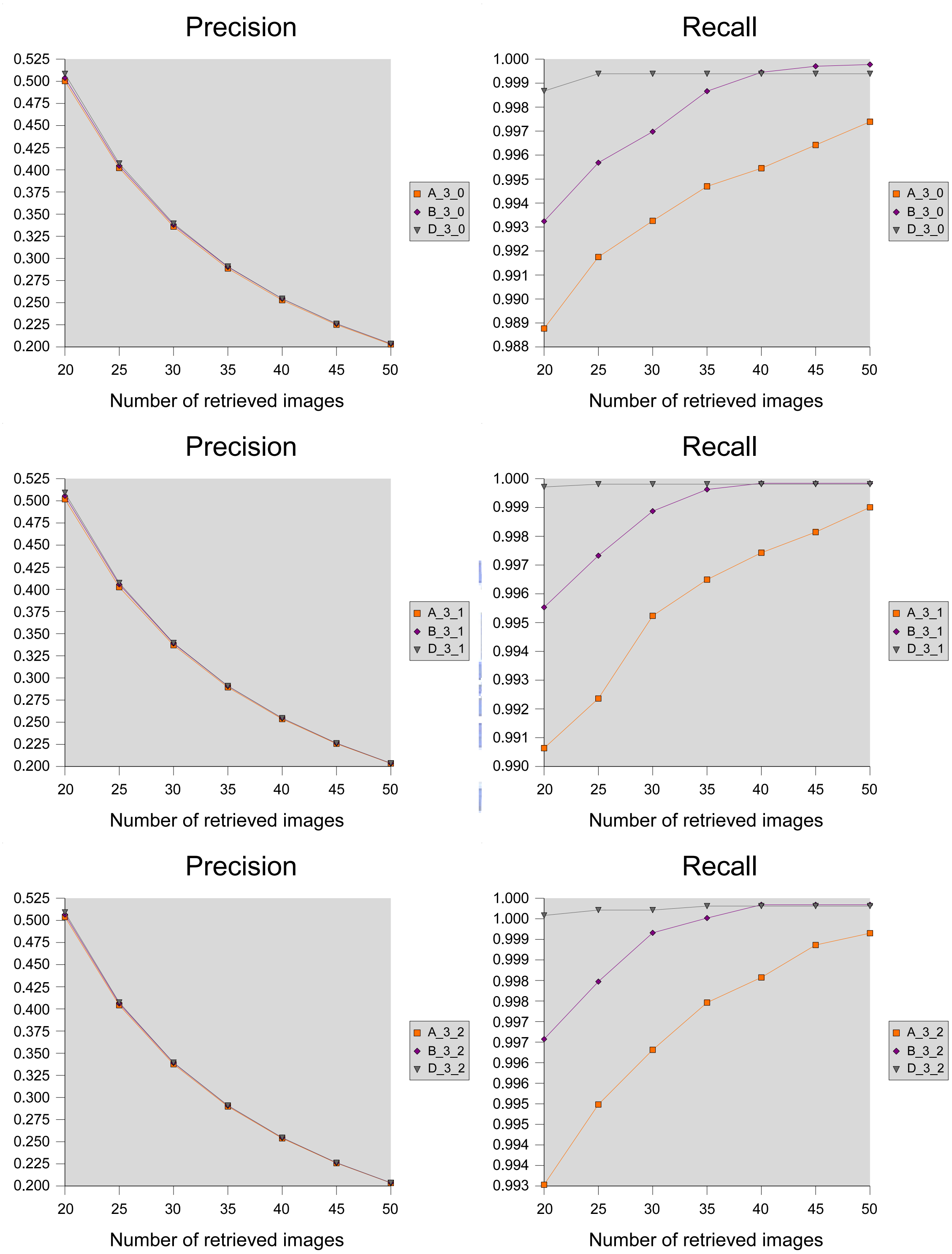


Figure 3.16: Precision and recall curves of positive-and-negative query of input/query=3. The curves are labeled as [scheme]_3_[pseudo/input].

3.10 Summary

Based on the above simulation results, we briefly summarize our observations below.

- Distance-based weight estimation outperforms when multiple input instances are available.

- Pseudo images improve query accuracy in many cases, especially when our method is used with SNR scalability.

- Experiments show that one pseudo image per input image gives significant performance boost in most cases.

- Negative instances used as a pruning criterion produce better results than those used as negative samples in weight calculation.

- When input instances are few, negative feedback may even degrade the performance of the MARS-like and the Gaussian-normalized schemes.

- When sufficient input instances are available, the Gaussian normalized feature distance does not provide as precise estimation as our method.

- The SNR multi-scaled pseudo images provide better ANMRR values; they also lead to more consistent improvements in accuracy.

An overall comment about the performance of our scheme is as follows:

- When only one input image is available, our scheme looses about 0.8 in $log(ANMRR)$. However, with the assistance of pseudo images, the gap shrinks to about 0.25.

- In the case of two input images, our scheme improves. Without pseudo images, ours looses about 0.2; with pseudo images, ours may win or loose in the average of 0.05.

- When we have three input images, our scheme wins. The figure in $log(ANMRR)$ is about 0.3 to 0.5.

- From the above results, we can summarize that Scheme C is good for sufficient query images. For small-sample cases, though not as good as other schemes, it produces comparable accuracy by including pseudo images.