Table 1: **Left**: SUN gZSL results, **Middle**: ImageNet results, **Right**: Imagenet results with different unknown K.

	$accy_u$	$accy_s$	H
ALE[1]	21.8	33.1	26.3
VCL	10.4	63.4	17.9
CDVSc	27.8	63.2	38.6
BMVSc	29.9	62.9	40.6
WDVSc	**30.5**	**63.1**	**41.1**

Test data	Method	Hit@K (%) 1	2	5	10	20
ILSVRC2010	**DIPL**	–	–	31.7	–	–
2-hop	CONSE	8.3	12.9	21.8	30.9	41.7
	SYNC	10.5	16.7	28.6	40.1	52.0
	EXEM	12.5	19.5	32.3	43.7	55.2
2-hop	**VCL**	12.3	19.3	31.3	40.3	48.7
	WDVSc	**17.6**	**26.7**	**38.8**	**47.5**	**57.9**

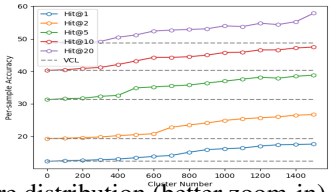
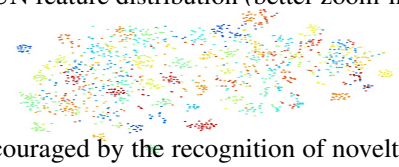


Table 2: **Left**: AUSUC evaluation, **Middle**:Center distances, **Right**: SUN feature distribution (better zoom-in).

	AwA2	CUB	SUN
VCL	0.47	0.31	0.20
CDVSc	0.74	0.55	0.35
BMVSc	0.76	0.51	0.37
WDVSc	**0.79**	**0.61**	**0.38**

	AwA2	CUB	SUN	ImageNet
VCL	0.07	0.07	0.04	0.13
CDVSc	0.04	0.04	0.03	–
BMVSc	0.03	0.04	0.02	–
WDVSc	**0.01**	**0.03**	**0.02**	**0.05**



We thank all the reviewers for their valuable comments. We are very encouraged by the recognition of novelty and performance boost from R1, R2. For each question from reviewers, we give strong experiment results and clarifications. Missing references, minor errors, and rephrasing introduction (R1) will be fixed in the revision because of page limit.

R1:Clarification of step 3 in Sec 3.5. Because some unrelated images will make the approximated centers deviate from the real centers. Therefore, we regard images whose distance to the original approximated centers is below one threshold as reliable ones (denote as "a new target domain") and get new better-approximated centers based on them.

R1:SUN results. Due to limited space, SUN results are not given in the original submission, which is shown in the left of Tab.1 now. Note that ALE is the best among SOTA methods on this dataset, but is still far behind our results.

R1: Comparison with [29] in Table 3. Will add it. And our harmonic means of gZSL on AwA2,CUB,and SUN (**76.4, 57.5, 41.1**) are all better than [29] (59.6, 49.7, 39.4) except a little lower unseen accuracy (43.3) on CUB.

R2:ImageNet results. WDVSc Results are shown in the middle of Tab.1, which outperform previous SOTA and baseline VCL by more than 5 points. Even without knowing K value (Right), our results consistently improve over the VCL. This further demonstrates the superiority of the proposed visual constraints. Due to limited computation resources in such a short period, the results of BMVSc and CDVSc are not reported here but will be included in the final version.

R2:Comparison with [A,18,31,34]. In supp Fig.2, we have provided the convergence comparison with [31]. On AwA1 and CUB dataset, our result (**96.2, 74.2**) is much better than [31]'s result (86.7, 58.3). Because [34] has not reported other results nor released their code, comparison with [34] is only given in the Tab.1 of our original submission. Since we only use naive nearest neighbor based label assignment, our superiority only comes from better learned projection function with the proposed visual structure constraints. For DCN[18], our results on AwA1,CUB,SUN (SS: **96.2, 74.2, 67.8, PS: 87.3, 73.4, 63.4**) are better than theirs (SS: 82.3, 55.6, 67.4, PS:65.2, 56.2, 61.8). For DIPL[A], our results on AwA1,CUB, SUN (SS: **96.2, 74.2**, 67.8, PS: **87.3, 73.4**, 63.4, gZSL(no SUN): **81.8, 57.5**, Avg: **75.2**) are also overall better than their results (SS: 96.1, 68.2, **70.0**, PS: 85.6, 65.4, **67.9**, gZSL: 75.6, 43.2, Avg: 71.5) except the SUN dataset.

R2:Submission checklist e.g."error bars". Sorry, we check it "yes" in the system by mistake and will change it.

R3:Key differences with [34]. Though our motivation is superficially similar, the key ideas are definitely different and complementary. ZSL methods often have two steps: projection function learning and label assignment. [34] is to improve label assignment over naive NN using a fixed project function while we aim to learn better projection function by only using naive NN assignment. Our better results also verified our key idea(i.e., the proposed projection learning objective). Besides, rather than using hard matching in previous methods including [34], our WDVSc is the first to use soft matching (Line 199-202) with probability, which brings the extra gain of WDVSc over CDVSc and BMVSc.

R3:Key differences with [A]. [A] is an Arxiv paper and not published yet. After reading it, we find it is just a special case (single direction version) of our CDVSc. On AwA1, CUB and SUN, its performance (88.64,58.8,86,16) is worse than our CDVSc (89.6, 69.9, 90.6) let alone WDVSc(**92.9, 71.0, 91.2**).

R3:Dependence on discriminative clusters and known cluster number K? For fair comparison in the traditional setting, our method shows that it can handle indiscriminative clusters and unknown K. These are already explained in **Line 303 -317 and Tab.6** very clearly. On fine-grained datasets CUB and SUN, we have provided their feature distribution in the supplemental materials and the right of Tab.2 respectively. Though their clusters are not perfectly separable, our method still achieves consistent performance gain. This is also validated on the large-scale ImageNet dataset, which has more than 1500 unseen classes. On the right of Tab.1, the per-sample results of different K (guessed values) are also provided. For Tab.4, it is the experiment results of the new setting where many noisy and unrelated images are manually added. In this new setting, most existing methods leveraging unseen center priors will fail, which also demonstrate the importance of this setting. By using the proposed simple strategy, our method works well again.

R3:Evaluation with AUSUC. Our AUSUC results are shown on the left of Tab.2, which are much better than the best-reported results EXEM [C](AwA2:0.559, CUB:0.366, SUN:0.251) by a very large margin.

R3:Quantitative evaluation of domain shift. We have calculated the distances between the projected and the real centers in the middle table of Tab.2. By using the proposed visual structure constraints, the distances are reduced significantly on all the datasets including ImageNet, which indicates the domain shift problem is improved quantitatively.

R3: New setting is ad-hoc and its evaluation. It is indeed a very common and important setting for real industry applications but never studied before. Quantitative evaluation is already given in Tab.4 of the original submission.